Text Classification with Support Vector Machine and Back Propagation Neural Network

Wen Zhang[1], Xijin Tang[2], and Taketoshi Yoshida[1]

[1] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Ashahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
{zhangwen,yoshida}@jaist.ac.jp
[2] Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, P.R. China
xjtang@amss.ac.cn

Abstract. We compared a support vector machine (SVM) with a back propagation neural network (BPNN) for the task of text classification of XiangShan science conference (XSSC) web documents. We made a comparison on the performances of the multi-class classification of these two learning methods. The result of an experiment demonstrated that SVM substantially outperformed the one by BPNN in prediction accuracy and recall. Furthermore, the result of classification was improved with the combined method which was devised in this paper.

Keywords: text classification, SVM, BPNN, Xiangshan Science Conference.

1 Introduction

Automated text classification utilizes a supervised learning method to assign predefined category labels to new documents based on the likelihood suggested by a trained set of labels and documents. Many studies have been taken to statistical learning methods to compare their effectiveness in solving real-world problems which are often high-dimensional and have a skewed category distribution over labeled documents.

XiangShan Science Conference (XSSC) is well known in China as a scientific forum for the invited influential scientists and active young researchers about frontiers of scientific researches. The major discussions are summarized and posted at XSSC Web-site (http://www.xssc.ac.cn) and then aggregated into a large and valuable repository with all kinds of information which are related to scientific research and development in China. Some studies about XSSC have been undertaken from the perspective of enabling knowledge creation and supporting it by information tools [1-2]. Augmented information support (AIS) is one of those tools, for which Web text mining technologies are applied and basic tasks such as Web crawler, feature selection and indexing, and text clustering [3].

Performance examinations of statistical learning methods are usually carried out on the standard data set. Generally speaking, there is no superior algorithm in the statistical learning area for text classification problems. Even with the same classifier, different

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 150–157, 2007.

performances may be revealed with different types of data sets because no statistical analysis was conducted to verify the impact of difference in the data on the performance variation of these classifiers [4]. For this reason the practical XSSC datasets are applied to examine the performance of SVM and BPNN.

The rest of this paper is organized as follows. Section 2 describes XSSC Web document representation and clustering of the datasets for XSSC data preprocessing. Section 3 describes the experiment design. The experimental results are presented in Section 4 and a comparison between SVM and BPNN is also conducted on the multi-class texts classification in the same section. The combined method is devised to integrate the results of SVM and BPNN and its performance is also demonstrated. Finally, concluding remarks and further research are given in Section 5.

2 XSSC Data Preprocessing

This section describes the preprocessing for the performance evaluation of both SVM and BPNN.

2.1 XSSC Web Documents Representation

Based on our prior work, 192 Web documents were collected from XSSC Website and a set of keywords for the collection of the whole documents was created to represent the Web documents. That is,

$$\text{Doc(i)} = (k_{i,1},...,k_{i,j},...,k_{i,m}),\ \text{let } k_{ij} = \begin{cases} 1, \text{ if keyword j exists in the } i\text{th document} \\ 0, \text{ if keyword j does not exist in the } i\text{th document} \end{cases} \quad (1)$$

m is the total size of the combined keywords collection.

Thus, 192 Boolean vectors were obtained to represent the 192 Web documents mentioned above initially. Secondly, a cosine transformation was conducted with these Boolean vectors to represent the Web documents more accurately and objectively. That is, let $\bar{k}_{i,j} = \dfrac{Doc(i) \bullet Doc(j)}{|Doc(i)||Doc(j)|}$ ($\bullet$ describes an inner product of vectors) and the representation vectors for the 192 Web documents were replaced with the newly generated immediate vectors using cosine transformation $\overline{Doc(i)} = (\bar{k}_{i,1}, \bar{k}_{i,2},..., \bar{k}_{i,192})$. Here, $\overline{Doc(i)}$ is the final and newly adopted representation vector for ith document. All following data preprocessing and the latter performance examination are all carried out on these transformed representation vectors.

2.2 Data Clustering

Clustering techniques are applied when there is no class to be predicted but the instances are required to be divided into natural groups. Two methods hierarchical clustering method and heuristic tuning method are applied to cluster the 192 documents into the predefined classes for a testing data set for the multi-class classification performance examination of SVM and BPNN in the next section. The testing data set is constructed by the following two steps:

Step 1: The similarity vectors representing the Web documents are processed by hierarchical clustering analysis in SPSS and a dendrogram is generated to describe the overall distribution of the documents on the given categories.

Step 2: Heuristic method is employed by manually adjustment on the references of documents clusters obtained in Step 1 to make them appropriately categorized i.e. to provide a standard classification for these documents.

Table 1 is the standard documents clustering generated by the above mentioned processing method. A skewed category distribution can be seen and a general trend of research focus currently existing among all the categories in XSSC can be drawn out. Here, 5 outliers were detected during clustering excluded from the following processing.

Table 1. Standard documents clustering on XSSC data set

Category ID	Subject of discipline	Total	Percentage
1	Life Science	60	31.25
2	Resource and Environment Science	31	16.15
3	Basic Science	21	10.94
4	Scientific Policy	16	8.33
5	Material Science	15	7.81
6	Transportation and Energy Science	11	5.48
7	Information Science	8	4.17
8	Space Science	6	3.13
9	Complexity Science	6	3.12
10	Outliers	5	2.60
11	Aeronautics & Astronautics	4	2.08
12	Micro-electronic Science	3	1.56
13	Safety Science	3	1.56
14	Other	3	1.56
Total	------	192	100.00

3 Experiment Design

The purpose of experiment design is to devise a controlled study on SVM and BPNN multi-class text classification performance. Regarding the skewed data distribution as shown in Table 1, the problem of the unbalanced data is dealt by assigning the number of training data and the number of test data in each class with same proportion.

3.1 Multi-class Text Classification Experiment Design

For multi-class text classification, four types of experiments are designed to test the performance of SVM and BPNN. Here three-class examination is conducted because the classification of more than three classes is very similar with three classes; on the other hand, the number of sample data is not enough to carry out classification more than three classes as well. The test strategy here is that the number of training sample is fixed as twice as the number of the testing sample whereas the categories from where the data set were selected are varied from each other. Table 2 shows the experiment design for the multi-class examination by using SVM and BPNN where

"30/20/20" means that 30 data samples were selected out randomly from "Life Science" as training data, 20 from category No.2 and 20 from category No.3 as training data. It can be seen that the numbers of training samples and test samples follow a decreasing trend because we also want to study on the performance of SVM and BPNN when the numbers of training and testing set are varying.

Table 2. Experiment design for multi-classification

Test No.	Test 1	Test 2	Test 3	Test 4
Selected Categories	No.1/ No.2/ Other classes	No.2/ No.3/ No.4	No.3/ No.4/ No.5	No.4 / No.5/ No.6
Numbers of Training data	30/20/20	20/14/11	14/11/10	11/10/8
Numbers of Testing data	15/10/10	10/7/5	7/5/5	5/5/3

3.2 SVM and BPNN Design

SVM is a classifier derived from a statistical learning theory by Vapnik and Chervonenkis (called VC theory hereafter) and it was firstly introduced in 1995 [5]. Based on VC theory and a kernel theory, SVM was proposed that was equivalent to solve a linearly constrained quadratic programming problem.

For multi-classification the one-against-all (OAA) method is adopted because of its same computation complexity with the one-against-one (OAO) in a SVM classifier and usually performs well [6]. Some deficiencies and improvement of this k-class (k>2) classification method were discussed in Ref. [7, 8] such as majority vote, SVM decision tree and so on. Here polynomial kernel $K(s,t) = ((s \bullet t)+c)^d$ (c=1,d=2) is used as the kernel function of SVM classifier because of its better learning ability compared with other kernels in the validation examination of our training data which is used to select the network of BPNN.

BPNN is a method known as back propagation for updating the weights of a multi-layered network undergoing supervised training [9, 10]. The back propagation algorithm defines two sweeps of the network: first a forward sweep from the input layer to the output layer and then a backward sweep from the output layer to the input layer. A three-layer fully connected feed-forward network which consists of an input layer, ahidden layer and an output layer is adopted here. With the previous basic training mentioned above, the "tansigmod" function is used in the hidden layer with 5 nodes and "purelinear" function for the output layer with 3 nodes [11]. The network of BPNN is designed as shown in Figure 1.

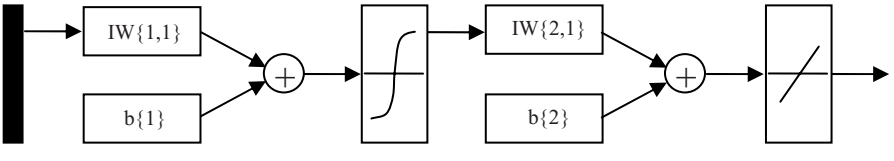

Fig. 1. BPNN with 5 nodes in hidden layer and 1 node in output layer

3.3 Combined Method

A combined method for the prediction of the unlabeled samples is designed to investigate if there are any improvements in prediction accuracy with combined results of SVM and BPNN. If the unlabeled sample is predicted with the same class by both SVM and BPNN, it would be then labeled by this predicted class. Otherwise, it will be given no label and cannot be assigned to any class. The accuracy of a combined method is calculated out by the following formula:

$$Accuracy(Combined\ Mehtod) = \frac{\left|S_{L(\mathrm{SVM})=L(\mathrm{BPNN})=L(\mathrm{Standard})}\right|}{\left|S_{L(\mathrm{SVM})=L(\mathrm{BPNN})}\right|} \tag{2}$$

Here, $S_{L(\mathrm{SVM})=L(\mathrm{BPNN})}$ is defined as the set of those samples that SVM and BPNN give the same predicted class label. By analogy, $S_{L(\mathrm{SVM,\,i})=L(\mathrm{BPNN,\,i})=L(\mathrm{Standard,\,i})}$ is defined as the set of those samples that SVM, BPNN and standard data set give the same class label. Then the *accuracy* means that to how much extent the prediction which the SVM and BPNN give the same label is reliable, i.e. the right answer for the unlabeled class.

4 Results of Experiments

Experiments were conducted according to the design described in section 4. We constructed the SVM and BPNN classifiers with the help of mySVM [11] and MatLab Neural ToolBox [12]. All the tests were conducted iteratively for 10 times and the average value of indicators was calculated to observe the performances of SVM and BPNN.

4.1 The Results of SVM and BPNN on Multi-class Text Classification

The results of SVM and BPNN on multi-class text classification are shown in Table 3. The indicators as accuracy and recall are employed to measure the classification of SVM and BPNN. Take the Test 1 for example, we obtained the accuracy as 0.7143 which comes from that 10 test samples from category No.1, 9 test samples from category No.2 and 6 test samples from category No.3 of all the 35 test samples designed in Test 1 mentioned in Section 3 were given the right labels. And recall as "10/16/19" means that 10 test samples were given label No.1, 16 as No.2 and 9 as No.3 in the multi-class classification in Test 1.

Table 3. Accuracies and recalls of SVM and BPNN on multi-class text classification

Classifier \ Test No.		Test. 1	Test .2	Test. 3	Test. 4
BPNN	Accuracy	0.7143 (10/9/6)	0.5909 (8/2/3)	0.4706 (2/3/3)	0.6923 (3/3/3)
	Recall	10/16/9	10/5/7	4/8/5	5/4/4
SVM	Accuracy	0.7714 (11/8/8)	0.6364 (9/3/2)	0.4706 (5/1/2)	0.8462 (4/4/3)
	Recall	14/11/10	11/7/4	9/3/5	5/4/4

From Table 3 it can be said that SVM has outperformed BPNN on the task of XSSC Web documents multi-class classification. The result from SVM classifier demonstrated convincingly better than that the one from BPNN on whatever the accuracies and recalls are.

4.2 The Result of the Combined Method

The combined method introduced in Section 4 was conducted with SVM and BPNN on multi-class text classification in order to examine the performance of our combined method. Table 4 shows the combined result of SVM and BPNN. Take Test 1 for example, the accuracy is 0.9200 which comes from that 25 test samples in this test is given the same label by SVM and BPNN and of them 23 test samples has the same labels as the standard data set. It can be seen that the combined accuracies were improved significantly in comparing SVM with BPNN with the contrast with the result shoed in Table 3. Also a particular comparison in accuracy between the combined method, SVM and BPNN is plotted in Figure 2.

Table 4. Accuracies of the combined method for multi-class text classification

Test No. / Classification	Test 1	Test 2	Test 3	Test 4
Multi-class classification	0.9200(23/25)	0.6875(11/16)	0.5714(4/7)	0.8889(8/9)

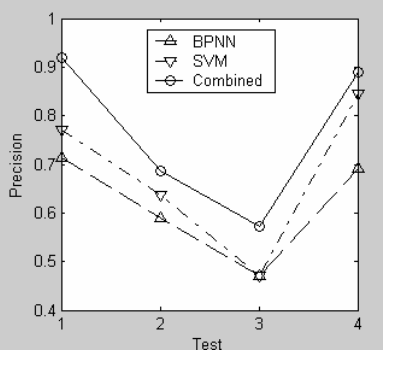

Fig. 2. Accuracies in the combined method, SVM and BPNN on multi-class text classification

5 Concluding Remarks

In this paper some experiments are carried out for multi-class text classification by using SVM and BPNN. Unlike the usual performance examinations the data sets of the experiments taken here are from a real practical application, that is, XSSC data set. A combined method of synthesizing the results from SVM and BPNN is developed to study whether there is an improvement in accuracy if the prediction result from different classifiers is combined. The experimental results demonstrated that

SVM showed a better performance than BPNN on the measure of accuracy and recall. The adaptation of the combined method achieved the improvement of accuracy for the multi-class text classification task.

Although the experimental results have provided us with some clues on text classification, a generalized conclusion is not obtained from this examination. Our work is on the initial step and more examination and investigation should be undertaken for more convincing work

One of the promising directions in text mining field is concerning the predictive pattern discovery from large amount of documents. In order to achieve this goal, we should introduce not only the required learning algorithms but also the semantics into the text mining field. More attention will be concentrated on the areas of semantic Web and ontology-based knowledge management, especially on the work that employs ontology to describe the existing concepts in a collection of texts in order to represent documents more precisely and explore the relationships of concepts from textual resources automatically.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant No.70571078 and 70221001 and by Ministry of Education, Culture, Sports, Science and Technology of Japan under the "Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project".

References

1. Tang, X.J., Liu, Y.J., Zhang, W.: Computerized Support for Idea Generation during Knowledge Creating Process. In: Khosla, R. J. Howlett, and L. C. Jain (eds.): Knowledge-Based Intelligent Information & Engineering Systems (proceedings of KES'2005, Part IV), Lecture Notes on Artificial Intelligence, Vol.3684, Springer-Verlag, Berlin Heidelberg (2005) 437-443.
2. Liu, Y.J., Tang, X.J.: Developed computerized tools based on mental models for creativity support. In: Gu, J. F. *et al*.(eds.): Knowledge and Systems Sciences: toward Knowledge Synthesis and Creation. (Proceedings of KSS2006), Lecture Notes on Decision Support, Vol. 8, Global-Link, Beijing (2006) 63-70.
3. Zhang, W., Tang, X.J.: Web text mining on XSSC. In: Gu, J. F. *et al*.(eds): Knowledge and Systems Sciences: toward Knowledge Synthesis and Creation. (Proceedings of KSS2006), Lecture Notes on Decision Sciences, Vol.8, Global-Link, Beijing (2006) 167-175
4. Yang, Y.M., Lin, X.: A re-examination of text categorization methods. In: Proceedings on the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley,C A, (1999) 42-49..
5. F., Mulier.: Vapnik-Chervonenkis (VC) Learning Theory and Its Application. IEEE Trans on Neural Networks. 10(5) (1999) 5-7
6. Christophier, J., C, Burges.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. Kluwer Academic Publisher, Boston (1998) 121-167

7. Rennie, J. D., Rifkin, R.: Improving Multi-class Text Classification with the Support Vector Machine. Master's thesis. MIT (2001)
8. Weston, J., Watkins, C.: Multi-class support vector machines. In Proceedings ESANN. Brussels (1999)
9. Rob, C.: Artificial Intelligence. Palgrave Macmillan. New York (2003) 312-315.
10. Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning internal representations by error propagation. In Parallel Distributed Processing, Exploitations in the Microstructure of Cognition, Vol. 1. Cambridge, MA: MIT Press (1986) 318-362
11. Stefan, R.: mySVM-Manual. Online: http://www-ai.cs.unidortmund.de/software/mysvm. (2000)
12. Neural Network Toolbox for MATLAB. Online: http://www.mathworks.com/products/neural-net/