World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 9, No. 1, 1-6, 2019

Customer Segmentation Based on GRFM: Case Study

Sahar Ghoreishi
Industrial Engineering Department K.N.Toosi University of Technology Tehran, Iran

Keyvan Khandestani
Information Technology Department, Electronic Branch, Islamic Azad University, Tehran, Iran

Abstract— in the last decades' firms which have directly or indirectly contact with a customer migrate from product-oriented to be a customer-oriented, hence, some products and customers are not profitable in the same way and some of them bring detriment to the firm. In this regard, firms should recognize loyal, profitable and potential customers with a glance of impressive product which brings added value for them. In order to distinguish profitable customers, they supposed to cluster customers and study their behavior's group for the sake of having the best investment in the best segment. In this paper, we utilize customize GRFM (Group RFM) to cluster customers based on proposed APC (account-pattern constraint clustering) algorithm. Hence, we calculate the cluster RFM value which could aid the bank to explore both profitable accounts and customers.

Keywords-Component; Data Mining; Constraint Clustering Algorithms; Segmentation; RFM.

I. Introduction

Over the past few decades, there have been lots of changes in marketing and commerce. What is particularly important is the ability of organizations to adapt to these changes. In today's world, electronic services are increasingly expanding. Many retailers are trying to create online stores [22]. However, the existence of significant technological innovations and the potential for more power to locate networks of online stores in this century created the e-commerce market [21], gaining new customers in the competitive market has become difficult and the profits from loyal customers have grown throughout business communications. Companies have shifted their marketing focus from pure satisfaction to fostering loyalty [23]. The backwardness of a business from the scope of such competition only leads to not overcoming competitors, losing market share, and in general, the loss of business in the true sense. Due to the growth of industries and businesses, over time, product-centric sales have turned to customer-centric. The reason is they have concluded that mass the production of goods and services is no longer a way of profit-making, and products or services producing it according to the needs of customers and at the same time recognizing its lucrative customers could be an effective relationship with them [1]. In modern banking, with the emergence of competition as well as the growth and diversification of the customers' demands, banks are paying special attention to their customers and trying to communicate, more effectively, with their beneficial customers [2].In this study, we tried to use data mining techniques to find an effective way of identifying profitable products and customers. In contrast past research, this study attempts to find products that have more profitable customers in their group instead of only grouping customers by their financial values. Therefore, in this paper, we use a novel Group RFM (GRFM for short) framework [18] and customized it for the bank customers to identify high loyal and contribution of customers and profitable accounts for the bank; moreover, it distinguishes potential customers for products promotion. Instead of calculating the customers' RFM values on all of the accounts they have, the GRFM calculates customers GRFM values what considers customers' accounts patterns as well as the characteristics of accounts in analyzing customers. The customized GRFM first discovers each of which presents a set of customer's accounts that are frequently in the transactional data set. Then, based on gained patterns, we use a novel account-pattern constraint clustering (APC) algorithm to create groups of accounts pattern to calculate RFM values. In this regard, the bank could rank their account and design of their market strategies based on the results.

II. Related Work

The concept of customer segmentation was developed by an American marketing expert, Wendell R. Smith, in the middle of 1950. It is a technology to cluster customers into groups that share similar characteristics and tend to display similar patterns. Later, the RFM model is first proposed by Hughes in 1994 and it is a model that differentiates important customers from large transaction data [15]. RFM method is very effective attributes for customer segmentation [16]. In general, different models have been proposed by researchers in the area of customer segmentation. In most of these studies, models are different in terms of input variables.

The main inputs for customer segmentation are RFM. Siemens used a SOM neural network to identify customer groups based on the behavior of reimbursement, payback, frequency, and monetary predictions. He also classifies the banks' customers into three major groups of profitable customer groups.

Cheng and Chen also proposed a new method of joining the Quantitative Characteristics of RFM and the K-Means Algorithm in the uneven series theory for extracting the rules of meaning [17]. The data from this electronic case study at Chang Hua Company includes 401 records of corporate exchanges, conducted in 2006. Indicators for customer segmentation were as follows: A: Region B: Country C: Credit value.

In addition, a combination of input variables mentioned above has also been used by researchers. For example, Chan and colleagues presented a new approach that combines customer targeting and customer segmentation for campaign strategies [17]. In this research, customer behavior is identified using an RFM model, then an LTV model is used to evaluate the customers of the proposed segments.

As noted earlier, some writers focused on the process of partitioning from a technical point of view. For example, Lee and colleagues developed a new method for cross-market segmentation. The authors proposed a two-stage approach integrating statistical and data mining techniques. To test the difference between clustering factors in the first stage, using statistical methods (multiple confirmatory factor analysis groups), and the second step is to develop real clusters by a two-level SOM method. Huang and colleagues have used vector support for clustering of marketing segment. Also a research is conducted on customer segmentation based on a two-stage clustering model [20].

Customer profiles for each group can serve as a starting point for managers to determine marketing strategies for the bank to provide services; however, clustering customers in to different groups does not show the most profitable products. Consequently, marketers could not precisely investigate to develop profitable products and services in order to bring more advantage for the firm. Our approach to diminish the above problems is to consider the characteristics of the accounts owned by customer. The RFM value measuring methods all adopt a single criterion to measure the RFM value of a customer, no matter what kinds of products were purchased. However, the characteristics and lifetimes of the purchased products are not always the same, grouping customers in this way cannot provide precise quantitative prediction [18].

Although, customer segmentation based on RFM values could conduct managers to find most profitable group of customers, it fails to address the effective products promotion. In bank industry, accounts known as a products. In this regard, some accounts, in the Post Bank, are known as low-cost products that assumed could bring more value for the bank since it does not cost for the bank (lottery or interest rate). Consequently, the proposed model is about finding the most profitable products and customers by applying the account-pattern constraint clustering (APC) algorithm and RFM method.

III. RESEARCH METHODS

The research based on the type of target is a part of descriptive-exploratory and applied research. The case study is the Post Bank of Iran. The collection of data is the records of customers transactions stored in the database of the Post Bank of Iran.

In terms of time, the present study is a cross-sectional study because it examines customers in a quarterly period of the first three months of the year. This section first presents the concepts of the constraint-based clustering. Then, explains about RFM measurement (TABLE I). Finally, describes method of setting accounts pattern for each customer.

TABLE I. RFM ATTRIBUTES

RFM attributes	RFM attributes Original Data
Recency	The interval between the last transaction date and last date of tracing time
Frequency	Frequency of transaction during the traced time
Monetary	Monetary or transaction (negative form shows withdrawal)

A. Constraint clustering

While it is possible that a fully unsupervised clustering algorithm might naturally a solution that is consistent with the domain knowledge, the most interesting cases are those in which the domain knowledge suggests that the default solution is not the one that is sought. Therefore, researchers began exploring principled methods of enforcing desirable clustering properties [24]. The reason is they provide flexibility to attach user-specified constraints while clustering [18]. In general, the constraints can be classified into the following two categories:

Must link (ML): Let M be the set of must-link pairs; then (xi,xj) 2 M implies the instances xi and xj must be assigned to the same cluster.

Cannot link (CL): Let C be the set of cannot-link pairs; then (xi,xj) 2 M implies the instances xi and xj should be assigned to a different cluster.

In fact, different constraints should have different penalty weights. However, the difference is not easy to be identified. Moreover, pair-wise constraint clustering cannot be used when the constraints focus on partial characteristics between the pairs [18].

B. RFM Method

RFM models have been used in direct marketing for more than 30 years. Given the low response rates in this industry (typically 2% or less), these models were developed to target marketing programs (e.g., Direct mail) at specific customers with the objective to improve response rates. Prior to these models, companies typically used demographic profiles of customers for targeting purposes. At this level, quality of data is acceptable as a transactional data in banks which couldn’t have outliers and errors. RFM attributes are shown and explained(TABLE I).

Min-max normalization method is used in this phase (TABLEII). This method performs a linear transformation of the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute A. Then min-max normalization maps a value v, of A to v’ in the range of $[newmin_A, newmax_A]$ by computing in Eq. (1) [19].

$$v' = \frac{v - min_A}{max_A - min_A}(newmax_A - newmin_A) + newmin_A \quad (1)$$

TABLE II. NORMALIZATION OF VALUES

	M	F	R
M	63/79	14/17	9/12
F	9/79	2/17	2/12
R	7/79	1/17	1/12

C. *Accounts pattern recognition*

The difference of the bank with other institution is about the low and specific number of its product, and new products may enter the banking industry in many years. Thanks to the nature of bank accounts, it is a logical way to promote specific accounts that are more profitable. Iranian banks have 5 types of accounts (shown in Table III) and each customer in the bank could have different accounts. As explained above, there is an assumption about accounts value for the bank (shown in Table III). In order to set accounts pattern binary value, we assign 5 bit base on account value and if the customer has the account, set 1, and if not, set 0. As explanation 5 mean most valuable and 1 means least valuable for the bank.

TABLE III. ACCOUNTS INDEX

Value index	Accounts Name
5	Current accounts payable by card only
4	Savings Account
3	Current account credit
2	Short-term savings account
1	Long-term savings account (six months, nine months, one year, two years and five years)

For example, a customer with only long-term savings account and a long-term savings account in the bank has a quantity value of 17 based on Eq. (2), which is equivalent to a binary number of 10001. As a result, the bigger a quantity value indicates the customer has a combination of products that assumed to be more profitable for the bank.

IV. ENVIRONMENT AND STATISTICAL SAMPLE

As stated above, the statistical society is the research of the information obtained from the Post Bank. The bank has two core banking systems due to its branches and supervision of private offices, which have about 450 branches of 15,000 offices in the country. In this study, only three months have been chosen since the transactions of customer accounts related to larger intervals are bulky and sometimes disrupt data mining operations.

V. RESEARCH MODEL

At this stage, a conceptual model is presented in the following way with library studies and the study of the factors in the research resources. According to a review that has been carried out on research literature, the following model (Fig1) can be a conceptual model of customer clustering using data mining.

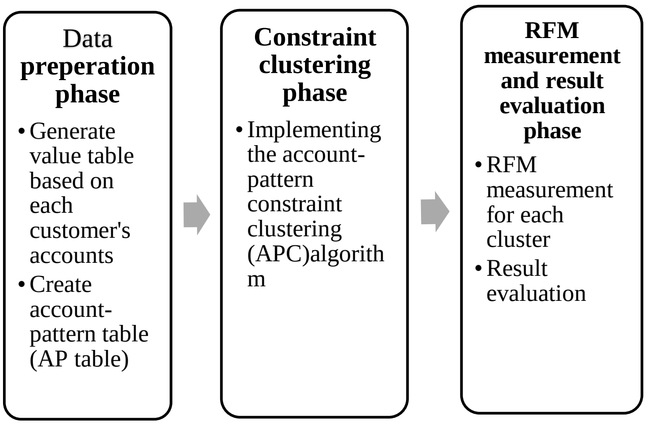

Figure 1.Research Model

$$\text{Quantity value of account'spattern} = \sum_{k=0}^{n} 2^k$$

n is customers' account (2)

A. *Data Preparation phase*

1) *Tables selection*

In this research, 10 tables of central bank system of branches were used. The results of these tables are used after applying different scripts.

Tables contain customers' accounts information. This table is extracted in the header of each column of the account type, and if the client has an account or product in the bank, one number will be placed in the corresponding field and if it does not have the account the number is zero in the desired field.

2) *Clearing the data*

Due to the fact that the information extracted from the databases is very accurate bank information, and the mistake and error do not occur, we were sure of the accuracy of the information and there was no need to clear the data.

3) *Implementing the AP table*

The algorithm in this phase is illustrated in TABLE IV. After examining customer accounts dataset, the output of employing, according to TABLE V, the algorithm output is twelve accounts patterns. As described before, the customer account-pattern table (AP table) is generated base on Eq. (2) and (3); the results are presented below (TABLE V).

TABLE IV. CREATION OF ACCOUNT-PATTERN TABLE (AP TABLE)

Data preparation and creation of account-pattern table (AP table)

Input: Customers' account dataset

Output: Value table, account-pattern table (AP table)

1. *Index accounts based on their cost for the bank*
2. *Classifying customers based on their accounts*
3. *do*
4. *{Set 2^i for each customer's account based on the type of account*
5. *inserting (or updating) value table records for each customer }*
6. *insert into account-pattern table the distinct binary value of value table*
7. *sort descending AP by its binary(quantity) value*
8. *end;*

TABLE V. ACCOUNT-PATTERN TABLE (AP TABLE)

Binary Value	Quantity Value
11100	28
11010	26
11000	24
10100	20
10011	18
10000	16
01100	12
01010	10
01000	8
00100	4
00010	2
00001	1

B. Constraint clustering phase

As described in the previous sections, the account-pattern constraint clustering (APC) algorithm performs the grouping of customers based on their accounts, as shown in TABLE VI. The difference of the algorithm to K-mean constraint clustering does not have initial class centers, and the cluster centers would be defined through the algorithm process by dislike function. The clustering result could satisfy the particular expectation. When a constraint triggers the related data records, the APC algorithm starts to adjust the corresponding binary value of AP table by using the dislike function into clusters. It then uses the (DV (i), DV ($i-1$)) function to measure the distance between two quantity records in order to decide whether they should be allocated into the same cluster or not. Eq. (5) defines the dissimilarity function. If the function value of Dislike (DV (i), DV ($i-1$)) is less than a predefined threshold, then the cluster i is a candidate cluster, otherwise record i is not clustered to Ci.

$$\mathrm{DV(i)} = \sum_{n=0}^{5} b_n 2^n \qquad (3)$$

$$Result\ (i) = AP\ (i) \land AP\ (i+1) \qquad (4)$$

$$Dislike\ (DV\ (i), DV\ (i-1)) = (i) - DV\ (i-1) \qquad (5)$$

TABLE VI. CONSTRAINT CLUSTERING PHASE: ACCOUNT PATTERN CONSTRAINT CLUSTERING (APC)

Account pattern constraint clustering (APC)

Input: Account pattern table (AP table)

Output: Clusters

1. *Sort AP table base on descending binary value*
2. *For i=0 to AP count*
3. *Fetch AP(i) from AP table*
4. *Set C(i)=AP(i)*
5. *Add C(i) as a candidate cluster*
6. *Fetch AP(i+1)*
7. *Set Result(i)=AP(i) ⋀ AP(i+1)*
8. *If result(i-1)=0*
9. *{set result(i-1)= result (i)}*
10. *If $Dislike$(DV(result (i)) , DV(result (i − 1))) = 0*
11. *{add AP(i+1) into C(i) cluster}*
12. *Else*
13. *Add AP(i) as a new cluster center*
14. *Next*

TABLE VII. SUMMARY OF THE NOTATIONS UTILIZED IN THIS PAPER

n	is a count of bits
b	is the digit
AP (i)	the ith record of AP table
C (i)	the ith cluster
Result (i)	the ith result of $(i) \wedge (i+1)$
DV (i)	the function of calculation of quantity value of binary data

$Dislike\ (DV\ (i), DV\ (i-1))$ the function is used to calculate ( ) $-$ $(i-1)$

As explained, the grouping of the AP table base on APC algorithm is presented with the results listed in Fig2.

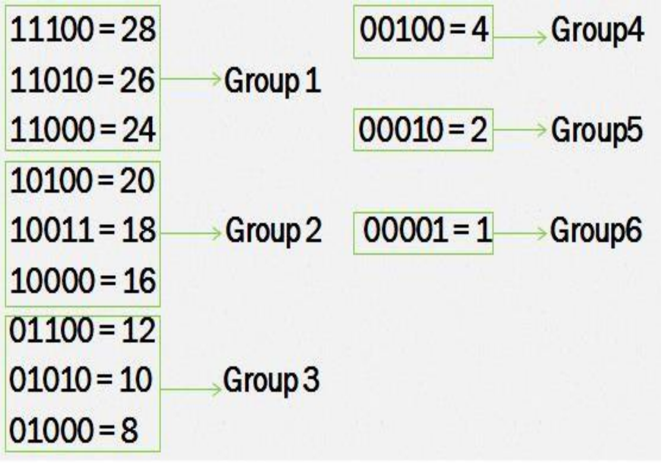

Figure 2. Account pattern Clusters

C. RFM measurement and result evaluation phase

In this section, in order to calculate the RFM, each of the values of R, F, and M must be multiplied by their weight in accordance with Eq. (6) Based on AHP. The weights of the AHP technique are used here to calculate the weights. In tables 4 and 5, weights are calculated.

$$CLVci = NRci \times WRci + NFci \times WFci + NMci \times WMci \quad (6)$$

1) AHP approach

The AHP is used to determine the relative importance (weights) of the RFM variables, WR, WF, and WM, respectively. The three main steps of the AHP are as follows.

Step1: Perform pair wise comparisons.

Step2: Assess the consistency of pair wise judgment.

Step3: Computing the relative weights.

In this study, the following three groups of evaluators judge the RFM weightings: (a) two administrative managers, (b) one business managers in marketing, and one marketing consultant and (c) ten customers who have made at least one account (TABLE VIII).

TABLE VIII. WEIGHING ACCORDING TO EXPERTS

WR	WF	WM
0. 076921	0. 132746	0. 790333

TABLE IX. WEIGHTED R, F, M

	M	F	R
M	1	7	9
F	1/7	1	2
R	1/9	½	1

As described in the previous sections, based on the presented model, we first clear the data, and then, based on the proposed model, we give a binary number to the customer accounts in the bank, then using the APC algorithm to grouping the provided clients of the bank based on their accounts. Then, RFM is calculated for each group of customers. By calculating RFM, it is determined which cluster of customers is with their RFM products. As a result, products that are more profitable for their customers are characterized by products that have higher RFMs, and then results can be used in relation to marketing. Finally, the results obtained are based on the proposed model is shown below (TABLE X)

TABLE X. GROUPING RESULTS ARE BASED ON THE PROPOSED ALGORITHM AND THE CALCULATION OF RFM

RFM	Account pattern's quantity value	Clusters
0.7965	28,26,24	C1
0.0546	20,18,16	C2
0.7961	12,10,8	C3
0.9823	4	C4
0.6845	2	C5
0	1	C6

2) Result Evaluation

According to RFM values of clusters (TABLE X), the study found accounts which were considered as valuable account for the bank were actually less profitable than the rest of the accounts. The reason is RFM values of clusters. Furthermore, as shown in TABLE X, cluster4 has the biggest amount of RFM and customers of the cluster are who only have Short-term saving accounts. In this regard, against the bank's managers assumption Short-term savings account should not be index as low-benefit account. On this subject, we explore new account index based on RFM values of clusters that as shown below (TABLE XI).

TABLE XI. NEW ACCOUNT INDEXES

Index(cost)	Account's Name
3	Current accounts payable by card only
4	Savings Account
2	Current account credit
5	Short-term savings account
1	Long-term savings account (six months, nine months, one year, two years and five years)

As shown above, the saving account and the Long-term savings account indexes do not changed; however, thanks to our output analysis other accounts value for the bank is changed.

VI. CONCLUSION

Customized GRFM applies APC clustering algorithm to profoundly analyze and utilize the RFM value of the customer. It aids cluster analysis from both aspects of customers account and their transactions. While the analysis takes into accounts transactions, the (R, F, M) values could reveal the true CLV value of customers. Customized GRFM is the same as the traditional RFM analysis in the sense that each cluster has the same loyalty and contribution. The customized GRFM difference is it allows the bank to evaluate their account values, and the account could belong to different clusters. Thus, it could associate with different contributions with respect to different characteristics of accounts. This difference allows GRFM to address marketing section to planning on account promotion. Furthermore, customized GRFM provides a clustering method that could use customers' transactions to distinguish the type of accounts they have and calculating RFM value. Consequently, values of account patterns promptly respond to the market-oriented demands. It converts the bank dataset into corresponding account-pattern table (AP Table), which can then be quickly and conveniently adjusted to cluster by APC algorithm.

REFERENCES

[1] Alizadeh, Somayeh, Malek Mohammadi, Samira. (2012) "Data mining and discovery of step by step with Clementine software (Sixth Edition)", Khaje Nasir al-Din Tusi University of Technology.

[2] Minaei, Behrooz, Nasiri, Mehdi, Hosni, Daniel, Ida, Abraham. (2012)" Step by step data mining with Clementine", Engineering & Research Group of Saher.

[3] Raymond Wong, Canon Ton and Anthony Wong, (2014), "Examine the Effects of Customer Satisfaction on Customer Loyalty: An Empirical Study in the Healthcare Insurance Industry in Hong Kong".4(3), pp.372-399

[4] Sanjay Puligadda, WilliamT.Ross, JinjieChen, ElizabethHowlett, (2012), "When loyalties clash purchase behavior when a preferred brand is stocked out: The tradeoff between brand and store loyalty". 19(6), pp.570-577

[5] Anne-Françoise Audrain-Pontevia, Gilles N'Goala, Ingrid Poncin, (2013), "A good deal online: The Impacts of acquisition and transaction value on E-satisfaction and E-loyalty". 20(5), pp.445–452

[6] Chih-Yuan Chang, (2010), "Exploring customers' store loyalty using the means-end chain approach".17(5), pp.395-405

[7] Tung Tsai, Hsien & Chiang Huang, Heng, (2009), "Online consumer loyalty: Why e-tailors should seek a high profile leadership position", Computer in Human Behavior, 25, pp. 1231-1240.

[8] Casalo, Lvis & Flavian, Carlos & Guinaliu, Miguel, (2008), "The role of perceived usability, reputation, satisfaction & consumer familiarity on the website loyalty formation process", Computer in Human Behavior, 24, pp.325-345.

[9] Thomas A. Stewar (1995), "After All you've done for Your Customers, Why Are They Still Not Happy?" Fortune, pp. 178-182.

[10] Philip Sugai, Donghum Kim (2008), "Difference in consumer loyalty & willingness to pay for service attributes across digital channels: A study of the Japanese digital content market", Telecommunication Policy, 32, pp.480-489.

[11] Mo Koo, Dong, (2006), "The fundamental reason of e-consumers' loyalty to an online store", Electronic Commerce Research & Applications, 5, pp.117-130.

[12] Vladimir Vapnik, S.Kotz (2006) "Estimation of Dependences Based on Empirical Data"

[13] Colin Campbell and Yiming Ying (2011),"Learning with Support Vector Machines, Morgan and Claypool".

[14] Xiaohui Hu, Haolan Zhang, Xiaosheng Wu1, Jianlin Chen1,Yu Xiao1, Yun Xue1 , Tiechen Li1, Hongya Zhao,"A New Customer Segmentation Framework Based on Biclustering Analysis",9(6),pp.1359-1366.

[15] Hughes, A. (1994). "Strategic database marketing. Chicago: Probus Publishing Company".

[16] Newell, F. (1997). "The new rules of marketing: How to use one-to-one relationship marketing to be the leader in your industry", New York: McGraw-Hill.

[17] Cheng, C.-H., & Chen, Y.-S. (2009), "Classifying the segmentation of customer value via RFM model and RS theory. Expert Systems with Applications".36(3),pp. 4176-4184

[18] Hui-Chu Chang, Hsiao-Ping Tsai (2011)," Group RFM analysis as a novel framework to discover better customer consumption behavior",38,pp.1499-14513

[19] Sahar Ghoreishi,M.J Tarokh(2015) ,"Evaluation and segmentation based on customer lifetime value: case study",7,pp.39-50

[20] Morteza Namvar, Mohammad R. Gholamian, Sahand KhakAbi (2010),"A Two Phase Clustering Method for Intelligent Customer Segmentation", International Conference on Intelligent Systems, Modelling and Simulation

[21] Wan-ILee,Chih-YuanChang,Yu-LunLiu, (2010), "Exploring customers' store loyalty using the means-end chain approach".17(5),pp.395-405

[22] Tung Tsai, Hsien & Chiang uang, Heng, (2009), "Online consumer loyalty: Why e-tailors should seek a high profile leadership position", Computer in Human Behavior, 25, pp. 1231-1240.

[23] Casalo, Lvis & Flavian, Carlos & Guinaliu, Miguel, (2008), "The role of perceived usability, reputation, satisfaction & consumer familiarity" 24, pp. 325-345.

[24] Sugato Basu, Ian Davidson, and Kiri L. Wagsta (2007) "CONSTRAINED CLUSTERING: Advances in Algorithms, Theory,and Applications".

AUTHORS PROFILE

Sahar Ghoreishi, University alumni

Keyvan Khandestani, University alumni