

Segmentation Study on Enterprise Customers Based on Data Mining Technology

Xin-an Lai
School of Business
Wenzhou University
Wenzhou, China
laixinan21@163.com

Abstract—In the current context of fierce market competition, Customer Relationship Management (CRM) has become the main means for enterprises to gain competitive advantage and an important option for enterprises to make strategic planning. As a core concept of CRM, customer segmentation can effectively lower the marketing costs of a company and help it achieve more visible and profitable market penetration. This paper begins by introducing the definition of data mining and data mining methods used in CRM; it then proceeds to analyze the advantage of customer segmentation based on data mining and establish the models for customer segmentation; in the end this paper probes into the application of clustering analysis technology of data mining in customer segmentation.

Keywords—customer relationship management; data mining; customer segmentation

I. INTRODUCTION

Customers are regarded as important strategic resources of an enterprise, and whether an enterprise can gain, retain and develop its customers has become the most critical factor of an enterprise's success, hence the customer-centric management concept—Customer Relationship Management (CRM) is generated. CRM stresses the importance of customer relationship, it mainly covers the following functions - managing the coming into being of customer-and-enterprise relations and promoting their rational development [1], coordinating and unifying the interaction with customers and attaining the purpose of retaining valued customers, identifying potential customers and winning customers loyalty by centering customers with enterprise organization, flow of work, technical support and customer service, and achieving the ultimate goal of harvesting the long-term values of customers [2].

The traditional ways of customer segmentation are mainly categorization methods based on experiences or simple partitioning [3]. Those methods are valuable to an enterprise's CRM in a certain way, however it cannot satisfy the requirement of some more complex analysis. Enterprises have accumulated a growing amount of customer data with the extensive use of management information system and the deepening development of electric commerce; the traditional customer segmentation methods fall short of need faced with such a great amount of customer data. The introduction of data mining technology into the CRM of helps an enterprise to analyze and handle a huge amount of complicated customer data, sort out the demand patterns of potential customers and their consuming behavior modes, and in turn

improve communication between enterprises and their customers, help enterprises to satisfy the needs of their customers realize the “customer-centric” business model. This paper primarily sets to make preliminary probe into the application of clustering analysis technology of data mining into the practice of customer segmentation.

II. DATA MINING METHODS FREQUENTLY USED IN CRM

A. Definition of Data Mining

Data mining is the process of sorting through large amounts of data and picking out relevant information and knowledge by using a series of modern techniques. Those information or knowledge is implicit, unknown before but potentially useful. The representation of those findings takes the form of concepts, rules, patterns and models etc [4]. Data mining is an emerging field of great value in the research of database; it integrates theories and techniques from the realms of database, artificial intelligence and machine learning, statistics etc, it performs data testing and modeling by selecting relevant data and thus digging out fundamental information and knowledge.

B. Commonly Used Methods of Data Mining

In CRM, some data analysis methods are utilized to integrate large amounts of superficial and unordered data so as to reveal the underlying relevance and regularity and guide the decision-making process. The following are some commonly used methods for data mining [5-7]:

1) Neural network method

This system is constituted by a series processing units similar to the nerve cells of human brain which are called nodes. Those nodes are interconnected via network; a certain data processing model is triggered at each data input. Structurally, a neural network can be divided into the input layer, the intermediate layer and the output layer. The intermediate layer is dotted with a great amount of nodes, which perform a large part of the networking. While the output layer outputs the execution results of data analysis.

2) Decision tree method

Decision tree is an inductive learning algorithm based on examples. It involves two phases while conducting data categorization: the formation and the pruning of the tree. At first, a test function is generated from training data, and the branches of the tree are established according to the specific needs. Each branch develops further into lower level nodes and sub-branches and a new decision tree is thus formed.

Pruning is applied to the established decision tree to transform the decision tree into rules, which are further utilized to classify new incidents.

3) Association rules

Association rules mining seeks inter-relations hidden between data entries, its general objects are transaction databases. The discoveries made through large amounts of association rules help enterprises to implement marketing planning, advertising planning, classified design, overlapping shopping and strategic analysis etc. The most well-known association rules based data mining algorithm is the Apriori Algorithm brought forward by Rakesh Agrawal et al. The majority of well-developed algorithms are all based on Apriori Algorithm.

4) Rough sets

The rough sets theory is both a new math tool handling ambiguous and uncertain issues and a theoretical method researching the expressing, learning and induction of incomplete and uncertain knowledge and data. This method attribute unidentifiable datum into the borderline area. It assumes the form of a definite mathematical formula, thus the number of ambiguous entries is calculable. Its main feature lies in its ability to handle ambiguous phenomenon using the rough sets method, namely its ability to precede data categorization based on the imprecise results from observation and measurement.

5) Genetic algorithm

Genetic algorithm is a combinatorial optimization method based on the evolutionary process of living creatures. It starts from an initial population which is specified or randomly chosen, and then constant iterative computation is performed according to a certain rule to retain good varieties and weed out substandard items based on the fitness value of each individual entry. In this way, the search process is guided to the optimum solution. The main characteristic is that it operates directly on the object structure without the limits on differentiation and the continuity of a function, providing intrinsic imperceptibility and preferable global optimization ability.

III. THE ADVANTAGES AND PROCESS OF CUSTOMER SEGMENTATION BASED ON DATA MINING

A. The Advantages of Customer Segmentation Based on Data Mining

Customer segmentation is the foundation of effective marketing and service of an enterprise. It categorizes its

large amounts of customers, with customers in the same category possessing similar attributes and customers in different categories possessing different attributes. Compared with traditional methods of customer segmentation, customer segmentation methods based on data mining are more advantageous in the following regards:

- The results of segmentation based on data mining are decided by the objectivity of the data, the subjectivity of the people who are processing them are avoided, resulting a more objective representation of the differences among different populations.
- It represents the categorization features among different customer categories more comprehensively, which facilitates marketing staff know their customers more thoroughly and in turn make more targeted and individualized marketing plans.
- The changes of customer behaviors can be tracked more easily by collocating clustering analysis models and updating the categorization of customers regularly.

In the implementation of CRM, the most valuable link to the enterprise itself is the customer segmentation part. It provides more comprehensive customer knowledge otherwise impossible from other segmentation methods, which give reference to the enterprise on the better implementation of precise marketing [8]. For instance, an enterprise divides its customers into three grades (high, middle, low) according to their contribution to the enterprise; however, information derived from this grouping cannot guide the enterprise to conduct precise marketing. It is the segmentation of customer behaviors that help find the demands and preferences of customers. By analyzing the information obtained from the segmentation of customer behaviors, an enterprise can provide its customers with products and services truly needed by them.

B. Process of Customer Segmentation Based on Data Mining

According to the characteristics of data mining and the requirement of enterprise customer segmentation, model of the process of customer segmentation based on data mining is established as shown in Fig. 1.

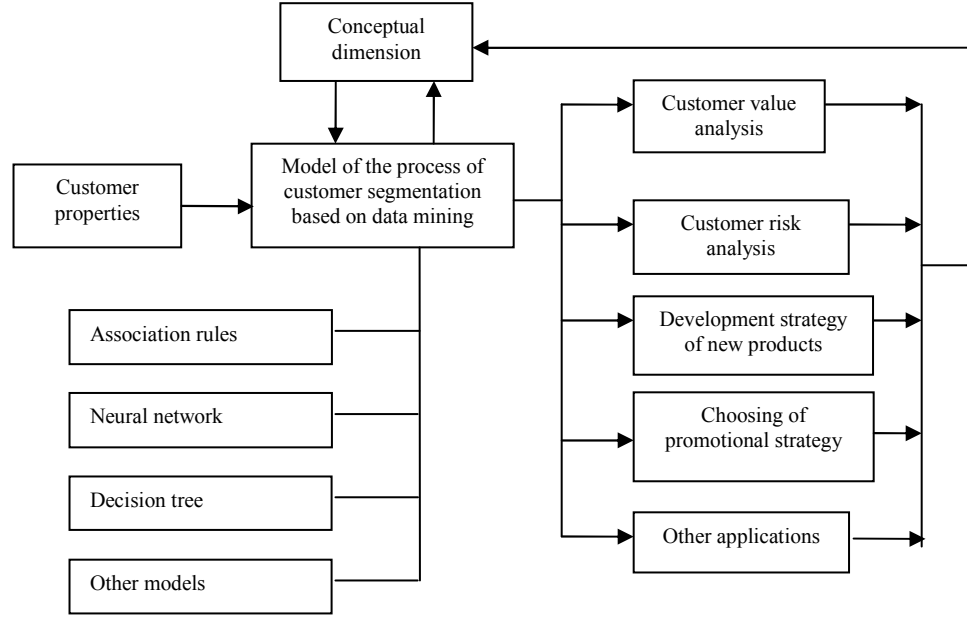


Figure 1. The model of the process of customer segmentation based on data mining

According to the above model, the customer segmentation process mainly includes the following three steps:

1) *Categorization model learning*

Use relevant algorithms to find categorization models or categorization functions which describe and differentiate data classes or concepts categories. The generated categorization models can take various forms such as association rules, neural networks, decision trees, etc.

2) *Application of categorization rules*

Use customer categorization models and categorization functions to categorize customers to be grouped. After the categorization models training, categorization models like rules or networks will be obtained, which can be used to categorize customers with unknown categorization numbers.

3) *Functional analysis*

Functional analysis includes the analysis of customer value, customer risk and promotional strategies. With the development of management practices, new functional demands will be created during the CRM process, those functional demands will be added to the concept dimension data.

IV. APPLICATION OF DATA MINING TECHNIQUES IN ENTERPRISE CUSTOMER SEGMENTATION

Customer segmentation using data mining techniques can be materialized through various means. The most frequently used customer segmentation technique in data mining is clustering analysis [9]. Clustering analysis can be used to categorize customers based on the differentiating features of their address, ages, sexes, incomes, occupations, education levels, etc. Meanwhile, clustering analysis can generate the different levels of importance associated with

different variables in the classifying process; those data can be used to assist decision-makers.

K-center algorithm is normally used when apply clustering analysis to customers. The fields involved include basic customer information (such as their ages, incomes, occupations and educational backgrounds etc.) and customers' related business information (such as customers' earning, transaction time, customer relations retention costs, etc.). While adopting K-center algorithm, enterprises can choose from primary-attribute clustering analysis and all-attribute clustering analysis according to their industrial characteristics and business environments [10].

A. Primary-attribute Clustering Analysis

Simply put, primary-attribute clustering analysis is when an enterprise chooses principle attributes from its calibration variables as the clustering analysis variable to make analysis based on its needs. Normally, the measurement units chosen by interval-valued variables have direct bearing on the outcome of clustering analysis: the smaller the measurement unit, the wider the possible range of values for variables, and thus the greater the effect on clustering analysis. Data should be standardized in order to avoid the dependency on measurement units, with its concrete steps as follows:

1) *Mean Absolute Error- s_f*

$$m_f = (x_{1f} + x_{2f} + \dots + x_{nf}) / n \quad (1)$$

$$s_f = (|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|) / n \quad (2)$$

In (1) and (2), m_f is the mean of the n measurement units of $x_{1f}, x_{2f}, \dots, x_{nf}$.

2) *Standardized Metric Values*

$$x_{if} = (x_{if} - m_f) / s_f \quad (3)$$

When calculate the mean absolute error, the purpose of using the metric value of $|x_{if} - m_f|$ instead of its square is to reduce the impact of isolated points. After the standardization of data, the dissimilarities of all objects are mainly measured by the space in between each object. The calculation of the space in between each object can resort to one of the following three measurements: Euclidean distance, Manhattan distance and Minkowski distance. Presently, the traditional data clustering is mainly measured by Euclidean distance. However, when using the Euclidean distance, the importance of each customer attribute should be considered; a solution to this is to give weight $w_1, w_2 \dots w_n$ to each attribute value respectively. The computational equation is as follows:

$$d(i, j) = (w_1 |x_{1f} - m_f| + w_2 |x_{2f} - m_f| + \dots + w_n |x_{nf} - m_f|)^{1/2} \quad (4)$$

B. All-attribute Clustering Analysis

The adoption of all-attribute clustering analysis is different than that of the primary-attribute clustering analysis in that the difference in the type of each attribute necessitates the combination of different types of variables into a single dissimilarity matrix and the conversion of all variables to a common branch and a common range. For instance, for a data set of 10 different types of variables, the dissimilarity between i and j is as follows:

$$d(i, j) = \frac{\sum_{f=1}^8 \delta_{ij}^{(f)} d_{ij}^f}{\sum_{f=1}^8 \delta_{ij}^{(f)}} \quad (5)$$

In (5), if x_{if} or x_{jf} is missing, or $x_{if} = x_{jf} = 0$, and the variable f is an asymmetric binary variable, then the designator $\delta_{ij}^{(f)} = 1$. The computing mode of the dissimilarity between variable f and i and j is associated with their concrete types. If f is a binary variable or a nominal variable: $d_{ij}^f = 0$ when $x_{if} = x_{jf}$, otherwise $d_{ij}^f = 1$. If f is an interval-valued variable, the following equation can be applied:

$$d_{ij}^f = \frac{|x_{if} - x_{jf}|}{\max h x_{hf} - \min h x_{hf}} \quad (6)$$

In (6), h can be any non vacant object of variable f . If f is an ordinal number or a ratio-scaled variable, then $z_{if} = (r_{if} - 1) / (m_f - 1)$, z_{jf} is treated as an interval-valued variable.

V. CONCLUSION

This paper mainly discusses the application of the clustering analysis techniques of data mining in the segmentation of enterprise customers. When an enterprise uses clustering analysis techniques to categorize its customers, comprehensive treatment should be done to the outcome of the mentioned two types of clustering and the customers should be marked accordingly. Then the decision tree method should be adopted to establish the customer segmentation model, thereby, it helps enterprises identify their valuable customers.

REFERENCES

- [1] Michael K Rich, "The Direction of Marketing Relationships", Journal of Business & Industrial, Vol.15, Mar.2000, pp.170-179.
- [2] Jia-yin Qi and Huai-zu Li, "Analysis on the System Framework of Customer Relationship Management", Industrial Engineering, Vol. 5, Jan.2002, pp. 42-45.
- [3] Guo-zheng Zhang and Jie Gao, "The Customer Segmentation Study Based on Data Mining of CRM", Journal of Marketing, May.2005, pp.24-25.
- [4] Qiao-yun Yuan and Gang Cheng, "Overview of the Data Mining Research in Recent Years in China", Intelligence Journal, Feb.2005, pp.252-254.
- [5] Xiao-hu Zhu, Zhi-wei Ni and ChaoWang, "Exploring the Application of Data Mining Techniques in CRM", Value Engineering, Dec.2007, pp.114-116.
- [6] Long Xue, "Knowledge Discovery and Data Mining and Their Application in CRM", Science & Management, Apr.2007, pp.39-41.
- [7] Kun-jiang Lin, "Research of the Application Data Mining Techniques in CRM", Wuhan: Wuhan University of Technology, 2007, pp.27-29.
- [8] Liu-wei Xing, "The Research of the Application of K-Means Algorithm in Customer Segmentation", Chendu: Southwestern University of Finance and Economics, 2007, pp.9-10.
- [9] Liang Tian, "The Research of CRM Based on Value Theory and Data Mining", Guangzhou: Jinan University, 2004, pp.29-31.
- [10] Wen-chang Liu, "The Application of Apriori Algorithm in CRM", Changchun University Journal, Vol.15, Apr.2005, pp. 25-27.