New approach clustering algorithm for customer segmentation in automobile retailer industry

Seyed Reza Bahari Saravi
M.S. Student of Industrial engineering, Amirkabir University of Technology, Tehran, Iran

*Corresponding Author Email: arezohamzei@yahoo.com

Abstract

This research presents a new method for automatic customer segmentation based on usage behavior, without using a human specialist. In addition, the research focuses on profiling customers and finding a relation between the profile and the segments. The customer segments were constructed by applying modified Gustafson-Kessel clustering algorithm. The customer’s profile was based on personal information of the customers. A novel Support Vector Machines was used to estimate the segment of a customer based on his profile. To demonstrate the efficiency of the proposed method, this work performs an empirical study of a Nissan automobile retailer to segment over 5000 customers. This led to solutions for the customer segmentation with six segments.

Keywords: Customer segmentation, automobile retailer and support vector machines.

Introduction

Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer Segmentation can be a powerful means to identify unmet customer needs. Companies that identify underserved segments can then outperform the competition by developing uniquely appealing products and services. Customer Segmentation is most effective when a company tailors offerings to segments that are the most profitable and serves them with distinct competitive advantages. This prioritization can help companies develop marketing campaigns and pricing strategies to extract maximum value from both high- and low-profit customers. By Having Customer Segmentation, marketers can decide which marketing actions to take for each segment and then allocate scarce resources to segments in order to meet specific business objectives.

One way for customer segmentation is to defining segmentations in advance with knowledge of an expert, and dividing the customers over these segmentations by their best fits. This research tries to develop a method for customer segmentation without knowledge of an expert and without defining the segmentations in advance. Usage behavior is used for customer segmentation. Modified Gustafson-Kessel clustering algorithm is used for this purpose. Once the segmentations are obtained, for each customer a profile will be determined with the customer data. To find a relation between the profile and the segments, Support Vector Machines (SVM) will be used. SVM is able to estimate the segment of a customer by personal information. Based on the combination of the customer profile, the segment can be estimated and the usage behavior of the customer profile can be determined.

Rest of paper is organized as follows: Section 2 reviews related works. Section 3 describes the proposed method. Therefore, at first the process of selecting the right data from the data warehouse is described. It provides information about the structure of the data and the data warehouse. Then the process of clustering is discussed and is analysed. Section 4 will evaluate the proposed methods and will discuss parameters. Conclusions are proposed in section 5.

Literature review

In this section, we review the theories and concepts relevant to this research. In traditional markets, customer segmentation is one of the most significant methods used in studies of marketing. This study...
classifies existing customer segmentation methods based on data mining technique. Each data mining technique can perform one or more of the following types of data modelling:

(1) Association;
(2) Classification;
(3) Clustering;
(4) Forecasting;
(5) Regression;
(6) Sequence discovery;
(7) Visualization.

The above seven models cover the generally mentioned data mining models in various articles (Ahmed, 2004; Carrier & Povel, 2003; Mitra, Pal, & Mitra, 2002; Turban et al., 2007). There are numerous machine learning techniques available for each type of data mining model. Choices of data mining techniques for customer segmentation should be based on the data characteristics and business requirements (Carrier & Povel, 2003).

In this study, Customer relationship management is defined as helping organizations to better discriminate and more effectively allocate resources to the most profitable group of customers through the cycle of customer identification, customer attraction, customer retention and customer development. Detailed knowledge must be built up systematically so as to obtain a deeper understanding of each customer’s behaviours, characteristics and needs. Customer relationship management begins with customer identification, which is referred to as customer acquisition in some articles. This phase involves targeting the population who are most likely to become customers or most profitable to the company. Moreover, it involves analyzing customers who are being lost to the competition and how they can be won back (Kracklauer et al., 2004). Elements for customer identification include target customer analysis and customer segmentation. Target customer analysis involves seeking the profitable segments of customers through analysis of customers’ underlying characteristics, whereas customer segmentation involves the subdivision of an entire customer base into smaller customer groups or segments, consisting of customers who are relatively similar within each specific segment (Woo, Bae, & Park, 2005).

Customer profiling provides a basis for marketers to ‘communicate’ with existing customers in order to offer them better services and retaining them. This is done by assembling collected information on the customer such as demographic and personal data. Customer profiling is also used to prospect new customers using external sources, such as demographic data purchased from various sources. This data is used to find a relation with the customer segments that were constructed before. This makes it possible to estimate for each profile (the combination of demographic and personal information) the related segment and vice versa. More directly, for each profile, an estimation of the usage behavior can be obtained (Amat, 2002; Giha et al., 2003; Virvou et al., 2007).

**Methodology**

At first we explain the general customer segmentation steps and then we describe the proposed methods. Customer Segmentation has four steps:

* **Divide the market into meaningful and measurable segments according to customers’ needs, their past behaviors or their demographic profiles;**
* **Determine the profit potential of each segment by analyzing the revenue and cost impacts of serving each segment;**
* **Target segments according to their profit potential and the company’s ability to serve them in a proprietary way;**
* **Invest resources to tailor product, service, marketing and distribution programs to match the needs of each target segment;**
* **Measure performance of each segment and adjust the segmentation approach over time as market conditions change decision making throughout the organization.**

In the net of this section we explain the proposed method:

**Data collection**

The first step (after the problem formulation) in the data mining process is to understand the data. Without such an understanding, useful applications cannot be developed. In general, database marketers seek to have as much data available about customers and prospects as possible. For marketing to existing customers, more sophisticated marketers often build elaborate databases of customer information. These may include a variety of data, including name and address, history of shopping and purchases, demographics, and the history of past communications to and from customers. For larger companies with millions of customers, such data warehouses can often be multiple terabytes in size.
To enhance campaign effectiveness, Empower company (a Nissan dealer) spent one and half a years conducting a project for customer relationship management. This project collected and sampled over 5000 customers. The present study sampled 5139 customers. The detail of customer segmentation is represented and discussed in this section. To segment customer data, Empower Corporation provided 4659 pieces of historical customer data from January 1991 to April 2010. Table 1 lists the customer distribution of the nine car models. The most popular car models include SENTRA M1, MARCH, CEFIRO, ALL NEW SENTRA, NEW CEFIRO and X-TRAIL. If car models are classified according to size, SENTRA M1, MARCH, ALL NEW SENTRA were classified as small and medium size vehicles. Meanwhile, CEFIRO and NEW CEFIRO are classified as luxury size vehicles. X-TRAIL is compact SUVs.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of customer</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SENTRA M1</td>
<td>985</td>
<td>19.17</td>
</tr>
<tr>
<td>MARCH</td>
<td>900</td>
<td>17.51</td>
</tr>
<tr>
<td>CEFIRO</td>
<td>658</td>
<td>12.80</td>
</tr>
<tr>
<td>ALL NEW SENTRA</td>
<td>345</td>
<td>6.71</td>
</tr>
<tr>
<td>NEW CEFIRO</td>
<td>985</td>
<td>19.17</td>
</tr>
<tr>
<td>X-TRAIL</td>
<td>45</td>
<td>0.88</td>
</tr>
<tr>
<td>QRV</td>
<td>345</td>
<td>6.71</td>
</tr>
<tr>
<td>TEANA</td>
<td>876</td>
<td>17.05</td>
</tr>
<tr>
<td>Total</td>
<td>5139</td>
<td>100.00</td>
</tr>
</tbody>
</table>

**Modified Gustafson-Kessel Algorithm**

The data are typically summarized observations of a physical process (call behavior of a customer). Each observation of the customers calling behavior consists of \( n \) measured values, grouped into an \( n \)-dimensional row vector \( x_k = [x_{k1}, x_{k2}, \ldots, x_{km}]^T \), where \( x_k \in \mathbb{R}^n \). A set of \( N \) observations is denoted by a matrix. In this matrix, the rows of \( X \) are called patterns or objects, the columns are called the features or attributes, and \( X \) is called the pattern matrix. In this research, \( X \) will be referred to the data matrix. The rows of \( X \) represent the customers, and the columns are the feature variables of their behavior. The purpose of clustering is to find relationships between independent system variables and future values of dependent variables.

Clusters can formally be seen as subsets of the data set. One can distinguish two possible outcomes of the classification of clustering methods. Subsets can either be fuzzy or crisp (hard). Hard clustering methods are based on the classical set theory, which requires that an object either does or does not belong to a cluster.

Hard clustering in a data set \( X \) means partitioning the data into a specified number of exclusive subsets of \( X \). The number of subsets (clusters) is denoted by \( c \). Fuzzy clustering methods allow objects to belong to several clusters simultaneously, with different degrees of membership. The data set \( X \) is thus partitioned into \( c \) fuzzy subsets.

The discrete nature of hard partitioning also causes analytical and algorithmic intractability of algorithms based on analytic functionalities, since these functionalities are not differentiable. Fuzzy partition can be defined as a generalization of hard partitioning, in this case \( \mu_{ik} \) is allowed to acquire all real values between zero and one. Consider the matrix \( U = [\mu_{ik}] \), containing the fuzzy partitions, its conditions are given by:

\[
\sum_{k=1}^{c} \mu_{ik} = 1 \quad 1 \leq i \leq N.
\]

\[
a < \sum_{k=1}^{c} \mu_{ik} < 1 \quad 1 \leq k \leq c
\]

Note that there is only one difference with the conditions of the hard partitioning. In remain of this section we proposed Modified Gustafson-Kessel Algorithm.
The Gustafson and Kessel algorithm is a variation on the Fuzzy c-means algorithm (Gustafson, and Kessel, 1979). It employs a different and adaptive distance norm to recognize geometrical shapes in the data. Each cluster will have its own norm-inducing matrix $A_i$, satisfying the inner-product norm. The matrices $A_i$ are used as optimization variables in the c-means functional. This implies that each cluster is allowed to adapt the distance norm to the local topological structure of the data. A $c$-tuple of the norm-inducing matrices is defined by $A$, where $c$. The objective functional of the GK algorithm can be calculated by:

$$f(X; U, V, A) = \sum_{i}^{c} \sum_{k=1}^{N} (u_{ik})^{m} D_{ik}^{2}$$

Validation

Cluster validation refers to the problem whether a found partition is correct and how to measure the correctness of a partition. A clustering algorithm is designed to parameterize clusters in a way that it gives the best fit. However, this does not apply that the best fit is meaningful at all. We consider two measures for validation that describe as follows:

- Classification Entropy (CE): measures only the fuzziness of the cluster, which is a slightly variation on the Partition Coefficient.

$$CE(c) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij} \log(u_{ij})$$

- Partition Index (PI): expresses the ratio of the sum of compactness and separation of the clusters. Each individual cluster is measured with the cluster validation method. This value is normalized by dividing it by the fuzzy cardinality of the cluster. To receive the Partition index, the sum of the value for each individual cluster is used.

$$PI(c) = \frac{\sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^m \|x_k - v_i\|^2}{N \sum_{k=1}^{N} \|x_k - v_i\|^2}$$

PI is mainly used for the comparing of different partitions with the samenumber of clusters. A minor value of a SC means a better partitioning.

- Separation Index (SI): in contrast with the partition index (PI), the separation index uses a minimum-distance separation to validate the partitioning.

$$SI(c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} (u_{ik})^m \|x_k - v_i\|^2}{N \min_{i,j} \|x_k - v_i\|^2}$$

The findings

In this section, the cluster algorithm will be tested and their performance will be measured with the proposed validation methods. The best working cluster method will be used to determine the segments.

The disadvantage of the proposed cluster algorithm is the number of clusters that has to be given in advance. In this research the number of clusters is not known. Therefore, the optimal number of clusters has to be searched with the given validation methods of previous sections. To find the optimal number of clusters, a process called Elbow Criterion is used. The elbow criterion is a common rule of thumb to determine what number of clusters should be chosen. The elbow criterion says that one should choose a number of clusters so that adding another cluster does not add sufficient information. Unfortunately, this elbow can not always be unambiguously identified. To demonstrate the working of the elbow criterion, the feature values that represent the call behavior of the customers are used as input for the cluster algorithms.

The optimum could differ by using different validation methods. This means that the optimum could be detected by the comparison of all the results. To find the optimal number of clusters, partitions with fewer clusters are considered better, when the difference between the values of the validation measure is small. The results of the Gustafson-Kessel algorithm will be shown. In Figure 1 the results of the Partition Index and the Classification Entropy are plotted.
The optimal number of cluster cannot be rated based on those two validation methods. For the PI and the SI, the local minimum is reached at \( c = 6 \). Therefore, number of cluster for the Gustafson-Kessel algorithm will be six.

For the situation with 6 segments, the customers in these segments can be described as follows:

- Segment 1: In this segment are customers with a relative low number of voice calls. Their average call duration is also lower than average. However, their sms usage is relative high. These customers do not call to many different numbers.
- Segment 2: This segment contains customers with a relative high number of contacts. They also call to many different areas. They have also more contacts with a Vodafone mobile.
- Segment 3: The customers in this segment make relative many voice calls. Their sms usage is low. In proportion, they make more international phone calls than other customers.
- Segment 4: These customers are the average customers. None of the feature values is high or low.
- Segment 5: These customers do not receive many voice calls. The average call duration is low. They also receive and originate a low number of sms messages.
- Segment 6: These customers originate and receive many voice calls. They also send and receive many sms messages. The duration of their voice calls is longer than average. The percentage of international calls is high.

A Support Vector Machine is an algorithm that learns by example to assign labels to objects (Nobel, 2006). In this research a Support Vector machine will be used to recognize the segment of a customer by examining thousands of customers of each segment.

How does a SVM discriminate between large varieties of classes? There are several approaches proposed, but two methods are the most popular and most used (Nobel, 2006). The first approach is to train multiple, one-versus-all classifiers. For example, if the SVM has to recognize three classes, A, B and C, one can simply train three separate SVM to answer the binary questions, “Is it A?”, “Is it B?” and “Is it C?”.

Another simple approach is the one-versus-one where \( \frac{k(k - 1)}{2} \) models are constructed, where \( k \) is the number of classes. In this research the one-versus-one technique will be used.

To avoid over fitting, cross-validation is used to evaluate the fitting provided by each parameter value set tried during the experiments. The optimal parameters for the Support Vector Machine will be researched and examined. Since this research use, the radial basis function, one variable must be set. Results of the Radial Basis function are given in Table 2. The best result with 6 segments is 78.5%.

<table>
<thead>
<tr>
<th>( c )</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>500</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>73.6</td>
<td>77.4</td>
<td>72.6</td>
<td>70.9</td>
<td>68.0</td>
<td>65.1</td>
<td>52.7</td>
<td>51.8</td>
<td>40.0</td>
</tr>
<tr>
<td>0.6</td>
<td>72.5</td>
<td>74.8</td>
<td>74.8</td>
<td>72.7</td>
<td>70.4</td>
<td>54.0</td>
<td>49.3</td>
<td>39.1</td>
<td>49.3</td>
</tr>
<tr>
<td>1.0</td>
<td>71.4</td>
<td>74.8</td>
<td>72.6</td>
<td>71.3</td>
<td>70.2</td>
<td>50.0</td>
<td>45.5</td>
<td>37.2</td>
<td>50.3</td>
</tr>
</tbody>
</table>

Table 2. Radial basis function, 6 segments
In this research, the features will be validated. The importance of each feature will be measured. This will be done, by leaving one feature out of the feature vector and train the SVM without this feature. The results are shown in Figure 2. The results show that Age is an important feature for classifying the right segment. Each feature increases the result and therefore each feature is useful for the classification.

![Figure 2. Results while leaving out one of the features with 6 segments](image)

**Summary and Conclusions**

The first objective of our research was to perform automatic customer segmentation based on usage behavior, without the direct intervention of a human specialist. The second part of the research was focused on profiling customers and finding a relation between the profile and the segments. The customer segments were constructed by applying Gustafson-Kessel Algorithm. The clustering algorithms used selected and preprocessed data from the Vodafone data warehouse. This led to solutions for the customer segmentation with respectively four segments and six segments. The customer’s profile was based on personal information of the customers. A novel data mining technique, called Support Vector Machines was used to estimate the segment of a customer based on his profile.

**References**


