Modern Applied Science; Vol. 10, No. 10; 2016
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Customer Segmentation of Bank Based on Discovering of Their Transactional Relation by Using Data Mining Algorithms

Neda Shokrgozar[1] & Farzad Movahedi Sobhani[1]

[1] Department of Management of Information and Technology, Faculty of Management and Economic, Science and Research Branch, Islamic Azad University, Iran

Correspondence: Neda Shokrgozar, Department of Management of Information and Technology, Faculty of Management and Economic, Science and Research Branch, Islamic Azad University, Tehran, Iran. E-mail: neda.shokrgozar@gmail.com

Received: January 15, 2015 Accepted: February 24 2015 Online Published: September 26, 2016

doi:10.5539/mas.v10n10p283 URL: http://dx.doi.org/10.5539/mas.v10n10p283

Abstract

In this research, based on financial transactions between bank customers which extracted from bank's databases we have developed the relational transaction graph and customer's transactional communication network has been created. Furthermore, using data mining algorithms and evaluation parameters in social network concepts lead us for segmenting of bank customers. The main goal in this research is bank customer's segmentation by discovering the transactional relationship between them in order to deliver some specified solutions in benefit of some policy about customers equality in banking system; in other words improvement of customer relationship management to determination of strategies and business risk management are the main concept of this research. By evaluation of Customer segments, banking system will consider more efficient and crucial factors in decision process to estimate more accurate credential of each group of customers and will grant more appropriate types and amount of loan services to them therefore it is expected these solutions will reduce the risk of loan service in banks.

Keywords: customer segmentation, transaction, data mining, social network, graph

1. Introduction

After you have introduced Using bank services and financial transactions by customers in various subsystems of banking, great amounts of data in databases of bank have been generated. These databases can be important source to detect knowledge and behavioral models of customers of bank to take decisions in development of customer relationship management (CRM), determination of strategies and business risk management. Also, banks are helped in policies as selection of effective advertisement method and encouraging systems or presenting important services for each group of customers. One of the most important parameters given to banks or each economic enterprise is increasing awareness and knowledge of bank/ institution to financial issues between customers over time. Following the behavior of customers as independent is one of the limitations of traditional CRM (Domingos & Richardson, 2001). Network –based CRM believes that there is a relationship between customers. Thus, relationship of customers is one of the most important information types and we store databases of information of customers and information of their relations and we consider mutual relationship among the customers (Hill, Provost & Volinsky, 2006). Banking industry is encountered with many risks. As facilities service is one of the most important business sectors of banking industry, loan risk or risk of non-payment has special position in banking. It is associated to the first role of bank in economy, collection of deposit and loan. For example, it is possible that legal or real entities delay in payment of installments or they don't pay the loan or it is possible the profit rate for loans is less than profitability of investment in other fields for bank. The financial institutions and banks should measure the risk of credits to avoid such problem.

Mina Asghari (2008) in a study "Application of data mining to detect scoring model and analysis of customers behavior, applied neural network techniques to classify the customers getting bank loan. After initial processing, modeling was performed. Also, finally, dependence rules techniques were used to detect behavioral models of customers in paying out the loans.

Danuta Zakrzewska (2007) conducted a study based on supervised and unsupervised classification integration. The proposed method allowed the use of different rules for various customers with high risk for bank (Lai et al.,

(2007) in a study proposed an intelligent CRM system and this system applied support vector machine technique to help the managers to avoid risk and effective CRM. Fang,Bin;Ma,Shoufeng (2009) in a study raised the application of data mining technology in CRM.

2. Method

CRM is an organizational approach to perceive significant relationship with customer to keep loyalty of customers (Swift, 2001, p. 12). Implementation of CRM requires moving to customer orientation and definition of market strategy from external organizational view. This trend is called life period of customer and it includes absorption of new customers, increase of customer value and keeping customers (Cotler and Armstrong). Indeed, CRM is a commercial process addressing all aspects of customer features. Customer knowledge is created, relations with customer are formed and their image of products and services is organized and are defined by four elements in a simple framework: Knowledge, goal, sale, service. In another definition, CRM is composed of three important terms (Swift 2001). Management is science and art of planning (prediction); organizing (separation of duties and labor division); guidance and leadership (guiding the subordinate and affecting them); supervision and control and finally coordination to achieve pre-defined goals. Relationship is a set of skills, the most important ones are perceiving the other views and understanding their views. The relationship between customer and organization can be short or long-term, continuous or discrete, frequent or sudden. Customer is final consumer as acting supporter in value creation relations. According to most of organizations, good customers are those with great share in profitability but now we cannot consider profitability the only criterion of customer definition. One of the great goals of CRM in analytical framework is segmentation of customers based on features, behaviors, needs and their value to determine good strategies and presenting suitable services to each class of customers in business. To improve CRM process, customer behavior is evaluated in bank system and as customer behavior in bank system is transactions by them in interval, the collected data of customers is the data of financial transactions by them. Loans are important assets of bank and the majority of revenue of bank is occurred via giving loans but money turnover and capital in society expose the bank to different risks. Credit risk is the one that the other party cannot fulfill the obligations of contract. Regarding the selection of indices and variables in segmentation, the banks try to absorb capital and loan with low risk and the stable deposit of people is used for computation and estimation of their solvency to give loans. Information detection ability with hidden value among the data helps the organization to use them as business knowledge. Data mining by finding the existing models in behaviors and features of customer helps customers segmentation. One of the important issues in customers' segmentation is the feature by which segmentation is performed. These features include demographic, psychological and behavioral or a combination of these four classes (Malhetra, 1993). Customer segmentation based on Customer Lifetime Value (CLV) is one of the new approaches regarding customers' segmentation and is one of the efficient methods in segmentation of customers. Data mining is extraction of valid, unknown, comprehensible and reliable information of great databases and using it in decision making in important commercial activities. In another definition, data mining is referred to semi-automatic process of analysis of big data to find good models. Data mining is searching in databases to find the models among data (Jeffery W. Seifert , December 2004). Simply, database is only storage and recovery of data but data mining is an analysis on the data to extract the rules or make prediction. The data warehouse includes different types of data as all of them are not required in data mining. When the required data were selected and the searching data were determined, we need specific conversion on data. The converted data are searched by data mining operation and techniques to detect the required models. Clustering process as one of the techniques of data mining is classification of heterogeneous population in homogenous type. In this type of classification, no pre-defined index is defined for classification. This method is dividing a heterogeneous group to some sub-groups. Automatically, clustering defines distinction features of sub-groups and sub-groups are formed as the databases are divided into some sections and some groups of records are created showing special attribute. The models are established in the databases and show sudden and valuable information (Ye, 2003). There are various models for clustering, k-means algorithm is used for segmentation of customers in databases with high data volume and high efficiency. In this study, we applied this algorithm

The study population is including 4880867 bank customers of one of banks of Iran. After performing ETL process on data, the data of 494045 customers was transferred to target data basis and the rest of customers didn't perform any valid transaction of intra-bank money transfer and they were excluded from the study. The hypothesis of study showed that type of service providing can be affected by relationship between customers and by recognition of quantitative features of people member of a network of relevant customers, we can predict financial sharing and other information of the same network and apply it in purposeful marketing.

Before paying credit to customers, the banks should consider rules for payment maximum value and these

limitations are determined based on risk and bank capital. For exact investigation, we should determine the sectors risk is focused and suitable decisions should be taken. Based on the increase of demand of loans and existing risk in this bank service, validity of loan recipients is one of the basic principles of risk management in banks and financial institutes. Using risk management tools enables the banks to decide regarding loans. The main goal of this study is segmentation of customers of bank by detection of relationship between them to present suitable solutions for target customers along CRM policies.

The data collection method is observation. To achieve information of customers of bank, the existing data basis in databases in which the data of customers are kept, a list of required data is provided. Also, to identify the effective variables on behavior of customers, free interview with experts of banking was used.

During the study, namely in identification of databases structure and formation of target database and in business complexities perception, the experts of e-banking sector were applied (management sector of databases).

Table 1. The features of bank experts

Experience in bank system (Year)	Education	Age	Position in database management office	No
9	MA	37	MA	1
7	MA	37	MA	2
7	BA	34	MA	3

The features of experts and database administrator of studied bank Table 1.

As the data of study population are investigated in terms of quality and are refined, based on the study model, after ETL process, only on 494045 transactions as final sample was performed.

Table 2. Some records of Transfers Table from target database

Id	TransfPairId	ATJHDAT	cifFrom	cifTo	ExactAmnt	IsSameBranch	TransfDelay
1	184025	02/06/2013	8	2477714	34000000	1	36.38163194
2	184025	04/28/2013	8	2477714	60666664	1	80.9902662
3	184025	06/16/2013	8	2477714	30000000	1	49.15585648
4	100337	02/06/2013	8	4186857	34000000	1	36.38163194
5	100337	04/30/2013	8	4186857	60666664	1	83.20142361
6	100337	06/16/2013	8	4186857	43000000	1	46.94523148
7	96231	02/23/2013	14	4000	650000	1	53.41349537
8	96231	05/27/2013	14	4000	2000000	1	93.04471065
9	89420	02/09/2013	17	2858	410000000	1	39.32247685
10	182350	01/09/2013	26	163	100000000	1	8.500138889

The general framework of target data basis and databases fields is shown in Table 2.

The data of study includes records of transactions of intrabank money transfer of customers. In other words, each record of target database includes a money transfer transaction in which the number of origin customer, number of destination customer, the fee for money transfer, date of money transfer and uniqueness or non-uniqueness of origin and destination branches of transfer . The type of data of this study are continuous numerical.

Table 3. Explanation of important fields of target database

Field	Explanations
TransferPayID	Transfer payment ID
AtjhDAT	Transaction and document date
CifFrom	Customer information from
CifTo	Customer information to
ExactAmount	Amount

IsSameBranch	It is zero or one and it determines whether from or to deposit are performed in this branch of studied bank or they are from two different branches of bank.
TransferDlay	The period from the final transfer from deposit and to deposit in day

In the next stage, we plot the graph of communication symmetrical matrix of customers Number including nodes with label of number of customers and edges with ability of the relationship between nodes (transaction as including a multiplicative of total transactions by the number of transactions in the study). After putting information in database neo4J of communication graph by Gephi software was plotted and general image of graph is shown in Figure 1.

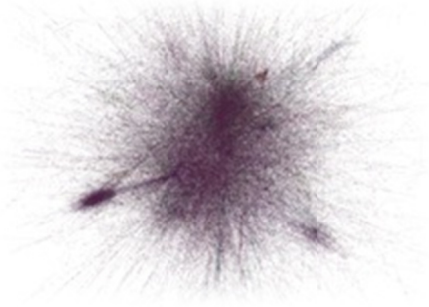

Figure 1. Graph of total communication of customers of bank

This graph includes 90412 nodes and 494045 edges. The important point in Graph Figure 1 is high density of communication (transaction of money transfer at central points) is one of the densest areas of a city in Fars province and it shows a money laundry in this area.

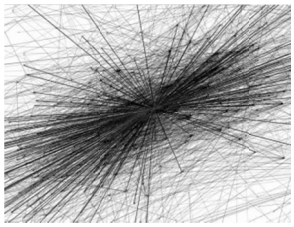

Figure 2. Zooming a part of graph of bank customers' relation in high dense areas. The graph image of relationship of one node with other nodes

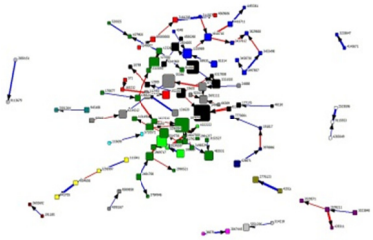

Figure 3. It is regarding financial transaction graph of legal customers of bank

In this stage of study, based on existing indices in social networks analysis and by data mining algorithms, customers' segmentation is performed. Closeness and betweenness centrality indices in complex and big networks including some sub-networks are not suitable indices. These indices are evaluated based on the shortest distance between nodes in network and they don't have good criterion for big networks with some sub-networks For example, if two nodes a, b are inside a big network, inside the network, a node is as it is closer compared to other nodes. The set in which node a is located has weak rank in terms of centrality and the number of existing nodes is low in network. Node b has average distance from other nodes but it is in a bigger set with number of nodes with high centrality degree. The evaluation of close and far distance of both nodes can be similar but node

b has high centrality compared to node a as node b can be available in network by relationship with nodes with high rank centrality. Due to high density of graph and numerous transactions in this study, to evaluate nodes, eigenvector centrality is used. In this index, centrality score of a member in network is affected by centrality score of its neighbor. Indeed, the points with high eigenvector centrality are those with many central neighbors and they have high power (Brandes and Al-Rebach, 2005). Based on extension of graph of bank and high volume of data, eigenvector centrality index is used and the numerical value of this index is ranging 0 to 9.98 for graph nodes of bank. For segmentation of customers based on financial communication graph on clustering communication graph, K-means algorithm is applied.

This algorithm is used due to high volume of data and its efficiency in similar previous studies.

3. Results

Total computed nodes in this clustering are 16389 customers. Indeed, after plotting the graph, some customers with centrality zero or those around the graph were excluded from the clusters.

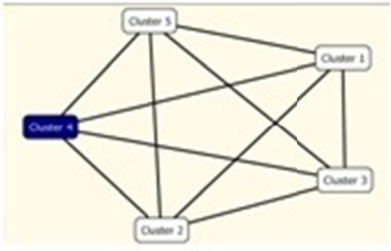

Figure 4. Clusters based on eigenvector centrality index

Table 5. The segmentation of customers result

Cluster Number	Number of customers in cluster	Cluster label	Eigenvector centrality
1	13830	Normal customers	=< 0.54
2	1531	Relatively valid customers	0.54< , =<2.74
3	887	Valid customers	2.74 < , =<3.21
4	129	Highly valid customers	3.21< , <6.91
5	12	Very valid legal customers	>= 6.91

Cluster 5 has only 12 legal companies as insurance companies, non-profit institute, payment companies, stock brokerage and leasing companies. These customers are legal customers of banks and the score of high centrality of them shows their validity.

These customers require much loan from banks and the banks consider that a consistency should be established between validity and price of loans and only the credit of these customers is not considered and risk of macro-loans should be also computed.

Clusters 2, 3 are relatively valid customers cluster. It is proposed that the bank on evaluation on each of clusters, considers the service of giving credit card with suitable price for each group to achieve profitability for bank and value creation for customer with the lower risk. Regarding the customers of cluster 1, 2, we can say as these customers didn't have high centrality and high validity, as high volume of internal transactions is performed and capital exiting from bank is avoided, customers are valuable and marketing policies are applied on these customers mostly that by keeping the loyalty of these customers, capital and money are not exiting from bank.

4. Discussion

After ETL process in the data of money transfer transactions, target database is formed and communication graph among customers is extracted from communication matrix and finally by K-mean algorithm, clustering on graph is applied and the customers are segmented into separated groups with different validity degree. Majority of bank revenue can be occurred via giving loan but based on limitation issue in bank resources, justice in giving loans is an important issues in business of each bank. Thus, the classification or segmentation of customers can lead to the fact that bank can define the priority of customers in receiving different loans and paying out methods and take good and low-risk decision regarding the loan of each class of customers. In addition, by this segmentation, the bank can give new services as INTERNET cartable service to customers. One of the important

issues of managers and policy makers is determining the maximum and minimum transfer of normal and continuous Paya and Santa. Based on the segmentation in this study, the bank can present normal or continuous transfer service to the customers who have continuous transactions to achieve two aims: 1- Via investigation of the relations between customers in the graph and evaluation of business of two nodes, money laundry is prevented, 2- By presenting new internet services, low cost continuous transfer is provided for the bank and customers.

References

Application of BP Neural Network in stock Market prediction (2009). Bin Fang and shoufeng Ma ,School of Management, tainjin University, china

Danuta, Z., & Jan, M. (2005). Clustering Algorithms for Bank Customer Segmentation- IEEE 0-7695-2286-06/05

Hill, Sh., Provost, F., & Volinsky, Ch. (2006). Network-Based Marketing: Identifying Likely Adopters via Consumer Networks. *Institute of Mathematical Statistics, 21*(2), 256-276.

Jeffery, W. S. (2004). Analyst in information science and Technology Policy, 'Data Mining An Overview' December 2004.

Lai, K. K., et al. (2007). An intelligent CRM system for identifying high-risk.

Malhotra, N. K. (1993). Marketing Research- an applied orientation, Prentice-Hall,International, Inc.

Philip, K. (2011). Garyarmstrong principles marketing 14th edition.

Richardson, M., & Domingos, P. (2002). Mining Knowledge-Sharing Sites for Viral Marketing. International Conference on Knowledge Discovery and Data Mining. (Proceedings of the eigth ACM SIGKDD international conference on knowledge discovery and data mining), pp. 61-70.

Swift, R. S. (2001). Accelarating customer relationships: Using CRM and relationship technologies. Upper saddle river. N.J.: Prentice Hall PTR.

Ye, N. (2003). The Handbook of Data Mining, Mahwah, NJ/London: LawrenceErlbaum Associates, Publishers.

Copyrights