Life Insurance Customers segmentation using fuzzy clustering

Gholamreza Jandaghi*, Hashem Moazzez, Zahra Moradpour
Faculty of Management and Accounting, Farabi College, University of Tehran, Tehran, Iran
E-mail address: jandaghi@ut.ac.ir

ABSTRACT

One of the important issues in service organizations is to identify the customers, understanding their difference and ranking them. Recently, the customer value as a quantitative parameter has been used for segmenting customers. A practical solution for analytical development is using analytical techniques such as dynamic clustering algorithms and programs to explore the dynamics in consumer preferences. The aim of this research is to understand the current customer behavior and suggest a suitable policy for new customers in order to attain the highest benefits and customer satisfaction. To identify such market in life insurance customers, We have used the FKM.pf.niose fuzzy clustering technique for classifying the customers based on their demographic and behavioral data of 1071 people in the period April to October 2014. Results show the optimal number of clusters is 3. These three clusters can be named as: investment, security of life and a combination of both. Some suggestions are presented to improve the performance of the insurance company.

Keywords: Market segmentation; customer segmentation; data mining; fuzzy clustering; life insurance

1. INTRODUCTION

In recent years, many companies are paying more attention to profit by expanding sales of existing customers. It is important to determine customer preferences in formulating market strategies (Lin et al. 2012). CRM or customer relationship management is a strategy that is
taken from the concept of business intelligence (BI) and its main objective is the establishment of a close relationship to meet customer needs and to obtain information needed for innovation and achieve a higher degree of productivity of the company. CRM as a technology tool is located at the whole organization. For this reason, it is possible that its measure affects the firm performance (Carmen & Cristina 2014). Nowadays, mass marketing approach cannot meet customers' needs and varied preferences and many companies need to communicate and manage their customers by offering various attractions, personalization and product or service. Market segmentation is thought to give a response to partitioning customers to groups of individuals with similar needs and buying patterns. With proper market segmentation, companies can target goals or field service to customers and thus improve productivity through marketing strategies. In order to better understand their customers, companies may integrate abundance data collected from multiple channels.

In general, variables can classified in two groups, variables associated with clients and variables related to the product (Wang 2010). The key to the survival in a competitive market lies in understanding customers better. A popular tool for segmentation of customers based on their profiles is clustering algorithm (Bose & Chen 2014). Data mining approach will lead organizations to move towards a customer-oriented one. Today, organizations are trying to implement CRM systems to lower their weakness. The key to success in CRM is having an effective strategy for data management and data warehouse and capabilities to analyze customer interaction data in the business environment. CRM is a core concept with multiple layers and data mining is thought to be an important layer in CRM (Behrouzian-Nejad, et al. 2012). Data analysis capability is a very important factor for the business. Data mining is an interactive and iterative process that involves multiple steps based on which many decisions are taken (Wegener & Ruping, 2010).

The insurance industry has been rapidly growing through technological advances and new ideas are driven (Goonetilleke & Caldera 2013). Data mining models has been applied to help insurance companies in the acquisition of new customers, retain existing customers (Devale & Kulkarni 2012). Insurance companies provide unique financial services for economic growth and development. World trade is unstable and risky without insurance since the variety of risks and uncertainties of the global economy is changing (Akotey et al. 2013). Life insurance is seen as the risk of certain individuals. Different from other industries, the sold products of life insurance business are invisible and untouchable. People have an important role in transferring knowledge and customer service in life insurance industry. On the other hand, most of life insurance policies are long-term. In addition, life insurance companies have sustainable services, sometimes for the whole life for their customers (Huang & Lai 2012). In this study, by using fuzzy clustering algorithm based on the demographics and behavioral characteristics of customers, we will recognize some clusters of customers.

2. LITERATURE REVIEW

Data mining techniques is used to identify patterns in large amounts of data. Data mining includes statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and knowledge from large databases. There are several different functions of data mining algorithms used in particular to
include the association, classification, clustering, modeling, sequential patterns (Kirlidog & Asuk, 2012). The steps of knowledge discovery is as below:

1) Data cleaning: Remove the confusing and inconsistent data
2) Data integrity: the combination of different data sources
3) Select the data: finding relevant data from the database
4) Transfer Data: conversion of data into a form suitable for exploring
5) Data Mining: Data mining models for the use of technology
6) Pattern evaluation: evaluation models really useful for current students
7) Provide knowledge of the current knowledge of exploration for users using technology such as visual presentation (Sithic & Balasubramanian 2013).

Life insurance is a contract whereby the insurer, in exchange for a premium, undertakes to pay insured in case of death or life or the end of the insurance capital as a lump sum to beneficiaries or users (Bakhshi 2011). Segmentation for the first time in 1956 was introduced by Smith in the marketing literature. Since then, segmentation has been used as an alternative concept instead of strategy. Cutler (2003) has introduced five criteria for effective segmentation, including the measurability, accessibility, nature, distinction and action (Hiziroglu 2013). Market segmentation is defined as a division of a market into distinct groups of customers with significant customer specifications (Ming-Chih, et al. 2011). Clustering is partitioning the objects into meaningful groups. Objects in a cluster are very similar to each other but very different with while objects in other clusters. Clustering also known as division or separation of data and clustering is taken into account as an unsupervised classification (Sithic & Balasubramanian 2013).

Clustering generally is classified in two ways: absolute clustering and fuzzy clustering. In final clusters, each data point is explicitly assigned to one and only one cluster. Since the boundary of clusters are hard, decisive and do not overlap. Fuzzy Clustering (FCM) is a strong and flexible approach to natural data sets that include poorly defined borders that should result in overlapping objectives. In contrast, classical clustering algorithm is certain, while in FCM each data point belongs to more than one cluster and associated with the concept of membership degree between 0 and 1 (Dai, 2011).

Kaffashpour et al (2011) in a study segmented customers based on their life cycle model using data mining based on RFM. They used indices like “the time between the last pickup at the end of the specified period”, “the number of customer purchases a specified time period”, and “the amount of customer purchases in Rials” and weighting them using analytic hierarchy process. In that study, 260 customer of East-Toos company have been studied in a one-year period resulting in 8 clusters. Morovvati Sharifabadi (2013) in their study of "clustering of bank customers using competitive neural networks" discussed and compared the use of artificial neural networks and statistical methods traditionally competitive with each other to cluster clients to search a database of 600 customers in 2010. For clustering customers, seven key characteristics (including account opening duration, the average six-month current account, gender, education, age, occupation, number of visits in a month) were used. According to the results show considerable competitiveness of artificial neural network over of statistical methods. Tavakoli et al (2009) in their study, "the use of data mining process to predict customer defections patterns in insurance" explored data mining capabilities in the management of customer defections, and using standard methodology of
CRISP- DM. They used 31616 record of fire insurance customers. A decision tree method was carried out on the data containing 36 variables. The results showed that the main predictive factor was customer acquisition channel in retention of customer.

Singh Rana (Singh, & Rana., 2013) focused on mining customer data in the automotive industry using clustering techniques. The aim of that study was that the customer segmentation. Based on their analysis the customers were classified loyal customers, very satisfied customer, excellent, above average, average and passive. They used K-means clustering to explore 631 records with 12 demographic variables such as age, expectations, utility, performance, mileage, economic environment, value for money, exchange Cars (trade) and employment.

Crespo and Weber (Crespo & Weber, 2005) studied a methodology for dynamic data mining based on fuzzy clustering. Customer traffic behavior for 16 months from 1999 to 2000 was used. They applied FCM clustering using the original data set consisting of 5822 customer, each is described by 86 features. With the implementation of the proposed system they showed benefits in two functional areas: customer segmentation and traffic management. Cheng and Chen (Cheng & Chen., 2009) in a study of customer value segmentation by RFM model and Rough set theory Rough, proposed a new method based on some of the features of RFM and k-means algorithm. The data set collected in 2006 from Taiwan's electronic industry. In that study, they used RFM model.

3. RESEARCH FRAMEWORK

This study, in terms of purpose, is an applied research and in terms of data collection is a cross-sectional research. The population used in this study consisted of 1071 individuals from Pasargad life insurance. The collected data is for the period from April to October 2014. In this study, demographic information and information about the items affecting customers insurance is examined. Figure 1 shows the flowchart of the study.

Phase 1: Determine the business purpose or understanding: This stage is usually done to identify the requirements and objectives of the client's data mining. The main objective of this study was to analyze customer data to the same pattern between the different segments of customers based on fuzzy clustering FKM. Finally, the clusters obtained thereby is used to classify and identify customers to achieve customer policy. It also results in improved customer relationship management and marketing strategy to be used for different groups of customers.

Phase 2: data collection (defined variables): In this section, based on objective of the research, characteristics of the customers are specified, then the data collection and review of data collection based on the profile of insured persons, will be discussed.

Phase 3: Clearing and preparation of data: information received from the insurance was loaded in a database. Records were examined in terms of missing values. Outliers have a bad impact on clustering. But in cases like segmentation of customers is not required to remove them because they may be rewarding to customers. In this section, fill in the void, paving the confusion and inconsistencies resolved to purge data.

Phase 4: Modeling: In this part of the construction of the proposed model will be done on the data set. Then you need to describe the individual components of the model. Here FCM
analytical method was used to build the model. The following steps have been performed in this phase.

\[ X = \text{data that is to be partitioned to } k \text{ clusters} \]

where:

\[ X_j = \{X_{1j}, X_{2j}, \ldots, X_{lj}\} \] indicates the j-th sample for \( j = 1,2,\ldots,d \)

\[ v_i = \{v_{1i}, v_{2i}, \ldots, v_{ti}\} \] indicates the center of the i-th cluster

\( u_{ij} \) indicates the degree by which \( X_j \) belongs to i-th cluster with the partitioning matrix \( U = (u_{ji})_{dxk} \)

The FKM algorithm is defined based on the following objective function:

\[ J_{fuzz} = \sum_{j=1}^{d} \sum_{i=1}^{k} u_{ji}^{m} d_{ji}^{2} \]  

(1)

where

\[ d_{ji} = \sqrt{\sum_{p=1}^{d} (v_{pi} - x_{pj})^{2}} \]  

(2)

and \( m \) indicates the fuzzifying parameter.

\( u_{ji} \) and \( v_{i} \) are defined as below:

\[ u_{ji} = \frac{1}{\sum_{p=1}^{k} \left(\frac{d_{ji}}{d_{jp}}\right)^{(m-1)}} \]  

(3)

\[ v_{i} = \frac{\sum_{j=1}^{d} u_{ji}^{m} x_{j}}{\sum_{j=1}^{d} u_{ji}^{m}} \]  

(4)

Therefore the FKM algorithm has the following steps:

Step 1: choosing a number of \( k \) clusters, the \( m \) fuzzification degree and threshold values \( \varepsilon \).

Step 2: calculation of cluster centers \( v_{i} \) \( i = 1,2,\ldots,k \) according to the above equation.

Step 3: Calculate the Euclidean distance \( d_{ji} \) of the sample \( X_{j} \) using equation (2). Then all \( u_{ji} \) calculated using equation (3) and update the fuzzy partition matrix \( U \).
Step 4: Calculate the objective function $J_{fuzz}$ using equation (1).

Step 5: Check the convergence of the algorithm using $\varepsilon$. If converged, return the results, otherwise go to step 2.

Finally the result of the FKM algorithm and cluster centers is a fuzzy partition matrix, a $t \times k$ matrix whose i-th column shows the vector $v_i$ as follows

$$C_k = \{ v_{1i}, v_{2i}, ..., v_{ki} \}$$

The parameter $m$ (fuzzy parameter) is an important parameter in FKM. With higher values of $m$, we will have softer partitions and lower values of $m$ yields in harder partitions. (Kumar 2011).

![Flow chart of the fuzzy clustering](image)

**Figure 1.** Flow chart of the fuzzy clustering.
Segmentation variables such as life insurance customers

To select the proper characteristics in this study, first we examined the characteristics of each customer. At this stage a number of characteristics such as (name, surname, place of birth, etc...) that is not related to their behavioral traits, were deleted.

Finally, based on the aim of this research segmentation of insurance customers and finding hidden relationships among the data for future projections, the total that were used is given in Table 1.

Table 1. Characteristics of the insured people.

<table>
<thead>
<tr>
<th>Insurance characteristics</th>
<th>Demographic characteristics of insured people</th>
</tr>
</thead>
<tbody>
<tr>
<td>●The insurer relationship to the insured person</td>
<td>●age</td>
</tr>
<tr>
<td>●User of insurance capital in case of being alive</td>
<td>●gender</td>
</tr>
<tr>
<td>●Payment method</td>
<td>●Number of children</td>
</tr>
<tr>
<td>●Insurance term</td>
<td>●Marital status</td>
</tr>
<tr>
<td>●The premium</td>
<td>●Job</td>
</tr>
<tr>
<td>●Additional coverage</td>
<td>●Place of living</td>
</tr>
<tr>
<td>●Life factor</td>
<td>●Annual premium increase</td>
</tr>
<tr>
<td>●Annual increase of life insurance capital</td>
<td>●Annual increase of life insurance capital</td>
</tr>
<tr>
<td>●The capital increase of fire insurance</td>
<td>●The capital increase of fire insurance</td>
</tr>
<tr>
<td>●Disease history</td>
<td>●Disease history</td>
</tr>
<tr>
<td>●Family medical history</td>
<td>●Family medical history</td>
</tr>
<tr>
<td>●Weight</td>
<td>●Weight</td>
</tr>
<tr>
<td>●Height</td>
<td>●Height</td>
</tr>
<tr>
<td>●Final Capital</td>
<td>●Final Capital</td>
</tr>
</tbody>
</table>

Data analysis

Data analysis is a multi-stage process in which the data summary, coding and classification, and finally processing to establish context and analysis of the relationship between the data. The data must be analyzed and transformed data into understandable information. In this study, the function Fclust in R statistical software packages was used for fuzzy clustering analysis.

Although in FCM method, a predetermined number of clusters is not clear, but in the beginning the number of clusters for segmenting customers and the appropriate number of clusters is determined by trial and error. It is advised that the number of clusters can be between 2 and 6. We chose the number of clusters as $K = 3$. The fuzzy parameter $m$ has been taken 1.5. Table 1 shows a part of output in which the case number and the cluster
number. Table 2 shows a part of the matrix of fuzzy membership of each client in each cluster. For example, the person with the highest degree of membership in cluster 2 owns 0.84. Total value of fuzzy membership (for a client) is also equals 1.

**Table 2.** Clustering of people in 3 clusters.

<table>
<thead>
<tr>
<th>Case number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster number</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Case number</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td>Cluster number</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Case number</td>
<td>39</td>
<td>40</td>
<td>41</td>
<td>42</td>
<td>43</td>
<td>44</td>
<td>45</td>
<td>46</td>
<td>47</td>
<td>48</td>
<td>49</td>
<td>50</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>55</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>Cluster number</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

**Table 3.** Membership degree of insured people for each cluster.

<table>
<thead>
<tr>
<th>Clus 1</th>
<th>Clus 2</th>
<th>Clus 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.13</td>
<td>0.84</td>
</tr>
<tr>
<td>2</td>
<td>0.98</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.24</td>
<td>0.67</td>
</tr>
<tr>
<td>4</td>
<td>0.39</td>
<td>0.57</td>
</tr>
<tr>
<td>5</td>
<td>0.21</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>0.17</td>
<td>0.79</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>0.15</td>
<td>0.78</td>
</tr>
</tbody>
</table>
4. RESULTS

According to the output of the clustering algorithm, the 3 clusters have the following specifications:

Cluster 1:
40 percent of individuals belong to this cluster.
61 percent of these individuals have age between 20 to 29
77 percent are men
55 percent are singles
80 percent have no children
64 percent pays their insurance on monthly base
58 percent have premium under 990000 Rials
61 percent have supplement coverage
41 percent are self-employed

Therefore, the first cluster of customers are single young men, most of whom are self-employed, as can be seen from 3 supplement coverage, they pay attention to safety of life, 30 years of insurance individuals, and 43% of the capital value of the investment is indicative of the final end user of the long-term care and self-insurance funds that invest their clients to the cluster. Since the largest number of customers falls in cluster 1, this cluster can be called cluster of customers with high profitability.

Cluster 2:
30 percent of the customers fall in this cluster
48 percent have between 30 to 39 years of age
51 percent are men
94 percent are married
77 percent have 1 to 3 children
41 percent are self-employed
60 percent are on monthly payment
63 percent have 3 supplement coverage

So the people in the second cluster are mostly self-employed clients and almost young and married. The male to female ratio is almost the same. The insurance term of these individuals show that they intend to reach the final capital earlier. This cluster can be called the cluster of low profitability customers.

Cluster 3:
This cluster 40 percent of individuals
54 percent are men
99 percent are singles
51 percent are students
55 percent are on monthly payment
41 percent have no supplement coverage

Therefore, the third cluster of customers of students who do not have any source of income to pay the premium. This cluster can be called short-term investment.

5. DISCUSSION AND CONCLUSION

The findings of this study identify useful characteristics when considering life insurance industry. These findings are consistent with other studies. The acquisition of new customers is an important business problem. Although the traditional method of simply trying to increase the customer base is expanding efforts of unit sales, but sales efforts associated with data mining methods will lead to a more successful results. Traditional sales methods aim to increase the number of insured simply by targeting those who have limited visitation. The disadvantage of this method is that a lot of marketing efforts may have little efficiency. Statistical method known as "cluster analysis" can be used to identify different market segments (Devale & Kulkarni, 2012). Insurance companies need to know the principles of decision making and data mining techniques for competition in the insurance market's life. Another study by Nagorno-race (2009) used data mining functionality embedded in insurance calculations and data mining process, various methods including linear and nonlinear calculations and functionality of the various branches of insurance.

References


( Received 26 August 2015; accepted 11 September 2015 )