2012 International Conference on Information and Network Technology (ICINT 2012)
IPCSIT vol. 37 (2012) © (2012) IACSIT Press, Singapore

Pareto Classification of Data Mining for Customer Relationship

Velu C. M.[1] and Kashwan K. R.[2 +]

[1] Anna University of Technology, Coimbatore, India.

[2] Department of Electronics and Communication Engineering – PG, Sona College of Technology (Autonomous), TPT Road, Salem-636005, INDIA. (Affiliated to Anna University of Technology, Coimbatore)

Abstract. This paper proposes a research investigation of a data mining based classification of customers into different clusters of specific interests. The specific interest groups may consist of high-profit making customers, high income customers, long term customers and modern life style customers. Classification is achieved by considering parametric variations of spending habit, life style based on type and choice of items they purchased, opting for brands and adopting for latest technology based items. We have buildup a software based intelligent model by using Artificial Neural Network (ANN), Genetic Algorithm (GA) and Fuzzy Logic to discover the current spending pattern of the customers. Input data to the model consists of only information of the transactions they have made at super market. This has also resulted in classification of the customers into class *A* of high income group, class *B* of medium income group and class *C* of low income group. We simultaneously performed real time survey by collecting actual status of the customers at super market counters by posing direct questions. We then compared the results of our software model with that of hard data collected by means of surveys. Results are quite encouraging as 98 % above classification is performed correctly. We employed Linear Regression and Rapid Association Rule Mining Algorithm (RARM) for preprocessing. Both techniques are fast and easy to apply on initial data processing for fewer attributes.

Keywords: cluster, data mining, intelligent model, life time value, customer relationship.

1. Introduction

Data mining techniques are quite often employed for market segmentations and subsequently these produce better forecasting results of sales statistics. Data mining is also comparatively new concept and a powerful software based tool adopted for collecting transact information of customers for analyzing Customer Relationship Management (CRM) in various industries such as airline [1]. Combining management skills with data mining can improve CRM analysis and prediction models greatly. This is the fundamental principle for using data mining for business support information system in many ways to enhance statistical models. Data mining theory has been quite successful for customer segmentation application models in retail industries to improve business opportunities [2]. Similarly banking sector is quite optimistic of data mining models in collection of customer preferences and segmentations information to have strategic planning for long term relationship and retention of valued customers. It is very critical for banks to continuously add on new customers while simultaneously keeping good confidence level among existing customers. The innovative models developed with the help of information technology and data mining techniques can be quite naturally useful for banks to forecast the behavior of customers [3].

Data mining techniques have been traditionally used for knowledge classification and search engines. These techniques have been very useful and handy in searching required information from huge information data collections. Customized and carefully selected data mining techniques reinforces information

[+] Corresponding author. Tel.: + 91 427 4099877; fax: +91 427 4099888.
E-mail address: drkrkashwan@sonatch.ac.in

management process to serve customers with better and efficient services. The models developed can readily find out customer preferences which may be very useful for various strategically important decisions. Traditionally and in good olden days, customer preference could only be known a little by surveys and posing direct questions to the customers. This was quite tedious and time consuming process. Data mining techniques have provided opportunity to determine customer preferences by only analyzing customer transaction and purchase information [4]. This can result in very useful and quick method. Yet another area of customer relationship management is customer lifetime value. The large organizations are quite often interested to know the future returns from their customers for long term planning and forecasting based on lifetime value [5]. This is very complex and has many challenges to achieve meaningful outcome. Many individual prediction models have shown little success. Data mining and software based techniques have proven quite a bit more accurate in predicting customer lifetime value [5]. Many recent research works have indicated that data mining using concept of 3-dimensional data visualization can be very helpful in many quick decision making and better understanding of complex business issues. The 3-dimensional analysis can be applied to slice, rotate and zoom in the data projections to obtain various minute details by visual perceptions. Indirectly, business management for marketing can be easily represented by customer knowledge management [6]. Song et al. had proposed a new method to determine the dynamic and ever changing customer behavior with the help of only customer sales data [7]. They have demonstrated to monitor the changes among customers and then formulated corresponding rules for future predictions.

Key factor for a business to be successful is to address customers in the clearest terms and the CRM is considered customer friendly. It takes customer life cycle into account as a very important and critical business parameter. CRM improves business environment by retaining customers for longer and enhancing their confidence [8]. Most of the successful business enterprises have mastered customer preferences by way of keeping their up to date knowledge and changing socioeconomic conditions [9]. All these efforts are made to promote business for particular group of customers. It appears to be challenging work which definitely requires modern information technology based tools and techniques for more accurate results. The most recent trends are indicating that hybrid systems of data mining are more powerful and intelligent for customer relationship management. The hybrid models may have combinations of fuzzy logic-genetic algorithms, neural-fuzzy systems, neural-genetic systems and fuzzy-genetic fields. Since the business is a very complex and dynamic entity, it needs quite intelligent systems to predict more accurately. In particular, CRM may be represented as the joint strategy of sales, marketing and service which have very stringent requirement of understanding the customers' psychology in contrasting situations [10]. Banks are increasingly using credit cards and other electronics media to collect the customer preference information in order to understand customer's changing psychology [11].

2. Methodology

2.1. Customer value matrix

The customer value matrix is widely used to analyze customer values based on customer data base. It was first time proposed by Marcus [12]. The customer value matrix contains different type of information related to customers, such as income range, buying preferences, gender information, age group, whether customer has interest in electronic media, whether a customer reads a particular type of magazine, why he prefers products from this market, what is most important point of this product, what is next alternative and so on. Firstly these parameters are divided into suitable range bound values and then, customers are classified in one of the ranges based on software tool prediction. The software tool makes decision based on certain transaction information available to it after customer has visited super market for purchase. Mostly the matrix is rearranged by converting certain information by using portfolio classifications. For an affirmative parameter a *1* is entered while for a negative parameter a *0* is entered. Subsequently data mining and neural network based intelligent techniques are applied to extract certain patterns of information from customer value matrix. The process is automatic and customer data base is updated by way of classification, association and cluster analysis as soon as he or she makes a purchase.

2.2. Rapid association rule mining

Rapid Association Rule Mining (RARM) is a tree structure based data mining process which can be used for finding patterns, associations, correlations among customer data base [13]. It is a very fast technique compared to many other similar techniques. It can generate multiple attributes by just one time scanning of customer transaction information from billing system. Our system is programmed using neural network and fuzzy logic so that the algorithm is automatically executed in order to update customer data base for further analysis. In this automatic software based model, RARM works first to find out all rules that apply. Subsequently it correlates the presence of absence of a set of attributes with that of another set of attributes. Thus, clusters obtained from this analysis are used as basis for segmentation of target customers. In the view of ever changing customer's behavior, an analysis is made on dynamic nature of attribute patterns. We have applied multiple-level association rules to formulate RARM execution sequences. Multiple-level association rules are based on following principles.

- Attributes have inherent property of hierarchy and thus suitable tree structure.
- Attributes with low values have low contribution and thus can be mostly ignored.
- Rules for finding of attributes are to be determined very carefully and appropriately at different levels.
- Dimensions of transaction data base can be reduced to represent only dominating attributes at different levels, so that algorithm execution time can improve.
- There is a possibility to incorporate feedback between multi-level mining to improve the accuracy.

2.3. Data transformation and enhancement

Firstly we have applied data cleaning by removing redundant data. Similarly it is also used to reduce data by eliminating very low contributing data sets as indicated in RARM principles. Customer value matrix is re-structured by converting *YES* (presence of attribute) as *1* and *NO* (absence of attribute) as *0*. We have also transformed the purchase days into months and then finally months into annual periods. Data enhancement is performed to infer certain conclusions based on indirect information available in transaction records. Through enhancement priority matrix, attributes can be classified as high priority and low priority to maintain appropriate customer relationship [14]. This can be illustrated as supposing that a customer has high income, consider electronic media as an important communication mean, show good confidence in newer technology and has interest in gaining knowledge. If this information is explicitly available from his or her past transaction records, then there is a very high possibility that he or she is a high valued potential future customer for new computer model being launched shortly by a firm. The determined attribute can be subsequently applied for all other customer's transaction records which ultimately results into a well defined customer segment size. The knowledge of such information can be extremely important for forecasting.

2.4. Linear regression model

Linear regression is a statistical tool used to model the relationship between independent and dependent variable by using linear mathematical function. Independent variables are also called scalable variables whereas dependent variables are called explanatory variables. If there is only one explanatory variable, it is simple regression else it is multiple regressions. Almost all real problems are solved by using multiple regressions. Linear regression is quite suitable for prediction and forecasting model where observed data are used to find model parameters. Subsequently prediction can be made even without observing data. Scalable variable is set arbitrarily and dependent variable is predicted b y using regression model developed. Normally regression model is expressed in the simplest form as $F_i(x) = x_iB + e_i$, where $i = 1, 2, ...n$, $F_i(x)$ is called regressand or measured variable or response variable, x_i regressors or explanatory variables or predictor variables, B is called parameter vector or regression coefficients, e_i is called error term or is unobserved random variable that can be treated as noise in model. Our model focuses on determination of B.

$$F(x) = \begin{pmatrix} F_1(x) \\ F_2(x) \\ . \\ . \\ F_n(x) \end{pmatrix}, \quad x = \begin{pmatrix} x_{11} \; . \; . \; . \; x_{1p} \\ x_{21} \; . \; . \; . \; x_{2p} \\ . \\ . \\ x_{n1} \; . \; . \; . \; x_{np} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \\ . \\ . \\ B_n \end{pmatrix}, \; e = \begin{pmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{pmatrix} \tag{1}$$

All variables in the regression model can be mathematically represented as shown in equation (1) below.

The model first identifies customer classes using clustering algorithm and then generate association rules using neural networks. Similarity and dissimilarity are measured among the rules. Depending upon chosen threshold, the rule matching is performed which results in determination of the degree of change. The degree of change knowledge can be used to determine trends of change in customer behavior.

3. Customer Segmentation Model

The data has been collected online directly from transaction billing counter through computing systems. We also collected small amount of data by direct questionnaire to the customer after they have made purchase. Thus the hard data collected are used for testing of the results produced by software based proposed model. Main objective is to classify the customers into one of the three classes *A, B* and *C* of high, medium and low income groups respectively as shown in figure 1. Based on the transaction information for last one year, model has done classification and also direct questionnaire result is also available for verifications. Our research study is mainly focused on the type and amount of purchase made by customers over the period of long time. Next we carried out data transformation as explained in previous section. Using software models we performed data enrichment. For this, classification technique based on supervised neural network is applied to find possible clusters. We implemented a supervised model for classification by using instances. The model consists of input layer, hidden layer and output layer of neurons as shown in figure 1. The back propagation algorithm is then used to validate the classification results obtained by supervised classification. Another software tools, genetic algorithm and fuzzy logic are used as a combination or hybrid approach. The idea is based on the multiple validation of the model and thus to improve further accuracy of customer segmentation. The steps of back propagation and genetic algorithms are listed below.

Back Propagation Algorithm	*Genetic Algorithm*
Step 1: Initialize weight matrix	Step 1: Generate initially random population
Step 2: Get inputs and generate output	Step 2: Determine fitness of population candidates
Step 3: Compare with actual output, estimate error	Step 3: Create new population by crossover and mutation
Step 4: Stop if error ≤ threshold, else go to Step 5	Step 4: Place new offspring in new population
Step 5: Propagate error back through network	Step 5: If population converged, stop, else next step
Step 6: Update the weight matrix of all the layers	Step 6: If end condition, Stop, return best solution
Step 7: Loop - Repeat Steps 2 through 6	Step 7: Loop – go to Step 2

The training of neural networks consists of updating weights automatically so as to minimize error in the desired output with reference to actual outputs. Back propagation algorithm is widely accepted method for finding errors. Genetic algorithm finds the customers with high potential and high income groups on the basis of survival of fittest rule. Figure 2 shows steps and sequences for achieving validation and testing of the developed model. Similarly, fuzzy logic is applied by generating rule based membership functions and degree of correctness.

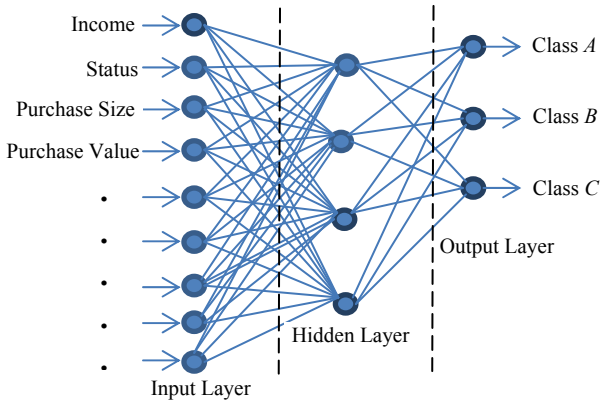

Fig. 1: Neural Network Model for customer Segmentation

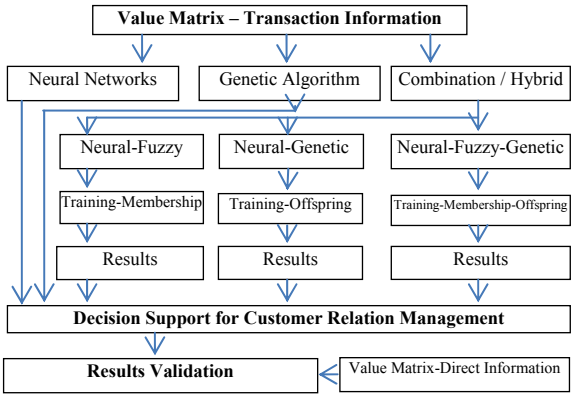

Fig. 2: Validation of Customer Segmentation Model

4. Result Analysis and Conclusion

The experimental work was carried out for a long time over a super market customer's daily purchasing activities. We posed direct questions and our software based model generated instant results by simply acquiring transaction informs online. A total of 3090 customer's transactions were analyzed and used as inputs to segmentation model and also direct questionnaire results were compiled manually for validation and testing purpose. Table 1 shows the results by various tools and finally results of combined techniques. As proven by the results available to us, the classification of customers was quite accurate.

Table. 1: Results of Pareto Segmentation classification Model

Classification Technique	Class-A Customer	Class B Customer	Class C Customer	Total Customer
Direct questionnaire (Reference Value)	716	978	1396	3090
Neural Network- Back Propagation	742	897	1451	
Genetic Algorithm	711	899	1480	
Fuzzy Logic	801	817	1472	
Neural-Genetic	784	906	1400	
Neural-Fuzzy	722	985	1383	
Neural-Fuzzy-Genetic	728	987	1375	
Validation Result with Combined algorithm results	98.32 %	99.08 %	98.49 %	

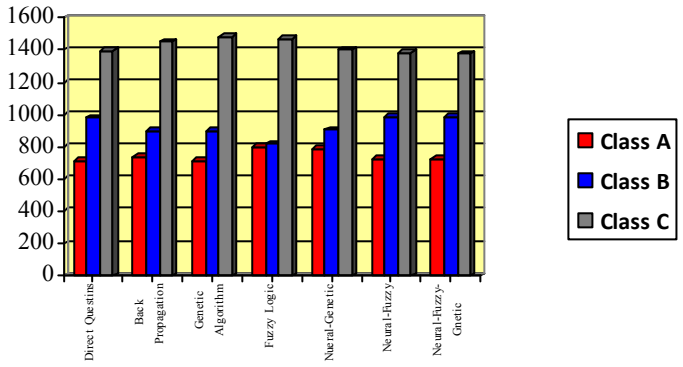

Fig. 3: Results of Pareto Classification of Segmentation Model

Direct questionnaire classification is considered as standard reference value. The classification is in comparison with standard reference considering total number of customers. We find that results of correct classification are above 98 % in all types of simulation techniques. The other analysis of resulted have led to the conclusion that the B-Class and C-Class customers collectively accounted for 76.82 % of the total customers may have contributed less proportionate revenue to the market. The A-Class customers have generated high revenue contribution, thus are considered more important and to be retained or probably increase in their number will be good news. But same time class B and class C are also very critical as their number is quite large. The Pareto classification as illustrated by figure 3 has just produced priority based classification. The super market could easily interpret the results and adopted to use professional knowledge to offer value added service to its customers and improve the customer relationship.

5. References

[1] Huirong Zhang Yun Chen, "An Analysis of the Applications of Data Mining in Airline Company CRM", Fuzzy Systems and Knowledge Discovery, 2009, Sixth IEEE International Conference on, Vol 7, Page(s) 290 – 293.

[2] Huaping Gong Qiong Xia, "Study on Application of Customer Segmentation Based on Data Mining Technology", Future Computer and Communication, FCC, 2009, IEEE International Conference on, Pages 167 – 170.

[3] Wu Dong Sheng, “Application Study on Banks’s CRM Based on Data Mining Technology”, Electrical Information and Control Engineering, ICEICE, 2011, IEEE International Conference on, Pages 5727 – 5731.

[4] Young Sung Cho Keun Ho Ryu “Implementation of Personalized Recommendation System Using Demograpic Data and RFM Method in e-commerce”, Management of Innovation and Technology, 2008, ICMIT, 2008, 4TH IEEE International Conference on, Pages 475 – 479.

[5] Lim Chia Yean and Khoo, V.K.T., “Customer relationship management: Computer-assisted Tools for Customer Lifetime Value Prediction”, Information Technology, ITSim, 2010, IEEE International Symp. Pages 1180 – 1185.

[6] Shaw M. J., Subramaniam C., Tan G. W. and Welge M. E., “Knowledge management and Data Mining for marketing”, Decision Support Systems, 2001, Pages 127 - 137.

[7] Song H.S., Kim J.K. and Kim S.H., “Mining the Change of Customer Behavior in an Internet Shopping Mall” Expert Systems with Applications, 2001, 21(3), Pages 157 - 170.

[8] Jill Dyche, “The CRM Handbook: A Business Guide to CRM”, Addison-Wesley Professional, 2002, First Edition.

[9] A. Berson, K. Thearling and S. Smith, “Building DM Applications for CRM,” McGraw-Hill, 2000.

[10] Alex Sheshunoff, “Winning CRM Strategies”, ABA Banking Journal, 1999, Pages 54 - 66.

[11] Alireza Fazlzadeh, Mostafa Moshiri Tabrizi and Kazem Mahboobi, “Customer relationship management in small-medium enterprises: The case of science and technology parks of Iran”, 2011, African Journal of Business Management, Vol. 5(15), pp. 6159-6167.

[12] Marcus C., “A Practical yet Meaningful Approach to Customer Segmentation”, 1998, Journal of Consumer Marketing, Vol 15 (5), Pages 494-504.

[13] Das A., Ng W. K. and Woon Y. K., “Rapid Association Rule Mining”, Information and Knowledge Management, 2001, 10th International Conference on, ACM Press, Pages 474 – 481.

[14] Chun-Cho Chen, Ching-Sung Wu and Rebecca Chung-Fern Wu, “e-Service Enhancement Priority Matrix: The case of an IC Foundary”, 2006, Elsevier Journal of Information & Management, Vol:43 (5), Pages 572 – 586.