IJCSN - International Journal of Computer Science and Network, Volume 7, Issue 3, June 2018
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Integrating Data Mining and Knowledge Management to Improve Customer Relationship Management in Banking Industry (Case Study of Caspian Credit Institution)

[1] **Mohammad Ordouei** , [2] **Dr. Touraj BaniRostam**

[1] Computer Engineering Dep., Islamic Azad University, Central Tehran Branch (IAUCTB)
Tehran, Iran

[2] Computer Engineering Dep., Islamic Azad University, Central Tehran Branch (IAUCTB)
Tehran, Iran

Abstract - Banks and financial - credit institutions by customer relationship management database analysis can manage to identify customers and allocate resources to profitable customers in a better manner. This study aims to find an index for customers, using customer characteristics to identify significant profitable customers of banking system. In this way we will be able to provide them with more adequate facilities. To do so, mix method of fuzzy clustering and imperialist competitive algorithm has been applied to accomplish customer data clustering; then Fuzzy C-Means and criteria of sum of intra – cluster distances were used to evaluate the method. Results display minimum and maximum as 39.270153 and 53.100917 respectively which at least are 6000 units less than the compared method. This indicates advantage of the suggested method compare to the other.

Keywords – Data Mining; CRM

1. Introduction

Some of the problems of banks include lack of identification of customers , lack of appropriate decision management, and moody treatment with customers. Customer selection is one of important issues of customer – based marketing. From this perspective, basis of value creation via customers is to find profitable or potential profitable customers. Nowadays capability of identification of profitable customers, creating endurable customer loyalty and developing customer relationships are among key competitive factors of an organization.

Organizations produce and save a large amount of data. Data management, data processing and obtaining a whole picture of produced data reflect challenges which are key of success in competition as well [1]. Knowledge management as a strategic tool to make maximum use of knowledge within an organization is an essential tool for every organization. Leading and successful organizations that outperform their competitors in competition field are those who rely on their knowledge and explore their invisible knowledge to gain competitive advantages, in contrary to traditional organizations who count on their physical and financial resources as advantage factors.

Strategies based on finding customers and customer retention in a correct way, create value significantly. So, classification of customers as one of the main approaches of marketing plays an important role in customer relationship management. Organizations by classifying customers will be able to adopt various strategies based on their customer characteristics to maximize customer potential value [2]. Data mining as a tool of discovering and exploiting invisible knowledge of organizations and transforming it to elicit new knowledge plays an important role. Agents of customer relationship management can predict future behavior and customer lifetime according to past behavior. Various data mining techniques can help organizations to elicit behavioral patterns and meaningful dynamics of customers and make it possible to identify and classify customers and predict of potential profitable customers. Clustering and classifying are two of data mining techniques. This study has made use of clustering method. Clustering method attempts to segment a data set into multitude clusters in such a way that data within a cluster are of most similarity with each other and of most difference from the other data clusters.

In the first section, literature of fuzzy clustering will be reviewed; in the section two the research model will be presented; In the section three, data analysis and finally in the last section, result and conclusion will be delivered.

2. Literature review

Multitude studies have been devoted to implementation of data mining in banking industry, some of which are as below:

In [3], customer behavioral patterns of bank customers have been categorized by means of neural networks and self – organizing method; exploring customer financial behaviors, a categorization of profitable customers have been identified. In this study, bank customers have been classified in to 3 main categories. This study shows that identifying customer characteristics by behavioral ranking model is useful and facilitates development of marketing strategies.

In [4], researchers have defined customer relationship management as business processes aiming to knowledge – based communicate with customer for creating added value. In this reference, using a data set of 153 Spain hotels, relations between successful knowledge management and customer relationship management has been explored by structural equation model. Results indicated that knowledge management is not adequate in itself for successful customer relationship management and other factors like organizational factors are effective as well.

[5] Is a review study. This reference has considered implementation of data mining and has noted that applying data mining in customer relationship management is not limited to marketing, risk management and fraud detection; wining customer and keeping customer are among other usages of data mining. Furthermore, this reference defines data mining as a powerful instrument for banks and retail industries.

In [6] evaluation of customers' needs by means of clustering algorithms and association rule learning have been suggested. The process of integrating classification with data mining provide valuable information to marketers; here the proposed method is to use clustering method and discovery of association rules to identify favorable factors of customers and distinguish their preferences. This method after gathering data by clustering method classifies customers based on their behaviors; then elicits association rules of each cluster to identify customer satisfaction factors.

In [7] a mixed framework of customer relationship management and data mining has been presented. To do so, simple biz categorization model and neural system have been used. According to the reported results, neural network accuracy is relatively better.

Reference [2], has been provided us with knowledge of bank customer ranking based on their shares or customer lifetime value (CLV). According to its results clustering customers in different groups not only yields customer ranking based on parameters of customer investment but also helps us to recognize different segments of market more explicitly; so that, more effective strategies based on production ownership and customer transaction will be developed. Customer lifetime value (CLV) has a lot of implications in marketing and CRM field.

3. Fuzzy Clustering

In classic clustering each sample belongs to one and only one cluster; it cannot be a member of two or more clusters. The main difference between classic and fuzzy clustering method lies in this fact that in fuzzy clustering one sample is allowed to be assigned to more than one cluster.

One of the most important and practical clustering algorithms is C mean algorithm. In fuzzy version of this algorithm, the number of clusters (c) has been recognized previously.

Below, fuzzy C- means has been explained in details:

In C clustering algorithm, fuzzy C- means of target function is like relation (1) [8, 9, and 10]:

$$J = \sum_{I=1}^{C} \sum_{k=1}^{n} {u_{ik}}^{m} {d_{ik}}^{2} = \sum_{I=1}^{C} \sum_{k=1}^{n} {u_{ik}}^{m} \|x_k - v_i\|^2 \quad (1)$$

In (1) formulation , m is a real number bigger than 1; in most cases 2 is assigned to m. if in the above formulation m is assumed equal to 1, we will have a non – fuzzy target function of clustering c mean (classic). In the above formulation xk is the sample no. k and vi is the representative or center of cluster no. i and n indicates number of clusters. u_{ik} is the belongingness degree of sample i in cluster k. ‖*‖ symbol is the sign of similarity (distance) of sample to (not to) cluster center. One can make use of any kind of function indicating similarity of sample and center of cluster. A U- matrix can by defined by uik possessing c rows and n columns which its features take an amount between zero to 1. If all features of U - Matrix adopt zero to 1 , the algorithm will be like classical c- mean. Though any amount between zero to 1 may be assigned to U -matrix, the sum of features of each column should be equal to 1 so we have [9]:

$$\sum_{i=1}^{c} u_{ik} = 1, \forall k = 1, \ldots, n \quad (2)$$

This term means that sum of belongingness of each sample to C cluster should be equal to 1. To define a formula related to uik and vi, we must minimize the specified target function. Using the above term and assuming deferential of target function as below we will have:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m-1)}} \quad (3)$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}} \quad (4)$$

4. Research model

Each Individual makes use of services in a different way but with similar patterns. This means that it is possible to distinguish various groups of customers.

The proposed algorithm in this study for customer classification is as follow:

1. Data collection
2. Data preprocessing
3. Applying imperialist competitive algorithm on data; obtaining suitable amount for cluster centers
4. Implementing fuzzy clustering algorithm
5. Evaluation of results

4.1 Data preprocessing

In data pre- processing stage to obtain better results the amount of each feature has been normalized from zero to 1; then rows of total matrix of data will be moved casually to change primary data order. Normalization is necessary to get more accuracy. To normalize amounts of each data set, relation (5) has been applied [11].

$$\text{Normalize}(x) = \frac{(\mathrm{x} - \mathrm{X}_{\min})}{(\mathrm{X}_{\max} - \mathrm{X}_{\min})} \quad (5)$$

$\mathrm{X}_{\max}$ and $\mathrm{X}_{\min}$ are maximum and minimum amounts of X feature domain. After data normalization, the amounts of all features will be embedded in interval [0,1] .

4.2 Integration of imperialist competitive algorithm and fuzzy C- mean

Smart algorithms always have been employed as a general search method for many optimization issues one of which is clustering issue. Clustering is a process of receiving input data set and dividing them into sub – groups. Clustering fuzzy algorithms like FCM are explicitly more privileged compare to definite samples. Though FCM is preferable in compare to definite method in determining categories cluster centers should be specified arbitrary so their possibility of getting stuck in local peaks is very high. As a result a new algorithm is presented which by using fuzzy logic and imperialist competitive algorithm avoids from getting stuck in local peaks and automatically finds the optimized comprehensive answer, means the optimized location of cluster center. In this combination using optimized amounts of cluster centers proposed by imperialist competitive algorithm better results and proper identification become more accessible. In this study to integrate fuzzy logic and imperialist competitive algorithm methods, first, cluster centers will be specified by imperialist competitive algorithm then clustering will be performed through fuzzy c – mean clustering method.

Applying imperialist competitive algorithm to determine optimized cluster centers we find them in such a way that:

1. Each member connects to the nearest similar cluster.
2. Similar data connects to each other in one point (sitting in one cluster)

We considered the cost function as below:

$$\min_{i} \sum_{i=1}^{n} \min_{j} \|x_i - c_j\| \quad (6)$$

Unknowns of c1, c2, …cm are cluster centers. Here m indicates number of clusters, n indicates number of data and xi indicates data.

After applying imperialist competitive algorithm, we found cluster centers; then FCM algorithm had been implemented to accomplish clustering.

The stages of implementing FCM algorithm are as below:

1. Cluster centers found by imperialist competitive algorithm are considered as primary amounts of cluster centers.
2. The degree of input membership in each class will be determined. The degree of belongingness of i input to j cluster (u_{ij}) is something between 0 and 1 (0≤ u_{ij} ≤1) which is determined via formulation below:

$$U_j(x) = \frac{1}{\sum_k \left[\frac{D(c_j, x)}{D(c_k, x)}\right]^{\frac{2}{m-1}}} \tag{7}$$

3. All c_j are updated by this formula.

$$c_j = \frac{\sum_i u_{ij} x_i}{\sum_i u_{ij}} \tag{8}$$

4. Repetition of stage 2 to 3 till no evident change has been detected.

To evaluate the results minimum and maximum of sum of intra – cluster distances in addition to degree of repetitions will be considered,

5. Analysis

5.1 Statistical population and sample

Statistical population includes recorded customer information bank of Caspian credit institution in Iran which is available via Informatics administrations of the above institution. The statistical population is about 40000 customers. Customer information bank consist of 11 features including account no., the passed period after opening date, age, education degree, income level, marital status , offspring no., people under care, account residual, housing status, account transactions.
Given that more data will yield more accurate data mining results, sampling includes a time period of one year, from April 2016 to March 2017 (Farvardin to Esfand 1395). Records of those customers are selected which have joined to the institution before the specified time period.
Sampling has been carried out by Morgan table method. Regarding the statistical population a sample of 380 records has been selected.

5.2 Research variables

Features under investigation in this study consist of:

- Bank account no.
- the passed period after opening date (months)
- age
- education degree (under high school, high school diploma, A.A or A.S. , B.A., M.A., PhD., and more)
- income level monthly (under 1 million , 1- 2 millions, 2-3 millions, 3-4 millions, 4-5 millions, more than 5 millions)
- Marital status
- Offspring no.
- People under care
- Account residual
- Residential place status
- Account transactions

5.3 Findings

Proper data, good pre- processing, appreciate data mining method yields good results for banking data. In this study we tried to observe the above in order to provide a well – developed model. To do so, a clustering - based mix method has been proposed. Ideal clustering is accessible when there is minimum similarity among discreet classes and maximum similarity among data within a cluster. Selection of primary amounts of cluster centers in fuzzy c–mean algorithm is of great effect on final finding. So, in this study we attempt to determine cluster centers by imperialist competitive algorithm and then consider these centers as input data for fuzzy c – mean algorithm.
MATLAB software has been applied to explore and extract results.
Data set consists of 11 variables from which “cycles of account transactions” has been considered as the target variable (output).
As mentioned before, stages of implementation of model include: data collection, data preprocessing, finding cluster centers by imperialist competitive algorithm, clustering by fuzzy c- mean method. Following is the description of each stage.
In preprocessing stage, data has been transferred into a format usable to the software. First, bank account no. variable has been excluded due to the fact that it does not effect on output; it have been assigned from 1 to 6 points to under high school to PhD and more, respectively, to education degree variable. Average monthly income from less than 1 million Tomans to 5 million Tomans and more adopted from 1 to 6 respectively; in case of marital status , 1 point has been assigned to married status and 2 to single. Furthermore, in case of residential place status , 1 point has been assigned to positive response and 2 to negative response; after normalization, the amount of data has been between 0 and 1.
In this study, first cluster centers have been determined by imperialist competitive algorithm then clustering is done through fuzzy c- mean method.
The primary population for imperialist competitive algorithm according to trial and error approach has been considered 100 and repetitions have been assumed as 100. Number of realms have been assumed as 10.
After implementing imperialist competitive algorithm, cluster centers are as below:

(1.8587,0.2529)
(3.3036, 0.3942)

Cluster centers found by imperialist competitive algorithm, then are dealt with FCM algorithm for clustering purpose.
The results of mix method for a sample of 50 records have been displayed in table (2) for example.

IJCSN - International Journal of Computer Science and Network, Volume 7, Issue 3, June 2018
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Table 1: Results of mix method applied in this study

	Percentage of belongingness to cluster1 *(class of people with high transaction)*	*Percentage of belongingness to cluster 2* *(class of people with low transaction)*
1	0.841741	**0.158259**
2	0.718638	**0.281362**
3	0.077073	**0.922927**
4	0.132456	**0.867544**
5	0.452228	**0.547772**
6	0.407355	**0.592645**
7	0.660103	**0.339897**
8	0.854377	**0.145623**
9	0.949186	**0.050814**
10	0.110743	**0.889257**
11	0.076166	**0.923834**
12	0.776714	**0.223286**
13	0.118004	**0.881996**
14	0.075492	**0.924508**
15	0.781011	**0.218989**
16	0.182622	**0.817378**
17	0.853225	**0.146775**
18	0.943182	**0.056818**
19	0.118004	**0.881996**
20	0.075492	**0.924508**
21	0.781011	**0.218989**
22	0.182622	**0.817378**
23	0.795073	**0.204927**
24	0.797334	**0.202666**
25	0.50571	**0.49429**
26	0.687569	**0.312431**
27	0.091994	**0.908006**
28	0.118004	**0.881996**
29	0.771784	**0.228216**
30	0.789677	**0.210323**
31	0.510323	**0.489677**
32	0.67774	**0.32226**
33	0.088664	**0.911336**
34	0.439821	**0.560179**
35	0.708304	**0.291696**
36	0.86476	**0.13524**
37	0.93249	**0.06751**
38	0.108374	**0.891626**
39	0.069469	**0.930531**
40	0.778825	**0.221175**
41	0.921179	**0.078821**
42	0.439821	**0.560179**
43	0.66361	**0.33639**
44	0.118004	**0.881996**
45	0.075492	**0.924508**
46	0.781011	**0.218989**
47	0.182622	**0.817378**
48	0.439821	**0.560179**
49	0.592638	**0.407362**
50	0.728531	**0.271469**

This algorithm after 20 repetitions has halted on cost function 39.270153 .

Beholding outputs one may conclude that people aged 30–50 years old have high transactions; furthermore education degree and average income are in direct relation with transaction numbers. Number of offspring and under care individuals are in converse relation with transaction numbers. In addition , single people owning residential

place , with stock more than 10 millions display high transactions.
To determine a new customer's performance the related cluster has been identified and the performance has been predicted. Results of proposed algorithm have been compared with fuzzy c- mean algorithm.
Minimum and maximum amounts of sum of intra – cluster distances of the proposed approach through repetitions, number of repetitions and the applied method has been displayed in table (3).

Table 2: comparing mix method with fuzzy c- means method

No. of repetitions	Sum of intra – cluster distance criteria		method
	Maximum repetition	Minimum repetition	
20	53.1009	39.2701	Proposed method
50	60.5485	45.73327	Fuzzy C-Means

Diagrams of comparing minimum sum of intra cluster distances with maximum amounts of the proposed method and compared method during repetitions have been displayed in fig. (1) and (2) respectively

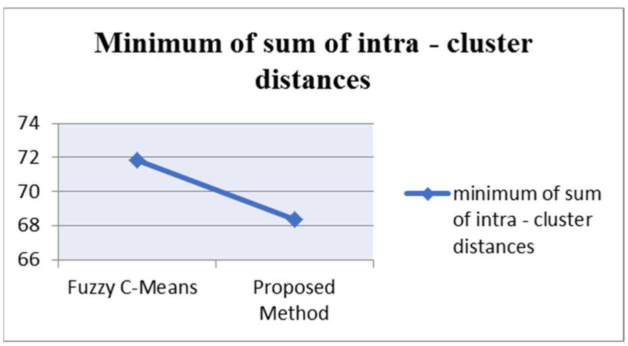

Fig. 1 minimum of sum intra – cluster distances

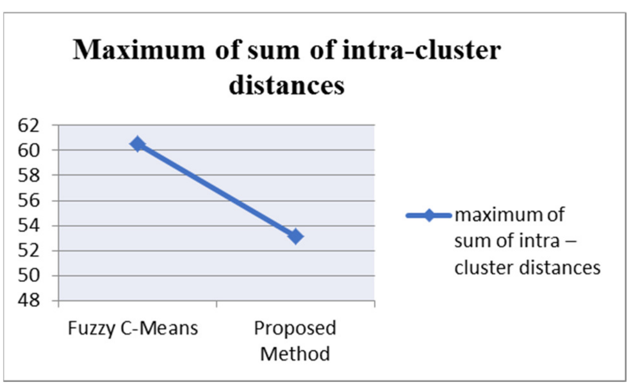

Fig. 2 maximum of sum of intra – cluster distances

According to table (1) and figures (1) and (2) it can be concluded that the proposed method has performed better than the compared algorithm.

6. Conclusion

In this study a clustering based mix method has been used to improve performance of customer relationship management in banking industry. After primary investigation on subject it was clarified that the combined algorithm of the study has not been used to improve customer relationship management and this is innovation of this study. To assimilate mix methods used in this research, MATLAB software has been applied. To evaluate the mix method, it was compared to sum of intra – cluster distances. Considering results revealed that the mix method had yielded more acceptable data mining results and it had performed better than fuzzy c- means. Furthermore it was concluded that people aged 30- 50 with B.A. degree or upper, average income higher than 3 millions, high account stock, less individual under care, had more transactions. Also single ones, owning residency with account stock more than 10 millions.
Lack of unified data set to evaluate results is one of limitations of this study.
Imperialist competitive algorithm views imperialism as one of stages of human
Social – political evolution , modeling this historical phenomenon mathematically to create a powerful algorithm for optimization. Fuzzy clustering is the product of integration of fuzzy approach with clustering to make it more practical and in accordance with real world.
For future researches one can make use of fuzzy approach in inputs. Furthermore , one can apply other evolutionary algorithms to combine with FCM algorithm . Other suggestion is to use more criteria other than those used in this study.

References

[1] M. D. Assunção, R.N. Calheiros, S. Bianchi, M.A. Netto, & R. Buyya, "Big Data computing and clouds: Trends and future directions", Journal of Parallel and Distributed Computing, 79, 2O15, pp. 3-15.
[2] M. Khajvand, & M.J. Tarokh, "Analyzing customer segmentation based on customer value components (Case Study: A Private Bank), 2011.
[3] N. C. Hsieh, "An integrated data mining and behavioral scoring model for analyzing bank customers", Expert systems with applications, 27(4), 2004, pp. 623-633.
[4] A. Garrido-Moreno, & A. Padilla-Meléndez, "Analyzing the impact of knowledge management on CRM success: The mediating effects of organizational factors", International Journal of Information Management, 31(5), 2011, pp. 437-444.

[5] A. M. Hormozi, & S. Giles, "Data mining: A competitive weapon for banking and retail industries", Information systems management, 21(2), 2004, pp. 62-71.

[6] S. Balaji, & S.K. Srivatsa, "Customer segmentation for decision support using clustering and association rule based approaches", International Journal of Computer Science & Engineering Technology, 3(11), 2012, pp. 525-529.

[7] T.F. Bahari, & M.S. Elayidom, "An efficient CRM-data mining framework for the prediction of customer behaviour", Procedia computer science, 46, 2015, pp. 725-731.

[8] M. Al-Ayyoub, M. Al-andoli, Y. Jararweh, M. Smadi, & B. Gupta, "Improving fuzzy C-mean-based community detection in social networks using dynamic parallelism", Computers & Electrical Engineering, 2018.

[9] Y. Tao, Y. Zhang, Y., & Q. Wang, "Fuzzy c-mean clustering-based decomposition with GA optimizer for FSM synthesis targeting to low power", Engineering Applications of Artificial Intelligence, 68, 2018, pp. 40-52.

[10] E. E. Nithila, & S. S. Kumar, " Segmentation of lung nodule in CT data using active contour model and Fuzzy C-mean clustering", Alexandria Engineering Journal, 55(3), 2016, PP. 2583-2588.

[11] B. Etzkorn, "Data Normalization and Standardization", 2011 [cited 2012 Jun 30]. BE BLOG [Internet]. Available from: http://www. benetzkorn. com/2011/11/data-normalization-and-standardization.