A Customer Segmentation Model Based on Affinity Propagation Algorithm and Improved Genetic K-Means Algorithm

Meiyang Zhang[1], Zili Zhang[1,2(✉)], and Shi Qiu[3]

[1] College of Computer and Information Science, Southwest University, Chongqing 400715, China
zhangzl@swu.edu.cn
[2] School of Information Technology, Deakin University, Locked Bag 20000, Geelong, VIC 3220, Australia
[3] Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China

Abstract. Customer Relationship Management System (CRM) has accumulated massive customer transaction data. Effective customer segmentation by analyzing transaction data can contribute to marketing strategy designing. However, the state-of-the-art researches are defective such as the uncertain number of clusters and the low accuracy. In this paper, a novel customer segmentation model, AP-GKAs, is proposed. First, factor analysis extracts customer feature based on multi-indicator RFM model. Then, affinity propagation (AP) determines the number of customer clusters. Finally, the improved genetic K-means algorithm (GKAs) is used to increase clustering accuracy. The experimental results showed that the AP-GKAs has higher segmentation performance in comparison to other typical methods.

Keywords: Customer segmentation · Affinity propagation
Genetic K-means algorithm

1 Introduction

With the rapid growth of the national economy, the competition in all walks of life has become fierce recently. In this competitive commercial framework, it is becoming more and more important for enterprises to analyze and understand the needs and expectations of customers. Accordingly, most enterprises establish CRM to accumulate massive customer data which can be analyzed and applied to effective targeting and predicting potential customers.

Customer segmentation [1] using clustering to discover intrinsic patterns of customer behaviour based on the transaction data. K-means has been widely used in customer segmentation because of its simplicity and fast convergence,

Z. Shi et al. (Eds.): IIP 2018, IFIP AICT 538, pp. 321–327, 2018.
https://doi.org/10.1007/978-3-030-00828-4_32

but K-means is apt to local optimum. Self-Organizing Map (SOM) can map high dimensional input space to low-dimensional topologies and intuitively display data structures. [2] used SOM to segment customer based on RFM model, which has solved characteristic parameter stagger and nonlinear distribution. However, SOM is difficult to accurately analyze the performance indicator. Genetic K-means algorithm (GKA) [3] combines the global optimization of GA and the local search ability of K-means so as to find the global optimal solution. [4] added greedy selection in GKA to solve clustering problem.

The previous researches indicated that the GKA can get more accurate results by solving the problem that K-means is apt to fall into the local optimum. And the uncertain number of clusters will cause the algorithm converge to an immature result. Moreover, most models are researching customer segmentation, which only focus on customer clustering problems. Consequently, this paper proposes the AP-GKAs model to completely analyze transaction data.

2 Related Work

The AP algorithm [5] uses all data points as potential clustering center which called exemplar. The similarity $s(i,k)$ indicates how well the data point k is suited to be the exemplar for data point i, each similarity is to a negative squared error: for point x_i and x_k, $s(i,k) = -\|x_i - x_k\|^2$. The reference degree p indicates the tendency of the data point chosen as an example. It is recommended that **all p are set as the median of $s(s_m)$ without prior knowledge** [6].

The responsibility $r(i,k)$, sent from a data point i to candidate exemplar point k, reflects the accumulated evidence for how well-suited point k is to serve as the exemplar for point i, taking into account other potential exemplars for point i. To begin with, the availability is initialized to zero.

The availability $a(i,k)$, from a candidate exemplar point k to point i, reflects the accumulated evidence for how appropriate for point i to choose point k as its exemplar, taking into account the support from other points. The following availability update gathers evidence from data points to decide whether the candidate exemplar can make a good exemplar.

When updating the two messages, it is important to add damping factor to avoid possible oscillations arises. Each message is set to γ times of its previous iteration value plus $1-\gamma$ times its prescribed updated value.

3 AP-GKAs

Based on the AP algorithm and the improved GKA (GKAs), a customer segmentation model AP-GKAs is proposed. The steps is: (1) **Extract feature**. Factor analysis constructs the feature attribute and weights the feature by the normalized multi-indicator RFM model; (2) **Determine the clustering quantity**. The AP algorithm is used to cluster the data obtained in step (1) and then get range of clusters, then two evaluation indicators are used to determine the number of clusters; (3) **Cluster**. GKAs clusters the transaction data obtained in step (1) with the number of cluster obtained in step (2).

3.1 Extract Feature

Extracting feature which firstly establishes muti-indicator RFM model, and then uses factor analysis to construct the implied feature and weight on each feature from the normalized multi-index RFM model.

Muti-indicator RFM Model. RFM model describes the customer overall trading behavior through recency (R), frequency (F) and monetary (M). However, customer behavior is a complex phenomenon, and it still has some drawbacks in using three indicators. For example, the transaction with the same attribute R can not determine the customers new or old property. The AP-GKAs uses the multi-indicator RFM model [7] to reflect customers transaction information.

As shown in Table 1. The multi-indicator RFM model uses ten indicators to describe the transaction information comprehensively. After obtaining indicators, it is unable to carry out unified measurement for big different between each indicator in magnitude. Therefore, ten indicators are normalized to improve the model accuracy.

Table 1. Muti-indicator RFM system.

Tradition indicator	Improved indicator
Recency (R)	Recent purchase (R_r), Farthest purchase (R_f), First quartile (R_{q1}), Median purchase (R_{q2}), Third quartile (R_{q3})
Frequency (F)	Sum frequency a month (F_{sum}), Max frequency a month (F_{max}), Min frequency a month (F_{min})
Money (M)	Sum money (M_{sum}), Average money (M_{avg})

Factor Analysis. Nevertheless, each weigh is different in empirical analysis. The widely used method is the analytic hierarchy, but it is more inclined to the plan decision than the weight determination. Accordingly, we use factor analysis method to find out implied feature and weight on each feature.

$$X = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \bullet factor \tag{1}$$

where X is multi-indicator RFM indexes, $factor$ is public factor.

3.2 Determine the Clustering Quantity

Traditional algorithms use K-means to determine the number of clusters, whose cluster centers are often initialized randomly for each number, therefore the results have poor comparability of validity indices and the inaccurate K.

Most researchers use rules $K_{max} \leq \sqrt{n}$ to determine the K_{max}, but this conclusion is based on the premise of uncertainty function, nothing that this assumption is not a sufficient condition [8]. The AP algorithm can get more exact cluster range, hence the upper limit of K changed to the result from the AP.

Sum of Square Errors: The sum of square errors reflects the similarity of the data points in same cluster. The higher similarity of data points in the same cluster displays, the smaller sum of square errors is.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{2}$$

where k is the number of cluster, and μ_i is the center of cluster i.

Silhouette Coefficient: The silhouette coefficient combines two factors of cohesion and resolution, which can be used to evaluate different algorithms using same original data or the influence of different operation modes on clustering results. The clustering results are excellent with larger Sil average values.

$$Sil(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

where $a(i)$ is the mean distance between point i and other points in same cluster. $b(i) = \min\{b_{i1}, b_{i2}, ..., b_{ik}\}$ where b_{ij} is the mean distance of point i to points of other clusters. $Sil(i)$ is from -1 to 1.

3.3 Cluster

The GKA [3] algorithm, overcomes the defect that K-means is easy to fall into local optimum and improves the search scope of GA. However, the algorithm still has the shortcoming of premature convergence. The AP-GKAs uses adaptive mutation rate to solve the problem of premature convergence. The steps is:

Coding: Real number coding based on clustering center.

Initialization: Randomly generated initial population.

Selection: Elitist strategy and roulette wheel strategy. The fitness function is:

$$F(s_i) = \frac{Between(s_i)}{1 + SSE(s_i)} \tag{4}$$

where $Between$ is the distance between each cluster using the average connection. SSE shown in Eq. (2).

Mutation: Taking a random value instead of the original gene.

Standard genetic algorithms have been confirmed that they fail in converging to the global optimal solution [9]. Therefore, the AP-GKAs uses the adaptive mutation rate to solve immature convergence.

$$P_m = \begin{cases} p_{m1} & f \leq f_{avg} \\ p_{m1} - \frac{(p_{m1} - p_{m2})(f_{max} - f)}{f_{max} - f_{avg}} & f > f_{avg} \end{cases} \tag{5}$$

where p_{m1} and p_{m2} are mutation rate; f is the fitness value of the individual to be mutated in the population; f_{avg} and f_{max} is the mean and the largest of fitness value in population. The mutation probability will be larger (lower) for individuals with small (larger) fitness, which makes the AP-GKAs algorithm keeping the population diversity and ensuring the convergence of the algorithm.

K-means operation: (1) The best individual as the cluster center and reassign each object. (2) Calculate new cluster center to replace the worst individual.

4 Experiments

The AP-GKAs model is suited to customer transaction data which mainly contains three attributes: **transaction ID**, **transaction time**, **transaction money**. The experiments are conducted on two kinds data sets. The first is the online retail business data with 8 attributes in the UK (ORDS) [10]. ORDS contains 4371 customers with 541909 trading data. The second is the card transaction data with 11 attributes from a bank in China (CBDS). CBDS contains 4500 bank customers with over one million transaction for five years.

In two experiments, the reference degree $p = s_m$ and damping factor $\gamma = 0.5$ in the AP. In AP-GKAs, the population size m is 20, the mutation probability $p_{m1} = 0.01$ and $p_{m2} = 0.001$, and the maximum number of iteration $T = 100$. In SOM-GKA, SOM is used to determine the initial cluster center, the learning rate is 0.25. GKA sets the same parameters except the mutation rate $p_m = 0.01$.

4.1 Results Based on ORDS

The maximum number of clusters is 35 through the AP, it greatly reduce calculation time to $\sqrt{n} = 66$. For cluster number 2 to 35, we calculate the SSE and Sil value 50 times with K-means and average. In theory, the clustering quality should be in proportion to the Sil and in inverse to the SSE. Figure 1 shows the

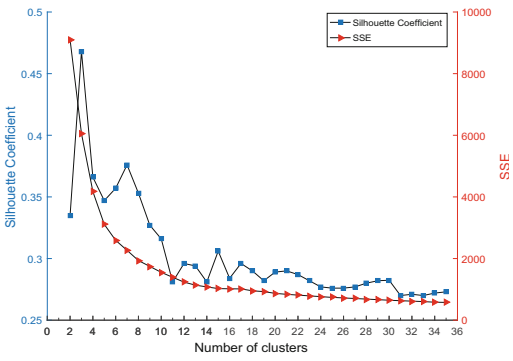

Fig. 1. SSE and silhouette coefficient value in different K of ORDS.

SSE value has converged to a minimum value with the larger Sil value when $K = 15$, hence the number of clusters is 15.

As can be seen from Table 2, AP-GKAs achieved lower SSE, higher Sil, $Between$ and F than the others. K-means has the lowest accuracy, GKA [3] has more improvement compared with K-means. SOM-GKA is better than GKA due to the initial cluster center determine by SOM [2]. Compared with GKA, AP-GKAs has 9.4% and 2.5% improvement in Sil and F.

Table 2. GKAs results compared with other algorithms in OBDS

Algorithm	SSE	Sil	$Between$	F	Iteration
K-means	1209.316	0.26345	10828.7	8.947	**29**
GKA	1138.269	0.29546	11215.8	9.853	50
SOM-GKA	1108.568	0.30812	11265.5	10.153	42
AP-GKAs	**1080.794**	**0.33734**	**11268.3**	**10.416**	35

4.2 Results Based on CBDS

The maximum number of clusters is 14 through the AP. From Fig. 2, the SSE value has converged to a minimum value with the larger Sil value when $K = 9$.

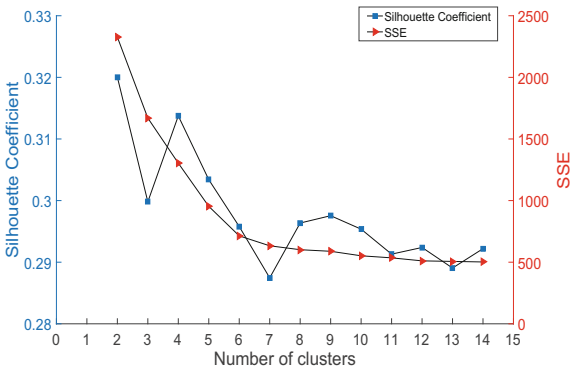

Fig. 2. SSE and silhouette coefficient value in different K of OBDS.

Table 3 shows that 7.36% improvement is displayed in Sil and 3.7% improvement is displayed in F compared with GKA algorithm. Moreover, AP-GKAs converges faster than GKA. Its convergence rate is less slower than K-means, but the clustering effect of AP-GKAs is much better than K-means.

Table 3. GKAs results compared with other algorithms in ORDS

Algorithm	*SSE*	*Sil*	*Between*	*F*	Iteration
K-means	694.441	0.1993	2198.377	3.161	**27**
GKA	590.38	0.2975	2507.573	4.24	34
SOM-GKA	584.961	0.2965	2514.966	4.292	35
AP-GKAs	**574.026**	**0.3194**	**2528.902**	**4.398**	29

5 Conclusions

In this paper we have shown how to completely process the customer transaction data for segmenting customer. We used muti-indicator RFM to describe the behaviour closely because the customer transaction is a complex phenomenon. It is always difficult to determine the number of clusters without expert knowledge for customer segmentation. To overcome this problem, we introduced AP algorithm to get range of cluster quantity. Moreover, we showed GKAs with adaptive mutation rate and one-step K-means operation to increase the accuracy of clustering result. We have successfully validated our model on standard data set and real transaction data, AP-GKAs has the fastest convergence rate and more accuracy than other three algorithms.

References

1. Kashwan, K.R., Velu, C.: Customer segmentation using clustering and data mining techniques. Int. J. Comput. Theory Eng. **5**(6), 856 (2013)
2. Hu, G., Yu, X., Huang, Q.: SOM neural network-based mobile client segmentation study. Microcomput. Appl. (2015)
3. Krishna, K., Murty, M.N.: Genetic K-means algorithm. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **29**(3), 433–439 (1999)
4. Girsang, A.S., Tanzil, F., Udjaja, Y.: Robust adaptive genetic K-means algorithm using greedy selection for clustering, pp. 1–5 (2017)
5. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(5814), 972–976 (2007)
6. Wang, K., Zhang, J., Li, D., Zhang, X., Guo, T.: Adaptive affinity propagation clustering. arXiv preprint arXiv:0805.1096 (2008)
7. Zeng, X., Xu, Q., Zhang, D.: New multi-indicator customer segmentation method based on consuming data mining. Appl. Res. Comput. **30**, 2944–2947 (2013)
8. Yu, J., Chen, Q.: The search range of the best clustering number in fuzzy clustering method. Sci. China Ser. E **32**(2), 274–280 (2002)
9. Rudolph, G.: Convergence analysis of canonical genetic algorithms. IEEE Trans. Neural Netw. **5**(1), 96–101 (1994)
10. Chen, D., Sain, S.L., Guo, K.: Data mining for the online retail industry: a case study of rfm model-based customer segmentation using data mining. J. Database Mark. Cust. Strat. Manag. **19**(3), 197–208 (2012)