ICIC Express Letters
Part B: Applications
Volume **7**, Number **5**, May **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **1077–1082**

APPLICATION OF AN IMPROVED CLUSTERING ALGORITHM IN CUSTOMER RISK PREFERENCE RECOGNITION MODEL

Lin Xiao[1,2,3], Shenghua Xu[1], Jeng-Shyang Pan[2,3] and Tonghua Yang[1]

[1]College of Information Management
Jiangxi University of Finance and Economics
No. 665, West Yuping Road, Nanchang 330013, P. R. China
xiaolin201@qq.com

[2]College of Information Science and Engineering
Fujian University of Technology
No. 3, Xueyuan Road, University Town, Minhou, Fuzhou 350108, P. R. China

[3]Key Laboratory of Big Data Mining and Applications of Fujian Province
Fuzhou 350108, P. R. China

Received October 2015; accepted January 2016

Abstract. *The main problem of this paper is to establish a recognition model for the risk preference of brokerage individual customer. It is based on transaction data. Its core algorithm is the use of improved K-means algorithm. Its process includes the principal component analysis and clustering operation of the data. In this paper, we improve the design of automatic merging approximation class in clustering algorithm. The main basis of customer risk is the investment subject of the customer. The contribution of this research is to promote the accuracy and success rate of marketing and provide a good support for the company to make reasonable marketing decision.*
Keywords: Customer classification, Risk preference, Clustering, Securities Company

1. **Introduction.** As securities and stock investors grow and mature, they gradually realized the securities market is difficult to forecast and hold, and reflected on their own investment behavior and concept. Especially individual customers began to think their own risk characteristics and focus on the products and services that is matching their risk attributes. On the other hand, in the actual marketing, brokerage firms can provide so many products for investors. For Securities Company, how to find the customer preference for the product and recommend the suitable products to him (or her) is the key to increase the success rate of the marketing. Therefore, Securities Company needs to understand customers about the customer characteristics and needs, to provide the appropriate products and services to customers. Securities Company can excavate client's inherent characteristics and needs by their external behaviors, and apply the data mining to classifying and management customers. This is also the Chinese stock market regulations and requirements, which is referred in [1].

In [2-6], many scholars expounded the application of data mining technology in the securities industry. For example, Wu and Shi used concept clustering to analyze the trading behavior of customers. They researched customers behavior of buying and selling, and concluded the general rule that affects clients' profit and loss; Kuo et al. researched new two-stage method that applied clustering method to classifying customer; Liang describes using fractal clustering method to classify customers in the aspect of financing, contribution and trading frequency; Qian and Wang put forward the multidimensional model of securities customer segmentation. The above research is mostly based on customer value or customer loyalty or customer life cycle to manage the customer classification. There is few to analyze portfolio risk appetite of investors based investment target. Therefore,

this paper puts forward a risk preference analysis model based on consumer's transactions, which has a practical significance.

2. **Problem Statement and Model Construction.** This article puts forward a model aimed at identifying the risk preference for individual customers of Securities Company. The main idea is to use data mining technology to excavate the data of individual customers trading behavior, so as to find the hidden relevant rules about risk preference of customers. The model framework discussed is shown in Figure 1.

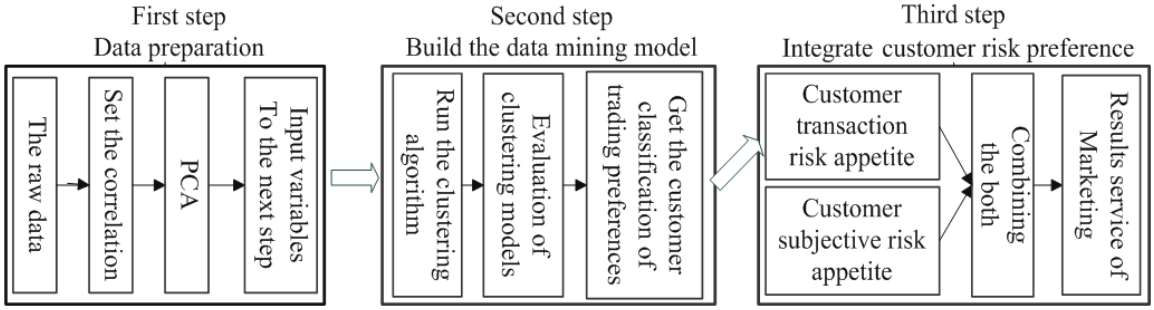

FIGURE 1. Model of the risk preference identification for individual customers

The first step is data preparation. This article uses Principal Component Analysis (PCA) to simplify the data. The raw data is the real transaction data of a securities company. The data matrix after PCA is input variables to the next step. The second step is to build a data mining model whose form is loop iteration. It is mainly based on modified K-means algorithm to achieve classification results with the customer's trade categories. The third step is to integrate customer risk preference. On the one hand, based on the clustering results, namely the customer's trading category according to the risk value of the portfolio in the category, customer's transaction risk preference can be identified. On the other hand, we can get customer's subjective risk appetite by securities through online questionnaire survey of the trading system. We can get the final result of customer's risk preference by combining the both.

2.1. **Data preparation process.** In this paper, the purpose of the PCA is to reduce the dimensions of the data set, which is referred in [7]. The PCA process follows three steps: firstly, it makes the original data to eliminate the dimension influence by transformation of Z standardization; then it calculates the correlation coefficient matrix; finally, composition analysis is conducted and obtains the main components matrix. The steps details are described by Procedures 2.1.1 to 2.1.3.

Procedure 2.1.1. *If there is an original matrix X as shown in the matrix*

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ & & \vdots & & \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}$$

that has n samples, and p is a variable number, the formula of Z transform of standardization for X is shown in Formula (1).

$$x^*_{aj} = \frac{x_{aj} - \bar{x_j}}{\sigma_j} \tag{1}$$

In Formula (1), $\bar{x_j} = \frac{1}{n}\sum_{a=1}^{n} x_{aj}$, $\sigma_j = \sqrt{\frac{1}{n}\sum_{a=1}^{n}\left(x_{aj} - \bar{x_j}\right)^2}$, $j \in 1, 2, \ldots, p$. The purpose of Formula (1) is eliminating the dimension influence in the matrix X.

Procedure 2.1.2. *The correlation coefficient matrix can reflect the correlation of data. This paper uses the traditional Formula (2) to calculate the correlation coefficient matrix for X, namely R that is an X co-variance matrix.*

$$R_{ij} = \frac{\sum_{i=1}^{n}\left(x_{ij} - \bar{x_j}\right)\left(x_{ik} - \bar{x_k}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ij} - \bar{x_j}\right)^2 \sum_{i=1}^{n}\left(x_{ik} - \bar{x_k}\right)^2}} \tag{2}$$

Procedure 2.1.3. *The next operation is to calculate characteristic value (λ) and orthogonal unit feature vector (a_i) for co-variance matrix R, among $i \in 1, 2, \ldots, p$. Then, according to Formula (3) we could calculate the comprehensive vector index (F_i) by the feature vector (a_i).*

$$F_i = a_{1i}x_1 + a_{2i}x_2 + \cdots + a_{pi}x_p, \quad i = 1, 2, \ldots, p \tag{3}$$

Finally, it needs to calculate the variance contribution rate of comprehensive vector, and can determine the principal components by comparing with the cumulative contribution rate. And by the principal components, its matrix (F) of the sample can be calculated out.

2.2. **The improved K-means clustering algorithm.** This paper applies the improved K-means algorithm to doing clustering analysis for the principal component matrix F. The algorithm is used in the second step of the model. In [8], it mentions that K-means algorithm is the classical clustering analysis method in data mining. Due to that the customer's sample data of Securities Company is large, K-means is more suitable for it.

The traditional K-means algorithm steps are showing below.

a. K points were randomly selected as the initial clustering center.

b. It calculates each sample's distance to the center of the cluster, and the sample is attributed to the cluster which has the nearest clustering center with it.

c. It calculates the new clustering center by the sample's average in each cluster.

d. Go back to step b, from b to d loop iteration until all cluster centers are stable, end the algorithm.

What the algorithm is improved in this paper has two aspects which are the selection strategy of initial cluster center and the deleting centers strategy about the small clusters that is near larger cluster.

The selection strategy of initial cluster center is as follows. First, because X has m components, it selected m samples as the cluster centers, in which each sample has a maximum of one component. Then find out a sample from the rest of the samples as a new clustering center, and the new one must satisfy the condition that the sum of its distance with all those who have been chosen as cluster centers must be the largest. Based on the rule, it constantly finds out new cluster center, until it has k number of cluster centers. The advantage is that the distribution of initial clustering center is relatively uniform, and the disadvantage is that a little running time is wasted. However, it, for the modern computer, is negligible.

The delete centers strategy is described below. The traditional K-means algorithm is not to cut the clustering center, which is referred in [2]. Improvement strategy is to count the number of objects in each cluster when every clustering ends. The cluster center which is satisfying Inequality (4) will be deleted.

Definition 2.1. *For any cluster C_i and its most adjacent clusters C_j, if they meet the following two conditions, C_i center will be deleted: (I) the objects number of C_i is less*

than the objects number of C_j; (II) the distance of their cluster centers is less than twice of the distance that is the C_i center to its farthest objects.

$$num\_C_i < num\_C_j \ And \ d\_C_{ij} < 2 * \max\_C_i \tag{4}$$

If C_i is deleted, its objects would be assigned to the near clusters based on the principles of the closest distance in the next clustering. It could make the number of clusters be combined to a relatively reasonable value based on the principles of adjacent. As shown in Figure 2, there have cluster C_1 and cluster C_2. One of their centers will be deleted and their objects will be combined into a large cluster in the next clustering. Some clusters that have fewer objects but far from other clusters own independent features and are not easy to meet Inequality (4). Thus, it will not be deleted. In order to implement algorithm, this paper designed cluster node information as shown in Table 1.

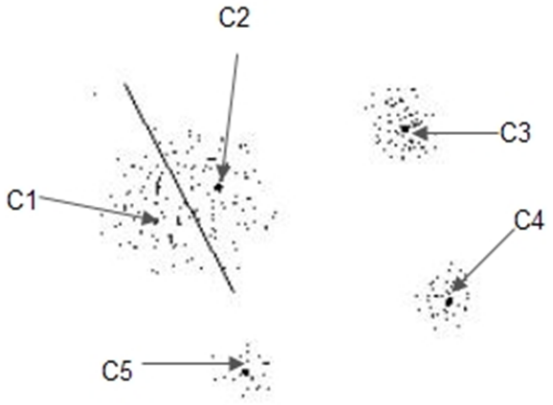

FIGURE 2. Deleting C_1 or C_2

TABLE 1. Cluster node information

Cluster No.	*Cluster center*	*total number of its samples*	*Nearest cluster No.*	*The max distance of its samples*	*Next cluster address*

Definition 2.2. *The improved K-means algorithm steps are as follows.*

a. It identifies the initial center of K clusters.

b. It identifies the most neighboring cluster of each cluster, and records the distance between the two centers.

c. All the sample object is assigned to its neighboring clusters. Then it updates the total number of objects and the farthest object distance for each cluster at the same time.

d. When the object allocation is over, adjust every cluster center according to the average distance between objects in one cluster after the object allocation process. If the cluster centers are not shifted, the algorithm is terminated.

e. According to Inequality (4), the algorithm examines each cluster and deletes the cluster centers which meet the conditions. When it deletes a cluster center each time, the value k reduces to 1. Algorithm needs to modify the value of the corresponding node pointer field to keep the list completed, and return to step b.

3. **Empirical Study.** This paper takes the example of X Securities Company in China, using the company data from April to June in 2013. This original data is calculated by using Formulas (1), (2) and (3), which can obtain the results that are the cumulative contribution of variance of main components. It is shown in Table 2 that the total variance what is explained by the 10 comprehensive indexes can represent 95.23% original information. This paper based the 10 comprehensive indexes to establish the principal

Table 2. The cumulative variance contribution of each component

The explanation of total variance							
component	*The initial Eigen value*			*component*	*The initial Eigen value*		
	combined	*variance %*	*cumulative %*		*combined*	*variance %*	*Cumulative %*
1	3.361	14.615	14.615	13	.075	.325	98.587
2	2.517	10.945	25.559	14	.052	.225	98.812
3	2.128	9.252	34.811	15	.051	.220	99.032
4	2.054	8.930	43.741	16	.049	.212	99.244
5	2.014	8.754	52.495	17	.046	.201	99.445
6	1.990	8.652	61.148	18	.031	.134	99.578
7	1.987	8.639	69.786	19	.030	.130	99.708
8	1.958	8.511	78.298	20	.027	.119	99.827
9	1.950	8.479	86.777	21	.021	.090	99.917
10	1.945	8.458	95.234	22	.011	.046	99.963
11	.453	1.969	97.203	23	.008	.037	100.00
12	.244	1.059	98.262	*NA*	*NA*	*NA*	*NA*

Table 3. The clustering results for X company customers

Groups	*Character description*	*Ratio of product configuration*
Class-1	*Customers Like new shares subscription*	*More than* 70% *of assets to purchase new shares*
Class-2	*A + B shares mixed customers*	*About* 41% *of the A shares,* 30% *of B shares*
Class-3	*A shares + ST stock investment clients*	*About* 53% *of A shares and* 28% *ST shares*
Class-4	*B shares customers*	*More than* 87% *positions at B stock*
Class-5	*Based on the fund investment clients*	*More than* 50% *on fund investment*
Class-6	*Clients like ST stock investment*	*More than* 78% *ST stock proportion*
Class-7	*Customer type is equilibrium configuration*	30% *a-shares,* 31% *internal funds,* 15% *open stock fund*
Class-8	*Customers preference GEM stocks*	*About* 67% *of the GEM,* 19% *of the A shares*
Class-9	*Customers like stock fund*	*More than* 90% *investment stock funds*
Class-10	*Mixed customers*	*About* 25% *of A-shares, open fund* 25%, 8% *of new shares subscription, money fund* 12%
Class-11	*A-shares customers* 1	*More than* 82% *A-shares positions*
Class-12	*Clients like internal fund*	*About* 90% *of internal fund investment*
Class-13	*A-shares customers* 2	53% *A-shares,* 24% *GEM stocks,* 3% *new shares subscription*
Class-14	*A-shares customers* 3	60% *A-shares*, 15% *internal funds*, 3% *open stock fund*
Class-15	*Priority to Bond investment*	*Bond funds on average* 80% *and* 8% *stock fund*

component matrix. We used the improved K-means algorithm. When customer groups are classified into 15 categories, a relatively stable result is achieved. The 15 categories character description is shown in Table 3.

4. **Marketing Strategy.** The risk preference of clients will change as they learn and grow. Therefore, risk preference recognition model should be able to do the continuous tracking and dynamic assessment of customer risk attribute and only by constant update, could the customer risk attribute be reflected more objectively. Therefore, based on the model results this research proposes some strategies for Securities Company services, as follows.

Strategy 4.1. *When Securities Company provides appropriate services, the service designing can base on the customer's risk attribute and the provided services must meet the needs of their risk appetite to help customer to control the risk within the range adapted*

to their own ability. This is the use of the contribution of this paper clustering algorithm to help customers avoid possible super expected risk.

Strategy 4.2. *In special marketing, the Securities Company can first evaluate risk characteristics of the product and find out the customer type that matches with the product risk. Based on this, Securities Company could generate target customer lists for financial advisers to have more targets to carry out one-to-one marketing. This is using the contribution of the algorithm to improve the success rate of product sales.*

5. **Conclusions.** In summary, this article puts forward a customer risk analysis model. To the data, the research applies PCA method to eliminating the redundant variables and utilizes the improved K-means algorithm to build the model of customer segmentation and achieve customer grouping purposes. Then, the research puts forward the corresponding marketing strategy. The innovation of this paper focuses on transactions subject to dividing the customer group, dynamic correction to reflect customer growth and preference changes and proposed for different target groups to provide matching product marketing strategy. In future research, we will further improve the accuracy of the algorithm, as well as from the perspective of value for customer classification.

Acknowledgment. This work is partially supported by the National Natural Science Foundation Project (71261009) and Jiangxi province graduate student innovation foundation project (YC2015-B052) and Fujian University of Technology Foundation Project (CY-Z15092). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] H. He, J. Zhu and B. Xie, The study on security investment attitude from the viewpoint of data mining, *Statistical Research*, vol.25, no.9, pp.49-54, 2008.

[2] Y. Wang and Z. Li, A scale invariant feature transform based method, *Journal of Information Hiding and Multimedia Signal Processing*, vol.4, no.2, pp.73-89, 2013.

[3] F. Wu and P. Shi, Mining method of conceptual clustering for customers trading behavior analysis, *Micro Computer Application*, vol.16, no.5, pp.26-28, 2000.

[4] R. J. Kuo, L. M. Ho and C. M. Hu, Cluster analysis in industrial marker segmentation through artificial neural network, *Computers and Industrial Engineering*, vol.4, no.2, pp.391-399, 2002.

[5] M. Liang, *Application Research on Fractal Clustering Analysis in Customer Segmentation of Security Market*, Hefei University of Technology, Hefei, 2009.

[6] W. Qian and Y. Wang, Empirical research on customer segmentation of securities based on clustering, *Journal of Computer Applications*, vol.30, no.2, pp.495-498, 2010.

[7] X. Wang, *The Application of Multivariate Analysis*, 3rd Edition, Shanghai University of Finance and Economics Publishers, Shanghai, 2009.

[8] J. Hilala and A. Rovshan, Applying K-means clustering algorithm using oracle data mining to banking data, *Proc. of the 9th International Conference on Management Science and Engineering Management*, vol.362, pp.809-816, 2015.