LECTURE NOTES IN ELECTRICAL ENGINEERING 56

Qi Luo (Ed.)

Advancing Computing, Communication, Control and Management

Springer

Lecture Notes in Electrical Engineering

Volume 56

Qi Luo (Ed.)

Advancing Computing, Communication, Control and Management

Springer

Qi Luo
School of Electrical Engineering
Wuhan Institute of Technology
Wuhan 430070
China
E-mail: ccnu_luo2008@yahoo.com.cn

ISBN 978-3-642-05172-2 e-ISBN 978-3-642-05173-9

DOI 10.1007/978-3-642-05173-9

Lecture Notes in Electrical Engineering ISSN 1876-1100

Library of Congress Control Number: 2009940122

Typeset & Cover Design: Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed in acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

A large 2008 ISECS International Colloquium on Computing, Communication, Control, and Management (CCCM 2008), was held in Guangzhou, August 2008, China. Just like the name of the Colloquium, the theme for this conference is Advancing Computing, Communication, Control, and Management Technologies. 2008 ISECS International Colloquium on Computing, Communication, Control, and Management is co-sponsored by Guangdong University of Business Studies, China, Peoples' Friendship University of Russia, Russia, Central South University, China, Southwestern University of Finance & Economics, China, and University of Amsterdam, Netherlands. It is also co-sponsored IEEE Technology Management Council, IEEE Computer Society, and Intelligent Information Technology Application Research Institute. Much work went into preparing a program of high quality. We received about 972 submissions. Every paper was reviewed by 3 program committee members, about 382 were selected as regular papers, representing a 39% acceptance rate for regular papers. The CCCM conferences serve as good platforms for the engineering community to meet with each other and to exchange ideas. The conference has also stroke a balance between theoretical and application development. The conference committees have been formed with over two hundred committee members who are mainly research center heads, faculty deans, department heads, professors, and research scientists from over 30 countries. The conferences are truly international meetings with a high level of participation from many countries. The response that we have received for the congress is excellent.

This volume contains revised and extended research articles written by prominent researchers participating in the conference. Topics covered include intelligent computing, network management, wireless networks, telecommunication, power engineering, control engineering, Signal and Image Processing, Machine Learning, Control Systems and Applications, The book will offer the states of arts of tremendous advances in Computing, Communication, Control, and Management and also serve as an excellent reference work for researchers and graduate students working on Computing, Communication, Control, and Management Research.

Qi Luo

Table of Contents

Study on MRF and REF to Semi-supervised Classification

Liang Jun, Xianyi Cheng, and Xiaobo Chen

School of Computer Science and Telecommunication Engineering,
Jiangsu University, China, 212013
liangjun@ujs.edu.cn

Abstract. We study the semi-supervised classifier with a decision rule learning from labeled and unlabeled data. A model to semi-supervised classification is proposed to overcome the problem induced by mislabeled samples. A new energy function based on REF (robust error function) is used in MRF (Markov Random Field). Also two algorithms based on iterative condition mode and Markov chain Monte Carlo respectively are designed to infer the label of both labeled and unlabeled samples. Our experiments demonstrate that the proposed method is efficient for real-world dataset.

Keywords: semi-supervised learning, classifier, Markov Random Field, Simulating.

1 Introduction

Semi-supervised learning has received considerable attention in the machine learning literature due to its potential in reducing the need for expensive labeled data [1]. Given a sample set $Y = \{y_1, y_2, ..., y_l, y_{l+1}, ..., y_N\} \subset R^m$ and a label set $L = \{1, 2, ..., C\}$, the first l samples $y_i (1 \le i \le l)$ are labeled as $f_i \in L$ and the remaining samples $y_i (l+2 \le i \le N)$ are unlabeled. The goal is to classify the unlabeled samples to its latent class.

Most formulations of semi-supervised learning approach the problem from one of the two ends of the unsupervised-supervised spectrum: either supervised learning in the presence of unlabelled data [2,3] or unsupervised learning with additional information [4,5].

Almost all current research has not considered the situation when there exist mislabeled samples in semi-supervised classification. Because of the little number of labeled samples in semi-supervised classification, the mislabeled samples will have severe influence on the final classification result. An example is show in figure 1 which contains 70 samples from two classes. One class is show in blue and the other in red. There are eight labeled samples in the data set which are show in green. Three of them are from class 1, four of them are from class 2 and especially one sample from class 1 is mislabeled to class 2 (show by the green diamond on top most).

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 1–6.
springerlink.com

According to the algorithm in literature [6], we can see that many samples in class 1 are misclassified to class 2 because of the existence of mislabeled sample.

In order to tackle the problem, we proposed a new model based on MRF and REF. The paper is organized as follows. In section 2, we proposed a MRF with REF. In Sect 3, we describe two different inference algorithm for the model and Section 4 reports the experiments on the above examples and its application in semi-supervised classification. Some conclusions are given in section 5.

2 MRF with REF

2.1 MRF for Semi-supervised Classification

We first recall MRF (MRF) models [7] which originated in statistical physics and have been extensively used in image processing. MRF constraints the spatial smoothness though puts high probability to the class labels which are consistent locally. In the context of machine learning, what we can do is create a graph with a node for each sample, and with undirected edges between them that are similar to each other. A typical MRF model is show in figure 1. The nodes in the upper level are that we can observe while the nodes in the lower level are unobserved. Because of its flexibility, MRF have been introduced into semi-supervised learning [8].In the context of semi-supervised classification, some hidden labels are known as prior which differentiate it from the conventional MRF as illustrated in figure 2. A natural energy function takes the form:

$$E_g(x) = -\sum_{\{i,j\}} \omega_{ij} x_i x_j \tag{1}$$

where $x_i \in \{-1,+1\}$ are binary labels and ω_{ij} is the weight on edge $\{i, j\}$, which is a measure of the similarity between the samples.

2.2 REF

Recently, many researchers have found that the use of REF can lead to improved results on applications such as super-resolution, CCD demosaicing, image in-painting, and image denoising. Examples of REFs include the truncated quadratic function and the Lorentzian error function [9] which is

$$E_l(x) = \ln(1 + \frac{1}{2}x^2) \tag{2}$$

The most important feature of REF is that unlike the quadratic error function, the magnitude of the derivative of the error function does not increase as the error rises. This feature is show in figure 3 which we can see the rate of increase in the Lorentzian error function actually decreases as the error rises. It is this property that makes this function robust to large outlier errors.

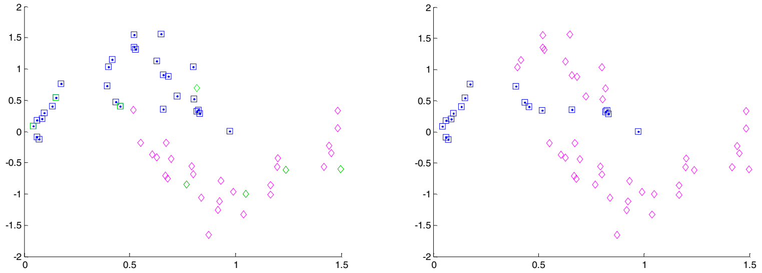


Fig. 1. Classification results according to the semi-supervised method of [6]

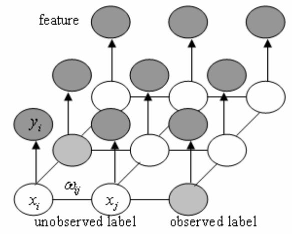


Fig. 2. MRF

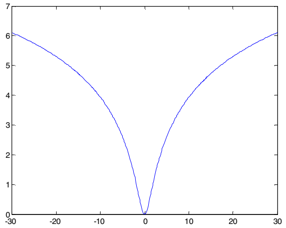


Fig. 3. Lorentzian error function

2.3 MRF with REF

Due to the analysis above, we incorporate the global energy function of MRF which means a good classifying function should not change too much between nearby samples with the local Lorentzian error function which means a good classifying function should not change too much from the initial label assignment into the follow energy function:

$$E(x)=\lambda E_g(x)+(1-\lambda)E_l(x)=\lambda\sum_{\{i,j\}}\omega_{ij}(x_i-x_j)^2+(1-\lambda)\sum_{1\le i\le l}\log(1+\frac{1}{2}(x_i-f_i)^2) \tag{3}$$

where $\omega_{ij}=\exp(-\|y_i-y_j\|^2/2\sigma^2)$, which is a measures of the similarity between samples. λ is a parameter which balance the global consistency and local consistency. To assign a probability distribution to the labels, we form a random field:

$$p(x)=\frac{1}{Z}\exp\{-\beta E(x)\} \tag{4}$$

where the partition function Z normalizes over all labels.

So our goal is to maximize $p(x)$ which is equivalent to minimize $E(x)$ as below:

$$x=\arg\min_{x} E(x) \tag{5}$$

Solving for the lowest energy configuration in this MRF will produce a partition of the entire (labeled and unlabeled) samples that maximally optimizes self-consistency, subject to the constraint that the configuration must agree with the labeled data.

3 Algorithm

Application of the model above requires a method to minimize the energy function $E(x)$ .We will suggest and implement two different inference schemes. One is based on iterated conditional modes or ICM [10], the other is based on Markov Chain Monte Carlo or MCMC sampling method [11]. The former has been widely used in machine learning community because of its easy to implement but often leads to a local optimized solution. The latter can generate exact and global optimal solution but is computationally demanding. We will see the difference between them in the experiment section.

3.1 ICM

The ICM based algorithm is as follows:

(1) Initialize

Assign the variables $\{x_i\}$ to initial values, which we do by simply setting x_i to a selected class randomly. Also we set λ to 0.05 and σ to 0.45.

(2) Repeat

We take one node x_j at a time and evaluate the total energy $E(x)$ for the two possible states $x_j = +1$ and $x_j = -1$, keeping all other node variables fixed, and set x_j to whichever state has the lower energy. The node is updated by choosing at random. Repeat the procedure until certain convergence criteria is satisfied and then go to the following process.

(3) Finally

Output the final class label x_i for each node in the MRF.

3.2 MCMC

The MCMC based algorithm is as follows:

(1) Initialize

The same as the step in ICM based algorithm above besides a temperature parameter T is set to 1.

(2) Repeat

Sample a random variable U from uniform distribution $U[0,1]$.

Sample a new label x^* based on the current label x according to a suitable proposal distribution $q(x^* | x)$ .In our current configuration, $q(x^* | x)$ adopt Gibbs sampler.

We accept x^* as our new label if it satisfies the following condition, otherwise we drop x^* and keep current label x unchanged.

$$U < \min\{1, \frac{E(x^*)^{\frac{1}{T}} q(x \mid x^*)}{E(x)^{\frac{1}{T}} q(x^* \mid x)}\} = \min\{1, \exp(\frac{E(x^*) - E(x)}{T}) \frac{q(x \mid x^*)}{q(x^* \mid x)}\} \tag{6}$$

Update temperature T according to a chosen cooling schedule.

Repeat the above process until certain convergence criteria is satisfied and then go to the following process.

(3) Finally
The same as the step in ICM based algorithm above.

4 Experiments

The proposed method is first test on the examples show in section 1 which we know the true classification. The above two algorithm are used to classify the samples and we compare both them and the literate [6] (referred as LGC). The results are show in figure 4. The decrease curve of energy is show in figure 5 from which we can see the MCMC based method can read lower energy then ICM but need more computational resource. Finally the misclassification rate is show in table 1.

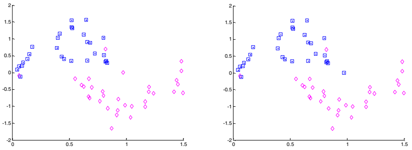


Fig. 4. Classification results

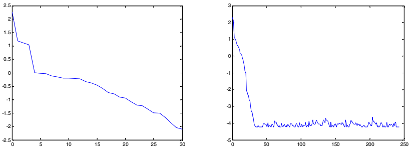


Fig. 5. Energy curve

Table 1. Misclassification rate

algorithm	misclassification rate (%)
LGC	21.6
ICM	5.0
MCMC	1.6

We have also illustrated the classification error rate when different percentage mislabeled samples are appeared in data set as in figure 6.

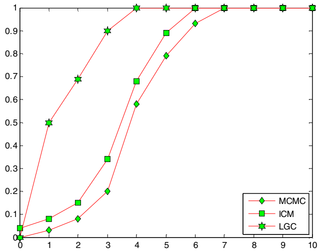


Fig. 6. Classification error rate

5 Conclusion

In this work we have introduced MRF that combine REF for semi-supervised classification of samples when there exist mislabeled data. The main idea is to introduce REF with the energy function of MRF to reduce the effect induced by mislabeled samples. Two inferring algorithms are given to infer the labels both of the labeled and unlabeled samples.

References

1. Zhu, X.: Semi-Supervised Learning Literature Survey. Technical Report, Computer Sciences Department, University of Wisconsin, Madison (2006)
2. Belkin, M., Niyogi, P.: Using Manifold Structure for Partially Labeled Classification. In: Proc. Neural Information Processing Systems, vol. 15. MIT Press, Cambridge (2003)
3. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised Learning Using Gaussian Fields and Harmonic Functions. In: Proc. ICML 2003, Washington DC (2003)
4. Law, M., Topchy, A., Jain, A.K.: Model-based Clustering With Probabilistic Constraints. In: Proc. SIAM Conference on Data Mining, Newport Beach (2005)
5. Lu, Z., Leen, T.: Probabilistic Penalized Clustering. In: Proc. NIPS, vol. 17. MIT Press, Cambridge (2005)
6. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Scholkopf, B.: Learning with local and global consistency. In: NIPS (2003)
7. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6(6), 721–741 (1984)
8. Bekkerman, R., Sahami, M.: Semi-supervised clustering using combinatorial MRFs. In: ICML 2006 Workshop on Learning in Structured Output Spaces (2006)
9. Black, M.J., Rangarajan, A.: On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. International Journal of Computer Vision 19(1), 57–92 (1996)
10. Kittler, J., Foglein, J.: Contextual classification of multispectral pixel data. Image and Vision Computing 2, 13–29 (1984)
11. Andrieu, C., de Freitas, N., Doucet, A., Jordan, M.I.: An Introduction to MCMC for Machine Learning. Machine Learning (2003)

An Extension Method of Space Syntax and Application

Xinqi Zheng[1,2], Lu Zhao[1], Yanjun Su[1], Guang Yan[1], and Shuqing Wang[1]

[1] School of Land Science and Technology, China University of Geosciences, Beijing, China
[2] Land Resources Information Development Research Laboratory, Beijing, China

Abstract. Space syntax has remarkable advantage in predicting the accessibility of urban road system, and has been applied in many other fields. However, as it is just the abstract of the reality, space syntax doesn't take the impact of road width on the accessibility into account, and there are many limitations in its practical application. This study discussed the relationship both between the total depth and width of the road, and between road width and integration degree by extending the space syntax and the calculation of integration degree. Through the case studies on 9 cities in China and other countries, we found out that the total depth is negatively correlated with the logarithm of road width. Moreover, the axis graph obtained with the extended formula performed better in displaying details than the one obtained with the original one. In addition, the empirical formula for space syntax extension has also been summarized. It is concluded that the results calculated with the extended methods are more accordant with the actual urban conditions and this study can provide a new approach to extend and apply the space syntax.

Keywords: space syntax, extension, road width, empirical formula, cities.

1 Introduction

At the end of 1970s, space syntax was first put forward and applied by Bill Hillier and his team from Bartlett College of the University of London [1-3]. Since then, many scholars have made a number of extension studies on space syntax [4-8], such as the studies on urban traffic [9-11], the characteristics of urban street layout [12-15], urban planning [16], and so on. As the development in the past three decades, the research on space syntax has been carried out in different study area . In 1997, the first International Space syntax Seminar was held in London. And the second and the third seminars were held in Brasilia (1999) and Atlanta (2001), respectively. The fourth seminar returned to London in 2003, and Defland in Holland and Turkey held the fifth and the sixth seminars in 2005 and 2007, respectively. By now, the focus of the research on space syntax abroad has gradually shifted from empirical research to the innovation in the theory and method. In contrast, domestic research and studies are largely focused on empirical research but rarely on the theory and method.

Despite its remarkable advantage in the application in urban system, such as predicting the accessibility, space syntax was found through studies to bear an inherent defect when applied in appraising the urban road network. It ignored the effects of road width on the accessibility. That is to say, the impact of road width was not considered

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 7–14.
springerlink.com

when calculating the integration degree. Besides, it is known commonly that the width of road affects the traffic condition significantly. Therefore, the application of space syntax in urban morphology and urban planning, etc. will be affected if the road width is not taken into account in space partition. Though the integration degree obtained with space syntax can well describe the distribution of traffic flow, it is inadequate to reflect the smoothness of traffic on roads. This study tried to add the road-width parameter into integration degree with analysis of great amounts of urban data and summarized the corresponding empirical formula. It is proved that the results of extended space syntax are more accordant with the actual condition.

2 Extension Method of Space Syntax

2.1 Taking Road Width into Account

We analyzed the basic principle of space syntax in order to eventually obtain the relationship between road width and integration degree. According to the principle, the depth value refers to the shortest distance between one node and the others. The distance mentioned is not the actual distance, but the step number. It is defined in space syntax that the step number between two adjacent nodes is one step. The depth value is not an independent morphological variable, but an intermediate variable for calculating the integration degree [17-18].

Suppose that d_{ij} is the shortest distance between any two points i and j in the connection graph, then the total depth D can be calculated with formula (1), where n is the total node number in a connection graph.

$$D = \sum_{j=1}^{n} d_{ij} \tag{1}$$

And the average depth can be calculated with formula (2).

$$MD_i = \frac{D}{n-1} \tag{2}$$

Integration degree refers to the convergent or discrete degree of one space and the others in the system. It is denoted by RA_i and RA_i, which can be calculated with formulas (3) and (4), where D_n ($D_n = 2\{n[\log_2((n+2)/3-1)+1]\}/[(n-1)(n-2)]$) is the standardized integration degree.

$$RA_i = \frac{2(MD_i - 1)}{n-2} \tag{3}$$

$$RRA_i = \frac{RA_i}{D_n} \tag{4}$$

It can be seen in formulas (1) and (2) that we can't obtain the relationship between road width and integration degree directly, for integration degree is related not merely to road width but also to the total node number and the average depth of the urban road network. In order to obtain the relationship between road width and integration degree, we researched the relationship between road width and the total node number of the road network or the relationship between road width and average depth at the first step. However, the total node number in the road network of a city bears no direct relation with road width. Therefore, we have to start with the relationship between road width and average depth. But it can be seen in formula (1) that the average depth is related to two variables, namely the total depth value of roads and the total node number of the road network. So we conducted the analysis on the relationship between the total depth value of roads and road width firstly to find out the effects of road width on integration degree. Suppose W represents road width and ε represents relation coefficient, then they may have the following functional relation with the total depth value D (see formula (5)).

$$D = f(W, \varepsilon) \tag{5}$$

2.2 Extending Integration Degree

In formula (3), different impact factors actually act invariably as a weight factor in the calculation of road integration degree. Therefore, we can consider the impact factor of road width in the total depth value when calculating the integration degree by multiplying the original total depth value with the weight factor p which reflects road width. According to formula (2), we can see that the above improvement means that it is equal to multiply the average depth with p . The improved total average depth can be calculated with formula (6). Then we can obtain the corresponding calculation formula for the improved road integration degree and the integral accessibility.

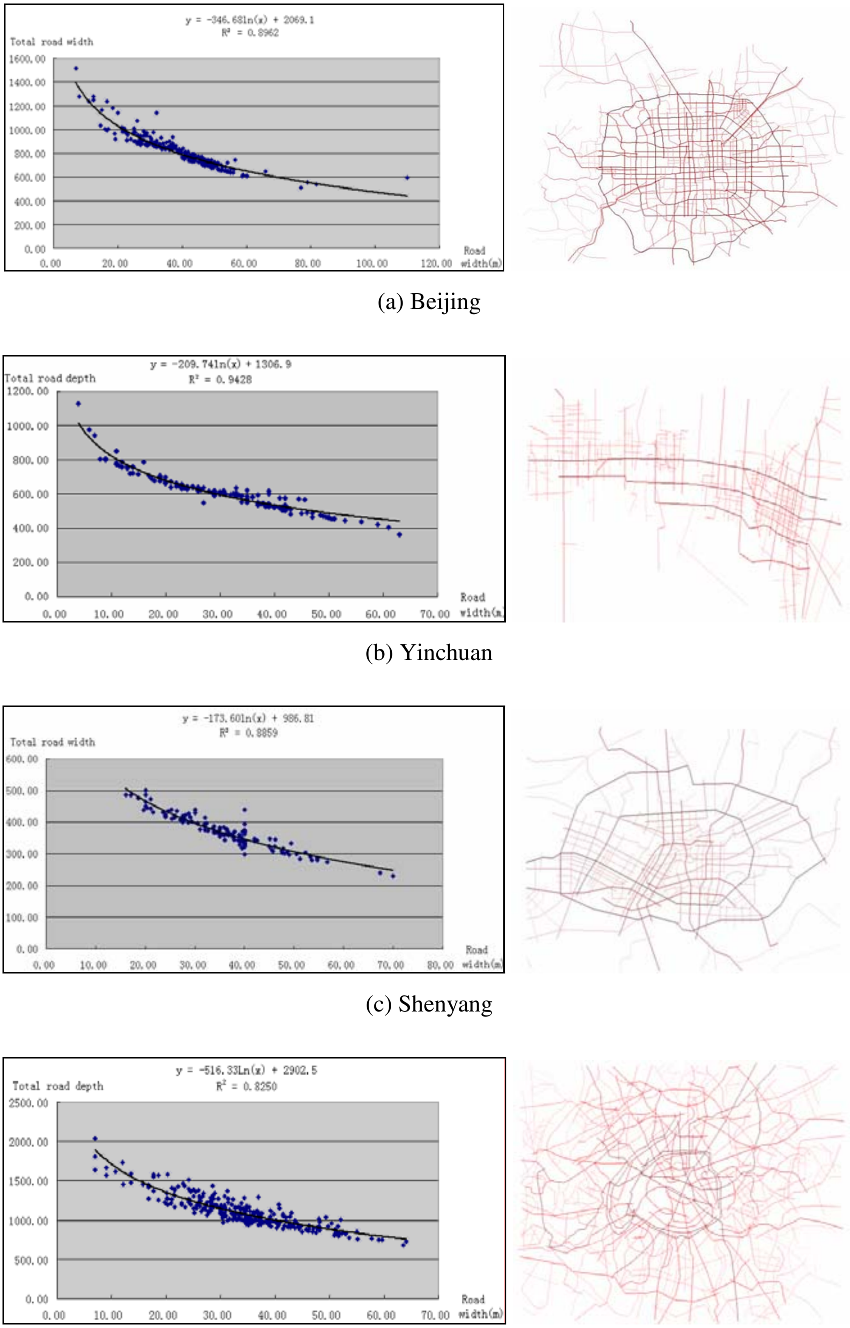
$$MD_i = \frac{f(W, \varepsilon)}{n-1} \tag{6}$$

On the basis of the above formula, we can obtain the calculation parameters for the extended space syntax reflecting road width through the statistical analysis on the data from typical cities in next step.

3 Extension Method of Space Syntax

3.1 Data Selection and Processing

We totally selected nine typical cities in China as well as other countries to comprehend the impact of road width on integration degree. They are Beijing, Shanghai, Tianjin, Shenyang, Zhuhai, Shaoxing, Yinchuan, Paris and Sydney. There are municipalities, provincial capitals and prefecture-level cities distributing in different areas of China. And the foreign cities are all international ones.



(a) Beijing

(b) Yinchuan

(c) Shenyang

(d) Paris

Fig. 1. Relationship between total road depth and road width and the road axis graph

The data of the cities came from official electronic map in *.jpg format on the Internet. With the aid of secondary development language—Avenue, we modified the Axwoman plug-in on the platform of ArcView 3.3 [19]. During the process of vectorization, the road-width attribute value for each axis was presumed. Through calculation of the vector axis graph, we obtained the value of each morphological variable and stored it in the attribute table which is in *.dbf format.

3.2 Data Calculation

Then we calculated the relevant data from the nine cities by the extended Axwoman plug-in on the platform of ArcView 3.3. The results of Beijing, Yinchuan, Shenyang and Paris are showed in figure 1, respectively. Then we imported the attribute data in *.dbf into Excel for the next regression analysis.

4 Results and Discussion

4.1 Negative Correlation between Total Depth and Logarithm of Road Width

We adopted the least square method and multi-functions to analyze the relationship between road width and total depth. Through mass analysis, we noticed that the total depth is negatively correlated with the logarithm of road width. This latent functional relation can be described by formula (7), where D is the total depth and W is the road width while a and b are the coefficients of the function.

$$D = -a \times \ln(W) + b \tag{7}$$

The calculation results of the nine cities are shown in table 1. It can be seen from table 1 that there is close correlation between the total road depth and the road width, and the correlation coefficient R^2 in each case study is more than 0.80.

Table 1. Functional relation and R^2 between road width and total depth for the nine cities

City Name	Functional Relation	R^2
Beijing	$D = -346.68 \times \ln(W) + 2069.1$	0.8962
Shanghai	$D = -311.45 \times \ln(W) + 1818.5$	0.8581
Tianjin	$D = -314.24 \times \ln(W) + 1893.0$	0.8574
Shenyang	$D = -173.60 \times \ln(W) + 986.81$	0.8859
Yinchuan	$D = -209.74 \times \ln(W) + 1306.9$	0.9428
Zhuhai	$D = -446.57 \times \ln(W) + 3010.0$	0.8153
Shaoxing	$D = -158.61 \times \ln(W) + 1021.9$	0.8754
Paris	$D = -516.33 \times \ln(W) + 2902.5$	0.8250
Sydney	$D = -349.36 \times \ln(W) + 2200.6$	0.8034

4.2 Comparison of Integration Degrees before and after Improvement

Compared with the original formula for integration calculation, the new one took the impact of road width into account. To verify the feasibility of the results, we took Beijing for example and conducted the calculation with the original formula and the improved one separately, and obtained their corresponding road axis graphs and their attribute tables (see figure 2 and 3).

It can be seen from the comparison that the axis graph obtained with the new formula performed better in displaying details. Meanwhile, the integration degree of urban roads grows higher than the original value, while that of roads in suburban area becomes much lower. This is more accordant with our general thinking and actual conditions, and enlarges of the contrast between them so that the details are more distinct.

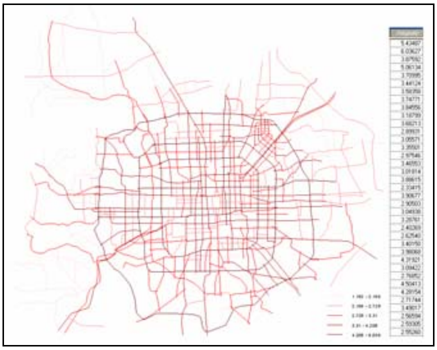


Fig. 2. Road axis graph and its attribute table in Beijing before integration improvement

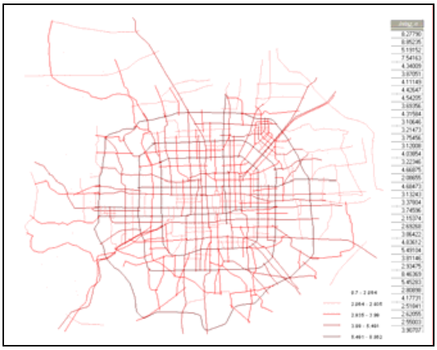


Fig. 3. Road axis graph and its attribute table in Beijing after integration improvement

5 Conclusion

Based on the above discussions and analysis, we found out that the space syntax becomes more reasonable than the original one after the road width is considered.

Moreover, the integration degree can illustrate road details better with the impact factor of road width being taken into account.

However, because the geographic maps of the ten cities are in different detail levels, it is unavoidable to neglect some roads when vectorization was to be conducted. In addition, we ignored the elements in very small values when deducing how to convert the impact factor of width for calculating integration degree, in order to determine the weight function. However, these defects won't exert too much influence on the final conclusion.

Acknowledgments

It is a project supported by the National Social Science Foundation of China under Grant (No. 07BZZ015), the National Key Technology R&D Program of China (No.2006BAB15B03 and 2006BAB15B06).

References

1. Hiller, B., Hanson, J.: The Social Logic of Space. Cambridge University Press, Cambridge (1994)
2. Hong, Z., Xinsheng, W., Ruilin, Y.: Space syntax and its research progress. Geospatial Information 4(4), 37–39 (2006)
3. Yu, Z., Jianguo, W.: A review on space syntax. Architect (3), 33–44 (2004)
4. Cardillo, Scellato, S., Latora, V., Porta, S.: Structural properties of planar graphs of urban street patterns. Physical Review 73, 66107 (2006)
5. Chen, Lu, F.: Analysis of spatial configuration of the palace museum: an application of axial-based space syntax. Geoinformatics, X4200–X4200 (2006)
6. Hiller, B.: The hidden geometry of deformed grids: or, why space syntax works, when it looked as though it shouldn't. Environment and Planning B-Planning & Design 26(2), 1–191 (1999)
7. Hiller, B.: Space and spatiality: what the built environment needs from social theory. Building Research and Information 36(3), 216–230 (2008)
8. Young, K.: Linking the spatial syntax of cognitive maps to the spatial syntax of the environment. Environment and Behavior 36(4), 483–504 (2004)
9. Ben, Alan, P., Hiller, B.: Spatial distribution of urban pollution: civilizing urban traffic. The Science of the Total Environment 189/190, 3–9 (1996)
10. Changxiu, C., Wenchang, Z., Jie, C.: Evaluating the accessibility about Beijing's subways in 2008 based on spatial syntax. Geo-Information Science 9(6), 31–35 (2007)
11. Xinqi, Z., Lu, Z., Meichen, F., Shuqing, W.: Extension and application of space syntax- a case study of urban traffic network optimizing in Beijing. In: 2008 Workshop on Power Electronics and Intelligent Transportation System, pp. 291–295 (2008)
12. Ruilan, D., Xinqi, Z.: The relation of the city road structure and the land price based on the space syntax. Science of Surveying and Mapping 29(5), 76–79 (2004)
13. Jiang, B., Claramunt, C.: A comparison study on space syntax as a computer model of space. In: Proceedings of Second International Symposium on Space Syntax. Brazil university, Brasilia (1999)
14. Bin, J., Claramunt, C.: Extending space syntax towards an alternative model of space within GIS. In: Proceedings of 3rd Symposium on Space Syntax, Atlanta, pp. 1–6 (2001)

15. Haijun, L., Xuejun, L., Jianquan, C., Ningrui, D.: Study on the accessibility of urban road network based on space syntax. China Water Transport 7(7), 131–133 (2007)
16. Jin, D., Hiller, B.: Space Syntax and Urban Planning. South University Press, Nanjing (2007)
17. Hillier, B., Stutz, C.: New methods in space syntax. World architecture (11), 54–55 (2005)
18. Bin, J., Bo, H., Feng, L.: Spatial analysis and geovisualization in GIS. Higher Education Press, Beijing (2002)
19. Jiang, B., Claramunt, C., Klarqvist, B.: An integration of space syntax into GIS for modeling urban spaces. International Journal of Applied Earth Observation and Geoinformation (2), 161–171 (2000)

A Feasible Registration Method for Underwater SLAM

Feng Sun[1], Wenjing Wang[2], Fuqiang Liu[3], and Wenfeng Wang[4]

[1] Automation College, Harbin Engineering University, Harbin, China
[2] Beijing Aerospace Control Device Institute, Beijing, China
[3] Beijing Institute of Spacecraft System Engineering, Beijing, China
[4] Electronic Components Reliability Center, Beijing, China
{sunfeng_hrbeu, cristinwwj, liufq407}@163.com, cocow2@sina.com

Abstract. Simultaneous localization and mapping (SLAM) algorithm could make up for the disadvantages of underwater navigation methods based on priori maps, hence makes underwater vehicles truly autonomous. A modified data association method is proposed to lighten the systemic computational burden and improve data association process. The method makes use of a two-step filtration to solve the ambiguities arisen by multiple observations falling into the validation gate of a single feature or an observation lying in the overlapping region of gates of multiple features. Simulation experiments results show that the method could achieve a satisfactory association result with O(mn) computation complexity even when dead-reckoning error is quite large, thus suitable to on-line data association for underwater SLAM implementations.

Keywords: underwater navigation, SLAM, data association, computational complexity.

1 Introduction

Underwater geophysical navigation methods such as terrain-aided navigation, INS/Gravity integrated navigation and geomagnetic navigation, which are based on various priori environmental information maps, could obtain a pretty high localization precision; however, these methods tend to be influenced by the mapping precision of priori maps. Besides, these methods are valid only when the vehicle is working within certain regions covered by priori maps. The simultaneous localization and mapping (SLAM) algorithm incrementally construct an environmental information map while simultaneously using this map to locate itself without the help of priori maps, hence could make up for the aforementioned disadvantages of underwater navigation methods based on priori maps.

In despite of the great improvements which have been made to the theoretical study of SLAM in the last two decades, there remains some critical problems unsolved, of which the most important two are computational complexity and data association. Since the computations required by SLAM scales as the square of the number of features maintained in the map, such huge computation complexity make people try every means to cut down the systemic computation burden. Data association, which is also called registration or correspondence, is another most

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 15–22.
springerlink.com

challenging problem in SLAM. It refers to the process of specifying the correct correspondence between sensed feature observations and map features. Data association is of great importance to SLAM implementations. False associations may destroy the consistence of map by updating map and vehicle states with mismatched observations, leading to filter divergence and estimation failure.

Therefore, taking computational complexity and association efficiency into account, a modified data association (MDA) is presented to achieve a more computational and robust data association for underwater SLAM implementations. The rest of the paper is organized as follows: the current state-of-the-art in data association for SLAM is reviewed in Section 2. Section 3 presents MDA algorithm and its application in SLAM. Section 4 shows the simulation results under different circumstances. Conclusions and future work are summarized in Section 5.

2 Current Data Association Methods

Generally, the data association methods in SLAM can be classified as Bayesian and non-Bayesian approaches. Non-Bayesian approach mainly refers to the conventional nearest neighbor (NN), which is also the most widely employed data association method in existing applications as in [1] & [2]. It associates a feature to the nearest observation falling in the validation gate centered about the predicted feature location. The marked advantages of NN are the conceptual simplicity and its $O(mn)$ computation complexity (m corresponds to the observations number available and n to the number of features maintained in the map), and it works quite well where clutter density is low and sensor precision is high. But obviously, the hypothesis that the nearest observation is the associated one isn't always the truth. As a result, its performance deteriorates rapidly when an observation falls into the overlapping area of multiple features gates or in high clutter environment.

Most of Bayesian approaches are based on probability density function (pdf). G. D. Huang [3] has proposed a hybrid independent/coupled sample-based joint probability data association filter (Hyb-SJPDAF) to solve multiple target tracking and data association problem. The more famous and complex multiple hypotheses tracking (MHT) method in target tracking field is also applied to SLAM data association process to achieve better correspondences by [4]. However, the hypothesis tree in MHT grows exponentially in time, which means that exponential memory and computational resources are required, thus has limited its applications in SLAM.

Besides, other matching algorithms are also introduced to improve the data association. Tim Bailey [5] uses a graph search to perform batch-validation gating by constructing and searching a correspondence graph in combined constraint data association (CCDA) method. The joint compatibility branch and bound (JCBB) method proposed by J.Neira & J.D.Tardos [6], which is based on branch and bound search, has taken the full spatial correlations between vehicle and features into account to make correspondence decisions, thus resulted in an exponential search space and a big computational burden for the real-time implementation in moderate-sized environments.

Almost all of aforementioned approaches tend to perform better associations than NN, but need far more computational and memory resources which make them impracticable and unsuitable for large scale real-time implementation. However, truly autonomous vehicle in completely unknown, large-area environment must be capable of on-line self-localization and mapping, so the systemic requirements to the computation complexity are rigorous. The computation complexity, with no doubt, should be an important criterion for data association methods besides the association efficiency. That's exactly why NN is prevalent although its performances are not satisfactory enough. How to achieve the best associations with the least computations is the final aim of all association methods. In this paper, we have proposed an efficient data association called MDA, which is suitable for the on-line implementation of SLAM with a computation complexity equal to NN.

3 Problem Formulation

The SLAM algorithm is typically implemented in a two-step recursive prediction correction form as in [7], which can be illustrated in detail as Fig. 1.

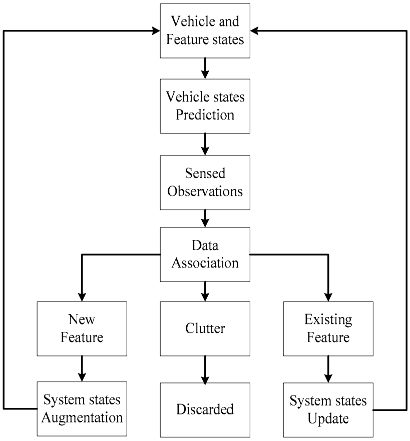


Fig. 1. SLAM recursive estimation circle

The vehicle change-in-pose estimation $\mathbf{\Delta X}$ with covariance $\mathbf{\Delta P}$ is always obtained by the on-board internal sensor such as gyroscope or a vehicle dynamic model and is used to predict the vehicle location. The vehicle process model in discrete time can be written as

$$\mathbf{X}_v(k+1|k) = f(\mathbf{X}_v(k|k), \mathbf{\Delta X}(k+1)) \; . \tag{1}$$

Since features in the environment are assumed stationary, states of features do not change during the prediction stage, that is,

$$\mathbf{X}_F(k+1|k) = \mathbf{X}_F(k|k) = \{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_n\} \; . \tag{2}$$

Hence, the augmented system vector prediction is

$$\mathbf{X}(k+1|k) = g\left(\mathbf{X}(k|k), \Delta\mathbf{X}(k+1)\right) = \begin{bmatrix} f\left(\mathbf{X}_v(k|k), \Delta\mathbf{X}(k+1)\right) \\ \mathbf{X}_F \end{bmatrix} . \tag{3}$$

with the covariance $\mathbf{P}(k+1|k)$

$$\mathbf{P}(k+1|k) = \nabla g_X \mathbf{P}(k|k) \nabla g_X^T + \nabla g_{\Delta X} \mathbf{\Delta P} \nabla g_{\Delta X}^T . \tag{4}$$

where $\nabla g_{\Delta \mathrm{x}}$ denotes the Jacobian of $g(\cdot)$ with respect to $\mathbf{\Delta X}$ evaluated at $(\mathbf{X}, \mathbf{\Delta X})$. Other denotations are similar and will not be mentioned again.

Once observations have been extracted from the sensor scan, the data association process is first passed to decide whether the observations and previously stored features are associated. Correct correspondence between available observations and map features are essential for consistent map building and vehicle pose estimation. Now we focus on the application of the MDA method to SLAM.

Almost every association method makes use of validation gate to reduce the number of candidate observations, so does MDA presented here. The validation gate defines a region in observation space centered about the predicted feature location $h(\mathbf{X}_v(k+1|k), \mathbf{F}_i)$ of feature $\mathbf{F}_i$. Only the observations falling inside this region are considered candidate observations of $\mathbf{F}_i$. Association ambiguity may arise when single observation falls into the gates of multiple features or multiple observations coexist in the gate of a single feature, thus false associations probably occur.

To settle the ambiguities, for a particular feature, we first compute the Mahalanobis distance or the normalized innovation squared (NIS) between every candidate observation $\mathbf{z}_i$ and the predicted feature location, then we get

$$\mathbf{NIS}_i = \mathbf{\upsilon}_i^T \mathbf{S}_i^{-1} \mathbf{\upsilon}_i . \tag{5}$$

Where, $\mathbf{\upsilon}_i = \mathbf{z}_i - \hat{\mathbf{z}} = \mathbf{z}_i - h(\mathbf{X}_v(k+1|k), \mathbf{F}_i)$ denotes the observation innovation and $\mathbf{S}_i$ represents the innovation covariance. Then the observation with the minimal NIS value is picked out to make a comparison with a threshold value r. The value of r is predetermined and the detail discussions on how to choose its value appropriately will appear in Section 4.

$$\mathbf{NIS}_{\min} < r . \tag{6}$$

If $\mathbf{NIS}_{\min}$ is larger than r, the observation will be dropped; If $\mathbf{NIS}_{\min}$ is less than r and $\mathbf{NIS}_{\min}$ corresponds to observation $\mathbf{z}_k$, then the association pair of feature $\mathbf{F}_i$ and observation $\mathbf{z}_k$ will be accepted temporally. By this means we make certain that there is only one observation in the validating gate of feature $\mathbf{F}_i$.

Next, all the observation and feature temporal association pairs available in this way, along with the corresponding $\mathbf{NIS}_{\min}$ values, are assembled to form a temporal association group. If environmental feature density is relatively high, single observation might assigned to multiple features in temporal association group, which means that the observation has fallen into the overlapping area of validation gates of these features. Again, for a single observation, we choose feature with the minimum

$\mathbf{NIS}_{\min}$ value in the group to associate. For instance, both $\mathbf{F}_1$ and $\mathbf{F}_2$ correspond to $\mathbf{z}_6$ in temporal association group, and $\mathbf{NIS}_{\min}$ for $\mathbf{F}_1$ is 0.051 while the one for $\mathbf{F}_2$ is 0.028, then a final decision is made that $\mathbf{z}_6$ corresponds to $\mathbf{F}_2$. So we set up an association rule when observation falls into multiple gates. The basic principle is to ensure an observation must correspond to a unique feature.

After these filtrations, the association pairs remained are assumed to be optimal associations, and will be used to update the vehicle pose and map states. Non-clutter observations not assigned to any existing features will be transformed into a new feature in the map using current vehicle location.

For instance, if $\mathbf{F}_i$ is associated with $\mathbf{z}$ of variance $\mathbf{R}$, then the predicted observation and the innovation can be written as

$$\hat{\mathbf{z}} = h(\mathbf{X}_v(k+1|k), \mathbf{F}_i), \ \boldsymbol{\upsilon}(k+1) = \mathbf{z}(k+1) - \hat{\mathbf{z}} . \quad (7)$$

where $h(\mathbf{X}_v(k+1|k), \mathbf{F}_i)$ is the observation model. The innovation covariance $\mathbf{S}(k+1)$ and Kalman gain $\mathbf{W}(k+1)$ are given by

$$\mathbf{S}(k+1) = \nabla h_X \mathbf{P}(k+1|k) \nabla h_X^T + \mathbf{R} , \ \mathbf{W}(k+1) = \mathbf{P}(k+1|k) \nabla h_X^T \mathbf{S}(k+1)^T . \quad (8)$$

The updated system state estimation and corresponding covariance matrix become

$$\mathbf{X}(k+1|k+1) = \mathbf{X}(k+1|k) + \mathbf{W}(k+1)\boldsymbol{\upsilon}(k+1) . \quad (9)$$

$$\mathbf{P}(k+1|k+1) = \mathbf{P}(k+1|k) - \mathbf{W}(k+1)\mathbf{S}(k+1)\mathbf{W}(k+1)^T . \quad (10)$$

4 Experiment Results

The data association algorithm presented is tested via simulations. Comparisons under the same initial errors and measurement perturbations are also made between NN and MDA. Limited by simulation conditions, it is assumed that there are ideally no clutters in the environment.

In the simulation, the association results acquired by data association method at time k are considered true if they are exactly the truth. The might occurring false associations can be classified into two categories: Fault I. several association pairs are missed. For example, at time k there should be 5 association pairs of features and observations, however, the process only confirmed 4 pairs, and all of the 4 pairs are correct correspondences. The unassociated observation will be considered as a new feature and added into the map for the second time, thus a single feature will appear at two different locations on the map; Fault II. Some pairs are correct, and the others are false, which means some observations have been assigned to wrong features. In such circumstance, the consistency will be destroyed by updating features with false observations, leading to the filter divergence.

4.1 Straight Sailing Mode

In experiments, vehicle are planned to start at (1m,40m) and navigate straightly forward to the destination (100m,40m) with a constant velocity of 1m/s, and both vehicle pose prediction and measurements sampling frequencies are 1 hertz. The external sensor has a semicircular field-of-view of 25m. 120 point features are randomly distributed in the environment. Vehicle pose comprises the vehicle positions in both east and north directions and vehicle bearing, and the initial vehicle pose $\hat{\mathbf{X}}_{v0} = [1.4 \quad 39.8 \quad 0.003]^T$. There are Gaussian noises in both distance and bearing measurements.

When implementing with NN, the estimation result is shown in Fig.2. Because of the dense distribution of features, NN quickly go wrong from time 7. Since the vehicle follows a straight line trajectory without looping back to permit re-observation of features detected earlier in the mission, all features become obsolete features after being observed for several times, and will no longer contribute to the system state updating. For this reason, the negative influence which false association should exert on filter has been weakened to a great extent. However, the precision is unavoidably reduced due to these false associations all the same.

When using MDA, different association results can be obtained as we gradually increase the value of r, and the corresponding results are listed as follows.

1. When $r = 0.08$, the first false association of Fault I, occurs at time 48. We can learn from this that the threshold has filtered out several correspondences which should be associated. Therefore, the threshold value is not large enough and we should continue increasing r;
2. When keeping on increasing r to be equal with 0.2, the occurring instant of the first false association of Fault I has been postponed to 88;
3. When $0.21 \le r \le 1.36$, all of the association pairs during whole mission are correct;
4. When $r = 1.37$, the first false association of Fault II, rose at time 87;
5. When the value of r becomes larger and larger, the occurring instant of the first false association is more and more advanced, meaning that the happening probabilities of Fault II are larger and the threshold is too large.

We can draw conclusions from above analyses that, the value of threshold has a prevailing influence on association results. The value of threshold depends mainly on the error characteristics of external sensor and prediction errors of vehicle. In the experiments, perfect associations can be achieved when $0.21 \le r \le 1.36$, so every r falling in the bounds is good for us and we have a vast space to choose in practice. In real world implementations, we can easily specify a threshold fit for current vehicle and sensor characteristics via repeated experiments before hand. Choose $r = 0.7$ randomly, and MDA results show in Fig.3.

As we see in Fig.2-Fig.3, when vehicle dead-reckoning error and measurement error are relatively large, results acquired by NN is quite poor with divergent map estimations, however, MDA is perfect all the time as long as proper r is chosen. Besides, the computation complexity of MDA is compared with NN by PC running time. Both algorithms are run on Pentium IV PC, 3.2GHz CPU, 1GB RAM, and the average running time for NN is 3.328s, while MDA 3.25s. MDA have shown an equal computation complexity to NN, thus can also be applied to on-line SLAM.

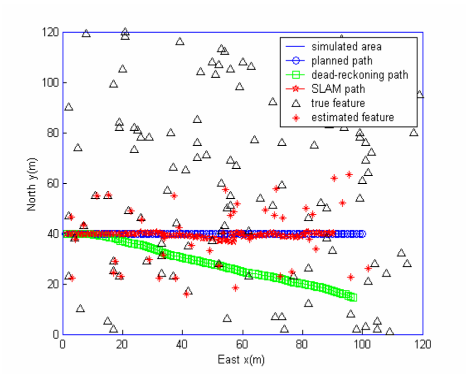


Fig. 2. NN Estimation in Experiment I

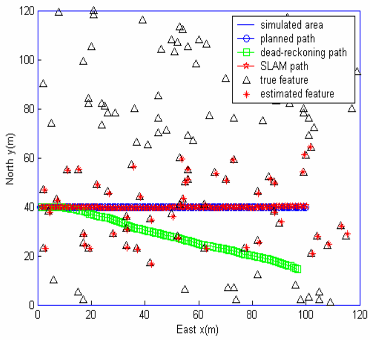


Fig. 3. MDA Estimation in Experiment I

4.2 Local Searching Mode

In Experiment II, the vehicle begins at (10m, 10m) and explores the environment three times following a square path anticlockwise, with a constant velocity of 1m/s. There are 16 uniformly spread features altogether. The view range of external sensor is 20m and the initial vehicle state is $[10.2 \quad 9.8 \quad 0.001]^T$. Other conditions are just as same as Experiment I.

When using NN to complete the associations, as we see in Fig.4 and Fig.5, during the beginning stage the associations are pretty good, however, the false association of Fault II occurred at time 21 for the first time, and all the following associations are false, leading to fast filter divergence. The maker symbols in Experiment II just have the same meaning with Fig.2.

After several trials, we choose $r = 1.5$ and perfect associations last for whole three loops exploration. Localization results are shown in Fig.6 and Fig.7. Although NN deteriorates rapidly due to large dead-reckoning and observation errors, MDA shows a long-term perfect association efficiency.

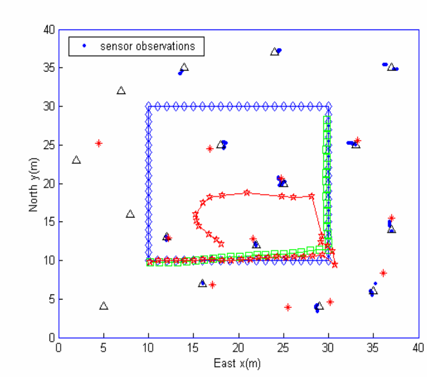


Fig. 4. NN estimation in Experiment II

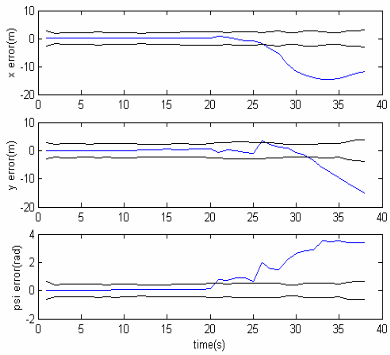
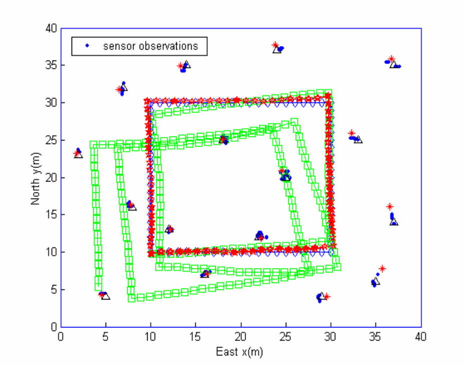


Fig. 5. NN vehicle pose estimation errors



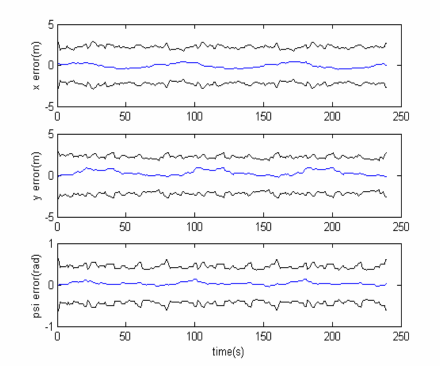


Fig. 6. MDA estimation in Experiment II

Fig. 7. MDA vehicle pose estimation errors

5 Conclusions

A modified data association has been presented in this paper, and experiments are simulated to test the efficiency of the algorithm. Experiments show that compared with the conventional NN method, MDA can achieve a comparable computation complexity and much better association performance, and have excellent advantages in terms of computations and association stability. MDA tends to help release the computation burden of SLAM algorithm and fit for real time data association. The efficiency of MDA without environmental clutters has been checked out, and future work will focus on study on the association robustness under different clutter density and the influence of sensor precision on the association performance.

References

1. Bar-Shalom, Y., Li, X.R., Kirubarajan, T.: Estimation with Applications to Tracking and Navigation. John Wiley and Sons, Malden (2001)
2. Nieto, J., Guivant, J., Nebot, E.: Real Time Data Association for FastSLAM. In: 2003 IEEE International Conference on Robotics & Automation, pp. 412–418. IEEE Press, Taipei (2003)
3. Huang, G.Q., Rad, A.B., Wong, Y.K., Ip, Y.L.: SLAM with MTT: Theory and Initial Results. In: 2004 IEEE Conference on Robotics, Automation and Mechatronics, pp. 834–839. IEEE Press, Singapore (2004)
4. Tena Ruiz, I., Petillot, Y., Lane, D.M., Salson, C.: Feature Extraction and Data Association for AUV Concurrent Mapping and Localization. In: 2001 IEEE International Conference on Robotics & Automation, pp. 2785–2790. IEEE Press, Seoul (2001)
5. Bailey, T., Nebot, E.M., Rosenblatt, J.K., Durrant-Whyte, H.F.: Data Association for Mobile Robot Navigation: a Graph Theoretic Approach. In: 2000 IEEE International Conference on Robotics & Automation, pp. 2512–2517. IEEE Press, San Francisco (2000)
6. Neira, J., Tardos, J.D.: Data association in stochastic mapping using the joint compatibility test. IEEE transactions on robotics and automation 17, 890–897 (2001)
7. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping (SLAM): Part I. IEEE robotics & automation magazine 13, 99–108 (2006)

A Promoted Global Convergence Particle Swarm Optimization Algorithm

Du Ronghua and Cai Yue

Changsha University of Science and Technology, Changsha, 410076
csdrh@163.com, Caiyue1234@163.com

Abstract. Particle Swarm Optimization (PSO) has premature convergence problem. Theory and experiment have proven that parameters of PSO establish the proportion relation of local search capabilities and global search capabilities, and have great influence to the convergence. In this paper, the existing parameter adjustment strategies are studied and analyzed and their existing problems are pointed out. Using the diversity and the mutation mechanism of the vertebrate immune system for reference, a new parameter adjustment strategy is presented. The new strategy, based on the affinity of antibodies and the aggregation level of particles, determines the optimal fitness value change rate and algorithm parameters. The test results of the classic function show that the global convergence capability of PSO is significantly improved, and the premature convergence problem of the PSO algorithm is effectively avoided.

1 Introduction

PSO (Particle Swarm Optimization, PSO) proposed by Kennedy and Eberhart in 1995, is a kind of evolutionary computation technique which simulating bird population [1, 2]. And it is an effective tool to solve nonlinear continuous optimization problem and combinatorial optimization problem because of its simple algorithm, easy realization and fewer adjustable parameters. The algorithm quickly closes to the optimal value at beginning of iteration, and then slowly converges for the monotone function, strictly convex function and unimodal function; the premature phenomenon is easier to appear for the multimodal function [3]. Many solutions, which mainly focus on adjusting parameters, were analyzed and put forward to solve these problems by domestic and foreign scholars, but there are corresponding defects. Using the diversity and the mutation mechanism of the vertebrate immune system for reference, a new parameter adjustment strategy is presented. The new strategy, based on the affinity of antibodies and the aggregation level of particles, determines the optimal fitness value change rate and algorithm parameters. The test results of the classic function show that the global convergence capability of proposed method is significantly improved, and the premature convergence problem of the PSO algorithm is effectively avoided.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 23–30.
springerlink.com

2 The Basic PSO

In the continuous space coordinate system, the mathematical description of PSO is: the size of particle swarm is S , the coordinate position of each particle in N dimensional space is $x_i = (x_{i1}, x_{i2} \cdots x_{iN})$, The velocity of particle i ($i = 1,2 \cdots S$) is defined as the moving distance in each iteration , $v_i = (v_{i1}, v_{i2} \cdots v_{iN})$. And individual extremum of particle i is $p_i = (p_{i1}, p_{i2} \cdots p_{iN})$, The global extremum of particle swarm is $p_g = (p_{g1}, p_{g2} \cdots p_{gN})$. So the velocity of particle i ($i = 1,2 \cdots S$) in d ($d = 1,2 \cdots N$) dimensional subsection space is adjusted as follows:

$$v_{id} = \omega v_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}) \quad (1)$$

$$\begin{cases} v_{id} = v_{\max}, if v_{id} > v_{\max} \\ v_{id} = -v_{\max}, if v_{id} < -v_{\max} \end{cases} \quad (2)$$

Where, c_1 and c_2 are acceleration constant, usually limited to 1.5 ; r_1 and r_2 are random number in $[0,1]$; ω is inertia weight. $v_{\max}$ determines the search precision of particle in solution space. If $v_{\max}$ is too large, particle may fly over the optimal solution, while too small, particle will easily fall into local search.

The coordinate position x_{id} of particle i is adjusted as follows:

$$x_{id} = x_{id} + v_{id} \quad (3)$$

The iterative algorithm composed by formula (1) and (3) is considered as basic PSO algorithm.

3 Convergence Problem of PSO

Similar to other evolutionary algorithm, PSO algorithm searches the optimal solution in complex space through collaboration competition between the individuals. PSO randomly initialize a particle swarm in the solution space. Each particle is a solution of optimization problem. The objective function determines a fitness value for a particle (the objective function value corresponding to coordinate position is defined fitness value.). Each particle moves in the solution space according to formula (1) and (3). In each iteration, the particle updates their position by tracking two extreme values: One is the optimal solution searched by itself, on behalf of their cognitive level of particle; the other is the global optimal solution, on behalf of the community awareness level [4]. However, during the running of PSO, if a particle found an optimal location, other particles would quickly close to it. If the optimal location is the local optimal, the PSO will not be able to re-search in solution space and will sink into local optimum. So the premature convergence phenomenon appears.

Since the PSO was put forward, its convergence problem becomes the point of research. The convergence of PSO concerns with parameter setting. It is important how to establish the proportion relation of local search capabilities and global search capabilities. Experiments show that, inertial weight ω will affect search capability, the larger ω, the stronger global search capability and the weaker local search capability. Inertial weight ω was regard as a constant in previous work by Shi; it was found that linearly decreasing weight could adjust local search capabilities and global search capabilities [5]. Namely:

$$\omega(t)=\frac{(\omega_{ini}-\omega_{end})\times(T_{\max}-t)}{T_{\max}}+\omega_{end} \qquad (4)$$

The dynamic inertial weight [5] is also proposed by Eberhart, $\omega=0.5+r(t)/2$, where, $r(t)$ is random number in $[0,1]$. The optimization result obtained by dynamic inertial weight is better than fixed, but there are two drawbacks: First, the local search capability is weak at the beginning of iteration, So it is possible to miss the global optimum even if the initial particles have been closer to it, the global search capability is weaken at the end of iteration, PSO is easy to fall into local extreme. Second, it is difficult to forecast the largest iteration $T_{\max}$, which will affect the regulatory function of PSO. In 2001, Shi used fuzzy system to regulate weight, but this method need to establish fuzzy rules based on expert knowledge [6], the domain knowledge is scarce and difficult to obtain before a complex system was optimized, so the realization is more complicated and not easy to use. The inertial weight which has influence on the convergence of PSO algorithm has been analyzed by ZHANG Li-ping [7] and a random inertia right strategy was proposed to improve the PSO algorithm performance.

In the early stage of the study, the algebraic analysis has been used as a method to study the convergence condition of PSO when p_i and p_g keep invariant[8], namely obtain the convergence condition which ω、 c_1 and c_2 satisfy. It has been proven that PSO converges to the weight center $\frac{\varphi_1 p_i+\varphi_2 p_g}{\varphi}$ of p_i and p_g , when p_i and p_g keep invariant and $\sqrt{2(1+\omega-\varphi)^2-4\omega}<2$ ($\varphi=\varphi_1+\varphi_2$, $\varphi_1=c_1r_1$, $\varphi_2=c_2r_2$). But in fact, p_i and p_g will continuously be renewed in accordance with their fitness value during evolution. Although the article proved that PSO is convergent as long as the parameters meet the above conditions, it cannot ensure that PSO converges to the global optimum.

The experiment proved that particles would congregated if premature convergence or global convergence appears [9]. The particles will gather in a particular location or several specific locations, which mainly depend on the own characteristics and fitness value function. Literature [9] proved in theory that the consistent position of the particles equivalent to the same fitness. So we can track the status of the particle swarm through studying the overall change of the particles fitness value. A new PSO, based on adaptive mutation scheme is presented. p_g satisfying certain conditions will mutate in probability p_m .

$$p_m = \begin{cases} k, \sigma^2 < \sigma_d^2 and f(p_g) > f_d \\ 0, others \end{cases} \tag{5}$$

Where, k takes value in $[0.1,0.3]$. σ_d^2 is related to actual problem, f_d can be set to the theoretically optimum value. But the method only considers the minimization problems, and cannot distinguish the premature convergence and the global convergence. It is not convenient to use.

The literature [10] set ω to zero and removed the speed term in equation (1), which lead to weaken the global searching ability. At each generation, at least one particle is at the optimal position. Using stopping evolution particle to improve the global searching ability is the basic thought of literature [10]. Although it is proved that the algorithm can converges to the global optimum in probability 1, In fact, PSO is a hybrid algorithm based on genetic algorithm and simulated annealing algorithm. The experiment proved that the convergence rate decreased significantly and the complexity increased, so that particles can not quickly and effectively escape local extreme point.

4 A Promoted Global Convergence PSO

In order to improve PSO algorithm, we focus not only on convergence, but also on adjusting the scope of the search algorithm, as well as the global and the local search capabilities. Aim of parameter adjustment is improving the convergence rate and global convergence capability in the early iteration and increasing the capacity of local convergence in the late iteration. The algorithm can flexibly adjust the overall search and local search capabilities if ω will be changed along with the optimal value.

Using the diversity and the mutation mechanism of the vertebrate immune system for reference, a new parameter adjustment strategy is presented.

According to the properties of biological immune system, in constructing artificial immune system, the artificial antigen and antibodies need be constructed firstly. The affinity is used to describe the similarity degree of antibodies and antigens, so that the artificial and the biological immune system have the similar self-regulatory mechanism. When artificial immune system is used to solve optimization problems, the optimal solution meeting the constraints is antigen, candidate solution is antibody, the affinity of antibody and antigen reflects the similarity degree of the candidate and the optimal solution. And it also reflects the satisfaction degree of the candidate solution meeting the constraints and the objective function. The affinity of two antibodies reflects their similarity degree and the diversity of antibodies. In evolutionary algorithm, maintaining the diversity of antibodies can prevent the algorithm into local optimal solution. Based on the affinity of antibody and antigen to choose effective antibody fully reflects the principle of selection - “the survival of the fittest ", the more obvious of the principle, the higher efficiency of search.

Definition 1: Affinity degree between antigen and antibody k is:

$$A_k = \frac{1}{1+t_k}, t_k = \frac{f(x_k) - f^{\min}(x)}{f^{\max}(x) - f(x_k) + \delta} \tag{6}$$

Where, t_k shows combined degree between antigen and antibody k, $f^{\min}(x), f^{\max}(x)$ are minimum and maximum of objective function values, and δ is a little Positive Number, so dividend is not zero.

Definition 2: For a given number N of the population, concentration of the population is:

$$C = \frac{\sum_{i=1}^{N} f(i)}{N} \tag{7}$$

$$f(i) = \begin{cases} 1, A_i > \sigma \\ 0, A_i \le \sigma \end{cases} \tag{8}$$

Where, C is concentration of the population, reflects the aggregation extent of particles, N is Total number of antibodies, $f(i)$ is concentration function of antibody, σ is concentration Threshold Value, also is concentration inhibition radius, $0 < \sigma \le 1$.

Definition 3: Based on Affinity degree and concentration of antibody, the optimal fitness value change rate p_k is:

$$p_k = \alpha \frac{A_k}{\sum_{j=1}^{N} A_j} + (1-\alpha)\frac{1}{N} e^{-\frac{c}{\beta}} \tag{9}$$

where, α, β are Regulation factor, $0 < \alpha, \beta < 1$, and A_k is affinity degree of antibody, the greater affinity degree , the greater choice probability; the greater the concentration of antibodies, the smaller choice probability, and which retain high affinity antibodies with antigen, and also reduce the choice of the similar antibodies to ensure population diversity in the evolution of population.

Referred to idea described in literature [7], based on the optimal fitness value change rate p_k, setting $\omega = 1 - p_k$ will improve convergence ability at the early iteration stage because of larger ω. While closing to the local optimum, the concentrations of k increases and p_k decreases because of clustering phenomena.

The optimal particle is mutated to improve the global searching ability in literature [9], but the method is not convenient. Referred to idea described in literature [9], by adjusting c_2 to mutate the best individual, the algorithm improves the ability of PSO algorithm to break away from the local optimum solutions effectively. If $c > \gamma$, where, $0.75 < \gamma < 1$, c_2 in formula(1) will be set according to formula(10):

$$c_2 = 1.5 \times (1 + 0.5 \times \eta) \tag{10}$$

Where, η is random variable obeying $Gauss(0,1)$.

From above analysis, we can conclude that parameters-setting methods of this paper are similar to methods in literature [6-9]. The calculation amount is increased, but setting is more convenient and the algorithm maintains the simplicity.

5 Performance Test

To verify the parameters-setting methods, A computer simulation program was developed and two classical test functions (Rosenbrock and Rastrigin) in literature [9] were selected to test and compare the searching performance of M1 (this paper), M2 (literature [7]), M3 (literature [5]) and M4 (literature [9]).

Rosenbrock is a unimodal function. There is a very strong coupling between variables of the function, its global minimum point is $x^* = (1,1\cdots1)$ and global minimum is $f(x^*) = 0$.

$$f(x) = \sum_{i=1}^{n} \left(100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right), \ -10 < x_i < 10 \tag{11}$$

Rastrigin is a multimodal function. There are many local minimum points, its global minimum point is $x^* = (1,1\cdots1)$ and global minimum is $f(x^*) = 0$.

$$f(x) = \sum_{i=1}^{n} \left(x_i^2 - 10\cos(2\pi x_i) + 10\right), -10 < x_i < 10 \tag{12}$$

Two classical test functions whose dimension is 10 are test, parameter settings are: initial population size is 15, the biggest evolution of the number $\lambda = 10000$, concentration threshold $\sigma = 0.5$, regulatory factors $\alpha, \beta = 0.5$, $c_1 = c_2 = 2.0$ in M1 and M2, $\alpha_1 = 0.5$, $\alpha_2 = 0.4$ in M2 and $\omega_{ini} = 0.9$, $\omega_{end} = 0.4$ in M4.

Two testing scheme for test functions are as follows:

Testing scheme 1: To test the evolutional generation of four algorithms for given optimization accuracy. So the variance of the population's fitness is defined as:

Definition 4: The number of the particles is n, f_i is fitness of particle i, f_{avg} is the mean fitness value of particle swarm, σ^2 is the variance of the population's fitness. So σ^2 is defined as:

$$\sigma^2 = \sum_{i=1}^{n} \left| \frac{f_i - f_{avg}}{f} \right|^2 \tag{13}$$

Where, f is normalized calibration factor, which limits the size of σ^2. In this paper, f is taken as follows:

$$f = \begin{cases} \max\left\{\left|f_i - f_{avg}\right|\right\}, \max\left\{\left|f_i - f_{avg}\right|\right\} > 1 \\ 1, others \end{cases} \tag{14}$$

From definition 4, we can conclude that the variance of the population's fitness σ^2 reflects the convergence degree of all particles. The smaller σ^2, the higher is the convergence degree; On the contrary, the particle swarm is at random search stage. This paper set $\sigma^2 \leq 0.1$, the whole experiments for 4 algorithms were repeated for 50 times. The even evolution generations of four algorithms are in Table 1.

Testing scheme 2: To test σ^2 of four algorithms for given evolution generation 2000. The whole experiments for 4 algorithms were repeated for 50 times, the even σ^2 of four algorithms are in Table 2.

The test results show that four algorithms have similar optimization performance in one-peak functions Rosenbrock. M2 has the least evolution generations at certain accuracy and has the highest accuracy at certain evolution generations, because M2 has good convergence, without considering the local optimal issues, can be quickly and effectively close to the optimal value. Meanwhile, M2 has the largest evolution generations, because too much attention has been given to global convergence, which affects the convergence rate. As for multimodal Function Rastrigin, M3 has the slowest convergence speed, 23 tests did not reach the accuracy at certain evolution generations; M2 has stronger global search capabilities and quicker convergence speed; M2 has stronger global search ability and slower convergence speed;M1 has the best effect to adjust local search capabilities and global search capabilities.

Table 1. Comparison of evolution generations among different methods at certain accuracy

Function	M1	M2	M3	M4
Rosenbrock	1420	1368	1569	2345
Rastrigin	1568	1943	5120	3276

Table 2. Comparison of accuracy among different methods at certain evolution generations

Function	M1	M2	M3	M4
Rosenbrock	0.0823	0.0965	0.1046	0.1157
Rastrigin	0.0675	0.0787	0.1239	0.1543

6 Inclusion

Theory and experiment have proven that PSO parameters establish the proportion relation of local search capabilities and global search capabilities, and have great influence to the convergence. In this paper, the existing parameter adjustment strategies are studied and analyzed and their existing problems are pointed out. Using the diversity and the mutation mechanism of the vertebrate immune system for reference, a new parameter adjustment strategy is presented. The new strategy, based on the affinity of antibodies and the aggregation level of particles, determines the optimal fitness value change rate and algorithm parameters. The test results of the classic function show that the global convergence capability of proposed method is significantly improved, and the premature convergence problem of the PSO algorithm is effectively avoided.

References

[1] Kennedy, J., Eberhart, C.: Particle swarm optimization. In: Proc IEEE international conference on Neural Networks, vol. 4, pp. 1942–1948. IEEE Press, USA (1995)

[2] Shi, Y., Eberhart, R.C.: Modified swarm optimizer. In: IEEE International Conference of Evolutionary Computation. IEEE Press, Anchorage (1998)

[3] Xie, X.-f., Hang, W.-j., Ang, Z.-l.: Overview of particle swarm optimization. Control and Decision 18(2), 129–134 (2003) (in Chinese)

[4] Eberhart, R.C., Simpson, P.K., Dobbins, R.W.: Computational Intelligence PC Tools. Academic Press Professional, Boston (1996)

[5] Shi, Y., Eberhart, R.C.: A modified particle swarm optimizer. In: Proceedings of the IEEE Congress on Evolutionary Computation, pp. 303–308. IEEE Press, Piscataway (1998)

[6] Shi, Y., Eberhart, R.C.: Fuzzy adaptive particle swarm optimization. In: Proceedings of the IEEE Congress on Evolutionary Computation, pp. 101–106. IEEE Press, Seoul (2001)

[7] Zhang, L.-p.: Analysis and Improvement of Particle Swarm Optimization Algorithm. Information and Control 33(5), 513–517 (2004) (in Chinese)

[8] Ozcan, E., Mohan, C.: Particle swarm optimization: Surfing the waves. In: Proc. of the Congress on Evolutionary Computation, pp. 1939–1944. IEEE Service Center, Piscataway (1999)

[9] Lu, Z.-s., Hou, Z.-r.: Particle Swarm Optimization with Adaptive Mutation. Acta Electronica Sinica 32(3), 416–420 (2004) (in Chinese)

A Web-Based Integrated System for Construction Project Cost Prediction

Huawang Shi and Wanqing Li

School of Civil Engineering, Hebei University of Engineering, Handan, 056038, China
stone21st@163.com

Abstract. Construction cost estimation and prediction, the basis of cost budgeting and cost management, is crucial for construction firms to survive and grow in the industry. The objective of this paper is to presented a novel method integrating fuzzy logic(FL), rough sets (RS) theory and artificial neural network (ANN) which inherent in. The particle swarm optimization (PSO) technique is used to train the multi-layered feed forward neural networks With this model integrating WWW and historical construction data to estimate conceptual construction cost more precisely during the early stage of project. Becouse there are many factors affecting the cost of building and some of the factors are related and redundant, rough sets theory is applied to find relevant factors to the cost, which are used as inputs of an articial neural-network to predict the cost of construction project. Therefore, the main characteristic attributes were withdraw, the complexity of neural network system and the computing time was reduced, as well. A case study was carried out on the cost estimate of a sample project using the model. The results show that the integrating rough sets theory and articial neural network can help understand the key factors in construction cost forecast, and it provided a way for projecting more reliable construction costs.

Keywords: fuzzy logic, rough sets, artificial neural network, particle swarm optimization, construction cost estimation.

1 Introduction

Cost prediction, the basis of cost budgeting, and cost management, not only play an essential role in construction project feasibility studies, but are fundamental to a project's ultimate success[1,2]. The development and use of models, able to predict failure in advance, can be very important for the firms in two different ways. First, as forecasting systems, such models can be very useful for those (i.e. managers, authorities, etc.) who have to prevent failure. Second, such models can be useful in aiding decision-maker of construction firms in charge of evaluation and selection of the sub-contractors.

The prediction of cost has two methods, qualitative method and quantitative method. Qualitative predicting method also called intuitive predicting method is a classical predicting method, laying emphasis on the analysis and prediction of things' nature. Qualitative analysis and prediction methods mainly have experts' prediction

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 31–38.
springerlink.com

method, Delphi method, the subjective probability method and so on. And the experts' prediction method includes personal judgement method, the expert conference method, and the brains storm method. Quantitative prediction is to predict the characteristics of object in quantitative aspect in future. It mainly depends on historical statistics, and uses the mathematical model established in a scientific way to predict the possible number of target. Combining historical statistics with scientific methods mathematical model which was used to predict the amount of the object possibly shown was established. The often-used quantitative methods include time-series prediction method, regression method, gray prediction method and so on.

This study proposes the use of an artificial intelligence approach, the rough sets- an articial neural network Model (RS-ANN), which joins together The Particle Swarm Optimization (PSO), Fuzzy Logic(FL) and Neural Networks (NNs) to improve cost estimation and prediction accuracy. The advantages inherent in rough sets and Neural Networks are incorporated into the RSANN, making this model highly applicable to identifying optimal solutions for complex problems. Furthermore, this paper presents novel method to forecast construction project cost obtained by integrating rough set (RS) theory and anarticial neural network (ANN), and historical construction data to assist in project cost management.

The rest of this paper is organized as follows: The qualitative factors affecting construction cost are quantitatived using Fuzzy Comprehensive Evaluation(FCE), and after introducing of rough set theory, the construction cost factors reduction algorithm, which has 2 steps, such as attribute reduction and attributes value reduction as follows is proposed in section 2, and. In section 3, cost application areas that requir pridiction could be implemented by PSO based ANN. In section 4. Web-base conceptual cost pridiction is proposed. In the final section, this paper focuses on that forecast results discussion.

2 Construction Project Cost Analysis Based on FCE and RS

2.1 Date Deal with FCE

The factors affecting the cost of construction project such as floor area, area, building purposes, standard layer area, story number, building height, door and window type, story height, foundation type, basis type, staircase structural type, shork strengh, structural type, basement area, period, project management level, field condition and so on are analyzed. Then the indexes are described based on rough sets. The table of original cost indexes of construction project is shown in Table 1.

In the construction business, project management level directly impact on the company's project cost. However, the level of project management is a generalization concept of abstract, it is difficult to quantify. In this paper, FCE model to achieve the assessment of the level of project management. the level of project management can be described in aspects such as the level of project manager, the situation of staff, the organization, management planning, and the project control level.

It will take one project manager candidates A in a construction dnterprise's as the evaluation object, and use the above method to evaluate the overall quality of candidates and sort them, then select a candidate with higher overall quality as the project manager lf an important project. First of all, a selection team composed of the experts from the human rdsources department and the leaders carry on the judgment to A candidate is shown in Table1 as follow:

Table 1. Marking table to object A by a expert

Index	Best	Better	Average	Bad	Index	Best	Better	Average	Bad
u_1	1	1.5	0	0	u_2	3	1.5	0	0
u_3	0	2	0.5	0	u_{5u_4}	0	0.5	2	0
u_5	3.5	1	0	0		3.5	1	0	0

According to comment set V of $u_1 \sim u_5$ to obtain the evaluation matrix of single factor, that is:

$$R_2 = \begin{pmatrix} 0.1600 & 0.8000 & 0.0400 & 0 \\ 0.8400 & 0.1600 & 0 & 0 \\ 0.1600 & 0.7600 & 0.0800 & 0 \\ 0.8000 & 0.2000 & 0 & 0 \\ 0.6400 & 0.3600 & 0 & 0 \end{pmatrix}$$

Index weight of A candidate is as follow:

$$A_1^1 = (0.0666,0.0671,0.0677,0.0398,0.0358)$$

The fuzzy comprehensive evaluation vector of candidate A can be got as follow:

$$B^1 = A^1 \bullet R^1 = (0.2904,0.1960,0.5076,0.0048)$$

According to the principle of the maximum membership degree, A's comprehensive level of bid competitive power is 0.5076, that is average. Similar other project' level of management of the fuzzy comprehensive evaluation results can be got as similar.

2.2 Date Based on RS

Pawlak (1982) first introduced rough set theory. Rough sets theory[4] is based on the undistinguished thought and knowledge reduction method. Objects characterized by the same information are indiscernible in view of the available information. It can be used to remove the redundant attributes affecting the project cost, greatly simplify the space dimension of project cost knowledge, depict the importance of different attributes in the expression of project-cost knowledge, simplify the expression space of project-cost knowledge, thus being able to quickly prepare for the prediction of the target project cost. The construction cost factors reduction algorithm has 2 steps, such as attribute reduction and attributes value reduction as in[5].

As rough set approach is concerned with discrete values, we have to transform quantitative attributes into qualitative terms. Even if an attributer epresents a continuous measure, such as floor area and standard layer area the expert usually interprets the values of this attribute in qualitative terms, i.e. low, medium or high.

In Table 1, for example, the results of original indexes is dispersed as Table 1. These indexes are divided into 4 grades {0,1,2,3} to represent {lower, low, average,high}. For example, project management level is divided into 4 grades {0,1,2,3} to represent {bad,average,good,best}. D is the desicion attribute, and is divided into 3 grades {0,1,2} to represent that the construction cost is graded. U is the construction project number.

Table 2. The table of original indexes

U	floor area	building purposes	Standard layer area	story number	project management level	…	*D*
1	0	1	1	1	0		0
2	0	2	2	1	1		2
3	1	1	1	1	2		2
4	0	1	2	0	2		2
5	0	1	2	0	2		1
6	0	2	2	1	0		1
7	0	2	1	0	0		1
8	0	2	1	0	0		1
9	0	1	1	1	0		0
10	0	1	2	0	2		0
…	…	…	…	…	…	…	…

From Table 2, the table of original indexes is reduced as Table 3 by using rough sets algorithm. Rough set analysis part of the experiment was performed with program we developed in JAVA. Table3 shows the results that the forecast index system of construction cost after rough set analysis was performed.

Table 3. The pridiction index system of building cost

Target level	Index
Building cost	Total height H_1
	Standard layer area H_2
	The type of structure H_3
	Project management level H_4
	Period H_5
	Basement area H_6

As we can see, we obtained one minimal reduct $\{H_1, H_2, H_3, H_4, H_5, H_6\}$. This minimal reduct of construction project cost is affected by only 6 important factors, including Total height H_1, Standard layer area H_2, The type of structure H_3. Project management level H_4, Period H_5 and Basement area H_6.

3 ANN Based on PSO

At present, the BP neural network is one of the most maturest, wide spread artificial neural network. Its basic network is three-layers feed-forward neural network such as input layer, hidden layer, and output layer. The input signals must firstly disseminate for- ward into the hidden node. The output information of the concealment node transmits into output node Via- function action. Finally the output variable result is obtained. It helps to discern patterns among input data, requires few ons, and achieves a high degree of prediction accuracy. These characteristics make neural network technolog a potentially promising alternative tool for recognition, classification, and forecasting in the area of construction, in terms of accuracy, adaptability, robustness, effectiveness, and efficiency. Therefore, cost application areas that requir pridiction could be implemented by ANN. PSO is an optimization algorithm, modeled after the social behavior of flocks of birds. Its basic principle and algrithms are in[6].

The demo chart of PSO-based ANN is show as follow:

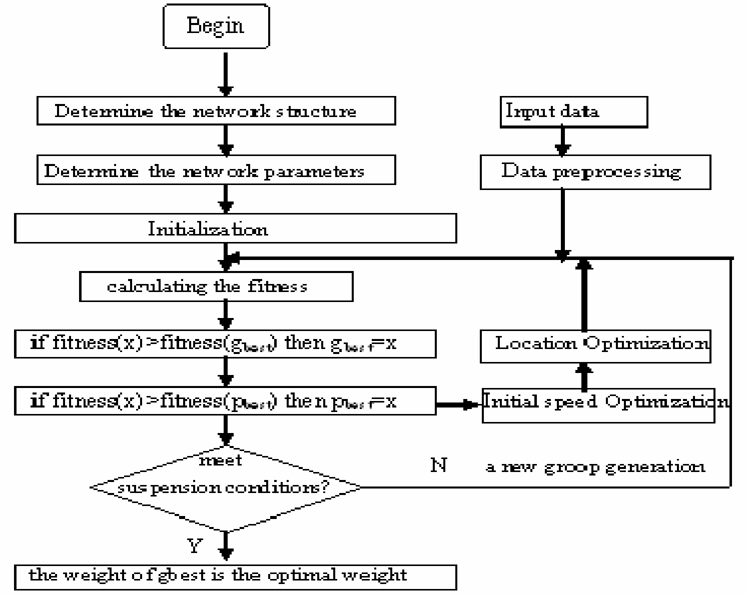

Begin
Determine the network structure
Determine the network parameters
Initialization
calculating the fitness
if fitness(x)>fitness(g_{best}) then g_{best}=x
if fitness(x)>fitness(p_{best}) then p_{best}=x
meet suspension conditions?
Y
N
a new groop generation
the weight of gbest is the optimal weight
Input data
Data preprocessing
Location Optimization
Initial speed Optimization

Fig. 1. The PSO-based network model

Table 4. The running times comparison of three algorithm for XOR problem

Algorithm	Minimum times	Maximum times	Average times
BP	4444	5891	5214
GA	1024	2760	2314
PSO	565	921	847

Table 5. The error comparison of three algorithm with 200 times

Algorithm	Best	Worse	Average
BP	0.4466	0.9981	0.7794
GA	0.0182	0.4071	0.1872
PSO	0.0079	0.1208	0.2437

The performance of the PSO-based multi-layer ANN is benchmarked with a conventional BP-based network and GA-based network. Table4 and Table5 shows comparisons of the results of network for the three different perceptrons. From Table3 and Table4, it can be observed that PSO-based ANN is best with the running times and the errors with the same running times. It can be concluded that the PSO-based perceptron performs better than the BP- based perceptron.

4 Web-Base Conceptual Cost Pridiction

4.1 Conceptual Design

Three critical concerns for planning Web-base conceptual cost pridiction proposed are[2]: (1) Lowest maintenance costs-the system should adopt cetralized data and knowledge base management; (2) globa and all-time accessibility-the system should adopt web-based application with Internet connection; and (3) real-time response-the system should minimize the online calclation requirements. As above concept, data base management and Internet environment were used intuitively to implement the proposed cost pridiction system, the concept of which is shown in Fig.4. Important data, stored in the database, can be handled in the system as developed via onscreen interface buttons. Users canaccess the Web-based system via the World Wide Web(WWW). The evolutionary neural inference system, is used for intelligent knowledge mapping.

4.2 Configuration and Simulation of Neural Network

In this paper, in order to forecast the cost of construction project, the 54 training samples data of Beijing city were collected as learning samples to input network, using the 6 attributes processed by rough set as a cost early-forecasting index system of construction project. The network structure of NN was 6-9-1 for input layer, hidden layer and output layerre spectively. The population size was 200, 200 particles for the swarm. The inertia weight w was 0.3. The acceleration constants c_1 and c_2 were the same 2.0. The maximum velocity was 20 and the minimum velocity was -20. The

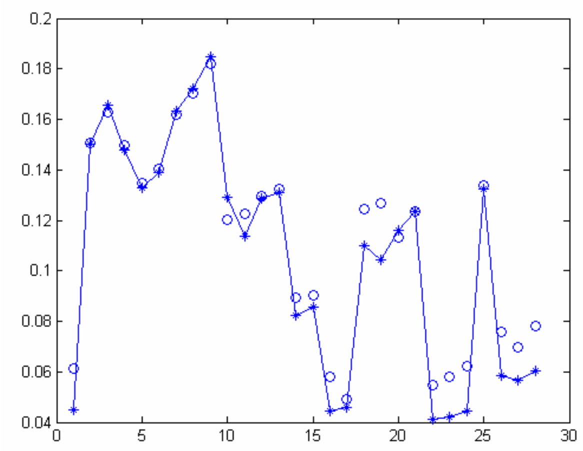


Fig. 2. The training result of net

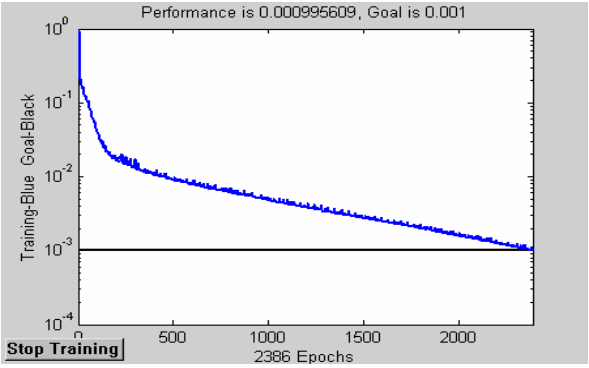


Fig. 3. The error curve of network training

maximum step of iteration was 500. There were 6 nodes in the input layer, 5 nodes in the hidden layer, and 1 node that indict the output value of the cost value in the output layer. The learning rate was 0.01, and expectative error was 0.001. Then the neural network was programmed by software Matlab7.1. The average variance EMS was 3.25461×10-5, and training time was only 0.405 second. The error curve of network after 2386 training times are shown in Figure2 and Figure3.

5 Conclusions

Construction project cost estimation and prediction is crucial for construction firms to survive and grow in the industry. This paper presents to combine the rough set theory with neural networks and forecast the cost of construction project. The experiment results show the effectiveness of rough set approach as a data preprocessor for neural network. The reduction of information system has a great meaning for neural network in that reduction of attributes prevents overftting problem and save straining time. If applying rough sets (RS) theory to ANN. Its powerful abilities of attributes reduction and pattern identification can help to find useful information, simplify the processing of information so as to improve ANN's learning accuracy and convergence speed. The result also shows that this way is not only intuitive and effective, easy to understand, but also the result that forecasted is the closest to the actual cost. This method, compared with the BP neural network, has the advantages of highly accurate prediction and rapidly convergent speed.

Researching on pridiction of construction project cost is a complicated systems engineering. But there are still some problems left which need to be studied completely, such as how to make the construction project cost index system more scientific, the difierent ways of dealing with data, the combination of cost prediction module with monitoring and control systems and etc.

References

1. Cheng, M.-Y., Tsai, H.-C., Hsieh, W.-S.: Web-based conceptual cost estimates for construction projects using Evolutionary Fuzzy Neural Inference Model. Automation in Construction (2008)
2. Yu, W.-d., Lai, C.-c., Lee, W.-l.: A WICE approach to real-time construction cost estimation
3. Moon, S.W., Kim, J.S., Kwon, K.N.: Effectiveness of OLAP-based cost data management in construction cost estimate (2008)
4. Shi, H., Li, W.: The Integrated Methodology of Rough Set Theory and Artificial Neural-Network for Construction Project Cost Prediction. In: Second International Symposium on Intelligent Information Technology Application, December 2008, pp. 60–64 (2008)
5. Pawlak, Z.: Rough Sets-Theoretical Aspects of Reasoning about Data. Klystron Academic Publisher (1994)
6. Kennedy, J.: The particle swarm: social adaptation of knowledge. In: Proceedings of the 1997 International Conference on Evolutionary Computation, Indianapolis, pp. 303–308 (1997)
7. Ahn, B., Cho, S., Kim, C.: The integrated methodology of rough-set theory and articial neural-network for business failure prediction. Expert Syst. Appl. 18(2), 65–74 (2000)
8. Arditi, D., Suh, K.: Expert system for cost estimating software selection. Cost Engineering 33(6), 9–19 (1991)
9. Chau, K.W.: Application of a PSO-based neural network in analysis of outcomes of construction claims. Automationin Construction 16, 642–646 (2007)

Research of Corporate Credit for Anhui Province's Listed Companies Based on Computer Technology

Li Yang[1] and Malin Song[2]

[1] School of Management, University of Science & Technology of China, Hefei, Anhui, China
yangli081003@163.com
[2] School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, Anhui, China
songmartin@163.com

Abstract. Aiming at the shortcoming that obtained from comprehensive evaluation by principal component analysis, this paper points out the method of based on the EXCEL/VBA one-click technology to achieve principal component comprehensive evaluation. Namely, according to the algorithm of Eigenvalue and Eigenvector with the function in Excel-macro, in this article, an easy-to-implement solution to comprehensive evaluation by principal component analysis with VBA in Microsoft EXCEL is provided, which makes users be completely free from manual operation. It is raising the comprehensive evaluation by principal component analysis method practical. Meantime, as the dishonesty violations of listed companies occur from time to time, the credit risk of listed companies has become an issue related to the financial system stability. By understanding the concepts of credit and credit risk, a quantitative research has been carried out on the credit taking listed company in Anhui Province as an example on the premise of considering the availability of data. According to analysis of the company's credit system, combined with 46 financial statements of listed companies in Anhui Province, 12 indicators are selected for analysis and ranks with the company credit. Then through the cluster analysis of listed companies, these companies will be divided into three types. As results, some conclusions have been drawn from the Empirical Analysis that being work along both effective regulation and institutional innovation is an effective solution to the problem for the way of enhancing the company credit.

Keywords: Principal component analysis, comprehensive evaluation, excel, credit risk, cluster analysis.

1 Introduction

On the premise of the world economic integration and China's entry into the WTO, Chinese enterprises are facing with a growing number of challenges, at the same time developmental opportunities and operation risk industry grow in almost the same

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 39–46.
springerlink.com

percentage. All they make the credit risk of business being increasingly serious, which makes listed companies suffer from serious credit risk, including breach of contract, bad debt losses, trade and financial fraud and other phenomenon. In the face of such circumstances, Chinese enterprises must carefully analyze the reasons and take effective measures immediately to do good jobs in management and prevention of credit risk.

When the number of listed companies is increasing in Anhui Province, the enterprise scale is also expanding. In the last five years, there are 24 new listed companies in Anhui Province, an increase of 82.76 percent. The cumulative total capital stock of listed companies is from 16,426,000,000 shares by the end of 2002 to 27,480,000,000 shares by the end of 2007, an increase of 67.3 percent; the accumulated total market value is from 88,643,000,000 yuan by the end of 2002, jumping to 501,893,000,000 yuan by the end of 2007, an increase of 5.66 times.

At present, listed companies cover nearly the economic pillar industries and backbone enterprises, and play a more distinctive role in economic development of Anhui province. The quality of listed companies is having a significant increase while the number and scale are expanding in Anhui. However, presently, there are still some problems and shortcomings in the development of capital market in Anhui Province, in particular, the imbalance of development situation and the polarization of performance are relatively serious. In the middle of 2007, the sum profits of 5 companies (Maanshan Iron and Steel Company, Conch Cement, Tongling Nonferrous Metal Company, Anhui High Speed and Jianghuai Automobile Company) having the highest net profit, accounts for 79.83 percent of the total net profit of listed companies in the province. There are 33 listed companies whose earnings per share are lower than the average of the province; some companies are even in huge losses. Therefore, it is particularly important to carry on the credit analysis of listed company in Anhui Province.

2 The Related Theories

2.1 Credit and Credit Risk

In general it is believed that the credit is a credit behavior. The characteristic of this economic behavior is to pay on the conditions of the resumption, or to get on the conditions of taking the return as the obligation: and the reason why the lender provides loans is that he is entitled to getting interest; the latter is likely to borrow because of a commitment to pay interest. In other words, the basic meaning of credit is repayment [1].

Credit risk, also known as default risk [2], refers to the possibility of leading banks, investors or other traders to suffer losses, because the borrower, the issuer of securities or the traders are unwilling or unable to fulfill the conditions for a variety of reasons. Risk research is helpful for Chinese listed companies with times.

2.2 Research Methods

On the study of credit risk measurement methods, each has its own advantages, but each of them has its shortcomings. Although now many banks are still using 5C element analysis, such method is facing with two major challenges of consistency and subjectivity [3]. For a similar borrower, a different person in charge of credit may have different evaluation results using completely different standards, and their judgments are affected easily by the feelings and outside interference to make an analysis with large deviations. As a result, in recent years, financial institutions have gradually given up purely qualitative analysis of 5C element analysis.

It is no doubt that developing countries are lack of sufficient preconditions to implement credit risk management in using of modern means. Lots of the modern credit risk measurement models require the stock data of listed companies, while in China, credit system is not sound in the stock market, credit system has not yet been established, a serious lack of the historical data of the company in default or bankruptcy, it is not mature enough to use modern credit risk measurement model. So this article is using the traditional credit risk measurement method - principal component analysis to evaluate the credit risk of listed companies [4]. Topically, Principal component analysis (PCA) is basically in six steps.

2.3 Index Selection

This article chooses 12 variables from financial indicator variables, where rate of return on total assets reflects the profitability; accounts receivable turnover, inventory turns and total asset turnover reflect the operating capacity; return on shareholder equity and income growth of main business reflect the growth capacity; current ratio, quick ratio and shareholder equity reflect the repayable ability; fixed ratio reflects the financial structure.

3 Realization of Principal Component Analysis

Principal component analysis is a statistical method widely used in various fields, which is a few less representative of the composite index to replace more of the original index. The comprehensive evaluation by principal component analysis can eliminate the impact of the various indexes with different dimension, but also can eliminate the correlation and information overlaps lying in multi-indexes. In particular, it has overcome the problem of the target weight by manual interventions. In general, we can select a few principal components as the composite index variables instead of the original target first when the total variance explained has contributed up to a given threshold. Secondly, with the principal component, we can calculate the total scores and ranks of them. So the method is simple in practice with intuitive meaning [5]. However, the application and development of this method are hindered as it is very cumbersome and inconvenience to calculate principal component manually. SPSS, which is the extensive use of statistical software in social statistics can only calculate factor analysis and can not calculate principal component. So it is disadvantage to the settlement of the issue [6]. It is easy to resolve the problem of calculating the principal component and comprehensive evaluation with using Excel VBA macro program. Thus, users are liberated from manual operation with increasing problem-solving efficiency accordingly.

Macro [7] program is a command set with a series of orders and directives, which is the VBA procedure code. Microsoft VBA is a shared common language automated in its development application. A combination of keyboard input makes a simple command when we create a macro. Calculation steps and algorithm based on Excel VBA are as follows: Step 1 is index selection; Step 2 is index standardization; Step 3 is calculating correlation coefficient matrix or covariance matrix; Step 4 is solving eigenvalue; Step 5 is solving eigenvector; Step 6 is determination of the number of principal components by calculating variance explained and total variance explained. Step 7 is calculating scores of principal component; Step 8 is calculating comprehensive evaluation; Step 9 is Results output.

4 Analysis of Corporate Credit Risk

Being in the central of China, Anhui Province’s domestic economic environment is in a disadvantage. So it is particularly important of credit analysis for listed companies in Anhui Province. The company's credit is an important basis for investors.

At present, in Anhui there are a total of 56 listed companies, with the number ranking 10 in China. The 56 listed companies issue a total of 61 stocks, refer to machinery, textile chemicals, automobiles, building materials, energy, transportation, household appliances, metallurgy, electronics, medicine and many other fields. All data in this article came from the financial reports of listed companies published by china securities regulatory commission (SRC).

To choose four-quarter data of 46 listed companies in Anhui Province from June 2007 to March 2008, which can be seen in table 2. Anhui Gujinggong B shares and Huangshan B shares which are two foreign-owned shares listed in China and several others listed for a shorter time are not in this analysis. According to the selected indicators of the index system to analyze the credit of 46 listed companies.

When EXCEL software is used to compute, 5 main elements could be got which are Y1, Y2, Y3, Y4, and Y5. They can explain up to 85.22 percent of the total, the result being better. Table 3 shows the situation of main components explaining total variance of original variables. Table 4 gives the approximate linear expression which uses the principal component to express the original standard variables.

The first principal component is determined by rate of return on total assets, earnings per share, growth rate of shareholder equity and proportion of shareholder equity, their loads are 0.93, 0.84, 0.86, 0.63, the principal component reflects the profitability, solvency and growth ability of enterprise.

Because the return on total assets including net assets and liabilities reflects the overall profitability of all the assets, it is used to evaluate the enterprise's overall profitability in use of all the assets, and is also an important indicator which evaluates the efficiency of the enterprise operating assets. And proportion of shareholder equity is the ratio of shareholder equity and total assets, it should be modest. If the ratio is too small, which shows the enterprise is over-indebtedness, it easily weakens the company's ability to withstand the external risk, while it is too large, which means that company does not actively use financial leverage to expand the operation scale.

The second principal component is determined by gross margin and inventory turnover rate, their loads in the main components are 0.85, 0.65, and the principal component reflects the profitability and operational capacity of the business. A direct reflection of the gross margin is the price difference level of all sales, the largest category sales and some sales of the enterprise, being the basis of rationality of enterprises' running effect and the price setting, by which it can help us to distinguish between fixed and variable costs to a certain extent. Inventory turnover rate reflects the level of inventory management, affecting short-term solvency of enterprises, is an important content of the whole business management. In general, the faster speed of inventory turnover means the lower occupancy levels of inventory and the stronger liquidity, the inventory can be converted to cash or account receivable faster. Therefore, improving the inventory turnover rate may increase the liquidity.

The third principal component is determined by fixed ratio, and its load is 0.76. When business assets are converted to cash, fixed assets are often difficult to be done even with more discounts, and intangible assets also have the risk of difficult liquidation and depreciation. Fixed ratio shows the number that is not easy to realize the total fixed assets of enterprises. This is a conservative indicator to measure the long-term solvency.

The fourth principal component is decided by quick ratio, with its load 0.66, which shows the principal component reflects the solvency of enterprises. Directly reflecting the level of short-term solvency of enterprises, quick ratio is a supplementary for current ratio, and is more intuitive and more credible than current ratio.

The fifth principal component is decided by accounts receivable turnover rate, whose load in the main components is 0.92. The principal component reflects the operational efficiency of enterprises and the length of time required that needed from the right to accounts receivable to its conversion, and it is the expression of the average times of converting accounts receivable into cash within the year. If the turnover rate is too low, it will impact the short-term solvency of enterprises.

As the first principal component which reflects the profitability, solvency and growth ability to the business, its correlation coefficient with each company is higher, to rank the companies in use of the first principal component, the profitability and

Table 1. Acronym of variables

Variables	Acronym
Equity Ratio	ER
Rate of Return on Total Assets	RRTA
Return on Equity	RE
Per Share Earning Ratio	PSER
Working Capital Ratio	WCR
Quick Ratio	QR
Rate of the Rights and Interests of Shareholders	RRIS
Inventory Turnover Ratio	IRAR
Turnover Rate of Accounts Receivable	TRAR
Rate of Gross Profit	RGP
Growth Rate of the Rights and Interests of Shareholders	GRRIS
Fix Ratio	FR

Table 2. Acronym of anhui province's listed companies

Listed Company	Acronym	Listed Company	Acronym
ST Keyuan	STKY	Kinmen Stock	JMGF
Anhui Feiya	AHFY	the Gold Seeds of Wine	JZZJ
Heli Anhui	AHHL	Jingda Shares	JDGF
Anhui Water	AHSL	Science innovation	KDCX
Ankai Bus	AKQC	Leiming Science and Chemistry	LMKH
Seiko Yangtze River	CJJG	Chemical Industry of Six Countries	LGHG
East Cao Shares	CDGF	Maanshan Iron and Steel shares	MGGF
Franshion Science and Technology	FXKJ	Meiling Electrical Appliances	MLDQ
Fengle Seed Industry	FLZY	Firewood Dynamic	QCDL
Fenyuan Biochemical	FYSH	Three Best Science and Technology	SJKJ
Fengyuan Pharmaceutical	FYYY	Mountain Eagle Paper Industry	SYZY
Gong Ancient Wine	GJGJ	Four-hit Electronics	SCDZ
Guofeng Plastic	GFSY	Copper Peak Electronics	TFDZ
Guotong Pipe Industry	GTGY	Tongling Nonferrous Metal Company	TLYS
Conch Cement	HLSN	Wanneng Power	WNDL
Conch profile	HLXC	Anhui High Speed	WTGS
Hefei Department Store	HFBH	Wanwei Gaoxin	WWGX
Hefei Sanyo	HFSY	Wuhu Port	WHG
Hengyuan Coal and Electricity	HYMD	Xin Branch Materials	XKCL
Huamao shares	HMGF	Xingma Automobile	XMQC
Huaxing Chemical Industry	HXHG	Shares of Yongxin	YXGF
Huangshan Tourism	HSLY	Middle Tripod Stock	ZDGF
Jianghuai Automobile	JHQC	Middle Steel Day Source	ZGTY

solvency of Anhui High Speed, Conch Cement, Wuhu Port are better, Feiya Anhui lower, ST Keyuan worst in 46 listed companies.

Wantong company belongs to the transportation and warehousing industry, "the top 100 best investor relations management in China in 2006" enterprises released, the company won the eighth, it shows that the company's credit risk is low. The production and sales of Conch Cement has been ranked first in China for 11 years, which is the largest supplier of cement and clinker in Asia. Today the company's three new cement clinker production lines at a daily output of 10,000 tons represent the most advanced level in cement industry in the world. Wuhu Port is a transport infrastructure sectors with total ranked as the No. 25, experts are optimistic about the stock with the growing trend.

Table 3. The situation of principal components explaining total variance of original variables

Principal Component	Total Variance	Variance Contribution	Accumulative Variance Contribution
1	4.05	33.78	33.78
2	2.11	17.57	51.35
3	1.86	15.51	66.86
4	1.20	10.00	76.86
5	1.00	8.36	85.22
6	0.61	5.07	90.29
7	0.42	3.47	93.75
8	0.34	2.84	96.60
9	0.18	1.50	98.10
10	0.13	1.06	99.16
11	0.09	0.71	99.87
12	0.02	0.13	100.00

Table 4. Factor loading matrix

Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
ER	0.60	-0.61	0.20	0.32	-0.01
RRTA	0.93	0.12	-0.09	0.10	0.10
RE	-0.79	0.34	0.05	0.07	0.07
PSER	0.84	-0.07	0.03	0.16	0.12
WCR	0.24	0.36	-0.78	-0.19	-0.03
QR	0.02	0.29	-0.53	0.66	0.29
RRIS	0.63	0.29	-0.15	-0.64	-0.18
IRAR	0.24	0.65	0.55	0.07	-0.05
TRAR	0.03	-0.07	0.07	-0.37	0.92
RGP	0.39	0.85	0.02	0.18	0.00
GRRIS	0.86	-0.28	-0.11	-0.01	-0.07
FR	0.29	0.26	0.76	0.02	0.09

The principal components originated from the above-mentioned indicators variables cannot fully reflect the characteristics of kinds of corporate credit risks, it would take a further improvement in the indicators to remove the low information extracted indicators, but to add the enterprise risk indicators which are not included in the above-mentioned indicators, the twice analysis will be a better analysis of business risks, which is more helpful to investor.

To use system cluster method for cluster analysis[10], a cluster tree will come to be seen, the 46 companies are divided into 3 categories: Anhui High Speed represent a class; ST Keyuan is a class, from which the profitability and operational capacity are poor; Wuhu Port, Leiming Science and Chemistry, and Conch Profile are a class, from which these companies have certain advantages in the short-term solvency; other companies are a class, the values of these indicators including inventory turnover rate, gross margin, fixed ratio have no significant differences. The results of principal component analysis are basically the same.

Acknowledgments. This article is funded Natural Science Research Projects of Universities of Anhui Province (KJ2009A59). Huiying Zhang, Xiaojun CAI, Chao LI and Bei HUANG, do numerous works for this paper.

References

1. Martin, D.: Early Warning of Bank Failure: A Logit Regression Approach. Journal of Banking and Finance 1(3), 249–276 (1977)
2. Ling, Z., Jialin, Z.: Development of Credit Risk Measurement Methodology. Forecasting 19, 72–75 (2000) (in Chinese)
3. Yatao, Z.: A comparative Analysis of the model for Measurement of Credit Risks. Journal of Shanxi Finance and Economics University 24, 107–108 (2002) (in Chinese)
4. Li, B.: Evaluation Model of Credit Risks of Chinese Listed Enterprises Based on Main-Composition Analysis. Journal of Xi'an University of Technology 21, 219–222 (2005) (in Chinese)
5. Zhengming, Q., Weiyan, C.: The appraisal of indicators of industry economic results and a case study in terms of method of principal components. Statistical Research (7), 49–52 (1999) (in Chinese)
6. Jiwei, X., Deyin, F.: The redevelopment based on SPPS to find principal component. Statistical Research (4), 73–75 (2006) (in Chinese)
7. Chao, L.: Gray forecasts: Excel / VBA programming easy to performance. Statistics & Information Tribune (5), 72–75 (2004) (in Chinese)
8. Shumei, W.: Applications of Excel in the related calculations of matrix. Computer Knowledge and Technology (1), 12–47 (2007) (in Chinese)
9. Iiulin, Y., Xuesong, R.: Multivariate statistical analysis, pp. 162–166. China Statistical Publishing House, Beijing (1999) (in Chinese)
10. Dun, L., Pei, H., Ping, H.: Evaluation Model of Scientific Diathesis Based on AHP and Clustering Analysis and Its Empirical Study. Science & Technology Progress and Policy 24, 66–69 (2007) (in Chinese)

Evaluation of Industrial Parks' Industrial Transformations and Environmental Reform Actualized by AHP Based on MatLab Software

Malin Song

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics,
Anhui Bengbu, P.R. China
songmartin@163.com

Abstract. In order to explore the study of regional industrial parks at home and abroad and objectively analyze problems that China industrial parks are facing, the symbiotic system along with technical and policy aspects of eco-industrial parks will be analyzed and observed by reviewing the researches on eco-industrial parks. Eco-industrial parks' construction in china should be continued by focusing on their industrial transformations and environmental reform processes under the premise of enhancing competitiveness of the existing industrial parks established based on circular economy and theories of eco-industries. It is the only way for China to build a model environment-friendly and resource-saving society by establishing circular economy based on particular mode of symbiotic manufacturers cluster network and the industrial transformations and environmental reform for its economic and technological development zones and high-tech industrial parks. The feasibility analysis, reform programs and reform models of transformation for industrial parks are focus problems scholars pay close attention to. At the same time, research of industrial parks' industrial transformations and environmental reform based on Complex Network technology will provide effective support which could speed up the process of transformation. Then a evaluation index system for industrial parks' industrial transformations and environmental reform is built which includes 3 levels index and Analytic Hierarchy Process based on MatLab Software is used to empowerment, by which it provide an effective way for performance evaluation of industrial parks' industrial transformations and environmental reform.

Keywords: industrial park, industrial transformations and environmental reform, complex network, evaluation index system, analytic hierarchy process, MatLab Software.

1 Introduction

Confronted with continuing deterioration of the ecological environment and people's growing concern about environmental issues, terminal control mode of environmental pollution based on the traditional linear industrial system, that is, using raw materials

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 47–54.
springerlink.com

free from the ecosystem and emitting a large number of by-products into the ecosystem in the form of unnecessary waste, has been difficult to support sustainable development time by time. The practice of some developed countries showed that the transformation from conventional economic activity to polycyclic process, involving the process of "resources - products - renewable resources - recycled products", will help to reduce resource consumption and waste emission at the maximum from the whole process of production to consumption [1].

Through the research of industrial symbiosis in Fort Danmaikalun, foreign scholars found a direct way of the practice for industrial ecology- Eco-Industrial Parks (EIPs) [2]. Since then, the practice of eco-industrial parks have been carried out around the world and almost all western countries and some developing countries in Southeast Asia set off a boom of the constructing of Eco-Industrial Parks. Up to 2006, the United States have started 40 EIPs at least. There have been similar 60 EIP projects at any rate in Asia, Europe, South America, Pacific Island, Africa and other places, among the rest Japan has established more than 30 EIPs. All these projects are involved with a variety of industries and have different characteristics [3].

Currently, ecological environment has become more fragile and continuous deterioration of ecological environment has not yet been under effective control [4]. The transformation from its economic and technological development zones and high-tech industrial parks to the third generation of industrial parks - Eco-Industrial Parks will establish special symbiotic manufacturers cluster networks essentially based on circular economy model, by which the core is economic benefit and basis is environmental benefits. This will lead the way to achieve a model of China's environment-friendly and resource-saving society.

2 Patterns of Industrial Transformations and Environmental Reform for Industrial Parks

Industrial Parks in China are some relatively independent regions developed with the reform of Chinese economic system. They are opening to outside world which guides the development of other areas, which have become active points of economic development in China. Chinese Industrial Parks, in accordance with the nature of the industry, can be divided into various types, from which they includes economic and technological development zones, high-tech industry development zones, bonded areas, export processing zones, border economic cooperation zones, tourist resorts. They have acted as an important driving force for economic development and outstanding carriers for regional development in China [6]. At the same time, their development has also brought a lot of problems [7-9].

2.1 Transformational Patterns for Industrial Parks

The author believes that industrial parks can be seen as complex networks under the prerequisite of meeting targets of traditional industrial parks' various stakeholders. Application of Small-World Network [10] and Scale-Free Network [11] technology and analysis from micro-level of industries and manufacturers would gain dynamic integrated solutions for industrial transformations and environmental reform.

Scholars found that a large number of real networks have small- world networks [10]. It is simple and intuitive to say that the small-world effect means that the average distance of networks is small; A more rigorous description is that a network had Small-world effect or effects of ultra-small world, if its average distance will increase by a speed of logarithm [10] or slower [12-13] than the speed of node growth, under a premise that the average peak time of a network is fixed. More reasonable layouts of distribution networks of industrial parks should be in line with the principle of small-world networks, of which the average distance of space is randomly slightly larger than the same number of links and nodes of the network [10]. Special attention should be paid to that, in the actual national distribution network of industrial parks, their different roles are not simply a quantitative difference. Some important industrial parks in China that are state-level Economic and Technological Development Zone play decisive roles. In response to environmental issues such as energy-saving and pollutions reduction, they should perform characteristics of self-organized criticality. The concept of self-organized criticality (SOC) was first proposed by Bak and others [14-15]. Bak imagined that the process of adding sand grains to sand slowly, although the addition of some new sand isn't likely to have any impact on the sand as a whole, to a degree it can have a similar avalanche effect which changes the shape of the sand. Inspired by this thinking, Bak proposed self-organized criticality concept which refers to a specific system, some conditions evolving on their own in particular. Through the process of dissipation of kinetics, it eventually reaches a critical state that the next perturbation will bring response with different sizes and it sizes submits to the power-law distribution. Though there is no creation of a more appropriate model to quantify these measures used for environmental issues industrial parks face, the frequency distribution of environmental issues should show characteristics of the power-law, in different measurement methods, which is in line with its scale-free in nature.

Therefore, we should pay attention on industrial parks that have various advantages for development in China and then stimulate the industrial transformations and environmental reform of other industrial parks. It should also be insisted that the providing of robustness for entire system of industrial Park cannot be exerted some industrial parks' disadvantages.

2.2 Formation of Symbiotic Effect in Industrial Parks

The key of industrial parks' industrial transformations and environmental reform is the formation of a symbiotic effect for manufacturers. Symbiotic effects can be divided into: by-product exchange effects, cluster effects, green effects and positive external effects. These four sub-effects can play roles in four different aspects in which the cost of manufacturers is reduced or the differentiation is formatted, which constitute direct sources of competitiveness of industrial parks [16-18].

3 Evaluation of Industrial Transformations and Environmental Reform Actualized by AHP Based on MatLab Software

Three-stage optimization model and economic evaluation model based on recycling economy paradigm could be used to transform traditional industrial parks [19].

Small-World Network and Scale-Free Network technology could be used, combining with time value and capital, to establish optimization model of symbiosis industrial chain; game model based on cooperation of interests' distribution of industrial symbiotic chains is important to stabilize symbiotic industrial chains. The introduction of Complementarities of symbiotic industrial chains and promotion of affiliated enterprises of eco-related industries are also important.

Industrial transformations and environmental reform is a systematic project and must be put in practice on long-term development to promote ecological modernization fundamentally. Systemic analytic methods can be used, mainly from the economic development efficiency, to establish the secondary targets of evaluation index system for industrial transformations and environmental reform of industrial parks and to construct the third evaluation index.

Analytic Hierarchy Process (AHP) method is a structured technique for helping people deal with complex decisions. Rather than prescribing a "correct" decision, the AHP helps people to determine one. Based on mathematics and human psychology, it was developed by Thomas L. Saaty in the 1970s and has been extensively studied and refined since then. The AHP provides a comprehensive and rational framework for structuring a problem, for representing and quantifying its elements, for relating those elements to overall goals, and for evaluating alternative solutions. It is used throughout the world in a wide variety of decision situations, in fields such as government, business, industry, healthcare, and education [20].

The evaluation of Industrial Parks' transformation is a multi-scale, multi-objective evaluation problem that involves subjective and objective criteria. The AHP method based on MatLab software divides various indicators of the transformation evaluation of the eco-industrial park involved into a number of levels. On the same level, with every indicator compared with each other, some subjective value judgments will be expressed and fixed by quantitative forms. It provides a scientific and effective method for the comprehensive measure of industrial parks' industrial transformations and environmental reform. So the evaluation index system of industrial transformations and environmental reform is composed of a multi-level indicators group. Analyzing empowerment MatLab-based, it determines matrix structure through the experts' opinion and determines weights through some form of mathematical treatment.

Therefore, combining with the empowerment of experts and quantitative methods to determine the weight, AHP based on MatLab is more scientific. According to the tables, a series of indicators are established to measure the structure, including: first-level indicators - A; secondary indicators - B1 ~ B8; the individual indicators index layer - C11 ~ C83; and then by using scale 1 to 9, matrixes are formatted based on the comparison of the goal layer. By using MatLab software, consistency test and single-level ranking for the matrix are determined, which in turn get results finally. Acronym of categories' denomination can be seen in Table 1 to 4; the results can be seen in Table 5. Therefore, the performance evaluation of industrial parks' industrial transformations and environmental reform is actualized by complex network analysis based on the environment-resources-ability dynamic performance evaluation index system and weight empowerment by AHP. From the aspects of governments, parks' managers, symbiotic industrial chains, and manufacturers, the static and dynamic way of transformational process can be explored and judgments can be decided.

Table 1. Acronym of categories' denomination

Denomination Of Category	Acronym
First-Level Indicators (Object Hierarchy)	OH
Numbering For Object Hierarchy	NOH
Secondary Indicators (Rule Hierarchy)	RH
Numbering For Rule Hierarchy	NRH
Third Grade Indicators (Indicators Hierarchy)	IH
Numbering For Indicators Hierarchy	NIH
Indicators' Weights	IW

Table 2. First-level indicators

Object hierarchy	OH
Evaluation of industrial transformations and environmental reform	A

Table 3. Secondary Indicators (Rule Hierarchy)

RH	NRH
Efficiency of economic development	B1
Situation environmental protection	B2
Regional energy saving	B3
Supervision of environmental protection	B4
The situation of capital market	B5
Science and technology education	B6
Manpower resource utilization	B7
Industrial agglomeration	B8
Efficiency of economic development	B1

Table 4. Comparison matrix' consistency test

Comparison matrix	CI	CR
A	0.0721	0.0511
B1	0.0193	0.0332
B2	0.0652	0.0693
B3	0.0046	0.0079
B4	0.0193	0.0332
B5	0.0000	0.0000
B6	0.0198	0.0220
B7	0.0853	0.0761
B8	0.0018	0.0032

Table 5. Indicators' Weights

IH	NIH	IW
GDP per capita	C11	0.0962
Fixed assets investment per capita	C12	0.0158
Total retailing of social consuming goods per capita	C13	0.0390
1/ industrial waste water discharge per square kilometer	C21	0.0095
1/industrial fumes emission per square kilometer	C22	0.0160
1/SO_2 emission quantity per square kilometer	C23	0.0294
1/industrial solid waste per square kilometer	C24	0.0041
1/annual water supply quantity per capita	C31	0.0041
1/power consumption of total output value unit area	C32	0.0134
1/energy consumption of total output value unit area	C33	0.0074
1/per break money for environmental pollution accident	C41	0.0107
Proportion of clean-process company	C42	0.0018
The number of sites for environmental monitoring per square kilometer	C43	0.0043
Asset-liability ratio	C51	0.0075
Hedging and proliferating ratios	C52	0.0075
Financial GDP per capita	C53	0.0226
The number of professional and technical personnel every 10000 people	C63	0.1904
The proportion of educational expenditure account for expenditure	C61	0.0692
Proportion of the college teacher numbers every 10000 people	C62	0.0316
The proportion of science expenditure account for expenditure	C64	0.1073
All labor productivity	C71	0.0480
Ratio of professional and technical person account for total population	C72	0.0151
The number of people affected by college education each thousand	C73	0.0205
Average salary	C74	0.0032
Ratio of urban population account for total population	C75	0.0060
The proportion of the number of high-tech manufacturing employee	C81	0.1427
The proportion of the number of modern service industry employee	C82	0.0506
Profit of the state-owned and designed size non-state-owned enterprises per capita	C83	0.0269

4 Conclusions

Industrial transformations and environmental reform of Industrial parks which involve eco-industrial materials, energy and information flows, implementation strategy, mutual relations of members, support mechanisms, operational efficiency and competitiveness, is bound to expand the areas of study of industrial ecology. The general paradigm of economic growth negatively correlating to environmental degradation could be given. It is the only way that could achieve win-win for social, economic and environment and promote the process of ecological modernization. At the same time, research in the theory of small world and scale-free network technology and their creative applications of industrial parks' industrial transformations and environmental reform will expand the areas of complex network's research and applications. It is considerable that in the face of objective contradictions between governments' subjective enthusiasm of industrial transformations and environmental reform and unable achievement of desired objectives, systematically discussion of ecological characteristics for successful transformation and their development programs would be the direction for further research. At the same time, empirical analysis of the relationship between environment and development for human society, being the actual solutions of inconsistency between unlimited economic growth desires of mankind and limited capacity of natural environment to support them, will provide strong support for conservation civilization construction.

AHP Based on MatLab Software provides efficient quantitative method to do it. The consistency of the judgment matrix should be tested when using the method of AHP to solve problems. The value of the consistency index of High0-Order judgment matrix involved in the algorithm is not easy to acquire, which sets back the putting of the method of AHP into practice. There gives the algorithm of the mean random consistency index based on studying the method of AHP, and gives its implementation based on MatLab software.

Acknowledgments. This article is funded by the Philosophy and Social Sciences Projects of Anhui province (AHSK07-08D25) and Natural Science Research Projects for Universities of Anhui province (KJ2008B135). Dr. Xiaojun CAI, from the Beijing Municipal Commerce Bureau, gives the author a lot of help. Yejun Yang, Xin LI and Bei Huang, from Anhui University of Finance and Economics, also do numerous works.

References

1. Song, M.: Empirical Research on Relations of China's Energy Consumption to Its Industrial Development. In: ICIII 2008, Taibei, China, December 19-21, pp. 104–107 (2008) (in Chinese)
2. A Report to Asian Development Bank: A Handbook for Eco-Industrial Parks in Asia Developing Countries (2006), `http://www.Indigodev.com`
3. Lowe, E.: A Handbook for Eco-Industrial Parks in Asia Developing Countries. A Report to Asian Development Bank (2001), `http://www.Indigodev.com`
4. Xinhuanet: Interpretation of China's environmental communiqué in 2007: happiness and worries after 'inflection point' (2008), `http://news.xinhuanet.com/` (in Chinese)
5. Ministry of Commerce of the P. R. China: Development and ideas of State-level economic and technological development zones. People's Forum(07/B), 7–8 (2006) (in Chinese)
6. Ding, F.: A Comprehensive Study on Management Patterns in Chinese Economic and Technological Development Zone. Huazhong University of Science and Technology, 1–5 (2004) (in Chinese)

7. Guo, Z., Hui, W.: Literature Review Of Chinese Urban Development Zones. City Planning Review 29(8), 51–58 (2005) (in Chinese)
8. Zhang, T.: Alert New Sources of Pollution: high-tech pollution. Science & Technology Progress and Policy 10, 10–13 (2000) (in Chinese)
9. Lan, A., et al.: Classes of small-world networks. Proceedings of the National Academy of Sciences of the USA 97(21), 11149–11152 (2000)
10. Cohen, R., Havlin, S.: Scale-free networks are ultrasmall. Physical Review Letters 90(5), 058701 (2003)
11. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998)
12. Zhou, T., et al.: Topological Properties of Integer Networks. arXiv: cond-mat/0405258 (2004)
13. Bak, P., Tang, C., Wiesenfeld, K.: Self-organized criticality: An explanation of the 1/f noise. Physical Review Letters 59(4), 381–384 (1987)
14. Bak, P.: How Nature works: The science of Self-Organized Criticality. Corpnicus, New York (1996)
15. Jenson, H.J.: Self-Organized Criticality. Cambridge Univ. Press, Cambridge (1998)
16. Song, M.: On the Causation of Eco-Industrial Park Manufacturers' Low Cost and Cost Diminution. Journal of Shandong University of Science and Technology 4, 69–72 (2008) (in Chinese)
17. Song, M.: A Technical and Economic Evaluation Model of Eco-Industrial Parks Project Based on Recycle Economy. International Institute of Applied Statistics Studies, Yantai, China, pp. 1398–1402 (2008)
18. Liu, X.-p., Lei, D.-a.: On the Connotation and Meaning of Externality Theory. Journal of the Northwest Normal University (Social Sciences) 39(3), 72–75 (2002) (in Chinese)
19. Song, M.: The Research Status Quo and Consideration of Industrial Parks' Ecological Transformation. In: FITME 2008, Leicestershire, UK, pp. 313–316 (2008)
20. Guo, J.-y., Zhang, Z.-b., Sun, Q.-y.: Study and Applications of Analytic Hierarchy Process. China Safety Science Journal 1, 83–85 (2005) (in Chinese)

Appendix: MatLab Program

This article uses MatLab software to actualize the comparison matrix' consistency test, and single-level sequencing as well. Due to the limitations of space, these programmers can not be listed thoroughly. Readers can ask them via the author's email if necessary.

A Model Integrated Development of Embedded Software for Manufacturing Equipment Control*

Zhumei Song[1] and Di Li[2]

[1] Department of Information Control and Manufacturing, Shenzhen Institute of Information Technology, Shenzhen, China
mesong@163.com
[2] College of Mechanical Engineering, South China University of Technology, Guangzhou, China
itdili@scut.edu.cn

Abstract. The state-of–art approaches for embedded control software development are costly. It is difficult for manufacturing domain engineers to develop equipment control software in general development environment. Therefore, we propose domain-extending technology of embedded software for equipment control. Model-integrated computing approach is introduced to build meta-model for manufacturing equipment control. Graphical domain specific modeling language is interpreted from the meta-model. Engineers in equipment control domain can use this modeling language to rapidly build application model for different embedded control system. A simple example, signal process, indicates that this technology can reduce development and maintenance costs significantly. Domain engineers can develop their own embedded system with facility. The architecture for embedded control system is explored in design space. It is viable to built modeling language for equipment control software.

Keywords: Embedded System, Model Integrated Computing, Software Development.

1 Introduction

As manufacturing equipments are required higher and higher performance, control system comes through relay, PLC and industrial PC (IPC). Whereas, the control system based on IPC use commercial operating system, which is resource consuming, costly and difficult to achieve high reliability. Embedded control system become to the main stream of control system. Embedded system is application-oriented and HW/SW pruned, performed by interacting components. It is uniquely suitable for implementing multifaceted performance, i.e., timing, reliability, robustness, resource consumption, et al.

* This work is supported by the 3rd Program of Shenzhen Institute of Information Technology (QN-08009) and the 11th High Tech Fair Program of Shenzhen Institute of Information Technology (GJH11-ZS005).

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 55–61.
springerlink.com

Current embedded control system development use generic language and environment, which is difficult for domain engineers to design their own application and maintain their system evolvement. In addition, many of the problems, i.e., functional and non-functional performance, must be considered at system design time.

One approach, model integrated computing (MIC), addresses these problems by extending the scope and use of models [1]. This technology facilitates system engineering by providing domain specific modeling language (DSML). Multiple-aspect models are created and evolved based on DSML. Embedded system is automatically synthesized through the models used in domain specific field. This approach has been successfully applied in several different applications, e.g., military affairs [2], automotive manufacturing [3], avionics [4], chemical process [5], et al.

Karsai [5] summarized the model integrated development process of embedded systems. The key parts of this process include:

1) Meta-model: is a domain independent abstraction from which domain specific modeling language is interpreted.

2) Model interpreter: generates executable programs from domain models and configuration files for tools.

In this paper, we introduce MIC to manufacturing equipment control software development. The paper has an overview on MIC, and then discusses the software architecture of the embedded control system. Next, a simple example shows the model based development process, which indicates that it is viable to facilitate embedded control system development by building domain specific modeling language for equipment control.

2 Overview on MIC

MIC has been developed over a decade at Vanderbilt University, USA, for embedded system development. The greatest advantage of this approach achieves in application domain includes [6]:

1) The specification allows hierarchical decomposition. System can be divided into subsystems and encompasses multiple aspects, e.g., software, hardware and environment.

2) Modularization is feasible. In MIC, models capture the relationship between interacting components and are formally defined.

3) Software and its associated physical system can evolve together. Meta model captures a class of problems in specific domain instead of a specialized application. Graphical DSML makes it easier for domain engineers to represent their understanding of the entire control system. Model interpreters automatic synthesis the applications. When the external requirements change, domain engineers can modify system models and generate the updated software.

The multi-graph architecture (MGA) [7] is the framework for MIC. MGA employs two-level modeling process, as shown in Fig. 1. Details can be found at Ref.[7]

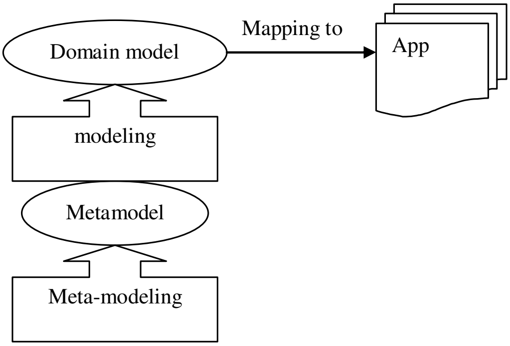


Fig. 1. System development process using MIC

The whole development process includes:

1) Meta-modeling. Meta-models use meta-modeling language based on UML/OCL to specify the objects, attributes and relationships, in meta-modeling environment. Meta-interpreter maps the meta-models to graphical DSML with well defined syntax and semantics.

2) Modeling. The models are constructed for a certain application utilizing DSML specified in meta-level process. Complex models can be built by composing simpler components.

3) Mapping. Interpreters map the domain models to frame code or data structure for tools. This technology increases programming productivity as demonstrated in section 4.

Programmers implement the concrete code for the application components. Executable application is created by the compiler.

The difference between MIC and model driven architecture (MDA) is: MIC emphasizes domain specific modeling technology and utilizes UML/OCL as meta-modeling language, MDA addresses application specific modeling and utilizes UML directly as modeling language.

3 System Architecture

In attempting to build the high level meta-model for equipment control domain, it is important to explore the system architecture in design space.

The whole system is divided into several subsystems or components, i.e., data collection, data analysis, control strategy, alarm & report, communication, et al. The data-centric environment provides the interoperation between subsystems. Control strategy subsystem provides control algorithms, methods and rules for specific events. The system employs these control strategies to determine what behavior should be taken, e.g., trigger an alarm mode, feed back the power, writing a log file, et al. Data analysis subsystem conducts data mining, display data trends and provides useful potential information for different people, i.e., engineers and managers. Database stores the basic data needed by other interacting components and the processed diagnostic data. The RTOS manages communication among the components of the software infrastructure. Communication component responds to remote request and transmit data according to specified protocol. For hard real-time communication, synchronous or time-triggered traffic is employed to guarantee its timeliness. For soft real-time communication, asynchronous event-triggered traffic is used to implement its flexibility.

4 An Example for System Development Using MIC

A core tool in MIC is the generic modeling environment (GME), a domain specific modeling environment that can be configured from meta-level specifications [9]. The successful use of GME for DSML development requires detailed knowledge of the underlying domain and programming experience. ESML [4] has been built for avionics application.

For equipment control system discussed in section 3, signal process module is absolutely necessary. In this section, application development from meta-level construction for signal process module is described as a simple example to illustrate how MIC facilitates the whole development process. See Ref. [10] for more details. Consequently, the viability of building DSML for equipment control is indicated.

4.1 Meta-modeling

The meta-model for signal process is shown in Fig. 2. The objects, relationships, attributes and constraints are specified utilizing meta-modeling language. The abstract signal object is inherited as InputSignal and OutputSignal atoms. The connection

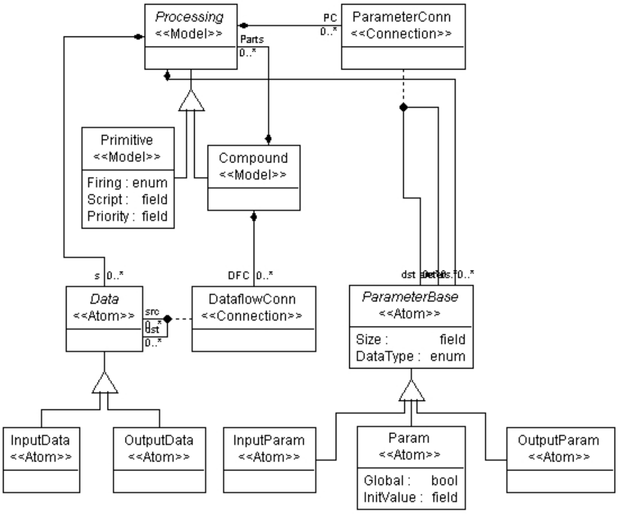


Fig. 2. Signal processing meta-model

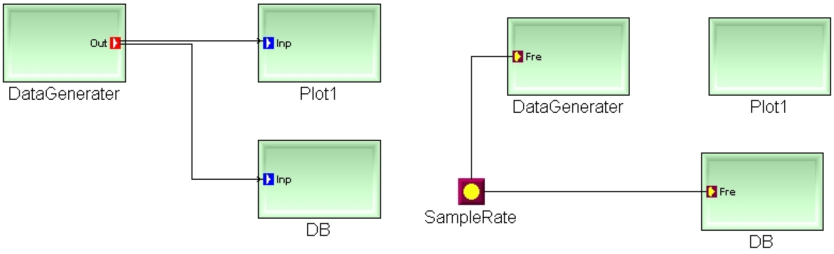
between them is defined as DataflowConn. Similarly, ParameterBase is inherited as InputParm, Param and OutputParam atoms. The connection between them is defined as ParameterConn. There are two processing models: PrimitivePart and CompoundPart. CompoundPart model can encompass PrimitivePart models, but PrimitivePart model can only encompass atoms and their connections. Two aspects are specified: SignalFlowAspect and ParameterAspect. PrimitivePart, CompoundPart, InputSignal and OutputSignal are visible in SignalFlowAspect. PrimitivePart, CompoundPart, InputParm, Param and OutputParam are visible in ParameterAspect.

4.2 Modeling

DSML for signal process is created by meta-interpreter from the meta-model built in subsection A. Then, domain engineers can construct models for their applications. Here, two application models are built to illustrate the rapid construction process using DSML.

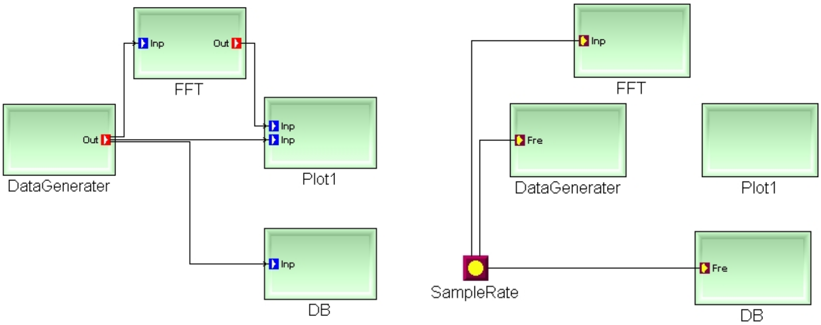
One simple model is created as shown in Fig. 3. This application generates a series of random data, stores them in database and plots the curve.

Another model is built based on the previous model. A FFT block is added specified, as shown in Fig. 4. The raw data and the processed data are both transmitted to plot model for display.



(a) Signalflow aspect (b) Parameter aspect

Fig. 3. Signal processing model



(a) Signalflow aspect (b) Parameter aspect

Fig. 4. Another signal processing model with FFT block added

4.3 Mapping

After the model is constructed, the model interpreter maps it to application code. The multi-graph kernel automatically synthesizes applications determined by their dataflow control graphs. The kernel allows the creation and propagation of data between nodes and provides a memory allocation facility with garbage collection. The dataflow node execution is controlled by its scheduling attributes: trigger mode and priority.

In this way, once DSML is built, a class of application model in this domain can be rapidly constructed and executable applications can be generated by the model interpreter, which facilitates the whole development process and work division. Further more, the meta-model can be evolved when the requirements change with technology advancement.

5 Conclusions

MIC technology is introduced to embedded system development for manufacturing equipment monitoring and process control. Signal process module is described as an example to illustrate how MIC facilitates the whole system development process and to discuss the viability to build DSML for embedded equipment control. Two signal process applications are built conveniently utilizing graphic DSML, which indicates the advantages of this technology. The control system architecture is explored in high-level design space. To satisfy the requirements for embedded control system development, this needs further efforts:

1) Build meta-model for subsystems and integrate them at meta-level. Detailed high-level abstraction of applications for this domain is required. Function and non-function performance must be considered and specified with syntax and semantics.

2) Modify the model interpreter according to the specified meta-model. GME provides an SDK for building interpreter.

3) Simulate the software and its physical environment at model level to guarantee the whole system performance at design time. GME provides XML facility for export and import configuration files for tools.

Acknowledgment

The authors wish to acknowledge the 3rd Program of Shenzhen Institute of Information Technology (QN-08009) and the 11th High Tech Fair Program of Shenzhen Institute of Information Technology (GJH11-ZS005). We also would like to express our gratitude to many people in ISIS, Vanderbilt University, for their suggestions and tools support.

References

1. Sztipanovits, J., Karsai, G.: Model-integrated computing. Computer 30(4), 110–111 (1997)
2. Nichols, J., Neema, S.: Dynamically Reconfigurable Embedded Image Processing System. In: Proceedings of the International Conference on Signal Processing Applications and Technology, Orlando, FL (November 1999)

3. Neema, S., Karsai, G.: Embedded Control Systems Language for Distributed Processing (May 12, 2004)
4. ESML documentation, `http://www.isis.vanderbilt.edu`
5. Gabor, K., Janos, S., Hubertus, F., Samir, P., Frank, D.: Model-embedded on-line problem solving environment for chemical engineering. In: Proceedings of the IEEE International Conference on Engineering of Complex Computer Systems, ICECCS, pp. 227–233 (1995)
6. Karsai, G., Sztipanovits, J., Ledeczi, A., Bapty, T.: Model-integrated development of embedded software. Proceedings of the IEEE 91(1), 145–164 (2003)
7. Ledeczi, A., Bakay, A., Maroti, M.: Model-Integrated Embedded Systems. In: Robertson, P., Shrobe, H.E., Laddaga, R. (eds.) IWSAS 2000. LNCS, vol. 1936, p. 99. Springer, Heidelberg (2001)
8. `http://www.isis.vanderbilt.edu/`
9. Sztipanovits, J., Karsai, G., Biegl, C., Bapty, T., Ledeczi, A., Malloy, D.: MULTIGRAPH: An Architecture for Model-Integrated Computing. In: Proceedings of the International Conference on Engineering of Complex Computer Systems, Ft. Lauderdale, FL, pp. 361–368 (November 6, 1995)
10. Karsai, G., Maroti, M., Ledeczi, A., Gray, J., Sztipanovits, J.: Composition and Cloning in Modeling and Meta-Modeling. IEEE Transactions on Control System Technology 12(2), 263–278 (2004)
11. GME 4 User's Manual, version 4.0 (March 2004), `http://escher.isis.vanderbilt.edu`

Two Stage Estimation Method in Data Processing and Simulation Analysis

Pan Xiong[1], Chunru Zhao[1], and Lihong Jin[2]

[1] Faculty of Information Engineering, China University of Geosciences, Wuhan, China
pxjlh@163.com, chunru0505@163.com
[2] Department of Basic, JiangCheng College China University of Geosciences, China
jlhpx@sina.com

Abstract. The new method presented in this paper shows an effective way of solving an estimation problem, the estimated values are nearer to their theoretical ones than in an adjustment with the method of least squares. Considering the semi-parametric adjustment models, firstly, the estimators of the parameters and the nonparametric are derived by using a kernel function and the least square method. Secondly, we discuss the approach of bandwidth choice in the semi-parametric adjustment models. Finally, a simulated adjustment problem is constructed to explain the method.

Keywords: semi-parametric models, kernel function, model error.

1 Introduction and Motivation

Stochastic factors and certain factors often coexist in survey data processing, conventional measurement adjustment adopts a stochastic model that just treats random error and does not treat certain systematic errors. However, systematic error surely exists in observation. If it is ignored, the adjusted result would be biased. When the model errors are minute as compared with random errors, omitting model errors does not influence the estimated value of the parameters seriously. When the model error is bigger, it will have a bad influence on parameter estimation, even resulting in false conclusions. It is generally assumed that genuine observations can be expressed by the functions depending on a group of parameters. Usually, the functional form of the regression model is assumed to be known. The problem is reduced to estimating a set of unknown parameters; in this case the observations are completely parameterized. Actually, it is difficult to describe the observations correctly by using several parameters.

Some measures can be taken to lessen or eliminate the model errors: Using existing compensation models of systematic errors to correct observations; When the relations between systematic errors and their influencing factors are linear, adding a few systematic parameters in the mathematical model can eliminate their influence efficiently; In survey adjustment, using a group of linear combination observations, instead of observations themselves, is another way to reduce systematic errors. For example, in the GPS data processing, the phase double difference technique is a

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 62–70.
springerlink.com

commonly used method to cancel out or significantly reduce most of clock errors, ionosphere and troposphere errors. Positioning precision is improved while some useful information is lost at the same time.

In many applications, there is not always evidence of a linear relationship. Therefore, research for more flexible models is needed. In the 1980s, statistics brought forward semi-parametric regression models comprising a parametric part and a non-parametric part. It has some significant merits over parametric models in dealing with complicated relationships between observations and estimated variables and offering useful information. This model has been discussed by a number of authors. e.g. Fischer and Hegland (1999),Green and Silverman (1994), Shi and Tsai (1999), Sugiyama and Ogawa (2002), Aerts and Claeskens (2002), Ruppert (2002), Wang and Rao (2002), David (2003).This paper introduce the semi-parametric regression method for surveying adjustment, and to develop a suitable method that can estimate the unknown parameters and can extract model errors simultaneously.

2 The Mathematical Model

The conventional survey adjustment is

$$L_i = b_i x + \Delta_i \qquad 1 \le i \le n \tag{1}$$

If the model include systematic error, in order to separate systematic errors form model errors. Assume that responses L_i include non-parametric $s(t_i)$, $i = 1,2,\cdots,n$, then Eq.(1) can be rewritten as:

$$L_i = b_i x + s(t_i) + \Delta_i \quad 1 \le i \le n \tag{2}$$

where $b_i = (b_{i1}, b_{i2}, \cdots, b_{ip})$ is a vector of non-random design points, $B = (b_1, b_2, \cdots, b_n)^T$ is of full rank with $rank(B) = p$, that is, its columns are linearly independent, $x = (x_1, x_2, \cdots, x_p)^T$ is a vector of unknown parameters, $t_i \in D$ (a subset of R^1), $s(\cdot)$ is an unknown smooth function on R^1, $s = (s_1, s_2, ..., s_n)^T$ is used to describe vector of model errors. $\Delta = (\Delta_1, \Delta_2, \cdots, \Delta_n)^T$ is an uncorrelated random error vector with $E(\Delta) = 0$ and $D_\Delta = \sigma^2 Q$, where σ^2 is the unit-weighted variance.

The error equation associated with Eq. (2) is

$$V = B\hat{x} + \hat{s} - L \tag{3}$$

where $\hat{x}$ and $\hat{s}$ denote the estimates of parametric x and non-parametric s ,$V \in R^n$ is the residuals.

3 Estimation Method and Models Solution

Two stage estimate method attempts to estimate the parameter x and non-parametric s separately. More precisely, it can be derived from the following two-stage procedure.

Stage 1. For fixed x, let $\alpha = \frac{1}{n}\sum_{i=1}^{n} s(t_i)$ and $\varepsilon_i = s(t_i) - \alpha + \Delta_i$. Then Eq. (2) can be written as

$$L_i = b_i x + \alpha + \varepsilon_i \tag{4}$$

To simplify the notation, we let $L = (L_1, L_2, \cdots, L_n)^T$, $1_n = (1, 1, \cdots, 1)^T$, $\hat{x}^*$ and $\hat{\alpha}^*$ be the least squares estimator of x and α.

The error equation associated with Eq. (4) is

$$V_1 = B\hat{x}^* + 1_n \hat{\alpha}^* - L \tag{5}$$

Based on ordinary least squares methodology, we construct the Lagrange function:

$$\Phi_1 = V_1^T P V_1 + 2\lambda_1^T \, (B\hat{x}^* + 1_n \hat{\alpha}^* - L - V_1) \tag{6}$$

where $P = Q^{-1}$ is a symmetric positive-definite matrix, and also the weight matrix of the observations L, λ_1 is a Lagrange constant.

Using $\frac{\partial \Phi_1}{\partial V_1} = 0$, $\frac{\partial \Phi_1}{\partial \hat{x}^*} = 0$ and $\frac{\partial \Phi_1}{\partial \hat{\alpha}^*} = 0$, we obtain

$$\lambda_1 = PV_1 \tag{7}$$

$$B^T \lambda_1 = 0 \tag{8}$$

$$\lambda_1^T 1_n = 0 \tag{9}$$

By simple calculus, it follows that Eq. (6) is minimized when $\hat{x}^*$ and $\hat{\alpha}^*$ satisfy the block matrix equation:

$$\begin{pmatrix} B^T PB & B^T P1_n \\ 1_n^T PB & 1_n^T P1_n \end{pmatrix} \cdot \begin{pmatrix} \hat{x}^* \\ \hat{\alpha}^* \end{pmatrix} = \begin{pmatrix} B^T PL \\ 1_n^T PL \end{pmatrix} \tag{10}$$

Eq.(10) forms a system of $p+1$ equations. This equation system is typically very large, and it may not be convenient, or even practical, to solve it directly. Fortunately, this is not necessary. One approach is to re-write Eq. (10) as a pair of simultaneous matrix equations

$$B^T PB\hat{x}^* + B^T P1_n \hat{\alpha}^* - B^T PL = 0 \tag{11}$$

$$1_n^T PB\hat{x}^* + 1_n^T P1_n \hat{\alpha}^* - 1_n^T PL = 0 \tag{12}$$

Form Eqs. (11) and (12), we obtain

$$\hat{x}^* = (B^T PB)^{-1} B^T P(L - 1_n \hat{\alpha}^*) \tag{13}$$

$$\hat{\alpha}^* = (1_n^T P1_n)^{-1} 1_n^T P(L - B\hat{x}^*) \tag{14}$$

We could alternate between Eqs. (13) and (14), solving repeatedly for $\hat{x}^*$ and $\hat{\alpha}^*$ respectively. This procedure is sometimes known as backfitting (Green and Silverman 1994). In practice, the convergence can be very slow. An alternative approach is a direct method. With $d_n = 1_n^T (P - PB(B^T PB)^{-1} B^T P)1_n$, and substituting Eq. (13) into Eq. (14), we get the first estimate $\hat{\alpha}^*$ of α

$$\hat{\alpha}^* = d_n^{-1} 1_n^T (P - PB(B^T PB)^{-1} B^T P)L \tag{15}$$

Substituting Eq. (15) into Eq. (13), we get a first estimator $\hat{x}^*$ of x

$$\hat{x}^* = (B^T PB)^{-1} B^T P(L - 1_n \hat{\alpha}^*) \tag{16}$$

Stage 2. Substituting $\hat{x}^*$ into Eq. (2), Eq. (2) can be considered as a non-parametric regression model

$$L_i - b_i \hat{x}^* = s(t_i) + \Delta_i \tag{17}$$

Therefore, $s(t_i)$ can be estimated by the kernel method

$$\hat{s}(t_i) = \sum_{j=1}^{n} K_h(t_i, t_j)(L_i - b_i \hat{x}^*) \tag{18}$$

where the weight $K_h(\cdot)$ is associated with a kernel function $K(\cdot)$ and has a bandwidth $h = h_n > 0$. For an asymmetric kernel function $K(\cdot)$, the weight $K_h(\cdot)$ is taken to be

$$K_h(t, t') = K\left((t - t')/h\right)/h \tag{19}$$

Substituting $\hat{s}(t_i)$ for $s(t_i)$ in Eq. (2) gives

$$L_i = b_i x + \hat{s}(t_i) + \Delta_i \tag{20}$$

Its error equation is

$$V_2 = B\hat{x} + \hat{s} - L \tag{21}$$

Based on the ordinary least squares methodology, we obtain the Lagrange function

$$\Phi_2 = V_2^T PV_2 + 2\lambda_2^T \ (B\hat{x} + \hat{s} - L - V_2) \tag{22}$$

Let $\frac{\partial \Phi_2}{\partial V_2} = 0$, and $\frac{\partial \Phi_2}{\partial x} = 0$, respectively. We obtain

$$\lambda_2 = PV_2 \tag{23}$$

$$B^T \lambda_2 = 0 \tag{24}$$

Substituting Eq. (23) into Eq. (24), and considering Eq.(21)

$$B^T PBx + B^T P\hat{s} - B^T PL = 0 \tag{25}$$

Since $B^T PB$ is a positive matrix, we get the final estimate of x as

$$\hat{x} = \left(B^T PB\right)^{-1} B^T P\left(L - \hat{s}\right) \tag{26}$$

Considering Eq. (18), we obtain the final estimate of s

$$\hat{s}(t) = \sum_{i=1}^{n} K_h(t,t_i)(L_i - b_i\hat{x}) \tag{27}$$

4 Fitted Values and Hat Matrix

If we now assume that

$$\tilde{P} = (I - W)P \tag{28}$$

has full column rank, the weights estimators for the parameters in model Eq. (2) are

$$\hat{x} = \left(B^T \tilde{P}B\right)^{-1} B^T \tilde{P}L \tag{29}$$

and

$$\hat{s} = W\left(L - B\hat{x}\right) \tag{30}$$

where I is an $n \times n$ identity matrix, and $W(t) = diag\left(K_h\left(t,t_1\right),\cdots, K_h\left(t,t_n\right)\right)$.

The ordinary least-squares fit to the data is

$$\hat{L} = B\hat{x} + \hat{s} \triangleq J(h)L \tag{31}$$

with

$$J(h) = W + (I - W)B(B^T \tilde{P}B)B^T \tilde{P} \tag{32}$$

In the semi-parametric model, $J(h)$ is usually called the hat matrix.

5 Bandwidth Selection

As is well known, the bandwidth parameter h controls the trade-off between the goodness of fit and the roughness, and it is important to choose an appropriate h.

The residual sum of squares (RSS) of a model is a measure of predictive ability, since a residual is the difference between an observation of a response and its fitted or predicted value:

$$e_i \mathrel{\hat{=}} L_i - \hat{L}_i \tag{33}$$

However, RSS is not satisfactory as a model selector. The problem is that $\hat{L}_i$ uses L_i as well as the other observations to predict L_i. The result is that the most complex model – that is, the model with the largest degrees of freedom and containing the other models as special cases – always has the smallest RSS. On the other hand, because the observation L_i is being used as part of its own predictor, a small amount of smoothing, which gives a large weight to L_i, appears optimal to RSS for prediction result.

There is a simple remedy to this problem: when predicting L_i, use all the observations except the *i*th one. Thus, define $\hat{L}^{-i}(t_i, h)$ is the semi-parametric regression estimator applied to the data but with (t_i, L_i) deleted. Then, let $e_{(i)} \mathrel{\hat{=}} L_i - \hat{L}^{-i}(t_i, h)$ be the *i*th deleted residual. The predicted residual sum of squares (PRESS) is defined by

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 \tag{34}$$

This method is the technique of model validation that splits the data into two disjoint sets, fits the model to one set, predicts the data in the second set using only the fit to the first set, and then compares these predictions to the actual observations in the second set. This “leaving one out” strategy is a way of guarding against the wiggly answer that RSS gives. The choice of h is the one that minimizes PRESS.

The PRESS criterion is not as difficult to calculate as it might first appear. One does not need to fit the model n times, thanks to an important identity. Let $J_{ii}(h)$ be the *i*th diagonal element of the hat matrix $J(h)$. Then the *i*th deleted residual is related to the *i*th ordinary residual by

$$L_i - \hat{L}^{-i}(t_i, h) = (L_i - \hat{L}_i) / \left(1 - J_{ii}(h)\right) \tag{35}$$

Using Eq. (35), we obtain

$$PRESS(h) = \sum_{i=1}^{n} \left((L_i - \hat{L}_i) / \left(1 - J_{ii}(h)\right)\right)^2 \tag{36}$$

Thus, the PRESS can be computed using only ordinary residuals and the diagonal elements of the hat matrix.

6 Simulation Experiment

We compared the semi-parametric approach with the ordinary least squares adjustment approach. For a simulated simple semi-parametric model

$$L = Y + S + \Delta \tag{37}$$

Where $Y = BX$, $X = [2,3]^T$, $B = (b_{ij})_{100\times 2}$, $b_{i1} = -2 + i/20$, $b_{i2} = -2 + (i/20)^2$, $S = (s_i)_{100\times 1}$, $s_i = 9\sin(t_i\pi)$, $t_i = i/100$, $i = 1,2,\cdots,200$. The observation error vector Δ is composed of 200 stochastic data obeying the standard normal distribution, which are $N(0,1)$ distributed. In Figure 1, the smooth curve shows the true values Y ; the saw-tooth curve is the observation L .

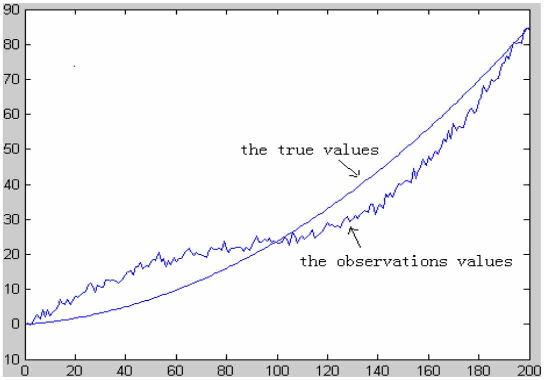


Fig. 1. The Observations and Their True Values

Choose the weighted matrix P is an identity matrix. Using the parametric model and the least squares adjustment approach, we obtain

$$\hat{X} = (B^T PB)^{-1} B^T PL = (-3.1194, 3.3611)^T$$

An incorrect result may be given with the conventional technique when the model error exits.

For the Eq. (37), we use the quartic kernel

$$K(t) = 15/16((1-t^2))^2 I(|t| \le 1)$$

The results for different bandwidth are presented in Table 1. Table 1 lists the bandwidth h and estimation of parametric component $\hat{X}$.

Table 1. Bandwidth and estimation of parametric component

h	$\hat{X}$	h	$\hat{X}$
0.11	(1.1938, 2.6599)	0.20	(1.9264, 2.8025)
0.12	(1.3695, 2.6545)	0.21	(1.9021, 2.8390)
0.13	(1.5196, 2.6553)	0.22	(1.8598, 2.8783)
0.14	(1.6451, 2.6620)	0.23	(1.8003, 2.9200)
0.15	(1.7468, 2.6742)	0.24	(1.7245, 2.9638)
0.16	(1.8254, 2.6914)	0.25	(1.6332, 3.0092)
0.17	(1.8819, 2.7132)	0.26	(1.5274, 3.0560)
0.18	(1.9171, 2.7392)	0.27	(1.4079, 3.1038)
0.19	(1.9316, 2.7690)	0.28	(1.2755, 3.1523)

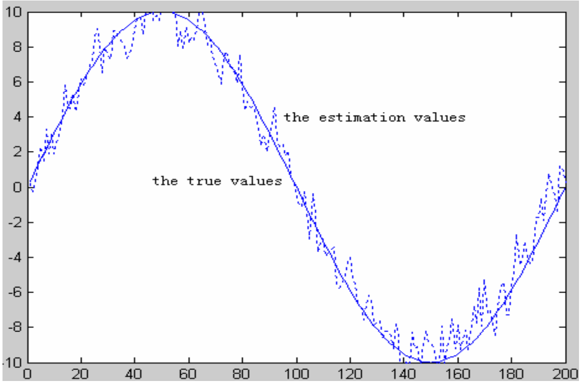


Fig. 2. The Estimation and True Values of Nonparametric Component

By calculation Eq.(31) and Eq.(36), the $PRESS(h)$ -selected bandwidth is $h = 0.2096$ and Estimation of parametric component is

$$\hat{X} = (1.9034, 2.8375)^T$$

the estimation $\hat{X}$ approximate to the true values of X , the model errors is very close to the value of S , which is shown in Fig. 2.Thus it can be seen that our method is successful.

7 Conclusion

In many applications the parametric model itself is at best an approximation of the true one, and the search for an adequate model from the parametric family is not easy. When there are no persuasive models available, the family of semi-parametric regression models can be considered as a promising extension of the parametric family. Semi-parametric regression models reduce complex data sets to summaries that we can understand. Properly applied, they retain essential features of the data while discarding unimportant details, and hence they aid sound decision-making.

From the above, it is difficult for the adjustment method of least squares to detect systematic errors in the data processing model, but semi-parametric regression model is valid. Added only a few additional parameters the model cannot express the complicated model errors, which will bring bad influence to the estimation of the parameters without considering the model errors shown in the examples. Introduce the semi-parametric regression theory for surveying adjustment, may be reduce the error of semi-parametric estimators by choosing an appropriate value for the kernel function and bandwidth.

On the other hand, there exit a lot of problems to discuss, for instance: how to choose kernel function and bandwidth? It is one of successful keys and needs to be further explored, and is under investigation.

References

1. Green, P.J., Silverman, B.W.: Nonparametric Regression and Generalized linear Models, pp. 145–189. Chapman and Hall, London (1994)
2. Fischer, B., Hegland, M.: Collocation, Filtering and Nonparametric Regression, Part, ZfV, 17–24 (1999)
3. Shi, P.D., Tsai, C.L.: Semiparametric regression model selections. J. Stati. Plan. & Infer. 77, 119–139 (1999)
4. Hong, S.Y.: Normal approximation rate and bias reduction for data-driven kernel smoothing estimator in a semiparametric regression model. J. Multi. Analy. 80, 1–20 (2005)
5. Sugiyama, M., gawa, H.O.: A unified method for optimizing linear image restoration filters. Signal Proce. 82, 1773–1787 (2002)
6. Aerts, M., Claeskens, G., Wand, M.P.: Some theory for penalized spline generalized additive models. J. Stati. Plan. & Infer. 103, 445–470 (2002)
7. Wang, Q.H., Rao, K.: Emprirical likelihood-based inference in linear errors-in-covariables models with validation data. Biometrika 89, 345–358 (2002)
8. Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression, pp. 271–320. Cambridge University Press, Cambridge (2003)

Research and Implementation of a Reconfigurable Parallel Low Power E0 Algorithm

Li Wei, Dai Zibin, Nan Longmei, and Zhang Xueying

Information Engineering University, Zhengzhou 450004, China
{try_1118, zhxy727, I}@163.com, nanlongmei@yahoo.com.cn

Abstract. A low power and dynamic reconfigurable parallel hardware architecture of E0 algorithm is presented, which can satisfy sixteen different LFSRs in the Bluetooth telecommunication systems. Moreover, the paper proposes the parallel realization of LFSR, which can support to generate N parallel sequence in each step. The new LFSR design techniques can be also useful in any LFSR. To reduce the conventional switching activity, we proposed the clock-gating technique to implement the LFSR. As to the different low power method, the paper performs detailed comparison and analysis. The design has been realized using Altera's FPGA. Synthesis, placement and routing of reconfigurable design have accomplished on 0.18μm CMOS process, the result proves Up to 40% power consumption was reduced compared with conventional E0 implementation. And the critical throughput rate can achieve 166Mbps.

Keywords: Reconfigurable, Low Power, Parallel, E0, Feedback shift register.

1 Introduction

Stream ciphers are the core mechanism of highly private and confidential communication. In the wireless communication Bluetooth system, the user information can be protected by the encryption of the packet payload[1]. The encryption of the payloads is carried out with a stream cipher called E0, which is resynchronized for every payload. The stream cipher system E0 consists of three parts. One part performs the initialization(generation of the payload key), the third part performs the encryption and decryption. The effective length of the encryption key can vary between 8 and 128 bits. Then with the use of E0, the key length is reduced by a modulo operation to the desire degree. The Bluetooth system specifies sixteen different polynomials for the implementation of this polynomial modulo operation.

In the paper, a new reconfigurable design for implementation of the LFSR is proposed, which can be programmed and configured in order to implement any of these previous mentioned sixteen polynomials[2].

By using the clock-gating technique, the proposed reconfigurable LFSR achieved to reduce the power consumption. So the reconfigurable design can be used effectively in portable telecommunication systems. In the following section 2 some

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 71–78.
springerlink.com

issues related to E0 algorithm are presented. The low power method related LFSR is introduced by section 3. The proposed architecture and its low power implementation are described in section 4. Experimental and simulation results are shown in section 5 and the paper conclusions are given in section 6.

2 Description of E0 Algorithm

The encryption of packet payloads in Bluetooth is performed by the E0 stream cipher[3], which consists of three components, as illustrated in Fig.1. The first component is the payload key generator, which performs the initialization. The second one, the keystream generator, generates the keystream bits, and uses for this purpose four Linear Feedback Shift Registers whose output is the keystream sequence or the randomized initial start value during the initialization phase. The length L of the four LFSRs are 25, 31,33 39 respectively.

For the LFSRs initialization[4], the keystream generator needs to be loaded with an initial value for the four LFSRs and with 4 bits that specify the values of registers in the summation combiner. The 132-bit initial value is derived form four inputs by using the keystream generator itself. The input parameters are the encryption key K_c, a 128-bit random number, a 48-bit Bluetooth address, and the 26 master clock bits. Within the payload key generator, the K_c is modified into another key denoted $K_c^{'}$, by using the polynomial modulo operation described in[1]. The maximum effective size of this key is factory preset and may be set to any multiple of eight, between one and sixteen.

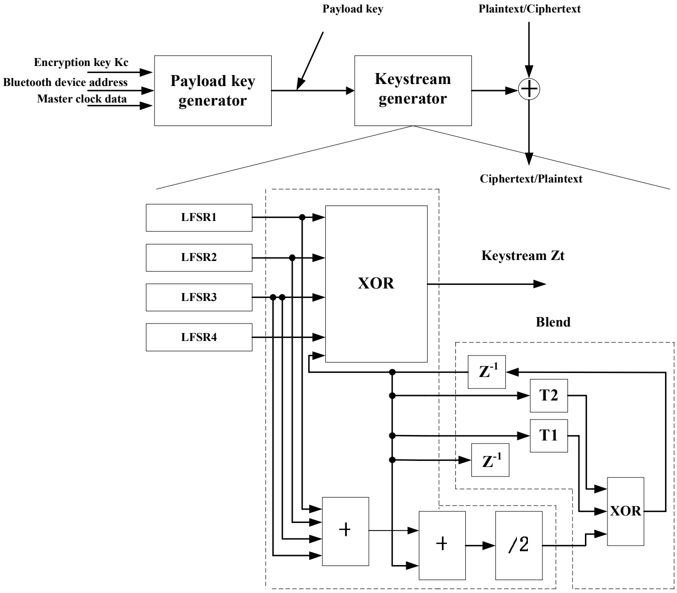


Fig. 1. The E0 stream cipher architecture

When the encryption key has been created, all the bits are shifted into the LFSRs, starting with the least significant bit. Then, 200 stream cipher bits are created by operating the generator. The last 128 of these bits are fed back into the keysteam generator as an initial value of the four LFSRs. The values of the state machine are preserved. From this point on, the generator produces the encryption sequence is bitwise XORed with the transmitted payload data, in the third component of the cipher.

3 Design for Low-Power Consumption

The total power consumption of a CMOS circuit is the sum of static and dynamic power consumption[5]. The static power consumption caused by the leakage current, mainly depends on the size of the chip. It is very small and can be more or less ignored for our setting, following equation presents the influences on dynamic power consumption. The design measures for lowering the power consumption result from minimizing the factors in this equation.

$$P_{dyn} = C_L \cdot V_{DD}{}^2 \cdot f_{CLK_{eff}} \cdot p_{sw}$$

The load capacitance C_L of the chip increases as more gates are placed on the die. This means that lowering the die size as well as reducing the supply voltage to a minimum directly reduces power consumption. These two factors tend to be predetermined by the low die-size constraint and the operating conditions of the chip. The switching activity p_{sw} of the circuit can be reduced by using a method called sleep logic. Thereby, unnecessary switching activity is eliminated by using AND gates at the input of combinational circuits. Assuming a fixed supply voltage and a minimum o switching activity in the combinatiorial paths, the best option for a low-power design is reducing the effective clock frequency $f_{CLK_{eff}}$ of the circuit. This reduces the power consumption linearly. Clock gating is a good measure for reducing the effective clock frequency. In our architectures, all datapath registers and nearly all control logic registers are clocked only when there is a potential signal change. This turned out to lower the lower consumption significantly. Thereby, parts of the circuit are virtually switched off when they are not in use.

Figure 2 shows a typical clock gating register which consists of a data latch and b flip-flops. The D-latch is used to prevent glitches in the clock signal of the flip-flops which build a register of b bits. The optimal bit-width of the registers depends on the implemented algorithm. Minimizing the number of memory elements(flip-flops and latches have nearly the same power consumption at the active clock edges) that are active at the same time, leads to a reduction of the total power consumption. Therefore we separate all N flip-flops of the design into registers with b bits. The number of active elements n depends on the total number of memory elements N required by the algorithm and the bit-width b of the registers and can be calculated using following equation. The first value N/b is the number of clock-gating D-latches which are active in every clock cycle(this is also the number of b bit registers in the

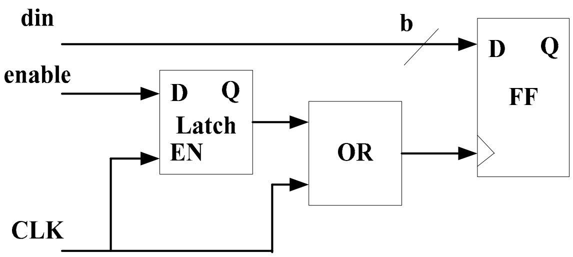


Fig. 2. Clock gating register

design) and b denotes the number of flip-flops in each register. When minimizing this equation it follows that the optimal bit-width of a design is $b = \sqrt{N}$.

$$n(b) = \frac{N}{b} + b$$

$$\frac{dn}{db} = -\frac{N}{b^2} + 1$$

$$-\frac{N}{b^2} + 1 = 0$$

$$b = \sqrt{N}$$

For the algorithms E0 the optimal number of flip-flops per register is 5.6, therefore we decide to implement algorithm using a 5-bit word size.

4 Reconfigurable Parallel Low Power Architecture of E0 Algorithm

The reconfigurable parallel low power architecture of E0 algorithm is shown as Fig.3. It can be seen the feedback shift registers LFSR shift 5 bits per clock cycle. Only a single register is clocked at the same point in time via clock gating technique. Additionally, all combinational circuits like the finite state machine, which have to be implemented in radix 5. The inputs of the function are selected bits from the registers and are not shown in detail in this figure. AND gates are used to enable and disable the appropriate inputs. The detail of the finite state machine is shown in 4.2.

During initialization, the key is loaded into the registers. In the next clock cycles, the EN1~EN8 is enabled in turn, then the finite state machine is used to generate final keystream. So producing a 5-bit encryption result after initialization requires 9 cycles.

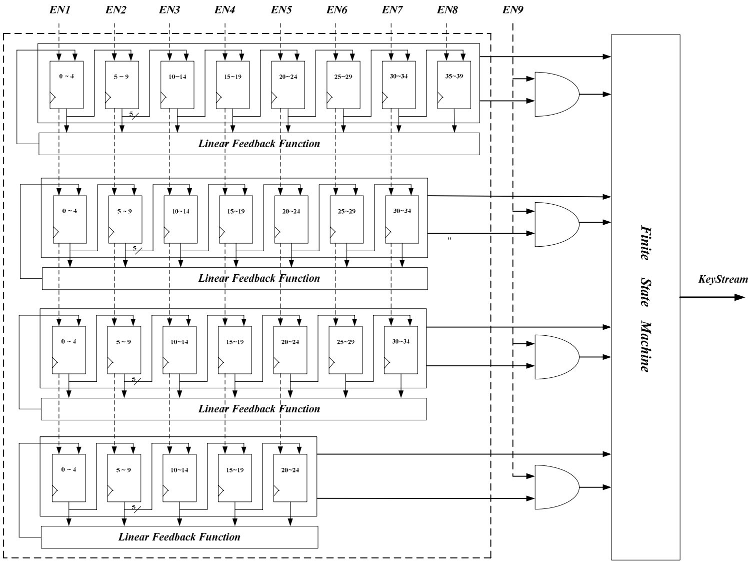


Fig. 3. The architecture of the reconfigurable low power E0

4.1 The Implementation of Finite State Machine

The keystream generator E0 used in Bluetooth belongs to a combination generator with four memory bits, denoted by $\sigma_t = (c_{t-1}, c_t)$ at time t, where $c_t = (c_t^1, c_t^0)$. The whole system uses four Linear Feedback Shift Registers denoted by $R_1, \cdots, R_4$ respectively. At clock cycle t, the four LFSRs' output bits $x_t^i, i = 1, \cdots, 4$, will be added as integers, the sum $y_t \in \{0, \cdots, 4\}$ is represented in the binary system. Let y_t^i denote its i-th least significant bit (i=1,2,3). A 16-state machine emits one bit c_t^0 out of its state $\sigma_t = (c_{t-1}, c_t)$ and takes the input y_t to update σ_t by σ_{t+1}. T_1 and T_2 are linear mapping, which can be implemented by XOR operation[4]. Finally, the keystream z_t is obtained by xoring y_t^1 with c_t^0. That is,

$$x_t^1 \oplus x_t^2 \oplus x_t^3 \oplus x_t^4 \oplus c_t^0 = z_t$$

More formally, by introducing two temporary bits $s_{t+1} = (s_{t+1}^1, s_{t+1}^0)$ each clock, the following indirect iterative expressions between s_{t+1} and c_{t+1} suffice to update c_t:

$$s_{t+1} = \left\lfloor \frac{y_t + 2 \cdot c_t^1 + c_t^0}{2} \right\rfloor$$

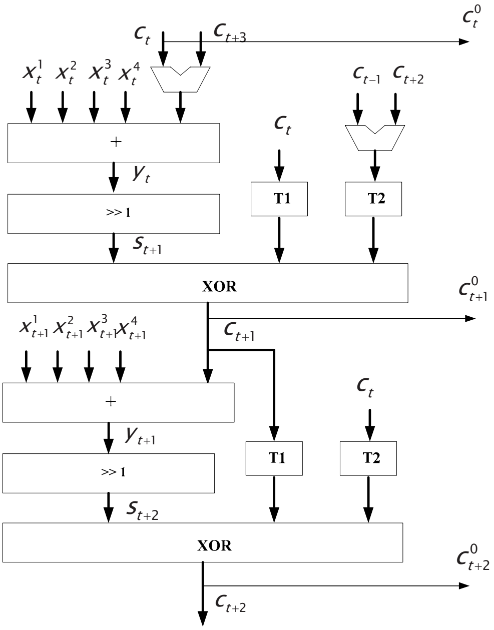


Fig. 4. The hardware of finite state machine

$$c_{t+1}^1 = s_{t+1}^1 \oplus c_t^1 \oplus c_{t-1}^0$$

$$c_{t+1}^0 = s_{t+1}^0 \oplus c_t^0 \oplus c_{t-1}^1 \oplus c_{t-1}^0$$

In the section 4, It can be seen the the feedback shift registers LFSR shift 5 bits per clock cycle, Based on the scheme above, we can calculate s_{t+2} and c_{t+2} in one clock, as shown in Fig.4. Obviously, this allows the speed to be easily multiplied by up to 5 if a sufficient amount of hardware is available, it is quite feasible to increase the throughput in this way.

5 Analysis and Comparison of Performance

5.1 The Realization of FPGA and ASIC

Based on the analysis above, the prototype has been accomplished RTL description using Verilog language. And the design continues with the synthesis using QuartusII 6.0 from Altera Corporation. The prototype has been verified successfully based on Altera's Cyclone EP1C12Q240C8. The performance is shown in table 1.

Table 1. The performance of reconfigurable E0 algorithm based on the FPGA

Maximum Frequency (MHz)	Logic Elements (LEs)	Throughput rate (Mbit/s)
174	320	96

The design has been synthesized under 0.18μm CMOS process using Synopsys' Design Complier to evaluate performance more accurately. Synthesis, placement and routing of the prototype based on architecture have accomplished. The performance results of the prototype have been shown in table 2.

Table 2. The performance of E0 algorithm based on ASIC

Maximum Frequency (MHz)	Area (μm^2)	Throughput rate (Mbit/s)
320	153673	177

Furthermore, The prototype using Magma's Blast Rail to evaluate the power consumption. The performance results of the design have been shown in table 3.

Table 3. The power consumption performance of E0 algorithm

Leakage consumption	Internal consumption	Dynamic consumption	Total consumption
247.1nw	87.8μw	1.03mw	1.11mw

Based on the synthesis result above, as to the two critical parameters including power consumption and throughput rate, we compare our design with references[6][7], Ref[6] is the specific circuit for E0 algorithm. Ref[7] is the low power design of E0 algorithm, which adopts Gray code representation. The result is shown in table 4.

Table 4. The comparison with other designs

Design	Power Consumption (mw)	Throughput rate (Mbit/s)
Our design	1.11	177
Ref[6]	--	189
Ref[7]	1.30	90

The result proved that our design has same latency with Ref[6], but Ref[6] can only support a kind of LFSR. So our design has more flexibility than Ref[6], besides, in the consideration of power consumption, our design has lower power consumption than Ref[7]. Especially, our design has an obvious advantage on throughput rate.

6 Conclusion

As a word, the paper proposes a low power consumption reconfigurable parallel architecture, which can implement sixteen different LFSRs in the Bluetooth telecommunication systems. Besides, we introduce the clock-gating technique to reduce the power consumption. Furthermore, the paper proposes the implementation of finite state machine, which can allow the speed to be easily multiplied by up to 5. Finally the paper makes the detailed comparison and analysis to low power design. Synthesis, placement and routing of reconfigurable design have accomplished on 0.18μm CMOS process. Compared with conventional designs, the result proved architecture has an obvious advantage in the power consumption and flexibility.

Acknowledgments

This work was supported by High Technology Research and Development Program of China(863) (No. 2008AA01Z103).

References

1. SIG Bluetooth. Bluetooth specification, `http://www.bluetooth.com` (accessed August 18, 2003)
2. Specification of the Bluetooth System. Specification, vol. 1, pp. 159–167 (December 1, 1999)
3. Ekdahl, P.: On LFSR based Stream Ciphers. Lund University (November 21, 2003)
4. Ekdahl, P., Johansson, T.: Some results on correlations in the Bluetooth stream cipher. In: Proceedings of 10th Joint Conference on Communications and Coding, Austria, pp. 83–93 (2000)
5. Feldhofer, M., Wolkerstorfer, J.: Low-power Design Methodologies for an AES Implementation in RFID Systems. In: Workshop on Cryptographic Advances in Secure Hardware 2005 (CRASH 2005), Leuven, Belgium, September 6-7 (2005)
6. Kitsos, P., Sklavos, N., Papadomanolakis, K., Koufopavlou, O.: Hardware Implementation of Bluetooth Security. IEEE Pervasive Computing 2(1), 21–29 (2003)
7. Kitsos, P., Sklavos, N., Zervas, N., Koufopavlou, O.: A Reconfigurable Linear Feedback Shift Register (LFSR) For the Bluetooth System, pp. 991–994. IEEE, Los Alamitos (2001)

A Model of Car Rear-End Warning by Means of MAS and Behavior

Liang Jun[1], Xianyi Cheng[1], Xiaobo Chen[1], and Yao Ming[2]

[1] School of Computer Science and Telecommunication Engineering,
Jiangsu University, China, 212013
liangjun@ujs.edu.cn
[2] School of Automobile and Traffic Engineering,
JiangSu University, Zhenjiang, JiangSu, 212013, China

Abstract. This paper proposes a Car Rear-end warning model based on MAS (Multi-Agent Systems) and driving behavior, which was consists of four different types of agents that can either work alone or collaborate through a communications protocol on the basis of the extended KQML. The algorithm of Rear-end alarming applies Bayes decision theory to calculate the probability of collision and prevent its occurrence real-time. The learning algorithm of driving behavior based on ensemble artificial neural network (ANN) and the decision procedure based on Bayes' theory are also described in this paper. Both autonomy and reliability are enhanced in the proposed system. The effectiveness and robustness of the model have been confirmed by the simulated experiments.

Keywords: rear-end warning, driving behavior, multi-agent, Bayes decision.

1 Introduction

According to the investigations and statistics, of all the traffic accidents ever took place, about seventy-five percent were caused by the misjudgment or steer miss of the drivers(see Fig 1), which, to a large extent, was due to the insufficiency of the drivers' cognitive and information processing abilities when steering more speedy and complicated vehicles. As a result, the driver-and-passenger centered intelligent active safety system for automobile, along with its new driving sensors and controllers, has become a greater concern to the manufacturers and mass consumers of automobiles and affiliated products in America, Japan, Europe and countries all over the world [1][2]. After attentive studying of different car rear-end warning models including mazda model on CW/CA system, honda model and an improved model by California Berklee College, the authors found out the disadvantages of most of the current alarm systems. One is that the distance between vehicles is taken as the only determinant factor of vehicle collision by current alarm systems, while the possible effects of the driver's behavior of driving is not taken into account [3]. The other is that the technology of measurement applied by most of the current alarm systems is active measurement which entails special hardwares such as radar, infrared ray and CCD camera, etc, which are, more often than not, expensive and there could be signal interference.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 79–87.
springerlink.com

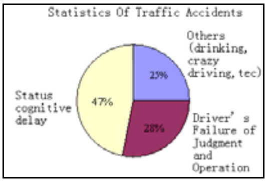


Fig. 1. Statistics of traffic accidents

As the latest development of artificial intelligence, the multi-agent system [4] is one of the most important branches in the studies of distributed artificial intelligence. This system is widely used in automatic control, distributed information processing system, meta-synthesis of complex system, intelligent expert system and so on. Also, it offers a new exciting way for the coming out of the car rear-end warning system [3][5].

The car rear-end warning system based on the multi-agent system is the new generation of the intelligent alarm support system [6]. Technically based on the multi-agent system theory in the field of artificial intelligence and theoretically guided by the meta-synthesis method from qualitative to quantitative, the subject alarm system is expected to have improved performance and superior functions.

Hence, taking advantage of the flexibility, autonomy, interaction and intelligence of the multi-agent system and taking into account both driving behavior and traditional alarm means of distance measurement, the paper intends to propose a model of car rear-end warning based on MAS and behavior (**MCRWMB**), which utilizes the Bayes Decision algorithm in car rear-end warning system to the calculation of collision probability in order to prevent the automobile accidents such as tailgating or collision.

2 MCRWMB

2.1 Driving Behavior

The history of driving behavior model can be traced back to the theory of field analysis of vehicles. The research of driving behavior didn't gain much progress in 50 years although some work still has been done. But it became a hot research area since 2000. After the comparison and analysis of the existing theory and model, Vaa[7] pointed out in 2001 that the cognition and emotion is good tools for prediction, avoidance and evaluation of dangerous situation. Then Salvucci, Krajzewicz and Delorme developed investigation on the cognitive structure of driving behavior and it turns out that the rapid development of cognition science and modeling methods are helpful to the research of driving behavior[3][8].

The model of driving behavior can be described as (E,H) abstractly, where E=(E_1,E_2,...E_n) is used to express external situation such as the road, weather and so on; H=(H_1, H_2…,H_n) representing the driving behavior which refers to the operation behavior on vehicles. It includes operations on steering wheel, accelerator (throttle), brake pedal, clutch pedal and related control unit like steering lamp, lighting lamp, windscreen wiper, etc. The driver usually follows his/her own habits or preference of

(E,H). The preference model of driver can be learned based on a lot of records of his/her driving behavior and can be used to predict his/her future behavior H_i. If the risk evaluation function y(.) can be built simultaneously based on this, then the risk of the predicted behavior can be assessed.

2.2 MCRWMB Structure

The multi-agent system theory provided a new way for the research of decision support system[6]. Car rear-end warning model based on multi-agent (see Fig 2) have four kinds of agents. Each agent is autonomous, and the cooperation among the agents could improve the accuracy and the real-time performance of the alarm system. For this end, it is necessary that a reasonable framework of multi-agent alarm system be designed with the ability of learning and dynamic coordination, which could give full play to its own superiorities and change the traditional routinization of the calculation mode so as to improve the warning efficiency and reinforce the driving security.

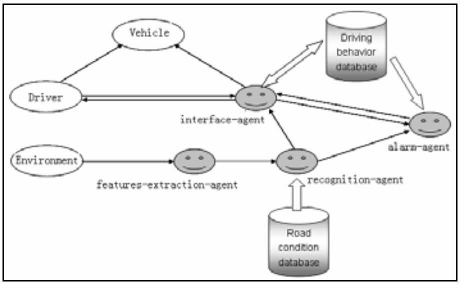


Fig. 2. MCRWMB

2.3 Multi-agent Communications Based on Extended KQML

Knowledge Query and Manipulation Language(KQML) is a kind of most generally used agent communication language [4][6][9], which could function as a coordinator or cooperator to help realize the real-time information share among agents through

Table 1. The semantics of extended KQML

Name	Weather Reserved Words	Meaning
Release	N	interface-agent send the driving message to all
Configure	Y	Alarm-agent need to be initialized before driving.
Announce	N	Interface-agent send messages about road and driving habit to other agents, and hope to get the message back
Precaution	N	alarm-agent send the result of alarm in “ content” to interface-agent
Accept	N	interface-agent accept the alarm message and begin to give an alarm to driver

information change. Moreover, KQML have the characteristics of openness, expandability, readability, convenience and platform independence. Therefore, on the foundation of original reserved words in KQML, we defined several new ones, such as <Announce>,<Precaution>,<Accept>, etc (see Tab 1) so as to realize real-time communication in our multi-agent system.

3 Bayes Decision with Driving Behavior

In this section we need to model the relationship between environment, driving behavior and danger which we abbreviate to env, beh and dang respectively.

3.1 The Modal of Environment and Danger

Considering the complexity of alarm environment, we introduce the fuzzy thought [10]. The dang is the measurement of criticality, which is a number in [0, 1], and 0 means having no danger and 1 means the severest danger. The env is also a number between 0 and 1, 0 shows that the environment is very good, such as clear whether, hollowness on road, no foot passengers and no vehicles, and 1 means that the environment is very scurviness (the complex road, many roadblocks and the bad weather). According to the linear model as follows, we can evaluate the environment.

$$env = \sum_{i=1}^{n} \alpha_i x_i \tag{1}$$

In formula (1), x_i is the factor to evaluate, such as weather condition, road grade and the number of roadblock in certain range; α_i is influencing factors' relative weight, which need to be appointed by the man who has great experience; beh is named as behavior tolerance, which, as mentioned earlier, is the result of risk evaluation y(H$_i$) of driving behavior, lower behavior beh means the rear-end accident has low probability to happen. We call P(env) as prior probabilities which express occurence probability of different environment, and regard it as uniform distribution when we have no prior information. P(env, beh | dang) expresses the occurrence probability of special environment and behavior conditioned on the danger happened, and it needs to be represented by a probability model such as GAUSS model[11]. P(beh | env) is the probability of adopting beh behavior under the current environment. P(dang | env) is the tolerance of danger degree toward this current environment. In this paper, P(beh | env) model and P(dang | env) model are all built by GAUSS model, which can be showed by:

$$\begin{cases} beh = a * env + \varepsilon \\ dang = b * env + \varepsilon' \end{cases} \tag{2}$$

Where a , b is called counterchange parameter, ε and ε' is noise data. P(dang | env,beh) is the probability of the danger of dang occurring under the current environment and driving behavior. When the value of P(dang | env,beh) is larger than the presented threshold, interface-agent will adopt many kinds of forms such as sound, photo electricity to send the danger's degree to driver.

3.2 Learning of Driving Behavior Based on Ensemble ANN

Now we focus on the relationship between driving behavior and environment as follows. The structure of our ensemble ANN is illustrated in Fig 3. It is composed of three layers: the input layer, the middle layer and the output layer. It is the middle layer which differ it from traditional ANN. In traditional ANN, the middle layer is comprised of neural which do mapping using a simple function like sigmoid function. But in our ANN, the middle layer (hidden layer) constitutes multiple individual NN referred to as sub-NN which is used to learn each behavior such as accelerator (throttle), brake pedal. The output of sub-NN is the probability that behavior happens and they are trained by traditional learning algorithm, etc BP. The output of sub-NN is collected in the output layer based on decision theory. Here P(beh$_i$) represents the probability that behavior happens and w$_i$ is the combination factor which affect the weight of each behavior. Now our goal is to learn both hidden neurons and factor w. According to the classical decision theory, the behavior that leading to the largest w$_i$P(beh$_i$) will be chosen as the output finally. This can be expressed as follows:

$$beh_i = \arg\max_i \frac{w_i P(beh_i)}{\sum_j w_j P(beh_j)} \tag{3}$$

Usually, P(beh$_i$) has a parametric form which can be written down explicitly P(a$_i$) and then, the above expression can be rewritten as:

$$beh_i = \arg\max_i \frac{w_i P(\alpha_i)}{\sum_j w_j P(\alpha_j)} \tag{4}$$

On the basis of (4), we can use training samples to learn both the parameters a$_i$ and w$_i$ simultaneously through optimization method such as conjunct graduation descent. Having learned the unknown parameters, we can obtain the relationship between environment and driving behavior through the first-half of the ensemble ANN (Fig 3.).

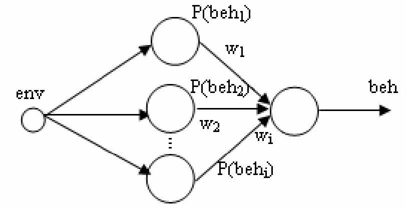


Fig. 3. The structure of ensemble ANN

3.3 MCRWMB Algorithm

Combining the information of the risk assessment function of y(H$_i$) and the grade of alarm sent by alarm-agent, we can obtain the result by bayes' formula of conditional probability.

$$\begin{aligned}
&P(dang \mid env, beh) \\
&= P(env, beh \mid dang)P(dang) / (\sum_{dang} P(env, beh \mid dang)P(dang)) \\
&\propto P(env, beh \mid dang)P(dang) \\
&= P(beh, dang \mid env)P(env) \\
&= P(beh \mid env)P(dang \mid env)P(env)
\end{aligned} \tag{5}$$

Each item of the last line of (5) has been studied in the previous sections which are now combined to infer the final probability of danger with regard to special environment and driving behavior.

4 Simulation Experiment and Analysis

In this paper, the authors venture to design a simplified experiment of virtual-single-lane and double-vehicles- following in order to obtain relevant data needed in MCRWMB model. In the experiment, we translated the 3D virtual models of vehicles and roads built by 3DMAX into 3DS format, and then input them into the virtual reality software Eon Studio. The position coordinates of those vehicles and roads were obtained from the “Property Bar” in Eon Studio. The main data recorded by this experiment was driver’s reactions while facing the acceleration and decelerating of the vehicle afront in the virtual vehicle flow, mainly including the records of the accelerator and the brake. The multi-group data were also obtained by measuring and recording the following distance and speed difference of the opposite vehicles.

Experiment 1 : verifying the effectiveness of MCRWMB.

When we input the speed difference (Δv) and the following distance (Δs) measured previously into the model, we obtained the output, i.e., the collision time ($\Delta t = \Delta s / \Delta v$) and the change process of the risk degree, which is illustrated in Fig 4. It is shown in Fig 4 that the interface-agent gave alarm twice during the driving process. The alarm degree for the first time is relatively higher than that of the second time, which is in accordance with the experiment design. Hence the effectiveness of this model could be verified.

Comparing two curves in Fig 4, we found that MCRWMB model gave right alarm when the degree of the dangerous circumstances reached class C or above. and the above degree. whereas, at about 1000 sample point, the output of the model was class C while the collision time curve showed only a minor degree of danger at this time, which means deviation exists in the estimation of the model. However, it could just be explained that MCRWMB model adopt quite cautious "attitude" towards danger.

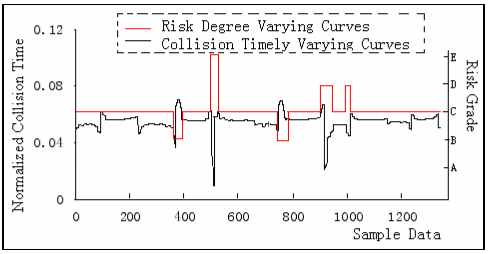


Fig. 4. The change process of normal state to collision and the degree of danger

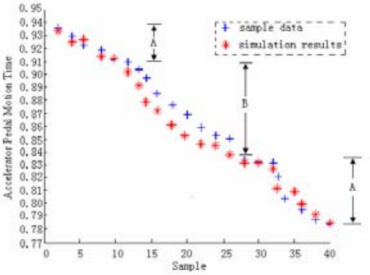


Fig. 5. The change procedure of accelerator pedal and the result of simulation

In the experiment, when the alarm-agent gave the first alarm, the following car's driver braked urgently, and then turned into another kind of normal driving state. According to the above analysis, we can see that MCRWMB model not only can tell the driver's driving preference and give effective alarm, but also has good robustness at noise data caused by the driver's unsuitable operation in urgent conditions.

Experiment 2 : use simulative brake sample data as a case to verify MCRWMB 's safety.

Here accelerator pedal motion time refers to the moment that the driver's feet begin to move from the acceleration pedal and the start point at which the driver judges whether braking is needed. It is expressed in the data that the opening of accelerate pedal change from an opposite and stable degree to 0 abruptly (See Fig 5). It is an important guideline to scale the driver's driving habits. With the neural network training and simulation, the change procedure of accelerator pedal and the result of simulation was shown in Fig 5. Obviously, we may notice:(1) Sample data and simulation results's trend is highly accordant;(2)Compared with sample data and simulation results, they all lie in rather safe area, so the time of applying the brake is brought forward and the distance of applying the brake is increased. It is not easy to obtain a global understanding of the procedure of accelerator pedal and the result of simulation. In fact, it is observed that several different characteristics exist during the progress. However, it is enough to verify MCRWMB 's safety. It is worth of pointing out that sample data and simulation results are fitting better in [0.91,0.94] and [0.78,0.83] section, which is marked A, but in section B there is stated warp. This calls for our further studies.

Table 2. The critical safety distance under different speed(km/h)

Speed(km/h)	50	60	70	80	90	100	110	120
ANFIS	40.6	50.6	65.4	79.2	98.1	115.3	142.2	168.4
Calculation	39.6	51.1	64.2	78.8	95.5	112.8	132.8	152.6
MCRWMB	40.9	50.7	65.5	79.5	98.1	116.1	143.6	171.3

Experiment 3 ：compared with other typical models.

According to the settings of ANFIS model and calculation model[12], the authors did 1500 experiments to MCRWMB model, recorded the results and did the analysis and comparison adequately(see Tab 2 to 4). Tab 2 shows that MCRWMB model has the longest safety-critical-distance and the least probability of rear-end collision when the speed is between 50 and 120km/h. Tab 3 shows that combining driver' driving preference, MCRWMB model has longer safety distance than other model within the same length of braking time. And Tab 4 shows that the forward misdiagnosis rate of MCRWMB, i.e., the ratio of the alarms which should but not be given to all the alarms that should be given, is the lowest among these models.

5 Conclusion

To sum up, the multi-agent-based car rear-end warning model has great value in reducing the number of traffic accidents and in raising the utilization ratio of roads. With the factors of driving behavior added in, this model's predominance is mainly embodied in the coordination and cooperation among agents, thus improving the reliability, correctness and autonomy of the alarm system to a large extent. Furthermore, MCRWMB model incarnate that driver have the leading position in driver-vehicle-road closed-loop system. In the meantime, the effectiveness and robustness of MCRWMB model has been verified by the simulated experiment, which is of great theoretical significance and practical value in the study of intelligent transportation system.

References

1. Schultheis, M.T., Simone, L.K., Roseman, E., Nead, R., Rebimbas, J., Mourant, R.: Stopping behavior in a VR driving simulator: A new clinical measure for the assessment of driving. In: Proceedings of the 28th IEEE EMBS Annual International Conference, New York City, USA, August 30-September 3 (2006); SaBP4.8
2. Guan, H., Zhang, L.C., Gao, Z.H.: Research of Driver Optimal Preview Acceleration Integrated Decision Model. In: CMESM (2006)
3. Sekizawa, S., Inagaki, S., Suzuki, T., Hayakawa, S., Tsuchida, N., Tsuda, T., Fujinami, H.: Modeling and Recognition of Driving Behavior Based on Stochastic Switched ARX Model. IEEE Transactions on Intelligent Transportation Systems 8(4) (December 2007)
4. Cheng, X.-y.: Agent computing. Heilongjiang Science and Technology Press, Harbin (2003)
5. Li, H., Dai, J.-h.: Study on Agent-based Modeling and Simulation Method for Satellite Space System. Journal of System Simulation (August 2007)

6. Li, y.: Multi-Agent System And The Application In Prediction And Intelligent Transportation System. East China University of Science and Technology Press, Shanghai (2004)
7. Liu, Y.-f., Wu, Z.-h.: Driver behavior modeling in ACT-R cognitive architecture. Journal of Zhejiang University (Engineering Science) 40(10) (2006)
8. Yu, W., Wu, T.-j., Zhou, J.-f., Luxu, Hong, L.: Describing driving behaviors based on decision optimization model. Journal of Zhejiang University (Engineering Science) 40(4), 704–707 (2006)
9. Mao, X.-j.: Agent oriented software development. Tsinghua Publishing House, Beijing (2005)
10. Chu, L.-l., Chen, S.-y., Zhou, M.: Computational Intelligence Mathematical Basis. Science Press, Beijing (2002)
11. Xu, Z.-w., Zhu, M.-l.: Color-based Skin Detection: A Survey. Journal of Image and Graphic (2007)
12. Hou, Z.-x., Wu, Y.-h., Liu, Z.-w.: A Model of Critical Safety Distance in Highway Based on ANFIS. Computer Engineering and Applications (2004)

Novel Hue Preserving Algorithm of Color Image Enhancement

Caixia Liu

Department of Computer Science, Zaozhuang University
Zangzhuang, P.R. China
cxsqz@126.com

Abstract. Color images usually converted to gray image firstly in traditional image enhancement algorithms. The detail information was easily lost and at the same time, these algorithms enhanced noise while they enhanced image, which lead to the descent of information entropy. With the combination of the characteristics of multi-scale and multi-resolution of wavelet transform, the predominance of histogram equalization and some denoising methods, a novel method of color image enhancement based on Hue Invariability with characteristics of human visual color consciousness in HIS color pattern was presented here. The experimental results showed that this algorithm can enhance color images effectively and cost less time.

Keywords: color image, enhancement, wavelet transform, histogram equalization[1].

1 Introduction

Images are often corrupted owing to channel transmission errors in transmission, faulty image acquisition devices in acquisition, engine sparks, a.c, power interference and atmospheric electrical emissions. Because of above reasons, the objectives of corrupted images are hardly to distinguish. So these images need enhancement processing before objectives recognition. The purpose of enhancement is to stand out the useful information and enlarge the difference of characters in different objects in order to improve the vision effect and stand out the characters. The traditional enhancement algorithms are usually based on the computation of the whole image and the low frequency, high frequency and the noise were transformed synchronously while calculating transformation of the whole image. These algorithms enhance noise signal in image while they enhance image, which leads to the descent of information entropy. At the same time, image enhancement in gray-level images is a well-established area, image enhancement in color images has not received the same attention.

With respect to gray-scale images, color images generally include a richer measurement information. It's important to use color in image processing for two

[1] Western naming convention.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 88–96.
springerlink.com

reasons [1]: firstly, color is a powerful describing tools in image analysis and it can predigest the processing of recognizing and extracting objectives. Secondly, the gray layers that human eyes can distinguish are about twenties kind of. However, human can recognize thousands upon thousands colors. One of the two key problems [2] of color image enhancement is that how to keep Hue Invariability and the other is which method is suitable for image enhancement when Hue Invariability is kept.

The enhancement methods generally divided into two kinds: space domain and frequency domain enhancement methods. The space domain enhancement methods can't enhance the definition and the number of the image feature points but for the uniformity of the whole pixels, which is useful to the further processing.

Image denoising is a hot research area and plays an important role in image procesing. The most intuitionistic methods is to adopt low-pass filter according the characteristics that the noise energy is mainly in high frequency and the image frequency is mainly in a finite area[3]. However, the low-pass filter can remove some useful high-frequency information while removing the noise. So the denoising methods are to find the balance of removing noise and reserving useful high-frequency information[4].

Aimed at these questions, a new algorithm is proposed. The HIS color space is divided into three channels H channel, S channel and I channel. We first make the H channel keep invariable and in the I channel, the image is divided into the low frequency part and the high frequency parts, then histogram equalization processing is only applied to the low frequency part. After that, the wavelet is reconstituted by the low frequency part which has been equalized. In the S channel, we adopt saturation enhancement by exponent stretching.

2 The Algorithm Based on His Color Space

2.1 HIS Color Space

In all of the color spaces, the most popular one is HIS space because of the two advantages: First, I component is independent on other color channels. Secondly, H and S components was closely related to the way that our eyes obtaining color [5]. These characteristics make HIS color space is very suitable for color image processing algorithms with visual system to apperceive color [6]. The characters of HIS color space make it easy to process color image [7].

$$H = \begin{cases} \theta & if \quad B \le G \\ 360-\theta, & if \quad B > G \end{cases} \tag{1}$$

$$\theta = \cos^{-1}\left\{ \frac{1/2[(R-G)+(R-G)]}{\left[(R-G)^2+(R-B)(G-B)\right]^{1/2}} \right\} \tag{2}$$

$$\begin{aligned} S &= 1-\frac{3}{(R+G+B)}[\min(R,G,B)] \\ I &= \frac{1}{3}(R+G+B) \end{aligned} \tag{3}$$

2.2 I Component Enhancement

The histogram equalization is one of good space domain enhancement methods. However, the detail information and noise are almost both exist in high frequency domain and the noise would be magnified and the detail information would be easily lost when we adopt histogram equalization on the whole image [8]. Wavelet analysis is a local analysis method with fixed window and changing window shape, time window and frequency window. The low frequency coefficients reflect the outline information and the high frequency coefficients reflect the detail information and noise after a digital image was decomposed with wavelet transform. At the same time, the visual feeling of the general image is dependent on the low-frequency information.

On this point, we can do histogram equalization just in low frequency domain. So the detail can avoid being blurred and the noise can't be magnified if we just process the low frequency.

1) Wavelet Decomposition

We take one-scale wavelet transform on the component *I*.

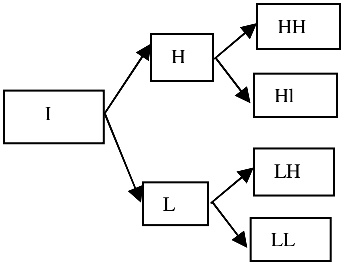
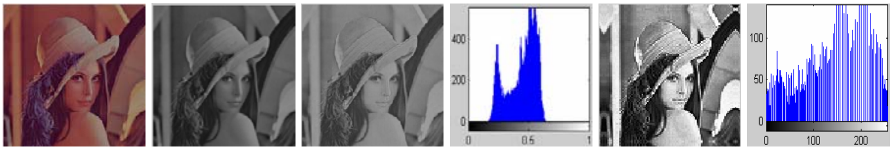


Fig. 1. One scale wavelet decomposition of component *I*

2) Histogram Equalization

The histogram equalization in space domain is used to improve the image contrast by evenly re-distributing the gray level of the image. Based on above analysis, we make histogram equalization on LL1 sub-image of figure 1.

Figure2 (b) shows the image is low light and the detail is illegible. From figure2 (e) we can see that the contrast of low frequency area in I component is enhanced after histogram equalization in figure (b).



(a) (b) (c) (d) (e) (f)

Fig. 2. (a) Original image; (b) gray-level image of (a); (c) LL1 component of (a) after wavelet decomposition; (d) The histogram of (b); (e) the image after histogram equalization in figure (b); (f) the histogram of (d)

3) Denoising

The signal mixed with noise can be expressed as[9,10]:

$$D = F + W \qquad (4)$$

The purpose of denoising is to renew the original signal F from the polluted signal D.

$$\langle D, gm \rangle = \langle F, gm \rangle + \langle W, gm \rangle \qquad (5)$$

Can be gained by decomposing $D = F + W$ in a group of orthogonal basis $B = \{gm\}, (0 \le m \le N)$·

Figure 3 shows that the images with noise. We can see that the Gaussian noise and salt &pepper noise make the image more blurred.

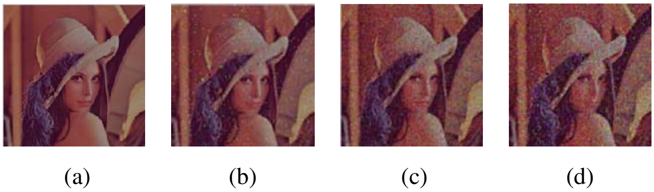

(a) (b) (c) (d)

Fig. 3. (a) Original image, that is figure 2(a); (b) Image (a) with salt &pepper noise; (c) Image (a) with Gaussian noise; (d) Image (a) with both Gaussian noise and salt &pepper noise

Image noise can be divided into gaussian noise and pulse noise according to the feature. Median filter method can well remove pulse noise, but it 's not suitable for removing Gaussian noise[11]. Wavelet transform can well remove Gaussian noise and associated with median filter, the mixed noise with Gaussian noise and pulse noise can be well removed. The noise and detail are mainly in high frequency area after the image is decomposed with wavelet. So we can remove the noise only in high frequency area.

a) Wavelet filtering

In the processing of denoising, the several factors such as the evanishment moment, the regularity and orthogonality, and so on usually can be considered during the wavelet base selection process. After the orthogonal wavelet transformation, the coefficients are non-correlated and such filter effect can be better. The result as follows:

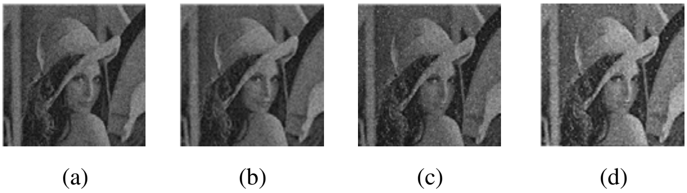

(a) (b) (c) (d)

Fig. 4. (a) The image with Gaussian noise. It's the grave-level of figure 3(c); (b) Image (a) after denoising with wavelet transform; (c) The image both Gaussian noise and salt & pepper noise. It's the grave-level of figure 3(d); (d) Image (a) after denoising with wavelet transform.

As to the Figure 3(d), the wavelet transform is carred in HH1, HL1, LH1 area and then associated with threshold method to denoise. As the scale increasing, the denoising effect will be better.

b) Median filter

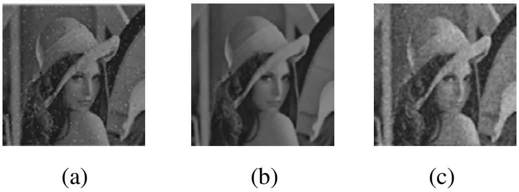
Median filter is one of nonlinear filter technologies. This essence of this method is the median method, the linear filtering method usually takes the smallest mean error of the noise and the useful signal as the standard of filter output, but the median filter are based on the smallest absolute error. The median filter can be constructed according to the basic principle of median method to realize denoising processing of the signal. This filter essence is one kind of sliding window filter, the filter operation is use the signal sampling value on the sliding window center position to substitute the all sampling median value on current window. The strict mathematics definition of median filter is as follows[12]:

Assusme that the length of filter window A is $n=2k+1$ or $n=2k$, the number of observation value is N, $N \geq n$. That is to say $x_1, x_2, \cdots x_N$. When the window A is shiftting on the observation value sequence, the output $med(x_i)$ of standard median filter is

$$med\left(x_i\right) = \begin{cases} x_{k+1}, & n = 2k+1 \\ \left(x_k + x_{k+1}\right)/2, & n = 2k \end{cases} \tag{6}$$

x_k stands for the median value of $2k+1$ or $2k$ number of data of the window.

The following figure shows that the result of median filter denoising. We can see that the noise has become little and the median filter is effective to Salt & pepper noise. (c)shows the result of median filter denoising of figure 4(b).



(a) (b) (c)

Fig. 5. (a) The image both Gaussian noise and salt & pepper noise. It's the grave-level of figure 3(d); (b) Image (a) after denoising with wavelet transform; (c) Image figure 4(b) after denoising with median filter.

4) Wavelet Reconstruction

Associate with histogram equalization, denosing method, the component I can be reconstructed use the LH1,HL1,HH1 and LL1' which is the low frequency domain after histogram equalization. LH1,HL1 and HH1 had been denosied. The result is shown in figure 7.

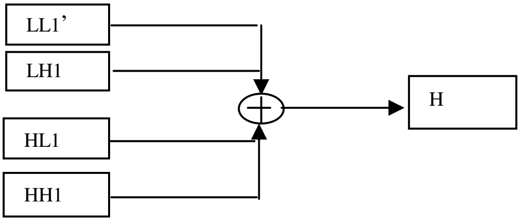


Fig. 6. Wavelet Recostruction

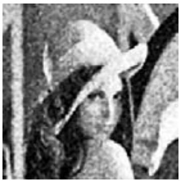


Fig. 7. The reconstructed I component image

2.3 S Component Enhancement

According to the I component enhancement, denoising technologies are used in S component. First, the wavelet transform applied in S component to remove Gaussian noise and median filter applied to remove salt &pepper noise. The results are shown in Figure 8. The results show the (c) is better than (a).

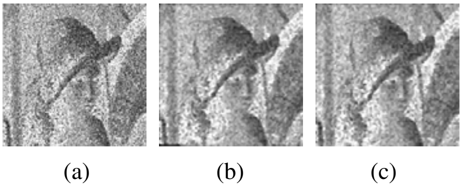


Fig. 8. (a)The S component of Figure 3(d) with both Gaussian noise and salt &pepper noise;(b) Image(a) after denoising with median filter; (c) The image (b) after denoising with wavelet transform

In order to make sure that the color more clearly, we can take Nonlinear exponent adjustment to extend the color changing dynamic range and Enhance its contrast. The exponent stretching formula is

$$S' = S^{\alpha} \qquad (7)$$

and α is the stretch factor, determining saturated degree of saturation component.

In figure 9, (a) is the exponent stretching figure, (b) is the original S component and the result image after stretching is shown in figure 9(c).

Normalize H, S, I component and convert HIS color space to RGB space.

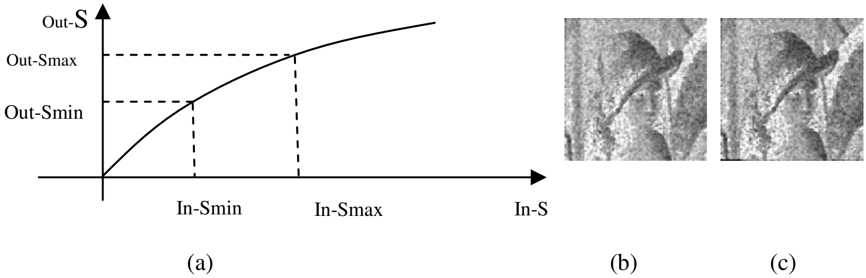


Fig. 9. (a) Exponent stretching. (b) the original S component. (c) The result image after stretching.

3 Result Analysis

We process two noise images and analyse the results.

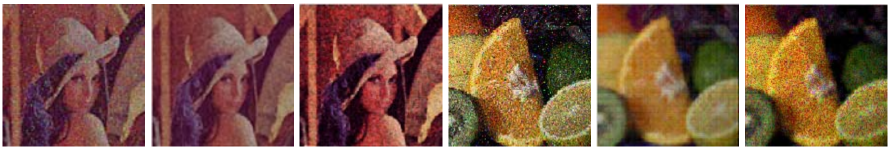


Fig. 10. (a) and (d) the original images with noise. (b) and (e) the enhancement results with original methods. (c) and (f) the enhancement results with ourl methods.

In figure10, (a) and (d) is the image with low lightness and both Gaussian noise and salt&pepper noise. (b) and (e) is use wavlet decomposition and filter to remove noise in RGB color space. We can see that the result is with less noise, but it's blurry and dark. The (c) and (f) is the result with our method and is more clear than (b). However, we can see the noise is more than (b) and it's because the Hue component is intact in order to hold the original hue of the image. The noise lies in H, S, and I component and we had just processed the S and I component. So the image still contains some noise. The situation is the same as in (d), (e) and (f).

As shown in Figure11, The (c1) is the result image with our algorithm and we can see that the (c1) has improved the image contrast and reserved the detail information.

The situation is the same as in (a2), (b2), (c2).

In order to validate the validity of our method, we calculated the image mean, variance, the change rates of luminance and contrast according to the measurement methods of luminance and contrast proposed in literature [13] on figure 13(a1).

$$\begin{aligned} C &= \frac{\overline{\mathrm{var}}(I_{out}(x,y)) - \overline{\mathrm{var}}(I_{in}(x,y))}{\overline{\mathrm{var}}(I_{in}(x,y))} \\ L &= \frac{mean(I_{out}(x,y)) - mean(I_{in}(x,y))}{mean(I_{in}(x,y))} \end{aligned} \tag{8}$$

$I_{in}(x,y)$ is the original image before processed, $I_{out}(x,y)$ is the result image, $\overline{\mathrm{var}}$ and $mean$ stand for calculating the mean of local variance and the mean of the whole image. C is the contrast change rate and L is the luminance change rate.

From the table 1, we can see the image variance changed greatly with our method and extended the dynamic range of the image. Comparing with traditional method, the luminance of (c1) increased by 22 percent and the contrast increased by 10 percent with our method.

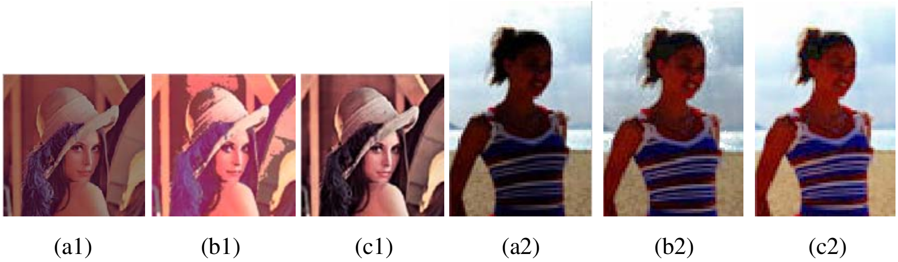


Fig. 11. (a1), (a2) is Original image. (b1), (b2) is the result image after histogram equalization directly on I componet. (c1), (c2) is the result with our algorithm.

Table 1. The results evaluating of several enhancement methods

Evaluate parameters	Original Image	Traditional histogram equalization method	The method with our method
Image mean	85.8092	149.3367	99.5776
Image variance	2097.4	5579.7	6368.1
C		0.7403	0.8605
L		1.6602	2.0361

4 Conclusions

In this paper, a new image enhancement algorithm was proposed based on color images. In the first, The HIS color space was divided into three channels H channel, S channel and I channel. The H channel kept invariable and in I channel, we used wavelet decomposition and histogram equalization in order to reduce noise and improve contrast. After that, the wavelet was reconstructed by the low frequency part which had been equalized. In the S channel, we adopt saturation enhancement by exponent stretching. The algorithm can effectively enhance the color images especially the fuzzy ones with low brightness. At the same time, this method is easy and cost less time. As to the image with noise, we adopt the median filter and wavelet decomposition to remove Gaussian noise and salt &pepper noise which are the main noise in images. The results are better than the original methods.

However, in order the retain the hue of image, the Hue component hasn't been denosing which dued to some noise in Hue component althought this method can improve the clearness and lightness of the noised image. So the next step is to find a good method to remove noise in Hue component.

References

1. Qiuqi, R.: Digital Image Processing. Publishing House of Electronics Industry, Beijing (2001)
2. Ping, W., Hao, C., Yingxin, L.: Color Image Enhancement Based on Hue Invariability. Journal of Image and Graphics 12(7), 1173–1177 (2007)

3. Xie, c., zhang, l., Xu, w.: Overview on Wavelet Image Denoising. Journal of Image and Graphics 7(3), 209–217 (2002)
4. huorong, R., pi, Z., jiali, W.: Novel wavelet image denoising method. Infrared and Laser Engineering 32(6), 643–646 (2002)
5. Shi, M., Li, Y., Zhang, J.: Novel method of color image enhancement. Computer Application 24(10), 69–71 (2004)
6. Yujun, Z.: Image Engineering, pp. 13–22. Tsinghua University Press, Beijing (1999)
7. Xiaowei, H.: The research on Color image processing key technology. Northeastern University (2005)
8. Yanhong, Z., Dewen, H.: An Image Enhancement Algorithm Based on Wavelet Frequency Division and Bi-histogram Equalization. Computer Application and Software 24(11), 159–161 (2007)
9. Donoho, D.L., Johnstone, I.M.: Ideal SpatialA dap tion by Wavelet Shrinkage. Biometrika 81, 425–455 (1994)
10. Zhang, y., Wang, x., Peng, Y.: Adaptive center weighted modified trimmed mean filter. Journal of Tsinghua University (science and technology) 39, 76–78 (1999)
11. Donoho, D.L.: Denoising by Soft thresholding. IEEE Trans on Inf. 41, 613–627 (1995)
12. Hongxia, N., Xinchang, Y., Hexin, C.: De-noising Method Based on Adap tiveWavelet Thresholding. Journal of Jilin University (Information Science Edition) 23(4), 643–646 (2005)
13. Jobson, D.J., Rahman, Z.U., Woodell, G.A.: The statistics of visual representation. In: Processing of SPIE Visual Information processing XI, pp. 25–35. SPIE Press, Washington (2005)

Artificial Immune System Clustering Algorithm and Electricity Customer Credit Analysis

Shu-xia Yang

School of Business Administration, North China Electric Power University
Beijing 102206, China
bjysx216@126.com

Abstract. The real encoding artificial immune system cluster analysis process was put forward firstly, and then the electricity customer credit analysis indexes were determined. At last, according to the customer data of a power company, it classified the electricity customer credit into high, medium and low three categories, and there were two customers with high credit, three customers with medium credit, and one customer with low credit. The results show that the artificial immune system cluster analysis method can obtain the solution once the concentration threshold and cluster number is determined, and its calculation is relatively simple. This method can minimize the requirements of professional knowledge and it is suitable to large volume of data while it is not sensitive to the different data order at the same time. So the artificial immune system cluster analysis has many advantages in obtaining the optimal solution, and this method is feasible to be used in cluster analysis.

Keywords: artificial immune system, cluster, customer credit.

1 Introduction

Cluster analysis can identify the sparse and dense regions, find out the global distribution pattern and the interrelationship between data attributes, so a further study with the cluster analysis is meaningful. Traditional cluster analysis mainly includes system clustering method, decomposition method, addition method, and dynamic clustering method etc. They all belong to the clustering of global comparison which needs to inspect all individuals to decide cluster division, so data needed should be given advance. They have no linear computational complexity and have difficulties in calculating large volume data. At the same time, traditional cluster analysis mainly is likelihood analysis with an assumption that the probability distribution of different date attribute is independent [1-3]. However, the assumption is difficult to stand up in the practical application under large number of data circumstances. Furthermore, a serious coupling may exit among data attributes. As a result of these problems, a further study on traditional clustering analysis is needed in their application conditions, computational complexity and result accuracy etc. Artificial immune system is a parallel and distributed self-adaptive system, it completes the identification and classification through learning, memorizing and associating search. So it has a strong robust information processing capability, regarded as a very important and very significant method [4-7]. The theory of artificial immune system cluster analysis was studied firstly

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 97–102.
springerlink.com

in this paper, the real encoding artificial immune system cluster analysis process was put forward, and then the electricity customer credit analysis indexes were determined. At last, according to the customer data of a power company, it classified the electricity customer credit into high, medium and low three categories and adopted the artificial immune system cluster analysis method to classify the electricity customers, studied the feasibility of using the artificial immune system in cluster analysis.

2 Artificial Immune System Cluster Analysis Principles

Artificial immune system is a global random probability searching method, has characteristics as follows: diversity, tolerance, immunologic memory and distributed parallel processing, self-organizing, self-learning, self-adapting and robustness etc [4]. The global optimal solutions can be converged rapidly through using the antibody represent the feasible solution, and the antigen represent the constraint condition and objective function, and adopting the expected breed rate which can reflect antibodies' promotion and inhibition to select the parental individual, then it can converge to achieve the global optimum solution rapidly. The artificial immune system is constituted by the following elements: coding, producing the initial antibody groups, calculating affinity and expected breed rate, genetic evolution operation and terminate conditions.

The process of artificial Immune system cluster analysis is shown in Fig.1.

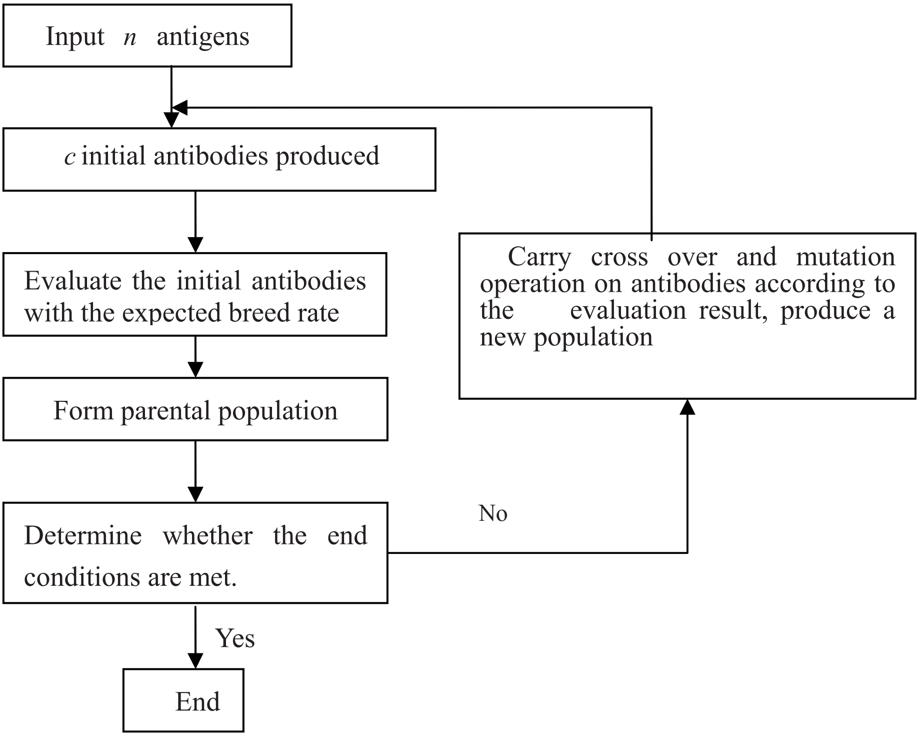

Input n antigens
c initial antibodies produced
Evaluate the initial antibodies with the expected breed rate
Form parental population
Determine whether the end conditions are met.
Yes
End
No
Carry cross over and mutation operation on antibodies according to the evaluation result, produce a new population

Fig. 1. Flow chart of artificial immune system cluster analysis

3 Process of Artificial Immune System Cluster Analysis

The calculation steps of real encoding artificial immune system cluster analysis are described as follows:

Step 1: Input n antigens. In the algorithm, n data objects $X_i(j=1,2,...n)$ are assumed as n arrested antigens input.

Step 2: Form the parental population. The antibodies could be produced by determining the number of categories with experience method according to the specific features and properties of the problem, and select the intuitively appropriate representative point in the data as the initial antibodies; another way is to randomly divide all data into c categories, calculate the mass center of each category, and these centers are taken as initial antibodies; the simplest method is randomly extract c data as the initial antibodies.

Step 3: Evaluate the initial antibodies, and the evaluation criteria is expectant breed rate e_v. The calculation method is shown as follows:

(1) Calculate the concentration of the antibody v [4]

$$c_v = \frac{1}{n}\sum_{w-1}^{n} ac_{vw} \tag{1}$$

Where, if $ay_{vw} \geq T_{ac}$, $ac_{vw} = 1$; otherwise $ac_{vw} = 0$, T_{ac} is the definitived concentration threshold. Similarity degree between antibody v and antibody w is

$$ay_{vw} = \frac{1}{1+H_{vw}} \cdot$$

Real encoding artificial immune system calculates the similarity degree of two antibodies with the application of European Space Second Pan $H_{v,w}$:

$$H_{v,w} = \left[\sum_{i=1}^{m}\left(w_{iv-}w_{iw}\right)\right]^{1/2} \tag{2}$$

Where, m is the number of decision variables, w_{iv} and w_{iw} represent the value of i the decision variable of antibody v and antibody w. ax_v represents the affinity between antigens and antibodies, it is used to show the antibodies' recognition degree to antigens. The affinity between antibody v and the antigen w is:

$$ax_v = J(u,c) \tag{3}$$

Where, $J(u,c)$ denotes the combination degree between antigens and antibodies [4, 8-10], it is generally represented by the value of objective function.

$$J(u,c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\left(X_j - c_i\right)^2 \tag{4}$$

Where u_{ij} is the degree of Vector Group X_j belonging to group $c_i\left(i=1,2,3,\ldots c\right)$, its value lays between 0 and 1. According to the affinity degree, n objects belong to separate groups $c_i\left(i=1,2,3,\ldots c\right)$ each time, in which:

Wherever $k \neq i$, if $X_j - c_i^2 \leq X_i - c_k^2$, $u_{ij}=1$, otherwise $u_{ij}=0$.

(2) Calculate the expectant breed rate of antibody v

$$e_v = \frac{ax_v}{c_v} \tag{5}$$

Antibodies' expectant breed rate [4] also reflects the immune system's promotion for high-affinity antibodies and its inhibition to high concentration antibodies, by which the diversity of antibodies could be maintained, and avoid the disadvantage of getting into local optimum solution prematurely.

Step 4: Form the parental population. Arrange the initial population in the descending order e_v , and select the first c individuals to compose the parental population; then according to descending order ax_v store the first n individuals in the memory base at the same time.

Step 5: Determine whether the end condition $\min J\left(u,c\right)=\sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\left(X_j - c_i\right)^2$ is met. If meet, end the process; otherwise continue the next step of operation, if a better antibody exist in the evaluation results, place it in the memory base, at the same time delete the antibody and the antigen with lower-affinity it in the memory base.

Step 6: Produce a new population. Carry the crossover and mutation operation on antibody population on the basis of the evaluation result in step 4, and then obtain the new population. Take out the stored individuals from the memory base, and constitute new generation population together. Then, turn to step 3.

4 Electricity Customer Credit Analysis Based on Artificial Immune System Cluster Algorithm

Internationally, tradition evaluation elements of enterprise credit were mainly as the following 5 factors: character, ability, capital, guaranty, and operating status. Based on the basic condition of electricity customer in China and according to relative literature [11-14], the credit analysis of electricity customers should be mainly established in the customers' comprehensive qualities including: (1) commercial credit. It was the ability and trust degree of various economic commitments and it was composed as these indicators: enterprise image, legal representative morality, business prospect, asset-liability ratio, profit volume, electricity charge paying proportion, return on net assets, net return on total assets etc. (2) Safety credit. (3) Legal credit. (4) Cooperation credit.

Took the electricity customer credit analysis by taking a power company as example. Relevant data were showed in Table 1.

Table 1. Electricity customer credit index value

Index		Electricity customer					
		A	B	C	D	E	F
commercial credit	enterprise image	bad	good	good	bad	general	general
	legal representative morality	good	general	good	good	general	general
	business prospect	bad	good	general	bad	good	bad
	asset-liability ratio （%）	42.30	43.17	35.23	43.89	36.85	30.9
	profit volume (Ten thousand)	129.9	153	136.8	127.9	140.7	145.3
	electricity charge paying proportion （%）	99.5	99.08	99.9	99.2	98.7	98.9
	return on net assets （%）	8.14	5.45	5.78	5.01	6.02	5.08
	net return on total assets （%）	4.62	4.13	4.11	3.78	3.91	5.3
Safety credit		general	general	good	good	general	good
Legal credit		general	good	general	general	good	general
Cooperation credit		general	general	good	general	general	good

Attention: grade according to literature [15-16]: good for 10, general for 6, and bad for 2.

In analysis, data of enterprise A, B, C, D, E, and F were input as the antigens captured. All data were divided into 3 categories randomly.

The concentration threshold T_{ac} was set as 0.8, the cluster kinds c is 3, that was, c_1 denoted the electricity customers with high credit; c_2 denoted the electricity customers with medium credit; c_3 denoted the electricity customers with low credit. Results were as follows:

$$C_1=(B,A)\,;C2=(F,C,E)\,;C_3=(D)$$

5 Conclusions

Results show that the artificial immune system cluster analysis method can obtain the solution once the concentration threshold and cluster number is determined, and its calculation is relatively simple. This method can minimizes the requirements of professional knowledge and it suitable to large volume of data while it not sensitive to the different data order at the same time. So the artificial immune system cluster analysis has many advantages in obtaining the optimal solution. The artificial immune system is feasible to be used in cluster analysis, but the value of the concentration threshold will affect the cluster result.

References

[1] Marcos, D., Leila, S., Maria, A.S., Cavalcanti, S.C.H.: Adaptive mean-linkage with penalty: a new algorithm for cluster analysis. Chemometrics and Intelligent Laboratory Systems 94, 1–8 (2008)

[2] Zhou, J.-c., Li, W.-j., Zhu, J.-b.: Particle swarm optimization computer simulation of Ni clusters. Trans. Nonferrous. Met. Soc. China 18(2), 410–415 (2008)

[3] Liu, W.: Research of data mining based on fuzzy clustering in comprehensive assessment. Mathematics in Practice and Theory 36(11), 88–92 (2006) (in Chinese)

[4] Zuo, X., Ma, G.-w., Liang, W.-h., Wu, S.-y., Tu, Y.-j.: The forecast of electric power market voidance price bases on manpower immune system small wave network. Generate Electricity by Water 32(1), 13–158 (2006) (in Chinese)

[5] Xie, G., Guo, H.-b., Xie, K.-m., Xu, X.-y., Chen, Z.-h.: The research and progress of artificial immune system algorithms. In: Proceedings of 6th International Symposium on Test and Measurement, vol. 7, pp. 6588–6591 (2005)

[6] Simon, T.P., He, J.: A hybrid artificial immune system and self organizing map for network intrusion detection. Information Sciences 178, 3024–3042 (2008)

[7] Tan, K.C., Goh, C.K., Mamun, A.A., Ei, E.Z.: An evolutionary artificial immune system for multi-objective optimization. European Journal of Operational Research 187, 371–392 (2008)

[8] Liao, G.-C., Tsao, T.-P.: Application embedded chaos search immune genetic algorithm for short term unit commitment. Electric Power Systems Research (71), 135–144 (2004)

[9] Timmis, J., Hone, A., Stibor, T., Clark, E.: Theoretical advances in artificial immune systems. Theoretical Computer Science 403, 11–32 (2008)

[10] Liu, T., Wang, Y.-c., Wang, Z.-j.: One kind of cluster-algorithm bases on manpower immune system. Computer Project and Design 25(11), 2051–2053 (2004) (in Chinese)

[11] Lee, T.S., Chiu, C.C., Chou, Y.-C., Lu, C.-J.: Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics & Data Analysis 50, 1113–1130 (2006)

[12] Li, X., Liu, G.-y., Qi, J.-x.: Fuzzy neural and chaotic searching hybrid algorithm and its application in electric customer's credit risk evaluation. J. Cent. South Univ. Technology 14(1), 140–143 (2007)

[13] Li, X., Yang, S.-x., Huang, C.-f.: Credit evaluation of electricity customers based on fuzzy multi-attribute decision making method. Power System Technology 28(21), 55–59 (2004) (in Chinese)

[14] Wu, W.-t.: Analysis on credit rating to electricity customers. Beijing Chang Ping North China Electric Power University (2004) (in Chinese)

[15] Wei, H.-l., Jiang, D.: Study of appraisal method of new product development project. Value Project (4) (2002) (in Chinese)

[16] Wang, L., Yang, J., Xiong, S.-w.: An multi-objective optimization algorithm based on artificial immune system. Journal of Wuhan University of Technology 30(2), 116–118 (2008) (in Chinese)

Virtual Space Sound Technique and Its Multimedia Education Application

Xinyu Duan

Media Research Institute, Anyang Normal University, Henan, China
dxy@aynu.edu.cn

Abstract. Generally research about acoustic perception starts late and not so mature like visual technique, when talking about whole virtual reality system. Paper introduced concept and principle of virtual space sound. Paper mainly studied VRML's sound processing technique, including AudioClip node and Sound node. AudioClip is used to provide sound source information. Its field's value is to designate where we can acquire a pre-record sound file, and how to control when broadcasting. Sound node is mainly used to produce sound field, as being sound transmitter. In Sound node, we can designate various broadcasting parameter, for example the position of sound source and sound's 3D form of expression. Here we demonstrate an Acoustic field ellipse sphere mechanism. Finally in combination with multimedia education, give a spacial sound effect testing example.

Keywords: space sound, auditory sensation, acoustical simulation, audio channel.

1 Introduction

It is well-known that auditory and visible sensations constitute the two most important kinds of information channel for human being to apperceive outside world. Theirs information carriers are separately sound and image media. In compare with vision system, research about acoustic perception starts late, when talking about the whole virtual reality system. Its relative manipulation technique is too not so mature like visual technique [1].

Even though it is, it doesn't mean existing virtual reality system must be silent. On the contrary, many virtual demonstrating systems possess quite abundant audio broadcast channel and exquisite audio play effect. But through carefully analyzing its formation mechanism, we understand these acoustic effects are quite at odds with system principle of virtual environment, where virtual reality system demands three dimension acoustical simulation. Stepping back, here we can think them as a kind of accompaniment music, just for fun.

2 About Space Sound Technique

Traditionally, acoustical simulation is to mean the physical process of sound producing, processing and broadcasting, by employing computer to proceed digital simulation or

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 103–110.
springerlink.com

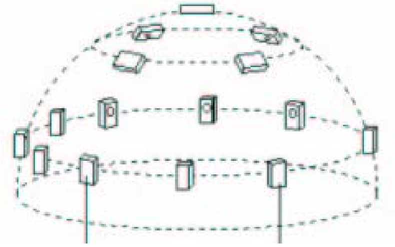


Fig. 1. Acoustical simulation system

auxiliary manipulation, such as environment noise analysis, hall tone quality analysis etc. However when applied to virtual environment this special domain, great changes have taken place to tradition acoustical simulation technique, no matter in its intension connotation or extension denotation.

The earliest research about sound perception and location can be restrospected to Duplex theory, putting forward by Rayleigh in 1907. The theory considered two major factors for human being to proceed sound location are ITD and IID. ITD means two ears' inter-aural time difference, while IID means two ears' inter-aural intensity difference.

For example, if sound source is located in the left region of head, so the sound source is near to left ear when comparative to right ear. The practical effect is uttered sound will first arrive left ear, and then a period later arrived right one. The existed inter-aural time difference received between left ear and right ear for a same sound source is called ITD. The fact is sound intensity will be attenuated with distance increased, when spreading through air. Here we call the inter-aural intensity difference arrived to two ears as IID.

There are a lot of limitations when applying above method, just simply locating sound source according to distance between sound source and dual-ear, in practical. To suppose people's dual ears is located in symmetrical position, so the distance differences of left and right two points which are mirrored by medial axis, from the midpoint of dual-ear link line while expanding outside to form a circular conical surface, are totally equivalent, in relative to dual-ears. According to Duplex theory, we will not be able to distinguish these tow sound source location and direction. But in reality, human's two ears can easily tell them apart accurately. So we are almost unable to correctly explain people's solid perceptive phenomenon about three dimension sound, by using Rayleigh's Duplex theory.

After that, peoples are mainly committing themselves to research of human ear's complicated structure. Many scholars agree with one accord. Just because many existed special and complicate structures of mankind ear, that makes people properly apperceive sound, coming from any locations or directions. Recent year's researches further demonstrate that the variation of sound wave, from sound source to eardrum, can be thought as a sound wave filtering activity of people's ears. We can define the transformation function, when sound wave is propagated from free field to eardrum, as a head-related transfer function, shorten for HRTF.

Under the basis of HRTF propagation function, scientists are successfully transform monophony sound to three dimension space sound, characteristic with dual ear perceptive effects. This can help us simulate the apperceive circumstances of human ear, for different sound frequency and different sound location.

Due to the reason of structure complexities of human ear and human body, it is also relatively difficult to accurately locate sound in technique. Another important factor is different people having different ear shape and physiological feature. This also makes sound location difficult and HRTF function non-normalization for multi users.

Now on theory, for any sound to exert ITD, IID function or HRTF coefficient, we can virtually locate sound source on any space position. Practically, it is far from enough just researching the simulation and broadcast process of sound signal. There is an instinctly analyzing and perception process when human being received sound, to identify every sound source contained in. This is so called auditory perception calculation (APC). APC started with Hilbert Breuman's work about acoustic spectacle analysis in 1990. It is now a relative new research area [2].

3 VRML Sound Effect Principles

In VRML built virtual environment, to produce 3D sound effect we should append following three kinds of nodes in scene structure. They are Sound node, AudioClip node and MovieTexture node. Among them, MovieTexture and AudioClip are used to create sound source. While Sound node is used to create sound field and to designate sound's broadcasting mode. Some of the digital sound files that can be cited in VRML scenes are WAV, MIDI and MPEG-1 [3].

3.1 Principle of AudioClip Nodes

AudioClip node is used to provide scene sound source information. Its field's value is to designate where we can acquire a pre-record sound file, and how to control when broadcasting. AudioClip node can only appear in Sound node's source field. Its syntax format is described below.

```
AudioClip {
   url [ ] #exposedField  MFString
   description " " #exposedField  SFString
   loop FALSE #exposedField  SFBool
   pitch 1.0 #exposedField  SFFloat
   startTime 0.0 #exposedField  SFTime
   stopTime 0.0 #exposedField  SFTime
   duration_changed #eventOut  SFTime
   isActive #eventOut  SFBool}
```

About some of the carried fields' function in AudioClip, we make following brief introduction.

url field's value designates a url address list for sound files. This file list is arrayed from high priority to low priority. This field's value describes a concrete location for needed sound file in WWW. When broadcasting, explorer will try to call a sound file from the first address of list. If this file not be found or can not be opened, explorer will try the second one, and so forth until found a file. Then system will read in this file and

Table 1. StartTime、stopTime、loop、pitch coaction relationship

loop field’s value	relation among startTime, stopTime and pitch field’s value	effect
TRUE	stopTime ≤startTime	unlimited cycle play
TRUE	startTime< stopTime	cycle play until stopTime time
FALSE	stopTime ≤startTime	stop after one play cycle
FALSE	startTime< （startTime +duration/pitch） ≤stopTime	stop after one play cycle
FALSE	startTime< stopTime<（startTime +duration/pitch）	less than one cycle, stoped in stopTime time

play it in the scene. If none file can be called and opened, or if url filed is empty, system will not produce any sound effect. In the same time when loading in a higher priority sound file, explorer maybe is now broadcasting a low priority file. AudioClip node supports WAV and MIDI sound format.

description field’s value is to setup a character string for describing cited sound. In practical, this string will be shown in net page, simultaneously when explorer is broadcasting relative file. It function is very similar to character explanation when explorer can not render or display an image on network page. description field’s default value is an empty string.

loop field’s function is to specify whether loop playback the cited sound file. If its value is TRUE, the cited sound will be unlimited cycle play or cycle play until stopTime time, according to the state of startTime or stopTime value. If its value is FALSE, the sound will stop after one play cycle or play duration is less than one cycle, stoped in stopTime time, according to the state of startTime or stopTime value.

pitch field’s value is to designate tone factor when sound broadcasting. Its value is to speed up or slow down sound’s play speed. If its value is between 0.0 and 1.0, sound tone will be dropped down and broadcasting speed slowed down. If its value is 1.0, sound will be played normally. If its value is greater than 1.0, sound tone will be raised up and playback be speeded up. Its default value is 1.0.

startTime field’s value is to set the beginning time of cited sound file. Its unit is second. Its reference start time is Greenwich Time 12 of midnight, in January 1, 1970. Its default value is 0.0 second. When a sound is being played back, pay attention to that the set_startTime event will be neglected.

stopTime field’s value is to set the stop time of sound play. It is corresponding to startTime field. Its default value is also 0.0 second.

All these four fields, loop, startTime, stopTime and pitch, jointly control the sound play mode of AudioClip node. Its resultant effect is shown in Table 1.

duration_changed output event is used to dispatch out sound’s broadcasting time (its unit seconds) after sound file be opened. So other nodes will know current sound’s duration of time. The output time has nothing to do with playback speed fixed by pitch field. That is to say the output time is its inherent broadcasting time. When a file can not be opened, or system is unable to determine its duration of time, duration_changed event will output event value -1, instead of broadcasting duration of time.

IsActive output event is used to transmit a TRUE event when this sound file starts to play, while a FALSE event when sound file being stopped. Through isActive event, system let other nodes know whether or not that current sound file is being broadcast.

3.2 Principle about Sound Node

AudioClip and MovieTexture nodes' function is to create sound source. But sound source must be source field's value of Sound node when program. Then we can obtain rendered music effect in VRML scene.

Sound node is mainly used in VRML environment to produce sound field as being sound transmitter. In Sound node, we can designate various broadcasting parameter, for example the position of sound source and sound's 3D form of expression. In VRML, virtual sound source can be located in any position bound to its local coordinates. Sound source will transmit sound out in a spherical surface, or in an ellipsoid surface. Sound can be dimensional sound form of manifestation, or either be monaural sound manifestation.

Sound node's syntax format is below [4].

```
Sound {
   source NULL #exposedField  SFNode
   location 0.0 0.0 0.0 #exposedField  SFVec3f
   direction 0.0 0.0 1.0 #exposedField  SFVec3f
   intensity 1.0 #exposedField  SFFloat
   maxBack 10.0 #exposedField  SFFloat
   maxFront 10.0 #exposedField  SFFloat
   minBack 1.0 #exposedField  SFFloat
   minFront 1.0 #exposedField  SFFloat
   priority 0.0 #exposedField  SFFloat
   spatialize TRUE #exposedField  SFBool}
```

source field's value is to provide sound source for broadcasting music in VRML scene. Its value can be any one node of AudioClip or MovieTexture. Its default value is NULL, indicating none importance of any sound source.

location field's value is to designate practical position in current coordinates for sound transmitter. Its default value is 0.0 0.0 0.0, indicating default sound source is in the origin of current coordinate system.

direction field's value is to setup the space orientation for sound transmitter, in VRML world. It is right the sound emission direction vector. This vector is composed by three floating point numbers, separately state a three dimension vector's X, Y, Z value. Its default value is 0.0 0.0 1.0. That is illustrating the sound source's default space orientation is Z axis's forward direction. It is also the user's facing direction.

intensity field's value is to assign sound transmitter's sound intensity, also called sound volume. This value can vary in the scope from 0.0 to 1.0. 0.0 indicates silence, while 1.0 indicates maximum volume. Pay attention that if the value is greater than 1.0, music maybe distortion. When necessary a high sound level, the best way is to rerecord the music in higher volume, and then reintroduction it into scene. Its default value is 1.0.

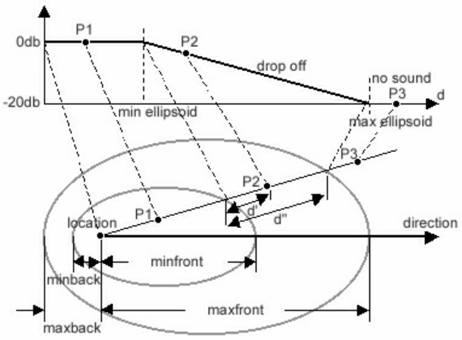


Fig. 2. Acoustic field ellipse sphere

maxBack field's value is to setup a hypothetical straight-line distance in current coordinates, from sound transmitter's location and along direction field's opposite direction. Beyond this distance, user will be unable to hear the sound. This field demands its value greater than or equal to 0.0. Its default value is 10.0.

maxFront field's value is to appoint a hypothetical straight-line distance in current coordinates, from sound transmitter's location and along direction field's positive direction. Beyond this distance, user will be unable to hear sound. This field demands its value greater than or equal to 0.0. Its default value is also 10.0.

minBack field's value is to setup a hypothetical straight-line distance in current coordinates, from sound transmitter's location and along direction field's opposite direction. Beyond this distance, sound volume is beginning to attenuate. And until the specifying position of maxBack field's value, here sound volume attenuated to 0.0. This field demands its value greater than or equal to 0.0. Its default value is 1.0.

minFront field's value is to setup a hypothetical straight-line distance in current coordinates, from sound transmitter's location and along direction field's positive direction. Beyond this distance, sound level is beginning to attenuate. And until the specifying position of maxFront field's value, here sound volume attenuated to 0.0. This field demands its value greater than or equal to 0.0. Its default value is 1.0.

priority field's value is to determine sound emitter's transmitting priority level, with the scope from 0.0 to 1.0. 1.0 indicates maximum priority, while 0.0 indicates minimum priority. Its default value is 0.0.

spatialize field's value is to make sure whether to execute sound's dimensional manipulation or not. So that user is able to practically identify sound transmitter's concrete location in a three dimension space. Its value determines sound broadcasting mode, as being a spatial point or being environment sound. If its value is TRUE, to be spatial point mode and user is able to ascertain sound source. If its value FALSE, to be environment sound mode and user unable to ascertain sound source. Its default value is TRUE.

In Sound node, above introduced six fields, location, direction, maxBack, maxFront, minBack and minFront are all coactive to configure two invisible elliptical spheres, representing VRML's sound spread scope. Of which, minBack and minFront fields defined an interior elliptical sphere, while maxBack and maxFront fields defined an exterior elliptical sphere. See the following figure 2.

These two elliptical spheres stand for a virtual sound transmitter’s broadcasting space and volume variation, when an explorer moving in VRML world. If located within minimum scope elliptical sphere (interior), user can hear the maximum intensity sound, moreover equivalent everywhere because of short distance to sound source. If located between minimum scope elliptical sphere (interior) and maximum scope elliptical sphere (exterior), user can hear different sound volume according to his location. The further the distance, the smaller level we can hear. Until to the margin of maximum scope elliptical sphere, sound volume is attenuated to zero. If explorer located beyond maximum scope elliptical sphere, user will be unable to hear any sound transmitted by this sound source, because of the considerable distance to transmitter.

The sound’s intensity variation regulation can be iconically illustrated with rectangular coordinate system in Fig.2. Among them, three points p1, p2, p3 separately represent a point at interior space within minimum scope elliptical sphere, a point at space between minimum scope elliptical sphere and maximum scope elliptical sphere and a point at exterior space beyond maximum scope elliptical sphere. Here, abscissa axis representing distance, while vertical axis representing sound volume.

Normally for producing reality attenuation effect, we should make maxBack fileld’s value as 10 times of minBack fileld’s value. Simultaneously make maxFront fileld’s value as 10 times of minFront fileld’s value. If the ratio greater than 10, so sound’s attenuation speed is slow. If the ratio smaller than 10, the sound’s attenuation speed is relatively fast.

4 Virtual Sound Application

Following is a concrete application example of stereo sound in multimedia education area. We give out its program method followed by specific scene comment.

```
#VRML V2.0 utf8
Shape {……
            geometry Sphere {radius 0.25}}
Transform {
   rotation 1 0 0 1.571
   children [
       Shape {……
                    geometry Cylinder { radius 1.0}}
       Shape {……
                    geometry Cylinder { radius 6.0}}]}
Sound {
   source AudioClip {
         url "wind.wav"
         loop TRUE}
   maxFront 6.0
   maxBack 6.0
   minFront 1.0
   minBack 1.0
   spatialize TRUE}
```

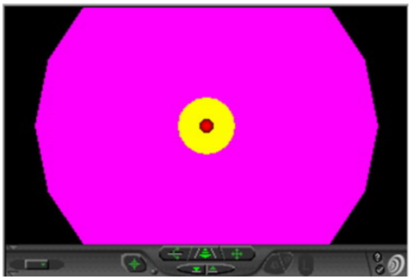


Fig. 3. Spacial sound effect testing

Here we all construct three geometry models in the scene. In the center it is a red ball, representing the position of sound source. Others are two cylinder top surfaces which have been rotated 90º around X axis, representing different sound field. The bigger one's radius is 6.0 (in pink colour). Its scope is same with Sound node's maxFront and maxBack fields' value. The smaller one's radius is 1.0 (in yellow colour). Its scope is same with Sound node's minFront and minBack fields' value.

This example's rendering effect in Cosmo Player is shown in Fig.3. When gradually moving the scene upwards to approach red ball (which representing sound source) by using ⇩ control button, the sound effect is little by little reflected and strengthened. Now if gradually moving the scene left side or right side to departure red source ball by using ✥ control button, we can obviously become aware of weakening sound effect. Here please pay attention to the sound volume transmitted by left and right speaker. We can definitely tell the sound source is on the right side or left side, relatively to current observation position. That is the efficacy when Sound node's spatialize field's value setup of TRUE.

References

[1] Philips, R.L.: Media View: A General Multimedia Digital Publication System. Communications of the ACM 34(7) (1991)

[2] Jiaoying, S.: Foundation of Virtual Reality and Applied Algorithm. Science Press, Beijing (2002)

[3] Xinyu, D.: Virtual Reality Foundation and VRML Program. Higher Education Press, Beijing (2004)

[4] ISO/IEC 14772-1:1997, VRML97 International Standard [EB/OL], The VRML Consortium, `http://www.vrml.org`

Survey on Association Rules Mining Algorithms

Mingzhu Zhang and Changzheng He

Sichuan university, Chengdu, China
mingzhu.86@gmail.com, hechangzheng@163.com

Abstract. In recent years, the association rule mining as an important component of data mining attracts many attentions. Up to now, there are many literatures on the association rules, scholars study the association rules mining deeply from improving the algorithm to proposing a new perspective, and thus, there is a great development in the field. In this paper, we perform a high-level overview of association rules mining methods, extensions and we also put forward some suggestions of the research in the future. With a rich body of literature on this theme, we organize our discussion into the following four themes: improving the algorithms to increase mining efficiency, proposing new algorithm to extend the notion of association rules, the integration of association rules and classification, the research on parameter such as support and confidence.

Keywords: association rules, survey, algorithm, efficiency, notion.

1 Introduction

Association rules mining, which is widely used in medicine, biology, business and so on, is presented in transaction database by R. Agrawal et al. in 1993[1]. From then on, association rules attracted a lot of interest, lots of researchers worked on it, from static rules to dynamic rules, from linguistics term rules to numeric rules, and so on, the study range of association rules is becoming larger and larger. In this paper, we perform a high-level overview of association rules mining methods, extensions and we also put forward some suggestions of the research in the future. With a rich body of literature on this theme, we organize our discussion into the following four themes: improving the algorithms to increase mining efficiency, proposing new algorithm to extend the notion of association rules, the integration of association rules and classification, the research on parameter such as support and confidence.

2 Basic Principles of Association Rules

Let $I = \{I_i, i = 1, 2, ..., m\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that T $\subseteq$ I. Associated with each transaction is a unique identifier, called its TID. We say that a transaction T contains X, a set of some items in I, if X $\subseteq$ T. An association rule is an implication of the form $X \Rightarrow Y$, where $X, Y \subset I$, and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ holds in

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 111–118.
springerlink.com

the transaction set D with confidence c if c% of transactions in D that contain X also contain Y. The rule $X \Rightarrow Y$ has support s in the transaction set D if s% of transactions in D contain $X \cup Y$ [2].

Given a set of transactions D, the problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support and minimum confidence respectively.

3 The Research Direction of Association Rules

In 1994, the famous association rules algorithm Apriori was presented by R. Agrawal et al. [2]. From then on, association rules were studied deeper. There are two ways to improve the algorithms to increase mining efficiency: Apriori-based algorithms and not Apriori-based algorithms.

3.1 Improving the Algorithm to Increase Mining Efficiency

Among all the association rules mining algorithms, Apriori played an very important role, it had great impact on the later researches. The basic principle of Apriori is this: The first pass of the algorithm simply counts item occurrences to determine the large 1-itemsets. A subsequent pass, say pass k, consists of two phases. First, the large itemsets Lk-1 found in the (k-1)th pass are used to generate the candidate itemsets Ck, using the apriori-gen function . Next, the database is scanned and the support of candidates in Ck is counted. The algorithm will be ended until the candidate large k itemset Ck is empty [2].

Apriori can extract the rules from the databases efficiently, but it costs too much time, and its efficiency is not too high. After that, many researchers put forward improved algorithms based this, that is, Apriori-based algorithm. Mannila et al. presented sampling algorithm, the method is based on careful combinatorial analysis of the information obtained in previous passes; in this way, it is possible to eliminate unnecessary candidate rules [3]. After that, Toivonen improved this algorithm [4]. Park et al. presented DHP algorithm which is hash-based. The candidate generated by this algorithm is mach smaller than previous methods, especially generated candidate 2-itemsets. And the generation of smaller candidate sets enables us to effectively trim the transaction database size at a much earlier stage of the iterations [5]. In 1995, Park produced an algorithm called Parallel Data Mining (PDM), which is the parallel edition of the DHP to mine sequential association rules [6]. In 1996, Agrawal presented mining association rules in the parallel way, and put forward 3 algorithms, which are called Count Distribution algorithm, Data Distribution algorithm, Candidate Distribution algorithm respectively to mine association rules [7]. Savasere et al. presented partition-based algorithm, which scans the database only twice, reducing the I/O operation observably, and thus improved the efficiency of the algorithm [8].

The Apriori heuristic achieves good performance gained by reducing the size of candidate sets. However, it has some disadvantages, such as: it is costly to handle a huge number of candidate sets when with a huge number of frequent patterns.

And there are other not Apriori-based algorithms. M.J. Zaki et al. presented an algorithm using cluster technique. The method can extract all the rules efficiently by scanning the database only once, and the time it costs is only 25% of the Apriori [9]. Nicolas et al. presented A-Closed algorithm in 1999, which using the closed itemset lattice framework. This approach is very valuable for dense and/or correlated data and it can lowering the algorithm computation cost by generating a reduced set of association rules without having to determine all frequent itemsets [10]. E. Yilmaz et al. presented Randomized Algorithm 1, which produces a small set of association of association rules in polynomial time. This method addressed the problem that most algorithms like Apriori may cause exponential computing resource consumption [11]. JIAWEI HAN et al. presented FP-tree algorithm, it develop an efficient FP-tree based mining method, FP-growth, for mining the complete set of frequent patterns by pattern fragment growth. FP-growth method is efficient and scalable for mining both long and short frequent patterns, and performs very well [12]. Ant colony is a newly algorithm, and it is able to provide more condensed rules than the Apriori method. The computational time is also reduced [13].

3.2 Proposing New Algorithm to Extend the Notion of Association Rules

At the early time, association rules studied the relationship among the data in the transaction databases only, at the later time, researchers extended the nation range of it, trying to extract interest information on different instance. Such these association rules as: generalized association rules, quantitative association rules, fuzzy association rules, rare association rules, and so on.

Generalized Association Rules. In 1995, Srikant et al. introduced the generalized association rules, which consists the items from any level of the taxonomy. A generalized association rules is an implication of the form $X \Rightarrow Y$, where $X \subset I, Y \subset I, X \cap Y = \varnothing$, and no item in Y is an ancestor of any item in X. Generalized association rules considered the presence of taxonomies, and allowed the items in the association rules to the leaf-level items in the taxonomy[14]. Andreas Myka et al. presented an approach called Prutax to improve the efficiency of mining generalized association rules. Prutax achieves an order of magnitude better performance than former approaches [15]. Takahiko Shintani et al. presented a parallel algorithm based a shared-nothing parallel machine for mining association rules with classification hierarchy, and the parallel algorithm improved the mining efficiency. And they proposed three parallel algorithms: NPGM、HPGM、H-HPGM to improve the mining efficiency[16]. According to the most of the commercially available mining systems integrate loosely with data stored in DBMSs, Shiby Thomas et al. present several alternatives for formulating as SQL queries association rule generalized to handle items with hierarchies on them and sequential pattern mining[17].

Quantitative Association Rules. Srikant, R et al. proposed the quantitative association rules. Quantitative association rules is that whose attributes are quantitative or categorical. They thought if all attributes are categorical or the quantitative attributes have only a few values, quantitative association rules problem could be mapped to the Boolean association rules problem [18]. Keith C. C. Chan et al. proposed APACS2 for mining quantitative association rules from very large databases. The technique has the advantage that it does not require any user-supplied thresholds which are often hard to determine, and it allows us to discover both positive and negative association rules [19]. Ansaf Salleb-aouissi et al. proposed a mining quantitative association rules system called QUANTMINER which based on a genetic algorithm. Thus, this system can dynamically discover "good" intervals in association rules by optimizing both the support and the confidence [20]. In 2007, RANJEET et al. proposed a algorithm for finding the boundaries of attributes domains dynamically. It can identify the boundaries of domains of (quantitative) attributes dynamically [21].

Fuzzy Association Rules. In 1997, Keith C.C. Chan et al. introduced F-APACS for mining fuzzy association rules. F-APACS employs linguistic terms to represent the revealed regularities and exceptions. The linguistic representation is especially useful when those rules discovered are presented to human experts for examination [22]. The fuzzy association rules is an implication that, 'If X is A then Y is B', to deal with quantitative attributes. X, Y are set of attributes and A, B are fuzzy sets which describe X and Y respectively. In 1998, Chan Man Kuok et al. proposed an algorithm which used fuzzy technique for mining association rules, and the algorithm has solved the problem of sharp boundary [23]. Hung-Pin Chiu et al. used cluster technique to extract fuzzy association rules, proposing CBFAR algorithm, and this approach outperforms a known Apriori-based fuzzy association rules mining algorithm [24]. In 2008, P. Santhi Thilagam et al. proposed a algorithm which extraction and optimization of fuzzy association rules using multi-objective genetic algorithm, the method is efficient in many scenarios [25].

Rare Association Rules. Association rule mining among frequent items has been extensively studied in data mining research. However, in recent years, there has been an increasing demand for mining the infrequent items (such as rare but expensive items). Hyunyoon Yun presented an association rule discovery technique called RSAA to mining significant data's association rules. That is, the significant rare data associated with specific data in a way that the rare data occur simultaneously with the specific data more frequently than the average co-occurrence frequency in the database, RSAA is more efficient than Apriori and MSApriori [26]. Paper [27] presented two efficient approaches for mining rare association rules, MBS and HBS, these two approaches extract the rules by scanning the database only twice, and they also can be applied to mine association rules efficiently among frequent items with limited length.

Other Association Rules. Besides the association rules presented above, there are other association rules. Such as: mining numeric association rules[28]; mining Simple association rules[29]; mining maximal association rules[30]; mining Temporal association rules[31]; and so on.

3.3 The Integration of Association Rules and Classification

Bing Liu et al. propose to integrate association rule mining and classification rule mining. The integration is done by focusing on mining a special subset of association rules, called class association rules. Its objectives are to generate the complete set of CARs and to build a classifier from the CARs. The classifier built this way is more accurate than that produced by the state-of-the-art classification system.C4.5 [32]. Paper [33] proposed the algorithm called GARC, which could filter out many candidate itemsets in the generation process, thus, the set generated by it is much smaller than that of CBA. Paper [34] presented RMR algorithm, the proposed algorithm resolves the overlapping between rules in the classifier by generating rules that does not share training objects during the training phase, resulting in a more accurate classifier. The literatures about association rules and classification are a lot, such as: paper [35] presented CARSVM model, which integrates association rule mining and support vector machine; paper [36] proposed a classification model that is based on easily interpretable fuzzy association rules and fulfils both efficiency criteria; paper [37] examine the effect that the choice of ARM parameters has on the predictive accuracy of CARM methods; paper [38] presents two algorithms that use decision trees to summarize associative classification rules. The obtained classification model combines the advantages of associative classification and decision trees, thus resulting in a more accurate classifier than traditional decision trees.

3.4 The Research on Parameter Such as Support and Confidence

Support and confidence play an important role in the association rules mining, most classical association rules mining algorithm base the two parameters for extracting association rules. Paper [39] addressed the problem about defining the minimum supports of itemsets when items have different minimum supports. The numbers of association rules and large itemsets obtained by the mining algorithm using the maximum constraint are also less than those using the minimum constraint. Paper [40] proposed a method of match as the substitution of confidence. The generated rules by the improved method reveal high correlation between the antecedent and the consequent when the rules were compared with that produced by the support–confidence framework. Paper [41] presented two algorithms, MMS_Cumulate and MMS_Stratify, and extended the scope of mining generalized association rules in the presence of taxonomies to allow any form of user-specified multiple minimum supports. Paper [42] presented Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support.

4 Future Works

Above all, at present, the association rules mining is studied very widely, we think it could from the following points to continue this task: (a) because of the huge database and the great number of the rules, the efficiency of rules mining attracts researchers most interest, they work hard on how to reduce the cost; (b) most algorithms mine so many rules that it’s impossible for the users to use the results, so it is important to mine fewer but more important rules; (c) so many fruit on association rules bringing another problem, that is the leak of information, so privacy preserving is put forward to protect the information while mining association rules, and it attract many researchers interest.

5 Summarize

This paper summarizes the various theories and technologies of the association rules mining over a decade, and has summed up the scattered fruits and the advantages and disadvantages of each method in the field. Different aspects of the association rules are used in different applications; it should be practical-oriented and come up with effective mining algorithm or association rules in new cases, to solve the practical problems.

Acknowledgement

The authors are grateful for the incisive and helpful comments of the anonymous referees and the editors.

References

[1] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in very large databases. In: Proceedings of the ACM SIGMOD Conference, pp. 207–216 (1993)

[2] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceeding of the 20th International Conference on Very Large Databases, pp. 407–419 (1994)

[3] Mannila, H., Toivonen, H., Verkamo, A.: Efficientalgorithm for discovering association rules. In: AAAI Workshop on Knowledge Discovery in Databases, pp. 181–192 (1994)

[4] Toivonen, H.: Sampling Large Databases for Association Rules. In: Proceedings of the 22nd VLDB Conference, Mumbai (Bombay), India (1996)

[5] Park, J.S., Chen, M.-S., Yu, P.S.: An Effective Hash-Based Algorithm for Mining Asociation Rules. In: Proceedings of the 1995 ACM SIGMOD international conference, pp. 175–186 (1995)

[6] Park, J.S., Chen, M.-S., Yu, P.S.: Efficient Parallel Data Mining for Association Rules. In: Proceedings of the fourth international conference, pp. 31–36 (1995)

[7] Agrawal, R., Shafer, J.C.: Parallel mining of association rules. IEEE Transactions on Knowledge and Data Engineering 8(6), 962–969 (1996)

[8] Savasere, A., Omiecinski, E., Navathe, S.: An Efficient Algorithm for Mining Association Rules in Large Databases. In: Proceedings of the 21st International Conference, pp. 432–444 (1995)

[9] Zaki, M.J., Parthasarathy, S., Ogihara, M., Li, W.: New Algorithms for Fast Discovery of Association Rules

[10] Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering Frequent Closed Itemsets for Association Rules. In: Proceedings of the 7th International Conference, pp. 398–416 (1999)

[11] Yilmaz, E., Triantaphyllou, E., Cken, J., Liao, T.W.: A Heuristic for Mining Association Rules in Polynomial Time. Mathematical and Computer Modelling 37, 219–233 (2003)

[12] Han, J., Pei, J., Yin, Y., Mao, R.: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery 8, 53–87 (2004)

[13] Kuo, R.J., Shih, C.W.: Association rule mining through the ant colony system for National Health Insurance Research Database in Taiwan. Computers and Mathematics with Applications 54, 1303–1318 (2007)

[14] Srikant, R., Agrawal, R.: Mining Generalized Association Rules. In: Proceedings of the 21st International Conference, pp. 407–419 (1995)

[15] Hipp, J., Myka, A., Wirth, R., Guntzer, U.: A New Algorithm for Faster Mining of Generalized Association Rules. In: Proceedings of the Second European Symposium, pp. 74–82 (1998)

[16] Shintani, T., Kitsuregawa, M.: Parallel Mining Algorithms for Generalized Association Rules with Classification Hierarchy. In: Proceedings of the 1998 ACM SIGMOD international conference, pp. 25–36 (1998)

[17] Thomas, S., Sarwagi, S.: Ming Generalized Assoiation Rules and Sequential Patterns Using SQL Queries

[18] Srikant, R., Agrawal, R.: Mining Quantitative Association Rules in Large Relational Tables. In: Proceedings of the 1996 ACM SIGMOD international conference, pp. 1–12 (1996)

[19] Keith, C.C., Au, C.W.-H.: An Effective Algorithm for Mining Interesting Quantitative Association Rules. In: Proceedings of the 1997 ACM symposium, pp. 88–90 (1997)

[20] Salleb-Aouissi, A., Vrain, C., Nortet, C.: QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules

[21] Kumar, R., Kumar, P., Ananthanarayana, V.S.: Finding the boundaries of attributes domains of quantitative association rules using abstraction- A Dynamic Approach. In: Proceedings of the 7th Conference, pp. 52–58 (2007)

[22] Keith, C.C., Au, C.W.-H.: Mining Fuzzy Association Rules. In: Proceedings of the sixth international conference, pp. 209–215 (1997)

[23] Kuok, C.M., Fu, A., Wong, M.H.: Mining Fuzzy Association Rules in Databases. ACM SIGMOD Record 27(1), 159–168 (2008)

[24] Jing, W., Huang, L., Luo, Y., et al.: An Algorithm for Privacy-Preserving Quantitative Association Rules Mining. In: Proceedings of the 2nd IEEE International Symposium, pp. 315–324 (2006)

[25] Santhi Thilagam, P., Ananthanarayana, V.S.: Extraction and optimization of fuzzy association rules using multi-objective genetic algorithm. Pattern Analysis & Applications 11(2), 159–168 (2008)

[26] Yun, H., Ha, D., Hwang, B., Ryu, K.H.: Mining association rules on significant rare data using relative support. The Journal of Systems and Software 67, 181–191 (2003)

[27] Zhou, L., Yau, S.: Efficient association rule mining among both frequent and infrequent items. Computers and Mathematics with Applications 54, 737–749 (2007)

[28] Chena, Y.-L., Weng, C.-H.: Mining association rules from imprecise ordinal data. Fuzzy Sets and Systems 159, 460–474 (2008)

[29] Chen, G., Wei, Q., Liu, D.: Geert Wets. Simple association rules (SAR) and the SAR-based rule discovery. Computers & Industrial Engineering 43, 721–733 (2002)

[30] Bi, Y., Anderson, T., McClean, S.: A rough set model with ontologies for discovering maximal association rules in document collections. Knowledge-Based Systems 16, 243–251 (2003)

[31] Li, Y., et al.: Discovering calendar-based temporal association rules. Data & Knowledge Engineering 44, 193–218 (2003)

[32] Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining

[33] Chen, G., Liu, H., et al.: A new approach to classification based on association rule mining. Decision Support Systems 42, 674–689 (2006)

[34] Thabtah, F.A., Cowling, P.I.: A greedy classification algorithm based on association rule. Applied Soft Computing 7, 1102–1111 (2007)

[35] Kianmehr, K., Alhajj, R.: CARSVM: A class association rule-based classification framework and its application to gene expression data. Artificial Intelligence in Medicine 44, 7–25 (2008)

[36] Pach, F.P., Gyenesei, A., Abonyi, J.: Compact fuzzy association rule-based classifie]. Expert Systems with Applications 34, 2406–2416 (2008)

[37] Coenen, F., Leng, P.: The effect of threshold values on association rule based classification accuracy. Data & Knowledge Engineering 60, 345–360 (2007)

[38] Chen, Y.-L., Hung, L.T.-H.: Using decision trees to summarize associative classification rules. Expert Systems with Applications (2008), doi:10.1016/j.eswa.2007.12.031

[39] Lee, Y.-C., Hong, T.-P., Lin, W.-Y.: Mining association rules with multiple minimum supports using maximum constraints. International Journal of Approximate Reasoning 40, 44–54 (2005)

[40] Wei, J.-M., Yi, W.-G., Wang, M.-Y.: Novel measurement for mining effective association rules. Knowledge-Based Systems 19, 739–743 (2006)

[41] Tseng, M.-C., Lin, W.-Y.: Efficient mining of generalized association rules with non-uniform minimum support. Data & Knowledge Engineering 62, 41–64 (2007)

[42] Yan, X., Zhang, C., Zhang, S.: Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. Expert Systems with Applications (2008), doi:10.1016/j.eswa.2008.01.028

Application of Rough Set Theory and Fuzzy LS-SVM in Building Cooling Load

Xuemei Li[1,2], Ming Shao[1], and Lixing Ding[2]

[1] School of Mechanical and Automotive Engineering, South China University of Technology,
Guangzhou, China, 510640
[2] Institute of Built Environment and Control, Zhongkai University of
Agriculture and Engineering, Guangzhou, China, 510225
aulimin@126.com

Abstract. Cooling load prediction models are expected to play an important role in building. Specifically, they will support advanced energy-saving optimal control systems. In this study, we develop an applicable predictor scheme, rough set theory (RST) and fuzzy least square support vector machine (LS-SVM) for building cooling load forecasting. Through the real-world building environment dataset experiment, we have proved that the reduction feature set exacted by RST has good subject-independence and intrinsic good separability. Fuzzy LS-SVM predictor demonstrated promising prediction accuracy, better generalization ability and more rapid execution speed than most of the all benchmarking methods listed in this study.

Keywords: Cooling load prediction, Rough set theory, Fuzzy LS-SVM.

1 Introduction

Because of significant amounts of energy consumed in heating, ventilating and air-conditioning (HVAC) systems, there have been efforts to study energy use performance and promote good operational practice in buildings. In order to improve the operation of HVAC systems, it is necessary to have reliable optimization routines [1]. Accurate prediction of the building dynamic cooling load is a key for the optimal control of a HVAC system.

Accurately predicting building cooling load is a challenging work. Several prediction methods have ever been adopted [2–4], such as the admittance and Fourier methods, the transfer function method, the neural networks method, the Monte Carlo simulation method, the Kalman filter methods, and so on. Although these methods have alleviated difficulties in air-conditioning load modeling and prediction to some extent, but there are still some problems. Support vector machine (SVM) overcomes the shortcoming that ANN structure relies on the experience of designer. It solves high dimension, local minimum, and small samples well. Least square SVM (LS-SVM) is developed from standard SVM. It substitutes inequality constraints of standard SVM to equality constraints, it substitutes quadratic programming to solving linear system of equation. Thus, it reduces calculation complexity, speed up solving, and enhanced interference-free ability.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 119–126.
springerlink.com

Since the available information in building cooling load is not always accurate and measurable but rather uncertain and incomplete, SVM doesn't effectively deal with vagueness and uncertainty information. In order to solve those problems, a hybrid prediction model composed of rough set component and LS-SVM component is presented in this paper. Our proposal aims to make great use of the advantages of Rough Set theory in pre-processing large data, eliminating redundant information and overcoming the disadvantages of slow processing speed caused by SVM approach.

LS-SVM can captures the geometric characteristics of feature space without deriving weights of networks from the training data and it is capable of extracting the optimal solution with small training data. Especially fuzzy LS-SVM prediction model based on time-domain membership can put different memberships according to the impact extent of history data. It reduces the impact of early data to the current producing process, and improves the accuracy of real-time prediction. The model was applied to cooling load prediction, the experiment results show that the ensemble of RST-based feature selection and Fuzzy LS-SVM-based predictor performs better than other combinations in terms of computational expense, classification accuracy and generalization performance.

2 Basic Principle of RST and Fuzzy LS-SVM

2.1 Basic Concept of Rough Set Theory

Rough set theory was introduced by Pawlak [5] to deal with imprecise or vague concepts. The main ideas of rough set theory are deducing decision or classification rules through reduction knowledge under the premise of maintaining the same classification ability. Rough set theory deals with information represented by a table called an information system. It is composed of 4-tuple as follow:

$$S = < U, A, V, f > \tag{1}$$

where U is a finite set of objects, called the universe, A is a finite non-empty set of attributes, $V=U_{\mathrm{a}}$ $_{\mathrm{A}}$, V_a is a domain of attribute a, and f: $U \times A \rightarrow V$ is the total decision function called the information function such that $f(x,a) \in v_a$, for information system is also seen as a decision table assuming that $A = C \bigcup D$ and $C \bigcap D = \phi$, where C a set of condition is attributes and D is a set of decision attributes.

In RST, the essential distinction is made between objects that may definitely be classified to a certain category and those that may possibly be classified. Considering specific attributes, objects are indiscernible if they are characterized by the same information. Let $P \subseteq A$ and $x_i, x_j \in U$. Then x_i and x_j are indiscernible by the set of attributes in S if and only if (1) holds.

$$f(x_i, a) = f(x_j, a), \forall a \in B \tag{2}$$

An elementary set is defined to be a set of all indiscernible objects with respect to specific attributes. Thus, given $B \subseteq A$, we define an equivalence relation on U, called the B indiscernibility relation, denoted by $U/IND(B)$, The equivalence relation

U/*IND*(*B*) will generate the *B* elementary sets in *S*, that contain the *B*-indiscernible objects.

Given $B \subseteq A$ and $x \subseteq U$, the *B*-lower approximation BY_* of set *Y* and the *B*-upper approximation BY^* are defined as follows:

$$\begin{cases} BY_* = \{x \in U : [x]_{U/IND(B)} \subseteq Y\} \\ BY^* = \{x \in U : [x]_{U/IND(B)} \bigcap Y = \phi\} \end{cases} \tag{3}$$

Furthermore, considering the attributes *B*, we can associate an index α_B of the accuracy of approximation for any set $Y \in U$ as follows:

$$\alpha_B = |BY_*| / |BY^*| \tag{4}$$

Where $|BY^*|$ indicates the cardinality of a set BY^*. Obviously $0 \leq \alpha_B \leq 1$, if $\alpha_B = 1$, then *Y* is and ordinary set with respect to *B*; if $\alpha_B < 1$, then *Y* is a rough set with respect to *B*. Let $B, Q \subseteq A$, the positive region of classification *U*/*IND*(*Q*) with respect to the set of attributes *B*, which is defined as

$$POS(Q) = \bigcup_{x \in U/IND(Q)} A \tag{5}$$

Thus, the dependency of *Q* on *B* is defined as

$$\gamma_B(Q) = |POS_B(Q)| / |U| \tag{6}$$

It was shown previously that the number $\gamma_B(Q)$ expresses the degree of dependency between attributes *B* and *Q* , it may be now checked how the coefficient $\gamma_B(Q)$ changes when some attribute is removed. In other words, what is the difference between $\gamma_B(Q)$ and $\gamma_{B-\{\alpha\}}(Q)$. Attribute important $\{\alpha\}$ about decision attribute is defined by

$$\sigma_{BQ}(\alpha) = \sigma_B(\alpha) - \sigma_{B-\{\alpha\}}(D) \tag{7}$$

Let $S = (U, A, V, f)$ be a decision table, the set of attributes $B(B \subseteq C)$ is a reduce of attributes C, which satisfies the following conditions:

$$\gamma_B(D) = \gamma_C(D), \gamma_B(D) \neq \gamma_{B'}(D), \forall B' \subset B \tag{8}$$

A reduce of condition attributes *C* is subset that can discern decision classes with the same accuracy as *C*.

2.2 Fuzzy Least Square Support Vector Machine

Least Square Support Vector Machine (LS-SVM) is a new technique for regression. When LS-SVM is used to model cooling load, the input and output variables should be chosen firstly. Hence, this paper takes history values as the input parameters. The hourly building cooling load is chosen as the model's output.

Given a training data set $\{(x_1,y_1),\ldots\ldots,(x_N,y_N)\}$ with input data $x_N \in R^n$ and output data $y_N \in R$. In order to get the function dependence relation, SVM map the input space into a high-dimension feature space and construct a linear regression in it. The regression function is expressed with

$$y = f(x) = w^T \varphi(x) + b \tag{9}$$

with $\varphi(\cdot): R^n \to R^{n_\varphi}$, a function which maps the input space into a so-called higher dimensional feature space, *w* and *b* are the regression parameters to be solved. LS-SVM regression estimation satisfies this inequality:

$$\min_{w,b,e} L_P(w,e) = \frac{1}{2}\|w\|^2 + \frac{\gamma}{2}\sum_{i=1}^{N} e_i^2 \tag{10}$$
$$s.t.\ \ y_i = w \cdot \varphi(x_i) + b + e_i,\ \ i = 1,2,...,N$$

where, positive real number *γ* is tuning constant. The loss function of LS-SVM changes inequality constraints to equality constraints, and it's different from the standard SVM. Lagrange function is introduced in as the following:

$$L_D(w,b,e_i,\alpha) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^{l} e_i - \sum_{i=1}^{l} \alpha_i (w^T \varphi(x_i) + b + e_i - y_i) \tag{11}$$

Here α_i, (i =1,2,…,N) are Lagrangian multiplier. The first order conditions for optimality are given by:

$$\begin{cases} \dfrac{\partial L_D}{\partial w} = 0 \to w = \sum_{i=1}^{l} \alpha_i \varphi(x_i), \\ \dfrac{\partial L_D}{\partial b} = 0 \to \sum_{i=1}^{l} \alpha_i = 0, \\ \dfrac{\partial L_D}{\partial e_i} = 0 \to \alpha_i = \gamma e_i, i = 1,2,...,N \\ \dfrac{\partial L_D}{\partial \alpha_i} = 0 \to y_t = w^T \varphi(x_t) + b + e_i, i = 1,2,...,N \end{cases} \tag{12}$$

Then, optimization can be transformed to:

$$\begin{bmatrix} 0 & \vec{1}^T \\ \vec{1} & \Omega + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{13}$$

where $y = (y_1, y_2, \cdots, y_l)^T$, $I = (1,1,...,1)^T$, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_N)^T$, Ω is a square matrix. The elements $\Omega_{i,j} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, Any function $k(x_i, x_j)$ satisfying Mercer's condition can be used as the kernel function. In this paper, Gaussian function is also selected as the kernel function, whose expression is shown as follows:

$$\varphi(x_i) \cdot \varphi(x_j) = k(x_i, x_j) \equiv \exp(\frac{\|x_i - x_j\|^2}{\delta^2}) \tag{14}$$

where δ^2 is the width parameter of Gaussian kernel.

To replace the dot product $\varphi(x_i)^T \varphi(x_i)$, let C = Ω + γ^{-1}I, then the formula (8) can be written as:

$$\begin{bmatrix} \vec{1} & C \\ 0 & \vec{1}^T \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix} \tag{15}$$

From formula (8), the regression parameters *α* and *b* can be solved as follows:

$$b = \frac{\vec{1}^T C^{-1} y}{\vec{1}^T C^{-1} \vec{1}} \tag{16}$$

$$\alpha = C^{-1}(y - b\vec{1}) \tag{17}$$

Thus, LS-SVM regression model is:

$$y = w^T \varphi(x) + b = \sum_{i=1}^{N} \alpha_i k(x_i, x_j) + b \tag{18}$$

Yu et al. introduce a least squares fuzzy LS-SVM model by formulating the classification problem as

$$\begin{aligned} &\min_{w,b,e} J_P(w,e) = \frac{1}{2}\|w\|^2 + \frac{\gamma}{2}\sum_{i=1}^{N} \mu_i e_i^2 \\ &s.t.\ \ y_i = w \cdot \varphi(x_i) + b + e_i,\ \ i = 1,2,...,N \end{aligned} \tag{19}$$

In this paper, we introduced the improved the fuzzy LS-SVM with bilateral-weight method proposed in [], which formulated the classification problem as follows

$$\begin{aligned} &\min_{w,b,e} J_P(w,e,\eta) = \frac{1}{2}\|w\|^2 + \frac{\gamma}{2}\sum_{i=1}^{N} [\mu_i e_i^2 + (1-\mu_i)\eta_i^2] \\ &s.t. \begin{cases} w \cdot \varphi(x_i) + b = 1 - e_i \\ w \cdot \varphi(x_i) + b = -1 + \eta_i \\ e_i, \eta_i \geq 0,\ i = 1,2,...,N \end{cases} \end{aligned} \tag{20}$$

Then the Lagrangian funcation can be constructed as:

$$L(w,b,e,\eta,\alpha,\beta) = J_P(w,e,\eta) - \sum_{i=1}^{N} \alpha_i [w \cdot \varphi(x_i) + b - 1 + e_i] - \sum_{i=1}^{N} \beta_i [w \cdot \varphi(x_i) + b + 1 - \eta_i] \tag{21}$$

Where $\{\alpha_i\}_{i=1}^{N} \geq 0, \{\beta_i\}_{i=1}^{N} \geq 0$ are the Lagrangian multipliers. The optimal point will in the saddle point of the Lagrangian function, i.e. $\max\limits_{\alpha,\beta} \min\limits_{w,b,e,\eta} L(w,b,e,\eta,\alpha,\beta)$

Thus, we can obtain

$$\begin{cases} \partial L / \partial w = 0 \to w = \sum_{i=1}^{N} (\alpha_i + \beta_i)\varphi(x_i) = 0 \\ \partial L / \partial b = 0 \to \sum_{i=1}^{N} (\alpha_i + \beta_i) = 0 \\ \partial L / \partial e_i = 0 \to c\mu_i e_i - \alpha_i = 0 \\ \partial L / \partial \eta_i = 0 \to c(1-\mu_i)\eta_i + \beta_i = 0 \\ \partial L / \partial \alpha_i = 0 \to w \cdot \varphi(x_i) + b = 1 - e_i \\ \partial L / \partial \beta_i = 0 \to w \cdot \varphi(x_i) + b = -1 + \eta_i \end{cases} \tag{22}$$

Then we can get the following linear equation by simple substitutions.

$$\begin{cases} \sum_{i=1}^{N}(\alpha_i+\beta_i)=0 \\ \sum_{j=1}^{N}(\alpha_i+\beta_i)K(x_i,x_j)+b=1-\dfrac{\alpha_i}{cu_i} \\ \sum_{j=1}^{N}(\alpha_i+\beta_i)K(x_i,x_j)+b=-1-\dfrac{\beta_i}{c(1-u_i)} \end{cases} \tag{23}$$

From the 2*N*+1 equations in Eq.23, we can derive the 2*N*+1 unknown variables $\{\alpha_i\}_{i=1}^{N},\{\beta_i\}_{i=1}^{N},b$, accordingly, we can obtain the following predictor.

$$z=\sum_{i=1}^{N}(\alpha_i+\beta_i)k(x_i,x_j)+b \tag{24}$$

3 Data Preparation Based on RST and Fuzzy LS-SVM Model

The hourly climate data and building cooling load from May to September is considered in this paper. The input parameters of SVM such as dry-bulb temperature, relative humidity and solar radiation intensity are taken from the climate database of Guangzhou in the typical meteorology year. DeST [7] is used to calculate the office building's hourly cooling loads, which are taken as the basic values to compare with the predicted values from LS-SVM.

Building hourly cooling load data were sampled per hour in every day. Since building cooling load data are always sampled with noise, the data preprocessing need to be implemented to correction errors, of which the threshold test and building HVAC theory-based check are two commonly used methods. At last, the data should be normalized treatment to improve the efficiency of computation.

3.1 Fuzzy LS-SVM Predictor Design

In this paper, Rough set theory and fuzzy LS-SVM is used to build a prediction model for building environment information. The algorithm in details is:

Step 1: Given a training sample set, we firstly discretize them if the sample attributes values are continuous, then feature extraction by rough set theory, thus we can get a minimal feature subset that fully describes all concepts by attribute reduction.

Step 2: Fuzzy membership is generated by linear transformation function in terms of initial score by expert's experience introduced in [6,8].

Step 3: Initialize the fuzzy LS-SVM algorithm, train the proposed predictor cooling load from attributes reduced samples as training data set, the parameters *α*, *β*, *b* are adapted by above mentioned formulas, Thus, the prediction value of cooling load *z* is:

$$z=\sum_{i=1}^{N}(\alpha_i+\beta_i)k(x_i,x_j)+b$$

Step 4: Test the prediction regression model along with the continuous dynamically acquired data.

The forecast mechanism can be seen in Fig.1.

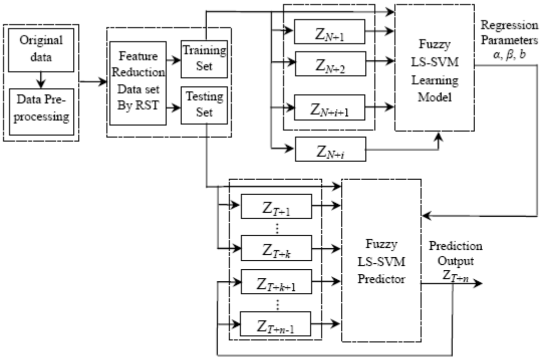


Fig. 1. The predictor scheme of hybrid model

3.2 Experiment Result

To evaluate the prediction applicability of building cooling load with RS-FLSSVM, for comparison purposes, BPNN, LSSVM, FLSSVM and RS-FLSSVM are also conducted in the experiments. We used Data in May and June as the training sample and those data in July, August and September as the testing sample in order to find the similarities and the instant dynamics from the hourly cooling load data.

To evaluate its adaptability to different predict methods, three error indicators: Relative Mean Errors (RME), Mean Absolute Relative Error (MARE) and Root Mean Squared Errors (RMSE) are applied as performance indices.

Table 1. Evaluation Indices between MLANN and AOSVR

Predictor	Evaluation Indices		
	RMERR%	MARERR%	RMSRERR%
BPNN	0.79	4.18	11.84
LSSVM	0.45	1.65	5.56
FLSSVM	0.42	1.36	4.47
RS-FLSSVM	0.37	1.24	3.29

Three evaluation indices of different predictors are listed in Tab.1, we can see that the RS-FLSSVM predictor gave higher accuracy than the other predictor. From the general view, the hybridized model of rough sets and FLSSVM dominates the other three predictors, revealing that the hybridized model is the most excellent predictor. Faced with variable building environment condition, the hybridized model shows strong powerful robustness.

4 Conclusion

Cooling load prediction models are expected to play an important role in building. Specifically, they will support advanced energy-saving optimal control systems. In this study, we develop an applicable predictor scheme, rough set component and fuzzy least square SVM for building cooling load forecasting. Through the real-world building environment dataset experiment, we have proved that the reduction feature set exacted by RST has good subject-independence and intrinsic good separability. Fuzzy LS-SVM predictor demonstrated promising prediction accuracy, better generalization ability and more rapid execution speed than most of the all benchmarking methods listed in this study.

References

[1] Bida, M., Kreider, J.F.: Monthly-averaged cooling load calculations-residential and small commercial buildings. J. Sol. Energy Eng. Trans. ASME 109(4), 311–320 (1987)

[2] Al-Rabghi Omar, M.A., Al-Johani Khalid, M.: Utilizing transfer function method for hourly cooling load calculation. Energy Convers Manage 38(4), 319–332 (1997)

[3] Ben-Nakhi Abdullatif, E., Mahmoud Mohamed, A.: Cooling load prediction for buildings using general regression neural networks. Energy Convers Manage 45(13-14), 2127–2141 (2004)

[4] Mui, K.W., Wong, L.T.: Cooling load calculations in subtropical climate. Build Environ. 42(7), 2498–2504 (2007)

[5] Pawlak, Z.: Rough set. Int. J. Comp. Inf. Sci., 341–356 (1982)

[6] Wang, Y.Q., Wang, S.Y., Lai, K.K.: A new fuzzy support vector machine to evaluate credit risk. IEEE Transactions on Fuzzy Systems 13, 820–831 (2005)

[7] DeST Development Group in Tsinghua University. Building environmental system simulation and analysis-DeST. China Architecture & Building Press, Beijing (2006)

[8] Yan, Z., Wang, Z., Xie, H.: Joint application of rough set-based feature reduction and Fuzzy LS-SVM classifier in motion classification. Medical and Biological Engineering and Computing (46), 519–527 (2008)

An Efficient Feature Selection Algorithm Based on Hybrid Clonal Selection Genetic Strategy for Text Categorization

Jiansheng Jiang[1], Wanneng Shu[2], and Huixia Jin[3]

[1] Faculty of Mechanical and Electronic Engineering, China University of Petroleum – Beijing, PRC. No.18, Fuxue Road, Changping Zone, Beijing 102249, China
johnjjs@sina.com
[2] College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China
shuwanneng@yahoo.com.cn
[3] Department of Physics and Telecom Engineering, Hunan City University, Yiyang 413008, China
jinhuixia1980@163.com

Abstract. Feature selection is commonly used to reduce dimensionality of datasets with thousands of features which would be impossible to process further. At present there are many methods to deal with text feature selection. To improve the performance of text categorization, we present a new feature selection algorithm for text categorization, called hybrid clonal selection genetic algorithm (HCSGA). Our experimental results, comparing HCSGA with an extensive and representative list of feature selection algorithms, show that HCSGA leads to a considerable increase in the classification accuracy, and is faster than the existing feature selection algorithms.

Keywords: Text categorization, Feature selection, Feature extraction, Hybrid clonal selection genetic algorithm.

1 Introduction

As the number of electronic documents available on the Internet daily increases, effective retrieval or filtering of text information has become an important in many information organization and management tasks. An increasingly useful tool for managing this vast amount of data is text categorization. Feature selection (FS) is commonly used to reduce dimensionality of datasets with thousands of features which would be impossible to process further [1]. One of the problems in which FS is essential is text categorization. A major problem of text categorization is the high dimensionality of the feature space; therefore, FS is the most important step in text categorization.

FS is important for clustering efficiency and effectiveness, because it not only condenses the size of the extracted feature set but also reduces the potential biases embedded in the original feature set. FS commonly has employed such feature

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 127–134.
springerlink.com

selection metrics as term frequency (TF) (which denotes the occurrence frequency of a particular term in the document collection), $TF \times IDF$ (in which IDF denotes the inverse document frequency measured by $\log(N / DF)$, where N is the number of documents in the collection and DF is the number of documents containing the focal term). The aim of FS methods is the reduction of the dimensionality of the dataset by removing features that are considered irrelevant for the classification. In this paper, we propose a new algorithm for FS, called hybrid clonal selection genetic algorithm (HCSGA). Proposed algorithm is applied to text features of bag of words model in which a document is considered as a set of words or terms and each position in the input feature vector corresponds to a given term in original document.

The remainder of the paper is organized as follows. Section 2 presents the related works. Section 3 describes the HCSGA algorithm. Section 4 presents the experimental evaluation of the HCSGA algorithm. Finally, we conclude with a summary and future research direction in section 5.

2 Related Works

There are numerous statistical classification methods and machine-learning techniques that have been applied to text categorization in recent years. In [2] three methods consist of centric, orthogonal certain, and LDA/GSVD, which are designed for reducing the dimension of clustered data are used for dimensional reduction in text categorization. In [3] proposes a subset search procedure based on ant colony optimization (ACO) for speech classification problem. In [4] propose a term frequency method to select the feature vectors for neural network document categorization.

Genetic algorithm (GA) was developed by Holland and works well on mixed and combinatorial problem. Applying GA to the FS problem is straightforward: the chromosomes of the individuals contain one bit for each feature, and the value of the bit determines whether the feature will be used in the classification. Using the wrapper approach, the individuals are evaluated by training the classifiers using the feature subset indicated by the chromosome and using the resulting accuracy to calculate the fitness. In [5] GA is due to find an optimal binary vector in which each bit corresponds to a feature. A ‘1’ or ‘0’ suggests that the feature is selected or dropped, respectively. The aim is to find the binary vector with the smallest number of 1’s such that the classifier performance is maximized. Adriana, P. et al. presents a GA, called Olex-GA for the induction of rule-based text categorization [6]. GA has been used to search for feature subsets in conjunction with several categorization methods such as neural networks, and k-nearest neighbors [7]. Besides selecting feature subsets, GA can extract new features by searching for a vector of numeric coefficients that is used to transform linearly the original features. In this case, a value of zero in the transformation vector is equivalent to avoiding the feature. Raymer et al. combined the linear transformation with explicit FS flags in the chromosomes [8], and reported an advantage over the pure transformation method.

The GA can provide an optimal solution but is possible for the algorithm to stick at local optimal solutions. Clonal selection algorithm (CSA) is one of the most widely employed artificial immune optimization (AIO) approaches [9]. It is based on the clonal selection principle, which explains how an immune response is mounted, when a non-self antigenic pattern is recognized by the B cells. Compared with the corresponding GA, CSA can enhance the diversity of the population, avoid the prematurely to some extent, and have a faster convergence speed. Thus, we propose to abstract the merit of GA and CSA, namely hybrid clonal selection genetic algorithm is proposed. The experiments demonstrate that, in most cases, the proposed HCSGA finds subsets that result in the best accuracy, while finding compact feature subsets, and performing faster than other common methods.

3 HCSGA Algorithm

In this subsection, we are providing a general description of the HCSGA algorithm.

Assume a set of documents $D=\{d_1,d_2,...,d_n\}$ and a set of classes $C=\{c_1,c_2,...,c_m\}$, where each document is labeled with one or more classes. In addition, each document contains one or more words, otherwise features, from a vocabulary $W=\{w_1,w_2,...,w_v\}$. For the sake of simplicity, we consider only the binary document representation, where documents are represented as vectors of binary attributes, indicating which words occur and which do not occur. Generally, a text categorization system consists of several essential parts including feature extraction and feature selection. After preprocessing of text documents, feature extraction is used to transform the input text document into a feature set(feature vector). FS is applied to the feature set to reduce the dimensionality of it. The performance of selected feature subsets is measured by invoking an evaluation function with the corresponding reduced feature space and measuring the specified classification result. The best feature subset found is then output as the recommended set of feature to be used in the actual design of the classification system. This process is shown in Fig. 1

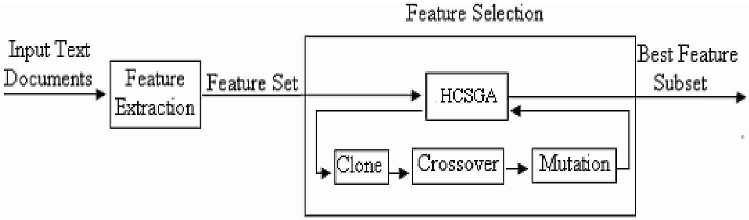


Fig. 1. The diagram of the HCSGA feature selection algorithm

The HCSGA algorithm starts by randomly generating an initial population ($popsize$) of antibodies in a given bounds for the problem (antigen) considered. Each antibody which means a candidate solution is represented by a binary string of bits. The length of bit string is suitably selected by the user to obtain a reasonable precision for the problem. The antibodies are evaluated over an affinity function and then cloned proportionally to their affinities [10]. A subpopulation is constructed with an antibody and its clones. The clones are subjected to a hyper mutation process inversely proportional to their affinity. The maturated clones are then evaluated over the affinity function, and the best antibody of each subpopulation is selected for surviving. The antibody population is updated by replacing the antibodies having m lowest affinities with the new ones generated randomly. With this replacement, the diversity of antibody population is maintained so that the new areas of the search space can be potentially explored. These processes are repeated until a termination criterion is attained.

The main steps of proposed feature selection algorithm are as follows:

Step 1 (Initialization): Set the termination criterion, and the algorithm parameters, such as the mutation probabilities P_m, antibody population size $popsize$, and the multiplying factor α. Then randomly initialize the antibody population $A(0) = \{a_1(0), a_2(0), ..., a_n(0)\}$. At last set evolution generation $k = 0$;

Step 2 (Calculate the affinity): calculate the affinity of initial population.

Step 3 (Clonal Operation): For each antibody, calculate the clone scale q, then apply clonal operator.

$$T^C(A(k)) = [T^C(a_1^{'}(k)), T^C(a_2^{'}(k)), ..., T^C(a_n^{'}(k))] \tag{1}$$

where $T^C(a_i^{'}(k)) = Q_i \times a_i^{'}(k)$, $i = 1, 2, ..., n$. Q_i are q_i dimension row vector. $T^C(a_i^{'}(k))$ is the q_i clone of antibody $a_i^{'}(k)$. We define:

$$q_i(k) = round(\alpha \times f(a_i^{'}(k)) / \sum_{j=1}^{n} f(a_j^{'}(k)), i = 1, 2, ..., n \tag{2}$$

where $f(a_i^{'}(k))$ is the affinity of the antibody $a_i^{'}(k)$, α is the multiplying factor and $round(.)$ is the operator that rounds its argument toward the closest integer. After the clone step, the population becomes:

$$B(k) = \{A(k), A_1^{'}(k), A_2^{'}(k), ..., A_i^{'}(k), ..., A_n^{'}(k)\} \tag{3}$$

where $A_i^{'}(k) = \{a_{i1}, a_{i2}, ..., a_{ij}, ..., a_{iq}\}, a_{ij} = a_i, j = 1, 2, ..., q$.

Step 4 (Mutation Operation): Since the affinity function is defined to be maximized, we sort the generic antibodies according to their affinity in an ascend order. Then the corresponding mutation probability of each antibody can be calculated:

$$p_m^i = \frac{f(b_i(k))}{f_{\max}(.) - f_{\min}(.)} \tag{4}$$

where $f(b_i(k))$ is the affinity of the antibody $b_i(k)$, $f_{\max}(.)$ is the maximum of affinity and $f_{\min}(.)$ is the minimum of affinity. The mutations operation can be embodied as follows:

$$b_i^{'}(k) = b_i(k) + p_m^i \times \exp(-f(b_i(k))) \times N(0,1) \tag{5}$$

where $N(0,1)$ is gauss variable, which mean value is 0 and mean square error equals $1, b_i^{'}(k)$ is the son antibody.

Step 5 (Selection Operation): $\forall i = 1,2,...,n$, if $b = \{a_{ij} \mid max f(a_{ij}), j = 2,3,...,N_C\text{-}1, a_{ij} \in A'\}$, and, $f(a_i) < f(b)(a_i \in A)$, then b replaces the antibody a_i in the aboriginal population. So the antibody population is updated, and the information exchanging among the antibody population is realized.

Step 6 (Determine the halt conditions): $k = k + 1$;Stop if the halt conditions are achieved. Otherwise, return to Step 2.

In HCSGA, each feature subset is represented by a binary vector of dimension m. If a bit is a 1, it means that the corresponding feature is selected. A value of 0 indicates that the corresponding feature is not selected. Classification accuracy and the feature cost are the two criteria used to design an affinity function. Thus, for the antibody with high classification accuracy and low total feature cost produce a high affinity value. We solve the multiple criteria problem by creating a single objective affinity function that combines the two goals into one. The antibody with high affinity value has high probability to be preserved to the next generation. We designed an affinity function as follows:

$$f(x) = accuracy(x) - \frac{\beta \times cost(x)}{accuracy(x) + 1} + cost_{\max} \tag{6}$$

where $f(x)$ is the affinity function of the feature subset represented by x; $accuracy(x)$ is the test accuracy; $cost(x)$ is the sum of measurement costs of the feature subset represented by x; and $cost_{\max}$ is an upper bound on the costs of candidate solutions. In this case, $cost_{\max}$ is simply the sum of the costs associated with all of the features. $\beta = 1$ represents that feature x is selected; otherwise, $\beta = 0$ represents that feature x is not selected.

4 Experiments

A series of experiments was conducted to show the utility of proposed FS algorithm. We implement proposed HCSGA algorithm and other three FS algorithms in Matlab7.0. We adjusted the parameters of the HCSGA by experiments, and finally selected the following combination of the parameters: the maximum number of generations $MaxGen = 150$, the mutation probabilities $P_m = 0.15$, antibody population size $popsize = 100$, and the multiplying factor $\alpha = 50$.

In order to make our evaluation results comparable to the most of the published results in text categorization evaluations, we have chosen as datasets the Reuters-21566. In Reuters-21566 dataset, we adopt the top ten classes, 5218 documents in training set and 2126 documents in test set. The maximum class has 2092 documents, occupying 38.2% of training set. The minimum class has 68 documents, occupying 1.13% of training set.

For evaluating the effectiveness of the text categorization algorithms, we have used the standard recall and precision. Given a category c_k , recall is defined as the probability that, if a random document d_i ought to be classified under c_k , this decision is taken. Analogously, precision is defined as the probability that, if a random document d_i is classified under c_k , this decision is correct.

Fig. 2-3 shows the performance of our proposed method against the ACO in [3], GA in [5] and Olex-GA in [6] for the ten most frequent categories. The precision of feature selection algorithm is shown in Fig. 2. The recall of the feature selection algorithm is shown in Fig. 3.

Analyzing the precision and recall show in Fig. 2-3, we see that on average, the HCSGA algorithm obtain a higher accuracy value than other three feature selection algorithms. The average precision for ACO, GA, Olex-GA and HCSGA are 82.8%,79.9%, 84.3% and 91.6% respectively. With the exception of categories

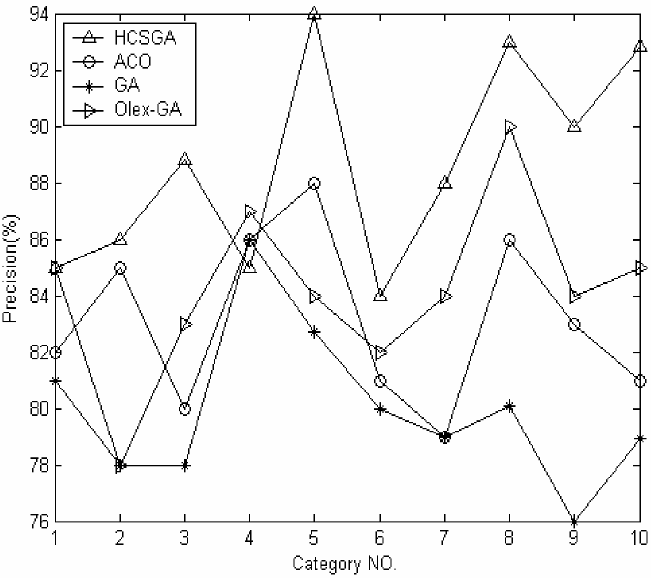


Fig. 2. The precision of four feature selection algorithms

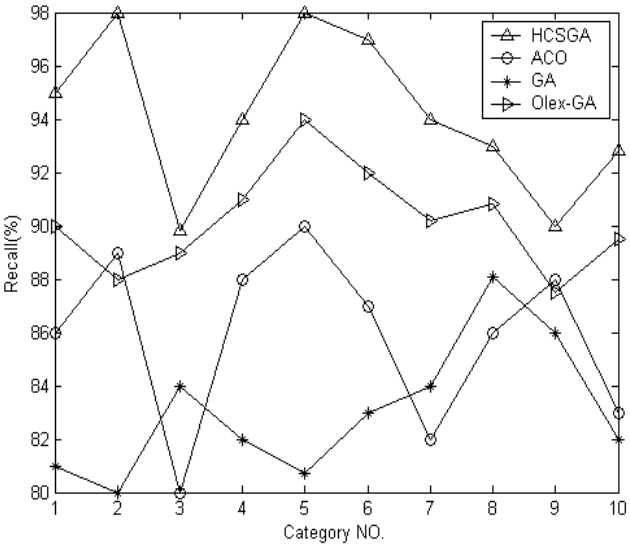


Fig. 3. The recall of four feature selection algorithms

Method, the precision of each category for HCSGA is higher than other algorithms. This indicates that the HCSGA algorithm perform at generally high precision. The average recall results for the ACO, GA, Olex-GA and QCGA are 85.6%, 82.3%, 90.6% and 95.1% respectively. The HCSGA can classify documents into the correct category mapping to precision, with a high recall ratio. This indicates that the HCSGA yields a better classification result than the other three methods. Because HCSGA can guide search to the optimal minimal subset every time.

5 Conclusions

In this paper, we have proposed a new algorithm for feature selection in text categorization, named hybrid clonal selection genetic algorithm. The experimental results show that the HCSGA yields the best result of these three methods, and also yields better accuracy even with a large data set since it achieved better performance with the lower number of features. In future research, we intend to investigate the performance of proposed feature selection algorithm by taking advantage of using more complex classifiers in that, and plan to combine proposed feature selection algorithm with other population-based feature selection algorithms.

References

1. Raymer, M., Punch, W., Goodman, E.: Dimensionality reduction using genetic algorithm. IEEE Transactions on Evolutionary Computing 12(4), 164–171 (2000)
2. Kim, H., Howland, P.: Dimension reduction in text classification with support vector machines. Journal of Machine Learning Research 12(6), 37–53 (2005)

3. Ani, A.: Ant colony optimization for feature subset selection. Transaction on Engineering, Computing and Tecgbikigt 12(4), 35–38 (2005)
4. Lam, W.: Automatic textual document categorization based on generalized instance sets and a met model. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(2), 628–633 (2003)
5. Jorng, T.H., Ching, C.Y.: Applying genetic algorithms to query optimization in document retrieval. Information Processing and Management (36), 737–759 (2000)
6. Punch, W.F., Goodman, E.D.: Further research on feature selection and classification using genetic algorithm. In: Proceedings of the Fifth International Conference on Genetic Algorithm, pp. 557–564. Morgan Kaufmann, San Mateo (1993)
7. Pietramala, A., Policicchio, V.L., Rullo, P., Sidhu, I.: A Genetic Algorithm for Text Classification Rule Induction. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 188–203. Springer, Heidelberg (2008)
8. Kudo, M., Sklansky, K.: Comparison of algorithms that select feature for pattern classifier. Pattern Recognition 33(2), 25–41 (2000)
9. Jiao, L.-c., Du, H.-f.: The prospect of the artificial immune system. Acta Electronic Since 31(10), 1540–1548 (2003)
10. Jiao, L.C., Wang, L.: A Novel Genetic Algorithm Based on Immunity. IEEE Transaction on Systems, Man, and Cybernetics-Part A: Systems and Humans 30(5), 552–561 (2000)

Power Demand Forecasting Based on BP Neural Network Optimized by Clone Selection Particle Swarm

Xiang Li[1] and Shi-jun Lu[2]

[1] Xiang LI, School of Business Administration, North China Electric Power University,
Beijing 102206, China
lxcbj@126.com

[2] Shi-jun LU, School of Business Administration, North China Electric Power University,
Beijing 102206, China

Abstract. Based on ordinary BP algorithm, firstly established the power demand forecasting model. Then the model's network structure was identified by using the power demand's influential factors as the input of the network, Repeated the optimization of the BP network's weight combination with the aid of clone selection particle swarm algorithm, and adopted the weight optimized as the initial value of the BP neural network, carried on the BP algorithm until the network met the training requirement. Finally recent years' annual data of relevant input and output variables were used to empirically forecast the power demand with the established model, mean absolute error is 156.8340, root mean square error is 160.9708, root mean square error rate is 0.0095. The results show that BP neural network based on clone selection particle swarm has both fast training speed and small error, the forecast precision also has been significantly improved.

Keywords: BP Neural Network, Clone Selection Algorithm, Particle Swarm Optimization, Power Demand.

1 Introduction

The macroeconomic system is a multi - variable, non-linear complex system with time variability and uncertainty. The economic cycle improves the linear models, such as the establishment of the linear model with time-varying parameters and the piecewise linear model, but the result is not very satisfied [1].

For above reasons, a lot of improvements have been made, such as the heuristic neural network, improved feed-forward network object model and prospective function with penalty term method [2-3]. But above methods only focus on the improvements of the network performance and do not fundamentally accelerate the speed of network learning and do not improve the forecast accuracy. Literature [4] used particle swarm algorithm's fast learning speed characteristic to introduce the particle swarm algorithm into the neural network and trained the neural network's parameters, and achieved certain results. However, fundamental particle swarm algorithm has fast convergence speed which still has shortcomings, such as the trend to local optimum, low precision accuracy [5]. To avoid above shortcomings, the clone

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 135–141.
springerlink.com

selection algorithm in the immune system to improve the particle swarm algorithm has been chosen which introduces the clone selection particle swarm algorithm into the BP neural network and establishes the neural network model based on the clone selection particle swarm algorithm (CSPSO-BP), and does the empirical analysis with the established model by using the recent years' related annual data.

2 Power Demand Modeling

2.1 Variable Data Selection and Pretreatment

Variables chosen in this paper include: GDP, the national population, industrial structure changes etc [6]. The annual data from1978 to 2005 [7-8] are chosen as training sample set of the neural network. Firstly make logarithm transformation with the selected variables, and then a differential change. Within, data from 1981-2000 used as network training, while data from 2001 – 2005 used as prediction test.

2.2 Determine BP Neural Network Structure

Optimize the BP neural network based on the clone selection particle swarm algorithm, using BP network structure in the process of its construction.

Literature [9] shows that, as long as the hidden nodes are enough, for the three-layer, the BP network (only one hidden layer) can realize any complex nonlinear mapping, which means that it can approach the function with random accuracy, achieving very precise curve fitting. And for multi hidden layer, multi-output BP network plays a same role. As long as the number of hidden layers and hidden nodes are appropriate, it can easily realize any complex nonlinear mappings.

(1) Determine the number of input and output layer neurons
According to the above variables selection, it can determine the input of the neural networks, current real GDP, prophase real GDP, population, the current industrial structural change and the prophase industrial structural change. Therefore, the final determined number of the input layer neurons is five. Using the power demand as the only output variable, the output layer has one neuron.

(2) Selection of the transfer function
The transfer function known as the activation function is an important part of the BP network. The transfer function must be continuously differentiable. After data pretreatment, the network's input vector is in the interval [-1, 1], while the network output vector is in the interval [0, 1]. Thus the transfer function of the network's middle layer neuron uses the S-tangent function tansig, and the output layer neuron's transfer function uses the S - logarithmic function logsig.

(3) The hidden layer and its neurons number determination.
Based on Kolmogorov theory [10], the hidden layer nodes number are determined. Kolmogorov Theory denote that, for any continuous function $f: U^n \to R^m$, $f(x) = Y$, where U is the closed unit interval [0, 1], f can be accurately realized by a three-layer forward network. This network's first layer (input layer) has n processing units,

the middle layer (hidden layer) has $2n+1$ processing units, and the third layer (output layer) has m processing units. After identifying the hidden layer nodes number's theoretical value k , simulate contrast with the data in the interval $(k-\delta, k+\delta), \delta \in N$ by using the trial-and-error method and get the best hidden layer nodes number of the neural network model.

The previously identified input layer nodes number is 5, so the theoretical value of the hidden layer nodes number is 11. Using the Matlab software, set the transfer function of the hidden layer neuron is tansig, while the transfer function of the output layer neuron is logsig, then using the trial-and-error method to train the network with the hidden layer nodes from 9-15, the training results are shown in figure 1. Different hidden layer nodes' network training errors are shown in table 1.

Table 1. Network errors and training times

Hidden nodes	MSE	EPOCHS
9	0.0253	169
10	0.0274	174
11	0.0262	146
12	0.0250	166
13	0.0227	86
14	0.0270	150
15	0.0260	176

In the table 1, MSE is the mean square error, EPOCHS is the training time which need to achieve the satisfied network error. It can be seen from the table 1, when the hidden nodes are 13, the network's excepted error can be obtained after 86 training times, and the mean square error is 0.0227. So modeling of hidden nodes the neural network structure with the hidden nodes number is 13.

3 Learning Process of Neural Network Model Based on Clone Selection Particle Swarm Algorithm (CSPSO-BP)

The learning processes are as follows [11-13]:

(1) Determine the parameters.
Assume the Learning factors (acceleration constants) $c_1 = c_2 = 2$, and 20 particles(antibodies)in the swarm group;

(2) Initialization.
Use particles as the weights and threshold of the BP neural network and generate random initialization particles (antibodies)'s initial position x_i and speed v_i , $i = 1, 2, \cdots, N$, and value in the range of [-1,1]. Meanwhile, in order to prevent the particles far from the search space, the speed v_i of the particles is usually limited

between $\left[-v_{i\max}, +v_{i\max}\right]$. If $v_{i\max}$ is too large, the particle will be swift away from the optimal solution, and if it is too small, it is easy to fall into a local optimum. Therefore, it is generally assumed $v_{i\max} = kx_{i\max}$, $0.1 \le k \le 1.0$ and the max iterative step $it_{\max}$ is 2000;

(3) Generate temporary clone particles (antibodies).
Calculate each current particle's individual extreme value and global extreme value, according to the formula:

$$F_i = 1/\exp(E_i) \tag{1}$$

Calculate affinity value F_i, where, $E_i = \frac{1}{2}\sum_{k=1}^{n}(\hat{y}_k - O_k)^2$, $\hat{y}_k$ and y_k are the expect and actual output of the node k.

Determine the best particle (antibody) based on the affinity value, and clone it into memory to judge whether it meet the conditions. If it satisfies the conditions, cease operation and output the result, otherwise continue to the next step;

(4) From the below formula:

$$v_i^{(t)} = v_i^{(t-1)} + \rho_1(x_{pbest_i} - x_i^{(t)}) + \rho_2(x_{gbest_i} - x_i^{(t)}) \tag{2}$$

$$\begin{cases} x_i(t) = x_i(t-1) + v_i(t) \\ t = t+1 \end{cases} \tag{3}$$

Generate new $N = 20$ particles (antibodies) and randomly generate new $M = \frac{N}{2}$ particles (antibodies). In the particles (antibodies) groups' updating process, it always hopes that these particles (antibodies) with high affinity value are saved. But if they are too concentrating which means excessive concentration, it is very difficult to ensure the diversity of the particles (antibodies), and also very easy to fall into a local optimum which looses these particles (antibodies) which have bad affinity value while maintaining a good evolution trend.

Thus it can adopt a diversity maintain strategy based on a concentration mechanism which make all particles (antibodies) at any affinity value maintain a certain concentration in the new generation groups. The concentration of the i th particle is defined as follows:

$$D(x_i) = 1\Big/\{\sum_{i=1}^{N+M}\left|f(x_i) - f(x_j)\right|\} \qquad i = 1, 2, \cdots, N+M \tag{4}$$

Probability selection formula based on the particle (antibody) concentration can be obtained from the below formula:

$$P(x_i) = \frac{1/D(x_i)}{\sum_{i=1}^{N+M}[1/D(x_i)]} = \frac{\sum_{i=1}^{N+M}\left|f(x_i)-f(x_j)\right|}{\sum_{i=1}^{N+M}\sum_{j=1}^{N+M}\left|f(x_i)-f(x_j)\right|} \quad i=1,2,\cdots,N+M \tag{5}$$

Where, x_i and $f(x_i)$, $i=1,2,\cdots,N+M$ are respectively denoted the i th particle and its affinity function value. By the formula, it can be seen that the more antibody similar to the i th particle, the smaller probability that i th particle is selected.

Whereas, the less antibody similar to the i th particle, the bigger probability that i th particle is selected. This makes the individuals with low affinity value can also get the opportunity to evolution, therefore the probability selection formula ensure the antibody diversity in theory.

Calculate the selection probability of the new generated $N+M=30$ particles (antibodies) with formula (5), sorting the $N+M$ particles (antibodies) according to the size of their probability, N particles with larger value are selected, generate a mature antibody group.

(5) Particle (antibody) Update: replace the low affinity value particles (antibodies) generated in step 4 with the memory ones, form a new generation of particle (antibody) group, then jump to Step 2.

Adopt the network weights and thresholds optimized by the clone selection particle swarm optimization algorithm as the initial weights the threshold for BP algorithm's network; go on with the BP algorithm until the network meets the performance indexes, save the network weights and threshold, calculate the output value. Then according to the training weights and threshold values respectively calculate the input and output value of the BP neural network hidden layer.

4 Power Demand Forecast

Based on the annual data from 1981-2005, use data from 1981-2000 for the network training, data from 2001-2005 for forecast test. Matlab7.0 was used in programmed simulation analysis. Adopt the current real GDP, prophase real GDP, population, the current industrial structural change, the prophase industrial structural change as the input of the neural networks, and the power demand as the network output, then began the empirical analysis.

In matlab7.0, assumed the cycle times was 2000, the training error was 0.001, adopted the newff function to form a BP neural network which had 13 hidden layer nodes and 1 output neuron. Initialize the network weights by the clone selection particle swarm algorithm, and then chose sim function to do the simulation forecast. The network training course is showed in figure 1, and the power demand curve simulation is showed in figure 2.

Meanwhile, used ordinary BP neural network in fitting the same sample data and the comparison of the results is shown in Table 2.

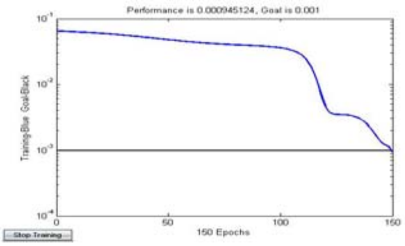


Fig. 1. Training of network

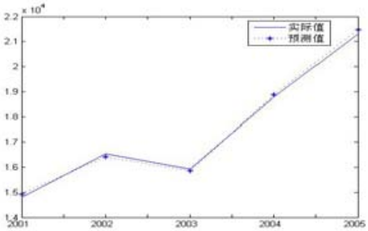


Fig. 2. Simulation fitting curve of power demand

The results can be seen from Table 2. that BP neural network model optimized by the clone selection Particle swarm is superior to the ordinary BP neural network model in many aspects, such as training speed and forecast precision etc. and the new method's validity and applicability is well illustrated.

Table 2. Comparison of two algorithms

Model	Time(s)	Mean absolute error	Root mean square error	Root mean square error rate
CSPSO-BP	15.6	156.8340	160.9708	0.0095
BP	52.6	691.6660	701.2059	0.0450

5 Conclusions

The ordinary BP neural network has some inherent drawbacks, such as slow learning speed, trend to local minimum and "over-study". Therefore, the clone selection particle swarm algorithm is introduced into the learning process of the BP neural network training. Then, establish the neural network power demand forecast model optimized by the clone selection particle swarm algorithm, and pass through the empirical analysis, and many advantages of this new model has been known, such as faster speed, better stability, higher forecast precision. So it's a better model for forecast power demand.

References

[1] Li, C.-b., Wang, K.-c.: A new grey forecasting model based on BP neural network and Markov chain. Journal of Central South University of Technology 14(5), 713 (2007)

[2] Mccall, J.: Genetic algorithms for modeling and optimization. Journal of Computational and Applied Mathematics 184, 205–222 (2005)

[3] Tong, C.-r., Li, M.-z., Wu, J.-c., Liu, D.-b.: Soft sensor model of sodium aluminate solution based on BP neural network with inverse mapping algorithm. The Chinese Journal of Nonferrous Metals 18(5), 917 (2008)

[4] Liu, L., Yan, D.J., Gong, D.C., et al.: New method for short term load forecasting based on particle swarm optimization and fuzzy neural network. Proceedings of the Chinese society of universities 18(3), 47–50 (2006) (in Chinese)

[5] Abou El-Ela, A.A., Fetouh, T., Bishr, M.A., Saleh, R.A.F.: Power systems operation using particle swarm optimization technique. Electric Power Systems Research 78(11), 1906–1913 (2008)

[6] Ling, B.: Structural changes, efficiency improvement and electricity demand forecasting. Economic Research Journal (5), 57–65 (2003) (in Chinese)

[7] National Bureau of statistics of China. China statistical yearbook. China Statistics Press, Beijing (2006) (in Chinese)

[8] "China electric power yearbook" Editiorial Board. China electric power yearbook: 2006. China Electric Power Press, Beijing (2006) (in Chinese)

[9] Jiang, W., Xu, Y.: A Novel Method for Nonlinear Time Series Forecasting of Time-Delay Neural Network. Wuhan University Journal of Natural Science 11(5), 1357 (2006)

[10] Wang, B.-x., Zhang, D.-h., Wang, J., Yu, M., Zhou, N., Cao, G.-m.: Application of neural network to prediction of plate finish cooling temperature. Journal of Central South University of Technology 15, 136–140 (2008)

[11] Nasseri, M., Asghari, K., Abedini, M.J.: Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network. Expert Systems with Applications 35(3), 1415–1421 (2008)

[12] Simon, T.P., Jun, H.: A hybrid artificial immune system and Self Organising Map for network intrusion detection. Information Sciences 178, 3024–3042 (2008)

[13] Tan, K.C., Goh, C.K., Mamun, A.A., Ei, E.Z.: An evolutionary artificial immune system for multi-objective optimization. European Journal of Operational Research 187, 371–392 (2008)

Research on Simplifying the Motion Equations and Coefficients Identification for Submarine Training Simulator Based on Sensitivity Index

Zhao Lin and Zhu Yi

College of Automation, Harbin engineering university, Harbin, Heilongjiang, China
zhuyi198215@gmail.com

Abstract. Submarine training simulator utilize the high technology devices to train the sailor in the safe, high-efficiency and realistic environment while reducing the dangerous quality and fund in actual training. Solving the six dimension motion equations of submarine plays a significant role in the training simulator. As the 108 hydrodynamic coefficients are hard to acquire completely, and the low solving speed because of the complex equations, it is necessary to simplify the equations in the range of accepted deviation. In this paper, the submarine motion model is analyzed, the motion equation is simplified, a sensitivity index is introduced, and three typical maneuver simulation tests are chosen as the research subjects. Furthermore, simplifying and identifying the model using the particle swarm optimization based on sensitivity index of the coefficients. We define the manipulate evaluation, determine the mathematical expression of the identification, restrict the type, range, field of definitions and evaluation function of the manipulate experiment. Improve the particle swarm optimization algorithm based on the sensitivity index. Through a mount of simulation calculation, this paper modifies the hydrodynamic coefficients of equations of submarine training simulator, and the influence of hydrodynamic coefficients for the maneuverability is evaluated. The simulation results prove the feasibility of the simplified and coefficients identification method.

Keywords: submarine training simulator, sensitivity index, particle swarm optimization, coefficients identification.

1 Introduction

For a long time, forecasting the hydrodynamic coefficients of the submarine motion accurately is the basis of the manipulate research and the hardest task. The main objective of the training simulator is training the sailor to be familiar with the submarine maneuverability. It is different to the real training in the actual submarine. The mathematical model [2] of the maneuver and control for submarine contains hydrodynamic forces and moments expressed in terms of a set of hydrodynamic coefficients. Hydrodynamic coefficients strongly affect the dynamic performance of submarine. The main objective of the training simulator is training the sailor to be familiar with the submarine maneuverability. It is different to the real training in the

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 142–149.
springerlink.com

actual submarine. The simulator gets the motion attitude according to solving the motion equation. The general equation [2] are very complex. It contains 108 hydrodynamic coefficients, so acquiring the whole coefficients is rather hard. Many researchers simplify the standard motion equation according to manipulate experiences and the hydrodynamic coefficients analysis. The experiments and simulation results[1][3][4] show that the main effects to the submarine space motion depend on some coefficients not the whole coefficients; therefore we could consider the particular primary coefficient influence on submarine motion and ignore the less effective coefficients. Meanwhile the accuracy of the simulation is lower than the complex motion equations after the simplification, and the training consequence is less effectively. This paper simplify the motion equation using the manipulate experiment simulation results of coefficients sensitive index. Recently, particle swarm optimization was proposed as a simple tool for solving function optimization problems. It is a revolutionary computation technique based on the social behavior of bird flocks and fish schools. It is has been shown to be effective in optimizing complex multidimensional problems [9][10][11]. This paper simplify the motion equation using the manipulate experiment simulation results of coefficients sensitive index and identify the coefficients that is simplified based on particle swarm optimization.

2 Submarine Space Motion Equation

The coordinate and symbol rules are comply with the well known general six-degrees-of freedom equations of Gertler [2]. The six degree of freedom equations of submarine is rather complex and non-linear which existing coupling between horizontal plane and vertical plane. The number of the hydrodynamic coefficients of the equations is 108.

$$\begin{aligned}
&m\left[\dot u - vr + wq - x_G\left(q^2 + r^2\right) + y_G\left(pq - \dot r\right) + z_G\left(pr + \dot q\right)\right] = X\\
&m\left[\dot v - wp + ur - y_G\left(r^2 + p^2\right) + z_G\left(qr - \dot p\right) + x_G\left(pq + \dot r\right)\right] = Y\\
&m\left[\dot w - uq + vp - z_G\left(p^2 + q^2\right) + x_G\left(rp - \dot q\right) + y_G\left(rq + \dot p\right)\right] = Z\\
&I_x\dot p + \left(I_z - I_y\right)qr + m\left[y_G\left(\dot w - uq + vp\right) - z_G\left(\dot v - wp + ur\right)\right] = K\\
&I_y\dot q + \left(I_x - I_z\right)rp + m\left[z_G\left(\dot u - vr + wq\right) - x_G\left(\dot w - uq + vp\right)\right] = M\\
&I_z\dot r + \left(I_y - I_x\right)pq + m\left[x_G\left(\dot v - wp + ur\right) - y_G\left(\dot u - vr + wq\right)\right] = N
\end{aligned} \tag{1}$$

In the above, u, v, w, p, q, r represent the translational and rotational velocities in the body system. x_G, y_G, z_G is the coordinate of the center of gravity G . m is mass and I_{xx} , I_{xy} , etc., denote the various principal and cross-moment moments of inertia of body. The left-hand side of above equations is exact. X, Y, Z, K, M, N , represent the resultant forces and moments. The right-hand side of above equations that contain external forces and moments are uncertain and approximate in the dynamic model. They are expressed in term of a set of hydrodynamic coefficients such as X'_{qq} $K'_{\dot p}$

which are hard to estimate or determine exactly. W, B , represent the weight and buoyancy of body. ϕ, θ, φ , represent the roll, pitch and yaw angle respectively.

As far as the training simulator is concerned, calculating the equations would cost a mount of time and system resources. However the main purpose in training sailor is that achieving the corresponding rudder effect according to manipulate the rudder, bow and stern in term of separately or combined steering with virtual instrument. In consequence, this paper simplifies the complex equations to improve the calculating efficiency of simulator and optimize the system resources.

3 Typical Steering Test

In this paper, typical steering test we decide to choose are overshoot maneuver in the vertical plane, overshoot maneuver in horizontal plane and turning circle maneuver in the horizontal plane. Through the test, we can get a set of maneuver parameters which denote essential characteristics.

Overshoot maneuver in horizontal: $10^\circ / 10^\circ$ rudder, the maneuver parameters are t_a , t_a , ψ_{ov} , T .

Turing circle maneuver in horizontal: 35° ruder, the maneuver parameters are D_s , D_T , A_d , T_r , T .

Overshoot maneuver in vertical: 20° stern, pitch angle 6° , the maneuver parameters are t_e , θ_{ov} , ζ_{ov} .

Maneuver in the vertical plane, two type of overshoot maneuver in horizontal plane which manipulate the bow or stern to steering the submarine and turning circle maneuver in the horizontal plane are the typical steering we choose.

Through the test, we can get a set of maneuver parameters which denote essential characteristics. The definition of maneuver parameter is the same as [2]. Each test has the condition that the submarine has the 15 knot at an appointed depth. The identification evaluations we choose in the steering test are 2 types in each test.

Overshoot maneuver in horizontal: $10^\circ / 10^\circ$ rudder, the maneuver parameters are ψ_{ov} , T . Turing circle maneuver in horizontal: 30° ruder, the maneuver parameters are D_s , T . Overshoot maneuver in vertical (bow): 20° stern, pitch angle 6° , the maneuver parameters are t_{ew} , θ_{ovw} . Overshoot maneuver in vertical (stern): 20° stern, pitch angle 6° , the maneuver parameters are t_{es} , θ_{ovs} .

4 The Sensitivity Index

The sensitivity index - S are introduced for measuring the sensitivity extent of manipulate parameters to the hydrodynamic coefficients. The definition is as follows:

$$S = \left(R - R^*\right) / R^* = \Delta R / R^* \tag{2}$$

R represents the manipulate parameters, R^* represent the corresponding manipulate parameters that is calculated on the standard hydrodynamic coefficients.

S represents the hydrodynamic coefficients sensitivity extent to the manipulate parameters. The more sensible in submarine motion, the higher value of S. That means hydrodynamic coefficients play a significant role in the manipulating simulation.

4.1 The Sensitivity Calculation of Hydrodynamic Coefficients

The procedures of experiment in this article are as follows:

(1) By choosing one of the typical manipulate test, defining the original situation, we simulated the motion in particular rudder degree within the initial hydrodynamic coefficients and calculate a serial of manipulate parameters as the standard values.

(2) Define the ith hydrodynamic coefficients as 0, do the same experiment in (1), get the changed manipulate parameters and calculate the corresponding sensitivity index. And then change the ith coefficient into the original coefficient. Repeat the simulation experiment until all the hydrodynamic coefficients were changed, the manipulate coefficients are acquired and the corresponding sensitivity index are calculated completely.

(3) Choose another typical manipulate test to do the same step in (1) and analysis in (2)

(4) Use the three types of coefficients to make a table that contains the original hydrodynamic coefficients, manipulate parameters and sensitivity index.

(5) Calculate the relative error of maneuver parameters.

4.2 Parts of Hydrodynamic Coefficients Results

The simulation of manipulate test base on the matlab7.0. The three table 1,2,3 show the some of the sensitivity index whose value is greater than 3% in the three typical test respectively. Inf represent infinite.

Table 1. Sensitive index results of overshoot maneuver in horizontal plane

coef	t_a	t_{ov}	ψ_{ov}	T
$Y'_{\dot{v}}$	0	0	-0.0914	-0.0488
$Y'_{\dot{p}}$	inf	inf	inf	inf
Y'_r	0.2666	-0.1111	-0.3658	0.122

Table 2. Sensitive index results of turning circle maneuver in horizontal

coef	D_s	D_T	A_d	T_r	T
$Y'_{\dot{p}}$	inf	inf	inf	inf	inf
Y'_v	0.5144	0.4366	0.4002	0.1346	0.2396
Y'_r	0.3484	0.3054	0.1141	0.2937	0.2099

Table 3. Sensitive index results of overshoot maneuver in vertical plane

coef	t_e	θ_{ov}	ζ_{ov}
$Z'_{\dot{w}}$	-0.2071	0.1245	-0.1991
$Z'_{q\|q\|}$	0.0357	0.0063	0.0561

Table 4. Simplified model manipulate evaluated index relative error

t_a	t_{ov}	ψ_{ov}	T	
0	0	0.05482	0	
D_s	D_T	A_d	T_r	T
0.011706	0.015834	0.006396	-0.00345	0.011438
t_e	θ_{ov}	ζ_{ov}		
0	-0.07685	-0.00308		

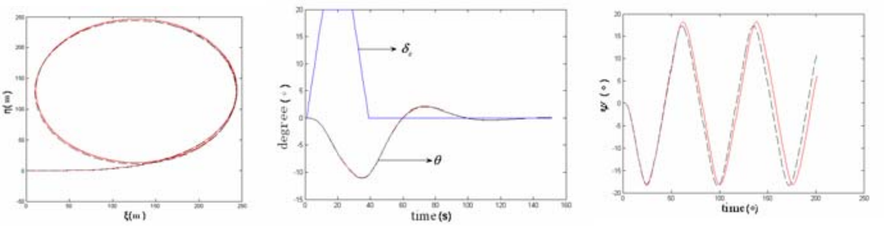


Fig. 1. Turning circle、overshoot maneuver in horizontal and vertical plane

4.3 Analyses of Results

In this paper, we reserve the sensitivity index whose value is greater than 3% in three tests. The 3% indicate that the maximum influence would not beyond 3%. We reduce the number of the hydrodynamic coefficients to 23 while the former number is 76. Figure 1 shows the simulation results (dotted line refer to simplified simulation curve, solid line refer to original simulation curve). Table 4 shows the relative error of the maneuver experiments.

5 Particle Swarm Optimization (pso)

Particle swarm optimization is similar to a genetic algorithm in that system is initialized with a population of random solutions [12]. The potential solution is assigned a randomized velocity, and the potential solutions, called particles, are flown through hyperspace. The original form of standard PSO is very simple, where the dimensions on random positions and velocities is N; the ith particle position is xi=(xi1, xi2……xiN), the ith particle velocity is vi=(vi1, vi2……..viN); the desired minimization function in N variables we should evaluated is also the stop meet of criterion; the ith best position value is pi=(pi1, pi2…….piN); the overall best position value is pg=(pg1, pg2…..pgN); Limit velocity: Vid=vmax, if vid>vmax; Vid=-vmax, if vid<-vmax; update in velocity and positions by following formula:

$$V_{id} = w * vid + c1rand(\ )(pid - xid) + c2rand(\ )(pgi - xid)$$
$$Xid = xid + vid \tag{3}$$

Each particle has achieved the best solution called pi that is keep track of its coordinates in hyperspace. The particle swarm also has the overall best value called pg that is obtained through any particle in the population. The initial weight is set as w=0.7298 and the learning factor is set as c1=c2=1.4962; $rand(\ )$ represent the random number between 0 and 1; the particle number is 10; the criterion function J is evaluated through the steering test based on sensitivity index of hydrodynamic coefficients.

$$S_{ij} = \frac{(R_j - R^*{}_j)/R_j^*}{(H_i - H_i^*)/H_i^*} = \frac{\Delta R_j / R_j^*}{\Delta H_i / H_i^*} \tag{4}$$

The sensitivity index - S_{ij} are introduced for measuring the sensitivity extent of manipulate parameters to the hydrodynamic coefficients. The definition [1] is as follows:

Where R_j represent the manipulate parameters, R_j^* represent the standard corresponding manipulate parameters that is calculated on the standard hydrodynamic coefficients. H_i represent the hydrodynamic coefficients, H_i^* represent the standard hydrodynamic coefficients.

S represent the hydrodynamic coefficients sensitivity extent to the manipulate parameters. The more sensible in submarine motion, the higher value of S . That means hydrodynamic coefficients play a significant role in the manipulating simulation. S we defines as the maximum of the S_{ij} :

The criterion function we define that contains 8 maneuver parameters Yi in four types steering test is as following; In this paper we try to find best solution of 28 hydrodynamic coefficients X^* which Y1~Y8 means the best manipulate parameters according to the criterion function J .

The range of varying coefficients H_i is $\Delta H_i / H_i^*$ =0.5～1.5, for each steering test, we calculate S for a 0.1change in the ith coefficients, we calculate it for 10 times. And then choose the maximum of them. For each coefficient in (2) we get the S_i in (4) and calculate sensitivity index of 28 coefficients. Get the maximum in (5). Calculate the J from every changing coefficient and steering test in (6), update in (3).

$$S \ = MAX(|S_{ij}|), j = 1 \sim 8 \tag{5}$$

$$J = ((1 - \frac{\varphi_{ov}}{\varphi_{ov}^*})^2 + (1 - \frac{T_Z}{T_Z^*})^2 + (1 - \frac{t_{ew}}{t_{ew}^*})^2 + (1 - \frac{\theta_{ovw}}{\theta_{ovw}^*})^2 + (1 - \frac{t_{es}}{t_{es}^*})^2 + (1 - \frac{\theta_{ovs}}{\theta_{ovs}^*})^2 + (1 - \frac{D_s}{D_s^*})^2 + (1 - \frac{T_H}{T_H^*})^2)^{1/2} \tag{6}$$

From the simulation of hydrodynamic coefficients sensitivity index, we inquire the manipulate parameters. The result we calculate is signed number, it reflect that changing (enlarge or reduce) a coefficient would effect the corresponding manipulate parameters. We can evaluate the trend and extent of deviation from the table. The last raw is the maximum. For example the parameters of Y_v' are 0.04952、0.025、0、0、0、0、5.46E-05 、0.00188 we can get the information that Y_v' has no effect in vertical plane and sensitive in Turing circle maneuver in horizontal., The value of w in this paper is changed from scalar to vector. The standard of the choose method is based on the sensitivity index. The larger the index, the less value in w. This paper gets three value of w according to the maximum result-S of the index in (5). This paper define w as 0.9 when S less than 0.1, define w as 0.7 when S between 0.1 and 1, define w as 0.5 when S is larger than 1. The stop criterion value of the PSO in this paper is 0.1414. We simulate the PSO test in (6) based on Matlab 7.0. Figure 2 show the different of iteration number and J between standard PSO and improvement PSO based on sensitivity index. We test the improvement PSO based on sensitivity index and standard PSO for ten times respectively, calculate J and the iteration number correspondingly.

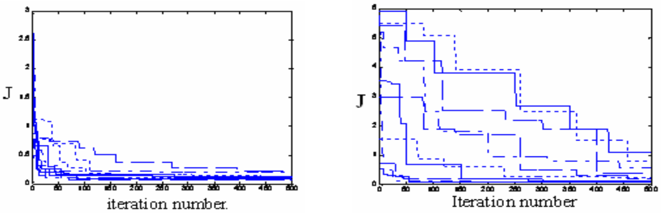


Fig. 2. J and iteration number of standard PSO and improvement PSO

6 Conclusion

This paper introduced a new method that simplified the submarine motion model based on the analyses of coefficients sensitivity. Three typical type of manipulate experiment and its experiments coefficients that evaluate the manipulation are discussed, determined the index and test procedure, simulate the algorithm in matlab, acquire the simulation curve and results, simplify the equations. In this paper, we also find a feasible way on the coefficients identification of simulator based on the improved PSO and sensitivity index.

References

[1] Sen, D.: A study on sensitivity of maneuverability performance on the hydrodynamic coefficients for submerged bodies. Journal of Ship Research 44(3), 186–196 (2000)

[2] Shi, s.: Controllability of Submarine. National Defence Industry Press, Beijing (1995) (in Chinese)

[3] Wang, y.f., Zhu, J., Zhang, z.s.: A method of evaluating the influence of hydrodynamic coefficients on controllability of submarine. Journal of Ship Mechanics 9, 561–568 (2005)

[4] Hwang, W.Y.: Application of system identification to ship maneuvering, MIT Ph.D. Thesis (1980)

[5] Le, M.D., Nguyen, D.-H.: Estimation of ship hydrodynamic coefficients in harbor maneuver and its applications. Proc. of IFAC Control applications of Optimization 1(1), 227–232 (2000)

[6] Fossen, T.: Guidance and control of Ocean Vehicles. John Wiley & Sons, UK (1994)

[7] Yoon, H.K., Rhee, K.P.: Estimation of roll related coefficients of a ship by using the system identification method. The Society of Naval Architects of Korea 41, 453–458 (2004)

[8] Kim, S.Y.: Estimation of maneuvering coefficients of a submerged body by parameter identification. Ph.D Thesis Seoul National University (2004)

[9] Eberhart, R.C., Kennedy, J.: Particle Swarm Optimization. Neural Networks 4, 1942–1948 (1995)

[10] Baskar, S., Alphones, A., Suganthan, P.N., Liang, J.J.: Design of Yagi-Uda antemmas using comprehensive learing particle swarm optimization. IEE (2005)

[11] Franken, N., Engelbrecht, A.P.: Particle Swarm Optimization Approaches to Coevolve Strategies for the Iterated Prisoner's Dilemma. IEE Trans. Evolutionary computation 9(6) (December 2005)

[12] Kim, S.Y.: Estimation of maneuvering coefficients of a submerged body by parameter identification. Ph.D Thesis Seoul National (2004)

Evaluating Quality of Networked Education via Learning Action Analysis

Bin Xu

College of Computer Science & Information Engineering, Zhejiang Gongshang University,
310018 Hangzhou, China
xubin@mail.zjgsu.edu.cn

Abstract. While networked education systems move at a very fast pace recent years, efficiency and quality of such education model receive much attention. On the basis of networked co-learning system establishment and experience in a large famous university, the authors present criteria to evaluate the quality of education and propose a learning action analysis method for the evaluation. Such method can be used to indicate the status of collaboration in co-learning environment and find out the potential problem during the learning progress.

Keywords: networked education, quality of education, learning action analysis, online co-learning system.

1 Introduction

Networked education originally provided the instructional courses to the country areas, but it moves at a very fast pace these years. There were only six universities provided networked courses in 1999. In 2005, there were a total of 31 network universities approved by the Ministry of Education and most of them were key institutions of higher learning and famous universities.

Networked education has been popularly used in many environments because computer-based programs become more popular and distance education has some significant advantages, including better visualization, more personalization, easier accessing to extra source, widely interaction, less language barrier, more creativity and less cost [1, 2, and 3]. Student writing abilities should be enhanced since writing is a critical element in communicating online; further, lifelong learning abilities should be enhanced since online education should be available throughout graduates' lifetimes [18].

Enhancing collaborative learning in smaller groups of students is a significant way of building community online [8]. Several researchers have reported on case studies at several universities that collaborative learning increased a sense of community among online learners [7, 10, 11, and 12].

Though cost is important for the education, quality of the education cannot be compromised in achieving the goal of education [13, 16]. The quality of asynchronous interaction in web-based conferencing among pre-service teachers had been studied [15]. Three different discussions in the learning were clarified from the

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 150–157.
springerlink.com

study, namely higher-level discussion, progressive discussion and lower-level discussion and the results showed that higher-level perspective taking was related to higher-level discussion. In order to evaluate and ensure the quality of a distance learning system, 6 scales questions has been brought forth, including relevance, reflection, interactivity, tutor support, peer support, and interpretation [4].

Moodle has been used in establishing the corporate education system [17], in university education [19 and 20]. But the quality of networked education is not evaluated in their practice. The most notable research mentioned about quality of networked education is presented by Vouk, M.A. et al. in [14]. They compared networked education with in-classroom education and stated the possible challenges for the networked education. They cared for the infrastructure and environment factors, but didn't mention the evaluation method about the networked education.

The faculty of our college established an online co-learning system to support networked education for computer technology. Moodle was used to build the course management system. On the basis of experience in the course management, the quality of education in co-learning environment is being evaluated and the college is to improve the education quality during the co-learning progress.

The main contribution of this paper is to present a learning action analysis method to evaluate the collaboration status in co-learning environment. The rest of this paper is organized as follows, the criteria to evaluate the quality of education will be brought forth in Section 2, and learning action analysis method will be introduced in Section 3. Section 4 presents the case study. Section 5 states the conclusion and the future work.

2 Quality Evaluation Criteria

According to the six scales stated by P. Taylor and D. Maor in [4], several criteria are used to evaluate the quality of networked education with the co-learning system.

2.1 Relevance

In order to ensure the online co-learning is high relevant to the students' professional practices, the content of the networked education system is required to benefit the information apperception, knowledge acquirement, skill enhancement, and talent development for the students' professional practices.

Table 1. Relevance evaluation form

Item	Low relevant	Relevant	High relevant
Information apperception	2	3	4
Knowledge acquirement	3	5	7
Skill enhancement	5	7	9
Talent development	6	8	10

A relevance evaluation form is designed as shown in Table 1. The value of relevance is rated from 1 to 10. Because the skill and talent development is most important to the student's professional practices, the value is much higher when the content is relevant to these items. The relevance to the information apperception is valued with low rate as it is not so much benefit the student's professional practices. This evaluation form is used to evaluate the relevance of each web page. When there are several pages, the relevant is measured as the maximum value of these pages.

Formally, vector Vs can be used to record the score given by the experts regarding different characteristic. For example, if an expert presented 2 to information apperception as he/she thought that the page is of low relevant regarding information apperception. He/she presented 5 and 7 to knowledge acquirement and skill enhancement, and presented 10 to talent development since he/she believed that the page is high relevant to the talent development. Vs = {2,5,7,10}. Matrix Mr can be used to store all the relevance value presented by the experts.

$$\mathrm{M_r} = \begin{bmatrix} Vs_1 \\ Vs_2 \\ Vs_3 \\ \cdots \\ Vs_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2c} \\ a_{31} & a_{32} & \cdots & a_{3c} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nc} \end{bmatrix},$$

where c is the number of characteristic of the relevance evaluation and a_{ij} is the value presented by expert i to characteristic j. Vector Ve refers to professional level of the experts to the domains related to the course. Vector Vimportance refers to importance degree of the different characteristic. The relevance value can then be calculated from Mr and Ve:

$$\mathrm{Re}\,levance\ value = M_r * V_{impor\tan ce}{}^T * V_e .$$

The web page with higher relevance value indicates the possible better education achievement.

2.2 Interpretation

Interpretation is hard to be measured formally and we value it by survey. The following measurement can be done to value the interpretation.

1. Randomly select *m* (m can be 100 or larger) students who have used the co-learning system as the candidates group A.
2. Randomly select *m* students who have never used the co-learning system but have used some other learning systems before as the candidate group B.
3. Randomly select *m* students who have never used any learning system as the candidate group C.
4. The students are required to view and operate the web page of a certain course for several minutes and present their score to value if the web page is of good interpretation. Regarding each web page in the learning system, two or three key functions will also be evaluated.
5. The final average score can then be used to value the interpretation.

The value from group A indicates the Interpretation value for the normal users in the system. Practically, if most of the members from group A used much time to value the functionality of the web page, then the web page should contain too much information or its interpretation is not so good.

The value from group B indicates the Interpretation value for the potential users in the learning system. When the students cannot operate the system efficiently, user guide, online help or even directly training will be helpful for them. If most of the members from group B couldn't operate the system easily with the assistance from the senior members, then the web page would lead in large amount of learning cost and it will be suggested to be improved.

The value from group C is similar as that from group B, but the value from group C indicates the interpretation value of the web page for those outside the university (maybe from some other secondary schools) while the value from group B indicates the interpretation value of the web page for those students inside the university. For those networked courses to be provided outside the university, the value from group C will be useful for the improvement of education quality.

3 Quality Evaluation with Learning Action Analysis

The students were asked to finish the exercise and their homework in a determined schedule. The student should go over the course in class-room, read the related material from the notebook, present material, or event networked education system, and think how to finish their work.

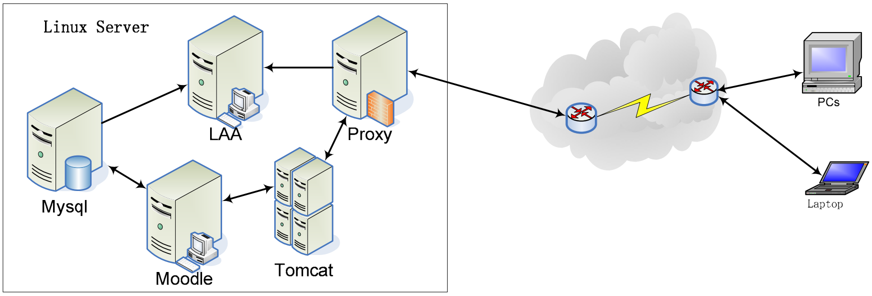


Fig. 1. Learning action analysis in the co-learning environment

When they meet problem, they may use the online chat-room in the co-learning system to get instant help. Some fellow student may help them when available. However, sometimes they cannot get the sufficient help from the chat-room; they may put the problem in the forum. The teacher will trace the forum and will provide some help enough for the student to finish their work.

Before analysis the detail learning actions, the role of the user should be identified so as to differentiate the teacher and student. As shown in Figure 1, the access actions to the Linux will be recorded and stored in the application of learning action analysis (LAA in Figure 1). For example, when a student logon to the system, there will be at least a message of password checking via a IP address and there will be some fetch messages when the web pages appear on the students' or teachers' screens. However, these messages contain no information about the role of the users. Therefore, the role identification should be made with the log captured from the MySQL database of the course management system. Action message from the proxy server can be defined as <time, IP address, Action Message, and Description> and the task message from Mysql database can be defined as <Time, IP address, Task, Role>. If conjoin these two messages, the user from same IP address at the sequent time will be identified to a certain role by the learning action analysis application deployed at LAA server [Figure 1].

3.1 Reflection Evaluation

For the critical part of a course, the reflection will be evaluated through the survey from experts and students. However, there can be many different web pages for an individual course and it's unreasonable to evaluate all the web pages before determining the reflection value of a course. Learning action analysis can be helpful in evaluating the reflection ratio of such web pages.

Table 2. Web page information in action messages

Time	IP address	Message Type	Description
10:00:01 2008-9-1	192.168.15.3	Password Checking	*URL for password checking*
10:00:12 2008-9-1	192.168.15.3	Password Ok	*URL for login welcome*
10:00:25 2008-9-1	192.168.15.3	Learning interface fetching	*URL for material loading*
…	…	…	
10:01:02 2008-9-1	192.168.15.3	Note add page fetching	*URL for note add*

As stated in Table 2, there will be URL information in the action message and each line contains the unique web page address. Hyperlinks analysis techniques [21, 22, and 23] can be used to identify all the hyperlinks contained in a certain web page. Students may view the relevant knowledge resource from the hyperlinks.

3.2 Tutor Support Evaluation

In the current co-learning framework, there will be support from tutor in the classroom. However, it's rather difficult to evaluate the tutor support online. Currently, we only evaluate the tutor support from two items: one is the mean time for the tutor to capture the problem which the student published and another is the mean time for the tutor to reply the problem.

Figure 2 shows the procedure of the tutor support evaluation. There will be a thread continues to search the activity of problem publishing. When it find such activity, it will keep the current location in MySQL database for the next searching and then get the keywords of the problem and the location of the problem. It analyzes the following log and find out the nearest action from the tutor which may possible capture the location of the problem and the nearest action from the tutor which reply the problem with the same keywords. The total problems published (*NoP*), the time for the tutor to capture the problems (*T2C*) and reply the problem (*T2R*) will be updated. The mean time for the tutor to capture the problem which the student published and then the mean time for the tutor to reply the problem can be calculated as,

$$MT2C = \sum T2C / NoP,$$
$$MT2R = \sum T2R / NoP.$$

MT2C indicates how fast will a tutor capture the problems and *MT2R* indicates how fast will a tutor bring forth their answer.

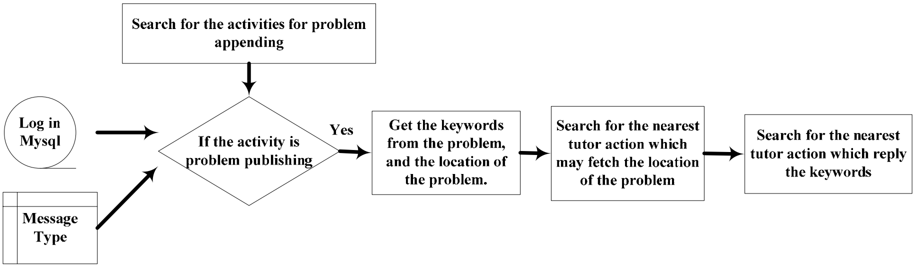

Search for the activities for problem appending
Log in Mysql
Message Type
If the activity is problem publishing
Yes
Get the keywords from the problem, and the location of the problem.
Search for the nearest tutor action which may fetch the location of the problem
Search for the nearest tutor action which reply the keywords

Fig. 2. Procedure of tutor support evaluation

4 Case Study and Discussion

The course management system in our college currently covers 26 courses and 2 contests. It has been used for 2 terms and more than 300 students have used this system. The students appraised the system very much, and they are willing to hand in their homework in time via the online system. The teachers were happy to find the students were more willing to learn the knowledge in the class-room, and the final score was a little improved with such enhancement. The system will try to cover the students from other major including e-business and e-logistics management majors. The courses will then include not only the computer science but also software engineering, e-commercial, and information technology. Some research projects mainly performed by the students will also be managed in this system.

There is a portal for the teachers and students to logon to the system. The main functions of the course management system in using are:

- Teachers upload teaching material.
- Students download the related material.
- Students hand in their exercise and homework to the system.
- Teachers review the students' exercise and homework and score them.

There is a main course page for the course "Software Design and Modeling". There are several columns for the courses management, including the glossary of the course, the public material of the course, the calendar & events, knowledge base, examination, and etc.

Having established the course management system using Moodle, the faculty started to design the framework of co-learning system. The co-learning system not only cover the scope of course management but also cover the students' learning path.

Such system should enhance the students' critical thinking capability, benefit the students' professional practices, and facilitate the communication between teachers and students.

The course manage system provides the web site as a tool for the student to navigate the course. While the co-learning system should carefully design the content, activities, and resources in the web site so as to ensure the quality of education. The participation and support of the students and the teachers are also very important and co-learning system is an integrated solution for the connection.

Currently, the co-learning system is deployed in campus environment, the connection speed to network is quite fast, and the teacher may get some rich material online to enhance the interactivity. But if we want to expand the scope to the internet, the connection speed will be reduced significantly, the facility is now thinking a way to deal with some problem. If succeed, distance education can be realized on this co-learning system.

5 Conclusion and Future Work

A course management system has been established using Moodle. The cost was very low as it only used the open source software including Linux, Apache2, Php, and MySQL. An online co-learning system is being designed and the learning path of the students is focused currently. The criteria to evaluate the quality of education are proposed.

The courses are high coupling in the university, however the relationship between these courses is not considered in current proposed quality of education. This should be researched in our future studies. Besides, in order to better sharing the education resource, the network scope should be expanded so as to gain more value. The network connection speed will become the bottleneck for such online learning system and rich online content turns out to be obstacle for the efficient learning. A potential solution is to dynamically reduce the rich online content when the connection speed is slow.

Acknowledgments. This research was financial supported by National Natural Science Foundation of China with No. 60873022 to Dr. Hua Hu, The Science and Technology Department of Zhejiang Province, China with No.2008C23034 to Dr. Hua Hu, No. 2008C11009 to Dr. Bin Xu, and No. 2008C13082 to Dr. Xiaojun Li, and Education Department of Zhejiang Province with No. yb07035 to Prof. Yun Ling and No. 20061085 to Dr. Bin Xu. The authors would like to express their special thanks to the development team and the students assisted in this research work.

References

1. Adams, J.: Then and now: Lessons from history concerning the merits and problems of distance education. Journal of Studies in Media & Information Literacy Education 7(1), 1–14 (2007)
2. Thrope, M., Godwin, S.: Interaction and e-learning: the student experience. Studies in Continuing Education 28(3), 203–211 (2006)

3. Watson, D., Tinsley, D. (eds.): Integrating information technology into education, pp. 169–184. Chapman & Hall, London (1995)
4. Taylor, P., Maor, D.: Assessing the efficacy of online teaching with the Constructivist On-Line Learning Environment Survey. Paper presented at the 9th Annual Teaching Learning Forum - Flexible Futures in Tertiary Teaching. Perth: Curtin University of Technology (2000)
5. Behan, C.: Context, Creativity and Critical Reflection: Education in Correctional Institutions. Journal of Correctional Education 58(2), 157–169 (2007)
6. Fisher, A.: Critical reflective thinking: an introduction. Cambridge University Press, Cambridge (2001)
7. Barab, S., Thomas, M., Merrill, H.: Online learning: From information dissemination to fostering collaboration. Journal of Interactive Learning Research 12(1), 105–143 (2001)
8. Bernard, R., Rojo-de-Rubaclava, B., St. Pierre, D.: Collaborative online distance learning: issues for future practice and research. Distance Education 21(2), 260–277 (2000)
9. Dede, C.: Distance learning—Distributed learning: Making the transformation. Learning and Leading with Technology 23(7), 25–30 (1996)
10. Fischer, M., Coleman, B.: Collaborative online learning in virtual discussions. Journal of Educational Technology Systems 30(1), 3–17 (2001)
11. Merriam, S.B.: Qualitative research and case study applications in education. Jossey-Bass, San Francisco (1998)
12. Murphy, K., Cifuentes, L.: Using web tools, collaborating, and learning online. TechTrends 45(1), 28 (2001)
13. Northrup, P.: Online learners' preferences for interaction. Quarterly Review of Distance Education 3(2), 219–226 (2002)
14. Vouk, M.A., Bitzer, D.L., Klevans, R.L.: Workflow and end-user quality of service issues in Web-based education. IEEE Transactions on Knowledge and Data Engineering 11(4), 673–687 (1999)
15. Järvelä, S., Häkkinen, P.: Web-based Cases in Teaching and Learning – the Quality of Discussions and a Stage of Perspective Taking in Asynchronous Communication. Interactive Learning Environments 10(1), 1–22 (2002)
16. Koper, R., Tattersall, C. (eds.): Learning Design: A Handbook on Modelling and Delivering Networked Education and Training. Springer, Heidelberg (2005)
17. Luther, C.: Moodle in Corporate Education, `http://moodle.org/mod/forum/discuss.php?d=3150`
18. Bourne, J., Harris, D., Mayadas, F.: Online Engineering Education: Learning anywhere, anytime. Journal of Engineering Education 94(1), 131–146 (2005)
19. Mallinson, B., Sewry, D.: eLearning at Rhodes University — A Case Study. In: Fourth IEEE International Conference on Advanced Learning Technologies, pp. 708–710 (2004)
20. Aguilar, D., Theron, R., Pealvo, F.G.: Understanding Educational Relationships in Moodle with ViMoodle. In: Eighth IEEE International Conference on Advanced Learning Technologies, pp. 954–956 (2008)
21. Dean, J., Henzinger, M.: Finding Related Pages in the World Wide Web. In: Proc. Eight Int'l World Wide Web Conf., pp. 389–401 (1999)
22. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. J. ACM 46, 668–677 (1999)
23. Bharat, K., Henzinger, M.: Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In: Proc. 21st Int'l ACM Conf. Research and Development in Information Retrieval, pp. 104–111 (1998)

Research on Face Recognition Technology Based on Average Gray Scale

Weihua Wang

School of Computer, Chongqing University of Arts and Science,
YongChuan, Chongqing, China, 402160
y2002ww@163.com

Abstract. The network of a face recognition system is very complex and therefore difficult to train. In order to reduce the complexity, a new face recognition algorithm based on average gray scale was proposed in this paper. Discussed are the data structures of the face feature vector and the average gray scale vector for face recognition. The paper also shows the readers the advantage of the new algorithm, Experiments have been conducted on ORL face database, the results show that the feature vector could be described easily by the gray-scale, the vector can be extracted in short time, also the neural network could decrease the training time, and this new method has higher recognition rate than Eigenface Algorithm in the same experiment conditions according to our practices.

Keywords: face recognition, gray scale image, neural networks, feature extracting, classifier, image pre-processing.

1 Introduction

Nowadays, face recognition system has become commonplace, it is getting more and more important in the position of security systems, criminal identification, image and film processing, and human-computer interaction.

In principle, the popular back-propagation neural network may be trained to recognize face images directly. For even an image with moderate size, however, the network can be very complex and therefore difficult to train. For example, if the image is 100x100 pixels, the number of inputs of the network would be 10,000. To reduce complexity, neural network is often applied to the pattern recognition phase rather than to the feature extraction phase.

A face recognition system based on average gray scale algorithm is introduced in the following. The system consists of five modules: face images pre-processing, feature extraction with average gray scale, face training, face recognition and database. In feature extraction module, average gray scale, a supervised learning algorithm that can compute low dimensional, neighborhood-preserving embeddings of high dimensional data which is used to reduce data dimension and extract features. The experiments show that the proposed method is superior to Eigenface algorithm.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 158–165.
springerlink.com

2 Image Pre-processing

Image pre-processing, including gray conversion and size normalization is helpful for the following face recognition system.

(1) Gray Scale Transformation

We use $f(x, y)$ to express the gray value of the position(x, y). The expression of the Gray-Scale conversion is:

$$f(x, y) = 0.299 \times r + 0.587 \times \mathrm{g} + 0.144 \times b + 0.5$$

Were r , g and b are the Red, the Green and the Blue values of the color image pixel.

(2) Standardization of Image Size

In order to define the numbers of the input nodes of the neural networks, the data of the input images must be passed the standardize process. The expression of the size normalization of the face image in this paper is:

$$RS = k * TS$$

WhereRS is the real value of the vehicle parameter, k is the ratio coefficient of the whole vehicle image, and TS is the value measured form the vehicle image.

3 Feature Vector

3.1 Face Feature Vector

Many features can be used to describe a human face but these feature data have large quantities of data and several interrelated variables acting together. Processing these large amounts of data has created new concerns with respected to data representation, disambiguation, and dimensionality reduction. Some traditional algorithms of feature vectors for face recognition are described in the following.

(1) NMF Algorithm

A Nonnegative Matrix Factorization (NMF) algorithm for feature extraction and identification in the fields of text mining and spectral data analysis is given in [8, 9, 10]. The nonnegative Matrix Factorization problem can be stated as follows:

Given a nonnegative matrix $A \in R^{m \times n}$ and a positive integer$k < \min\{m, n\}$, find nonnegative matrices $W \in R^{m \times k}$ and $H \in R^{k \times n}$ to minimize the functional

$$f(W, H) = \frac{1}{2} \| A - WH \|_F^2$$

Where WH is called a nonnegative matrix factorization of A , and it is an approximate factorization of rank at most k . An appropriate decision on the value of k is critical in practice, but the choice of k is very often problem dependent. Following shows an example of the application of NMF to represent the human faces.

The size of an ORL face image is 112×92 with 256 gray levels per pixel representing a front view of the face of a person. The image is transformed into face vector in R^{10304} (112×92=10304) to form the data matrix A of size 10304, so the size of the face vector is too large.

(2) PCA Techniques

Principal Component Analysis (PCA) is an optimal linear dimensionality reduction scheme with respect to the mean squared error (MSE) of the reconstruction. PCA is theoretically the optimal linear scheme, in terms of least mean square error, for compressing a set of high dimensional vectors into a set of lower dimensional vectors and then reconstructing the original set. It is a non-parametric analysis and the answer is unique and independent of any hypothesis about data probability distribution. However, the latter two properties are regarded as weakness as well as strength, in that being non-parametric, no prior knowledge can be incorporated and that PCA compressions often incur loss of information. The applicability of PCA is limited by the assumptions made in its derivation. These assumptions are:

- Assumption on Linearity.
- Assumption on the statistical importance of mean and covariance.
- Assumption that large variances have important dynamics.

(3) Discrete Cosine Transformation

The discrete cosine transform (DCT) is a technique for converting a signal into elementary frequency components. A DCT expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies.

The DCT attempts to decorrelate the image data. After decorrelation each transform coefficient can be encoded independently without losing compression efficiency. Figure 1 shows an example of DCT applied in image processing. (a) is the original image,(b) is the image transformed from (a) by DCT algorithm, and (c) is the image transformed from (b) by IDCT.

(a) (b) (c)

Fig. 1. The original image

There are multiple advantages to using the DCT for application purpose. The first main advantage of the DCT is its efficiency. Another advantage of the DCT is that its basis vectors are comprised of entirely real-valued components.

However, the process of discrete cosine transformation is quit complex. Although the Fast Discrete Cosine Transform (FDCT) has improved the compute speed, the computing process is still very time-consuming.

3.2 Feature Vector

In order to resolve the above problems, a new feature method describing a human face is proposed in this paper. Following is the particular discussion of this new method.

(1) Gray Scale Feature

A gray scale (or gray level) image is simply one in which the only colors are shades of gray. The reason for differentiating such images from any other sort of color image is that less information needs to be provided for each pixel. In fact a `gray' color is one in which the red, green and blue components all have equal intensity in RGB space, and so it is only necessary to specify a single intensity value for each pixel, as opposed to the three intensities needed to specify each pixel in a full color image.

Often, the gray scale intensity is stored as an 8-bit integer giving 256 possible different shade of gray from black to white. If the scales are evenly spaced then the different between successive gray scales is significantly better than the gray scale resolving power of the human eye. So, gray scale image is used to describe the face feature for face recognition.

(2) Average Gray Scale

If the whole gray scale image is used to the input vector for face recognition, the recognition network can be very complex and there difficult to train. For example, A 640 x 480 grey scale image requires over 300 KB of storage. In order to reduce the size of the input vector, the average of the gray scales was used in the new face recognition approach. And according to the direction of the image coordinate, the average gray scale was classified into the following categories:

- Vertical average gray scale
- Horizontal average gray scale

(3) Average Gray Scale Vector for Human Face

We assume that the height of an image is h , the width of the same image is w , the vector of vertical average gray scale is $avg[w]$, the vector of horizontal average gray scale is $ahg[h]$, and the value of the gray scale of a pixel of the gray scale image is $f(x, y)$, and $f(x, y) \in [0,255]$.

Then the value of the vector of vertical average gray scale is

$$avg[i] = \sum_{y=0}^{h} f(i, y) \qquad i \in [0, w)$$

The value of the vector of horizontal average gray scale is

$$ahg[j] = \sum_{x=0}^{w} f(x, j) \qquad j \in [0, h)$$

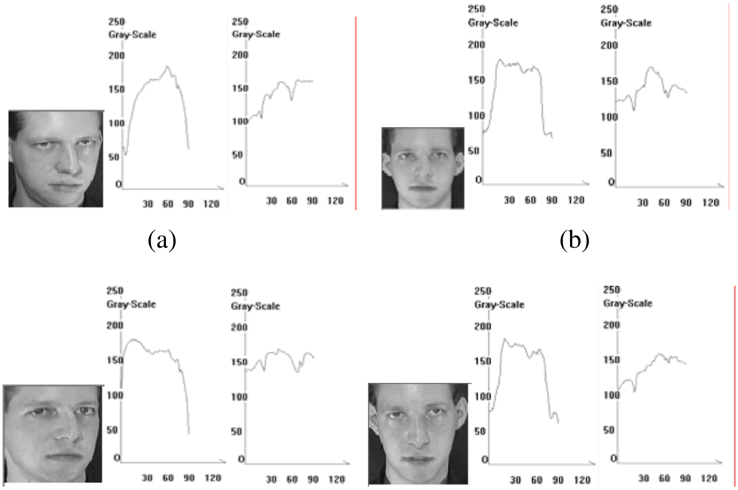
Therefore the whole expression of the feature extracting from image is:

$$p[j] = \begin{cases} avg[j] & j \in [0, h) \\ ahg[j-h] & j \in [h, j-h+w) \end{cases}$$

(4) Different Human Face Vectors

In order to illustrate the feasibility of the gray scale vectors for face recognition, following shows some examples of face vectors using this method.

Figure 2 shows the face vectors obtained by the same man with different view direction. The left image of each figure in figure 2 is the face image, the middle image of each figure in figure 2 is the vector of vertical average gray scale of the man, and the right image of each figure in figure 2 is the vector of horizontal average gray scale of the same man.

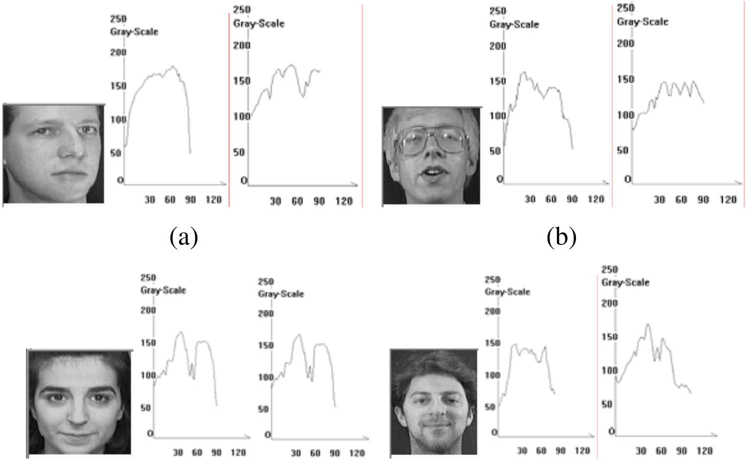


(a) (b)

(c) (d)

Fig. 2. Vectors of the same man

Figure 3 shows the face vectors obtained by the different men. The left image of each figure in figure 3 is the face image, the middle image of each figure in figure 3 is the vector of vertical average gray scale of the left man, and the right image of each figure in figure 3 is the vector of horizontal average gray scale of the same left man.



(a) (b)

(c) (d)

Fig. 3. Vectors of different humans

The face vectors of figure 2 shows that the discrimination of the vectors in the same individuals faces are very similar. And according to the face vectors of figure 3, the vectors of the different people's faces are very large different from each other.

Therefore, the vectors described above new method would be primly applied to express the face feature in the practical face recognition system.

4 Feature Extraction Algorithm Analysis

4.1 The Feature Extraction Algorithm

Feature extracting processing is the third step of the face recognition system. It is helpful for the following face recognition. We derive the whole expression of gray conversion as above method during the extract processing.

4.2 Complexity Analysis

In this paper, back-propagation neural network is used to be as the classifier of the above recognition. In principle, the popular back-propagation neural network may be trained to recognize face images directly. For even an image with moderate size, however, the network can be very complex and therefore difficult to train. We derive the expression of the image-size vector as follows during this processing. The size of the face image is:

$$v = N \times \mathrm{M}$$

Where N is the value of the image-height h, M is the value of the image-width w.

The size of the feature vector extracted by our new method is:

$$fv = N + \mathrm{M}$$

For example, the size of the feature vector in the following experiments is:

$$fv = N + \mathrm{M} = 100 + 100 = 200$$

However, the size of the face image in the following experiments is:

$$v = N * \mathrm{M} = 100 * 100 = 10000$$

Therefore, the complexity of the face recognition could be reduced effectively.

5 Experiment and Results

In order to prove the validity of our new approach introduced above, extensive experiments are conducted on two face database: the ORL database and our own database, which consists of 50 facial images of 5 individuals. Some examples of the test sets are shown in Figure 4.

(a) (b)

Fig. 4. Some Examples from ORL Face Database

Table 1 shows the times of the feature vectors extracting obtained by different method.

Table 1. Feature Extracting Time

Method	*Vector Extracting time(millisecond)*
FDCT	*157*
Average Gray Scale	*2.2*

A performance about the comparison of various image sizes is shown in table 2. This comparison adopted the public ORL face database.

Table 2. Training Time

Image size	*Training time(second)*
100×100	*543*
80×80	*319*
60×60	*186*

In Table 3, it demonstrates the result of our database by different approach.

Table 3. Recognition Rate

approach	*Recognition rate (%)*
Tradition Approach	*92.78*
Average Gray Scale Approach	*93.33*
Eigenface Approach	*76.08*

6 Conclusion

Face recognition is a both challenging and important technique. In this paper, a new approach based on average gray scale was introduced. The excellent performance of our approach demonstrated by extensive experimental results confirms the following:

- The novel feature vector model can make efficient use of the information efficiently in face subspace.

- The novel model can effectively reduce the input data and the input nodes of the network.
- The feature vector can be extracting in short time.
- The neural network could decrease the training time effectively.
- The recognizing time could be decreased effectively.

Therefore, an obvious extension of our work would apply the novel input model to model the practical face recognition system.

References

1. Moon, H., Phillips, P.J.: Computational and Performance aspects of PCA-based Face Recognition Algorithms. Perception 30, 303–321 (2001)
2. Chen, X., Gu, L., Li, S.Z., Zhang, H.-J.: Learning representative local features for face detection. In: 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 1126–1131 (2001)
3. van der Schaar, M., Chen, Y., Radha, H.: Embedded DCT and Wavelet Methods for Scalable Video: Analysis and Comparison. Visual Communications and Image Processing (January 2000)
4. Yang, Y., Xuping, Z.: General Theory Research on Morphological Correlation for Gray2Scale Face Recognition. Acta Photonica Sinica 35(2), 299–303 (2006)
5. Ryu, H., Yoon, J.-C., Chun, S.S., Sull, S.: Coarse-to-Fine Classification for Image-Based Face Detection. In: Sundaram, H., Naphade, M., Smith, J.R., Rui, Y. (eds.) CIVR 2006. LNCS, vol. 4071, pp. 291–299. Springer, Heidelberg (2006)
6. Hu, N.-p., Tian, J.-x.: Evaluation of the growth of real estate financial system based on BP neural network. In: Liu, D., Fei, S., Hou, Z., Zhang, H., Sun, C. (eds.) ISNN 2007, Part III. LNCS, vol. 4493, pp. 49–56. Springer, Heidelberg (2007)
7. Lin, S.-H.: An Introduction to Face Recognition Technology. Information Science Special Issue on Multimedia informing Technologies=Part 2 3(1), 1–7 (2000)
8. Hu, Y., Yin, B., Cheng, S., Gu, C., Liu, W.: Research on key technology in construction of a chinese 3d face database. Journal of Computer Research and Development 42(4), 622–628 (2005)
9. zhao, L., Liu, J., Xu, X.: A Survey of Human Face Detection. Application Research of Computers 21(9), 1–4 (2004)
10. Buciu, I.: Non-negative Matrix Factorization, A New Tool for Feature Extraction: Theory and Applications. Int. J. of Computers, Communications & Control (III) (2008); ISSN 1841-9836, E-ISSN 1841-9844, Suppl. issue: Proceedings of ICCCC 2008, pp. 67–74 (2008)
11. Berry, M.W., Browne, M., Langville, A.N., Pauca, P.V., Plemmons, R.J.: Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis 52(1), 155–173 (2007)
12. Liang, D., Yang, J., Chang, Y.: Relevance feedback based on non-negative matrix factorisation for image retrieval. IEE Proceedings-Vision Image and Signal Processing 153(4), 436–444 (2006)
13. Shlens, J.: A Tutorial on Principal Component Analysis. Version 2
14. Frigo, M., Johnson, S.G.: The Design and Implementation of FFTW3. Proceedings of the IEEE 93(2), 216–231 (2005)

The Active Leveled Interest Management in Distributed Virtual Environment

Jia Bei[1] and Yang Zhao[2]

[1] Nanjing University, China
beijia@software.nju.edu.cn
[2] Nanjing University of Sci. & Tech., China
yangzhao@mail.njust.edu.cn

Abstract. Active interest management manages interest with active routing method. It combines communication and interest management together to decide communication relationship between objects in distributed virtual environment (DVE). On analysis of active interest management, leveled interest management and some other related technologies, concept of level of interest (LOI) and an evaluating model of LOI between objects in DVE is put forward in active leveled interest management. LOI is used to control the detail of communication to reduce traffic load. Therefore, the system's scalability can be improved furthermore.

Keywords: distributed virtual environment, active interest management, level of interest, active leveled interest management.

1 Introduction

By introducing active routing technologies into interest management, Zabele [1] find a new method, Active Interest Management (AIM). Participators in DVE communicate in publish-subscribe pattern while Active Routers (ARs) manage interest instead of servers by delivering datagram according to subscriptions. Since communications are filtered on ARs, the network traffic is controlled. Instead of SBT (source-based tree) by Zabele [1], we built an AIM system, AIMNET, using CBT (core-based tree) [2]. In this system, less multicast addresses are consumed, multicast tree is easy to manage, and DVE can even be expanded online.

However, the researches on AIM still focus on how to find and achieve communications between objects in DVE. How to find the proper level of detail or frequency of communication according to the objects' statuses and features remains a new topic. In this paper, we combine the leveled filtering with AIM and put forward a method of active leveled interest management.

2 AIMNET

As shown in Figure 1, ARs and hosts are organized as a CBT, and the root is the core AR. Every application in DVE can build its own CBT separately.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 166–173.
springerlink.com

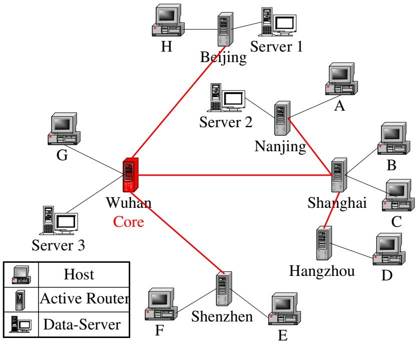


Fig. 1. Communication Architecture of AIMNET

In order to be compatible with HLA [3], the subscribers maintain Subscription Regions (SRs) and the publishers maintain Publication Regions (PRs). The interfaces of AR and data structures related are abstracted as Virtual Interfaces (VIFs), which stand for connections to other ARs or hosts. If a VIF connects the upstream AR, it is called upstream VIF. Otherwise, it's called downstream VIF. The VIF's SR is the summation of subscriptions of hosts connected to this VIF, directly or indirectly.

We denote the $n+1$ ARs in system to be AR_0, $AR_1 \ldots AR_n$, and AR_0 is the core AR. If there are $m+1$ VIFs on AR_i, they are denoted as VIF_{i0}, $VIF_{i1} \ldots VIF_{im}$, and VIF_{i0} is the upstream VIF. The SR of VIF_{ij}, $0 \le j \le m$, is denoted as $SR(VIF_{ij})$. Since AR_0 is the core AR, VIF_{00} and $SR(VIF_{00})$ are both empty. For a non-core AR, AR_i, $1 \le i \le n$, its upstream AR is AR_k, then among the p downstream VIFs on AR_k, there is a VIF_{kq} connecting VIF_{i0}. If the set of all VIFs in system is denoted as $VIFSET$, we can define the character function, $isAR$: $VIFSET \rightarrow \{true, false\}$. If VIF_{ij} connects AR, $isAR(VIF_{ij})$ is *true* and the AR is denoted as AR_{ij}. If VIF_{ij} connects host, $isAR(VIF_{ij})$ is *false* and the host is denoted as H_{ij}. Now we can define the SRs of VIF and AR as follows.

$$SR(VIF_{ij}) = \begin{cases} SR(AR_{ij}) & j \neq 0 \wedge isAR\ (VIF_{ij}) \\ SR(H_{ij}) & j \neq 0 \wedge \neg isAR\ (VIF_{ij}) \\ \bigcup_{l=0}^{q-1} SR(VIF_{kl}) \cup \bigcup_{l=q+1}^{p} SR(VIF_{kl}) & j = 0 \wedge i \neq 0 \\ \varnothing & j = 0 \wedge i = 0 \end{cases} \tag{1}$$

$$SR(AR_i) = \bigcup_{j=1}^{m} SR(VIF_{ij}) \tag{2}$$

As shown in Figure 2, after the acceptance of subscription from downstream VIF, SRs of this VIF and the other VIFs connecting this AR will be combined and updated sequentially.

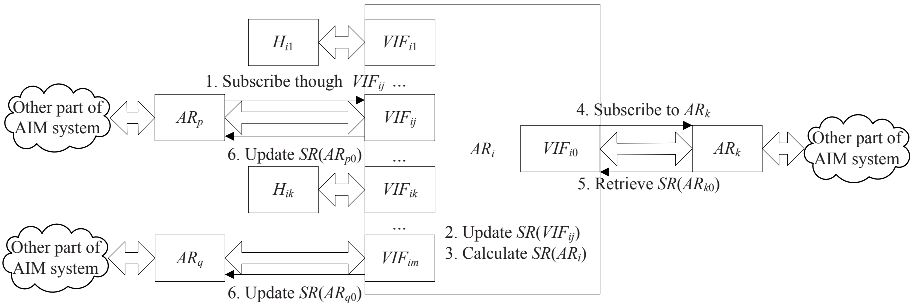
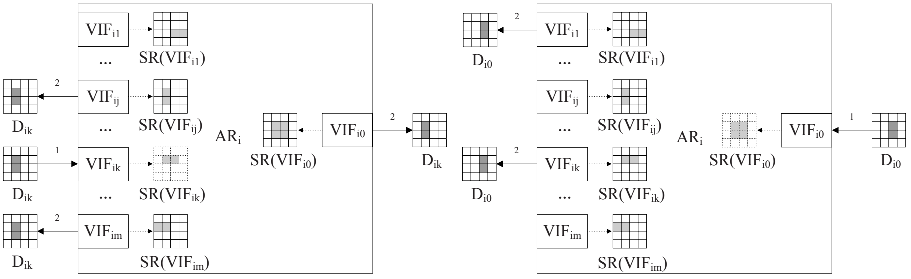


Fig. 2. Subscription Job on an AR

Hosts publish their datagram to AIM system though their agent ARs. When datagram arrives AR_i, $0 \leq i \leq n$, though VIF_{ik}, which have m+1 VIFs, the datagram is denoted as D_{ik} and its PR is denoted as $PR(D_{ik})$. We can define the match function between D_{ik} and VIF_{ij}, $0 \leq j \leq m$ as

$$match(D_{ik}, VIF_{ij}) = (j \neq k \wedge PR(D_{ik}) \cap SR(VIF_{ij}) \neq \varnothing) \tag{3}$$

The results of match function determine whether D_{ik} should be delivered though VIF_{ij}. If none of VIFs accepts D_{ik}, D_{ik} will be discarded. Two examples of datagram delivery are shown in Figure 3.



(a) Datagram From Downstream VIF (b) Datagram From Upstream VIF

Fig. 3. Examples for Datagram Delivery

In AIM system, the delivery of datagram is content-based. Neither AR nor host needs to know about the globe status. They just maintain their own SRs, publish and deliver datagram with PRs. This non-status method insures the robustness and scalability.

3 Leveled Interest Management

In order to save resources and network bandwidth, many technologies in DVE system divide objects to layers and offer different kinds of services. These technologies include Graphics LOD [4], motion LOD [5], multi-thresholds in dead reckoning [6], contend-based method in stream transmission [7], spatial interaction model in DVE [8], the extend HLA with layered priority[9] and so on.

From the point of view of interest management, if we use Level of Interest (LOI) to measure the participator‘s interest to a certain object, the above-mentioned technologies maintain LOIs between objects, and use them to determine the details of communications. We categorize these technologies to leveled interest management. Since the bottleneck of DVE is hosts' process ability and network bandwidth, leveled interest management mainly focuses on how to reduce host's traffic load and rendering time. Publishers in system must use extra multicast addresses and network resources to publish copies of data in different levels.

In these "mirror" methods, subscribers receive less data, but publishers publish more. In DVE, participators often act as subscribers and publishers at the same time, so leveled interest management brings little benefit.

4 Active Leveled Interest Management

In AIM system, in order to hide communication from publishers and subscribers, ARs have the power to handle and deliver datagram. If subscribers express their requirement of LOI in subscriptions and ARs evaluate LOI and perform leveled filtering, we can expand AIM to active leveled interest management. Then traffic load can be reduced and scalability can be improved furthermore. In active leveled interest management, besides what we have discussed above, there are some new problems. The models and parameters to evaluate LOI, the expressions of LOI calculation, leveled filtering methods and tasks related will be discussed in this section.

The parameters to evaluate LOI in traditional leveled interest management are mainly based on motion or awareness models, such as distance, size, eccentricity and velocity [10]. For a single participate in DVE, these parameters evaluate LOI very well. For example, if an object is closer, slower and in front of a participator, more details should be observed. But the sense ability and the information radiation ability of objects in DVE can be different. In order to simulate this status, participators in *Zhou*'s system [9] publish and subscribe on 5 different priority levels. After the division, 25 communication relationships are confirmed. However, in this method, number of division is fixed and the division is related to application. It's hard to modify after being settled.

If we regard objects' SRs and PRs as the representations of their sense ability and information radiation ability, then objects with larger SRs and PRs tend to observe more objects and be observed by more objects. Generally speaking, the sharp of SR and PR stands for the spatial extension of sense and information radiation. And the size stands for the ability. If we define the set of space region to be *Region*, we can define function $Volume : Region \rightarrow \Re$, $\Re$ is the set of real number. This function gives the size of region which is called volume function. If we denote the LOI from object A to object B to be $LOI(A,B)$, the SR of object A to be $SR(A)$, and the PR of object B to be $PR(B)$, then $SR(A) \cap PR(B)$ stands for both the part in the sense ability of object A that observes object B and the part in the information radiation ability of object B that is observed by object A. If we denote the result divided by $Volume(SR(A) \cap PR(B))$

to *Volume*(*SR*(*A*)) and *Volume*(*PR*(*B*)) as *PerS*(*A*,*B*), *PerP*(*A*,*B*), then they should be positively related to *LOI*(*A*,*B*). Besides, there are parameters from subscribers and publishers effecting LOI which stand for their statuses and features. We denoted the subscription leveled parameters (SLPs) from object *A* to be *SLP*(*A*), the publication leveled parameters (PLPs) from object *B* to be *PLP*(*B*). If there are *i* attributes in *SLP*(*A*), $SLP(A)_1$, $SLP(A)_2$…$SLP\ (A)_i$, *j* attributes in *PLP*(*B*), $PLP(A)_1$, $PLP(A)_2$…$PLP(A)_j$, they are *i*-dimensional and *j*- dimensional parameters, |*SLP*(*A*)|= *i*, |*PLP*(*B*)|= *j*. Then we can evaluate LOI from object *A* to object *B* with the following function.

$$LOI(A,B) = f(SLP(A), PLP(B), PerS(A,B), PerP(A,B)) \qquad (4)$$

In order to be compatible with the evaluation of awareness by *Greenhalgh* [8], generally we define the range of LOI to be [0, 1].

Size, eccentricity, velocity, the role and the organization of participators and any other characters that effect LOI can be expressed by leveled parameters. They can be any attributes in PR or SR, or they can be independent attributes. If the locations of objects are carried in leveled parameters, we can use them to calculate distance. Besides, if we define the volume function to be geometrical, then when *SR*(*A*), *PR*(*B*) are fixed, the bigger *PerS*(*A*,*B*) and *PerP*(*A*,*B*) are, the closer objects are. So we can also use them to estimate distance. In this model, everything in traditional leveled interest management can be covered appropriately.

In active leveled interest management, AR is the unit to evaluate LOI and perform leveled filtering. AR and VIF have SLPs. For AR_i, $0 \le i \le n$, its SLPs are denoted as $SLP(VIF_{ij})$. If AR_i has *m* downstream VIF, for a certain VIF_{ij},$1 \le j \le m$, its SLPs are denoted as $SLP(VIF_{ij})$ and the SLPs of H_{ij} are $SLP(H_{ij})$. After the acceptance of subscription on AR_i, the SLPs also need combining and updating. If we define the combining operation to be ∇, similar to SR, we can define the SLPs of VIF and AR as follows.

$$SLP(VIF_{ij}) = \begin{cases} SLP(AR_{ij}) & j \neq 0 \wedge isAR\ (VIF_{ij}) \\ SLP(H_{ij}) & j \neq 0 \wedge \neg isAR\ (VIF_{ij}) \\ \mathop{\nabla}\limits_{l=0}^{q-1} SLP(VIF_{kl}) \nabla \mathop{\nabla}\limits_{j=q+1}^{p} SLP(VIF_{kl}) & j = 0 \wedge i \neq 0 \\ \varnothing & j = 0 \wedge i = 0 \end{cases} \qquad (5)$$

$$SLP(AR_i) = \mathop{\nabla}\limits_{j=1}^{m} SLP(VIF_{ij}) \qquad (6)$$

If we combine *p* subscription, A_1, A_2…A_p, and denote the combined result to be *A*, then $SR(A) = \bigcup\limits_{i=1}^{p} SR(A_i)$ and $SLP(A) = \mathop{\nabla}\limits_{i=1}^{p} SLP(A_i)$. For any publication *B* and integer *k*, $1 \le k \le p$, $\forall k, B(SR(A_k) \cap PR(B) \neq \varnothing \rightarrow SR(A) \cap PR(B) \neq \varnothing)$ should be satisfied. So, as long as needed by any hosts connecting through this VIF directly or indirectly, datagram will be delivered through this VIF. At the same time, when leveled filtering are performed by VIF, we must ensure that LOI calculated from the SLPs of this VIF is not smaller than any LOIs of hosts connecting through this VIF directly or indirectly. With this insurance, subscribers can get enough details. For certain LOI function *f* and combining operation ∇, any publication *B* that matches publication *A* and integer *k*, $1 \le k \le p$, $\forall k, B(LOI(A,B) \ge LOI(A_k, B))$ should be satisfied. We can prove that the operation ∇ is related to *f*, and such an operation may not exist. But we can find a kind of functions *f* and operations ∇ which satisfy the condition listed above.

If the LOI function *f* takes *q*-dimensional SLPs, that is $|SLP(A)|=q$, for any subscription A_1, A_2 and publication *B*, if $SR(A_1)=SR(A_2)$, $SLP(A_1)_k \ge SLP(A_2)_k$, and $SLP(A_1)_i=SLP(A_2)_i$, where $1 \le i \le q$, $i \ne k$, $\forall B(LOI(A_1,B) \ge LOI(A_2,B))$ or $\forall B(LOI(A_1,B) \le LOI(A_2,B))$, then *f* is positively or negatively related to the *kth* SLP. We denote this to be $f\uparrow_k$ or $f\downarrow_k$. If *f* is positively or negatively related to all *q* SLPs, we call it monotone function about SLPs.

For a monotone function *f*, we can define a proper combining operation ∇ as follows. For any *q*-dimensional SLPs, $SLP(A_1)$, $SLP(A_2)$ and any integer *i* ,$1 \le i \le q$, if $SLP(A)=SLP(A_1) \nabla SLP(A_2)$, then

$$SLP(A)_i = \begin{cases} \max(SLP(A_1), SLP(A_2)) & f\uparrow_k \\ \min(SLP(A_1), SLP(A_2)) & f\downarrow_k \end{cases} \tag{7}$$

AR is responsible for calculating *PerS*(*A*,*B*) and *PerP*(*A*,*B*) from subscription *A* and publication *B*, so the volume of SR and PR should be carried in subscription and publication. According to definition, *Volume*(*SR*(*A*)) is negatively related to LOI. If we combine *p* subscriptions A_1, A_2...A_p, then

$$\mathop{\nabla}_{i=1}^{p} Volume(SR(A_i)) \le Volume(\bigcup_{i=1}^{p} SR(A_i)) \tag{8}$$

Upon all the definitions, active leveled interest management system can perform leveled filtering with LOI. If the content of datagram is ordered by priority, we can perform detail of content-based filtering (DOCBF). If the datagram is periodically, we can perform frequency-based filtering (FBF). In DOCBF, only segments needed of datagram are delivered. In FBF, datagram are picked out to deliver according to a certain frequency, and the others are discarded. We can perform one or both filtering method in different applications.

Actually, LOI function should de determined by application. We can't ensure it to be a monotone function. For function *f*, if $f\uparrow_k$, we add prefix "*i_*" to this attribute, if $f\downarrow_k$, we add prefix "*i_*", otherwise, we add prefix "*n_*". Table 1 shows an example of leveled parameter. Since the volume function of DOCBF and FBF can be different, there are different attributes to carry the volumes.

Table 1. An Example of Level Parameter

Type	Attributes	Value or Range
string	n_user agent	"bei.jia"
string	n_application	"e-classroom"
float	d_vol_DOCBF	10
float	d_vol_FBF	8
float	i_size	7
float	d_ velocity	15

In this rule, as long as there is no attribute with prefix "*n_*", the LOI function is monotone. After the acceptance of subscription, AR does not only combine and update SRs but also combine and update SLPs. After the acceptance of datagram, if AR decides to deliver it though a certain VIF, LOI will be evaluated by LOI function and be used to perform leveled filtering.

If the LOI function isn't a monotone function, only VIFs connecting hosts evaluate LOI and perform leveled filtering. The jobs of the other VIFs are similar with those in AIM.

Table 2. Examples of LOI Subscription

(a) A Complete LOI Subscription

Type	Attributes	Value or Range
string	volume_DOCBF	return spatial_x*spatial_y;

string	LOI_DOCBF	float temp_f= max(PerS,PerP);
		if (0< temp_f <=0.2) {
		    return 0.2;
		}
		else if (0.2<temp_f<=0.6) {
		    return 0.6;
		}
		else {
		    return 1.0;
		}

string	volume_FBF	return sizeof(medium_set);

string	LOI_FBF	if (subscriber.n_application
		    == publisher.n_application)
		return 1;
		return max(PerS, PerP);

(b) A Renew LOI Subscription

Type	Attributes	Value or Range

string	volume_DOCBF	return
		spatial_x*spatial_y*spatial_z;

string	LOI_FBF	return 1;

LOI function and the way to perform filtering can also be regarded as a kind of special interests from subscribers. So it ‘s supported naturally in AIM system. In DVE, with the passage of running time, the participators’ interests constantly change, but volume function and LOI function are relatively stable. So it is unnecessary to check them in every subscription. If we want to change them, we just send out a LOI subscription. Table 2 (a) shows a LOI subscription with both DOCBF and FBF. And volume function and LOI function are assigned separately. If some items remain unchanged in a new LOI subscription, they can be omitted. Table 2 (b) shows a renew LOI subscription which shuts FBF and changes the volume function of DOCBF. As we can see, the subscribers use simple programming language to express leveled interests. This process is similar to the ejection and running of routing code on AR in active network [11].

When publishers publish datagram, they don't need any information from subscribers. They just send datagram to system according to their own ability. After the acceptance of datagram on AR, we must know about the filtering performed before on other ARs and perform the proper filtering according to LOI calculated. So, two attribute *detail_DOCBF* and *detail_FBF* are introduced into datagram to record LOI of the DOCBF and FBF executed.

To sum up, in active leveled interest management system, according to different LOI functions, the LOI calculated can either be continuous or discrete. Instead of giving LOI, we define the function to evaluate it and the method to perform filtering. Every pair of subscription and publication may have different LOI and perform different filtering. So participators in DVE have dynamic LOI together with dynamic communication. The publishers publish datagram independently and there are no additional expenses for publishers. Besides, instead of a single server, the leveled filtering tasks are distributed on many ARs, which avoid the bottleneck. Comparing to traditional leveled interest management, active interest management is more flexible and easy to scale.

5 Conclusion

Active leveled interest management system uses ARs to perform leveled filtering in DVE. This paper puts forward a LOD evaluation model suitable for multiple objects in DVE and discusses some interesting problems related to active leveled interest management. By experiments of avatar's joint data transformation, we confirm that active leveled interest management can reduce network traffic and improve scalability of DVE furthermore. However the LOI evaluation and the filtering of datagram also introduce additional jobs to ARs. It impacts the system's performance negatively. How to measure this impact, the workload balancing, and QOS strategy for active leveled interest management would be our future work.

References

1. Zabele, S., Dorsch, M., Ge, Z., et al.: SANDS: Specialized Active Networking for Distributed Simulation. In: Proc. of the 2002 DARPA Active Networks Conference and Exposition, pp. 356–365. IEEE Computer Society, Washington (2002)
2. Sun, Y.H., Gong, Z.Y., Li, H., et al.: Research on Scalable Active Interest Management. Journal of Image and Graphics 8A(spec), 771–775 (2003) (in Chinese with English abstract)
3. Simulation Interoperability Standards Committee (SISC) of the IEEE Computer Society. IEEE standard for modeling and simulation (M&S) high level architecture (HLA)-IEEE std 1516.1-2000. Institute of Electrical and Electronics Engineers, Inc., New York (2000)
4. Chrislip, C.A.: Level of Detail Models For Dismounted Infantry In NPSNET-IV.8.1 [Ph.D. Thesis]. Naval Postgraduate School, Monterey, California (1995)
5. Carlson, D.A., Hodgins, J.K.: Simulation Levels of Detail for Real-time Animation. In: Davis, W.A., Mantei, M., Klassen, R.V. (eds.) Proc. of the conference on Graphics interface 1997, pp. 1–8. Canadian Information Processing Society, Toronto (1997)
6. He, L.Y., Li, S.K., Zeng, L., et al.: Hierarchical Interest Management in Large-Scale Distributed Virtual Environment. Journal of Computer Aided Design and Computer Graphics 12(9), 711–714 (2000) (in Chinese with English abstract)
7. Eide, V.S.W., Eliassen, F., Michaelsen, J.A.: Exploiting Content-Based Networking for Fine Granularity Multi-Receiver Video Streaming. In: Proc. of the 12th annual ACM international conference on Multimedia, pp. 104–105. ACM Press, New York (2004)
8. Greenhalgh, C.: Large Scale Collaborative Virtual Environments [Ph.D. Thesis]. The University of Nottingham, Nottingham (1997)
9. Zhou, Z., Zhao, Q.P.: Extend HLA with Layered priority. In: Proc. of the Spring Simulation Interoperability Workshop, Orlando (2003)
10. Berka, R.: Level of Motion Detail in Virtual Reality [Ph.D. Thesis]. Czech Technical University in Prague, Prague (2002)
11. Tennenhouse, D.L., Smith, J.M., Sincoskie, W.D., et al.: A Survey of Active Network Research. IEEE Communications Magazine 35(1), 80–86 (1997)

The Research of the Intelligent Fault Diagnosis Optimized by ACA for Marine Diesel Engine

Peng Li[1,3], Lei Liu[2], and Haixia Gong[1]

[1] Postdoctoral for Control Theory and Control Engineering, Harbin Institute of Technology,
150001 Harbin, China
[2] Department of Computer Science and Engineering, Harbin Institute of Technology,
150001 Harbin, China
[3] College of Automation, Harbin Engineering University,
150001 Harbin, China
lipeng@hrbeu.edu.cn, liulei8174@yahoo.com.cn,
gonghaixia@hrbeu.edu.cn

Abstract. The marine diesel engine has the important function to guarantee the marine security and reliability. It is a strong coupling relationship's system with multi-fault attributes. In this paper an advanced method of intelligent fault diagnosis based on fuzzy neural network (FNN) optimized and trained by ant colony algorithm (ACA) is proposed. The model, structure and parameters learning of intelligent fault diagnosis based on FNN were described concretely. The weight and the threshold value of this FNN are optimized and trained by the ant colony optimization algorithm. By simulation that has been carried out to evaluate the performance of proposed method and to compare with conventional FNN fault diagnosis method for this marine diesel engine's combustion system, the results show good quick convergence performance. The knowledge expression and the precision of fault diagnosis also can be improved effectively. Therefore, this method has the good application prospects in other similar system.

Keywords: diesel engine, intelligent fault diagnosis, fuzzy neural network (FNN), ant colony algorithm (ACA), optimization.

1 Introduction

The marine diesel engine is a complex system, the normal operation of which has the important function toward guaranteeing the marine security. It involves many domains of the machinery, the thermal energy, the signal examination, the safekeeping of security, the control and so on, and closely correlates with fuel oil, turbo-charged, combustion, cooling, lubrication and a series of subsystem. Therefore the question of its fault diagnosis needs synthetic research for each different aspect. It is a system of multi-fault attributes, namely input and output of fault pattern attribute are the multi-mapping relations. In recent years, the advanced intelligent fault diagnosis theory and method get swift and violent development. They provided the powerful guarantee for solving fault diagnosis of complex system. Now, some

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 174–181.
springerlink.com

researchers have studied the intelligent fault diagnosis on the turbo-charged subsystem of the marine diesel engine with RBF (Radial Basis Function) neural network [1] and the cooling subsystem of the marine diesel engine BP (Back Propagation) neural network [2]. Some good results were got relatively. These methods were easy and parameters were not used to optimize.

In the development of the intelligent fault diagnosis, the mapping process from the fault indication vector to the fault vector continuously may use the fuzzy method of reasoning to realize through the establishment fuzzy logic system. But the traditional fuzzy diagnosis rules lack association and self-study's capability, which could not meet the object change's need, thus the diagnosis effect will be affected seriously. With the neural network technology widely used, to establish method of fuzzy rules has attracted universal attention through self-study. The fuzzy neural network is mutually combined with the fuzzy logic system and the artificial neural networks, which has the merits of fuzzy logic system effective using fuzzy information and the neural network parallel processing, and has characteristic highly self-organization in studying information[3]. Currently, the fault diagnosis systems based on FNN have been used in many industrial fields [4], [5]. The train of algorithm for parameter usually uses BP algorithm. This algorithm exist shortcoming just like convergence rate is slow, easy to fall into local minimum.

Ant colony algorithm is a kind of stochastic search algorithm which is proposed by the Dorigo [6]. It is the same as the other simulated biological evolution algorithms, like genetic algorithm, simulated annealing algorithm, and so on. Ant colony algorithm can be used to solve combinatorial optimization problem. It has obtained some inspiring results [7], [8]. It is widely applied in many hot topic researches, such as the questions of the travel peddler, the optimized dispatch, the optical fibre in network route and so on [9], [10]. Recently the ant colony algorithm carries on the optimized training to the neural network has caused the scholar's recognition [11]. This paper uses the ant colony algorithm to optimize and train the connection weight and the function threshold value of FNN. Then this FNN based on the ant colony algorithm optimization training is applied to the intelligent fault diagnosis of marine diesel engine's combustion system. The simulation results show the feasibility and validity of the proposed method.

2 The Structure of Fuzzy Neural Network

The fuzzy neural network in this paper is a serial structure network of multi-layered forward feed. The model and structure of FNN is shown in Fig. 1. The FNN for fault diagnosis has two inputs and six outputs relative to marine diesel engine's combustion system. The input variables are the most highly explosive pressure and the exhaust temperature of the marine diesel engine's combustion system, which is measured by the corresponding sensors used in the fault diagnosis system. Respectively, they are P_z, T_r and are divided three fuzzy subsets namely {normal height low} = $\{N\ H\ L\}$. The Gaussian function of the membership function is as follow:

$$\mu_j(x)=\exp[-\frac{(x_i-a_{ij})^2}{\sigma_{ij}^{\ 2}}] \quad i=1,2 \quad j=1,2,3 \tag{1}$$

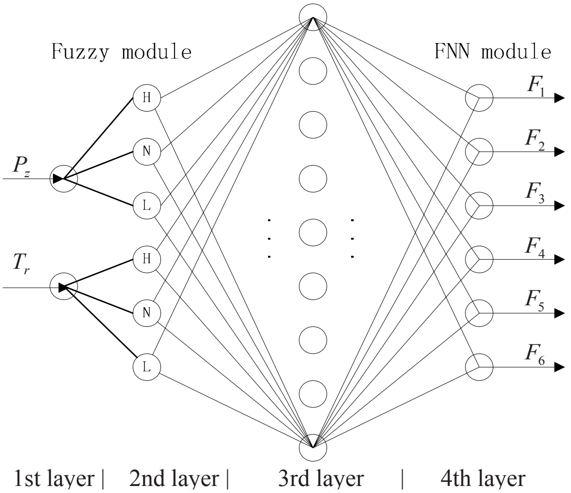


Fig. 1. The structure of fuzzy neural network

Where, the x of each membership function is real normalized variable. a_{ij} is the centre of the membership function. σ_{ij} is the width of the membership function. The corresponding fuzzy output can be calculated according to the different input.

The fuzzy neural network of the serial structure of multi-layered forward feed in this paper is composed by two parts. The frontal part is fuzzy quantification. The latter part is neural network part. In the network the temperature and pressure of the fault indication signal are connected with the network input layer. The fuzzy vector of fuzzy processing fault indication will input to the neural network for studying and training. It will obtain the fault confidence of fuzzy rules. Then the fault confidence will be determined by the rule logic operation. In fuzzy neural network the nodes of implicit layer number mostly are related with the information of input and output, and usually determined by the experience. Here we use 9 implicit layer nodes. So, the structure of fuzzy neural network is 2-6-9-6.

(1) The first layer is input layer which express fuzzy input signal of variable and has two nodes, namely real normalized variable input of temperature and pressure.
(2) The second layer is the fuzzy layer and has six nodes. The output of each node is the corresponding value of membership function, given by above membership function formula (1).
(3) The third layer is the fuzzy regular level and has nine nodes, namely the of number fuzzy reasoning rule strip for system diagnosis study. The non-linear function of this layer is tanh-function. That is $\sigma(x)=1-e^x \big/ 1+e^x$. Supposed the input of implicit layer is ($x_1, x_2, ..., x_6$), output is ($s_1, s_2, ..., s_9$).

$$s_i = \sigma(\sum_{j=1}^{6} w_{ij} x_j + b_i) \qquad 1 \le i \le 9 \tag{2}$$

The w_{ij} is connection weight of current input layer and implicit layer. The b_i is threshold value of implicit layer neuron.
(4) The fourth layer is the output layer, which non-linear function of this neuron is the *S* form function. That is $f(x)={1}/{1+e^{-x}}$. Supposed the output is ($y_1, y_2, ..., y_6$).

$$y_k = f(\sum_{j=1}^{9} s_{kj} x_j + b_k) \qquad 1 \le k \le 6 \tag{3}$$

The s_{kj} is connection weight of current implicit layer and output layer. The b_k is threshold value of output layer neuron.

In this fuzzy neural network, the adjustable parameter is the weights of 2 to 3, 3 to 4 layer and the threshold values of 3 and 4 layers. They are optimized training objects of ant colony algorithm in this paper for fault diagnosis simulation.

3 The Optimization and Study Algorithm of FNN Parameters

Ant colony optimization algorithm simulates the behavior of ant colony in nature when they are foraging for food and finding the most efficient routes from their nests to food sources. Biologist have found by a lot of espial and study that ants leave some chemical substance on the route that they have passed, which we call “pheromone”, and the quantity of pheromone is in inverse proportion to the length of the route. They would select a route with a probability in proportion to the concentration of the pheromone. With this way, the good effect can been obtained in solving complex combination optimization. For a certain kind of machine failure, the feature vector is selected as a training data. The accuracy of fault diagnosis on the nature will become the accuracy identification of fault regional boundaries. The basic thought of the ACA optimizing and training FNN in this paper is from reference [12]: the parameters need to train and optimize in the neural network have *M* parameters relative to intelligent fault diagnosis for combustion subsystem of marine diesel engine. They are 108 weights and 15 threshold values, and therefore *M* is 123. They are $p_1, p_2, ..., p_M$, where $p_i (1 \le i \le M)$ was establish by *N* stochastic non-vanishing number in the possible value scope composing a convene I_{p_i}. Then, the ant seeks food from the ant's nest. Supposed the ant number is *h*, each ant embarks from the first set. Based on information element $\tau_j(I_{p_i})$ of each element $p_j(I_{p_i})$ in set, a element from each set I_{p_i} ($1 \le i \le M$) is chosen stochastically, and the information element in the choosing element is adjusted correspondingly. All ants only can choose one element in each time step of ant search food. And to the different ant, in each time step the choosing element should belong to the different set. After the ant completes the choosing element in all sets, it arrives food source (a group of fuzzy neural network weight and threshold value is designated), then adjust the information element in set. This process carries on repeatedly, until the evolution tendency is not obvious or achieves the restraining iteration number of times. The train process of this FNN fault diagnosis system optimized by ACA can be completed on off-line. After FNN is optimized and trained, it can be applied to fault diagnosis system in real time. The process of the most optimal parameters algorithm through the ACA search satisfying error limits is as follows:

(1) Initialization: Making the time *t* and the cycle-index N_c is zero. Establish the biggest cycle-index $N_{c_{\max}}$ and make $\tau_j(I_{p_i})$=*C*, $\Delta\tau_j(I_{p_i})$=0. Set all ants in the ant's nest and *C* as constant.
(2) Starting all ants. Calculate transition probability based on the equation (4) in setting of I_{p_i} and ant *k* (*k=1, 2..., h)* .

$$\Pr(\tau_j^k(I_{p_i})) = \frac{\tau_j^k(I_{p_i})}{\sum_{g=1}^{N} \tau_g(I_{p_i})} \tag{4}$$

(3) Repeat (2), until all ants arrive at the food source.
(4) Then $t \leftarrow t+m, N_c \leftarrow N_c+1$. Compute neural network output value and error by ants choosing weight and record current optimal solution. After m unit of time, the ant arrives at food source from the ant's nest. The information content in various ways is renewed according to the equation (5) and (6).

$$\tau_j(I_{p_i})(t+m) = (1-\rho)\tau_j(I_{p_i})(t) + \Delta\tau_j(I_{p_i}) \tag{5}$$

$$\Delta\tau_j(I_{p_i}) = \sum_{k=1}^{40} \Delta\tau_j^k(I_{p_i}) \tag{6}$$

$\Delta\tau_j^k(I_{p_i}) = \frac{Q}{e^k}$. The ant k chooses the element $p_j(I_{p_i})$ in this cycle; otherwise its value is zero. In the formula, e^k set the output error of neural network weight by the ant k choosing group of weights. That is $e^k = \left|O - O_q\right|$. The O and O_q separately express actual output and the expectable output of this FNN. The smaller the error e^k is, the more the corresponding information element increases. $\rho(0 \le \rho < 1)$ expresses the information element durability. Q expresses the information element adjusting speed that is a constant.
(5) If the ant colony restrains completely to a way or $N_c \ge N_{\max}$, the cycle-index ends and outputs the computed result. Otherwise skips to step (2).

4 The Simulation Results of Intelligent Fault Diagnosis System

The simulation and the result analysis of optimizing and training the FNN fault diagnosis system based on the ACA for combustion subsystem of the marine diesel engine will be explained in this section. In succession, the structure model of intelligent FNN fault diagnosis is shown in Fig. 1. The most highly explosive pressure P_z and the discharge temperature T_r of the marine diesel engine are two inputs of this intelligent fault diagnosis system. The system takes six common faults of the diesel engine's combustion system as the diagnosis confidence level of output variable. The specification and the data originate of marine diesel engine system come from reference [13]. Its output vector elements are shown in Table 1.

Table 1. The physics significance of output vector element

Output vector	Physics significance
F_1	Fill-out of spray hole nozzle
F_2	Jamming of spray hole nozzle
F_3	Leaking of spray hole valve seat
F_4	Excessively late fuel injection timing
F_5	Premature fuel oil ejector timing
F_6	Jamming of exhaust pipe

Table 2. The fault diagnosis rules of system

Input		Output(Fault confidence level)					
Tr	P_Z	F_1	F_2	F_3	F_4	F_5	F_6
H	H	B	D	D	D	D	B
H	L	D	D	D	D	A	D
N	H	D	D	B	C	D	A
N	L	D	C	D	D	D	D
L	H	A	D	A	A	D	D
L	N	D	D	D	B	D	D

The related knowledge of intelligent fault diagnosis is needed to express by the form of the fuzzy rules after diagnosis system for the marine diesel engine based on FNN is built. Then, the training and study samples of fault diagnosis system are come into being according to these fuzzy rules. The fuzzy subsets of input characteristic parameters for diagnosis in fuzzy rules are divided three which are normal, height and low in this paper. The basic rules of intelligent fault diagnosis system for the marine diesel engine are set up and shown in Table 2 corresponding the above-mentioned method and theory. Where, A expressed "frequently", B expressed "sometimes", C expressed "very accidentally", D expressed "basic impossibility".

Table 3 is the representative samples which are for the fault diagnosis system training. The expression significance of partial value is 0.9 expressed "frequently", 0.6 expressed "sometimes", 0.2 expressed "very accidentally", 0.05 expressed "basic impossibility".

Table 4 is the confidence results using the BP algorithm to learn and train the fault diagnosis of FNN. Table 5 is the results using the ant colony algorithm to learn and train the fault diagnosis of FNN. In order to contrast the results, their learning and training sample all use table 3. Contrasting the results of table 4 and table 5 discovers following results in the process of simulating and learning to this system. When actual data converge to studied sample, to obtain the same fault diagnosis effect the fuzzy neural network by the BP algorithm training needs 239 steps to converge and mean squared error for representative diagnosis data is 0.0331. But the fuzzy neural network by the ACA optimizing and training needs 180 steps to converge. When new data deviate sample data, the fuzzy neural network by the ant colony algorithm optimizing and training can give the more accurate fault diagnosis result. Mean squared error for representative diagnosis data is 0.0137. So, compare to the fuzzy neural network using the BP algorithm to learn and train, the FNN by ACA optimization and train can get more precise diagnosis results. The parameter obtains that needs to adjust and select in the suitable range. In this simulation example, the parameter ρ is 0.4. The Q is 20. The ant quantity chooses h=60. Finally, it needs to be noted that the training results can be well enhanced by increasing the possible value set of weight and threshold value and ant colony quantity in certain scope.

Table 3. The training and study representative samples of fault diagnosis system

Input		Output(Fault confidence level)					
Tr	P_Z	F_1	F_2	F_3	F_4	F_5	F_6
358	138.0	0.60	0.05	0.05	0.05	0.05	0.60
358	123.2	0.05	0.05	0.05	0.05	0.90	0.05
341	138.0	0.05	0.05	0.60	0.20	0.05	0.90
341	123.2	0.05	0.20	0.05	0.05	0.05	0.05
322	138.0	0.90	0.05	0.90	0.90	0.05	0.05
322	131.4	0.05	0.05	0.05	0.60	0.05	0.05

Table 4. The simulation results of fuzzy neural network based on BP algorithm

Input		Output(Fault confidence level)					
Tr	P_Z	F_1	F_2	F_3	F_4	F_5	F_6
356	137.0	0.5645	0.0138	0.0126	0.0561	0.0398	0.6103
358	123.2	0.0448	0.0589	0.0249	0.0642	0.8779	0.0192
341	138.0	0.0967	0.0962	0.6248	0.2106	0.0751	0.8639
343	124.4	0.0192	0.1711	0.0579	0.0840	0.1005	0.0376
323	137.0	0.8788	0.0154	0.8581	0.9210	0.0280	0.0657
322	131.4	0.1296	0.0565	0.0720	0.5656	0.0280	0.0016

Table 5. The simulation results of FNN based on ACA optimizing and training

Input		Output(Fault confidence level)					
Tr	P_Z	F_1	F_2	F_3	F_4	F_5	F_6
356	137.0	0.6260	0.0583	0.0485	0.0474	0.0682	0.5899
358	123.2	0.0533	0.0513	0.0743	0.0483	0.9283	0.0579
341	138.0	0.0601	0.0397	0.6742	0.2361	0.0005	0.9137
343	124.4	0.0390	0.2033	0.0711	0.0396	0.0492	0.0578
323	137.0	0.9422	0.0508	0.9299	0.9122	0.0447	0.0422
322	131.4	0.1051	0.0270	0.0926	0.6327	0.0487	0.0215

5 Conclusions

In this paper, an advanced intelligent fault diagnosis system for the marine diesel engine using fuzzy neural networks and ant colony algorithms to optimize and train is developed to improve the performance of conventional intelligent fault diagnosis systems. The basic theory and method of the ant colony algorithm optimizing and training fuzzy neural network is introduced, comparing with the traditional intelligent fault diagnosis system of fuzzy neural network based on the BP algorithm. The fuzzy neural network given by the ant colony algorithm optimizing and training have the characteristics of quick convergence rate, enhance the accurate knowledge expression, and has the very good application prospects in other intelligent fault diagnosis system.

References

1. Huang, J.-l., Weng, Z.-m., Zhang, J.-d., Sun, P.-t.: The study of fault diagnosis for turbocharging system in low-speed marine diesel engine using RBF neural network. Journal of Dalian Maritime University 26(1), 9–13 (2000)
2. Wu, l., wanghua: Application of the BP Neural Network and Expert System in Fault Diagnosis. Information Technology 27(2), 66–68 (2003)
3. Lin, C.T., Lee, C.C.G.: Neural network based on fuzzy logic control and decision system. IEEE Trans. on computers 40(12), 1320–1336 (1991)
4. Ma, L., Lee, K.Y.: Fuzzy neural network approach for fault diagnosis of power plant thermal system under different operating points. In: IEEE Power and Energy Society 2008 General Meeting: Conversion and Delivery of Electrical Energy in the 21st Century, Pittsburgh, PA, United States, pp. 1–7 (2008)
5. Zhao, Y., Chen, L., Yang, Q.: The research on the fault diagnosis for boiler system based on fuzzy neural network. In: Proceedings of the 7th World Congress on Intelligent Control and Automation, Chongqing, China, pp. 8552–8556 (2008)
6. Dorigo, M., Maniezzao, V., Colorni, A.: The Ant System: optimization by a colony of cooperating agents. IEEE Transactions on systems, Man and Cybernetics Part B 26(1), 29–41 (1996)
7. Cheng, Q.-m., Wang, Y.-h.: The Study on Fuzzy Neural Network Controller Based on ACO Algorithm and Its Simulation. Journal of Shanghai University of Electric Power 22(2), 105–108 (2006)
8. Kopuri, S., Mansouri, N.: Enhancing scheduling solutions through ant colony optimization. In: IEEE International Symposium on Circuits and Systems, Vancouver, Canada, pp. 257–260 (2004)
9. Li, Y.-h., Wang, Z.: Use and reaching of ant algorithm forecasting daily water demand. Journal of Harbin Institute of Technology 37(1), 60–62 (2005)
10. Dorigo, M., Gambardella, L.M.: Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Trans. on Evolutionary Computation 1(1), 53–66 (1997)
11. Zhang, Z.-s., Sun, Y.-m., Zhang, S.-y.: Assessment on performance using two kinds of optimized fault-tolerance neural network in fault diagnosis models of transmission and distribution networks. Journal of Tianjin University 39(B06), 115–120 (2006)
12. Duan, H.-b.: Theory and Applications of ant colony algorithm. Science Publishing House, Beijing (2005)
13. Zhou, D.-h., Ye, Y.-z.: Modern Fault Diagnostics and Tolerance Control. Tsinghua University Press, Beijing (2000)

Extraction and Parameterization of Eye Contour from Monkey Face in Monocular Image

Dengyi Zhang[1], Chengzhang Qu[1], Jianhui Zhao[1,*], Zhong Zhang[1], Youwang Ke[1], Shizhong Han[1], Mingqi Qiao[2], and Huiyun Zhang[2]

[1] Computer School, Wuhan University, Wuhan, Hubei, China, 430079
jianhuizhao@whu.edu.cn
[2] Basic Medicinal College, Shandong University of Chinese Traditional Medicine,
Jinan, China, 250355

Abstract. This paper proposes one approach for eye contour extraction and parameterization from monkey face in the monocular image. The possible face regions are first segmented using facial skin color model based on YCbCr color space. The color model is constructed from many collected sample pixels of face regions, and Gaussian Mixture Model is used to represent the distribution of color points. Then the segmented regions are further processed with the help of mathematical morphology operations to locate the face region. The face is size normalized according to face width, and rotation normalized according to two located eyes. Taking the boundaries of eye as initial point set, ACM is adopted to search for more smooth and continuous eye contours. Points on the extracted eye contours are used to fit the quadratic curve with least-squares fitting method. Based on the parameters of quadratic function, angry and calm expressions of monkey face are recognized. Our method can be used for facial expression recognition and emotion decision, which is very helpful for medical science research often taking monkey as the experimental object.

Keywords: Color Model, Face Normalization, ACM, Least-Squares Fitting.

1 Introduction

Image based facial expression recognition has received more and more attentions from researchers in the world. As an interdisciplinary research problem, it has close relations with computer science, cognitive science, psychology, physiology, etc. It can be applied in the new generation of human-computer interface design, affective computing, automatic identification, and many other fields [1]. In traditional Chinese medicine, face observing is an important approach, i.e. studying the facial expressions and taking it into consideration for doctor's diagnosis. However, expression is very hard to be measured automatically. Until now, most of the published techniques for facial expressions are proposed for human being, while monkeys are often used as the experimental objects in medical science [2].

[*] Corresponding author.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 182–189.
springerlink.com

Emotion has relations with health and disease, thus it is very important to study the causes of different emotions and search for the suitable intervention approaches. In such research monkey can be used as it is close relative of human in biology, and it has similar responses to the same interventions. To capture the variations of facial expression in real time, images of monkey faces are recorded and then the frames are analyzed automatically by computer. Based on the image features on monkey face such as the contours of eye and mouth, facial expressions are detected and thus the corresponding possible emotion can be concluded accordingly.

Therefore an automatic approach is proposed in this paper to extract and then parameterize the eye contour of monkey face from monocular image, which is the prerequisite procedure for the recognition of monkey's emotion. Our method includes the following steps: face segmentation, face normalization, eye contour extraction and eye contour parameterization.

2 Face Segmentation

Facial skin color is often used for face segmentation, thus the color model is set up based on sample images [3-5]. As brightness affects the decision of skin color, YCbCr (YUV) color space is adopted because brightness value Y is independent with color value Cb and Cr. From the sample images, regions of monkey faces are manually marked, and then Y, Cb and Cr value of the pixels within face regions are used to statistically construct the color model. Fig. 1 shows about 388,000 color points from facial regions of sample monkeys, and Y value is within the range of [100, 200], Cb is within the range of [110, 140], Cr is within the range of [125, 150].

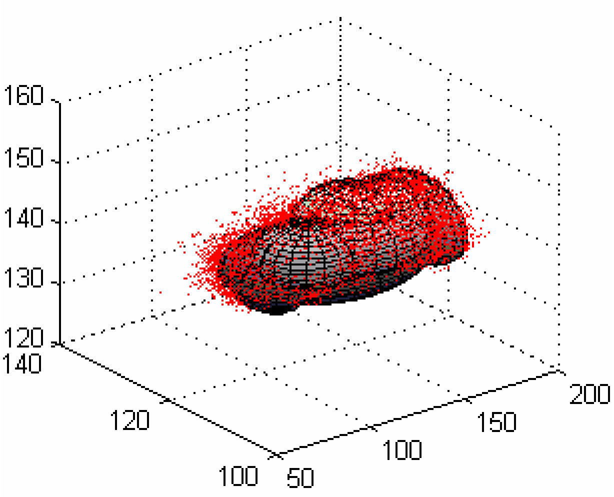


Fig. 1. Color model from sample pixels and 12 Guassian models

To define a more precise color model, 3D shape of the sample point cloud can be represented by Gaussian mixture model. First, we use expectation maximization (EM) algorithm to train the parameters of Gaussian mixture model [6-8]. Then, whether one pixel belongs to facial region of the monkey image can be decided by calculation of its probability with the following formula:

$$p(x)=\alpha_1 g(x;\mu_1,\sum\nolimits_1)+\alpha_2 g(x;\mu_2,\sum\nolimits_2)+\alpha_3 g(x;\mu_3,\sum\nolimits_3) \ . \qquad (1)$$

$$g(x;\mu,\sum)=\frac{1}{\sqrt{(2\pi)^d\left|\sum\right|}}\exp\left[-\frac{1}{2}(x-\mu)^T\sum\nolimits^{-1}(x-\mu)\right] . \qquad (2)$$

The weighting value, center and covariance matrix of each Gaussian model are α, μ and $\sum$ respectively. The probability p(x) shows how close of the point x to the facial region represented by the samples. The number of Gaussian mixture models can be manually assigned or automatically computed. We tried the automatic method and the calculated best number of Gaussian mixture models is 12. The 12 computed Gaussian models are used to approximate the density of sample points from facial regions in YCbCr color space. As illustrated in Fig. 1, one ellipsoid represents one Gaussian distribution function, the center of each ellipsoid is the average value of the corresponding function, while the focal length of each ellipsoid is the double square root of the diagonal data from the related covariance matrix.

Based on the obtained color model, each pixel of one image is checked and face region of the image can be automatically detected. As shown in Fig. 2, pictures in the 1st row are monocular images with monkey faces, pictures in the 2nd row are the segmented regions with the help of facial skin color model.

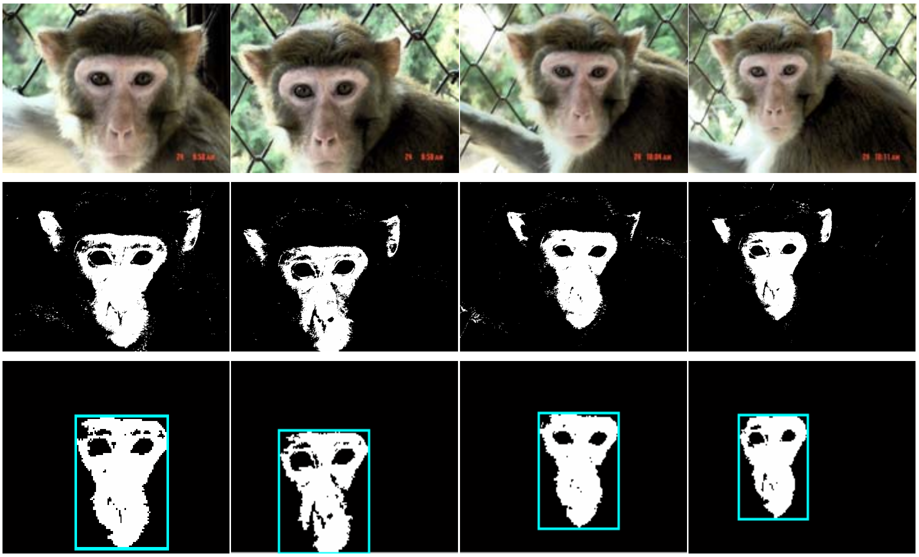


Fig. 2. Face segmentation using color model

The color model based method can segment the pixels with the colors in the space defined by the collected sample points, but at the same time it can obtain some unwanted regions with the same colors. For example, ears of monkey have the similar colors as face, and there may be scattered pixels in the ears' regions of the image happen to have facial skin color. Thus the segmented results are further processed by the following steps to refine the extracted monkey face.

Step 1. For each pixel with no skin color, if there are more than 4 pixels of its 3*3 neighbors having skin color, it is converted into the pixel with skin color, and such operation is used to reduce the affects caused by monkey fur on face.

Step 2. To strengthen the connectivity, the extracted face regions are processed by the dilation operation of the mathematical morphology.

Step 3. From the obtained possible face regions, find the continuous one with the largest area, and take it as the extracted monkey face.

Step 4. The extracted region of monkey face is processed by the erosion operation of the mathematical morphology, and then the face is marked with rectangle box, as shown in the 3rd row of Fig. 2.

3 Face Normalization

Monkey face may have many different postures, thus the image should be normalized before expression recognition. Face normalization includes both size and rotation normalization. Size normalization is performed based on the width of monkey face since the height of face may obviously change when the mouth is opened. After size normalization, the approximate region with two eyes can be located according to the basic structure of monkey face. Rotation normalization makes use of the linked line between the centers of two eyes. As shown in Fig. 3, (a) is the region of monkey's eyes, and the procedure for rotation normalization has the following steps:

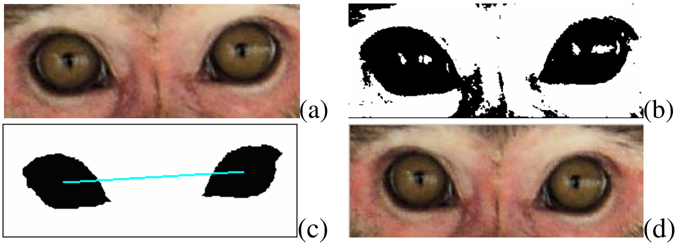


Fig. 3. Rotation normalization of monkey face

Step 1. Convert the color image (a) to gray image, and then take the gray value as the input of Otsu algorithm [9] to calculate the segmentation threshold automatically.

Step 2. Divide the gray image into two clusters with the threshold, and the segmented image is shown in (b) with clusters in black and white.

Step 3. Remove the small and discrete black regions, then the remained black areas are processed with erosion and dilation operations of mathematical morphology.

Step 4. Take the two large and continuous black areas as monkey eyes, and the center points of two eyes are linked by one line as shown in (c).

Step 5. Rotate the line around one of its end points to be horizontal, and the rotated image is the result of rotation normalization, as shown in (d).

4 Eye Contour Extraction

The boundary of segmented eyes in Fig. 3 is not smooth and continuous, thus not good enough to be taken as the contour of monkey eye. Therefore, points on the obtained boundary are used as the initial locations, and active contour model (i.e. snake model) [10-14] is adopted to move them to their better positions, and then the adjusted points construct the eye contour [15].

As shown in Fig. 4, the 1st row shows the pixels on the extracted eye boundary with initial positions, and the 2nd row shows their corresponding adjusted locations after ACM. From the experimental results it can be found that the obtained eye contour is relative smooth and continuous, which is the more precise presentation.

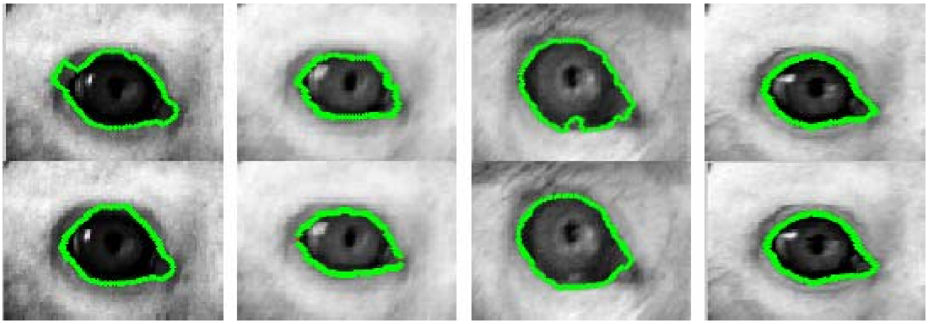
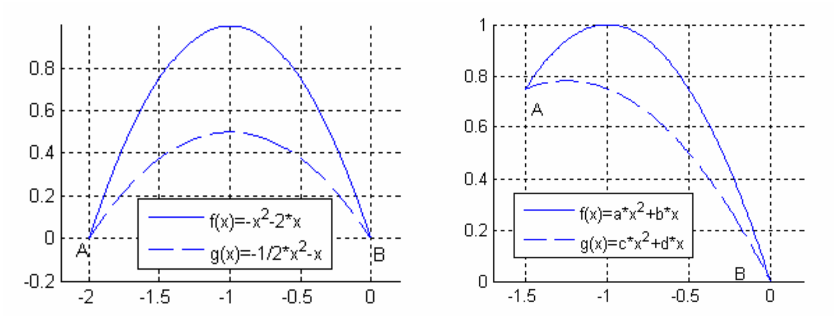


Fig. 4. Eye contour's approximation by ACM

5 Eye Contour Parameterization

Monkey eyes and human eyes perform similar in facial expressions, the length of the eyes shows little change in general, but the shape of two contours of eyelids changes obviously. Thus the contour characteristics are parameterized and analyzed to help determine whether there are some facial expressions such as "angry" and "calm". For these two kinds of facial expressions, the upper and lower eyelids opening to the largest can express the emotion of "angry", and the not apparently protruding eye shape can express the emotion of "calm". Therefore we use least-squares method to fit the quadratic curve for the pixels on the upper and lower eye contours. With the quadratic parameter, the wide opened and normal opened eyes are quantified and the related expressions are indicated. We call this quadratic parameter "curvature" of the eye, i.e. the curvature data is collected and analyzed in our experiment.

Different with human being, monkey cannot keep still in the experiment, thus it is almost impossible to capture a series of images of monkey faces with only changing eye regions. Therefore, we transform the two different images using corner points of eyes before comparing the curvatures of eyes. As shown in Fig. 5(a), the eye contours are rotated to make the line between corner points A and B be parallel. Then we take one as the basic image, based on which the other one is transformed by translation and scale. To be more convenient for the calculation of curve fitting, the eye contours are then translated to make the corner point B lie on the origin of the normal coordinate system. The solid curve represents the fitted function f(x) for the upper eyelid of the angry expression, while the dotted curve represents the fitted function g(x) for the upper eyelid of the calm expression. Obviously, the quadratic parameter's absolute value of f(x) is higher than that of g(x), thus the curvature of the eye contours can be used as the parameter for expression detection.



(a) (b)

Fig. 5. Quadratic curve fitted from upper eye contour

From the experiments we also find that even if the line between corner points A and B is not transformed to be parallel, curvature of the fitted quadratic curve from eye contours can still be used for facial expression. For two contours of the upper eyelids with their corner point B on the origin of the coordinate system, there is another intersection point A based on the fact that there is little change for the length of eye in different facial expressions, as shown in Fig. 5(b). Suppose the coordinates of point A is (m, n), m * n < 0, the solid fitted quadratic curve for the angry contour of the upper eyelid and the dotted fitted curve for the calm contour are

$$\begin{aligned} f(x) &= ax^2 + bx \\ g(x) &= cx^2 + dx \end{aligned} \quad . \tag{3}$$

Within the range of (m,0), f (x)>g (x), so

$$ax^2 + bx - cx^2 - dx > 0 \,,\ x \in (m,0) \ . \tag{4}$$

After transformation, there is

$$(a-c)x + b - d < 0 \,,\ x \in (m,0) \ . \tag{5}$$

For the intersection point A(m, n), there is always the equation f (m)=g (m), so

$$am^2 + bm = cm^2 + dm \ . \tag{6}$$

After transformation, there is

$$(a-c)m+b-d=0 \ . \tag{7}$$

With (7)-(5), there is

$$(a-c)m>(a-c)x, \ x\in(m,0) \ . \tag{8}$$

After transformation, there is

$$(a-c)(m-x)>0, \ x\in(m,0) \ . \tag{9}$$

Since x>m, we can conclude that

$$(a-c)<0 \ . \tag{10}$$

And finally there is

$$a<c \ . \tag{11}$$

From Fig. 5(b) we can find that both a and c are negative value, thus the relations between the absolute values of a and b are

$$fabs(a)>fabs(c) \ . \tag{12}$$

Therefore, absolute value of the quadratic parameter from angry contour is larger than that of calm contour, which proves that curvature can be used to describe facial expression even in the situation of Fig. 5(b).

6 Conclusion

Emotion is studied in medical science since it is related with health and disease. As the often used experimental object, monkey is studied and its emotion can be detected from facial expressions, during which eye contour is taken as one important image feature. Our paper works on the techniques for extraction and parameterization of monkey's eye contour from monocular image, and then uses them for the recognition of facial expressions. In our approach, facial skin color model is constructed from collected samples and represented by the Gaussian mixture models. Based on the color model, the regions with facial skin color are obtained, and the face region is segmented from them with the help of mathematical morphology. Monkey face is size normalized using face width and rotation normalized using the center points of two eyes. Taking the points on the boundary of eye as initial points, ACM method is adopted to adjust the points and obtain the more smooth and continuous contour for monkey eye. Then the contours of upper and lower eyelids are used to fit the quadratic curve with the least-squares method. The value of curvature, the quadratic parameter's absolute value, can be utilized for the decision of facial expressions, such as angry or calm expression. The experiments on monkey's contours have proved the efficiency of our method, which will be helpful for medical science research when monkey is taken as the experimental object.

In the future work, we will study the other image features related with the facial expression, such as mouth, nose, or the whole structure of the face. If the accuracy of the system for automatic facial expression recognition is high enough, it may provide an alternative way to avoid the low efficiency of manual observations used in the current research on emotions in the field of medical science.

Acknowledgments. This work was supported by a grant from the Major State Basic Research Development Program of China (973 Program) (No. 2006CB504804).

References

1. Pylyshyn, Z.W.: Computation and Cognition: Towards a Foundation for Cognitive Science. MIT Press, Cambridge (1984)
2. Elizabeth, S.P., Emma, J.H., Michael, M.: Measuring emotional processes in animals: the utility of a cognitive approach. Neuroscience and Biobehavioral Reviews 29, 469–491 (2005)
3. Jones, M.J., Rehg, J.M.: Statistical Color Models with Application to Skin Detection. International Journal of Computer Vision 46(1), 81–96 (2002)
4. Phung, S.L., Bouzerdoum, A., Chai, D.: Skin segmentation using color pixel classification: analysis and comparison. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(1), 148–154 (2005)
5. Belongie, S., Carson, C., Greenspan, H., et al.: Color-and texture-based image segmentation using EM and its application to content-based image retrieval. IEEE Int. Conf. Computer Vision 6(1), 465–473 (1998)
6. Yang, M.H., Kriegman, D.J., Ahuja, N.: Detecting Faces in Images: A Survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 34–58 (2002)
7. Park, J., Seo, J., An, D., et al.: Detection of Human faces using skin color and eyes. In: IEEE International Conference on Multimedia and Expo. (ICME), 30 July- 2 August, vol. 1, pp. 133–136 (2000)
8. Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A survey of skin-color modeling and detection methods. Pattern Recognition 40(3), 1106–1122 (2007)
9. Otsu, N.: A threshold selection method from gray-level histogram. IEEE Trans. Syst. Man Cybernet. 9, 62–66 (1979)
10. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. Int. Journal of Computer Vision 1(4), 321–331 (1987)
11. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models: Their training and application. CVGIP: Image Understanding 61, 38–59 (1995)
12. Davies, R.H., Twining, C.J., Allen, P.D., Cootes, T.F., Taylor, C.J.: Building optimal 2D Statistical Shape Models. Image and Vision Computing 21, 117–128 (2003)
13. Roberts, M.G., Cootes, T.F., Adams, J.E.: Robust Active Appearance Models with Iteratively Rescaled Kernels. In: Proc. British Machine Vision Conference, vol. 1, pp. 302–311 (2007)
14. Matthews, I., Baker, S.: Active Appearance Models. International Journal of Computer Vision 60(2), 135–164 (2004)
15. Zhang, D.Y., Qu, C.Z., Zhao, J.H., Zhang, Z., Ke, Y.W., Cai, B., Qiao, M.Q., Zhang, H.Y.: Eye Contour Extraction Method from Monocular Image with Monkey Face. In: Proceedings of International Symposium on Intelligent Information Technology Application, Shanghai, China, December 21-22, pp. 636–639 (2008)

KPCA and LS-SVM Prediction Model for Hydrogen Gas Concentration

Minqiang Pan[1], Dehuai Zeng[1,2], Gang Xu[2,3], and Tao Wang[1]

[1] School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou, China, 510640
[2] School of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, China
[3] Shenzhen Key laboratory of mould advanced manufacture, Shenzhen, China, 518060
mexmpan@126.com

Abstract. Hydrogen gas concentration forecasting and evaluation is very important for Bio-ethanol Steam Reforming hydrogen production. A lot of methods have been applied in the field of gas concentration forecasting including principal component analysis (PCA) and artificial neural network (ANN) etc. this paper used kernel principal component analysis (KPCA) as a preprocessor of Least Squares Support Vector Machine (LS-SVM) to extract the principal features of original data and employed the Particle Swarm Optimization (PSO) to optimize the free parameters of LS-SVM. Then LS-SVM is applied to proceed hydrogen gas concentration regression modeling. The experiment results show that KPCA-LSSVM features high learning speed, good approximation and generalization ability compared with SVM and PCA-SVM.

Keywords: KPCA, LS-SVM, prediction, hydrogen gas concentration.

1 Introduction

The need for alternative energy, which is clean without emission of pollutants and has high energy efficiency, gradually increases due to the exhaustion of fossil-fuel. Hydrogen is abundantly available in the universe and possesses the highest energy content per unit of weight compared to any of the known fuels, but today near to 95% of hydrogen is produced from fossil-based materials such as methane and naphtha [1-3]. In order to support hydrogen economy, renewable and clean energy source for hydrogen production is demanded. Bio-ethanol Steam Reforming (BESR) is considered as a promising route for Hydrogen production from renewable sources. Numerous studies have focused on the accurate prediction of hydrogen gas concentration produced by Bio-ethanol Steam Reforming method by means of statistical approaches and artificial intelligence approaches. However, neural network modeling will lead to the multiple local minima problem and the danger of over fitting.

Recently, a novel type of learning machine, called support vector machine (SVM), has been receiving increasing attention. SVM established on the unique theory of the structural risk minimization principle, usually achieves higher generalization performance than traditional neural networks that implement the empirical risk

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 190–197.
springerlink.com

minimization principle in solving many machine learning problems. Least squares support vector machine (LS-SVM) is employed to forecast hydrogen gas content. In the development of LS-SVM, all available indicators can be used as the inputs, but irrelevant or corrected features could adversely impact the generalization performance due to the curse of dimensionality problem [4~6]. Thus, it is critical to perform feature selection or feature extraction in SVM.

Principle component analysis (PCA) is a well-known method for feature extraction, which reduce the dimension of data under the information loss minimization by selecting maximal variance samples as the projected direction. PCA performs well on linear problems. But with nonlinear problem, PCA does not perform well. Kernel Principal Component Analysis (KPCA) can efficiently extract the nonlinear relationship between original inputs, which maps the original input data *x* into a high dimensional feature space *F* using a nonlinear mapping *Φ*, and then performs a linear PCA in the high dimensional feature space *F* [7, 8]. Therefore, we applied kernel-based feature extraction methods prior to learning in order to improve forecasting performance. First, nonlinear features were extracted by means of KPCA to evaluate its effect on a subsequent predictor in combination with learning algorithms such as LS-SVM. Furthermore, we examine whether the combination of KPCA and LS-SVM produce a superior forecasting performance than other forecasting systems.

The rest of this paper is organized as follows. In Section 2, KPCA for feature extraction is described. In Section 3, the theory of LS-SVM for regression estimation is presented. Section 4 shows the experimental results, and conclusion is drawn in the last section.

2 Kernel Principal Component Analysis (KPCA)

The mapping function is represented as follows:

$$\Phi : x_i \rightarrow \Phi(x_i) \in F \;\; i = 1, \dots l$$

Where Φ (x_i) is sample of *F* and $\sum_{i=1}^{l} \Phi(x_i) = 0$

The sample of covariance matrix *C* is formulated as:

$$C = \sum_{i=1}^{l} \Phi(x_i)\Phi(x_i)^T \tag{1}$$

Decomposing the *C* eigenvalue, (1) can be transformed to the eigenvalue problem

$$\lambda_i u_i = C_i u_i, \;\; i = 1, \dots, l \tag{2}$$

λ_i denotes one of the non-zero eigenvalue of C, u_i denotes the corresponding eigenvector of $\lambda_{i.}$

As all solutions lie in the span of Φ(x_i),...,Φ(x_m), we can consider the following equivalent equation:

$$\lambda(\Phi(x_i) \cdot U) = \Phi(x_i) \cdot CU \tag{3}$$

And denote *U* with expansion coefficients α_i as

$$U = \sum_{i=1}^{l} \alpha_i \Phi(x_i) \tag{4}$$

Combining Eq.(4) with Eqs.(1) and (3) and defining an *l*×*l* kernel matrix *K* yields

$$\lambda K\alpha_i = K^2\alpha_i, i = 1,2,...,l \tag{5}$$

By solving the following kernel eigenvalue problem, we solve Eq.(5)

$$\lambda\alpha_i = K\alpha_i \tag{6}$$

K is the *l*×*l* kernel matrix. The value of each element of *K* is equal to the inner product of two vectors x_i and x_j in the high-demensional feature space $\Phi(x_i)$ and $\Phi(x_j)$. That is K=$\Phi(x_i)\cdot\Phi(x_j)$. This means that the mapping of $\Phi(x_k)$ from x_k is implicit, any function satisfying Mercer's condition can be used as *K*.

Subsequently, we can calculate the orthonormal eigenvectors α_1, α_2, …, α_l of *K* corresponding to the L largest positive eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_l$. Furthermore, for assuring the eigenvectors of $\Phi(x_i)$ is the unit length $u_i \cdot u_i$=1, each α_i must be normalized using the corresponding eigenvalue by

$$\tilde{\alpha}_i = \frac{\alpha_i}{\sqrt{\lambda_i}}, i = 1,\ldots,l \tag{7}$$

Finally, based on the estimated $\tilde{\alpha}_i$, the principal components for input vector x_i is calculated by:

$$s_t(i) = u_i^T\Phi(x_t) = \sum\nolimits_{j=1}^{l}\tilde{\alpha}_i(j)\cdot K(x_j, x_t), i = 1,2,...l \tag{8}$$

The sample data in real do not always satisfy $\sum\nolimits_{k=1}^{l}\Phi(x_k) = 0$, in order to make the input samples centerlizing, the kernel matrix on the training set *K* and on the testing set K_t are respectively modified by:

$$\tilde{K} = \left(I - \frac{1}{l}1_l1_l^T\right)K\left(I - \frac{1}{l}1_l1_l^T\right) \tag{9}$$

$$\tilde{K}_t = \left(K_t - \frac{1}{l}1_l1_l^T K\right)\left(I - \frac{1}{l}1_l1_l^T\right) \tag{10}$$

where *I* is *l* dimensional identity matrix. l_t is the number of testing data points. 1_l and 1_{lt} represents the vectors whose elements are all ones, with length *l* and l_t respectively. K_t represents the $l \times l_t$ kernel matrix for the testing data points.

Next suppose that we select the first *d* principal components from the original feature space. As a criterion of selecting these components, the following accumulation ratio is often adopted:

$$A_c(d) = \sum\nolimits_{i=1}^{l}\lambda_i \Big/ \sum\nolimits_{i=1}^{n}\lambda_i \geq \theta \tag{11}$$

The accumulation ratio $A_c(d)$ shows how much information remains in the eigen-feature space after the *d* components are selected. The dimensionality *d* is selected such that the accumulation ratio for the *d* dimensional eigen-feature space is larger than a certain threshold *θ*.

3 LS-SVM Forecasting Model

Least Square Support Vector Machine (LS-SVM) is a new technique for regression. When LS-SVM is used to hydrogen gas concentration forecasting model, the input and output variables should be chosen firstly. Hence, this paper takes history data as the input. The minute hydrogen gas concentration is chosen as the model's output.

Given a training data set $\{(x_1,y_1),\ldots\ldots,(x_l,y_l)\}$ with input data $x_l \in R^n$ and output data $y_l \in R$. In order to get the function dependence relation, SVM map the input space into a high-dimension feature space and construct a linear regression in it. The regression function is expressed with

$$y = f(x) = w^T \varphi(x) + b \tag{12}$$

with $\varphi(\cdot): R^n \to R^{n_\varphi}$, a function which maps the input space into a so-called higher dimensional feature space, *w* and *b* are the regression parameters to be solved.

LS-SVM regression estimation involves primal and dual model formulations. Given the training data set $\{(x_1,y_1),\ldots\ldots,(x_l,y_l)\}$, the goal is to estimate the model, where *f* is parameterized as in Eq.12, we can formulate the following optimization scheme to infer our parameters

$$\begin{aligned} &\min_{w,b,e} L_P(w,e) = \frac{1}{2}\|w\|^2 + \frac{\gamma}{2}\sum_{i=1}^{l} e_i \\ &s.t.\ \ y_i = w\cdot\varphi(x_i) + b + e_i,\ \ i = 1,2,\ldots,l \end{aligned} \tag{13}$$

Where error variables $e = (e_1, e_2, \ldots, e_l)^T, e_i \in R$, the regularization constant $\gamma > 0$ is included to control the bias-variance trade-off. we perform the computations in another space, called the dual space of Lagrangian multipliers after applying Mercer's theorem. Consider the Lagrangian of Eq.13 given by

$$L_D(w,b,e_i,\alpha) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^{l} e_i - \sum_{i=1}^{l}\alpha_i(w^T\varphi(x_i) + b + e_i - y_i) \tag{14}$$

Here $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_l)^T, \alpha_i \in R$ are Lagrangian multipliers. The first order conditions for optimality are given by:

$$\begin{cases} \dfrac{\partial L_D}{\partial w} = 0 \to w = \sum\limits_{i=1}^{l}\alpha_i\varphi(x_i), \\ \dfrac{\partial L_D}{\partial b} = 0 \to 0 = \sum\limits_{i=1}^{l}\alpha_i, \\ \dfrac{\partial L_D}{\partial e_i} = 0 \to \alpha_i = \gamma e_i, i = 1,2,\ldots,l \\ \dfrac{\partial L_D}{\partial \alpha_i} = 0 \to y_t = w^T\varphi(x_t) + b + e_i, i = 1,2,\ldots,l \end{cases} \tag{15}$$

That is:

$$\begin{bmatrix} I & 0 & 0 & -Z \\ 0 & 0 & 0 & -\vec{1} \\ 0 & 0 & \gamma & -I \\ Z & \vec{1} & I & 0 \end{bmatrix} \begin{bmatrix} w \\ b \\ e \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ y \end{bmatrix} \tag{16}$$

Here

$Z = (\varphi(x_1), \varphi(x_2), ..., \varphi(x_l))^T$, $y = (y_1, y_2, \cdots, y_l)^T$ $I = (1,1,...,1)^T$， $e = (e_1, e_2, ..., e_l)^T$， $\alpha = (\alpha_1, \alpha_2, ..., \alpha_l)^T$

From Eq.15, combining the first and the last condition yields

$$y_i = \sum_{i=1}^{l} \alpha_i \varphi(x_i)^T \varphi(x_i) + b + e_i \tag{17}$$

Replacing the third expression from (15) into (17) gives

$$\begin{aligned} & y_i = \sum_{i=1}^{l} \alpha_i k(x_i, x_j) + b + \frac{\alpha_i}{\gamma},\ i = 1,2,...,l \\ & \sum_{i=1}^{l} \alpha_i = 0, \end{aligned} \tag{18}$$

that is:

$$\begin{bmatrix} \vec{1} & ZZ^T + \gamma^{-1} I \\ 0 & \vec{1}^T \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix} \tag{19}$$

Through selecting the appropriate kernel function, the nonlinear relation between the hydrogen gas concentration and its correlative influence parameters based on SVM is established.

Any function $k(x_i, x_j)$ satisfying Mercer's condition can be used as the kernel function. In this paper, Gaussian function is also selected as the kernel function, whose expression is shown as follows:

$$\varphi(x_i) \cdot \varphi(x_j) = k(x_i, x_j) \equiv \exp(\frac{\left\| x_i - x_j \right\|^2}{\sigma^2}) \tag{20}$$

where σ^2 is the width parameter of Gaussian kernel.

To replace the dot product $\varphi(x_i)^T \varphi(x_i)$, let $C = \Omega + \gamma^{-1} I$, then the Eq.19 can be written as:

$$\begin{bmatrix} \vec{1} & C \\ 0 & \vec{1}^T \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix} \tag{21}$$

From Eq.19, the regression parameters α and b can be solved as follows:

$$b = \frac{\vec{1}^T C^{-1} y}{\vec{1}^T C^{-1} \vec{1}} \tag{22}$$

$$\alpha = C^{-1}(y - b\vec{1}) \tag{23}$$

Thus, the regression function is expressed in dual form

$$y = w^T \varphi(x) + b = \sum_{i=1}^{l} \alpha_i k(x_i, x_j) + b \tag{24}$$

Kernel width parameter sigma σ and regularization parameter γmay affect LS-SVM generalization performance. In this paper, these parameters are automatically tuned using the PSO in the training phase.

4 KCPA-LSSVM Based Hydrogen Gas Concentration Forecasting

Hydrogen gas concentration is the dominate quality index used in the Bio-ethanol Steam Reforming based hydrogen production. Currently, no suitable on-line analysis exists to monitor the hydrogen gas concentration. Therefore, accurately forecasting the hydrogen gas concentration online is an urgent problem in the Bio-ethanol Steam Reforming based hydrogen production.

The basic idea is to use KPCA to extract feature and apply LS-SVM for forecasting Hydrogen gas concentration. Fig. 1 shows how the model is built.

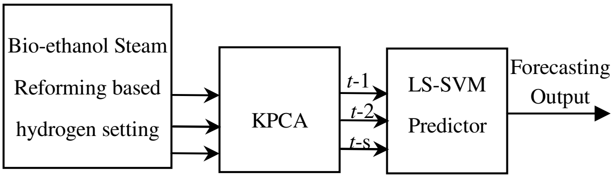


Fig. 1. KPCA-LSSVM based forecasting Model structure

We collected 480 groups of data of above 12 variables and corresponding hydrogen gas concentration from the production. The collected data were preprocessed by three-delta rule and sliding mean method and 442 groups of data remained. KPCA was used to extract the nonlinear principal components from the 442 groups of data. Then 330 groups are used as validating data set for finding the optimal arameters of LS-SVM, while the remaining 112 groups are used as testing data set for testing the predictive power of the model.

The detailed step of KCPA-LSSVM algorithm is illustrated in Fig.2:

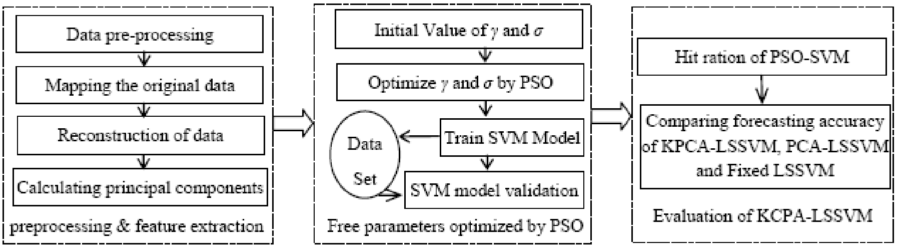

Data pre-processing
Mapping the original data
Reconstruction of data
Calculating principal components
preprocessing & feature extraction
Initial Value of γ and σ
Optimize γ and σ by PSO
Data Set
Train SVM Model
SVM model validation
Free parameters optimized by PSO
Hit ration of PSO-SVM
Comparing forecasting accuracy of KPCA-LSSVM, PCA-LSSVM and Fixed LSSVM
Evaluation of KCPA-LSSVM

Fig. 2. KCPA-LSSVM forecasting algorithm flow chart

① Applying kernel principal component analysis to reduce the noisy data that has interfered the performance of predictor and extract the principal characters of original data which provides the same quality of predictor as the set of all attributes.

② Implementing the Particle Swarm Optimization (PSO) to optimize the LS-SVM free parameters. Namely, the PSO tries to search the optimal values to enable SVM to fit various datasets.

③ Implementing the proposed KPCA-LSSVM to perform prediction task using these optimal parameters.

④ Compare the prediction accuracy of KPCA-LSSVM, PCA-LSSVM and Fixed LSSVM.

To validate the effect of the model based on KPCA-LSSVM, the model based on PCA-LSSVM and LSSVM were also established. Root Mean Square Error (RMSE) and and the normalized mean square error (*NMSE*) were used as performance indexes.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{25}$$

$$NMSE = \frac{1}{\delta^2 n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \delta^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{26}$$

Where n represents the total number of data points in the test set, $y_i, \hat{y}_i, \bar{y}_i$ are the actual value, prediction value and the mean of the actual values respectively.

The predicting performance of three models is summarized in table 1. The experiments results show that the predictive error of KPCA-LSSVM is the smallest. LSSVM by feature extraction using KPCA performs much better than that without feature extraction. And there is also superior performance in the KPCA than the PCA. KPCA-LSSVM features high learning speed, good approximation and generalization ability compared with SVM and PCA-SVM. The model based on KPCA-LSSVM may efficiently guide production.

Table 1. Comparision of different predictor

Model	PCs	RMSE	NMSE
KPCA+LSSVM	28	0.0142	0.0348
PCA+LSSVM	21	0.0368	0.0725
LSSVM	—	0.1654	0.2531

5 Conclusions

This paper describes a novel methodology, a LS-SVM based on combining KPCA to model and forecast hydrogen gas concentration. By employing kernel principal component analysis to attain the principal hydrogen gas production features and Particle Swarm Optimization algorithm to optimize the free parameters of Support Vector Machine, the proposed PSO-SVM model is applied to to predict the hydrogen gas concentration with Bio-ethanol Steam Reforming. The theoretical analysis and the simulation results show that KPCA can efficiently extract the nonlinear feature of initial data. KPCA-LSSVM has powerful learning ability, good generalization ability and low dependency on sample data. The contribution of this study demonstrate that the proposed model performed well when applied in the holdout sample, revealing the generalization of this model to forecast hydrogen gas concentration with Bio-ethanol Steam Reforming in renewable and clean energy source field.

Acknowledgment

This work is supported by the National Nature Science Foundation of China with grant #50805052 and Shenzhen Scientific Program in 2009.

References

[1] Vizcaino, A.J., Carrero, A., Calles, J.A.: Hydrogen production by ethanol steam reforming over Cu–Ni supported catalysts. Int. J. Hydrogen Energy 32, 1450–1461 (2007)

[2] Haryanto, A., Fernando, S., Murali, N., Adhikari, S.: Current status of hydrogen production techniques by steam reforming of ethanol: a review. Energy Fuels 19, 2098–2106 (2005)

[3] Yu, C.-Y., Lee, D.-W., Park, S.-J., Lee, K.-Y., Lee, K.-H.: Study on a catalytic membrane reactor for hydrogen production from ethanol steam reforming. Int. J. Hydrogen Energy 34(7), 2947–2954 (2009)

[4] Suykens, J.A.K., Van Gestel, T., De Brabanter, J., DeMoor, B., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)

[5] Lai, K.K., Yu, L., Wang, S.Y., Zhou, L.G.: Credit risk evaluation with least square support vector machine. In: Wang, G.-Y., Peters, J.F., Skowron, A., Yao, Y. (eds.) RSKT 2006. LNCS (LNAI), vol. 4062, pp. 490–495. Springer, Heidelberg (2006)

[6] Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Processing Letters 9(3), 293–300 (1999)

[7] Li, Y., Wang, Z.: An Intrusion Detection Method Based on SVM and KPCA. In: Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition, Beijing, China, pp. 1462–1466 (2007)

[8] Takeuchi, Y., Ozawa, S., Abe, S.: An Efficient Incremental Kernel Principal Component Analysis for Online Feature Selection. In: Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA (2007)

Random Number Generator Based on Hopfield Neural Network and SHA-2 (512)

Yuhua Wang[1], Guoyin Wang[2], and Huanguo Zhang[3]

[1] College of Information Science and technology, Henan University of Technology,
45001, Zhengzhou, China
yuhua.w@tom.com
[2] Henan Electric Power Survey & Design Institute,
450007, Zhengzhou, China
wanggyzz@163.com
[3] School of Computer Science, Wuhan University,
430079, Wuhan, China
Liss@whu.edu.cn

Abstract. With the rapid development of cryptography and network communication, random number is becoming more and more important in secure data communication. The nonlinearity of backward propagation neural network (BPNN) is used to improve the traditional random number generator (RNG). SHA-2 (512) hash function can ensure the unpredictability of the produced random numbers. So, a novel and secure RNG architecture is proposed in the presented paper, which is BPNN based on SHA-2 (512) hash function. The quality of random number generated by this proposed architecture can well satisfy the security of cryptographic system according to results of test suites standardized by the U.S. The proposed architecture can be used to improve performances such as power consumption, flexibility, cost and area in network security and security for cryptographic systems.

Keywords: Random number, Hopfield Neural Network, Security, SHA-2 (512).

1 Introduction

The wide use of computer networks and wireless devices of digital communications results in a greater need for the protection of transmitted information by using cryptography. The security of cryptographic systems depends on the unpredictable and irreproducible digital key streams generated by random number generator (RNG) [1-3]. Random numbers are also used for key management and authentication protocols in cryptographic system and smart cards [4]. Moreover, RNG can defeat traceability attack in RFID Systems [5] Random numbers of high quality becomes more and more significant in the security of communication. Generation of random number is an important primitive in many cryptographic mechanisms. Unpredictability of random sequence is the foundation of security in many cryptographic systems [6]. Generators suitable for cryptographic applications need to meet stronger requirements than for other applications [7]. Random number generation presents challenging issues.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 198–205.
springerlink.com

A novel PRNG based on Hopfield Neural Network techniques and SHA-2 (512) is proposed in this paper. It is well known that Hopfield Neural Network posses very interesting function approximation capabilities making them a very powerful tool in many scientific disciplines [8]. Their most important and intriguing property are their generalization capabilities. This operation is a kind of nonlinear operation, which is needed in generation of random number. The outputs of network will be unpredictable when the search process doesn't converge. The problem of convergence is considered of the most important theoretical issues in the study of neural network. In the general problem of optimization, the convergence of search is the objective of solving problem [9]. In the proposed architecture, we make the process not convergence. Hopfield recurrent Neural Network exploits this to produce random numbers.

However, some statistic properties of random numbers outputted by Hopfield neural network may not be good enough for the data security. SHA-2 (512) is suitable to improve the statistical properties of random number. In the proposed architecture, the raw bit stream is produced by Hopfield neural network RNG, SHA-2 (512) removes non-idealities of output bits from Hopfield neural network RNG. The SHA-2 (512) enhances the unpredictability of random numbers by highly compressing the raw bits. The word length of produced random number is 512 bits, which is acceptable in all the related applications. So the Hopfield neural network RNG is combined with SHA-2 (512) to improve the security of the proposed architecture. A VLSI implementation for FPGA device of the proposed architecture is described.

2 Architecture of Random Number Generator Based on Hopfield Neural Network

The proposed architecture is shown in Fig. 1. The system includes two main parts: the pseudo Hopfield neural network RNG and SHA-2 (512) hash function. Pseudo Hopfield neural network RNG produces raw random bits. SHA-2 (512) is chosen to be a well-designed decorrelating algorithm in the system. SHA-2 (512) hash function can remove the non-idealities of the raw sequence through high compression ratio.

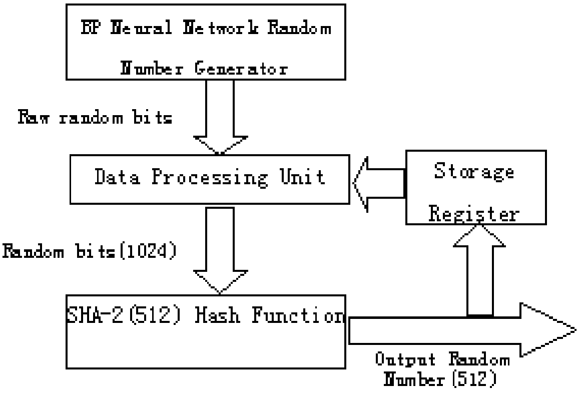


Fig. 1. Architecture of Random Number Generator Based on Hopfield Neural Networking

3 Hopfield Neural Network

Since Hopfield introduce in 1982 and 1984, the Hopfield neural networks have been used in many different applications. The important property of the Hopfield neural network is decrease in energy by finite amount whenever there is a change in puts [10]. Thus, the Hopfield neural network can be used for optimization. However, during the optimization, the output of network will be unpredictable when the process of search doesn't convergence [11]. Hopfield neural network is shown in Fig 2.

Each neuron i is connected to other neurons j through connection T_{ij}. The matrix is called the weight matrix of the network. The output, also called state, of the neuron i is V_i. The outputs evolve during time according to the following equations, where N is the number of neurons:

$\forall \quad i \in [1, N]$

$$dU_i(t)/dt = \sum_{j=1}^{N} T_{ij} V_j(t) - I_i \tag{1}$$

$$V_i(t) = g(U_i(t)) = \tanh(b \cdot U_i(t)) \tag{2}$$

At time t, neuron i receives signals $V_j(t)$ from other neurons j. It performs a weighted sum of these entries to calculate its internal potential $U_i(t)$, which can be viewed as an intermediate calculus variable. Then a nonlinear sigmoid function, defined by $g(x) = \tanh(b.x)$, is applied to produce new output $V_j(t)$. Hopfield introduced a global measure of the network dynamics he called energy function E of the net work:

$$E = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} T_{ij} V_i V_j + \sum_{i=1}^{N} I_i V_i \tag{3}$$

We have the remarkable theorem: If the matrix T is symmetric, the Hopfield network converge with time toward the minimum of energy E. If the matrix doesn't depend on V_i values, we have:

$$\partial E / \partial V_i = -\sum_{j=1}^{N} T_{ij} V_j(t) \tag{4}$$

$$dU_i(t)/dt = -\partial E / \partial V_i \tag{5}$$

Equation (5) shows the fundamental property of the network's dynamics: it consists of a parallel descent of energy E 's gradient. The formula (5) represents how the network minimizes the function.

A Hopfield network minimizes a cost function involving their weights and neuron activations under their weight matrix, which realize the process of optimization. In fact, the convergence of conditions is closely related to network architecture, network initial conditions as well as updating rule modes. To realize optimization is to obtain

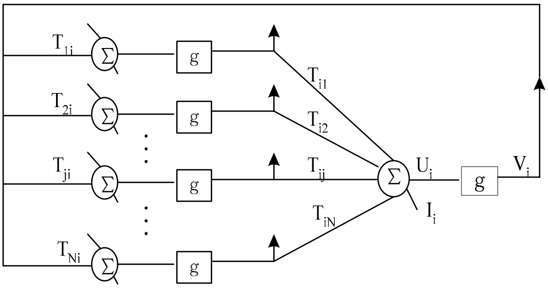


Fig. 2. Hopfield neural network overall architecture

convergence. However, these conditions will degrade in the design of random number generator. The important characteristics of a Hopfield network are shown as follow:

a) If the weight matrix of a Hopfield neural network is symmetric with zero valued diagonals and only one neuron is activated per iteration of the recurrent recall scheme, then there exists a Liapunov type cost function involving its weights and neuron activations, which decreases after each iteration until a local optimum of this objective function is found.

b) The final output vector of the Hopfield network, after the convergence of the above mentioned recurrent recall scheme, has minimum distance or is exactly equal to one prototype stored in the network during its weight matrix definition (learning phase) provided that the prototypes stored are orthogonal to one another and their number $M \le 0.15N$, where N is the number of neuron in the network.

c) if the prototypes stored in the Hopfield network are not orthogonal or their number $M > 0.15N$, then the recurrent recall scheme converges to a linear combination of the prototypes stored when it is fed with a variation of one of these prototype vectors, provided that the weight has the properties discussed in (a) above.

d) Hopfield network outputs are given by the following formula, which is precisely the update formula for the single neuron activated during the iteration of recurrent recall scheme mentioned in (a) above. The formula $g = \tanh()$ is nonlinear.

$$O_k = g(U_i(t)) = \tanh(b \cdot U_i(t)) \tag{6}$$

These properties lead us intuitively to the principles of the proposed random number generation methodology involving such recurrent neural network, summarized as follows.

1) If we impose a perturbation to the recurrent network weight matrix so that its symmetry is broken and its diagonal units obtain large positive values, then the convergence property of the recurrent recall scheme will be lost. This can be achieved, for instance, by adding a positive parameters δ to every unit in the upper triangle of the matrix, including diagonal units, and adding the negative quantity $-\delta$ from every unit in the lower triangle of the matrix.

2) Moreover, if we let a large number of neurons (in our experiments $N/2$ neurons) update their activations by following the formula of (d) above, then the recurrent recall scheme will loose its convergence property to a local optimum of suitable Liapunov function associated to the network.
3) If the recurrent recall scheme is not guaranteed to converge to a network output that corresponds to the local optima of a cost function, then the behavior of the network becomes unpredictable.
4) If the network is large and the patterns stored in it are orthogonal and thus uncorrelated (that is, they have maximum distance from one another), the possibility of obtaining predictable outputs after several iteration of the recurrent recall scheme is minimum compared to the one associated with storing non-orthogonal prototypes, which are correlated to one another. In our experiments we use binary valued orthogonal patterns.
5) If the history of the network outputs during its recall phase is considered for L iterations of the recurrent recall scheme, predicting the sequence of these output vectors is much harder than trying to predict a single output vector.

The above principles lead us to use the following function of network outputs over L iterations of the recurrent recall scheme as a pseudorandom number generator. To obtain better quality pseudorandom number, we have considered the Unix-function mod f , which outcomes the non-integral part of a real number, as the required mechanism for aiding Hopfield net output to acquire the desired properties, since the first digits of its decimal part are predictable, due to the fact that the sigmoidal nonlinearity g is a mapping on the (0,1) interval. Consequently, the formula of the Hopfield neural network proposed random number generator is as follows:

$$O = \mathrm{mod} f(1000 * (1/LN) \sum_{i=1\cdots T} \sum_{k=1\cdots N} (g(\sum W_{ki} O_i(t))^2) \tag{7}$$

The above discussion determines all the steps of the approach adopted here for designing pseudo random number generators employing recurrent recall scheme of Hopfield neural networks.

4 SHA-2 Hash Function

Hash function (one-way function) is a computationally feasible function, which converts binary strings of arbitrary length into binary strings of fixed length (hash-value). Hash function is very important in modern cryptography because of the property of one-way. On the one hand, hash function can output the hash value with fixed length for input with arbitrary length, and on the other hand, it's computationally infeasible to restore the input value for a given hash value. So, hash function can insure the security of one cryptographic system with sound unpredictability.

The SHA-1 hash function was designed by NIST as the digital signature standard. Three new hash functions SHA-2 (256,384,512) are used wildly. The basic principle of SHA-1 hash function is to produce the 160 bits message digest for inputted any message by dividing message into 512 bits blocks. SHA-1 hash function is good to design a better RNG. But with the higher security need, this design is not enough. So the higher security hash function is needed urgent. The hash function of SHA-2 (512) can provides the best security for inputted data. So in the proposed system, SHA-2 (512) is employed to remove the effect of non-idealities and improve the statistical properties of the outputted random numbers.

The hash function architecture is shown in Fig.3. The PADDER pads the input data to 1024-bit blocks to keep the equal input data block length. The padded message is generated with the following process.

The algorithm basic transformation round as the specification of the standard definition has to be processed 80 times. The 80×64-bit ROM Blocks are used for predefined K_i constants support. In the last phrase, the modification function processes the data and outputs them.

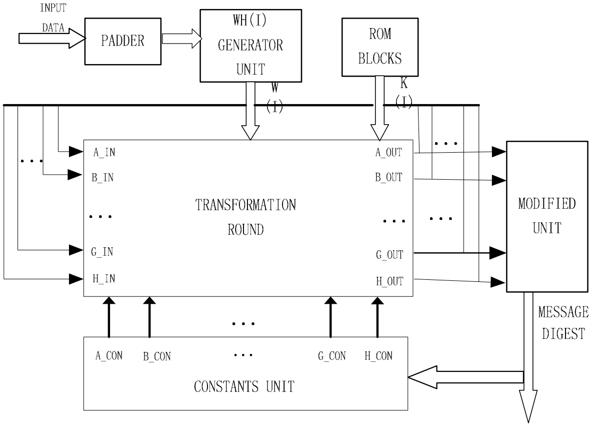


Fig. 3. SHA-2 (512) Hash Function Architecture

5 Experimental Results and Conclusion

The quality of random number is the key to security communication. In the proposed architecture, Hopfield neural network produces the raw random bits and SHA-2 (512) hash function in this proposed system can ensure the unpredictability of the outputted random numbers. The security of system is improved by SHA-2 (512). Two tests Standard, FIPS140-1 and SP800-22 are adopted to test the statistical qualities of random bit sequences. All the tests are performed on a 10^6bits long sequence. After these tests, the bit sequence of only pseudo Hopfield neural network RNG and the bit sequence processed by SHA-2 (512) can completely pass FIPS140-1 test. They are only different in the test results by SP800-22 test.

Table 1 shows that the random bits have better statistical properties than the results only applying pseudo Hopfield neural network RNG [11]. That is to say, the random bit sequence after SHA-2 (512) has better statistical properties than only outputted bits by pseudo Hopfield neural network RNG. Additionally, statistical experimental tests show that the efficiency of this system is obviously better than the results in the system only applying pseudo Hopfield neural network RNG.

Table 1. Test results after SHA-2(512) by SP800-22 test

	P-value Low	P-value High	P-value Avg	Pass Ratio
Frequency	0.770	0.932	0.861	0.986
Block-Frequency	0.801	0.856	0.842	1.000
Cusum-Forward	0.889	0.939	0.924	0.976
Cusum-Reverse	0.585	0.656	0.618	0.990
Runs	0.360	0.372	0. 366	0.954
Long Runs of ones	0.202	0.406	0.351	0.874
Rank(32×32)	0.845	0.896	0.859	0.979
Spectral DFT	0.522	0.625	0.570	0.928
Non-overlapping	0.602	0.747	0.677	0.975
Overlapping	0.188	0.196	0.188	0.946
Universal	0.522	0.657	0.593	0.976
Approx Entropy	0.617	0.769	0.733	0.965
Lempel-ziv	0.742	0.856	0.790	0.946
Linear Complexity	0.179	0.453	0.401	0.957
Serial (m=5)	0.702	0.779	0.753	0.993

For the proposed architecture, there will be more research to be done, and the system can be modified for other applications. For example, other better RNGs can be replaced for the pseudo Hopfield neural network RNG to improve the random numbers, and other better algorithms can be employed to process the raw random bits of pseudo Hopfield neural network RNG to improve the statistical properties of the random sequence. The proposed architecture can be used to improve the performances such as power supplying, flexibility, cost and area in network communication and improve the security of cryptographic system.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China under Grant No. 60373087, 69973034, 60673071.

References

[1] Huichin, T.: Modulus of Linear Congruential Random Number Generator. J. Quality & Quantity 39(4), 413–422 (2005)

[2] Callegari, S., Rovatti, R., Setti, G.: Embeddable ADC-based True Random Number Generator for Cryptographic Applications Exploiting Nonlinear Signal Processing and Chaos. J. IEEE Transactions on Signal Processing 53(7), 793–805 (2005)

[3] Katz, O., Ramon, D.A., Wagner, I.A.: A Robust Random Number Generator Based on a Differential Current-Mode Chaos. J. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 16(12), 1677–1686 (2008)

[4] Topaloglu, U., Bayrak, C., Iqbal, K.: A Pseudo Random Number Generator in Mobile Agent Interactions. In: IEEE International Conference on Engineering of Intelligent Systems, pp. 1–5. IEEE Press, Los Alamitos (2006)

[5] Zhang, H., Liu, Y.: Introduction of Cryptography. Wuhan University Press, Wuhan (2003)

[6] Zhaoyu, L., Dichao, P.: True Random Number Generator in RFID Systems against Traceability. In: IEEE Consumer Communications and Networking Conference (CCNC 2006), Las Vegas, NV, pp. 156–161 (2006)

[7] Tokunaga, C., Blaauw, D., Mudge, T.: True Random Number Generator With a Metastability-Based Quality Control. J. Solid-State Circuits, IEEE Journal 43(1), 78–85 (2008)

[8] Shouhong, W., Hai, W.: Password Authentication Using Hopfield Neural Networks. J. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions 38(2), 265–268 (2008)

[9] Li, H., Chen, B., Zhou, Q., Qian, W.: Robust Stability for Uncertain Delayed Fuzzy Hopfield Neural Networks With Markovian Jumping Parameters. J. Systems, Man, and Cybernetics, Part B, IEEE Transactions 39(1), 94–102 (2009)

[10] Shaoshuai, M., Huijun, G., Lam, J., Wenyi, Q.: A New Criterion of Delay-Dependent Asymptotic Stability for Hopfield Neural Networks With Time Delay. J. Neural Networks, IEEE Transactions 19(3), 532–535 (2008)

[11] Yuhua, W., Zhidong, S., Huanguo, Z.: Pseudo Radom Number Generator Based on Hopfield neural network. In: Proceedings of the Fifth International Conference on Machine Learning and Cybernetics(ICMLC 2006), Haerbing, pp. 2810–2813 (2006)

A Hybrid Inspection Method for Surface Defect Classification

Mengxin Li[1], Chengdong Wu[2], and Yang Cao[1]

[1] Faculty of Information and Control Engineering, Shenyang Jianzhu University,
Shenyang, China, 110168
limengxinf1972@yahoo.com.cn
[2] Northeastern University, Shenyang, China, 110004
wuchengdong@neu.edu.cn

Abstract. A vision-based inspection method based on rough set theory, fuzzy set and BP algorithm is presented. The rough set method is used to remove redundant features for its data analysis and procession ability. The reduced data is fuzzified to represent the feature data in a more suitable form as input data of a BP network classifier. The classifier is optimised using uniform design. By the experimental research, the hybrid method shows good classification accuracy and short running time, which are better than the results using BP network and neural network with fuzzy input.

Keywords: Vision-Based Inspection, Classification, Fuzzy, Rough Set, Neural Network, Uniform design.

1 Introduction

In the wood mill, the veneer sheets are placed on a conveyor which runs at a speed of 2.2m/s and the sheets appear at one to tow second intervals for human inspection. This task is extremely stressful and a little disturbance or loss of attention will result in a misclassification. Huber et al [1] made a series of experiments and found an accuracy of 68% with human inspection of boards. Similar experiments carried out by Polzleitner and Schwingshakl [2] indicated an accuracy of 55%. It is imperative to develop an automatic visual inspection system to relieve the human inspector and improve the classification accuracy, thus improving the productivity and profitability of the plywood factory.

Various pattern recognition methods have been researched during the last four decades, such as statistical decision theory, syntax and neural networks. Hybrid methods are of great interest due to their proven adaptability and advantages. This research presents a method using fuzzy rough neural network method. The rough set method is proposed to remove redundant features for its data analysis and processing. The reduced data is fuzzified to represent the feature data in a more suitable form for input to a BP network classifier. The BP neural classifier is considered the most popular, effective and easy-to-learn model for complex, multi-layered network.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 206–213.
springerlink.com

2 Rough Set for Feature Selection of Wood Veneer Defect

2.1 Basic Concepts

The following terms or concepts are introduced in order to facilitate and understand the proposed algorithm.

Information System: It is assumed that the given set of training samples represents the knowledge about the domain. In the approach described here, the training set is described by a classification system. The objects in a universe U are described by a set of attribute values.

Formally, an information system S is a quadruple $<U, A, V, f>$, where $U=\{x_1, x_2, \ldots x_N\}$ is a finite set of objects, which in this case are states of the environment; A is a finite set of attributes and the attributes in A are further classified into two disjoint subsets, condition attributes C and decision attributes D, such that $A=C\cup D$ and $C\cap D=\varnothing$; $V=\bigcup_{a\in A} V_a$ is a set of attribute values and V_a is the domain of attribute a (the set of values of attribute a); and f: $U\times A\rightarrow V$ is an information function which assigns particular values from domains of attributes to objects such that $f\ (x_i, a)\in V_a$ for all $x_i\in U$ and $a\in A$.

Decision Table: An information system can be designed as a decision table if the attribute set is divided into two disjoint sets - condition attribute set C and decision attribute set D, and $C, D\subset A$.

Indiscernibility Relation: For every set of attributes $B\subset A$, an indiscernibility relation IND(B) is defined in the following way: two objects, x_i and x_j, are indiscernible by the set of attributes B in A, if $\mathrm{b}(x_i)=\mathrm{b}(x_j)$ for every $b\subset B$. The equivalence class of IND(B) is called the elementary set in B because it represents the smallest discernible group of objects. For any element x of U, the equivalence class of x_i in relation IND(B) is represented as $[x_i]_{\mathrm{IND}(B).}$ The equivalent class of relation IND(C) and relation IND(D) are called condition class and decision class respectively for condition attribute set C and decision attribute set D.

Lower and Upper Approximation: Let X denote a subset of elements of the universe U ($X\subset U$). The lower approximation of X in B ($B\subseteq A$), denoted as $\underline{BX}$, is defined as the union of all these elementary sets which are contained in X. More formally:

$$\underline{BX}=\{x_i\in U|\ [x_i]_{\mathrm{IND}(B)}\subset X\neq 0\} \tag{1}$$

The above statement is to be read as: the lower approximation of the set X is a set of objects x_i, which belongs to the elementary sets contained in X. The upper approximation of the set X, denoted as BX, is the union of these elementary sets, which have a non-empty intersection with X:

$$BX=\{x_i\in U|\ [x_i]_{IND(B)}\cap X\neq 0\} \tag{2}$$

For any object x_i of the lower approximation of X, it is certain that it belongs to X. For any object x_i of the upper approximation of X, it may belong to X.

$BNX=BX-\underline{BX}$ is called a boundary of X in U.

CORE: The set of all attributes indispensable in *C* is denoted by CORE(*C*).

$$\mathrm{CORE}(C) = \cap \mathrm{RED}(C) \tag{3}$$

where RED(*C*) is the set of all reducts of *C*.

Accuracy of Approximation: An accuracy measure of the set *X* in *B*⊆*A* is defined as

$$\mu_B(X) = card(\underline{BX}) / card(BX) \tag{4}$$

The cardinality of a set is the number of objects contained in the lower (upper) approximation of the set *X*, $0 \le \mu_B(X) \le 1$ This round of checking takes place about two weeks after the files have been sent to the Editorial by the Contact Volume Editor, i.e., roughly seven weeks before the start of the conference for conference proceedings, or seven weeks before the volume leaves the printer's, for post-proceedings. If SPS does not receive a reply from a particular contact author, within the timeframe given, then it is presumed that the author has found no errors in the paper. The tight publication schedule of LNCS does not allow SPS to send reminders or search for alternative email addresses on the Internet.

In some cases, it is the Contact Volume Editor that checks all the pdfs. In such cases, the authors are not involved in the checking phase.

2.2 A Rough Sets Feature Selection Method

For an information system for wood veneer classification, there are 17 condition attributes including mean grey level, mode grey level, median grey level, standard deviation, skewness, kurtosis, lower number of pixels, higher number of pixels, lower grey level, higher grey level, dark grey level, bright grey level, number of edge pixels (threshold=μ), number of edge pixels (threshold=μ-2δ), number of edge pixel for feature 14, number of edge pixels (threshold=μ+2δ), number of edge pixel for feature 16, and 1 decision attribute expressed by 1-13 that means 13 defects including holes, pin knots, rotten knots, roughness, splits, strips, discoloration, coloured strips, barks, worms holes, curly grain, clear wood and sound knots. The task is to find out the optimal attributes and acquire decision rules.

2.2.1 Data Discretization

It is necessary to process the attribute values with a discretization algorithm to express and simplify the decision table. Clustering is a useful tool for analysing the structure of attribute spaces, and deducing similarity and dissimilarity among the observations. In terms of its high dimensionality, discovering clusters of arbitrary shapes, and dealing with different types of attributes, a hierarchical clustering method is presented [3] and adopted for data discretization before attribute reduction. This method has advantages such as embedded flexibility regarding the level of granularity and ease of handling any forms of similarity or distances. It can be divided into agglomerative and divisive algorithms. The agglomerative algorithm usually produces a sequence of clustering schemes of a decreasing number of clusters at each step, which results from merging the two closest clusters. The agglomeration schedule can be visualised by a dendrogram that shows which samples are combined. Nevertheless, it is not certain how many clusters are in the data for a group can be merged into different clusters. So the number of clusters further comes from the idea about statistics such as PST2, PSF, CCC which can judge how many classes should be suitable. Chosen clusters should make the number of clusters as small as it could be. All the optimal values constitute the information system.

The 17 defect features of wood veneers, used as condition attributes $C=\{X_1, X_2 \ldots X_{17}\}$, are discretized using hierarchical clustering and the values of decision attributes are expressed by 1-13.

2.2.2 Attribute Reduction

The aim of attribute reduction is to find a minimal subset of related attributes that preserves the classification accuracy of the original attributes of the decision table. It is therefore necessary to identify important attributes. There are many reducts, but in most cases it is not necessary to find all the reducts. The reduct with the least number of combinations of its attributes is selected [4]. In this research, 17 condition attributes are reduced according to the consistency principle in the following steps.

Step 1: The repetitive samples in the decision table are merged.

Step 2: The data in the decision table is further processed through attribute reduction based on the consistency principle. Important attribute sets thus remain.

The principle of rough sets for reducing redundant attributes can be expressed as the following:

Supposing $C=\{X_1, X_2, \ldots, X_n\}$ is an attribute set, if $\mathrm{POS}_C(D) = \mathrm{POS}_{(C-\{Xi\})}(D)$, then X_i in C is omissible or superfluous; otherwise attribute X_i in C is indispensable.

This has been implemented, as shown in Fig. 1. The reduct of condition attributes determines whether there are different decision values when the attribute values are the same in the decision table. Based on consistency principle, if an attribute set is removed and harmony still remains unchanged, the attribute set is removable. Among the 17 attributes of wood veneers, {c}, {g}, {i}, {l}, {m}, {q} are omissible after the attribute reduction, and the remaining attributes are sufficient. Mode grey level {c} and Kutosis {g} are insensitive to defining the changes of pixels. Because only splits and holes belong to white defects among the 13 defects, {i}, {m} and {q} are considered omissible. Dark grey level {l} contributes little to classification decision. After the attribute reduction, the optimum attributes are combined to obtain the decision rules for classification.

3 A Neural Network with Fuzzy Input for Inspection

3.1 Fuzzifier

A fuzzy set [4]can be represented as membership function μ_A that associates with each element x of the universe of disclosure X, a number $\mu_A(x)$, i.e. membership grade, in the interval [0, 1]. In particular, μ_A: $A \rightarrow [0, 1]$, where set A can also be treated as a subset of X. The main function of the fuzzifier maps a crisp input point $x \in X$ into a fuzzified value in $A \in U$ (the universe). There are two types of fuzzifier:

- Singleton fuzzifier: fuzzy set A with support x_i, where $\mu_A(x_i)=1$ for $x= x_i$ and $\mu_A(x_i)=0$ for $x \neq x_i$, for which the input measurement x is perfect crisp, $i=1,2,....n$.
- Non-singleton fuzzifier: $\mu_A(x_i)$ reaches maximum value 1 at $x= x_i$ and decreases from 1 to 0 while moving away from $x= x_i$.

The determination of the fuzzy membership function is the most important issue in applying a fuzzy approach. No common approach is available for determining such a function. In some cases, the fuzzy membership function is attained subjectively as a model.

3.2 A Neural Network Algorithm with Fuzzy Input

Taking advantage of data processing using fuzzy sets, fuzzy theory is considered to combine with a BP algorithm. Suppose that we have a recognition problem of *m* classes, which has *m* nodes in the output layer. The weighted distance is first defined between the *i*th class and the *j*th class:

$$z_{ji} = \sqrt{\sum_{k=1}^{n} (\frac{x_{jk} - m_{ik}}{\sigma_{ik}})}\ , \qquad i = 1, 2, \ldots, m \tag{5}$$

where x_{jk} is the *k*th vector of the *j*th pattern vector; $1/\sigma_{ik}$ is a normalisation factor, which results in small class weights for high variance, and m_i and σ_i are the mean value and standard dispersion respectively. The ambiguity of the *j*th pattern belonging to the *i*th class is then defined as follows:

$$\mu_{ij} = \frac{1}{1 + (Z_{ji} / \alpha)^{\beta}}, \qquad i = 1, 2, \ldots, m \tag{6}$$

where α and β are parameters used for controlling the fuzzy degree and α, β>0. According to equation 6.5, there is a low attributive degree if there is a large distance between a pattern and a class. If all elements satisfy $\mu_{ij} \neq 0$, a high fuzziness exists. If only one element satisfies $\mu_{ij} \neq 0$, no fuzziness exists. Under the condition of a high fuzziness, there is a need to modify the ambiguity factor in order to enlarge the difference of membership function.

$$\mu_{ij,INT} = \begin{cases} 2(\mu_{ij})^2 & 0 \le \mu_{ij} \le 0.5 \\ 1 - 2(1 - \mu_{ij})^2 & others \end{cases} \tag{7}$$

For the *j*th pattern, x_j, the *i*th subvector of the desired output, y_j is defined as:

$$y_{ij,INT} = \begin{cases} \mu_{ij,INT} & high \quad fuzziness \\ \mu_{ij} & others \end{cases} \tag{8}$$

where $0 \le y_{ij} \le 1$. All the input and desired output vectors (x_j, y_j) can be used for training with the improved BP neural network.

4 A Classifier Using Rough Sets Based Neural Network with Fuzzy Input

The input data is dealt with using rough sets and fuzzy sets for their powerful function of disposing infinite and incomplete information. This will decrease the number of input nodes and complexity of neural network. A classifier using rough sets based neural network with fuzzy input is proposed (Figure 1).

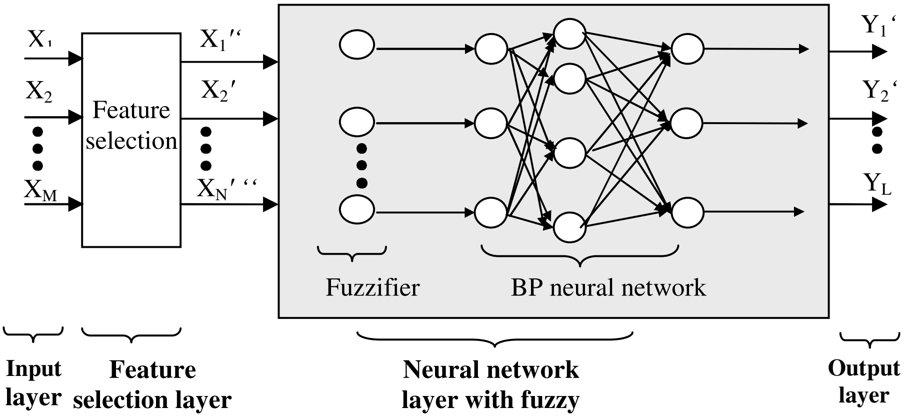


Fig. 1. A classifier using rough sets based neural network with fuzzy input

The system includes input layer, data reduction layer, neural network layer with fuzzy input and output layer.

- Input data

This raw data is preprocessed to suppress noise and normalise the input, and the processed data is input to the feature reduction layer. The normalisation formula is

$$Z = \frac{x - \mu}{3\delta} \tag{9}$$

where x is raw data, μ is the sample mean, δ is the sample deviation and Z is the normalised data which is restricted to [-1, 1][5].

- Feature reduction layer

The normalised data is processed with a hierarchical clustering discretization method. The data reduction is performed using rough sets, and important features for classification remain in the data.

- Neural network layer with fuzzy input

This layer includes a fuzzifier for the reduced data and the improved BP neural network. The crisp input data is converted into fuzzy data through fuzzification, and the fuzzified data is input to the improved BP neural network for defect classification. It is expected that a high classification accuracy and rapid running speed will be achieved through fast data processing.

- Output layer

Output results are obtained from this layer. The maximum coding method is used for the classification decision, which sets the highest output value to 1 and the others to 0. In other words, the defect class corresponds to the output neuron with the largest value.

4.1 Classifier Optimization

Taking into account the difficulties of determining the neural network parameters, uniform design (UD) is introduced to solve parameter optimisation of the neural network. UD is an experimental design method proposed by Fang [6]. It has been recognised as an important space-filling design, which plays a key role in large systems engineering design. UD is equivalent to generating a set of design points that are uniformly scattered in the experiment domain, which reflects the main features of the system. It can solve optimisation problems by finding the maximal or minimal value for the fitness or an error function.

The uniform design method is employed to optimise the parameters, and the 'best' level-combination is obtained to further improve the performance of the neural network classifier [6].

4.2 Experiments

Sample data from wood veneer defects is used for testing, of which 80% is used as training data and 20% as testing data. The experiments are carried out in the 3 groups of data. The results of feature reduction are used by the neural network classifier with fuzzy input.

The fuzzified data is then input to the BP network for training and testing. It should be noted it is not necessary to change the learning rate by large amounts. Especially, the closer the training to convergence is, the smaller the learning rate change should be. Therefore a learning coefficient that is very close to 1, but less than 1, 0.999, is chosen for adjusting the weight slowly. The learning rate η is expressed below:

$$\eta(k+1) = 0.999\eta(k) \tag{10}$$

where k is the number of epochs and maximum of k is 6000, and initial learning rate is set to 0.5. The experimental result is shown in Figure 2, which takes Group 2 of sample data as an example.

For comparison, the BP method with fuzzy input and BP network are adopted and trained with the original sample data, and comparison results are provided in Table 1.

Table 1. Comparison of overall accuracy and running time

	Accuracy	**Average times**	**Average epochs**
BP network	**86.01%**	**324.7s**	**44323**
BP network with fuzzy input	**89.12%**	**14.79s**	**5773**
A fuzzy rough neural network	**96.97%**	**9.67s**	**4322**

The results are further improved with each method proposed in that the average accuracy of classification increases from 86.01, 89.13% to 96.97% and the average running time drops from 324.7, 14.79s to 9.67s. Obviously, the rough fuzzy neural network method has the best classification performance.

Using the fitted model and constrained condition, optimal settings for each of the design parameters are found to give the best predicted performance. for the neural network classifier: Learning rate=0.01, Number of neurons in hidden layer = 57. the UD method has optimised the network parameters with effective algorithms adopted to further improve the accuracy to 98.10%.

5 Conclusions

The hybrid method incorporating a neural network, fuzzy sets and rough sets can achieve a high classification accuracy and rapid speed taking advantage of the complementary characteristics of these techniques even in situations where data is imprecise, noisy, inconsistent and huge. There is no need for extra hardware to deal with uncertainties.

This paper has presented the rough sets based neural network with fuzzy input for pattern recognition as a more effective hybrid approach. In the feature selection process, redundant features are reduced significantly without losing essential information using rough set. In the feature classification, data fuzzification is used to deal with imprecise data and shorten the running time, and the improved BP neural network tackles the local minimum problem to achieve a good accuracy. The hybrid method has taken the advantages of all the techniques incorporated.

Experimental results have shown that the rough fuzzy neural network classifier has a high classification accuracy of wood veneer defects and a short running time. The method is considered general and can be applied to inspection of other products such as ceramic tiles.

References

1. Huber, H.A., Mcmilin, C.W., Mckinney, J.P.: Lumber Defect Detection Abilities of Furniture Rough Mill Employees. Forest Products Journal 35(11/12), 79–82 (1985)
2. Polzleitner, W., Schwingshakl, G.: Real-time Surface Grading of Profiled Wooden Boards. Industrial Metrology, 283–298 (1992)
3. Li, M.X., Wu, C.D., Yue, Y.: A Hierarchical Clustering Method for Attribute Discretization in Rough Set Theory. In: Proceedings of International Conference on Machine Learning and Cybernetics (ICMLC), Shanghai, China, August 26-29, vol. 6, pp. 3650–3654 (2004)
4. Zadeg, L.A.: Fuzzy sets. Information and Control 8, 338–352 (1965)
5. Kjell, B., Woods, W.A., Freider, O.: Information Retrieval Using Letter Tuples with Neural Network and Nearest Neighbour Classifiers. In: IEEE International Conference on Systems, Man and Cybernetics, Vancouver, Canada, pp. 1222–1226 (1995)
6. Li, M.X., Wu, C.D., Yue, Y.: An Automatic Inspection System Based on a Neural Network and Uniform Design. In: International Conference on Machine Learning and Cybernetics, pp. 245–248 (2005)

Study on Row Scan Line Based Edge Tracing Technology for Vehicle Recognition System

Weihua Wang

School of Computer, Chongqing University of Arts and Science,
YongChuan, ChongQing, China
y2002ww@163.com

Abstract. In the process of the feature extraction of vehicle recognition system, the edge vector must be extracted first. However, there are many regions in an image, and it is difficult to extract the edge vectors of these regions of the objects of the image. Therefore a novel edge tracing method based on row-scan-line of image for vehicle recognition tasks is proposed. It was applied to the pixel set formed from the vehicle image in order to obtain the characteristic vector which is very import in the vehicle recognition. The paper shows the reader the feature vector for vehicle contour, the principle of this contour tracing algorithm, the structure of the recognition system, and the implement of this new approach Experiments have been conducted on the videos obtained from a real time monitor, the results show that the contour feature could be obtained in short time, the vector extracting method has good sensitivity to noise and local edge distortions, the edges of the objects could be determined easily, In particular, all of the edge vectors are extracted from the regions of the image at the same time, and a short time in computation can be achieved in the system.

Keywords: tracing technology, feature extraction, row scan line, vehicle recognition, feature vector.

1 Introduction

Feature extraction is a crucial step in vehicle recognition system. It is to evaluate the information including in the appearance of the vehicle object, and to eliminate unnecessary information causing time losses in recognition of the vehicle object. Edge tracing is one of the most fundamental subjects of vehicle recognition. Many vehicle images do not include concrete objects and to understand these objects depends on their structural features. The detection of these features depends on edge tracing, edge line of the vehicle object is accepted as the starting point and the edge is traced.

An edge can be defined as a boundary between two homogeneous areas of different luminance. Local luminance changes and edges corresponding to them are one of characteristic image features providing information necessary in the process of scene analysis and objects classification [14]. Most of contour extraction algorithms consist of the two basic steps: the edge detection and the thinning and linking. They are efficient when applied to the image of nearly homogeneous objects differing significantly from

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 214–221.
springerlink.com

the background if the image is not contaminated by noise. When the level of noise increases the obtained contours are often broken and deformed. That makes the process of interpretation and recognition more difficult. More sophisticated methods should be then implemented. This paper proposes a new algorithm of edge tracing for vehicles recognition system. In this method, instead of detecting all elements of the pixel set, only the pixels with contour lines were traced by scanning the set for a given order. Therefore, it was shown that the edges of basic non-convex and convex objects could be determined easily.

2 Pre-processing for Recognition

2.1 Binarization Processing

Any image from a scanner, or from a digital camera, or in a computer, is a color image. However, the aim of vehicle recognition is to recognize the vehicle according to the contour feature vector of the vehicle. The feature vector used to the image processing is to find a relation between an input image and models of the real world. And the contour feature can be presented by the vehicle boundary. Therefore, binarization image may be used in this vehicle recognition. Following is the binarization processing procedure.

(1) Gray-Scale Transformation Process
We use $f(x,y)$ to express the gray value of the position(x,y). The expression of the Gray-Scale conversion is:

$$f(x,y)=0.299\times r+0.587\times \mathrm{g}+0.144\times b+0.5\,.$$

Where r , g and b are the Red, the Green and the Blue values of the color image pixel.

(2) Image Binarization Process
The simplest way to use image binarization is to choose a threshold value, and classify all pixels with values above this threshold as white, and all other pixels as black. The problem then is how to select the correct threshold. In many cases, finding one threshold compatible to the entire image is very difficult, and in many cases even impossible. Therefore, adaptive image binarization is needed where an optimal threshold is chosen for each image area.

2.2 Edge-Based Segmentation

Image segmentation is one of the most important steps leading to the analysis of processed image data—its main goal is to divide an image into parts that have a strong correlation with objects or areas of the real world contained in the image. Image data ambiguity is one of the main segmentation problems, often accompanied by information noise. Segmentation method can be divided into three groups according to the domain features they employ:

- Global knowledge
- Edge-based segmentations
- Region-based segmentations

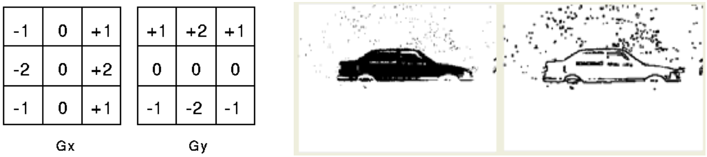
Edge-based segmentation may be applied to vehicle recognition system. It relies on edges found in an image by edge detecting operators, which will be discussed in the following.

2.3 Edge Detection and Thinning

Edge detection is a problem of fundamental importance in image analysis. In images, edges characterize object boundaries and are therefore useful for segmentation, and identification of object in vehicle recognition.

There are many methods for edge detection. The traditional Sobel operator is used in this paper. It is a discrete differentiation operator, computing an approximation of the gradient of the image intensity function. The Sobel operator is based on convolving the image with a small, separable, and integer value filter in horizontal and vertical direction and is therefore relatively inexpensive in terms of computations.

The operator use two 3×3kernels which are convolved with the original image to calculate approximations of the derivatives, one for horizontal changes, and the other for vertical. It is shown in Figure 1(a), and Figure 1(b) shows an example of the boundary obtained by Sobel operator in this paper. The left image is the binarization image obtained above, and the right is the boundary image, which was obtained by upper method.



-1	0	+1
-2	0	+2
-1	0	+1

Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

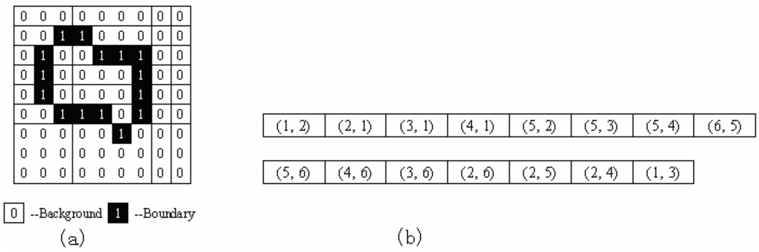
(a) Convolution kernels (b) Edge detection

Fig. 1. Sobel algorithm

3 Edge Tracing Algorithm

3.1 Edge Tracing

The next process after edge detection step is edge tracing. The most important information from the edge tracing is the boundaries of the object. Statistical or geometrical characteristic vectors which are the input data of the image recognition system are obtained from the information about boundaries. A queue can be used to save the vectors of the region boundaries. First, let us assume that the first coordinate of an image is (0, 0). A region boundary of an object is shown in figure 2(a), which was detected by the upper edge detection approach. Figure 2(b) shows the vector queue of the object shown in figure 2(a).



0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0
0	1	0	0	1	1	1	0	0
0	1	0	0	0	0	1	0	0
0	1	0	0	0	0	1	0	0
0	0	1	1	1	0	1	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

0 --Background 1 --Boundary

(a)

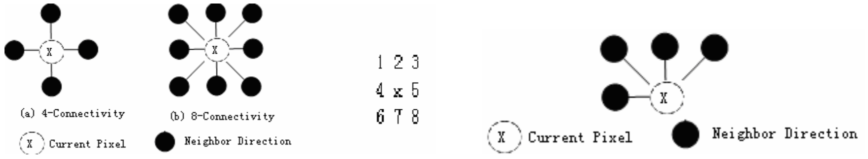
(1, 2)	(2, 1)	(3, 1)	(4, 1)	(5, 2)	(5, 3)	(5, 4)	(6, 5)
(5, 6)	(4, 6)	(3, 6)	(2, 6)	(2, 5)	(2, 4)	(1, 3)	

(b)

Fig. 2. Region Boundary of an Object and the Vector of the Object

3.2 Principle of Neighbor of Tracing Algorithm

There are many approaches for edge tracing. In this study, a new developed edge tracing method based on row scan line was applied to the pixel set formed from the numerical image of an object in order to obtain the characteristic vector which is very important in the recognition of an object. The algorithm can trace all of the region edges of the image easily and extract all of the region edge vectors at the same time, a queue can be used to save the vector of the region of the image. This new edge tracing algorithm tracing in neighborhood of the pixel, there are two methods to define direction of the neighbor of a pixel: 4-connectivity and 8-connectivity. Figure 3(a) shows the 4-connectivity and 8-connectivity direction. 8-connectivity method is used in this paper. Let the pixel neighborhood be labeled as Figure 3 (b).



(a) (b) (c)

Fig. 3. Neighbor of a pixel

In the algorithm, we note that the edge pixels are selected by scan line, and the order of the scanning is from line 0 to line n (n is the last line of the image), and the order of the scanning is from column 0 to column m (m is the largest column) in every line. So the next pixel only has the direction shown in Figure 3 (c) to adjoin.

3.3 Implement of the Algorithm

First, we assume that the image is two-value image, the gray-level value of pixels on the region edge is 0, and the gray-level value of the other pixels is 255. The edges of the regions of the image can be traced by a three-step process:

Step 1: Scan the image from top left to bottom right.
Step 2: Repeat until the pixel to be last.

- Step 2.1: Search the image by the scanning order until a pixel of a region edge is found, and then the gray-level value of this pixel f(x, y) is 0.

- Step 2.2: Select the border which the pixel f(x, y) belongs to. Search the neighborhood of the current pixel f(x, y) in an anti-clockwise direction, beginning the neighborhood search in the pixel positioned in the direction
  - f(x-1, y)
  - f(x-1, y-1)
  - f(x,y-1)
  - f(x+1,y)

  The first pixel found width the same gray-level value as the current pixel f(x, y) is the boundary which the pixel f(x, y) should belongs to, and the coordinate of the pixel f(x, y) is put into the vector of the neighborhood. If the pixel width the same gray-level value as the current pixel is not found, the current pixel f(x, y) is a new boundary, then a new vector is created, and put the pixel f(x, y) into the new vector.

Step 3: Return the vectors which are established by the step 2.

3.4 Performance Analysis

The advantages of this new edge tracing algorithm are those: simultaneity, speediness, good sensitivity to noise and local edge distortions.

(1) At the same time

An example of some feature vectors obtained by edge tracing is given in Figure 4. There are three boundaries in the image, and with the traditional method, these boundaries would be traced three times. In the new algorithm introduced above, the edge pixels are selected by scan line, when the scanning has finished, the whole vectors would be obtained together.

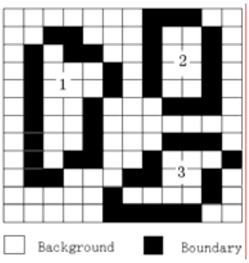


Fig. 4. Some Boundaries of Objects

Therefore, the first and main advantage of the above edge tracing algorithm is that all of the feature vectors of the object boundaries of an image can be tracing at the same time.

(2) Good Sensitivity to Noise and Local Edge Distortions

As a vision based algorithm, image filters are widely used in vehicle recognition algorithm for noise elimination, edge extraction and some other feature extraction. For example, figure 1 (b) left shows the noise of the binarization image, and figure 1(b) right shows the noise of the boundary image obtained by Sobel operator.

From figure1 (b) left, we can see that the main noise of vehicle recognition system is the isolated point, and the noise is expanded by the Sobel operator in figure 1 (b) right. In this study, a new filter algorithm based on isolated point and the length of boundary was proposed In order to filter the noise of the vehicle object.

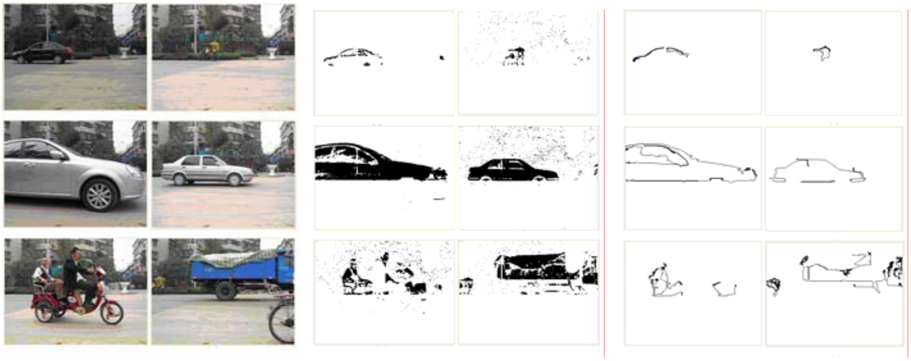
The noise of the vehicle object can be filtered by a two-step process:

- Scan the image from top left to bottom right.
- Repeat until the pixel to be last
  - Step 1: If current pixel f(x, y) is an isolated point, throw off it.
  - Step 2: A threshold value of the length of the border is chosen, hold the borders with value above this threshold, and all other borders are throw off.

4 Computer Simulation and Results

4.1 Image Data

In order to evaluate the performance of the new edge tracing technique, extensive experiments are conduced on many vehicle images, which are obtained from a video produced by a real time monitor. Figure 5(a) shows some vehicle image data obtained above. Figure 6 is the background image, and the others are the test data.



(a) Test Image Data (b) Binarization Images (c) Boundaries Images

Fig. 5. Test

Fig. 6. Background Image

4.2 Experiments

The experiment of the method in this study was done by following process:

- Pre-processing: Vehicle image pre-processing, including size normalization, gray-scale transformation and binarization transformation is helpful for the vehicle recognition system.
- Edge detection.
- Edge tracing.
- Feature vectors extracting.
- Recognition processing.

The following image in Figure 5(b) shows an example of the binarization image obtained from figure 5(a). And Figure 5(c) shows an example of the boundaries image obtained from figure 5(a).

4.3 Results

Table 1 shows a performance about the comparison of various image sizes.

Table 1. Performance

Image size	Tracing time(millisecond)
320×240	3.193
320×120	1.454
320×80	0.905
320×40	0.491

In Table 2, it demonstrates the tracing time of the above tracing algorithm by different regions.

Table 2. Performance

Region number	Tracing time(millisecond)
4	3.121
5	3.121
6	3.120

5 Conclusion and Future Work

With this new approach discussed above, all boundaries of regions of an image can be traced at the same time in the object recognition process, thus getting the characteristic vectors in shorter times. Therefore, this method can resolve the problem discussed above, and it is fit for the vehicle recognition system. The disadvantages of the approach are those: not to be able to distinguish the edges of the regions which have intersections, not filter the noise efficiently, and not to be able to sense the inner region of a region of the image, in the future, we will extend this work to solve the issues above in the future.

References

1. Wang, C.-C., Huang, S.-S., Fu, L.-C.: Driver assistance system for lane detection and vehicle recognition with night vision. In: Intelligent Robots and Systems (IROS 2005), pp. 3530–3535 (2005)
2. Chertov, M., Maev, G.: Extraction of the Straight Line Segments from the Noisy Images as a Part of Pattern Recognition Procedure. In: Vth International Workshop, Advances in Signal Processing for Non Destructive Evaluation of Materials, pp. 25–31 (2005)
3. Hwang, W., Ko, H.: Real-time vehicle recognition using local feature extraction. Electronics Letters 37(7), 424–425 (2001)
4. Yang, C., Wen, X., Yuan, H., Duan, B.: A Study on Data Parallel Optimization for Real-time Vehicle Recognition Algorithm. In: IEEE Intelligent Transportation Systems Conference Seattle, WA, USA, September 30 - October 3, pp. 661–665 (2007)
5. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision. Posts & Telecom Press, Beijing (2002) (in Chinese)
6. Foley, J.D., van Dam, A., Feiner, S.K., Hughes, J.F., Phillips, R.L.: Introduce to Computer graphics. China Machine Press, Beijing (2005) (in Chinese)
7. Kazemi, F.M., Samadi, S.: Vehicle Recognition Based on Fourier, Wavelet and Curvelet Transforms – a Comparative Study. In: IEEE, International Conference on Information Technology (ITNG 2007), pp. 939–940 (2007)
8. Bagciogullari, F.: A Edge detection algorithm developed for object recognition system. In: The Symposium of 2. Signal Processing and Applications, pp. 47–52 (1994)
9. Kang, C.W., et al.: Extraction of Straight Line Segments Using Rotation Transform: Generalized Hough Transform. Pattern Recognition 24(7), 633 (1991)
10. Marefat, K.: Geometric reasoning for recognition of 3-D object feature. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 950–964 (1990)
11. Rahati, S., Moravejian, R., Mohamad, E., Mohamad, F.: Vehicle Recognition Using Contourlet Transform and SVM. In: Fifth International Conference, April 2008, pp. 894–898 (2008)
12. Malik, M.R., Balakumar, P.: Acoustic receptivity of Mach 4.5 boundary layer with leading-edge bluntness. Theor. Comput. Fluid Dyn. 21, 323–342 (2007)
13. Webber, R.E., Samet, H.: Linear-Time Border-Tracing Algorithms for Quadtrees. Algorithmica 8(1-6), 39–54 (1992)
14. Genov, R., Cauwenberghs, G.: Kerneltron: Support Vector Machine in Silicon. In: Lee, S.-W., Verri, A. (eds.) SVM 2002. LNCS, vol. 2388, pp. 120–134. Springer, Heidelberg (2002)
15. Herng Wu, E.J., De Andrade, M.L., Nicolosi, D.E., Pontes, S.C.: Artificial neural network: border detection in echocardiography. Medical and Biological Engineering and Computing 46(9), 841–848 (2008)
16. Kyo, S., Okazaki, S., Arai, T.: An Integrated Memory Array Processor Architecture for Embedded Image Recognition Systems. ACM SIGARCH Computer Architecture News 33(2), 134–145 (2005)

Study on Modeling and Simulation of Container Terminal Logistics System

Li Li* and Wang Xiaodong

Hebei University of Technology, Tianjin, China, 300131
danmama95@sina.com

Abstract. Modeling and simulation of container terminal logistics system is rapidly developed and widely applied. This paper describes the process of the modeling and simulation of container terminal logistics system, which includes Logistics system modeling technology, Verification Validation and Accreditation and Screening Test and Robust Design, and then the research approach and target, the primary coverage and the key problem is given after summarized and selected. This study is devoted to determine the key problems and the best solutions in the process of modeling and simulation, which contributes to reduce the difference between the simulation and reality system and increase the capacity and efficiency of container terminal logistics system.

Keywords: Modeling, Simulation, Logistics, Testing, Verification Validation and Accreditation, Robust Design.

1 Introduction

Containerization has increasingly facilitated the transportation of goods since 1970s and it is still spreading all over the world. On the basis of the present trends, it is expected that the containerization ratio will be over 70% of all general cargo by 2015.

With the growth of container transportation demand, the expansion of port size and the growing number of external factors of uncertainty, there is a huge number of port operators who are meet with an increasingly severe crisis from their management. For such complex systems, the authors use simulation as a tool for supporting some strategic decisions involving too many different variables, which is not possible to manage otherwise. It is a broad applicable prospect for Container Terminal Logistics System (CTLS) in Modeling and simulation optimization theory.

The CTLS mainly consists of three subsystems, which conclude ship berthing, yard operation and transportation in port. The CTLS is affected by much uncertain factors, such as the Estimated Time of Arrival of ships, which is a discrete -event system, and dynamic characters is used to be presented as a set of discrete equations. And the ship berthing subsystem is also a random service system.

As known to all, it is Part of the international research in the field of the hot and difficult problems for CTLS in Modeling and simulation optimization. Projects focus

* Corresponding author.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 222–230.
springerlink.com

on how to create a simulation model, firstly uncertainty factors is analyzed, which is about the impact of efficiency of the port logistics system operating, and to utilize screening technology branch of the order, the various subsystems container logistics is re-optimized, thereby optimizing the system.

The paper is organized as follows. The technology about modeling of CTLS is given in Section 2, together with some definite questions in this stage. The Verification Validation and Accreditation is presented in Section 3, where the authors focus on main problems in this stage. The technology about screening test and core substance are introduced in Section4. Finally, some conclusions are given in section 5.

2 Logistics System Modeling Technology

2.1 Oversea Review

In this part, the substance and the technology of CTLS modeling research would be illustrated.

The substance of CTLS modeling research can be summarized as: operation and Handling Technology System. Operation: In this part, Resource allocation, Daily operation administration and Handling Technology System is presented.

Resource allocation: A mixed integer programming formulation for the container bridge scheduling of the problem has been proposed and a branch-and-bound algorithm and heuristic algorithm for the optimal solution is constructed [1]. On the basis of the container bridge scheduling research, with the aim of getting minimum waiting time of installation working, the mathematics model was established and a branch-and-bound algorithm is constructed [2]. Studying the landing and offing time of single container vessel and the service situation of traveling bridge given, the solution of minimizing operation time was obtained by using the branch-and-bound algorithm and greedy algorithm[3].

Daily operation administration: The container port simulation system is developed by Mitsui Engi-neering & Hipbuilding Cmpany, which could be used to simulate daily operation and management [4]. The main viewpoint is that highly efficient operation could optimize containers flow [5].

The specific part of operation: The main content is about the distribution of the storage space in yard, which concludes two parts, firstly the workload of containers in storage space is balanced and then the load is connected with vessels, at last transport distance of container is calculated to up to minimum [6]. The problems about an efficiently yard organization for handling and moving containers sequentially, and the dynamic model is introduced [7].

Handling Technology System: The typical containers handling technology system is divided into Rubber Tyred Gantry Cranes (RTGC), Railway Moving Gantry (EMC), RTGC- EMC, automatic containers terminal system and straddle truck, and the last two is more popular in recent years.

The main simulation technology abroad concludes Petri Net, Simulation-Modeling Technology and Analytical Model.

Petri Net: Petri Net is a combined model expressed by a set of simple graphics. The container port is divided into several operated spaces of containers handling and moving and the transportation which is consist of different vehicles are established [8].

This field is mainly about the three subsystems concluded ship berthing, yard operation and transportation. The schematic diagram is as follow.

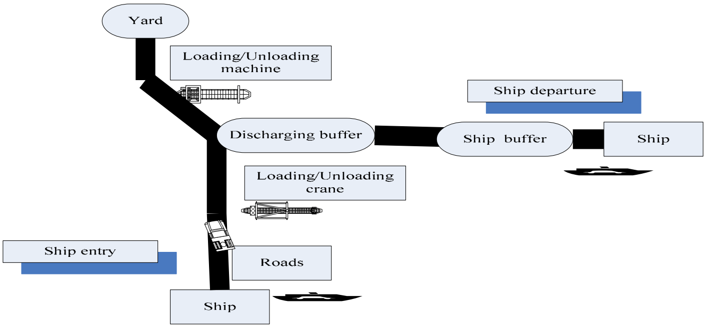


Fig. 1. Schematic diagrams of yard operation

Simulation-Modeling Technology

Ot-oriented Simulation and Modeling: The proprietary software about port logistics simulation has been invented. In [9], the flexism simulation environment is introduced, which is developed by C++ and used to develop, modeling, simulate and monitor the dynamic process and system.

Modeling and Simulation via Special Simulation Language: In [10], the simulation system of handling operation of container vessels is presented, which consists containers, vessels, vehicles, yard and handling bridge, and the constraint condition of system is defined by users.

2.2 Review in China

In china, the research in the distribution of resource mainly includes the ratio and the optimization of the allocation of equipments between the connections of operation parts. The throughput of container port and the equipment optimization is analyzed based on Shanghai Waigaoqiao Port Area [11]. According to the principle of storage and retrieval, the container yard layout is designed in terms of heuristic algorithm in order to minimize the frequency of equipment operation in yard [12].

The research of handling technology system is barely automatic containers terminal system and straddle truck, but simulation and modeling of the design of technological scheme based on the Rubber Tyred Gantry Cranes (RTGC), Railway Moving Gantry (EMC), RTGC-EMC.

2.3 Models

Though models in different reference are distinct, the theoretical basis of the staple methods in common could be divided into operational research, cybernetics and system simulation.

OR based modeling methods that be used to describe problem in general are integer optimization, integer optimization, queuing theory, evaluating strategies model, network flow, game theoretical model, statistical analysis and traditional optimization method, in which theory of the models referred to every major in logistics management, and the solvers are branch and bound, Lagrange multipliers, Bender decomposition method and so on.

In the control theory based upon process system engineering method, the differences between the status variable and the control quantity is extremely discriminate in CTLS, as well as the perturbation and output variable.

The quantity of containers in yard and vessels berthing are status variables that ought to monitor emphatically, while the plan of operation and the amount of containers among the subsystems stands for the systemic control strategic decision, namely control quantity. Therefore the design of CTLS is similar to design a reasonable control decision for a dynamical system, in which the system disturbance corresponds in variations of systemic output. And the disturbance is possible to be random variable, or be the variability of special mode.

In addition, the output of system is conveyance between subsystems. The CTLS modeling according to control theory is used to generally make decisions of daily operation, which also is limited, especially in the event linear hypothesis. If the system is unusual complicated, the models is difficult to describe, what's more solve.

The modeling method based on system simulation applied to management science is a newer field. The method is used to analyze CTLS in the round, since optimum relation isn't exit in the complicated system, and considering the all kinds of factors, the simulation is particularly adequate for CTLS which is great scale and difficult to be mathematical analysis and physical test.

2.4 Primary Coverage and Key Problem

In the stage of system modeling, the necessary data is collected and the key influence factors that impact the operational efficiency of any subsystem is analyzed for the purpose of the establishment of the physical model; and the next is that based of the simulation software Arena, the simulation system of any subsystem of CTLS would be set up. The main problems of this stage as follows:

How to define and describe the resource input/output that happen in the system boundary of subsystems.

How to define and distinguish the random factors, that impact efficiency of subsystem, and then discovered the parameter attributes and the statistical rules of these factors.

How to build the subsystems such as conclude ship berthing, yard operation and transportation in port.

How to establish the main functional modules with a three-layer structure, the following modules are included: the terminal layout, resource plan and job scheduling.

How to find out the logic connection among these functional modules, moreover, discover the logical relation of the behavior characteristic and strategy.

In this stage, the key issue is to distinguish the random factors and their numeric area, which impacts the efficiency of subsystem and operational efficiency index.

3 Verification Validation and Accreditation

3.1 Review

In the research area of reliability assessment above simulation system, a large amount of problems should be taken account, such as the accuracy of systems, the dependability of software and hardware of computer, the performance of simulator equipment, and the correctness of experimental results that used to achieved the fitness of purpose of analysis and decision. Verification, Validation and Accreditation (VV&A) is foundation of reliability assessment.

The statistical methods have application in verifying simulation model to verify [13] based on the data acquisition from actual system, the verification will be divided into the following three conditions.

No input and output data. The data in actual system is difficult in collecting or non-existent, thus data is produced via the simulation model and Design of Experiments (DOE) is used to check data. In the model verification of seabed mining detecting, the bilinear regression prediction model is established for estimating the parameter value, and which compared with the specialists' experience, through the method the correctness of model is verified at last [14]. Described a swine fever case in food service industry, in which regression model is also taken advantage since without the real input and output data [15].

Only output data. The statistical methods are used to verify simulation and the actual output data. Under the circumstance, through statistical hypothesis the real and simulative output data are integrated together for verifying and analyzing the simulator result.

Output and input data. In this case, the Tracking simulation is applied to validate and verify models. In reality, the most of data is non-normal distributed, therefore, Bootstrapping is further developing used to certify the relation between the simulator and the real, which only focus on the statistic progress of the data without regard to the distribution.

The simulated system about production line of Ericsson Company in Sweden is given in [16], the partial truthful data is collected and the relation between the simulation and reality is tested in terms of the historical experience, then the optimized system is design yet the system could be unverified. Therefore, there are requirements for the further investigation in the field of the method about the VV & A of CTLS.

Whether general manufacturing system or rare CTLS, the research about VV & A is not a lot, thereby the analysis about simulated system is difficult and obstructive. For the simulation of port logistic system, the premise is the establishment of truthful system model, meanwhile which also the difficulty of simulation.

3.2 Primary Coverage and Key Problem

In this stage of VV & A, the simulation model is verified whether the real system is correctly described via model. The main problems as follows:

How to verify the simulation model of CTLS based on the VV & A specification of DOD.

How to choose the proper static method in terms of the nature of data and the character of data distribution, and then certify the correctness of simulation system.

How to take advantage of experience design approach and regression analysis to estimate parameter thus variation trend of the parameter is compared between the reality and simulation.

The key problem is to verify the validity of simulation, where the parameter is compared between the reality and simulation. And the other is to verify the validity of simulation system because of choosing the proper static method in terms of the nature of data and the character of data distribution.

4 Screening Test and Robsut Design

4.1 Screening Review

Screening can be defined as a process of searching for a relatively important one from all the factors that need to be considered. In generally, there are various kinds of machinery in CTLS, and they have relation to each other. For this reason, so as to establish the simulation model about complex CTLS, factors (also known as input variables) and the roles among related factors must be considered.

Therefore the option of screening test technology often is a very critical matter, which including One factor at a Time (OAT), Iterated Fractional Factorial Design (IFFD), Sequential Bifurcation (SB) and so on. Only if the smoothness and the monotonous of input-output function of simulator models considered, the different test methods is definite. SB is proved highly active if the real input-output function could generate monadic regression.

So as to establish the simulation model about complex CTLS, factors (also known as input variables) and the role of factors related must be able to consider, therefore actually should choose which kind of screening test technology, often is a very critical matter, which often including One factor at a Time (OAT), Iterated Fractional Factorial Design (IFFD), Sequential Bifurcation (SB) and so on.

The Screening of factors is very importantly practical; there are still a lot of problems. The project will focus on SB method, which is applied to screen complex system factors in CTLS.

4.2 Robust Review

Robust design is also called risk analysis, which originated a design engineer named Taguchi who devoted to devise stable auto product line adapted to the change of market requirement. The elements in design are divided into two types: controlled variable and interfere factor. The stable value of controlled elements is generated through the variation of interfere factors. The robustness of CTLS is a solution that exist stable and less cost of system in the event that external environment smoothly change, which is different from optimal decision case and is more practical.

Because in practice optimal relation is not easily adaptive to environment change which compromised weather variations, marine risk, vessel anchored waiting, vehicle malfunction and road maintenance.

In the CTLS simulation, the accident probability is just above abrupt high-cost, if the unexpected probability caused by the variations of external environment is in excess of a special value, such as the worthless cost cannot be accepted by managers, the controlled variable of simulator ought to be adjusted. Therefore, the range of controlled factors is determined, and the system output variable is restricted under a specific value in the strategy of robust.

In this project, the concept of Taguchi is applied, but without regard to the statistic method since it is more applicable to design the merchandise, nevertheless the investigative complicated system involved more factors and relation.

Supposed a particular environment, then the optimal solution is discovered, at last Latin Hypercube Sampling(LHS) is used to estimate the robust degree of optimal plan after changed environmental elements, in which the late period stability is analyzed. On the opposite, the early stage robust analysis is more simulating before the project implement.

4.3 Primary Coverage and Key Problem about Screen

In this stage, the approach of SB is used to screen the main factors that impact the operation efficiency of these subsystems. The SB is that the relatively important factors are selected from the simulation system that includes a deal of factors. The main problem as follows:

Whether the factors could be efficaciously screened form any numbers of output variables or any other screen approach could be applied, because SB is always applied in only one output variable yet CTLS contains a great deal variables.

How to verify the correctness of the output result after the end of SB, and check the unimportant factors is whether used to properly describe CTLS.

The key problem is that in the progress of SB, how to determine the span of all data and the numeric area of one data, how times the various combinations are calculated are proper, and how to define the rule of screen ending.

How to verify the correctness of the output result after the end of SB, and check the unimportant factors is whether used to properly describe CTLS.

The key problem is that in the progress of SB, how to determine the span of all data and the numeric area of one data, how times the various combinations are calculated are proper, and how to define the rule of screen ending (if the distinct interaction between factors, the simulation of combination would be stop.)

4.4 Primary Coverage and Key Problem about Robust

In the last period of robust design, the optimized proposal would be designed by mean of changing the character of parameter of the screened main influence factor. The experimental data is produced by array center approach that is one of design of experiments and fitted via the surfaced response-models, finally the optimized parameter is generated. The main as follows:

How to get the optimized value of controllable factor, with the help of the optimization software Opt-Quest that is complement of Arena.

How to determine the result is optimality. Even though the optimized result is definite in terms of the satisfaction of KKT when the output data changed if is not one but many data, how to purpose the optimization of the complete random system.

How to determine the systemic optimality on the events that the parts of systems are optimal.

How to determine the robust value via the optimal value. Combining the optimal design theory with robust design, the mathematical optimization algorithm would apply to in CTLS.

In this stage, the key is if all subsystem is optimum how to design an optimal-zing solution of the entire CTLS. In addition, the other is how to determine the robust value via the optimal value.

5 Conclusion

The papers of this field were published in high-level journal and conference in late two years, which indicated that research is advanced and theoretical significance.

This study contributes to reduce the difference between the simulation and reality system and develops the capacity and efficiency of container terminal logistics system. The simulated system promotes to discover and decrease the uncertainties' affection for the CTLS from the standpoint of the container flow and research into the modeling and simulation of CTLS is directed toward improving the ability to analyze and predict these uncertain factors and thus to minimize transportation cost and expand the performance of container terminal completely.

Acknowledgments. We wish to thank Wan Jie, since she kindly helped me accomplish our idea and research. Her research extremely promoted this project. This work was supported in part by a grant from Natural Science Project of Hebei Province (No.F200900123).

References

1. Park, Y.-M., Kim, K.H.: A scheduling method for Berth and Quap cranes. OR Spectrum 25, 1–23 (2003)
2. Ng, W.C., Mak, K.L.: Yard crane scheduling in port container terminals. Applied Mathematical Modelling 29, 263–276 (2005)
3. Kim, K.H., Moon, K.C.: Berth scheduling by simulated annealing. Transportation Research Part B 37, 541–560 (2003)
4. Park, Y.-M., Kim, K.H.: A scheduling method for berth and quay cranes. OR Spectrum 25(1), 1–23 (2003)
5. Kwashima, K., Ksahara, K., Sato, S.: On-demand container terminal system. Mitsui Zosen Technical Review (178), 81–86 (2003)
6. Zhung, C., Liu, J., Wan, Y.-w., et al.: SLorage space allocation in container terminals. Transportation Research–B 37(10), 883–903 (2003)

7. Kim, K.H., Park, K.T.: A note on a dynamic space-allocation method for outbound container. European Journal of Operational Research 148(1), 92–101 (2003)
8. Degano, C., Di Febbraro: A Modelling automated material handling in inter-modal terminals. In: 2001 IEEE/ASME International Conference on Advanced Intelligent Mechatronics. Proceedings (Cat No01 TH86), vol. 2(2), pp. 1023–1028. IEEE, Piscatway (2007)
9. Nordgren, W.B.: Flexsim simulation environment (Item 7 from file2) DIALOC (R) File 2 INPEC. In: Winter Simulation Conference, Cat No U2CH37393. IEEE, Piscatway (2002)
10. Nehrling, B.C.: Container ship loading and unloading simulation. Michigan University. Ann Arbor Department of Naval Architecture and Marine Engineering, Ann Arbor MI - 48704
11. Guansheng, Y., Yizhong, D.: The handling capacity and equipped devices in Container port of Shanghai. China Forts (2002)
12. Shuqin, Y., Yunjie, Z.: A model and its algorithm on container yard problem. Journal of Dalian Maritime University 28, 115–117 (2002)
13. Kleijnen, J.P.C.: Validation of models: statistical techniques and data availability. In: Farrington, P.A., Nembhard, H.B., Sturrock, D.T., Evans, G.W. (eds.) Proceedings of the 1999 Winter Simulation Conference, pp. 647–654 (1999)
14. Kleijnen, J.P.C., Bettonvil, B., Persson, F.: Robust Simulation Optimization: Methodology and Supply-Chain Case Study. Working Paper, Tilburg University (2005)
15. De Vos, C., Saatkamp, H.W., Nielen, M., Huirne, R.B.M.: Sensitivity analysis to evaluate the impact of uncertain factors in a scenario tree model for classical swine fever introduction. Working Paper, Wageningen University (2005)
16. Persson, F., Olhager, J.: Performance simulation of supply chain designs. International Journal of Production Economics 77, 231–245 (2002)

Digitalized Contour Line Scanning for Laser Rapid Prototyping

Zeng Feng[1,2], Yao Shan[2,*], and Ye Changke[2]

[1] Computer School, JiaYing University, Meizhou, Guangdong, China, 514015
[2] School of Materials Science and Engineering, Dalian University of Technology, Dalian, Liaoning, China, 116024
yaoshan@dlut.edu.cn

Abstract. The process principle of Selected Laser Sintering, Stereo Lithography and Laminated Object Manufacturing are described in detail. A novel laser rapid prototyping approach based on Digitalized Contour Line Scanning (DCLS) is present which combine several merits of SLS and LOM. Thermosetting powder material is used in DCLS. Laser beam scans the 3-D contour line of physical prototype layer-by-layer, and the material scanned by laser heats up to lose thermosetting property. After being sintered by heating furnace, the whole sintered part divides into two separate parts, and the physical prototype is obtained. For the case study, impeller mould was made by DCLS used coated sand, and further, the metal impeller was cast. The products of DCLS have high density and hardness.

Keywords: Laser Rapid Prototyping, Digitalized Contour Line Scanning, Thermosetting Powder Material.

1 Introduction

Rapid Prototyping (RP) is based on computer numerical control, computer graphics, materials science and technology, computer aided design, digital manufacturing, and laser technology [1-3]. It is a new manufacturing technology which appeared in the late 1980s [4]. Rapid Prototyping has the high manufacturing flexibility [5]. CAD model with complex 3-D shapes is converted to a physical prototype quickly, without any fixture [6-7]. Different from the previous manufacturing technology, RP is a novel processing which is based on “addition” principle. The material is accumulated step-by-step and layer-by-layer in 3D space under the control of digital information, and the rapid prototyping is called the “digital forming technology”.

Rapid prototyping is usually divided into two broad groups: RP based on laser technology and RP based on micro drop technology. Selected Laser Sintering (SLS), Stereo Lithography (SL) and Laminated Object Manufacturing (LOM) are the most three common approaches. As shown in Fig. 1, Rapid Prototyping processing has the following steps:

[*] Corresponding author.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 231–238.
springerlink.com

Step 1. The 3-D CAD model of product is designed by computer.

Step 2. Layered software slices the 3-D CAD model with certain thickness after analysis and selection of forming direction.

Step 3. Under the control of computer, the current layer is processed by the laser beam or the micro drop nozzle.

Step 4. Worktable falls down with the distance of one layer thickness and next layer of material is processed. Repeat the step 3 and step 4 until the physical prototype of part which is developed by CAD system is obtained.

Step 5. The physical prototype is post-processed according to the product requirement.

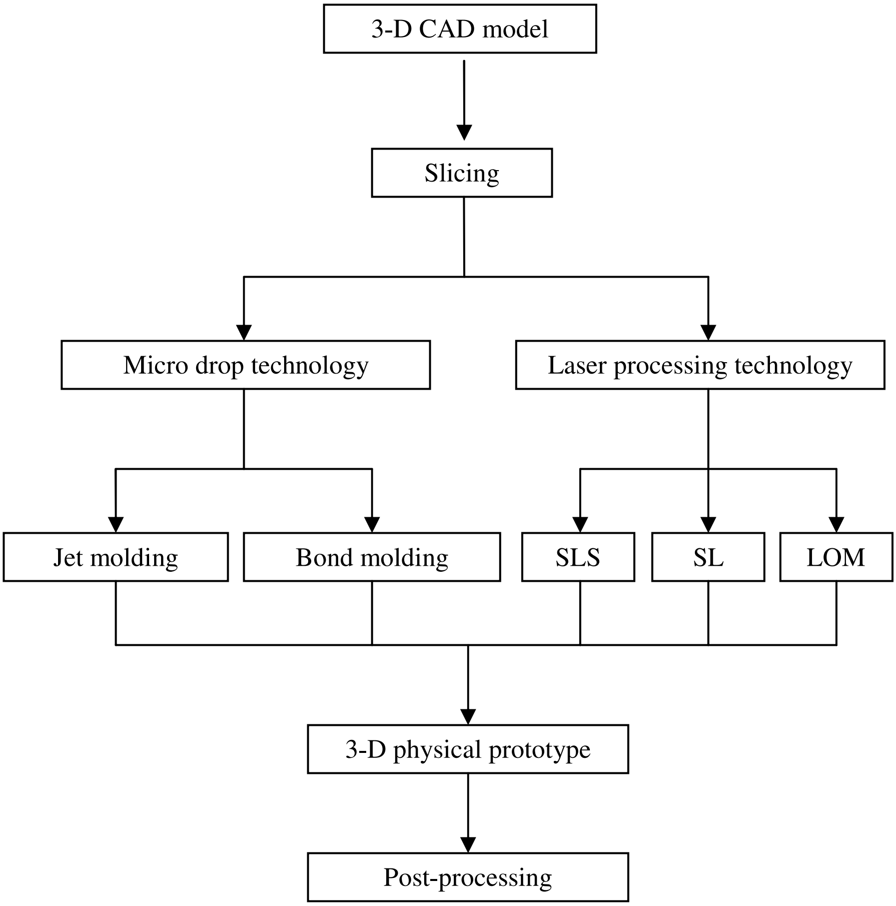


Fig. 1. Processing flow chart of rapid prototyping

2 Principle Comparing of SLS, SL and LOM

Powder materials are used in Selected Laser Sintering. Under the control of computer, the powder material scanned by the laser beam heats up and consolidates [8].

Worktable falls down with the distance of one layer thickness, and next material layer was processed, until the physical prototype of part which is designed by CAD system is obtained. The process principle of SLS is shown in Fig. 2.

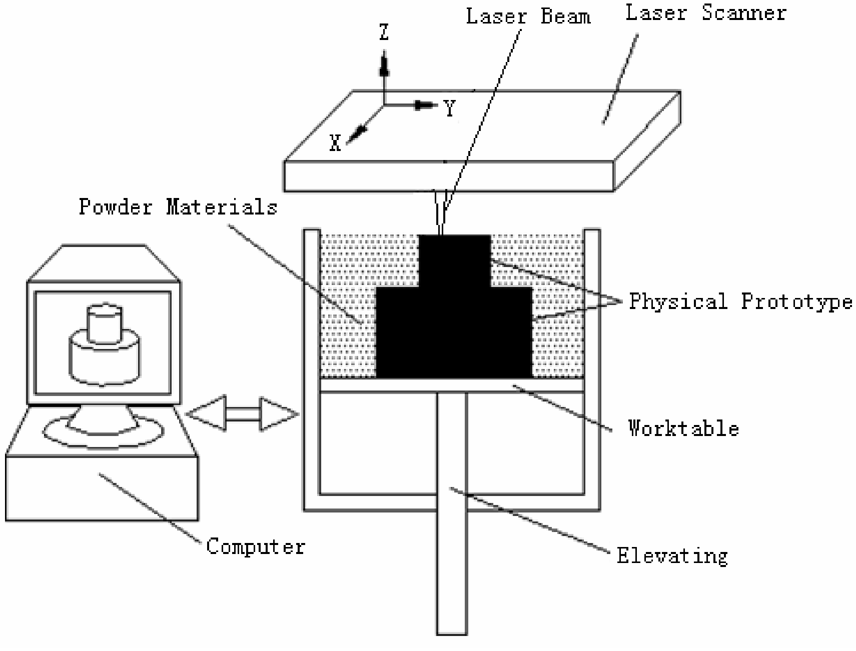


Fig. 2. Process principle of SLS

Using the powder materials, the physical prototype with complex shapes can be made by SLS [9-10]. Wax, engineering plastics, polymer coated powder, metal powder, PA, and ceramic powder can be molded by SLS [11-13].

Liquid photosensitive resin is used in Stereo Lithography process (SL). Photopolymerization would occur rapidly when the liquid photosensitive resin is irradiated by UV laser, and the material shift from liquid to solid. SL is developed on the basis of the liquid photosensitive resin with the UV laser curing characteristic. SL process is similar to SLS. The UV laser beam scans the Liquid photosensitive resin layer-by-layer until the physical prototype of part is obtained. In the process of SL, solidified depth of photopolymer resin controlling is a vital factor for affecting the accuracy of the shape and dimension of physical prototype. The process principle of Stereo Lithography is shown in Fig. 3.

Layered materials are used in Laminated Object Manufacturing. Under the control of computer, current layer of paper or other material is cut by laser beam. After the contour line of cross section is scanned, next new layer of paper is bonded to the previous layer by using a heated roller. And the next layer of paper is also cut by laser beam. In order to stripping the physical part easily, the excess paper is cut to grids by laser beam. Repeat this process until all layers of CAD model are cut. At the end of the LOM process, the excess material is removed, and the physical prototype is obtained. Laminated Object Manufacturing is especially suitable for foundry patterns, vacuum-form tooling, and larger models. The process principle of Laminated Object Manufacturing is shown in Fig. 4.

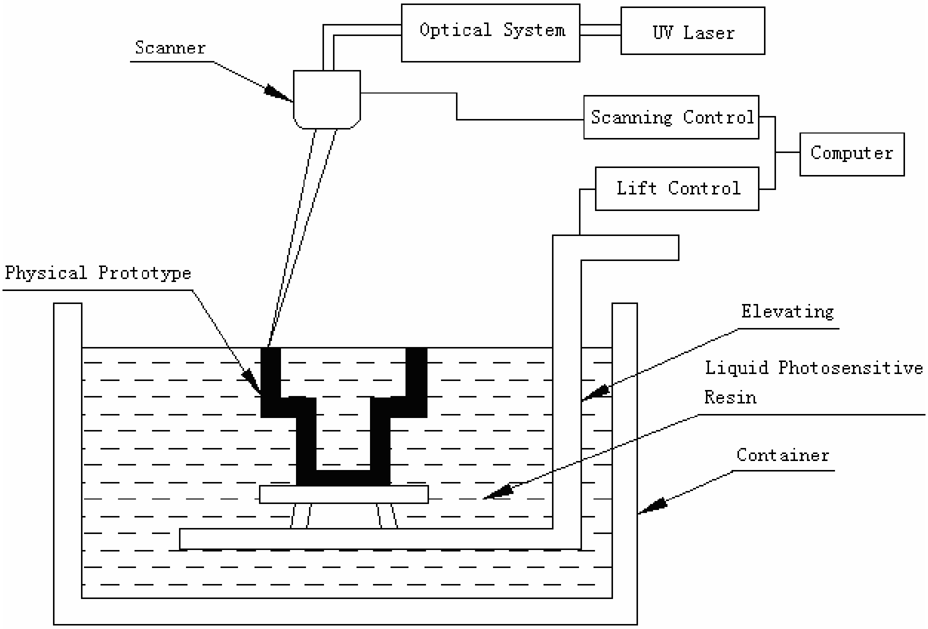


Fig. 3. Process principle of SL

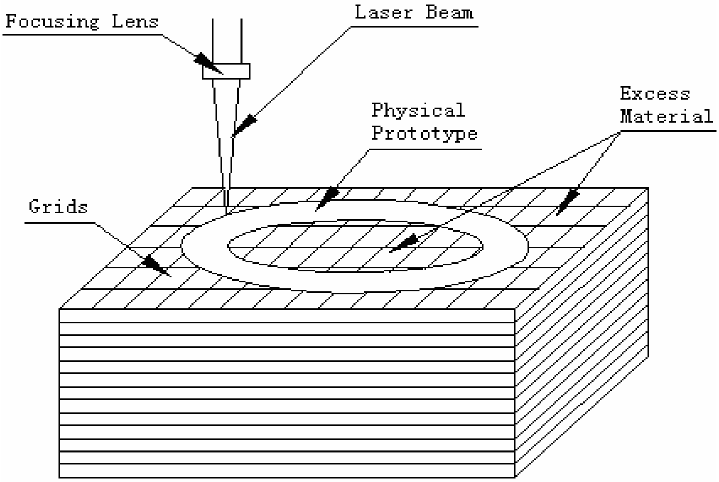


Fig. 4. Process principle of LOM

SLS, SL and LOM have been successively developed and commercialized, and they have different merits. But what is difficult to be overlooked is that, SLS, SL and LOM have their demerits which affect the hardness and precision of products because of the short developing history and the limitations of process themselves. The merits and demerits of SLS, SL and LOM are given in table 1.

Table 1. Merits and demerits of SLS, SL and LOM

RP Process	Merits	Demerits
SLS	1. A large variety of choices for SLS materials 2. Need no support structure 3. Some materials can be sintered into functional parts.	1. Low processing speed 2. Low hardness and low density for product
SL	1. High material utilization rate 2. High accuracy and surface quality for product 3. Directly manufacturing plastics parts	1. Low processing speed 2. High price for materials 3. Need support structure
LOM	1. High processing speed 2. High hardness and low density for product 3. LOM is suitable for manefacture lage models.	1. Few choices for LOM materials 2. The thickness of each layer can't change freely.

3 Digitalized Contour Line Scanning

Digitalized Contour Line Scanning (DCLS) for laser rapid prototyping is present based on the merits and demerits of SLS, SL and LOM. Thermosetting powder materials are used in DCLS, and the thermosetting property of these materials would be destroyed if they are heated to a certain high temperature. When DCLS works, laser beam scans the current layer of material along the contour line of 2-D cross section. The material scanned by laser beam heats up and loses thermosetting property. When scanning 3-D contour of physical prototype finished, the whole powder material is sintered by heating furnace or infrared hot plate. Finally, the sintered part is divided into two separate parts which are physical prototype of CAD model and the excess part.

3-D contour information is extracted digitally in DCLS. The contour line of CAD model is set to “0” and the rest of model is set to “1” by software. Data whose flag is “0” are sent to computer control system which drive the laser beam to scans the contour line, while the rest of material which is set to “1” by software are not any processed. The process of DCLS is shown in Fig. 5.

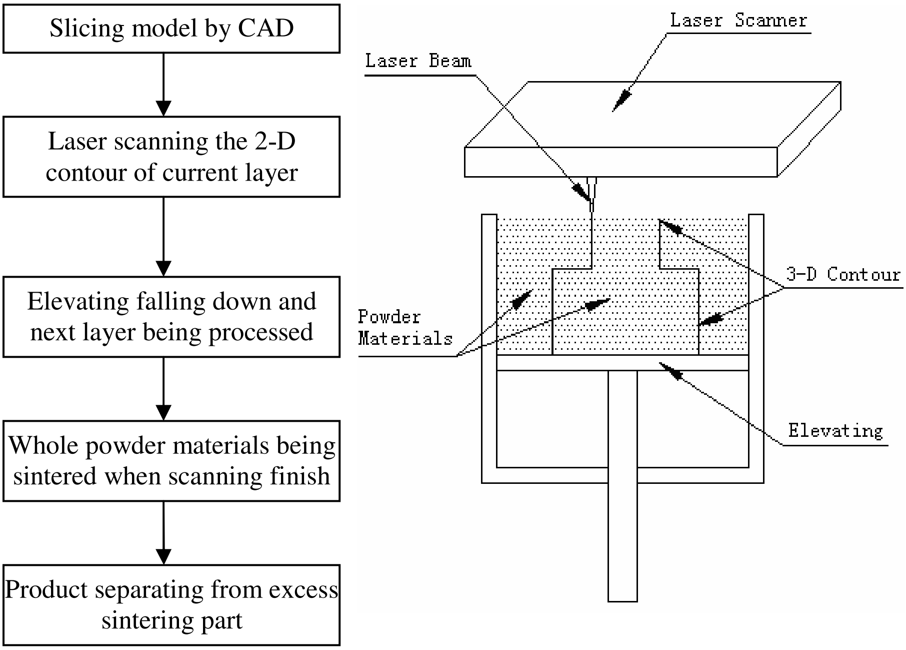
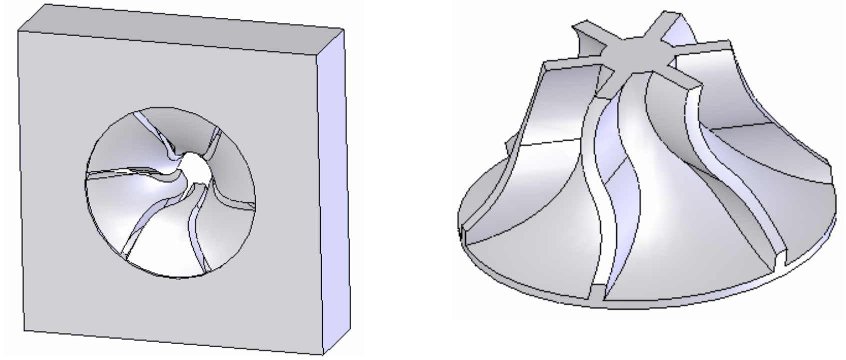

Slicing model by CAD
Laser scanning the 2-D contour of current layer
Elevating falling down and next layer being processed
Whole powder materials being sintered when scanning finish
Product separating from excess sintering part
Laser Scanner
Laser Beam
3-D Contour
Powder Materials
Elevating

Fig. 5. Process principle of DCLS

4 Case Study

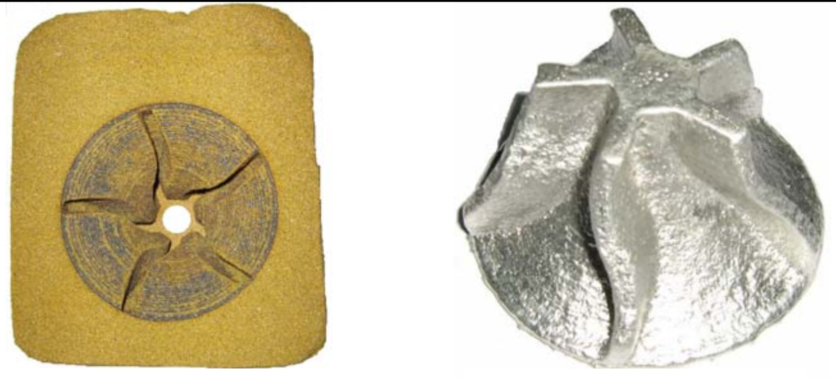
There are a large variety of choices for DCLS materials. Most of the thermosetting powder materials can be used for DCLS. The laser beam scanning is just an effect that powder material is divided into two parts, and the product is sintered by heating furnace, so, the physical prototypes of DCLS have good mechanical properties. And, only the contour lines are scanned by laser beam, the processing speed is high.

Coated sand was used in DCLS case study. According to experiments, the suitable process parameters are $P = 30(W)$, and $v = 0.007(m/s)$, where P is laser power and v is scanning speed. Models of impeller and the mould were designed by CAD software, as shown in Fig. 6, while the physical prototype of mould and final casting part are shown in Fig. 7.



(a) CAD model of impeller mould (b) CAD model of impeller

Fig. 6. CAD model



(a) Physical prototype of impeller mould (b) Finish casting impeller

Fig. 7. Products

5 Conclusion

The most common laser rapid prototyping approaches were introduced, and a novel laser rapid prototyping approach, Digitalized Contour Line Scanning (DCLS), is present, based on the merits and demerits of SLS, SL and LOM.

This approach uses thermosetting powder materials. The physical prototype's 3-D contour line which is scanned by laser beam heats up to lose thermosetting property, and divides the material into two separate parts. Finally, the whole material is sintered by heating furnace, and the physical prototype which is designed by CAD software is obtained.

Digitalized Contour Line Scanning for laser rapid prototyping has high processing speed. There are a large variety of choices of the materials for DCLS, and the products of DCLS have high density and high precision. After making some experiments and analyses, the suitable process parameters were obtained which $P = 30(W)$ and $v = 0.007(m/s)$. The impeller and its mould were designed by CAD software, and the physical prototype of impeller mould was made by DCLS used coated sand, and further, the impeller of metal was cast.

References

1. Choi, S.H., Samavedam, S.: Modelling and optimisation of Rapid Prototyping. Computers in Industry 47, 39–53 (2002)
2. Kruth, J.P., Wang, X., Laoui, T.: Progress in Selective Laser Sintering. Annals of the CIPP 42(3), 21–38 (2001)
3. Youmin, H., Hsiangyao, L.: CAD/CAE/CAM intergration for increasing the accuracy of mask rapid prototyping system. Computers in Industry 56, 442–456 (2005)
4. Xue, Y., Peng, G.: A review of rapid prototyping technologies and systems. Computer-Aided Design 4, 307–318 (1996)
5. Risheng, L., Weijun, L., Xiaofeng, S.: Numerical Simulation of Transient Temperature Field for Laser Direct Metal Shaping. International Federation for Information processing (IFIP), vol. 207, pp. 786–796 (2006)
6. Tang, Y., Fuh, J.Y.H., Loh, H.T.: Direct laser sintering of a silica sand. Materials and Design 4, 623–629 (2003)
7. HongSeok, B., Kwan, L.: Determination of the optimal build direction for different rapid prototyping processes using multi- criterion decision making. Robotics and Computer-Intergrated Manufacturing 22, 69–80 (2006)
8. Yang, H.-J., Hwang, P.-J., Lee, S.-H.: A study on shrinkage compensation of the SLS process by using the Taguchi method. International Journal of Machine Tools & Manufacture 42, 1203–1212 (2002)
9. Williams, J.D., Deckard, C.R.: Advances in modeling the effects of selected parameters on the SLS process. Rapid Prototyping Journal 4(2), 90–100 (1998)
10. Simchi, A., Pohl, H.: Effects of laser sintering processing parameters on the microstructure and densification of iron powder. Materials and Engineering 359, 119–128 (2003)
11. Tolochko, N.K., Arshinov, M.K., Gusarov, A.V., et al.: Mechanisms of selective laser sintering and heat transfer in Ti powder. Rapid Prototyping Journal 9(5), 314–326 (2003)
12. Hanninen, J., Rusko: Direct metal laser sintering. Advanced Materials & Processes 5, 33–36 (2002)
13. Wirtz, H., Freyer, C.: Investment casting shell in 1 day using selective laser sintering (SLS). Foundryman 93(1), 63–65 (2000)

3D Surface Texture Synthesis Using Wavelet Coefficient Fitting

Muwei Jian[1,*], Ningbo Hao[2], Junyu Dong[3], and Rong Jiang[4]

[1] School of Space Science and Physics, Shandong University at Weihai,
108 Wenhuaxi Road, Weihai, China
jianmuwei@gmail.com
[2] International College, Huanghuai University, Zhumadian, Henan, China
[3] Department of Computer Science, Ocean University of China, Qingdao, China
[4] Department of Information Engineering, Weihai Vocational College, Weihai, China

Abstract. Texture synthesis is widely used in virtual reality and computer games and has become one of the most active research areas. Research into texture synthesis is normally concerned with generation of 2D images of texture. However, real-world surface textures comprise rough surface geometry and various reflectance properties. These surface textures are different from 2D still texture as their images can therefore vary dramatically with illumination directions. This paper presents a simple framework for 3D surface texture synthesis. Firstly, we propose a novel 2D texture synthesis algorithm based on wavelet transform that can be efficiently extended to synthesis surface representations in multi-dimensional space. The proposed texture synthesis method can avoid joint seams during synthesis by first fitting wavelet coefficients in the overlap texture images, and then performing an inverse wavelet transform to generate new textures. Then, Photometric Stereo (PS) is used to generate surface gradient and albedo maps from three synthesized surface texture images. The surface gradient maps can be further integrated to produce a surface height map (surface profile). With the albedo and height or gradient maps, new images of a Lambertian surface under arbitrary illuminant directions can be generated. Experiments show that the proposed approach can not only produce 3D surface textures under arbitrary illumination directions, but also have the ability to retain the surface geometry structure.

Keywords: Texture Synthesis, Wavelet Transform, 3D Surface Texture, Photometric Stereo.

1 Introduction

Texture synthesis is an active research area in recent years. Most of recent work on texture synthesis focus on generating a new texture that looks identical with the sample texture and the synthesizing process is like a kind of "growing" process, which means the size of new texture is larger than that of the sample texture and with same pattern,

[*] Corresponding author.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 239–246.
springerlink.com

not simply copy the sample to result texture. Many approaches are based on statistic model or assumption [1, 2, 3, 4, 5, 6]. Zhu at. Al. [3] presents a mathematical definition of texture – the Julesz ensemble, which is the set of all images that share identical statistics. From the statistics, texture can be synthesized by matching statistics has been put on a mathematical foundation. A Markov chain Monte Carlo algorithm is proposed for sampling Julesz ensembles. Portilla and Simoncelli[4] analysis several statistical methods in texture synthesis. They model texture images by a statistical characterisation in the context of an overplete complex wavelet transform. Heeger and Bergen[5] decomposed the input texture by the frequency, and use Laplacian Pyramid and Steerable Pyramid to get different frequency layers. De Bonet[6] uses multi-resolution images to represent the sample texture and synthesise new texture by match a set of features. Efros and Leung [7] present a very simple method, which is under the assumption of Markov random field and assigns the pixel value in synthesised texture by matching the most similar neighbourhood between the sample and synthesised texture. It can produce realistic result in many textures, but the speed is very slow. Wei and Levoy [8] also use multi-resolution images to represent both input sample and output result textures, and from lower resolution to higher resolution, they develop an algorithm to accelerate the synthesis process. Zalesny[9] extracts statistical properties including first order statistics and second-order statistics which draw upon the concurrence principle to form a clique type from the example texture. Textures are synthesized by mimicking the statistics of the example texture for the different clique types. Bar-Joseph [10] represents texture by a hierarchical multi-scale transform of the signal using wavelets. From the tree data structure, new random trees are generated by learning and sampling the conditional probabilities of the paths in the original tree. An inverse transformation is taken to get the new texture. Nevel [11] presents a method that first uses wavelet decomposition on the sample texture and then matches the first (mean and histogram) and the second order statistics (correlation) between the each scale of the sample texture and the synthetic texture. An inverse wavelet transform is taken to get the result synthetic texture. More recently, Xu et al. [12] present a theoretical framework for designing and analyzing texture sampling algorithms by using admissibility, effectiveness, and sampling speed. From the theory, a new texture synthesis algorithm is present by pasting texture patches from the sample texture to synthetic texture.

Research into texture synthesis is normally concerned with generation of 2D images of texture. However, real-world surface textures comprise rough surface geometry and various reflectance properties. These surface textures are different from 2D still texture as their images can therefore vary dramatically with illumination directions. These surface textures are different from 2D texture as their images can therefore vary dramatically with illumination directions. Because the 3D three-dimensional surface texture can display the texture information of the object better than the two-dimensional lamination and can vary with scene illumination and the view angles, it is widely used in virtual reality and computer games. This paper presents a simple framework for 3D surface texture synthesis. Firstly, we propose a novel 2D texture synthesis algorithm based on wavelet transform that can be efficiently extended to synthesize surface representations in multi-dimensional space. In contrast to previous work, we propose a novel fast texture synthesis method that can avoid joint seams during synthesis by first fitting wavelet coefficients in the overlap texture images and then performing an inverse wavelet transform to generate new textures. Then, Photometric Stereo (PS) is used to generate surface gradient and albedo maps from three synthesized surface texture images. The surface gradient maps can be further integrated to produce a surface height map (surface profile). With the albedo and height or gradient maps, new images of a Lambertian surface under arbitrary illuminant directions can be generated.

This paper has two major novelty, one is we propose a novel 2D texture synthesis algorithm based on wavelet transform which can be efficiently extended to synthesise surface representations in multi-dimensional space. The other is we extend texture synthesis from two-dimensional to three-dimensional space.

The rest of the paper is organized as follows. In section 2, wavelet decomposition and novel texture synthesis using wavelet coefficient fitting are briefly described. Section 3 introduces the rendering and relighting the 3D synthesis surface texture. In section 4, we present experimental results. Finally, we conclude the paper in section 5.

2 Novel Texture Synthesis Using Wavelet Coefficient Fitting

2.1 Wavelet Transform

Wavelet transform is a multi-resolution analysis that represents image variations at different scales[13, 14]. A wavelet is an oscillating and attenuated function and its integrals equal to zero. The computation of the wavelet transforms of a 2D signal involves recursive filtering and sub-sampling as shown in Fig.1. At each level, there are three detail images. Following [13], we denote these detail images as LH (containing horizontal information in high frequency), HL (containing vertical information in high frequency), and HH (containing diagonal information in high frequency). The decomposition also produces one approximation image, denoted by LL, which contains the low frequency information. The wavelet transform can recursively decompose the LL band. Since two level wavelet decomposition yields 6 detail images, we use LH1, HL1, HH1, LH2, HL2, HH2, and an additional approximation image LL2 to denote all the subband images.

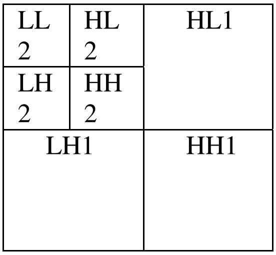


Fig. 1. 2-level wavelet decomposition

2.2 Novel Texture Synthesis Using Wavelet Coefficient Fitting

In allusion to the pixel-based texture synthesis algorithms often damage texture structure and require a large amount of computation, while patch-based algorithms usually produce obvious repetitiveness and seams. This section propose a fast texture synthesis method that can avoid joint seams during synthesis by first fitting wavelet coefficients in the overlap texture images and then performing an inverse wavelet transform to generate new textures. The following paragraph describes the synthesis algorithm in detail.

First, 1-level wavelet transform is performed to the sample textures, after the decomposition obtains four sub-images, respectively denoted by LL1, LH1, HL1 and HH1. The algorithm considerate the neighboring two sub-images that mutually overlaps in the same place part. In order to make the mutually overlap region to unify in together, the overlapping two pieces of the sub-images should be match in the structure. Supposing the overlap part of sample texture is vertical, there are *K* wavelet coefficients in the overlap regions, as Fig.2 shows. Then we carry on the wavelet coefficient fitting in the horizontal direction. Otherwise, if the overlap part is horizontal, then carries on the wavelet coefficient fitting in the vertical direction, respectively.

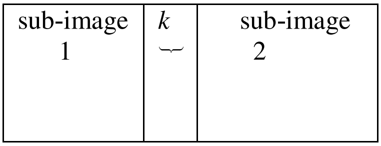


Fig. 2. Vertical overlap part of the sample textures

Here we considered the first kind situation, and the concrete algorithm is described as follows:

Step 1. 1-level wavelet transform is performed to the sample textures, after the decomposition obtains four sub-images, denoted by LL1, LH1, HL1 and HH1, respectively;
Step 2. Set the value *K* of the overlap region. In view of each overlap part of the sub-image, calculates difference of each wavelet coefficient absolute value;
Step 3. Regarding every pixel in the synthesis texture overlap region, calculates the sum of the four sub-images' comprehensive weighting wavelet coefficient's absolute value difference. Each weight of the sub-image is the wavelet coefficient standard difference's reciprocal. In view of each line in the overlap region, select fitting pixel of the wavelet coefficient according to the sum of the comprehensive weighting wavelet coefficient's absolute value.
Step 4. The selection set of the wavelet coefficient's fitting pixel through an error threshold to control its size. The element in the selection set is the minimum value sequence of the sum of the weighting wavelet coefficient's absolute value. Random selecting the element in the selection set in each line as the fitting pixel.
Step 5. After fitting all the wavelet coefficient in the overlap part of the four sub-images, by performing the inverse wavelet transform, we can obtain the new synthesis texture image.

If the overlap part is horizontal, then carries on the texture synthesis algorithm in the (row) vertically direction [20].

3 Rendering and Relighting the 3D Synthesis Surface Texture

3.1 Mathematical Framework of the Gradient Method

The Gradient method is based on the Lambertian reflectance model and uses surface gradient and albedo maps derived from photometric stereo techniques to generate new images under arbitrary illumination directions [1, 19]. The method uses a set of images as input in order to extract surface representations for relighting.

The framework expresses the image data matrix as a product:

$$\mathbf{I} = \mathbf{M}_1\mathbf{M}_2 \tag{1}$$

where $\mathbf{M}_1$ and $\mathbf{M}_2$ are two matrices. $\mathbf{M}_1$ is the surface relighting representation matrix that we want to extract. Thus, if we know $\mathbf{M}_2$ and assume a certain reflectance/lighting model, we can solve $\mathbf{M}_1$ by using SVD. The Gradient method falls into this category.

Thus, the relighting process can be expressed as a product of the surface representation matrix $\mathbf{M}_1$ and a vector **c** related to the required illumination direction:

$$\mathbf{i} = \mathbf{M}_1\mathbf{c} \tag{2}$$

where $\mathbf{i} = (i_1, i_2, \ldots, i_m)^T$ is the image data vector and $i_1, i_2, \ldots, i_m$ are pixel values.

3.2 Linear Combination of Three Synthesized Images from “a”, “b” and “c” Image

Initially we take three images from a texture illuminated at tilt angle of 0°, 90° and 180° respectively by a fixed camera. We name them “a” image, “b” image and “c” image. More details can be seen in [1, 17, 18, 19, 21].

It is a linear combination of “a”, “b” and “c” images. We then can render the synthesized images to produce new textures under arbitrary illuminant directions.

4 Experimental Results

Experiment 1

In order to validate the effectiveness of our scheme, we carried out a large number of experiments to verify the effectiveness of our proposed method.

In our experiments, all the sample texture images are taken from the PhoTex texture Lab [15] and Brodatz Textures database [16]. We synthesize corresponding larger textures from small sample textures using the proposed synthesis scheme.

A variety of experiments are performed to verify the effectiveness of the new texture synthesis algorithm. The experimental results are promising and satisfying. But for the sake of the limited space, only 2 groups of experiment results are illustrated in Figure 3. The experimental results show that this method is fast and can produced promising synthesized results.

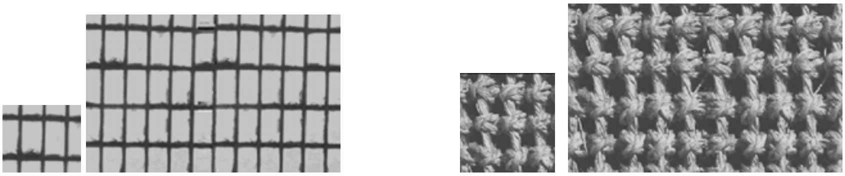


Fig. 3. Texture synthesis results of five groups of real-world textures. (a: sample texture, b: texture synthesis results).

Experiment 2

In order to verify the effectiveness of our proposed framework for 3D surface texture synthesis, many experiments are carried out. In this paper, we experiment on fifteen sets of images taken from the PhoTex texture Lab (http://www.cee.hw.ac.uk/texturelab/database/photex) and six groups of real-world texture images got by ourselves.

Many experiments have directly proved that our scheme can complete the task of 3D surface texture synthesis, and protect the texture's surface geometry structure well. The experimental results are promising and satisfying. Experimental results show that our proposed approach can not only produce 3D surface textures under arbitrary illumination directions, but also have the ability to retain the surface geometry structure. But for the sake of the limited space, for illustration purpose, only two groups of experiment results are illustrated in Figure 4. As illustrated in Figure 4

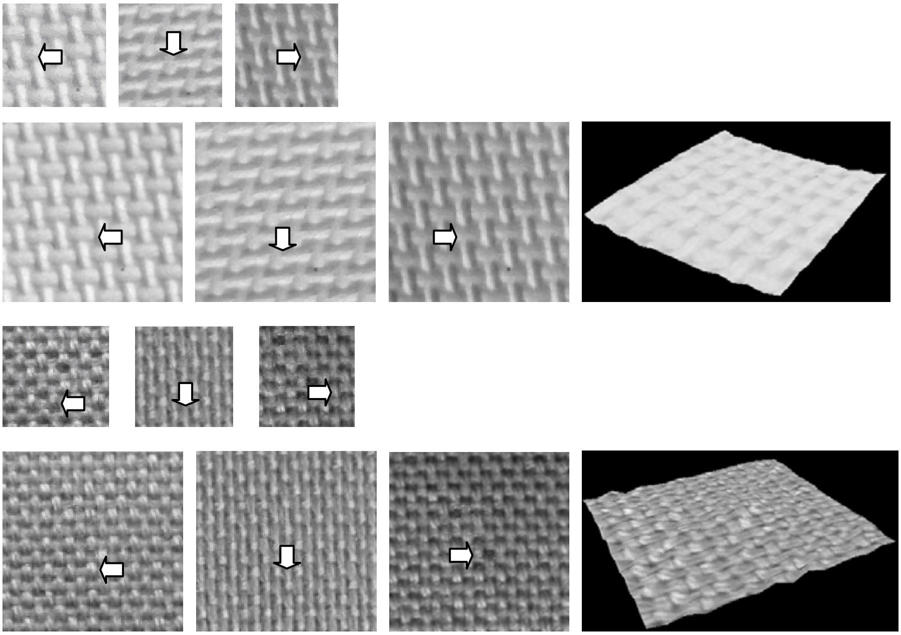


Fig. 4. 3D texture synthesis results of four groups of real-world textures
(The white block arrows indicate illumination directions. slant=45°, tilt=0°, 90°, 180°, 270°)

(The white block arrows indicate illumination directions. slant=45^o, tilt=0^o, 90^o, 180^o, 270^o), it can be seen that the proposed approach can produce promising results.

5 Conclusion

Combined with the Photometric Stereo, this thesis applies the 2D surface texture synthesis method to 3D surface synthesis domain. Firstly, we present in this report a novel 2D texture synthesis algorithm based on wavelet transform, which can be efficiently extended to synthesize surface representations in multi-dimensional space. Then we synthesize corresponding larger textures from small sample textures using the proposed synthesis scheme. Then, Photometric Stereo (PS), as one of the effective technologies for capture of three-dimensional surface texture information, is used to generate surface gradient and albedo maps from three synthesized surface texture images. We use many experiments to directly prove that our scheme can complete the task of 3D surface texture synthesis, and protect the texture's surface geometry structure well. Experimental results show that our proposed approach can not only produce 3D surface textures under arbitrary illumination directions, but also have the ability to retain the surface geometry structure.

In the future, more experiments based on different texture database will be investigated.

References

1. Dong, J., Chantler, M.: Capture and synthesis of 3D surface texture. International Journal of Computer Vision (IJCV) 62(1-2), 177–194 (2005)
2. Dong, J., Chantler, M.: Comparison of five 3D surface texture synthesis methods. In: Proceeding of the 3rd International Workshop on Texture Analysis & Synthesis, Nice, France, October 17 (2003)
3. Zhu, S.C., Liu, X.W., Wu, Y.N.: Exploring texture ensembles by efficient Markov chain Monte Carlo-Toward a "trichromacy" theory of texture. IEEE Transactions on Pattern Analysis & Machine Intelligence 22(6), 554–569 (2000)
4. Portilla, J., Simoncelli, E.P.: A parametric texture model based on joint statistics of complex wavelet coefficients. International Journal of Computer Vision 40(1), 49–71 (2000)
5. Heeger, D.J., Bergen, J.R.: Pyramid-based texture analysis/synthesis. In: Proceedings International Conference on Image Processing (Cat. No.95CB35819), vol. 3, pp. 648–651. IEEE Comput. Soc. Press, Los Alamitos (1995)
6. De Benet, J.S.: Multiresolution sampling procedure for analysis and synthesis of texture images. In: Proceedings of Computer Graphics, SIGGRAPH 1997, pp. 361–368. ACM, New York (1997)
7. Efros, A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proc. ACM Conf. Comp. Graphics (SIGGRAPH), Eugene Fiume, August 2001, pp. 341–346 (2001)
8. Wei, L., Levoy, M.: Fast texture synthesis using tree-structured vector quantization. In: Computer Graphics Proceedings, SIGGRAPH 2000. Conference Proceedings. Annual Conference Series, pp. 479–488. ACM, New York (2000)

9. Zalesny, A., Van Gool, L.: A compact model for viewpoint dependent texture synthesis. In: Pollefeys, M., Van Gool, L., Zisserman, A., Fitzgibbon, A.W. (eds.) SMILE 2000. LNCS, vol. 2018, pp. 123–143. Springer, Heidelberg (2001)
10. Bar-Joseph, Z., El-Yaniv, R., Lischinski, D., Werman, M.: Texture mixing and texture movie synthesis using statistical learning. IEEE Transactions on visualization and computer graphics 7(2), 120–135 (2001)
11. Van Nevel, A.: Texture Synthesis via Matching First and Second Order Statistics of a Wavelet Frame Decomposition. In: Proceedings of the 1998 IEEE International Conference on Image Processing (ICIP 1998), Chicago, Illinois, ICIP (1), pp. 72–76 (1998)
12. Xu, Y., Zhu, S.C., Guo, B., Shum, H.Y.: Asymptotically admissible texture synthesis. In: Proceedings of Second International Workshop of Statistical and Computational Theories of Vision, Vancouver, Canada, pp. 1–22 (2001)
13. Daubechies, I.: The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. on Information Theory 36, 961–1005 (1990)
14. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Analysis and Machine Intelligence 11, 674–693 (1989)
15. Heriot-Watt University, PhoTex texture Lab, `http://www.cee.hw.ac.uk/texturelab/database/photex`
16. Brodatz, P.: Textures: A Photographic Album for Artists & Designers. Dover, New York (1966)
17. Dong, J., Chantler, M.: Estimating Parameters of Illumination models for the synthesis of 3D surface texture. In: Proceedings of the 2004 International Conference on Computer and Information Technology (September 2004)
18. Jian, M.-W., Dong, J.-Y., Wu, J.-H.: Image capture and fusion of 3d surface texture using wavelet transform. In: International Conference on Wavelet Analysis and Pattern Recognition, ICWAPR 2007, November 2-4, vol. 1, pp. 338–343 (2007)
19. Dong, J., Chantler, M.: On the relations between four methods for representing 3D surface textures under multiple illumination directions. In: Proceedings of the 2004 International Conference on Computer and Information Technology (September 2004)
20. Jian, M., Liu, S., Dong, J.: Fast Texture Synthesis Using Wavelet Coefficient Fitting. In: 2008 International Symposium on Intelligent Information Technology Application Workshops, pp. 491–495 (2008)
21. Jian, M., Liu, S., Dong, J.: 3D Surface Texture Synthesis Based on Wavelet Transform. In: International Symposium on Computer Science and Computational Technology, 20-22, vol. 2, pp. 230–233 (2008)

Building Service Oriented Sharing Platform for Emergency Management – An Earthquake Damage Assessment Example

Ying Su[1,2,*], Zhanming Jin[2], and Jie Peng[1]

[1] Institute of Scientific and Technical Information Beijing, China, 100038
[2] School of Economics and Management, Tsinghua University, Beijing, China, 100084
suy.rspc@istic.ac.cn

Abstract. In this paper we study the feasibility of using services offered by a Spatial Data Infrastructure as the basis for distributed service oriented sharing. By developing a prototype we demonstrate that a Spatial Data Infrastructure facilitates rapid development of sharing platform that solves image sharing problems. The prototype provides clients with a distributed application that enables the assessment of earthquake damage areas based on house collapsed data in a given area. We present the architecture of the application and describe details about implementations specific issues. We conclude that the OGC specifications provide a sound basis for developing service oriented architectures for disaster sharing platform. The prototype was validated by enhancing image quality (I-Q) generated by remote sensing through the technologies of visualization.

Keywords: Service Oriented, Emergency Management, Spatial Data Infrastructure, prototype, earthquake damage.

1 Introduction

In recent years the technology trend within information technology has made it possible to move towards service oriented architectures and distributed computing. Service oriented architecture (SOA) is essentially a collection of services, which communicate with each other. The SOA approach also applies within the GIS domain where several initiatives have been launched [1]. This has created a technology evolution that moves from standalone GIS sharing platform towards a more loosely coupled and distributed model based on self-contained, specialized, and interoperable GI services [2].

Emergency management is a multifaceted process aimed at minimizing the social and physical impact of these large-scale events. It is thus not surprising that IT has become a critical tool for facilitating the communications and information-processing activities in managing disasters [3]. Especially important is the role of Web Services and the Service Oriented Architecture (SOA) for computer to computer communication in helping to support this capability [4]. SOA is a conceptual architecture that specifies

[*] Corresponding author.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 247–255.
springerlink.com

interoperable, modular, loosely-coupled, self-contained and self-describing sharing platform, systems or services that interact only at well-defined interfaces [5].

The remaining paper is as follows. Section 2 reviews related work on emergency management. Section 3 describes the proposed architecture of the use case application. Section 4 describes details of the various components and provides details about specific implementations and how they relate to relevant OGC standards. Finally, in Section 5 we conclude on our findings and briefly outline future research topics.

2 Emergency Management

Disasters are normally categorized according to the cause. For example natural disasters are caused by naturally occurring phenomena e.g. earthquakes, landslides, etc. Similarly, technological disasters are caused by design and management failures in technological artifacts. Other categories of disasters can similarly be defined. Nonetheless, disasters share a number of common features:

- Disasters are a threat to life, property and livelihoods
- Disasters are rapid onset events i.e. the time between the moment it becomes apparent that a disaster event is eminent and the onset of the event is rather short
- Disasters occur with intensities that demand emergency response and external intervention
- A greater proportion of the direct loss occasioned by a disaster is suffered within a relatively short time after onset of the disaster event.

Emergency management concerns the organized efforts focused on eliminating or reducing the risk of a disaster and minimizing the impact of the disaster when it happens. The process of emergency management consists of four broad phases; disaster mitigation, disaster preparedness, disaster response, and disaster recovery [3]. These phases can generally be grouped into pre-disaster and post-disaster phases. Pre-disaster phases are disaster mitigation and preparedness and generally concern activities that take place before a disaster event happens. In contrast, post-disaster phases concern activities that take place after a disaster event. The post-disaster phases are disaster response and disaster recovery.

3 Architecture of Prototype

An overview of the components in the overall architecture is depicted on Fig. 1. There are various data sources which need to be accessed via WMS, WFS, and Gazetteer (WFS) services. For the Wenchuan earthquake application we have access to the following thematic data, which are potentially important for house damaged statistics:

- Resources Satellite Application Center, which has created a single corporate spatial data warehouse.
- InfoAgent is an Instant Messaging client facilitating lightweight communication, collaboration, and presence management built on top of the instant messaging protocol Jabber.

A reason for doing statistics on a WFS instead of WMS is that a WFS offers possibilities to access attributes, which in further development of the application could be useful. Of other data relevant for our application are:

- Remote Sensing Image, which provides backdrop satellite images to the application.
- Place names used for locating specific geographic area based on geographic name input.

In addition to the data services, the application comprises a catalog service and two different sharing services (an area statistics service and coordinate transformations service), which are necessary for the application:

- The catalog service provides metadata of the various thematic data services.
- The area statistics service is responsible for calculating the areas affected by earthquake
- The coordinate transformation service is responsible for transforming coordinates into requested coordinate reference system.

The reason for including a coordinate transformation service is that the thematic data services provides data in geographic coordinates, which are not usable for area statistics, hence, the coordinates transformations service transforms the coordinates into projected ones.

3.1 Basic Components

The prototype was developed using ESRI ArcSDE software driven by a multiple clients interface developed in Java. This client interface adapts and integrates the mapping and database technologies required to suit the needs of the proposed assurance model. Fig. 1 illustrates the basic component view of an application scenario.

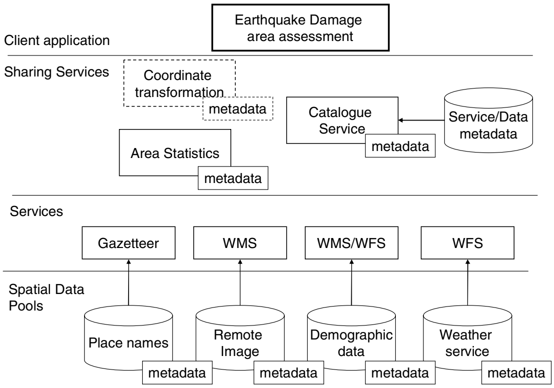


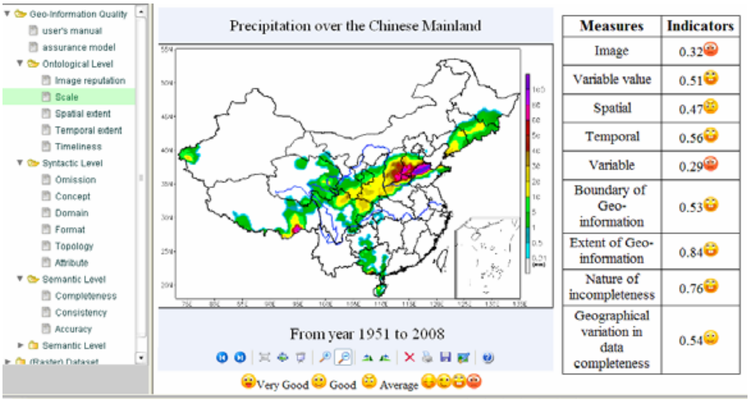
Fig. 1. Basic Components of Prototype Architecture

It is seen here that the client requests the remote sensing image for visualization. Then the area of interest and year is selected in order to get parameters for selecting appropriate data and then, the Image data showing the area of interesting is visualized.

The catalog is used to search and select those data used as source data and target data. In the scenario, the source data is a specific layer of damaged area and the target data could be satellite or aerial photograph data. In theory any type of thematic data layer could be used, but these are the ones we have access to and that are potentially important when assessing earthquake damage. After selection of data a parameter for statistics, which specifies whether statistics should be done per classes or as a whole, is selected by the client.

3.2 Indicator Management

Each indicator definition is stored within the database, including a description of what it represents, the way it is calculated, warnings related to its interpretation, and its importance as defined by the client (expressed in terms of weight), etc. The client can eventually adapt some items further or add more metadata about the indicators. One may select among different graphical representations to illustrate each indicator (e.g. traffic light, smiley, speed meter). (See Fig. 2 and Table 1)



Measures	Indicators
Image	0.32
Variable value	0.51
Spatial	0.47
Temporal	0.56
Variable	0.29
Boundary of Geo-information	0.53
Extent of Geo-information	0.84
Nature of incompleteness	0.76
Geographical variation in data completeness	0.54

Fig. 2. Indicators selection tool with Image with Disaster area

Indicator values are based on the spatial extent of the map being displayed to the quality expert. Indeed, if the client zooms in or pans towards a particular region of interest, quality indicators are recalculated for objects located within the disaster area.

3.3 Navigation into IQ Data

Using the prototype described in the previous section, information quality experts can improve their knowledge of IQ through the use of different navigation tools. Table 1 illustrates the benefits of such a system through different questions a client may have regarding IQ and the different tools offered by the system to help in answering those questions.

Image quality information is communicated through indicators using various color representations (e.g. red, yellow, or green). Quality indicator values can be represented using interactive thematic maps displaying quality values on each feature instance or geometric primitive. Using Geospatial On-Line Analytical Processing (SOLAP) operators, it is then possible to drill the data directly on these maps to access another level of detail of the information.

Based on our previous definitions [6, 7], we define the following quality metrics for a cube C: Accuracy of *C*, measured as $\alpha_C = |L_A|/|L|$; Inaccuracy of *C*, measured as $\beta_C = |L_I|/|L|$; Mismembership of *C*, measured as $\gamma_C = |L_M|/|L|$ and Incompleteness of *C*, measured as $\chi_C = |L_C|/(|L| - |L_M| + |L_C|)$.

Table 1. Map quality problems

Problems	%	Quality problem incidents counted
Ambiguity	89	What is the avarage quality of the image displayed on your screen?
Inaccuracy	25	What image quality characteristics can be a problem according to the task defined for emergency management
Incompleteness	100	Positional accuracy looks problematic, but is it spatally heterogeneous?
Inconsistency	26	Inconsistent formatting or representation of the same elements
Resolution	92	How good is the positional accuracy of this specific object class?
Expert	86	What about the quality of the school building in particular?

From Table 2, the IQ metric values of different layer for assured image were shown in the prototype system. In addition to the levels of detail within the image, this approach also allows clients to explore image quality along a quality indicator hierarchy. For instance, in the example of Fig. 3, a client looks first at the higher-level indicators.

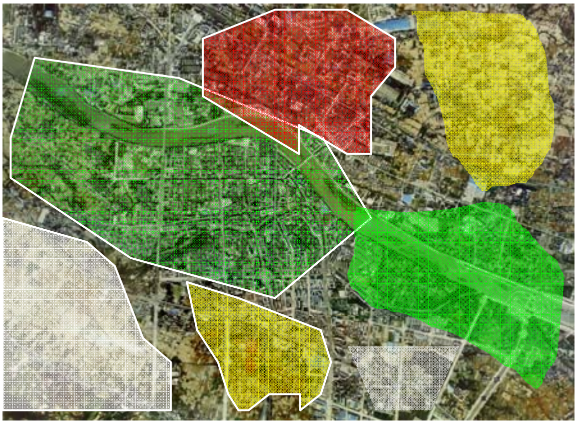


Fig. 3. Different Color Representations for Image Quality in Emergency management

He realizes that ‘General Quality’ is only average (i.e. yellow) because of the lower ‘Internal Quality’. He can then drilldown into the ‘Internal Quality’ to see its sub-indicators At this second level, he can wonder why the ‘Positional Accuracy’ indicator is only average, and then drill-down on ‘Positional Accuracy’ to obtain more detail. He finally arrives at the last level of detail available in our prototype and sees that the problem comes from the ‘Spatial Resolution’. He can then decide if this aspect of image quality is important for his application or not and then decide to either absorb the residual uncertainty or reduce it by, for instance, looking for another dataset.

Table 2. IQ metrics of assured image

Dimensions	Dataset	Band	Attribute Value	Object Instance
accuracy	0.51	0.32	0.54	0.46
		0.19	0.62	0.38
Resolution	0.68	0.66	0.78	0.83
			0.88	0.83
		0.89	0.48	0.94
completeness	0.67	0.53	0.59	0.74
		0.84	0.63	0.56
	0.62	0.76	0.36	0.64
		0.54	0.81	0.39
consistency	0.79	0.86	0.88	0.77
		0.64	0.73	0.66

4 Component Design

In this section the various components in the architecture are presented. More specifically the statistics service, the mapping and feature services, the catalog service and client, and the earthquake client.

4.1 Quality Service

We have developed indicator-based approach for spatial data quality risk and implemented the resulting approach in a geographical information system [8]. To analyze the fitness for use of image for a given area, we designed the IQ evaluation tool such that quality indicators would be displayed on a dashboard embedded within a cartographic interface, acting as a quality service that could support quality experts in the assessment of the fitness of datasets for an emergency management.

4.2 Statistics Service

The interface of the statistics service follows the WPS specification discussion paper (version 0.3.0) which is continuously evolving and is implemented in VB. Net using the ArcObject. A conceptual model of the service and its interface is shown in Fig. 4.

The model is simplified and does not show detailed implementation aspects. The OGC WPS specification specifies three operations as mandatory: *getCapabilities*, *describeProcess*, and execute. The *getCapabilities* operation (which is common for all OGC web services) simply allows clients to retrieve service metadata from the service. The *describeProcess* describes a specific process (operation) that is supported by the specific WPS. We have not (yet) implemented this operation in the statistics service. The process supported by a WPS can be called via the execute operation, which carries out the specific operation requested.

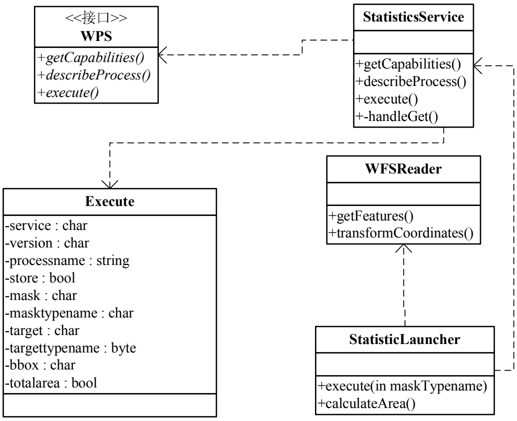


Fig. 4. A model of the statistics service

4.3 Catalog Service

As Catalog we use con terra terraCatalog, which is an implementation of the OGC Web Catalog Service (OGC, 2004a) specification and makes it possible to store and retrieve information about spatial data and services. In particular, this implementation supports the ISO 19115/19119 profile for CSW 2.0 catalog services (OGC, 2005a). In order to access the catalog from the earthquake client (and not the standard con terra client) we used the standard catalog interface to develop a client, which can access metadata stored in the catalog. A simplified model of the client is shown in Fig. 5.

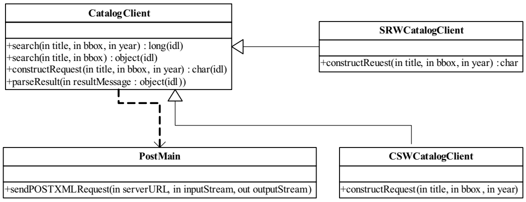

CatalogClient
+search(in title, in bbox, in year) : long(idl)
+search(in title, in bbox) : object(idl)
+constructRequest(in title, in bbox, in year) : char(idl)
+parseResult(in resultMessage : object(idl))
SRWCatalogClient
+constructReuest(in title, in bbox, in year) :char
PostMain
+sendPOSTXMLRequest(in serverURL, in inputStream, out outputStream)
CSWCatalogClient
+constructRequest(in title, in bbox, in year)

Fig. 5. A model of the client for accessing the catalog services

What the client basically offers is a search operation which takes title, bounding box, and year as parameters. The client supports two different protocol bindings using HTTP as transport mechanism (the Z39.50 protocol binding and the Catalog Services for the Web (CSW)). For the communication with the *terraCatalog* we use the

CSWCatalogClient, which implements the CSW protocol binding. The title of the data set and a service URL is returned for the catalog and then the preferred data set can be selected in the forest fire client. The service URL is used as parameter for a request to the statistics server.

4.4 Earthquake Client Application

The client was built using Dynamic HTML (DHTML) and the RedSpider Studio 3 which is a geospatial portal development solution for distributed OGC web services. More specifically, the ASP 'geotag' library allows for easy access to remote services which implement OGC specifications. As depicted on the screenshot in Fig. 2 users can zoom and pan or locate an area via a gazetteer service. Then, after selecting a year and keywords for searching burned areas (only this is shown) and target data, a damage area statistics report is generated (the bottom part in Fig. 2).

5 Conclusion

In this paper we have reported on the development of an application that enables the assessment of earthquake damage areas based on remote sensing data in a given area. We have done this using and implementing important components in an SDI. As stated in the introduction, most SDI initiatives are still in an initial state and just starting to offer geo portals that integrate on-line map viewers and search services for their data. However, we have demonstrated here, that apart from this initial step, SDIs can be used to develop sharing platform solving real problems in a more flexible and scalable manner than ad-hoc and stand-alone sharing platform.

Some studies carried out in the Ministry of Science and Technology's project for the spatial data sharing platform for emergency management that could better express qualitative presentation of data quality. Finally, it is worth mentioning that once quality data is stored in such a structured database with different levels of detail, quality data then becomes easily accessible and can be used to enhance many other aspects of a geo information application.

Acknowledgments. We would like to thank NNSFC (National Natural Science Foundation of China) for supporting Ying Su with a project (70772021, 70831003), and China Postdoctoral Science Foundation (20060400077).

References

1. Fisher, P.F.: Multimedia Reporting of the Results of Natural Resource Surveys. Transactions in GIS 7, 309–324 (2003)
2. Anderson, G., Moreno-Sanchez, R.: Building Web-Based Spatial Information Solutions around Open Specifications and Open Source Software. Transactions in GIS 7, 447–466 (2003)
3. Rao, R.R., Eisenberg, J., Schmitt, T.: Improving Disaster Management: The Role of IT in Mitigation, Preparedness, Response, and Recovery. National Academies Press, Washington (2007)

4. Bahler, L., Caruso, F., Chung, C., Collier, G., et al.: Improving emergency management with Web services, IIIS, Orlando, FL, USA, pp. 162–167 (2004)
5. Shen, Y., Zhang, J., Fan, Y.: Multi-index cooperative mixed strategy for service selection problem in service-oriented architecture. Journal of Computers 3, 69–76 (2008)
6. Su, Y., Peng, J., Jin, Z.: Modeling Information Quality Risk for Data Mining in Data Warehouses. Journal of Human and Ecological Risk Assessment 15, 332–350 (2009)
7. Su, Y., Jin, Z., Peng, J.: Modeling Data Quality for Risk Assessment of GIS. Journal of Southeast University (English Edition) 24, 37–42 (2008)
8. Su, Y., Yang, L., Jin, Z.: Evaluating Spatial Data Quality in GIS Database. In: 2007 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2007, pp. 5967–5970. IEEE Xplore, Shanghai (2007)

An Interactive Intelligent Analysis System in Criminal Investigation

Ping He

Department of Information, Liaoning Police College, Dalian 116036, China
heping2000@163.com

Abstract. On purpose of improving the research in interactive intelligent analysis system (IIAS) when the knowledge in hand is not sufficient, an intuition inversion learning model (IILM) based on experience and knowledge is presented. The paper introduces intuitionistic fuzzy mapping inversion (IFMI) method to the criminal investigation, and poses a skeleton of intuitionistic fuzzy reasoning. Through the relationship construction of practical crime model and on-the-spot model, it sets up a couple of mapping models intuitionistic fuzzy information acquisition. The study shows that the premise of automatic reasoning is to set up patterns of intuitionistic fuzzy relationship. The paper views that the reliability of the automatic reasoning depends on the man-computer interaction results. Simultaneously, choosing the case-cracking clue should be determined by comprehensive evaluations, and self-learning of intuition or fuzzy logical judgments are essentially needed. A simple example on how to create and apply the model is give. The presented model can be applied conveniently by selecting suitable IIAS in accordance with the give intuitive judge and computing the best decision from the rules in those IIAS.

Keywords: intuitionistic fuzzy set, experience mapping, intuitive inversion, relationship construction, interaction intelligent model.

1 Introduction

Intelligent decision support systems (IDSS) is a method of intelligent system design. many IDSS do not provide intuitive analysis capabilities for human, but instead rely on scenario evaluation as a means for developing solutions [1-9]. What kinds of intellectual tasks do we have? Who is more intelligent or smarter: a scientist or a wood-maker (human or machine), a metal-maker or a wood-maker? How to design an intuitionistic fuzzy system with reasoning as the most powerful intellectual function? What is intuition-learning? Can we design system with interaction intelligence model? All these topics are subjects of discussion are research hot spots in recent years [4].

To mimic the problem solving capacity of human being is one of the most basic and important task of artificial intelligence (AI). Such capacity in case solving is neither merely a pure reasoning algorithm, not completely relies on some formatters. The establishment of cooperative relationship can be regarded to be the identification and evaluation of fact inversion and evolution.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 256–263.
springerlink.com

From the research achievements of IDSS of criminal investigation in recent years [3,4,5], the original intention of researchers is that computers can substitute for the intelligence of human beings, thus acquire the decision-making capacity of investigation experts and also overcome the limitation of experts in the field, so that to reach the level of true experts. However, to study the intelligent investigation system (IIS) as an issue in computer science has hampered the system development. No matter in the aspects of knowledge acquisition and expression, or in uncertain reasoning, though great research achievements have been obtained (especially the introduction of artificial neural network and fuzzy system provides many new tools for development of IIS) few successful IIS are available [1, 2]. Many scholars believe that the key to build IIS is the selection and effective use of knowledge [3,4,5]. The "effective use" means whether the rule in the system synchronize with the thinking of the actual users, which is also the difficulty in IIS development. IDSS has achieved great success in formatted reasoning, but in reality, there are too few cases with fixed format. In the research and development of criminal investigation intelligent system [6,7,8,9], it has been found that the formation of specific technique and method comes from the knowledge and experience of people in dealing with routine duties, and this experience and knowledge is nonlinear. In addition, knowledge and experience are different from each other. Do all problems in reality correspond to some complete knowledge? Experiences in the field for different objects are obviously inconsistent. Accordingly, in the research of crime knowledge management [8,9], the first thing is the self-learning of knowledge and experience.

The rest of this paper is organized as follows: The next section provides a brief overview of the underlying method taken for evidence investigation, setting the background for the present research. Section 3 shows the architecture of the proposed interaction analysis system. Section 4 introduces the specific concepts employed and describes how different intuitionistic fuzzy set can be created systematically given evidence and generic domain knowledge. The final section concludes the paper and points out further work.

2 Related Works

2.1 Knowledge Mapping Inversion Principle

In fact, there are two components of intelligence: experience-based intelligence (basic intelligence) that is inherited at birth, and knowledge-based intelligence that can be improved by learning. All kinds of intellectual activities in the specific area are based on knowledge, but intelligence is not knowledge. Knowledge is a "tool" of intelligence. If you don't understand a goal, you are not capable to reach it. An ability to learn is an important intellectual ability that can improve knowledge. Knowledge reinforces intellectual activities.

Knowledge mapping inversion principle (KMIP) refers to a general method or criterion in knowledge discovery [7]. It belongs to a learning principle of knowledge system. See figure 1.

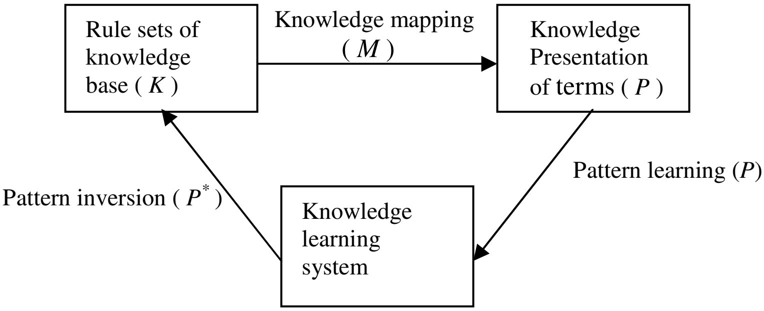

Rule sets of knowledge base (K)
Knowledge mapping (M)
Knowledge Presentation of terms (P)
Pattern learning (P)
Pattern inversion (P^*)
Knowledge learning system

Fig. 1. Knowledge mapping inversion principle

Abstraction of this principle can be described as follows: Let K denotes the rule sets of knowledge base of a group of terms attributes, if M denotes a kind of mapping, then the Knowledge Presentation P of the object attributes can be determined by M, if new knowledge presentation P^* can be decided by learning, then the pattern of attribute can be decided by new knowledge (inverse mapping). This is the basic framework of knowledge mapping inversion (KMI).

2.2 Intuition Learning System

At present, most Intuition learning (IL) [6] is designed manually based on past experience of their intuition. Since the number of possible Intuition is very large for realistic applications of reasonable complexity, heuristics designed manually may not work well when applied in new problem instances. Further, there is no systematic method to evaluate the effectiveness of IL designed manually. For these reasons, an automated method for discovering the proper IL for a particular application is very desirable. This leads to the development of our system for automated learning of intuition.

Definition 1. [8] Let A is an intuitionistic fuzzy set and there is judgment-based experience level (Experience degree) $E_A(x)$, then $E_A(x) = E_A^o(x) + E_A^{no}(x)$, where $E_A^o(x)$ is an experience degree of optimum, $E_A^{no}(x)$ is an experience degree of non-optimum and $E_A(x) \rightarrow [0,1]$. The definition with nature comes to following:

(1) If $E_A(x) = 1$, then with maximum experience degree for intuition judgment that is $E_A^o(x) = 1, E_A^{no}(x) = 0$. Thus, the intuition judgment is optimization.

(2) If $E_A(x) \in (0,1)$, then experience degree (experience level) with interval in (0, 1), for $\forall \theta \in (0,1)$, IDSS with the θ-level of experience judgment.

In the non-optimum analysis of knowledge management [6], we introduced the concept of sub-optimum degree, through which we can describe any factor in the fuzzy system and tell whether it belongs to optimum experience, non-optimum experience or border zones. Meanwhile, the factors belonging to one zone can be put into different layers according to the interval-optimum degree. According to quantitative expressions, we can give the following definition:

Definition 2. [8] Let $E = \{e_1, \cdots, e_n\}$ be a set of experience attribute of optimum in fuzzy system, optimum degree is $\mu(e) = \{\mu(e_1), \cdots, \mu(e_n)\}$, and there be a set $\overline{E} = \{\overline{e}_1, \cdots, \overline{e}_n\}$ of experience attribute of non-optimum, where non-optimum degree is $\mu(\overline{e}) = \{\mu(\overline{e}_1), \cdots, \mu(\overline{e}_n)\}$ then must there be a set $I = \{I_1, I_2, \cdots, I_n\}$ and $I \subseteq E \times \overline{E} = \{(e, \overline{e}) | e \in E \wedge \overline{e} \in \overline{E}\}$. Then I is called the intuitive feature set.

Definition 3. Let I be a set of intuitive feature, There be a $I = \{I(r) | r \in R(e, \overline{e})\}$, where $R(e, \overline{e})$ be a relationship of experience with optimum and non-optimum, $I(r)$ be a intuitive feature index (IFI), and $I(r) \to [0,1]$, without loss of generality, we have

$$I(r) = \begin{cases} \frac{1}{2}[1 + \mu(e)] & \mu(\overline{e}) = 0, \mu(e) \neq 0 \\ \frac{1}{2}[1 + (\mu(e) - \mu(\overline{e}))] & 0 < \mu(e) \neq \mu(\overline{e}) < 1 \\ 0.5 & \mu(e) = \mu(\overline{e}) \end{cases} \quad \begin{array}{l} (0 \leq \mu(e) \leq 1 \\ 0 \leq \mu(\overline{e}) \leq 1) \end{array}$$

The definition with nature comes to following:

(1) If $I(r) = 1$, then with maximum certain degree for intuition judgment, that is with maximum membership degree for intuitionistic fuzzy sets.

(2) If $0 < I(r) < 1$, then a degree of intuition judgment with θ -level, where $\theta \in (0,1)$, and there be a intuitionistic fuzzy set

$$A = \{I_A^1(r), I_A^2(r), \cdots, I_A^n(r)\} = \{\theta_1, \theta_2, \cdots, \theta_n\}$$

based on IFI, thus, the IDM is interval-optimum with intuitionistic fuzzy degree $I(r) \in (0,1)$.

3 Interactive Intelligent Analysis Systems

The key to develop the interactive intelligent system is to make it operable, so that to satisfy the practical needs of real criminal investigation. Accordingly, the application of KMI combined with intuition learning in the IDSS of criminal investigation is realized as follows:

3.1 Establishment of Cooperative Relational Database

Build database of criminal attributes, knowledge rule base of criminal cases and intuitionistic fuzzy relational database of case solving in the computer. These three bases are both independent information sources and interrelated organic whole, which can be called cooperative relational base. Database of criminal attributes selects all the characteristic data sets of the social crime attributes. It is stored in the computer in the form of data warehouse. Knowledge base of criminal cases selects all the characteristic expressions of the social crime. It is written into the attribute base in the form of IF AND THEN. intuitionistic fuzzy relational base of criminal cases selects the occurrences and intuitive processes of solving of all criminal cases. Intuitive reasoning learning selects experience analysis of various cases and it is stored in the computer in the form of human-computer interaction.

The above discussion reveals that KMI principle actually accomplishes a kind of reasoning. And the way of this reasoning should conform to the human thinking. Previous researches on IDSS have made efforts to enable the computer to make decisions like human beings with the assistance of certain algorithms, which has been the goal of research in this field. But the results have been unsatisfying. In the decision making of actual investigation, with certain and limited information, the investigators try to find out the case-solving clues by intuition. Previous works show that intuitive reasoning for decision making is actually a Similarity Inference, which is the repetitive mapping of the nerve stimulus inherent in human brain, and finds the fixed point of the suspect system from judgment of the disordered information by finite self-organization and self-learning. Thus, obtain the ordered objective initial image by self-organizing process of the relationship mapping, that is, build expressions of various relational matrixes, and find the characteristics of criminal attributes from in the scene information, and then find out the range of possibility of the criminal suspect. And determine the criminal suspect according to the additional particular information of the criminal scene.

3.2 Establishment of Intuitionistic Fuzzy Learning System

We use a novel model based intuition reasoning technique, derived from the existing technology of compositional modelling, to automatically generate intuition of crime data from the available evidence. Consistent with existing work on reasoning about evidence the method presented herein employs adductive reasoning. That is, the intuition are modelled as the fuzzy causes of evidence and they are inferred based on the evidence they may have produced.

The goal of the IIAS described in this paper is to find the intuitionistic fuzzy set (IFS) of hypotheses that follow from intuitive experience that support the entire set of available evidence. These IFS of hypotheses can be defined as:

$$H_{IFS} = \{h \in H : \exists s \in S, (\forall e \in E, (S \to e)) \wedge (S \to h)\}$$

where H is the IFS of all hypotheses (e.g. accident or murder, or any other important property of an intuition of case) S is the intuition concept space of crime case), our mini-stories in the example E is the set of all collected pieces of evidence.

The main construction within system of crime intuition relation lies in case cracking: the clues and suspect-to-make-sure, and can think from clues to the suspect type, i.e the constancy of the two mappings: the experience shaping and knowledge mapping.

Firstly, the knowledge mapping to suspect-to-make-sure, secondly, the experience clues to suspects. The experience clue means the conjunction of the intuition information and the similar case information. Thus, the intuition learning system could be founded. Primarily, the following two mappings are to be founded. F_{I}: $S \to K$, F_{II}: $A \to I$. Notes: $S = \{s_1, s_2, \cdots, s_m\}$ is the total collect of the main clues, $s_m (m = 1,2,\cdots,g)$ are the total collect of the specific clues, K ={ $k_1, k_2, \cdots, k_n$} k_j (j=1,2,…,n) is the specific knowledge suspect-to-make-sure, $E = \{e_1, e_2, \cdots, e_u\}$ is the total collect of the experience clues, e_u (u =1,2,…, r) are the intuition clues, $I = \{I_1, I_2, \cdots, I_v\}$ is the total collect of the suspects, I_v {v=1,2,…,l} are the specific suspects. If we represent the total collect with $W = \{K, I\} = \{W_1, W_2, \cdots, W_{g+r}\}$.

When we input a series of clues $W_i \subseteq W$, we accordingly get the output of the function of the two mappings: suspect-to-make-sure $k_j \in K$ and suspect-to-make-sure $I_j \in I$. Before making sure the two mappings' characteristics, we should divide the statistical clue group into main clue group and experience group.

When recognize a crime, the given information may be not unanimous, so as computer may be found difficult to tell it apart. To solve the problem and enable the computer to reason automatically and find the most valuable clues, then use the man-computer system to go on with the judgment. Automatic checking function is as follows. When inputting a series of clues E_0={e_1,e_2, ⋯,e_q}, if the computer reads $I(r) =$ max { $I(A_v), j = 1,2,\cdots,l\}$, let

$$\varepsilon = \frac{I(e_0) - I(k_0)}{I(r)} (i \neq j)$$

$$\theta(e_0, I) = \left| e_0 - \frac{I(e_0) + I(k_0)}{2} \right| - \frac{I(e_0) - I(k_0)}{2}$$

If $\varepsilon < \theta$ should be adjusted, here θ is the experience region value, $0 < \theta < 1$ [3,6].

4 An Examples

A scene of a specific criminal case will show us the implementing process of building interactive intelligent analysis system by the previous investigative reasoning method (See ref. 4, 9).

The [Ref. 4, 9] dialogue is a combination of knowledge reasoning and experiential reasoning, namely, cooperative reasoning, derived from medico legal knowledge and social experience.

An integration of the output information indicates that the murderer might have an amour with or be in love with the deceased.

The interactive intelligent analysis system gives clues to solve the case as follows:

- To investigate whether the handwriting on the note in the pocket is the victim's.
- To investigate who had a close relationship with the victim before her death.
- To learn when the victim was murdered and where those closely related with the victim has gone.
- To check whether the handwriting on the note was that of one of her acquaintances.

Investigation of this case reveals, eventually, that Wu (a married man) in the same community as the victim had been in an intimate relationship with the victim. The writing on the note was his handwriting. And his whereabouts around the time of the murder was unknown. Therefore, Wu was considered as a prime suspect of this crime. And it was finally confirmed that Wu had an adultery affair with Guo (the victim), and because of failure in abortion, he killed Guo to cover up the affair.

The above analysis shows that the cooperative reasoning in the IIAS can obtain the effect of the investigative expert in solving a crime, which indicates that the reasoning process in the IIAS of investigation conforms to reality.

5 Conclusion and Future Work

This paper discusses the critical issues in establishment of interactive intelligent analysis system that should be paid attention to through practice of criminal investigation work. The development of the interactive intelligent analysis system must be grounded on identification, otherwise this work is of little significance or value. Simultaneously, the knowledge reasoning should be distinguished from experiential and intuition reasoning. For different cases, experiential reasoning is variable. Only by combining the two together with intuition to reach cooperative reasoning can they possibly play their roles in reality. Besides, the operation mechanism of the IIAS should apply the DMI principle, a very useful intellectual system, which is certain to play a guiding role in the development of automatic reasoning computer.

In future work, the method presented here will be expanded upon. Firstly, the representation formalisms employed to describe states and events in intuitive process of criminal investigation will be elaborated. As described earlier, the intuitionistic fuzzy set of states and events that constitute a scenario are restricted by the consistency requirements. This paper introduced a generic means to represent when inconsistencies occur and to prevent inconsistent experience and knowledge from being considered when hypotheses are generated and evidence collection strategies are constructed. When reasoning about related events that take place over experience and intuition, the experience process of the intuition are an important source of such inconsistencies. To avoid overcomplicating this paper, the important issues of knowledge and intuition reasoning were not considered, but will be addressed in future work. Secondly, methods are under development to assess the relative likelihoods of alternative learning system. Several methods to expand the intuition entropy based decision making techniques employed by model based intuitionistic fuzzy diagnosis techniques have been presented in other papers. The application of these methods requires a means of generating an intuition concept space and a way of computing the relative intuitive degree of the experience. Thirdly, an extensive knowledge base will be developed to enable the deployment of IIAS. Currently, a prototype implementing the algorithms described here has been developed. This has enabled the validation of the theory and the example used in this paper. However, it is clear that a proper evaluation of the approach requires its application to a real-world domain problem.

Acknowledgments

I gratefully acknowledge and thank Professor Liu Wei and Dr. He Jiantong for their guidance throughout this project, along with Liu and He Jiantong for their reviews. This work was completed with the support of a research grant from the Ministry of Public Security.

References

1. He, P.: Crime Pattern Discovery and Fuzzy Information Analysis Based on Optimal Intuition Decision Making. Advances in Soft Computing of Springer 54(1), 426–439 (2008)
2. He, P.: The Learning System of Intuition Optimum Based on Hesitancy Set. In: Shi, Y. (ed.) Third International Conference on Innovative Computing Information and Control, pp. 578–582. IEEE Computer Society, Los Alamitos (2008)
3. He, P.: Crime Knowledge Management Approach Based on Intuition Concept Space. In: Zhou, Q. (ed.) Intelligent Information Technology Application, pp. 276–279. IEEE Computer Society, Los Alamitos (2008)
4. Qu, Z., He, P.: Interactive Intelligent Analysis Method: An Application of Criminal Investigation. In: Shi, Y. (ed.) International Symposium on Intelligent Ubiquitous Computing and Education, pp. 578–582. IEEE Computer Society, Los Alamitos (2009)
5. Li, J., He, P.: Extended Automatic Reasoning of Criminal Investigation. In: Shi, Y. (ed.) International Conference on Industrial Mechatronics and Automation, pp. 356–359. IEEE Computer Society, Los Alamitos (2009)
6. He, J., He, P.: Fuzzy Relationship Mapping and Intuition Inversion: A Computer Intuition Inference Model. In: IEEE International Conference on MultiMedia and Information Technology, pp. 298–301 (2008)
7. He, P.: Fuzzy Relationship Mode Mapping Inversion and Automatic Reasoning of Crime Detective. Journal of Pattern Recognition and Artificial Intelligence 16(1), 70–75 (2003)
8. He, P.: Intelligence Theory and Practice Means of Criminal Investigation. Journal of Liaoning Police Academy 17(3), 1–6 (2001)

Research on Functional Modules of Gene Regulatory Network

Honglin Xu* and Shitong Wang

Information Engineer School, Jiangnan University, Wuxi, Jiangsu, China, 214222
Xuhonglin_9@hotmail.com

Abstract. This paper reports the research development on modular organization of gene regulatory network. Gene regulatory network is an efficient measure to describe correlation and inference between genes in a dynamical and systematic way. Considering of the large number of genes, the complexity of model's structure and the practical significances of biological research, we discuss the necessity and positive meaning of analyzing and finding of the functional modules of gene regulatory. We take the Probability Boolean network as an example; decompose the PBN in a topological method, so as to confirm the feasibility of the issue: the gene regulatory network can be decomposed into independent functional models.

Keywords: Gene Regulatory Network, Boolean Network, Gene Functional Module.

1 Introduction

In essence, life activity is a complex system contains thousands of genes, proteins and other chemical signal molecules, which interact and organize with each other forming constantly changing life vital activity phenomenon. To understand the nature of cellular function, it is necessary to study the behavior of genes in a holistic rather than in an individual manner, because the expressions and activities of genes are not independent of each other [1]. Therefore it is more pertinent and practical to put forth effort to study gene regulatory mechanism for reveal the formation and function rather than gene itself.

Recently, a significant amount of attention has been focused on the inference or identification of the model structure from gene expression data. Gene regulatory networks (GRN) [2] modulate the action of metabolic networks, leading to physiological and morphological changes. The main purpose of research on genetic regulatory network is to analyze the mechanism of the birth, grow function, death of genes. In order to simulate the biology adjustment of most reality, choosing an appropriate model is a quite critical. So far, several kinds of genes regulatory network models have been proposed, including linear model, Bayesian network, neural network, differential equations, Boolean network, etc.

* Corresponding author.

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 264–271.
springerlink.com

The model system that has received the most attention may be the Boolean network originally introduced by Kauffman [3]. In an easy Boolean gene regulatory network, it is simple to find that some special status can be frequently got, which represents the stable state of the cell cycle. However, in a usual network model like Probability Boolean network [4] contains n genes, there are 2^n initial states, it is an obviously NP problem to analyze the attractors for all these states for the high computational complexity.

Therefore an approach is proposed in this paper to analyze the function modules partition methods. Take the Probability Boolean Network as an example; our method includes the following steps: abstractive the PBN to be a weighted digraph; find essential genes from the network; take these genes for the core decomposing the PBN.

2 Discussion on Functional Modules of Gene Regulatory Network

Models of gene regulatory networks mentioned above reveal correlations between the complex network structure and biological functions of cells, so a lot of implicit information biological related can be mined. Except for stimulating genomic biological activities, exploring the organization mechanism of each signal molecular or micelle is another important aim of gene regulatory network research.

For example, in a dynamic gene regulatory expression network, we often describe the status of genes in a binary logical language such as “on and off”, then quantized to two levels (1 or 0), Spell man etc. [11] constructed a dynamical evolution networks to stimulate the cell cycle. During the regulating process, they found the network status converge to some attractors. Just like in the simple Escherichia coli gene regulatory network, we observed that network finally get to a certain state, and meanwhile, the topological structure of the network is determined and brief. This phenomenon illustrates the stability of gene regulatory network under cell cycle. What’s more, it means gene regulatory network can describe biology mechanism in topological structure.

However, for networks with large number of genes (such as the yeast transcriptional regulatory network), calculating complexity of models mentioned above is too high to discuss the structure of the network simply (show as Fig.1). In complex networks, if we say the dynamical analysis of network system is the horizontal study, then we can consider the research on function modules division as longitudinal study of gene regulatory network [10]. Like Ernst and other researchers pointed out that, view in Artificial Intelligence, an obviously copy behavior between modules is a biological innovation, research on decomposing of network into functional has a very important significance for the understanding of biological regulation and evolution.

Some biological researcher consist that functional and evolutionary inference in gene networks do no matter with the topological structure [8]. Actually, we can get the functional modules of gene regulatory network by graph theory method and mathematic inference. A recent explosion of research papers related to the structure of complex networks has led to important results related to the topological properties of biological networks [1]. These networks, which include metabolic networks and protein interaction networks (PINs) [4], share important structural features with other real-world networks in disparate fields ranging from the Internet to social networks. According to gene expression data already know and mathematical theory, we can complete the extraction of functional modules; realize the search for genes and proteins with similar functions or high influence.

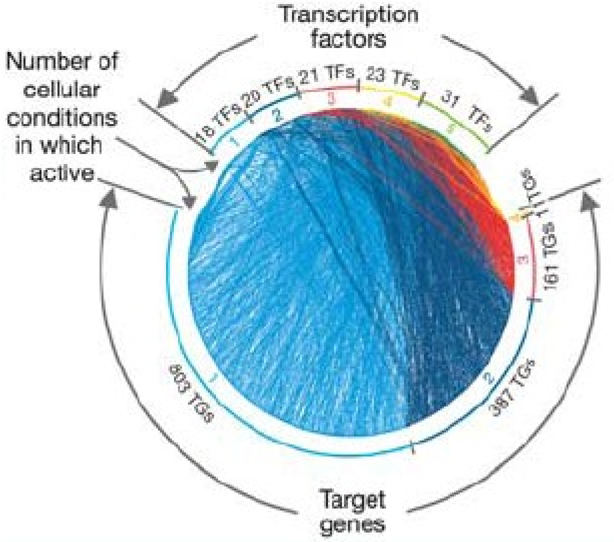


Fig. 1. The yeast transcriptional regulatory network

3 Probability Boolean Network

The recently introduced model Probability Boolean Network (PBN) [1] attracts more and more attention due to its distinctive characteristics. The PBN is based on the Boolean Network by adding probability dynamic characteristic. According to [1], structure of PBN is defined as follow: A deterministic gene regulatory network $G(V,F)$ consists of a set $V=\{x_1,x_2\cdots x_n\}$ of nodes representing genes and a list $F=(f_1,\cdots,f_n)$ of functions, where a function $f_i(x_{i_1},\cdots x_{i_k})$ with inputs from specified nodes $x_{i_1},\cdots x_{i_k}$ is assigned to each node x_i. In general, each node takes value from a finite set of values. Takes a binary example, $x_i\in\{0,1\}$, $i=0,1,\cdots n$. $x_i=1$ means this gene x_i is expressed, $x_i=0$ means not be expressed. The state of pattern of the network is composed of the values of all nodes. The expression pattern ψ_{t+1} at time $t+1$ is determined by the function in F from the expression pattern ψ_t at time t. To every, node x_i, there corresponds a set

$$F_i=\{f_k^{(i)}\}_{k=1,\cdots,l(1)}\ ,\ i=1,2,\cdots,n\ . \tag{1}$$

Where each $f_k^{(i)}$ is a possible function determining the value of gene x_i, $l(i)$ is the number of possible functions of gene x_i. These functions $\{f_k^{(i)}\}$ are also referred to as predictions.

If there are N possible realizations, then there are N vector functions $f_1, f_2, \cdots, f_N$ of the form $f_j = (f_{j_1}^{(1)}, f_{j_2}^{(2)}, \cdots f_{j_n}^{(n)})$, for $j = 1,2,\cdots,N$, $1 \le j_i \le l(i)$ and where $f_{j_i}^{(n)} \in F_i$ $(i = 1,2,\cdots n)$. In other words, the vector function f_j acts as a transition function represents a possible realization of the network. Let $f = (f^{(1)}, f^{(2)}, \cdots f^{l})$ be a random vector taking values from $F_1 \times F_2 \times \cdots F_n$. That is, f can take all possible realization of the network. Then the probability that predictor $f_k^{(i)}$ $1 \le k \le l(i)$, is used to predict gene i is:

$$c_k^{(i)} = P\left\{f^{(i)} = f_k^{(i)}\right\} = \sum_{j: f_{ji}^{(i)} = f_k^{(i)}} P\{f = f_j\} \,. \tag{2}$$

Obviously, $\sum_{k=1}^{l(i)} c_k^{(i)} = 1$.

According to the regulate and predict regulation between genes, mutual influence of genes can be calculated by following method, it is important to distinguish those genes that have a major impact on the predictor from those that have a minor impact. This method is proposed in [3].

Influence $I_j(f)$ of variable x_j on the function f with respect to the probability distribution $D(x)$, $x \in \{0,1\}^n$ can be defined as following

$$I_j(f) = E_D[\frac{\partial f(x)}{\partial x_j}] \tag{3}$$

Where $E_D[\bullet]$ is the expectation operator with respect to distribution D $\left({\partial f(x)} \big/ {\partial x_j} \right) = f(x^{(j,0)}) \oplus f\left(x^{(j,1)}\right)$ is the partial derivative of the Boolean function, symbol $\oplus$ means addition modulo 2, and $x^{(j,k)} = (x_1, x_2, \cdots, x_{j-1}, k, x_{j+1}, \cdots, x_n)$, $k = 0,1$. Simply, (3) gives the influence of the *j*th variable changes the value of function as the probability. So in the context of PBN, the influence of x_k on gene x_i is given as following:

$$I_k(x_i) = \sum_{j=1}^{l(i)} I_k(f_j^{(i)}) \cdot c_j^{(i)} \ . \tag{4}$$

Where $f_j^{(i)}$, $j = 1,\cdots,l(i)$, are the all possible predictions of gene x_i, The influence matrix Γ contains every pair of genes: $\Gamma_{ij} = I_i(x_j)$.

By calculating the mutual influence between genes, combined with the regulatory of network, we can consider the influence matrix Γ as an adjacent matrix of a digraph with weight. In other words, compute on genetic regulatory networks can be turned into a simple way: there is a digraph $G(X,F)$, has nodes set $X = \{x_1, x_2, \cdots x_n\}$ corresponding to the genes of the PBN $V = \{v_1, v_2, \cdots v_n\}$, with the adjacent matrix Γ, F is the prediction function. In this way, the PBN can be simply expressed with a digraph, which is profile to be divided into modules represent deferent functions.

4 Functional Modules of PBN

We define these following genes as “essential genes” according to the above paragraphs in two ways:

(1) Combined with the knowledge of medicine and biology we already known, basing on the research proposal, we can define some special genes as essential gene, it’s easy and straightforward in virtual experiments.

(2) Starting from the genetic regulatory network, essential gene can be inferred from the structure theoretically. By computing the influence input and output of each gene in the digraph, genes with strong effect and sensitivity can be found, which makes the hinge of the regulatory in the context of the network.

In this paper, we mainly introduce the second method. According to the influence matrix of genetic regulatory network, input and output influence of each gene can be defined as Effect $E(i)$ and Sensitivity $S(i)$, We let

$$E(i) = Output\ \ Influence\ \ of\ \ x_i = \sum_{j=1,2,\cdots n}^{i \neq j} a_{ij}\ . \tag{5}$$

$$S(i) = Input\ \ \ Influence\ of\, x_i = \sum_{k=1,2,\cdots n}^{i \neq k} a_{ki}. \tag{6}$$

Given the Effect $E(i)$ and Sensitivity $S(i)$, we define a gene Y as essential gene satisfying

$$E(i) > kS(i)\ . \tag{7}$$

And

$$Y = \arg\ \max\,(\alpha \cdot I_{\mathrm{i}} + \beta \cdot S_{\mathrm{i}})\ . \tag{8}$$

Where k , α , β are user-chosen parameters. By adjusting k , genes of high effect and low sensitivity are easy to be found. In the application shown in this paper, we let $\alpha = \beta = 1$. This is an easy and direct way to calculate the genes with strong interaction on neighbors, to achieve network decompose, we mark the genes off two kinds: essential and ordinary. Considering characterize of genetic regulation: a gene only haves influence on several adjacent genes, we set these essential genes as Source S, and the ordinary genes as Destination D in the following steps,, using an searching Dijkstra's algorithm to find the shortest path between them. Particular algorithm is given in [5].

The interaction between genes descend by the passing along of the path, only shortest path could not describe the mutual influence of genes, so we calculate the average information passing between the source nodes and the destination nodes along the shortest path, which is given by:

$$\mathrm{C}_{sd} = \begin{cases} 0 & \text{when } \mathrm{d_{sd}} = 0 \\ a_{sd} & \text{when } \mathrm{d_{sd}} = 1 \\ C_{sd_{i-1}} + {}^{1}\!/_{d_{sd}} \cdot a_{d_{(i-1)}d_i} & \text{when } \mathrm{d_{sd}} > 1 \end{cases} . \tag{9}$$

In (9), we use the multiplicative inverse of the distance d_{sd} as the weight to increase the information passing alone the shortest path.

Basing on the front analyze of genes interaction and the influence trace, we have the follow algorithm to decompose the PBN as model of genetic regulatory network into functional modules. If M is the number of modules maybe contained:

(1) Calculate the essential genes $\mathrm{Y} = \{y_k \mid 1 \le k \le M\}$ by using (7), mark these genes as Sources, and the left genes be Destination.

(2) Start with node from Y , search the shortest path to all destination nodes; record the trace in an array $\mathrm{Y}_k[\,]$.

(3) Using (8), along with the path from S to D, calculate the average information passing, form the passing information matrix $C_{s\times d}$.

(4) Under the same distance from S to D, gene can be divided to the functional modules to which that essential gene with the maximum passing information.

5 Example

To illustrate the algorithm proposed in the front, consider the influence-based network in Figure 2, this network has resulted from an artificial PBN constructed specifically to produce functional modules [3]. Using the method we can calculate the essential gene x_5 , Start from gene x_5 , shortest path to leaving genes can be found as following: when $d = 1$, result obviously given; when $d = 2$, shortest path with maximum information passing can be calculated by (8), the trace is given as

5→0→3,5→0→1. Comparing the algorithm with [3], the complexity of computing is obviously reduced.

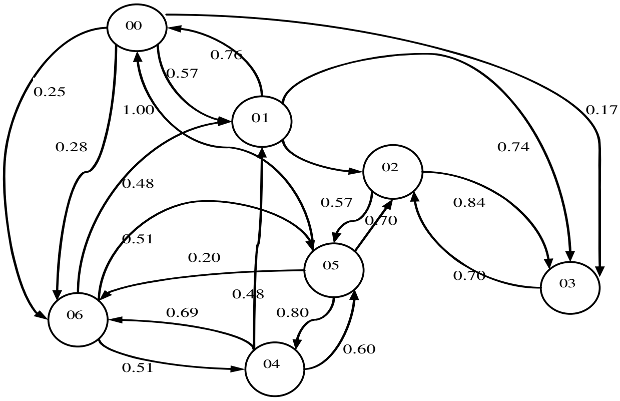


Fig. 2. Artificial influence –based network

6 Conclusion

Gene regulatory network is one of the basic methods for further understanding the mechanism of gene regulation. Decomposition of the large-scale gene regulatory networks into numbers of functional modules, helps mining more biological information from the microscopic. Our discussion and example of PBN confirm that the gene regulatory networks can be analyzed by the topological structure.

In the future work, combining with biological process and molecular reactivity, we will do further study in research the biological significance of gene functional modules. Based on the given definition of network modules, given evaluation criterion of module partition method, organization manners of gene regulatory networks can be through studied. For example, artificial intelligence methods can be used for calculate the cohesion and coupling of modules. In conclusion, analysis on modules' organization manner of gene regulatory network has great theoretical and practical signification.

Acknowledgments. This work is partially supported by the Hong Kong Poly-technic University Grant (Grant no. Z-08R), National 973 Key Project (Grant no. 2006CB705700), National Science Foundation of China (Grant nos. 60773206/F010106 and 60704047/F030304), New century Outstanding Young Scholar Grant pf Ministry of Education of China (Grant no. NCET-04-04960, National KeySoft Laboratory at Nanjing University, the Key Laboratory of Computer Science at Institute of Software, CAS, China.

References

1. Zhou, X., Wang, X., Dougherty, E.R.: Construction of genomic networks using mutual-information clustering and reversible-jump markov-chain-monte-carlo predictor design. J. Signal Process., 745–761 (2003)

2. Shmulevich, I., Zhang, W.: Binary analysis and optimization-based normalization of genes expression data. Bioinformatics, 555–565 (2004)
3. Ronaldo, F., Hashimoto, Kim, S., et al.: Growing genetic regulatory network from seed genes. Bioinformatics, 1241–1247 (2004)
4. Shmulevich, I., Zhan, W.: From Boolean to probabilistic Boolean Networks as model of genetic regulatory network. Proceedings of the IEEE, 1778–1791 (2002)
5. McGraw-Hill: Introduction to Algorithms, 2nd edn. Section 24.3: Dijkstra's algorithm, pp. 595–601. MIT Press, Cambridge (2001)
6. Kauffmans, S.A.: The Origins of Order, Self-Organization and Selection in Evolution. Oxford University Press, Oxford (1993)
7. Wang, P.: A study of 3-genes regulation networks using NK-Boolean network model and fuzzy logic network. Studies in Fuzziness and Soft Computing. Springer, Heidelberg (2006)
8. Siegal, M.L., Promislow, D.E., Bergman, A.: Functional and evolutionary inference in gene networks: does topology matter? J. Genetica 129(1), 83–103 (2006)
9. Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., Gerstein, M.: Genomic analysis of regulatory network dynamics reveals large topological changes. Nature, 308–312 (2004)
10. Wang, Z.-H., Liu, Q.-J., Zhu, Y.-P.: Research on modular organization of gene regulatory network. Hereditas (Beijing) 30(1), 20–27 (2008)
11. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., Friedman, N.: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetic 34(2), 166–176 (2003)

Block-Based Normalized-Cut Algorithm for Image Segmentation

Haiyu Song[1,2], Xiongfei Li[1], Pengjie Wang[2,3], and Jingrun Chen[2]

[1] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China
[2] Institute of Computer Graphics and Image Processing, Dalian Nationalities University, Dalian, China
[3] State Key Laboratory of CAD & CG, Zhejiang University, Hangzhou, China
songhaiyu@gmail.com

Abstract. Automatic image segmentation remains a challenging problem in the fields of computer vision, image analysis and understanding. A lot of algorithms and technologies have been proposed and developed for image segmentation, among which the N-cut has been a promising method and active research area due to better quality. But the N-cut can't process medium and large image dataset online due to incredible memory and time complexity. Moreover, the N-cut algorithm can't utilize texture information since its basic unit of clustering algorithm is pixel. To overcome the drawback and improve the description power, the paper proposes a block-based N-cut algorithm for image segmentation, whose basic unit of clustering is image block instead of individual pixel. Block-based algorithm not only decreases the time and storage complexity, but also improves the discrimination power of visual feature vector. The experiments demonstrate that proposed algorithm outperform the original N-cut algorithm in efficiency. Experiments show that the proposed block-based N-cut algorithm is more suitable for online processing large-scale data.

Keywords: Normalized-cut, Block-based, image segmentation, graph partition.

1 Introduction

Image segmentation is a key step to extract shape information and region information in the field of image analysis and understanding. Traditionally, segmentation is a preprocessing step for object recognition. While object recognition is a difficult task, every algorithm of object recognition always has strictly limited domain. Today, image segmentation is widely applied to image retrieval, image annotation, image analysis and understanding. The goal of image segmentation is to cluster pixels into image regions, which always corresponds to individual surfaces, objects, blobs, or natural parts of objects. In image retrieval, region-based signature and retrieval has been an active research area[1]. Image segmentation is a key step to acquire a region-based signature. Shape signature or shape similarity is meaningless without reliable segmentation. There are many kinds of segmentation approaches, such as Canny operator, local variation algorithm, edge-augmented mean-shift, k-means clustering[2][3]. All operators such as

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 272–278.
springerlink.com

Canny operator utilize local information to filter, which couldn't ensure a continuous closure border. To construct closure border shape signature, the most widely used segmentation approach is k-means clustering, whose advantage is high speed.

Cut segmentation is a new advance in this field, which is an application of spectral clustering to image segmentation. The algorithm treats image segmentation as a graph partition problem. The Cut algorithm maps image segmentation problem to graph partition. Any image can be represented as a weighted undirected graph G= (V, E), where the nodes of the graph are the points in the feature space, and an edge is formed between every pair of nodes. A graph G=(V, E) can be partitioned into two disjoint sets by simply removing edges connecting the two parts. The degree of dissimilarity between these two pieces can be computed as total weight of the edges that have been removed. In graph theoretic language, it is called the cut:

$$cut(A,B)=\sum_{u\in A,v\in B} w(u,v) \cdot \tag{1}$$

The optimal partitioning of a graph is the one that minimizes the cut value. The minimum cut criteria favors cutting small sets of isolated nodes in the graph [4] [5]. To avoid unnatural bias for partitioning out small sets of points, Jianbo Shi proposed a new measure of disassociation between two groups [6]. Instead of looking at the value of total edge weight connecting the two partitions, the measure computes the cut cost as a fraction of the total edge connections to all the nodes in the graph, and the disassociation measure was called the Normalized Cut (N-cut):

$$N-cut(A,B)=\frac{cut(A,B)}{assoc(A,V)}+\frac{cut(A,B)}{assoc(B,V)} \cdot \tag{2}$$

where assoc(A, V)= $\sum_{u\in A,t\in V} w(u,t)$ is the total connection from nodes in A to all nodes in the graph and assoc(B, V)is similarly defined. Compared with previous clustering segmentation algorithms, the N-cut aims at extracting the global impression of an image, and measuring both the total dissimilarity between the different groups as well as the total similarity within the groups. Because N-cut comprehensively considers global and local information as well as its robustness, it can achieve better performance than previous clustering algorithms for image segmentation. Currently, spectrum clustering liking N-cut has been an active research topic [7] [8] [9] [10].

2 Block-Based Normalized Cut

Every vertex of graph is corresponding to a pixel of image. The N-Cut looks like considerably appealing because it considers region segmentation from global perspective. It can be used to pattern recognition and other domain, but it is infeasible for online image retrieval due to memory and time complexity as considering medium and large datasets. When an image is mapped to a graph, the graph always is described by matrix, whose element number is (M*N)*(M*N) if the image size is

M*N. As a result, the incredible computational cost makes it lost usefulness. Moreover, the experiment results show that original N-cut algorithm is suitable for non-textured image instead of textured image, because the clustering basic unit is individual pixel that can't contain texture information.

To improve speed, we propose block-based normalized cut based on original normalized cut image segmentation algorithm. To improve performance, we propose combine texture and color as visual feature of basic unit for N-cut. To construct discriminative texture feature, we adapt weighted centroid pixel gray value as one of the region feature values by convolution with normal distribution for textured feature.

We propose an improved N-Cut algorithm, which is not based on single pixel but 7*7 pixels image block when image problem converted to graph. In the proposed N-Cut algorithm, each node of graph is corresponded to an image block instead of an individual pixel. The weight on each edge, w_{ij} between node i and j of adjacency matrix W, is a product of a feature similarity term and spatial proximity term: $w_{ij}=\exp\{-\frac{\|F(i)-F(j)\|_2^2}{\delta_I}\}*\exp\{-\frac{\|X(i)-X(j)\|_2^2}{\delta_X}\}$, if $\|X(i)-X(j)\|_2 <r$; otherwise, $w_{ij}=0$, where X(i) is the spatial location of node i, and F(i) is a feature vector based color information at the node i. When the formula is used in the original N-cut algorithm, the F(i) always represents the intensity value of corresponding pixel. While when it applied in our block-based N-cut, the F(i) represents a visual feature vector of corresponding image block. Because the N-cut algorithm is clustering of adjacent pixels, the weight of node i and j is 0 when the distance of two nodes is larger than the threshold r. We select the color feature as feature vector of image block, due to the block is homogeneous through N-cut segmentation.

Our proposed block-based N-cut algorithm is as follows.

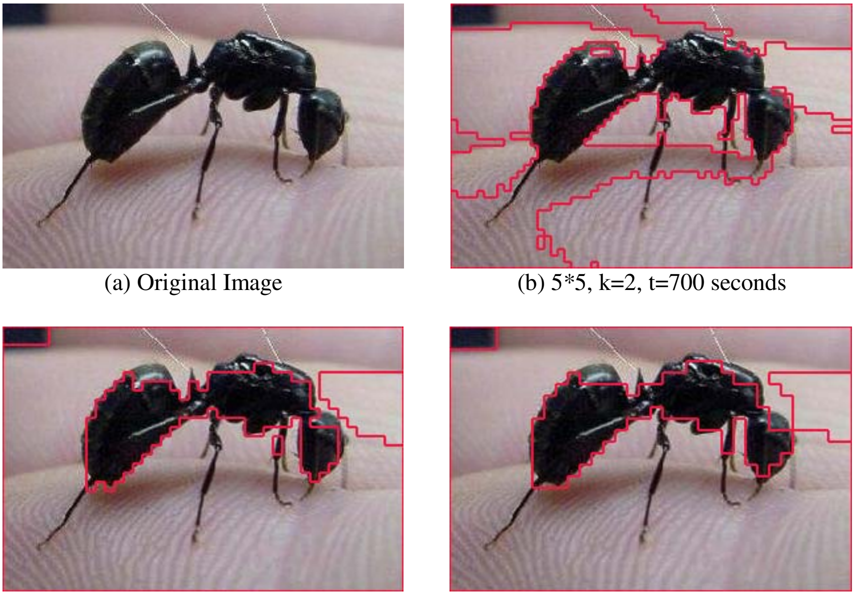
1. Segment image into several 7*7 pixels blocks;
2. Construct color feature vector for image block with 72-dimension gray feature vector;
3. Construct texture feature vector of image block, including average, variance, maximum, minimum, and weighted centroid pixel. Weighted centroid pixel is calculated by convolution product of pixels value distribution and normal distribution;
4. Construct graph G, whose vertex is corresponding to image block;
5. Compute weighted adjacency matrix W, whose edge weight w_{ij} is similarity and distance metric of node i and j corresponding to block i and j respectively;
6. Compute the unnormalized Laplacian L;
7. Compute the first m eigenvectors $v_1, \ldots, v_m$ of the generalized eigenproblem $Lv = \lambda Dv$;
8. Let $V \in R^{n\times m}$ be the matrix containing the vectors $v_1, \ldots, v_m$ as columns;
9. For $i = 1, \ldots, n$, let $y_i \in R^m$ be the vector corresponding to the i-th row of V;
10. Select the optimal parameters k and m;
11. Cluster the points $(yi)_{i=1,\ldots,n}$ in R^m with the k-means algorithm into clusters $C_1, \ldots, C_k$.

3 Experimental Results

We have implemented the improved block-based N-cut algorithm in Matlab, and the experiments are performed on Caltech101 dataset [11], Corel dataset and images downloaded from Internet. The experiments are performed on a PC of Pentium IVs with 3.0 GHz and 512 Mb memory, running Windows XP OS. In this section we will discuss the details of parameters selection and show experimental results with different parameters. We compare the results of block-based N-cut with original N-cut.

3.1 Experimental Design

We use different block size as basic unit for N-cut clustering, including 3*3, 5*5, 7*7, and 9*9, and compare the effect produced by different block size [12]. To compare the difference caused by algorithm itself, we don't adapt any optimization method for matrix storage and computation, which make the algorithms complexity comparable. As a result, the algorithm code in Matlab will cost much more time than other optimized computational methods implemented by the others. How to optimize the representation and computation of matrix is not the focus of this paper.



(a) Original Image

(b) 5*5, k=2, t=700 seconds

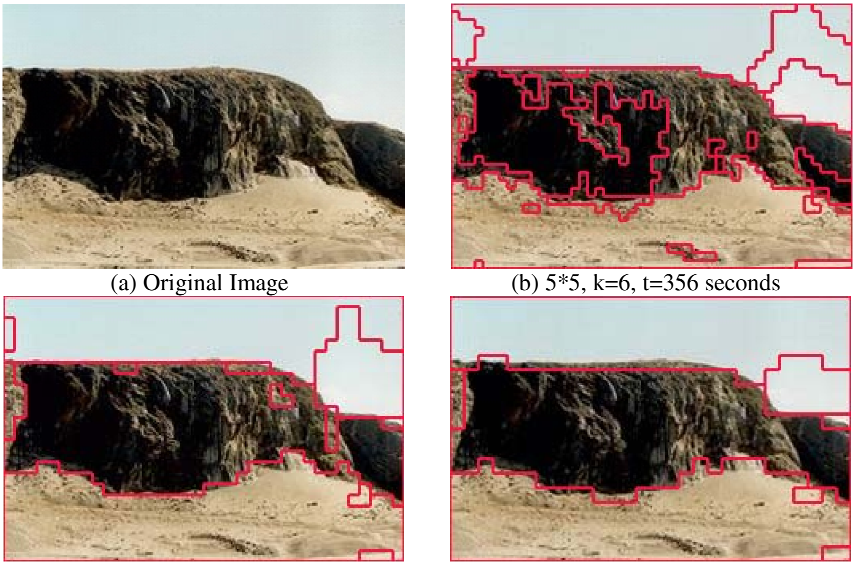
(c) 7*7, k=3, t=118 seconds

(d) 9*9, k=3, t=43 seconds

Fig. 1. Segmentation with different parameters

As the time overhead of original N-cut algorithm is too much, we use the smaller size images as experimental images. When the size is 90*120, the average time overhead is 11340 seconds for original N-cut, while only 63 seconds for block-based N-cut in average. When the image size is 300*250 with three difference block size including 5*5, 7*7, and 9*9, their time overhead of the block-based N-cut is 700, 118, and 43 seconds respectively. The effects of segmentation are shown in Fig.1.

We select many kinds of texture images as test images from dataset. The experiments show that the proposed block-based N-cut can achieve better segmentation performance for texture image. When the image is 250*165, and the block size is 5*5, 7*7, and 9*9, the block-based N-cut segment the image as shown in Fig.2, whose time overhead is 356, 59, and 7 seconds respectively. From Fig.2, we can conclude that the bigger block size, the better performance for texture image, this is because that bigger block has more discriminative visual feature especially texture information.



(a) Original Image

(b) 5*5, k=6, t=356 seconds

(c) 7*7, k=4, t=59 seconds

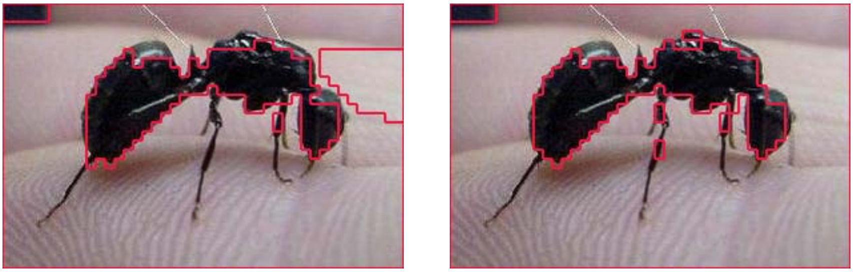
(d) 9*9, k=4, t=7 seconds

Fig. 2. Texture image segmentation with different parameters

3.2 Parameter Selection

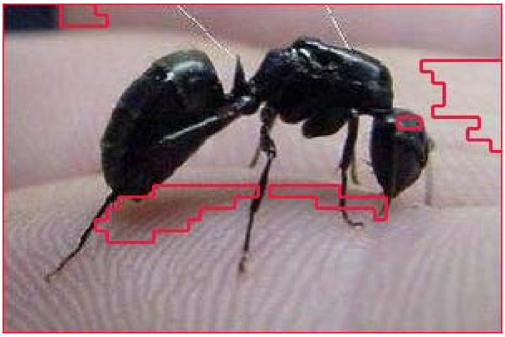
This section will discuss the parameters k and m setting and their effect. As shown in Fig.3, differeent parameters can produce different segmentation result. In Fig.3, there are two kinds of image segmentation result, the parameters in first image setting by clustering while the parameters in second setting by manual. The Fig.3(b) is much more simular with human's. We can infer that the "optimal" k produced by clustering algorithm is not always the best selection from human perspective.

We always think that the more m, and the more effect, but as shown in Fig.4, when the m is larger than threshold, the more m, and the less effect. When the block size is 7*7, k is 3, m is 10, the time overhead of segmentation is 115 seconds. It may be explained by the fact that the more rigor and accurate of individual feature with the bigger m, but the discriminative power will decrease if the m is overfull. As a result, the bigger m will lead to poor quality of segmentation when the m is larger than k. The appropriate m should approximate k, but no bigger than k.



(a) 7*7, k=3, m=2, t=118 seconds (b) 7*7, k=2, m=2, t=120 seconds

Fig. 3. Parameters selection comparison with the same image



7*7, k=3, m=10, t=115 seconds

Fig. 4. Poor quality of segmentation with larger m

4 Conclusions and Future Work

In this paper, we propose block-based N-cut algorithm for image segmentation so as to increase speed and improve the performance. The experimental results have shown that the proposed N-cut can accelerate the speed of segmentation greatly with scarifying a little boundary flatness. We determine the 7*7 as block size for clustering basic unit according to the experimental results. The block-based N-cut is a good choice for segmenting huge amounts of image dataset online.

To further improve the segmentation performance with smooth boundary, we can apply some fitting curve to approximately represent boundary of region. We can adapt variant size or different size block as clustering unit instead of fixed size block. We can apply supervise, semi-supervised learning, or adaptive algorithm to determine the parameters such as m and k that require user to specify manually. Segmenting image into regions as human-level is a hard problem to solve. Currently, every segmentation algorithm is difficult to attain ideal performance for unlimited domain images. We believe that combining more machines learning and psychology knowledge, cognitive science and feedback mechanism into image segmentation algorithms will be the research focus in the future.

Acknowledgements

We thank Xiaoli Zhang and Lintong Huang for their helping in experiment and evaluation. This work was supported in part by National Natural Science Foundation of China under grants 60675008, and Youth Foundation of Dalian Nationalities University under grants 2006A203.

References

1. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Transactions on Computing Surveys 40, 5:1–60 (2008)
2. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision 59(2), 167–181 (2004)
3. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transaction Pattern Analysis and Machine Intelligence 24, 603–619 (2002)
4. Wu, Z., Leahy, R.: An optimal Graph Theoretic Approach to Data Clustering: Theory and Its Application to Image Segmentation. IEEE Transaction on Pattern Analysis and Machine Intelligence 15, 1101–1113 (1993)
5. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 147–159 (2004)
6. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Transaction on Pattern Analysis and Machine Intelligence 22, 888–895 (2000)
7. Veksler, O.: Image segmentation by nested cuts. In: 2000 IEEE Conference on Computer Vision and Pattern Recognition, pp. 339–344. IEEE CS Press, Los Alamitos (2000)
8. Wang, S., Siskind, J.M.: Image segmentation with ratio cut. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 675–690 (2003)
9. Gdalyahu, Y., Weinshall, D., Werman, M.: Stochastic image segmentation by typical cuts. In: 1999 IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 596–601. IEEE CS Press, Fort Collins (1999)
10. Cour, T., Benezit, F., Shi, J.: Spectral Segmentation with Multiscale Graph Decomposition. In: 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 2, pp. 1124–1131. IEEE CS Press, San Diego (2000)
11. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: 2004 International Conference of Computer Vision and Pattern Recognition 2004, Workshop on Generative-Model Based Vision, pp. 178–187. IEEE CS Press, Washington (2004)
12. Chen, J.: Research and Implementation of Spectral Clustering in Image Segmentation. Dalian Nationalities University (2009)

Semantics Web Service Characteristic Composition Approach Based on Particle Swarm Optimization

Zhou Xiangbing

Dep. of Computer Science, Aba Teachers College,
611741 Pixian, China
studydear@gmail.com

Abstract. Service composition is one of the main behavior in the SOC(Service-Oriented Computing) process, which direct and indirect influences effectiveness and precision of service computing; But at present, relation researches mainly focus on semantics recognition and QoS(Quality of Service). In the paper, according to semantics characteristic classification, we proposed a semantics web service characteristic composition approach based on particle swarm optimization, and set up a characteristic selsection mechanism of semantics web service, and adopt charecteristic distance relation to implement service characteristic classification, and use the distance relation to build characteristic tendency degree, sufficiency and characteristic extractor computing formula of semantics web service, at the same time, according to the formula, to implement service characteristic composition algorithm. Then, we set up a optimal mathematical model via characteristic extractor formula. And employ particle swarm to optimize the model and Amazon service set to make experiment, which showed that it is feasible and effective.

Keywords: Particle Swarm Optimization, Semantics Web Service, Service characteristic Composition.

1 Introduction

Web service discovery, service selection and service composition are the most important characteristic of SOC(Service-Oriented Computing); And service demand adopt service composition to implement bussiness fuction, which employ WSDL(Web Services Description Language) to obtain service fuction, but WSDL lacks semantics recognition. Therefore, semantics web service is proposed by W3C, which is composed of semantics web and web service. And employ OWL-S(Ontology Web Language for Services) to describe semantics web service, which defines ServiceProfile, ProcessModel and ServiceGrounding to describe semantics web service. Where, ServiceProfile is similar with service's yellow page, which describes serivce fuction and relation property; ProcessModel facets service's process model, which describes service's work method; ServiceGrounding express that process model get in contact with communication protocol, message format and other information, what describe a service. But different bussiness systems possess many services, and these services are different or similar. Which is loosecoupling and dynamic, to compose services are very different, Therefore,

Q. Luo (Ed.): Advancing Computing, Communi., Control and Management, LNEE 56, pp. 279–287.
springerlink.com

many researchers have proposed some approaches of service computing, such as change in demand of services[1], intelligent computer[2], description logic[3], based on QoS(Quality of Service)[4]. And service selection mainly focus on sermantics recognition and QoS[6]. However, we proposed a service characteristic composition approach among services distance relation, and defined computing formula of tendency and adaptation degree. Firstly, we employ separable criterion to express among services distance[9], and defined similarity to implement service characteristic extraction among services, then, according to these formula, we set up a mathematical model of service characteristics composition. Secondly, we adopt Particle Swarm Optimization algorithm to optimize the model; which effectually composed services, and meet demand of service request. Finally, we employ Amazon services set to make experiment showed: it is a good result.

2 Semantics Web Service Characteristic Composition Approach

Web service is composed of service request, service response and UDDI(Universal Description Discovery and Integration), and adopt SOAP(Simple Object Access Protocol) to exchange information and WSDL(Web Services Description Language) to describe service fuction. However, semantics is infuseed into WSDL, which let web service possess semantics, namely: semantics web service. Which adopt OWL-S based Ontology to describe semantics, it is aim to compose services via semantics regulation and recognition rule. By now, web service possess itself of semantics. Showed in Fig.1.

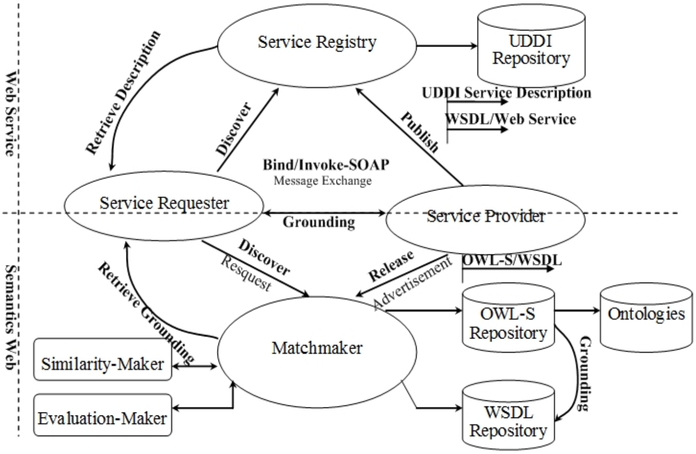


Fig. 1. Semantics Web Service Architecture

Definition 1 [7]**.** Ontology is defined by tuple O=(C,$Root$,A^C,H,R,I). Where, C={C_1,C_2,…,C_n} denote concept set, |C| denote concept's number, namely: |C|=n, $Root$ is root concept of Ontology, one and only one root $Root$, A^C denote concept $c\in C$'s property set, and A^C={$a^C(i)$|i=1,2,…,|A^C|}, |A^C| denote concept property number, H denote layer relation of concept, and $H\subseteq C\times C$, $(c_i,c_j,\omega_{ij})\in H$, $c_i\in C$, $c_j\in C$, which denote

c_i is c_j 's subconcept, ω_{ij} is weight of the relation, H is directed acyclic graph; R denote no-layer relation set among concepts, $R=\{r(c_i,c_j,\omega_{ij})| \; c_i\in C,c_j\in C \}$, $r(\cdot)$ denote relation between c_i and c_j, I denote instance set of concpet C, concept $c\in C$'s instance set is $I(c)$, $| I(c)|$ denote instance number.

Definition 2. Sermantics web service is defined by tuple

$$SWS=(O.owl\text{-}s,Sim(o_i,o_j),WS(r,q),wsdl.O,Uddi,Soap,Scope,Fc_D).$$

Where, *O.owl-s* denote description constraint in the Definition 1, $Sim(o_i,o_j)$ denote similarity between Ontology, and $o_i\in O,o_j\in O$, $WS(r,q)$ denote service type, namely: request and response service; *wsdl.O* denote mapping relation between WSDL and OWL-S via Ontology O, which employ the literature[8] to implement mapping between WSDL and OWL-S; *Uddi* is SWS's registration center; *Soap* is SWS's information transmission format, which adopt SOAP based on XML to express; *Scope* is a given scope of sermantics web service; *Fc* denote characteristic of sermantics web service, many semantics web services *Fc* can make up of *D dimensions* characteristic vector, which are denoted Fc_D.

Definition 3. Semantics web service characteristic selection is defined by tuple

$$SWSCS=(d(SWS.WS(r,q)),A(d)),$$

which separately denote distance and average distance between service request and response, thus, according to the literature[9]:

For $\forall sws_i,sws_j\in SWS$, if they possess distance among service request sws_i and response sws_j. Simultaneously, all of these distances add and average, and use the result to measure characteristic distribution situation. This moment, we suppose $x_k^{(i)},x_l^{(j)}$ separately denote *D dimensions* characteristic vector of sws_i and sws_j. Namely: $d(SWS.WS(r,q))$'s distance can be denoted to $|d(SWS.WS(x_k^{(i)},x_l^{(j)}))|$, $\delta(x_k^{(i)},x_l^{j})$ denote distance between vectors, then the average distance of service request and response is

$$A(a)=J_d(x)=\frac{1}{2}\sum_{i=1}^{c}P_i\sum_{j=1}^{c}P_j\frac{1}{n_i n_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}\delta(x_k^{(i)},x_l^{(j)}) \qquad (1)$$

Where,

c denote SWS's categories number, n_i denote sws_i class's sample number, n_j denote sws_j class's sample number; P_i is sws_i class's probability, P_j is sws_j class's probability, which can be defined:

$$P_i=\frac{|sws_i|}{|sws_i|+|sws_j|},P_j=\frac{|sws_j|}{|sws_i|+|sws_j|}$$

$|\;|$ denote service request and response sample set number. And two vectors distance adopt euclidean distance to compute in the hyperspace, namely:

$$\delta(x_k^{(i)},x_l^{(j)})=(x_k^{(i)}-x_l^{(j)})^T(x_k^{(i)}-x_l^{(j)}) \qquad (2)$$

Simultaneously, we use m_i to denote ith mean vector of sample set

$$m_i = \frac{1}{n_i}\sum_{k=1}^{n_i} x_k^{(i)} \tag{3}$$

We use m to express average vector each class sample set

$$m = \sum_{i=1}^{c} P_i m_i \tag{4}$$

This moment, (2),(3),(4) were embedded into (1):

$$A(a) = J_d(x) = \sum_{i=1}^{c} P_i [\frac{1}{n_i}\sum_{k=1}^{n_i} (x_k^{(i)} - m_i)^T (x_k^{(i)} - m_i) + (m_i - m)^T (m_i - m)] \tag{5}$$

All kinds of average distance of mean vectors is:

$$\sum_{i=1}^{c} P_i (m_i - m)^T (m_i - m) = \frac{1}{2}\sum_{i=1}^{c} P_i \sum_{j=1}^{c} P_j (m_i - m)^T (m_i - m) \tag{6}$$

Definition 4. To define semantics web service characteristic tendency degree

$$Ori(x_k^{(i)}, x_l^{(j)}) = \frac{|\log_2(\min(\sum_{i,j\in D} d(SWS.WS(x_k^{(i)}, x_l^{(j)}))))|}{|\sum_{i,j\in D} A(a)_{(i,j)}|} \tag{7}$$

Definition 5. To define semantics Web service characteristic sufficiency

$$Suf(x_k^{(i)}, x_l^{(j)}) = \frac{w_1 \sum_{i,j\in D} d(SWS.WS(x_k^{(i)}, x_l^{(j)})) + w_2 \sum_{i,j\in D} A(a)_{(i,j)}}{D(w_1 + w_2)} \times Ori(x_k^{(i)}, x_l^{(j)}) \tag{8}$$

Where

$0 < w_1 + w_2 \le 1$, which stand for weight.

Definition 6. Semantics web service quality is defined by tuple $A\text{-}QoS=(T,C,A,SR,AR)$, which separately denote time, cost, availability, success rate and aware characteristic extractor. If sws_i and sws_j start computing, then AR have to do with T,C,A,SR can be define

$$AR(sws_i, sws_j) = \frac{\sum_{i=1}^{D\cdot sws_i}\sum_{j=1}^{D\cdot sws_j} (T_{ij} + C_{ij} + A_{ij} + SR_{ij})}{D(|sws_i| + |sws_j|)} \times Suf(x_k^{(i)}, x_l^{(j)}) \tag{9}$$

Where i,j=1,2,…,n.

In order to compute, we set three service counters (Q_1, Q_2, Q) and one timer, which separately record service request number, service response number, service composition number and time. T_{ij} set a multi-thread timer to record, therefore

$$C_{ij}(T)=\frac{Q}{Q_1+Q_2}T\,;\; A_{ij}=\frac{Q_1}{Q_2}\,;\; SR_{ij}=\frac{|Q_1\leftrightarrow Q_2|}{Q}$$

Definition 7. Semantics web service composition is define by tuple

$$SWSC=(RS,SWS,SWS.Fc(Ori,Suf),A\text{-}QoS(AR),SWSR,SWSF,Fun)$$

Where

RS denote service demand, *SWSR* denote service relation, showed in literature[10], *SWSF* denote service composition workflow, showed in literature[11], *Fun* denote bussiness fuction of service composition later. And other items showed in definition 2-6.

Algorithm 1. Semantics web service characteristic composition algorithm

```
1 $\forall sws_i,sws_j\in SWS$, $\exists rs\in RS$ cause $rs\rightarrow sws_{(i,j)}$;
2 To compute characteristic distance $\delta(x_k^{(i)},x_l^j)$ and average distance $A(a)$;
3 if($A(a)$ meet given characteristic value)
    while($rs\rightarrow sws_{(i,j)}$)
      To compute endency degree $Ori$ and sufficiency $Suf$ of service characteristic
          extraction;
      if($Suf$ meet $rs$)
        To compute $A\text{-}QoS(AR)$, and to obtain a group of rs's service;
      else
        return 3 and recompute;
  else
  return 2 and recompute
4 while(select $SWSR$)
    Call $SWSF$ approach to run a bussiness fuction $Fun$;
5 if($Fun$ meet rs's demand)
    return 2 and recompute;
  else
    exit
```

3 Semantics Web Service Characteristic Composition Optimization

Semantics web service characteristic selection is a group of optimization characteristic(amount: $d(D>d)$) from *D dimensions* service characteristic. Therefore, we build a optimization model to implement semantics web service composition, and employ Particle Swarm Algorithm to optimize the model.

For one service demand x, which cause semantics web service aware extraction optimization, namely:

$$y=\max_{i,j\in D\cdot d}\{x\rightarrow AR(sws_i,sws_j)\} \tag{10}$$

And average distance of sample set employ Bhattacharyya[9] to limit

$$0<A(a)\le -\ln\int\sqrt{[p(x\,|\,\omega_1)p(x\,|\,\omega_2)]}dx \tag{11}$$

Where

ω_1, ω_2 is probability of P_i,P_j,

Now, we employ Particle Swarm Algorithm to optimize the y in the (10), which is a very simple, more powerful search evolutionary algorithm[12], therefore, which have been applied to function optimization, neural network trace training, semantics recognition and fuzzy systems and control, etc[7]. This moment, we supposed the swarm is composed of m particles, m is also known as population size, and m's size direct influence velocity and astringency of Particle Swarm Algorithm; if supposed z_i=(z_{i1},..,z_{iD}) is *D dimensions* position vector of ith(i=1,2,…,m) particles, and according to z_i 's adaptive value of y computing, namely: which can measure quality of particle position vector; v_i=(v_{i1},v_{i2},…,v_{id},…,v_{iD}) is flight velocity of particle i, namely, distace of particle move, but the distace must be less than Bhattacharyya distance; so far, p_i=(p_{i1},p_{i2},…,p_{id},…,p_{iD}) is optimization semantics Web service of particles search;p_g=(p_{g1},p_{g2},…,p_{gd},…,p_{gD}) is so far optimization Web service characteristic group of all particles search.

At a iteration time, particles update velocity and position following formula:

$$v_{id}^{k+1}=wv_{id}^{k}+c_1r_1(p_{id}-z_{id}^{k})+c_2r_2(p_{gd}-z_{id}^{k}) \tag{12}$$

$$z_{id}^{k+1}=z_{id}^{k}+v_{id}^{k+1} \tag{13}$$

Where

i=1,2,…,m, d=1,2,…,D, k is iterations, w is inertia factor (0.8≤w≤1); r_1, r_2∈[0,1], which keep diversity of population; c_1, c_2 is learning factor (0.5≤c_1,c_2≤1) or acceleration factor.

Algorithm 2. Particle Swarm Algorithm optimize semantics web service characteristic composition

```
1 To initialize semantics web service domain., which can optimize semantics web service domain;
2 To select a threshold value $\varepsilon$ and maximum iteration: $N_{max}$;
3 To initialize particles's position $z_i^{(0)}=(z_{i1},z_{i2},\cdots,z_{iD}),i=1,2,\cdots,n$ and each particle's velocity $v_i^{(0)}=(v_{i1},v_{i2},\cdots,v_{iD})$
4 To test each particle's adaptive value $z_i^{(0)}$, and denote: $y_i^{(0)}$, $p_i^{(0)}=z_i^{(0)}$;
5 for(k=0;k<n;k++)
  {
    According to $y_i^{(0)}=\max\{y_1^{(0)},y_2^{(0)},\cdots,y_m^{(0)}\}$ seek global optimum $p_g^{(0)}$;
    According to formula (9) update $v_i^{(k)}$;
    According to formula (10) update $z_i^{(k)}$;
  }
6 To measure y's adaptive value, and denote $y_i^{(k)}$, namely: $y_i^{(k)}=\max\{y_1^{(k)},y_2^{(k)},\cdots,y_m^{(k)}\}$;
```

```
7 To update $p_i^{(k)}$ and $p_g^{(k)}$;
8 if if( $((y^{(k-1)} - y^{(k)}) / y^{(k)} > \varepsilon)$ and $(k < N_{\max})$ )
   return (5);
  else
  exit.
```

4 Experiment Result

Experiment environment set: JDK1.4+Tomcat5.0+OracleXE+Eclipse(MyEclipse), experiment data edit tool(Semantics Web service set) is open source software :Axis(edit service)+ Protégé(ontology edit tool)+ OWL-S editor (http://www.semwebcentral.org/); And we employ Amazon service set as basic experiment data, at the same time, according to property of service set; and develop a experiment procedure set(Exp.test.*) to make experiment, which included semantics web service characteristic, Algorithm 1 and Algorithm 2 in the paper. Then we deploy mapping relation of these procedure set and Amazon service set in the Web.xml. And set three service counters (Q_1, Q_2,Q) and one timer as Definition 6, and start making experiment. Finally, we employ books online service set to analyze in the Amazon. And adopt login service, authentication service, order service, retrieval service, reputation service and clearing service to analyze. According to experiment advance, we add services number to analyze experiment result, showed in Fig.2 and Fig.3. Particle parameter set as literature [13](w=0.7, c_1=c_2=1.5), and Amazon service set population size is 2 times *D dimensions*, evolution generations set 500 times. Namely: particle number is servcie set's number, and according to Definition 6 can obtain service number and time.

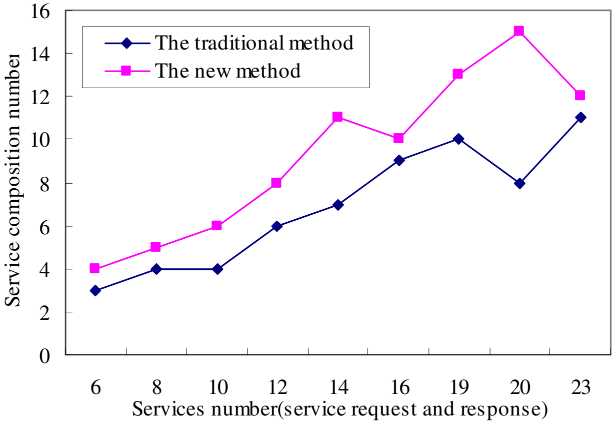


Fig. 2. The service composition number of the traditional approach compared with the new approach

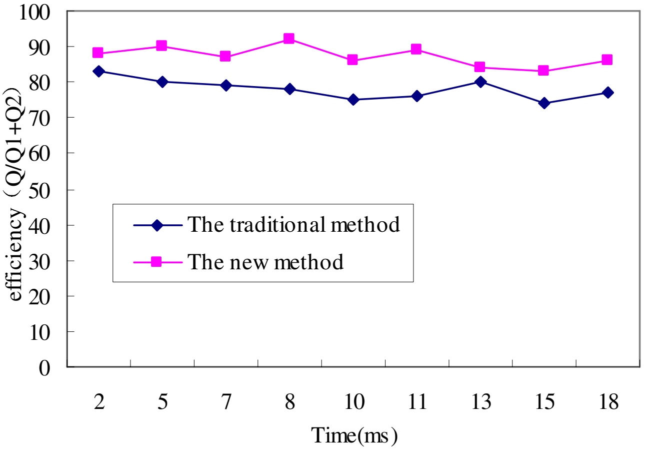


Fig. 3. The efficiency of the tradition is compared with the new approach

5 Conclusion

We proposed a semantics web service characteristic composition approach based on Swarm Algorithm in the paper. Firstly, we sum up current relation service composition approach; Secondly, to build semantics web service characteristic extractor approach via separable criterion distance, and define tendency degree, sufficiency and semantics web service characteristic extractor computing formula, at the same time, according to the formulas, which set up semantics web service characteristic composition algorithm and optimization model, and employ inertia factor's Particle Swarm Algorithm to optimize the model. Finally, we empoly Amazon service set to make experiment and showed: it is feasible and effective.

Acknowledgement

This work is supported by the Natural Sciences of education Bureau of Sichuan Province of China under Grant No. 07SC002.

References

1. Blessa, P.N., Klabjan, D., Chang, S.Y.: Heuristics for automated knowledge source integration and service composition. Computers & Operations Research 35, 1292–1314 (2008)
2. Liang, W.-Y., Huang, C.-C.: The generic genetic algorithm incorporates with rough set theory - An application of the web services composition. Expert Systems with Applications 36(3), 5549–5556 (2009)
3. Wang, J.S., Li, Z.J., Li, M.J.: Compose semantic web services with description logics. Journal of Software 19(4), 957–970 (2008)

4. Myoung, J., Ouk, C., Ick-Hyun: Quality-of-service oriented web service composition algorithm and planning architecture. Journal of Systems and Software 81(11), 2079–2090 (2008)
5. Ai, W.-h., Song, Z.-l., Wei, L., Wu, L.: Web Service Discovery Based on Domain Ontology. Journal of University of Electronic Science and Technology of China 36(3), 506–509 (2007)
6. Muhammad Ahsan, S.: A framework for QoS computation in web service and technology selection. Computer Standards & Interfaces 28, 714–720 (2006)
7. Qiang, X., Lei, Z., Liang, Z.: Ontology Partition Method Based on ImprovedParticle Swarm Optimization Algorithm. Journal of South China University of Technology (Natural Science Edition) 35(9), 118–122 (2007)
8. Patil, A., Oundhakar, S., Sheth, A., et al.: Meteor-S Web Service Annotation Framework 2008. In: Proc. of the 13th International Conference on World Wide Web, pp. 17–22. ACM Press, New York (2004)
9. Bian, S., Zhang, X.: Pattern recognition, 2nd edn. Tsinghua University Press (2001)
10. Xu, M., Chen, J.L., Peng, Y., Mei, X.: Service relationship ontology-based Web services creation. Journal of Software 19(3), 545–556 (2008)
11. Xiangbing, Z.: Semantic Web Services Component Automata Based Ontology 2008. In: Proceedings of the 27th Chinese Control Conference, Kunming, Yunnan, China, pp. 719–723. IEEE Press, Los Alamitos (2008)
12. Eberhart Russell, C., Yuhui, S.: Comparison between genetic algorithms and particle swami optimization. In: Porto, V.W., Waagen, D. (eds.) EP 1998. LNCS, vol. 1447, pp. 611–616. Springer, Heidelberg (1998)
13. Ji, Z., Liao, H., Wu, Q.: Particle Swarm Optimization and Application. Science Press (2009)

Author Index