

Customer Value Analysis with Fuzzy Data Mining

Jay B. Simha¹ and S.S. Iyengar²

¹ Siemens Information Systems Limited, Bangalore, India
Email: jaybharatheesh.simha AT siemens.com

² Department of Computer Science, Louisiana state university
Baton Rouge, LS, USA
Email: iyengar AT bit.csc.lsu.edu

Abstract

Customer Relationship Management (CRM) aims at optimizing the customer value in an organization by data analysis and communication. One important task of CRM is customer value analysis, which is used to identify the value of customers in different segments. In this paper an attempt has been made to develop a framework for value analysis using fuzzy data mining. The proposed method uses fuzzy segmentation approach to accommodate the domain knowledge and segment the customer base into homogeneous segments. Once the segments are identified, the important customers will be identified using outlier analysis. It is assumed that, the core remaining after the elimination of outliers is the real segment, which may have a stable behavior. Once the core members of a segment are identified, their value over a period is analyzed thru pre-defined measures. Preliminary investigation on a real world data set has shown positive results. The research is under progress to fine-tune the framework. The authors hope that the proposed framework will be useful for CRM, in customer retention and promotion related activities.

1. INTRODUCTION

CRM is a technology, which depends heavily on database and analytics including cognitive capabilities of humans to identify interesting and profitable patterns from data. The main idea behind CRM is to target profitable customers to increase the net revenue to the organization. Organizationally, CRM is a strategic focus on behavior of customers and communication with them. Technologically, CRM is based on analytics to identify the behavior and preferences using the historic data.

Mobile telecom market in India is very hot due to globalization and increased competition. It has been observed in this industry as in other customer intensive ones, that the retention of good customers is more effective in the long run than acquiring new customers.

Telecom marketers, in order to succeed, must first, identify market segments containing customers or prospects with high profit potential and, second, build and execute campaigns that favorably impact the behavior of these individuals.

The first task, identifying market segments, requires significant data about prospective customers and their buying behaviors. In theory, the more data the better will be the insight. In practice, however, massive data stores often impede marketers, who struggle to sift through the large data repositories to find the nuggets of valuable information. This is where the role of database technology and data mining will play an important role.

One of the important tasks of CRM is to go beyond segmentation and build value model, which is to identify the value of the customers in the organization. In this paper an attempt has been made to develop a framework for customer value analysis using fuzzy data mining. The rest of the paper is organized as follows. In section 2 the problem of customer value analysis is formally defined. In section 3 the methodology adopted in the research has been discussed. Section 4 provides the description of experimental results and the conclusions are given in section 5.

2. PROBLEM DESCRIPTION

Data mining as popular today has more inclination towards predictive modeling. Data mining, applied to value analysis has somewhat different focus. Here, aspects such as, e.g., the *understandability*, gain in importance. In fact, the goal in value

attrition analysis is not necessarily to induce *global* models of the system under consideration (e.g. in the form of a functional relation between input and output variables) or to recover some underlying datagenerating process, but rather to discover *local* patterns of interest, e.g. very similar (hence typical) or very rare (hence atypical) events. Customer Value analysis is of a more explanatory nature, and patterns discovered in a data set, are usually of a *descriptive* rather than of a *predictive* nature. Customer value analysis also puts special emphasis on the analysis of very large datasets and, hence, on aspects of scalability and efficiency.

Let $D = \{d_{ij} \mid 1 < i < n, 1 < j < m\}$ be the database, where, d is the database tuple, n is the number of records in the database and m is the number of attributes considered in analysis.

Let $C = \{c_{kj} \mid 1 < k < c, 1 < j < m\}$ be the cluster or the segment set, where c is the number of clusters and m is the number of attributes considered for analysis. Let $V = \{(v_i, x) \mid 1 < i < n, x \in R\}$, where n is the number of records in the database and x is the value attrition index defined by domain expert on the basis of attributes used in the analysis. Now the problem is to develop the relation $V = f(D, C)$. i.e. prepare the customer attrition matrix (V) using clusters (C) and database (D).

The customer value analysis is a typical datamining problem and the solution to the above problem is given in two stages. In the first stage, the data will be segmented using fuzzy data mining and in the second stage, the outliers in each segment will be identified and removed to understand the segment value. The details of the solution are discussed in the subsequent sections.

3. METHODOLOGY

This section describes the methodology followed in this research. The following three objectives were set for customer value analysis. They are

- Identification of patterns of customer behavior across segments.
- Determination of the customer value in each segment after removing outliers.
- Assessment of the effectiveness of fuzzy segmentation approach vs. other segmentation strategies.

The methodology adopted in the research is shown in fig 1. It consists of four stages – data modeling, data preprocessing, data mining and value analysis. Each of these phases are briefly discussed in the following paragraphs.

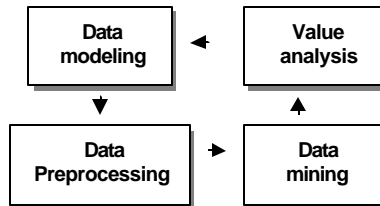


Fig. 1: Customer Value Analysis process

3.1 Data Modeling

Data modeling is a very important step in the analysis process. In this phase, the data to be used for analysis are identified along with the source and the required transformation. The variables chosen in the study are,

- Age on network – the association of the customer with the company is an indicator of recency.
- Number of recharges - This is an indicator of frequency and is more important in case of pre paid customers.
- Average revenue – is the monetary measure and is the major factor (along with age on network) in value analysis.
- Average usage - is the behavioral measure and is the major factor (along with age on network) in homogenizing the segments.
- Number of package changes – this is an indicator of frequency and customer satisfaction. More number of times the customer changes his package is a symptom of unfulfilled customer needs.

The above features are not exhaustive but are sufficient to analyze the value of the customers across the segments.

3.2 Data Preprocessing

Data preprocessing is the most time consuming phase in the entire process. It has been observed that 60% to 80% of the time in the process is spent on preparing the data for analysis [7]. This phase consists accessing data from heterogeneous sources, integrating them, analyzing for outliers, imputation of missing values, aggregation/summarization to the desired granularity, sampling and preparing the data in the standard spreadsheet format. In the present research the data was accessed from the data warehouse and was lightly summarized for the features used. The sampling was done based on Pareto analysis [8] and the data in standard format was developed.

3.3 Data Mining

Data mining is the process of automatically discovering useful patterns from data [5]. Clustering is one of the major data mining functions often used in analysis. There are different techniques of clustering. Since fuzzy c-means clustering [2] is known to give better quality clusters it has been chosen as the clustering algorithm. The number of useful and manageable clusters depends on the business situation and ability to meaningfully manage the cluster-categories. In most cases three to seven groups are enough to capture useful cluster properties. However, experiments can be done on the number of clusters to ensure correct number is chosen.

3.3.1 Fuzzy Clustering

Even though explanation of the particular fuzzy clustering algorithm used in this work [2] is not the intention of the paper, a brief summary of the considerations and major steps will aid in understanding the subsequent discussions.

The algorithm first posits a given number ‘c’ of clusters and an initial membership value (from zero to one) for each point (a customers attribute vector) in each of the ‘c’ clusters.

The pseudopartition cluster membership values for each point are chosen as adding to one, with the membership values not all equal at first. The algorithm then successively adjusts the membership values of each point in each of the various clusters, based on the point’s distance from the cluster’s center, compared to the distances from the other cluster centers.

The algorithm then uses the new membership degrees to iteratively move the cluster center points toward mutually better locations. The Euclidean distance based “center” of each cluster will be calculated from all the customers’ attribute vectors weighted by their membership degrees in the cluster. The weighting will also be recomputed based on the membership values.

The algorithm stops when the pseudo partition memberships collectively stop changing by a determined amount on successive iterations. The mathematical treatment of the algorithm can be found in [2].

The clusters after final iteration can be linguistically profiled using fuzzy logic, for comprehension to the business users. This output is then used to analyze the value of customers in each segment after removing the outliers. The algorithm used in the research is given in fig 2.

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
2. At k -step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m}$$
3. Update $U^{(k)}, U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$
4. If $\|U^{(k+1)} - U^{(k)}\| < \epsilon$ then STOP; otherwise return to step 2.

Fig. 2: Fuzzy c-means algorithm

3.3.2 Outlier Detection

The statistical definition of an “outlier” depends on the underlying distribution of the variable in question. Thus, Mendenhall et al. [6] apply the term “outliers” to values “that lie very far from the middle of the distribution in either direction”.

The importance of outlier frequency is emphasized in a slightly different definition, provided by Pyle [7]: “An outlier is a very low frequency occurrence of the value of a variable that is far away from the bulk of the values of the variable”. A more general definition of an outlier is given in [1]: “an observation (or subset of observations), which appears to be inconsistent with the remainder of that set of data”. The real cause of outlier occurrence is usually unknown to data users and/or analysts. However the presence of outlier will skew the profile. In order to detect the outliers a fuzzy logic based algorithm proposed in Simha et.al[9] is used in this research.

It is a median based algorithm, which detects the outliers, based on preset thresholds. A brief description of the algorithm is given below. Consider the data given in fig 3.

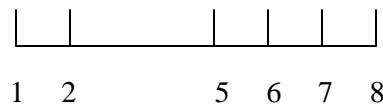


Fig. 3: A sample dataset

Let us assume we have the data in the sorted form. The difference between two successive numbers will be compared with the user provided threshold to identify the split point. For example if the threshold difference is set as 2, then the difference between third point (5) and second point (2), which is three, is a cut point for the cluster. Since most of the outliers are analyzed on univariate space, the proposed algorithm can be easily applied to detect and remove the outliers. The algorithm is given in fig 4.

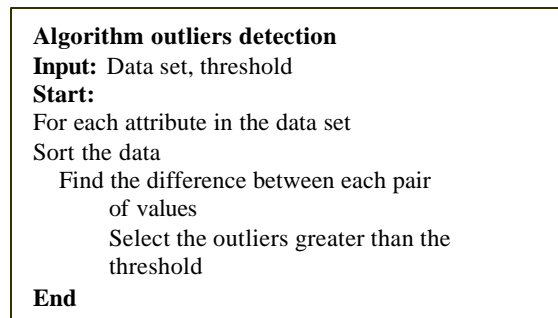


Fig. 4: Outliers detection algorithm

3.4 Customer Value Analysis

Once the segments are identified and the outliers are detected, they are removed from the segments. The core of the segments will now have approximately homogeneous members, whose value can be computed using a pre-defined method. One of the methods used in this research is to find the average revenue per call, computed over a 6 months period. This gives an insight into type of customer in each segment. This can be further used by the marketing department to identify the opportunities and better segments to serve.

5. EXPERIMENTAL RESULTS

Data mining applied to customer value analysis has been carried out for a major telecom operator. The data for the analysis was taken from a data warehouse and was lightly aggregated to a suitable granularity. The number of clusters was set to 5 after discussions with the domain experts. A sample of 50,000 customers was selected using stratified sampling based on the business logic provided by the domain experts. The value analysis process was applied to the dataset using k-means [4], EM [3] and fuzzy emeans [2] with the same outliers detection algorithm. The initial cluster centers were approximated using

domain knowledge for each cluster to reduce the run time complexities. The results of the experimental run are shown in table 1 and fig 5 to 7.

It can be observed from the table that, the performance of fuzzy data mining approach identifies and eliminates more non-representative customers. This is in line with the existing domain theory concept of “20% of customers give 80% of the revenue”. However the elimination of the outliers in all the segments may have the side effect of neglecting extremely rare cases. Work is under progress to treat these outliers separately.

Table 1: Results of a sample run

	<i>Fuzzy c-means</i>	<i>k-means</i>	<i>EM</i>
Variation in average customer value	-	22% (-)	18% (-)
Number of outliers	13,205	12,355	12,862
Run time (seconds)	3000	1800	3200

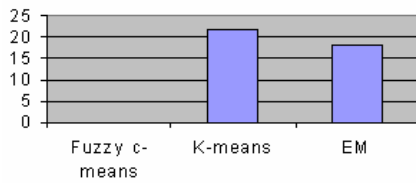


Fig. 5: Percentage variation

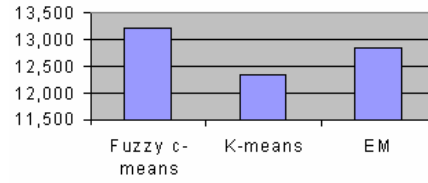


Fig. 6: Number of outliers in segments

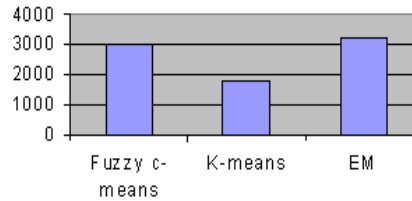


Fig. 7: Approximate run time

It can be seen that the run time for k-means algorithm is minimum, compared to other two. However the local optimization in k-means will not guarantee a good quality segments.

The variation of average value per customer computed over a period of six months with the logic provided by the domain experts was high with k-means algorithm. It is also considerably more in EM algorithm.

Further even though EM algorithm appears to be at par with fuzzy c-means approach, in case of much larger dataset, the assumption of Gaussian distribution may be violated leading to poor quality segments. This has to be tested in future, when a full-scale project is implemented.

6. CONCLUSIONS

Customer value analysis is an important CRM task. It goes beyond segmentation to remove outliers in the segments and to find the average value of stable segment. In this paper an attempt has been made to develop a fuzzy data mining based approach to value analysis. It appears that removing outliers from the segments will reflect the stable behavior of the segment. The results are encouraging and are accepted by the end users. The future work involves the analysis of outliers and investigation of the methodology for scalability.

REFERENCES

- [1] Barnett, V., Lewis, T., 1995, *Outliers in Statistical Data*, Wiley, 3rd Edition.
- [2] Bezdek J.C., *Pattern recognition with fuzzy objective function algorithms*, New York, USA: Plenum, 1981.
- [3] Dempster A.P., Laird N.M., and Rubin D.B., 1977, "Maximum Likelihood from Incomplete Data via the EM algorithm", *Journal of the Royal Statistical Society Series B*, vol. 39, 1:1 -38
- [4] Dunn J.C., 1973, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics* 3: 32-57
- [5] Fayyad Usama M. , Piatetsky-Shapiro Gregory, and SmythPadhraic., 1996, "From Data Mining to Knowledge Discovery: An Overview", In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/ The MIT Press.
- [6] Mendenhall, W., Reinmuth, J.E., & Beaver, R.J, 1993, *Statistics for Management and Economics*, Belmont, CA: Duxbury Press.
- [7] Pyle, D., 1999, *Data Preparation for Data Mining*, San Francisco, CA: Morgan Kaufmann.
- [8] Simha, Jay B., Iyengar S.S., 2005, "Fuzzy data mining applications in Customer Relationship Management", Working technical Report, Department of Computer science, Louisiana state university, Baton Rouge, LS, USA