Customer segmentation model based on two-step optimization in big data era

Wei Gao[1, a *], Huiting Jia[1, b] , Ruzhen Yan[2, c]

[1]Business School, Sichuan Agricultural University, Chengdu, China

[2]Business School, Chengdu University of Technology, Chengdu, China

[a] Email: gaowei@sicau.edu.cn

Keywords: Big data, customer segmentation, two-step optimization model, data mining

Abstract. With the advent of the era of big datasets, real-time data is becoming increasingly important in assisting the decision making process for commercial banks. In this paper, we develop a two-step optimization model (FSGA-FCEN) based on genetic algorithm (GA) and cluster ensemble (CE) for customer segmentation. Firstly, the key attributes are selected using GA. Then FCEN algorithm is used to segment customers into small groups. Taking 3544 customers in a commercial bank as samples, empirical results show that, compared with K-means, FCM and MAJ models, two-step model is an efficient and practical tool for customer segmentation.

Introduction

In recent years, the credit card business, as one of the important benefit for commercial banks, is developing rapidly. Although commercial banks owe to mass data in big data, most of commercial banks in China have not carried out the effective analysis of credit card customers, and the homogenization of competition means among the banks are outstanding. In order to obtain a higher efficiency, commercial banks must pay attention to customer relationship management and maximize customer value based on customer segmentation [1]. The so-called customer segmentation, according to the differences among the characteristics of customer demand and purchasing behavior, buying habits, reputation and other aspects of the situation, is the classification process of dividing the customers into a plurality of groups of customers [2].

Generally speaking, the method of customer segmentation includes experience description, statistical method and non-statistical method, etc. [3, 4]. Customer segmentation models based on statistical methods divide categories of customers according to the customer statistical characteristics, for example, sex, age. Non-statistical method mainly is to use data mining technology into customer segmentation [5-8]. These are the growing customer segmentation method in recent years. For example, Boone and Roehm [6] studied Hopfield-Kagmar (HK) clustering method of customer segmentation using Hopfield's artificial neural network technology. The study has shown that, each neuron in HK clustering method is connected with other neurons , and information can flow between neurons in multiple directions, which is more suitable for customer segmentation than the K-means clustering method; Kim et al. [7] used neural network clustering method to segment the customers of tourism; contrasting K-means, self-organizing map neural network and particle swarm optimization for three kinds of clustering algorithm, Deng et al. proposed hybrid clustering algorithm which was used for segmentation problem of catering industry customer. The above data mining customer segmentation method is constantly applied to the research of customer segmentation, and has achieved good effective in some empirical studies([9], [10]). Therefore, the choice of effective variables becomes the key problem in the research of customer segmentation.

Two-stage model of customer segmentation based on FSGA-FCEN

Compared with the researches of Feature Selection Based on Supervised Learning, the researches of Feature Selection Based on Non-supervised Learning are less. Fowlkes et al. [10] proposed the forward selection algorithm to solve the variable selection problem of clustering, starting from a single feature, and then adding a good feature to the current feature subset. Some researchers ([11, 12]) proposed the method of clustering based on GA. These methods use GA to find the best center of clusters and the number of the clusters, rather than be used to solve the problem about irrelevant variables. Milligan and Copper[13] has compared 30 overall types of clustering criterion. This research indicated that any criterion is better than other criterion in all cases. Reference [14] proposed the FCEN and applied to customer segmentation. The main idea of FCEN is regarding Fuzzy Clustering Algorithm (FCM) as basic cluster , using a method which is similar to Bagging to produce multiple basic clusters.

The processing of two-stage model FSGA-FCEN is as follows:

Step1: Input data set $X_{n´m}$, and encode the feature which length by Binary code;

Step2: Using formula (1), to calculate the fitness value, then using the roulette selection method to select chromosomes which will inhe$m$ rit to the next generation;

Step3: Using single point crossover operator with cross probability P_c to operate the cross calculation;

Step4: Using basic bit mutation operation with P_m to change some gene values in individual coding series and form a new individual;

Step5: Repeat step2, 3, 4 until the number of valid features m¢ is selected;

Empirical analysis

Data in this research roots in the source data of a Chinese city business bank’s credit card system, which include basic information about customers and trading records details. Basic information about customers consists of age, gender, education background and marital status, etc. Trading records details contain trading date, transaction amount and transaction type, etc. This paper selects the credit cards’ data from June 1st 2008 to November 30th 2010.There are 3,544 customers and 157,756 trading records involved in total.

Considering the various factors of bank customer segmentation, the related literature about domestic and foreign bank customer segmentation, and the availability of index data, this paper adopts FSGA to select variables from above based on the demographic information. On the progress of selection, the scale of genetic algorithm’s initial population is set as 10, crossover probability 0.6, mutation probability 0.001.

Table 1 Likelihood ratio checking of multiple regression analysis

Variables	Chi-square	df	Sig
Gender	4.2079	3	0.2389
Age	5.9835	1	0.1217
Education background	15.823	4	0.0004
Marital status	9.096	3	0.027
Dependents	10.377	1	0.0288
Housing conditions	38.413	2	0.0002
Job characters	1.1025	6	0.7935
Industry type	13.4009	3	0.0038
Annual income	8.648	3	0.0089
Customer type	1.759	2	0.6239
Line of credit	11.689	4	0.0041

Note: significance level 0.5

Furthermore, Logistic, a multiple regression analysis is put to use to prove the validity of FSGA, and the result is showed below (see table 1). Sig values suggest that the 4 variables including gender, age, job characters and customer type change not so notable, almost accord with the 4 variables selected. FSGA is proved to be capable of excluding most of the insignificant elements.

After selection of variables, final 7 customers demographic information based variables and 5 customers' value based variables are pinned down. Altogether 12 variables are chosen to be customers' value segmentation models. This paper uses clustering ensemble to subdivide credit card customers. During clustering ensemble, setting ensemble size as 20, cluster number 4, and take the average value among these 10 experiments data as clustering result(see table 2).

Table 2 FCEN segmentation results

C1	C2	P	L	F	A	C3	R
1	9.00	319	13.00	54.25	72960.97	23	1
2	41.86	1484	32.12	42.93	47394.75	15.61	0.98
3	13.76	488	28.72	43.65	38515.09	16.07	0.96
4	35.38	1254	38.61	30.28	33708	15.13	0.95
avg.			31.02	43.07	40785	15.7	0.97

Note: C1 represents customers subdivision; C2 represents customer number (%); P represents proportion; L represents latest purchase interval (day); F represents frequency (time); A represents amount (Yuan); C3 represents customer relationship length (month); R represents repayment capacity.

To clarify the superiority and inferiority of FCEN in customer segmentation in this paper, comparing FCEN with simplex clustering algorithm k-means, FCM and the most commonly used MAJ (major voting based) in demonstration. In this experiment, setting ensemble model as 20, clustering number 4, repeated experiment 10. Without clear cluster labels, the practical data tank is not so compatible with ideal data structure, so, clustering variance (ocq) [15] based on clustering distribution are chosen as a means to compare within the 4 means above in demonstrations. It is conspicuous that the smaller clustering variance is, the higher ocq will be, which means the result of clustering will be better as well. Chart 3 shows 4 different results from 4 respects. (See table 3).

Chart3 shows the clustering variance of FSGA-FCEN is definitely lower than model k-means, FCM and MAJ, the former is 2269.7, the latter 3 are respectively 6179.1, 3038.1 and 2624.8, but its ocq is apparently higher than the 3, the former is 0.8636, the latter 3 are accordingly 0.7823, 0.8156 and 0.8437. It turns out that model FSGA-FCEN in this paper applying on credit card customer segmentation is effective and reliable.

Table 3 Results Comparing

Means	k-means	FCM	MAJ	FCEN
Clustering variance	6179.1	3038.1	2646.8	2269.7
ocq	0.7823	0.8156	0.8437	0.8636

Conclusions

Customer segmentation is the prior task to make an efficient marketing strategy and also the cornerstone of customer relationship management. Affected by factors like social environment and customers' psychology, actions taken by customers are always with complexity, which increase the difficulty of segmentation, and get key attributes of customer segmentation on the premise of losing no information, which deplete the number of FCEN model's input end data greatly and get rid of the influences on seg-mentation results by irrelevant variables. Further-more, FCEN as a positive customer segmentation model optimizes validity and robustness, and can be utilized to subdivide the simplified sample data. The result shows that compared with other 3 single clustering algorithms, k-means, FCM and MAJ, FSGA-FCEN method in this research has performed the best, that is, FSGA-FCEN is the ideal means both efficiency and practicability.

It cannot be denied that there are indeed some shortcomings in this model, for instance, the low value customers, whose quantity is big, are treated as their high value counterparts will lead to a hefty cost on service providing. So, this is a significant part in our follow-up works on how to enhance the ability over customer segmentation.

Acknowledgements

This paper is supported by the Key Projects of Educational Commission of Sichuan Province of China (13ZA0140), the Program of Undergraduate Thesis Foster of Sichuan Agriculture University (34009114), and the Project supported by the National Natural Science Foundation of China (Grant No. 71501018).

References

[1] X.Q. Zeng, Q. Xu and D Zhang, New multi-indicator customer segmentation method based on consuming data mining, Application Research of Computers, 30 (2013), 2944-2947.

[2] S.Y. Kim, T.S.Jung, E.H. Suh, Customer segmentation and strategy development based on customer lifetime value: a case study, Expert systems with Applications, 31 (2006), 101-107.

[3] P.V.Freytag, Business to business market segmentation, Industrial Marketing Management, 30 (2001), 473-486.

[4] L. Chen, K. Soliman, S.E. Mao, Measuring user satis-faction with data warehouse: An exploratory study, In-formation & Management, 37 (2000), 103-110.

[5] H.V. Shashidhar, V. Subramanian, Customer segmentation of bank based on data mining security value based heuristic approach as a replacement to k-means algorithm, International Journal Computer Application, 19 (2011), 13-18.

[6] D.S. Boone, M. Roehm, Retail segmentation using arti-ficial neural networks, International Journal of Re-search in Marketing, 19 (2002), 287-301.

[7] J Kim, et al. Segmentation the market of West Australi-an senior tourist using artificial neural network [J]. Tourism Management, 2003, 24 (1): 25-34.

[8] X.Y. Deng, C. Jin, Q.P. Han, KSP: a hybrid clustering algorithm for customer segmentation in mobile E-commerce, Journal of Management Science, 24 (2011), 54-61.

[9] A. Strehl, J. Ghosh, Cluster Ensembles: a knowledge reuse framework for combination multiple partition, Journal of Machine Learning Research, , 3 (2002), 583-617.

[10] E.B. Fowlkes, R. Gnanadesikan, J.R, Kettenring, Vari-able selection in clustering, Journal of Classification, 5 (1998), 205-228.

[11] S. Dudoit, J. Fridlyand, Bagging to improve the accura-cy of a clustering procedure, Bioinformatics, 19 (2003), 1090-1099.

[12] Z.H. Zhou, W. Tang. Cluster ensemble, Knowledge-Based Systems, 19 (2006), 77-83.

[13] G.W. Milligan, M.C. Cooper, An examination of pro-cedures for determining the number of cluster in a data set, Psychometrika, 50 (1985), 159-179.

[14] W. Gao,C.Z. He, X.Y. Jiang, Customer segmentation study based on fuzzy clustering ensemble, Journal of Intelligence, 30 (2011), 125- 129.

[15] Y. Yang, F. Jin, M. Kamel, Survey of clustering validity evaluation, Application Research of Computer, 25(2008), 1630-1633.