A Novel Intelligence Recommendation Model for Insurance Products with Consumer Segmentation

Wei XU

School of Information, Renmin University of China, Beijing 100872, China
E-mail: weixu@ruc.edu.cn

Jiajia WANG

School of Information, Renmin University of China, Beijing 100872, China
E-mail: jiajiaunbreakable@126.com

Ziqi ZHAO

School of Information, Renmin University of China, Beijing 100872, China
E-mail: zhaoziqi@ruc.edu.cn

Caihong SUN

School of Information, Renmin University of China, Beijing 100872, China
E-mail: chsun@ruc.edu.cn

Jian MA

Department of Information Systems, City University of Hong Kong, Hongkong, China
E-mail: isjian@cityu.edu.hk

Abstract   As one of the financial industries, the insurance industry is now facing a vast market and significant growth opportunities. The insurance company will generate a lot transaction data each day, thus forming a huge database. Recommending insurance products for customers accurately and efficiently can help to improve the competitiveness of insurance company. Data mining technologies such as association rules have been applied to the recommendation of insurance products. However, large policyholders' data will be calculated when it being processed with associate rule algorithm. It not only requires higher cost of time and space, but also can lead to the final rules lack of accuracy and differentiation. In this paper, a recommendation model for insurance products based on consumer segmentation is constructed, which first divides consumer group into different classes and then processed with associate rule algorithm. The empirical results show that our proposed method not only makes the consumption of association rules analysis reduced, it has also got more effective product recommendation results.

Keywords   product recommendation; consumer segmentation; data mining; insurance

1   Introduction

Financial product recommendations were financial marketing problems, the purpose is to provide products and services met the personalized needs according to the characteristics of financial customers. Traditional financial marketing used cross-selling concepts and methods. Starting in the 1990s, the domestic and foreign scholars have made certain definitions to cross-selling. Jarrar and Neely[1] believed that cross-selling is a new way of marketing, CRM is used to find a variety of purchase requirements of current customers, then selling a wide range of products and services for them accordingly. Kamakura[2] put forward that cross-selling was referred to the efforts to make customers buy products at the same company. Because the company has established relationships with the customer, cross-selling products and services to the cost of the current customers are much lower than to find a new customer. Important cross-selling forecast model with latent trait model[3], acquisition model[4], survival analysis[5] and NPTB model[6]. Basic data for the customer of this kind of model makes the demographic characteristics and behavior characteristics, uses the method of statistics or econometrics model, the results can be achieved certain cross-selling opportunities.

But there are massive amounts of customer data and information in modern context, such cross-sell products recommended results from research are a bit too simple, and not easy to large-scale operations. Personalized recommendation methods and recommended system, established a link between the user and the product information of digging discoveries have been made by each user's potential interest in, so as to implement product recommendations. Personalized recommendation attracted much attention as an independent field of study since the 1990's[7]. "Collaborative filtering"[8] was first invented and applied in 1992 by researchers. Then 1997 ACM international conference put forward a automated collaborative filtering system called Grouplens[9], can realize the recommendation of the movie and news. In 1997 e-commerce recommendation system was defined by Resnick and Varian[10], then it was being widely used. Later, recommendation system theory and application research has been further deepened at home and abroad. Zhao et al.[11] put forward to increase two process of the personalized recommendation algorithm. Ronald Yager et al. studied the rough set theory to recommend improvements of the system rules. Liu[12] added customer life cycle theory to the study of product recommendations.

The financial sector has studied the application of personalized product recommendations. As early as 1996, Fletcher[13] and others studied Britain's major Banks marketing system and came to the conclusion that IT technology be applied into the marketing strategy to keep market advantage of the bank. Scholars study in our country, bank and insurance are both important industry of application products recommend. Li et al.[14] using collaborative filtering to construct a system of bank product recommendations. Wang and Wu[15] studied the application of collaborative filtering in cross-selling, using collaborative filtering for a set of simulated customer purchase information to forecast. Wang et al.[16] studied the application of clustering in implementation of cross selling life insurance.

This paper makes recommendation of insurance products using known customer buying behavior, having two major aspects of the different to previous research results. First of all, this paper selects the association rules as a way of product recommendation. Because similar

collaborative filtering recommendation according to the user to complete the personalization recommendation, need to know the user's purchasing behavior and the assessment for each product, the initial recommendation effect is poorer. As for the insurance products, we can directly get the customer's purchasing behavior information, but could not get the product ratings of customers. Secondly, set up the model of product recommendations based on customer segmentation, determining the classification of the customer first, and then we will recommend products to customers. At the same time of lower association rules analysis cost, recommend more targeted products for customers and improve the accuracy of the recommendation for insurance products.

2 Theoretical backgrounds

2.1 Introduction to neural network

Neural networks (NN) is a typical method of classification analysis. It sets a learning model through the training sample data , then applied to the test data set, to test the classification accuracy of the test model. Neural network is very suitable for the classification of the traditional statistical model problems, such as nonlinear problem. At the same time, it has a better ability to adapt in processing data and unknown noise data. At present the most commonly used Neural Network algorithm is the back-propagation Neural Network(BPNN)[17], which was put forward in the 1980's. Learning process is to carry out on the training sample by iterative processing, comparing network classification of predicted value and actual label. Modify the connection weights after each iteration, from each hidden layer to output layer, and finally reach the first hidden layer, to make the minimum mean square error of the predicted values and actual values. This iterative process keeps calculating until the training set is correctly classified so that the samples reaches a certain proportion, or the error reduces to an acceptable range.

BPNN can be divided into three steps: weights initialization, transmission input forward and error transmission backward.

Step 1: Weights initialization. Network's weights are initialized to small random numbers, usually every units also need to set up a small bias.

Step 2: Transmission input forward. Calculate each layer unit of the input and output. As to the units for input layer, the input and output are equal to the sample properties. For the hidden layer and output layer unit, input I_j and output O_j are expressed by the following formulae.

$$I_j = \sum_i W_{ij}O_i + \theta_j \tag{1}$$

$$O_j = \frac{1}{1 + e^{-I_j}} \tag{2}$$

Among them, W_{ij} is the connection weight between i and j, θ_j is the bias of unit j.

Step 3: Error transmission backward. NN puts the error between the predicted and actual values back to the first hidden layer, and modify the corresponding weights. The error

of the output layer and hidden layer units can be achieved by the following formulae.

$$Err_j = O_j(1 - O_j)(T_j - O_j) \tag{3}$$

$$Err_j = O_j(1 - O_j)\sum_k Err_k W_{ijk} \tag{4}$$

Weights and bias also updates accordingly by the following formulae.

$$\Delta W_{ij} = (l)Err_j O_i; W_{ij} = W_{ij} + \Delta W_{ij} \tag{5}$$

$$\Delta \theta_j = (l)Err_j; V_j = \theta_j + \Delta \theta_j \tag{6}$$

2.2 Introduction to association rules

Association rules are valuable connections among data set items. Support and confidence are two important indicators which measure the value of association rules. They represent the rule of the usefulness and the certainty. They have been defined by the following formulae.

$$support(A \Rightarrow B) = P(A \cup B) \tag{7}$$

$$confidence(A \Rightarrow B) = P(A \mid B) \tag{8}$$

When the rule meets the minimum support threshold and minimum confidence threshold, it is valuable strong association rules. Meaningful rules, can produce guiding role to the real market.

Association mining research regards every goods in the retail market as a data item. Item set is a collection of data items. K-itemsets is a set which contains k sets of items in data. Itemsets meet minimum support that occurrences to set the threshold of itemsets are called frequent itemsets. Association rules mining is mainly divided into two steps. The first step is to mine frequent itemsets according to the minimum support, then to find strong association rules by frequent itemsets and minimum confidence.

Apriori algorithm is put forward by Agrawal et al's[18] which is the most famous algorithm of association rules. The data model of association rules algorithm is a "market-basket" Boolean matrix. Lines represent the shopping baskets, and columns represent the commodity items. As a result, we get a sparse matrix. Operation method is to recycle $(k-1)$ itemsets to generate k itemsets. It uses the Apriori nature, a subset of the set of frequent itemsets is also frequent itemsets.

After digging out all the frequent itemsets, it can produce strong association rules which is in accordance with the minimum support and minimum confidence. In order to produce strong association rules in the frequent itemsets l, firstly need to produce all the nonvoid subset. The confidence of subset is calculated by the following formulas. If confidence $(s \Rightarrow (l - s))$ meets the minimum confidence threshold, an association rule is created.

$$confidence(s \Rightarrow (1 - s)) = \frac{support(1)}{support(s)} \tag{9}$$

In the process of generating frequent itemsets, Apriori algorithm will produce a large number of redundant candidate sets, leading to slow computing speed. So a lot of improved Apriori algorithms are designed to increase speed and improve the efficiency. Park et al.[19], using a hash table can reduce the number of candidates and the occupied space. Agrawal reduced

scanning the number of records of $k + 1$ itemsets according to k itemsets. Savasere used data partition technology to mine the local frequent itemsets[20], then merging to the global frequent itemsets. Brin proposed a dynamic counting method, adding candidate sets at different scanning moments[21].

3   The proposed method

In this part, this paper proposes a recommendation model for insurance products based on consumer segmentation. Firstly, we need to do the data preprocessing to the daily transaction records of insurance company. Then, the clustering analysis is used to determine the classification characteristics of the policyholders. Next, customer segmentation model is constructed using Neural Network, analyzing the influence degree of factors on the classification results made by policyholder personal information and vehicle information etc. Then, we use Apriori algorithm to obtain association rules and the propensity to buy of consumers. The product recommendation model is built. Finally, the precision, recall rate and F values which are widely used in the search engine evaluation are adopted to evaluate the performance of recommendation. We compare the product recommendation based on customer segmentation with general product recommendation, and come to the conclusion. The framework of the proposed method is illustrated in Figure 1, and the details will be discussed in the following subsections.
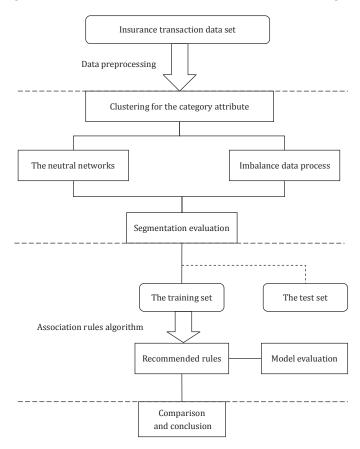

Figure 1   The framework of the proposed method

3.1 Data preprocessing

This paper mainly studies the customer segmentation and product recommendations, so the data comes from daily transaction records of an insurance company. This company's business is focusing on the property of vehicle insurance, and the trading records include the basic information of the customer, purchase products, insurance premium, and claims. According to characteristics of the database, each customer code (PK) can determine a single client, so the objects of study are 25,368 customers and 181,450 insurance records. After the original data is cleaned, integrated, and standardized, we integrate 25,317 policyholder purchase information as a matrix of $25317 * 22$. Horizontal represents insurance customers, vertical represents insurance products. The matrix is filled with binary values, and 1 means customers have bought this product, "0" means not to buy. In addition, we also construct a database of customer's purchase order , which records all the policyholder order of purchased products.

3.2 Clustering analysis for the category attributes

According to the customer value matrix, the value of the customer is decided by the current value and potential value. With reference to the policyholder's characteristics, insurance company's operating profits come from the insurance amount paid by policyholder, and losses come from the damages to the policyholders. We select two variables, policyholder's total premium and claims scoring, and use these two variables on behalf of the policyholder's current value and potential value.

This paper chooses clustering method based on classification. WEKA software has two kinds of clustering algorithm, SimpleKMeans and FarthestFirst. We use them to divide the data into four categories. We select more accord with actual situation of clustering result as the customer classification result, to do further research on the influencing factors of classification.

3.3 Constructing customer segmentation model

In the previous section we have got the categories sign of each customer by clustering, then we will use the typical classification method to determine the influencing factors of customer classification. Through the above data processing, we get a total of 11 classification indicators which can be used to describe the behavior of policyholder, as listed in Table 1.

Table 1 The descriptions of classification property

Attribute Name	Type	Value range
Gender	discrete	-
Location	discrete	-
Area	discrete	-
Use of years	discrete	-
Seats	discrete	-
Tons	discrete	-
Country	discrete	-
Color	discrete	-
Age	continuous	[19,80]
Price	continuous	[7000,300000]
Depreciation	continuous	[0,1]

The experiments use Neural Network as the main classification method, at the same time using logistic regression, decision tree and support vector machine (SVM) as a reference. Because data is not balanced, so using some algorithms of unbalanced data processing may lead to better results. By comparison with the result of the experiment, the model for customer segmentation will be built after choosing classification method with higher classification accuracy.

3.4 Constructing product recommendation model based on customer segmentation

The process of constructing product recommendation model is divided into two steps. Firstly, we need to get rules of the purchase by the training sets, then inspect the rules by the test set and make evaluation to the model. We use customer segmentation results in the previous section, make product recommendations as the classification of customers.

This paper selects the Apriori algorithm of WEKA software. The training sets randomly come from 80% of all kinds of policyholders' information, and the remaining 20% are conducted in the test sets. According to the definition of confidence, support and lift measure, the minimum confidence and support of the Apriori algorithm is set to 0.6 in order to choose the valuable rules which can produce guidance to the market. According to the condition that the lift measure is greater than 1, we extract the valuable rules from the candidate. After we get appropriate recommended products, it is necessary to make reasonable evaluation and comparison for the result. Evaluation indicators such as precision, recall rate and F value which are widely used in the field of search engine evaluation are used in this paper.

3.5 Comparison with general product recommendation model

In order to verify the validity of the insurance product recommendations model based on customer segmentation, it is necessary to construct general product recommendation model. Then we can compare the accuracy of the two models on the product recommendation.

For general product recommendation model, it is need to do association analysis of all the policyholder purchase information directly. Similarly, we select the Apriori algorithm of WEKA software. The training sets randomly come from 80% of all kinds of policyholders' information, and the remaining 20% are conducted in the test sets. The minimum confidence and support of the Apriori algorithm is set to 0.6. According to the condition that the lift measure is greater than 1, we extract the valuable rules from the candidate. Finally, we get precision, recall rate and F value of the experimental results. By comparing the experimental results and analysis, we draw the conclusion.

4 Empirical analysis

First of all, the experiments use SimpleKMeans and FarthestFirst which are the clustering algorithms in the WEKA software. The results of cluster analysis are listed in Table 2.

We can get specific distribution of four types of customers from the scatter diagram. With reference to the customer value theory, we can know that the result of K-means algorithm is more in line with the actual situation. So we use the result of K-means algorithm as the result of the customer classification, and do further research on the influencing factors of classification. The result of K-means clustering is shown in Figure 2.

Table 2 The results of cluster analysis

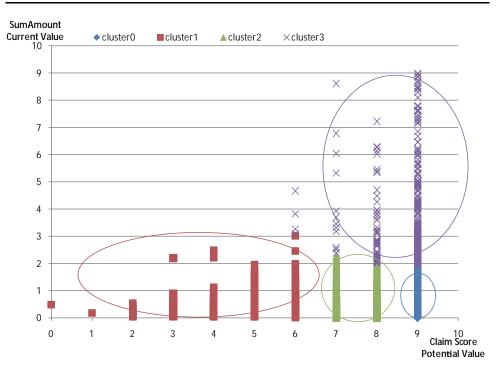| Cluster	SimpleKMeans			FarthestFirst		
	Number of customer	percentage	Value of customer	Number of customer	percentage	Value of customer
Cluster0	18553	3%	Class II	25033	99%	Class II
Cluster1	600	2%	Class I	128	1%	Class I
Cluster2	5798	23%	Class III	41	0%	Class IV
Cluster3	366	1%	Class IV	115	0%	Class III

Figure 2 Result of K-means clustering

Policyholders can be divided into four categories according to premium and the number of claims. This accords with customer value matrix. Policyholders with high premium and less claims, who are the most valuable, only account for 1% of the total number of customers. More than 70% of the customers are regular customers with low premium.

In the process of classification to the customer, we use some algorithms such as Back-propagation Neural Network, Multiple logistic regression, C4.5 of decision tree, support vector classification by software WEKA. At the same time, some algorithms such as Price sensitive function, Oversampling, Sub-sampling, Bagging, Boosting are used to dispose the unbalanced data. The experimental results are shown in Table 3 and Figure 3.

Table 3 The results of classification performance

Classification	Method	Accuracy	Recall	Precision	F-value	AUC
Neural Network	Multilayer Perceptron	0.733	0.537	0.733	0.620	0.503
Contrast	Logistic	0.601	0.588	0.601	0.594	0.501
	J48	0.601	0.588	0.601	0.594	0.501
	SMO	0.733	0.537	0.733	0.620	0.503
Unbalanced Data	Price sensitive function	0.801	0.888	0.801	0.842	0.742
	Oversampling	0.503	0.594	0.503	0.545	0.695
	Sub-sampling	0.542	0.544	0.542	0.543	0.712
	Bagging	0.882	0.882	0.882	0.882	0.749
	Boosting	0.803	0.894	0.803	0.846	0.795

Figure 3 The comparison of classification performance

From the perspective of the overall accuracy of classification, the accuracy of the Neural Network classification is above 70%, which is better than the other algorithms. After using some algorithms of unbalanced data processing, Price sensitive function and Bagging have a good effect.

Through the product recommendation model based on customer segmentation, the number of valuable rules can be seen in the Table 4.

After the valuable rules are applied in the test sets, some evaluation indicators are shown in Table 5.

Table 4 The number of valuable rules

Cluster number	Frequent 2-itemset	Frequent 3-itemset	Frequent 4-itemset	Frequent 5-itemset	Frequent 6-itemset	*Effective rules*	*Valuable rules*
Cluster0	5	2	/	/	/	22	16
Cluster1	3	/	/	/	/	6	4
Cluster2	5	2	/	/	/	22	16
Cluster3	26	42	36	16	3	1474	1313

Table 5 The evaluation of the product recommendation model

Cluster number	Value of customer	Recall	Precision	F-value
Cluster0	Class II	0.436026	0.232264	0.302022
Cluster1	Class I	0.494854	0.292955	0.349141
Cluster2	Class III	0.428662	0.225484	0.294669
Cluster3	Class IV	0.095891	0.047945	0.063927
Weighted average		0.430816	0.229485	0.298013

As can be seen from Table 5, first of all, under the condition of the same confidence and support the more rules means the more similar purchase characteristics these customers of the same class have. So it is more convenient for insurance companies to make product recommendations. Secondly, when recall value is high, the recommendation result contains the more hidden products. The product recommendation result of Class I customer is generally superior to other classes, and only the recommendation effect of Class IV is not ideal. Customers of Class IV who make biggest contribution to the insurance company are the most valuable customer resources, however, the recommendation result of this class is not ideal. To investigate its reason, on the one hand the quantity of Class I customer is minimum, which accounts for only about 1% of all the customers. On the other hand, Such customers' purchase characteristics may be to purchase a large amount of product portfolios.

To contrast, for general product recommendation model, it is need to do association analysis of all the customer purchase information directly. The methods and related parameters setting are as the same as the recommendation model based on Consumer Segmentation. Table 6 describes the contrast between two models in terms of data size.

Table 6 The contrast between two models in terms of data size

Group number	Total number of samples	Number of training sets	Number of test sets
General group	25317	20253	5064
Cluster0	18553	14843	3710
Cluster1	600	480	120
Cluster2	5798	4638	1160
Cluster3	366	293	73

The experimental results are in line with the expectations of our study, product recommendation based on customer segmentation can obtain more accurate results. Recall rate is

increased by 4.6%, accuracy is increased by 6.3% and F value is increased by 5.5%. This shows that the model based on customer segmentation make insurance product recommendation more targeted.
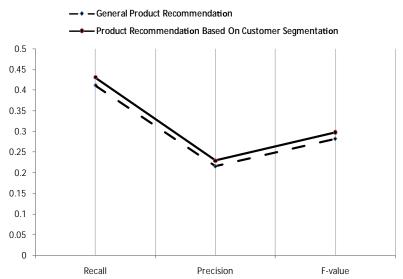

Figure 4 The comparison of performance between two recommendation model

Compare the different categories of customers, product recommendation for customers except Class IV get more accurate results. The reasons for this result are discussed above. Although the recommendation result of Class IV is not ideal, but the insurance company still need to spend a lot of power and resources to maintain the good relationship with the customers of Class IV, to avoid the loss of such class of customers.
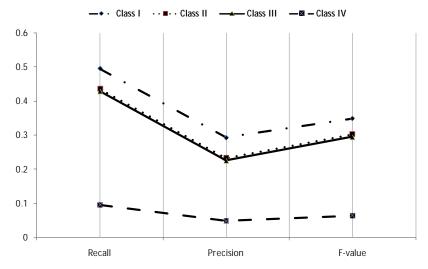

Figure 5 The comparison of performance between four classes of customers

5   Conclusion

This paper proposes a recommendation model for insurance products based on consumer segmentation. On the foundation of customer segmentation model constructed by the clustering

analysis and Neural Network, we use Apriori algorithm to build the product recommendation model. Through the experiments on real policyholder data by the insurance company, we mainly obtain the following conclusions. Firstly, the insurance premiums and claims can clearly reflect the value of policyholders when constructing the consumer segmentation model. Policyholders with high premium and less claims, who are the most valuable, only account for 1% of the total number of customers. More than 70% of the customers are regular customers with low premium. Secondly, the information of policyholders from insurance company can be divided into two types, personal insurance information and vehicle information. We use these two types of information to do the classification forecast in the experiments, and the results we obtain are good. After using some algorithms of unbalanced data processing, Price sensitive function and Bagging have a good effect. Finally, through the comparison in the experiments product recommendation based on customer segmentation can obtain more accurate results. This shows that the model based on customer segmentation not only makes the consumption of association rules analysis reduced, it also makes insurance product recommendation more targeted.

In addition, there are also some problems to do further discussion. Firstly, due to the limitation of data size, this paper chooses the premiums and claims on behalf of the customer's current value and potential value in process of constructing customer segmentation model. To a certain extent, it affects the accuracy of the classification model. So in future studies we should attempt to add more segmentation indicators to distinguish different categories of customers more effectively. Secondly, the method of products recommendation this paper propose only use the classic Apriori algorithm, and what we need to do is to improve it according to the actual situation of this study. So that it can lead the recommendation model has better recommendation results.

Finally, in the experiments of customer classification, we use the default parameters in the WEKA when apply Back-propagation Neural Network, Multiple logistic regression, C4.5 of decision tree, support vector classification. Future research will focus on looking for the standard procedures or methods to make the parameters of model achieve the overall optimization.

References

[1] Jarrar F Y, Neely A. Cross selling in the financial sector: Customer profitability is key. Journal of Targeting, Measurement and Analysis for Marketing, 2002(3): 282–296.

[2] Kamakura W A, Wedel M, Rosa F, et al. Cross selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. International Journal of Research in Marketing, 2003(20): 45–65.

[3] Kamakura W A, Ramaswami S N, Srivastava R K. Applying latent analysis in the evaluation of prospects for cross selling of financial services. International Journal of Research in Marketing, 1991(8): 329–349.

[4] Paas L, Kuijlen T. Acquisition pattern analyses for recognizing cross-sell opportunities in the financial services sector. Journal of Targeting, Measurement and Analysis for Marketing, 2001(3): 230–240.

[5] Harrison T, Ansell J. Customer retention in the insurance industry: Using survival analysis to predict cross selling opportunities. Journal of Financial Services Marketing, 2002(3): 229–239.

[6] Knott A, Hayes A, Neslin S A. Next product to buy models for cross selling applications. Journal of Interactive Marketing, 2002, 16(3): 59–75.

[7] Hill W, Stead L, Rosenstein M, et al. Recommending and evaluating choices in a virtual community of use. Proceedings of SIGCHI 1995. Denver, Colorado: ACM New York, 1995: 194–201.

[8] Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992, 35(12): 61–70.

[9] Reaniek P, Iacovou N, Suchak M, et al. Grouplens: An open architecture for collaborative filtering of netnews. Proceedings of the ACM CSCW'94 Conference on Computer-Supported Cooperative Work, 1994: 175–186.

[10] Reanick P, Varian H R. Recommender systems. Communications of the ACM, 1997, 40(3): 56–58.

[11] Zhao L, Hu N J, Zhang S Z. Algorithm design for personalization recommendation systems. Journal of Computer Research and Development, 2002, 39(8): 986–991.

[12] Liua D R, Shiha Y Y. Integrating AHP and data mining for product reconnnendation based on customer lifetime value. Information & Management, 2005, 42: 387–400.

[13] Fletcher K, Wright G. The strategic context for information systems use: An empirical study of the financial services industry. International Journal of Information Management, 1996, 16(2): 119–131.

[14] Li X L, Liu L. Modeling of recommendation system for financial products based on collaborative filtering. Computer and Engineering Lnstitute, 2007, 35(9): 6–8.

[15] Wang W P, Wu L. Application of collaborative filtering in CRM-Based cross-selling. Chinese Journal of Management, 2007, 4(4): 436–441.

[16] Wang X J, Hu M. Data mining technology application about cross-selling in life insurance market. Insurance Studies, 2009: 90–99.

[17] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. Nature, 1986, 323(6088): 533–536.

[18] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD Conference on Management of Data, 1993: 207–216.

[19] Park J S, Chen M S, Yu P S. An effective hash based algorithm for mining association rules. Proceeding of the 1995 ACM SIGMOD International Conference on Management of Data, 1995: 175–186.

[20] Agrawal R, Srikant R. Fast algorithms for mining association rules. IBM Almaden Research Center, San Jose, CA, 1994.

[21] Brin S. Dyanmic itemset counting and implication rules for market basket analysis. Proceeding of the 1995 ACM SIGMOD International Conference on Management of Data, 1997: 255–264.

[22] Sarasere A, Omiecinsky E, Navathe S. An efficient algorithm for mining association rules in large databases. Proceedings of the 21st International Conference on Very Large Database, 1995: 432–443.