Data mining using RFM Analysis

Divya D. Nimbalkar, Asst Prof. Paulami Shah

Abstract— The competitive world of today demands for having good marketing policies to attract the customers as well as retain the old customers .Organizations hence use strategies that would give the best customer satisfaction and which will return all their investments in their products with profit . It therefore becomes necessary to classify different segments of customers so that one can provide them with a personalized service . This paper serves this purpose of managing customer relationship by performing customer segmentation using RFM analysis then performing clustering on the values obtained from this analysis and then classification to get the applicable rules for each customer segment obtained after clustering.

Index Terms— association rule mining,clustering, classification, customer relationship management, customer value, FAPH ,K-means Algorithm , RFM model , Rough Set Theory , TOPSIS .

1 INTRODUCTION

The large amount of information present in organizations if used correctly can help generating important patterns and trends. These patterns provide useful insights of customer buying patterns.

But since this data is very complex it becomes even more difficult to know the customer needs, improve their satisfaction and further lead to their retention. Customer relationship management deals with this recognition and retention of potential customers.

The techniques of customer value analysis give details of only the future buying patterns of customer from their past purchasing records. But to know the customers value in future a new technique of RFM analysis is used . This techniques uses the three parameters of recency , frequency and monetary value to get the customer loyalty value which is further used in the clustering phase to cluster similar customers and finally develop rules for clusters generated in the previous step. The details of the terms customer relationship management, customer value analysis, RFM analysis are as follows:

1.1 Customer Relationship Management (CRM)

CRM is devoted to improve relationship with customers, it focuses on how to integrate customer value , requirements , expectations and behaviors by analyzing data from customer transactions .CRM is hence the manner in which both existing and new customers are retained in the organization. Enterprises apply some methods to effectively enhance customer relationships which include customer relationship management, customer value analysis , enterprise strategy and positive service mechanisms . An effective CRM is the one that has this capability of having both new and retaining old customers.

1.2 Customer Value Analysis

Customer value analysis is a kind of analytical tool which identifies customer future buying patterns from the existing patterns in large databases. The enterprises apply these methods to know the target customers whose contribution is outstanding. The RFM model is one of the well known customer value analysis method which extracts characteristics of customers using fewer criterions as cluster attributes so as to reduce the complexity of the model.

1.3 RFM Model

RFM analytical model differentiates important customers from large database by three variables as interval of customer consumption, frequency and the amount. The RFM model hence considers Recency, Frequency and Monetary value as the three criteria for getting loyalty value of customers. The detailed definition is as follows:

Recency of last purchase (R) : It represents the interval between the latest buying and the present time of a customer.The lower the interval the higher is the recency value of that customer .

Frequency of purchase (F) : It represents the number of times a customer buys within a particular period like once in a month , thrice in a year and likewise.The higher the value number of transactions in an interval the higher is the value of F.

Monetary value of purchase (M) : It represents the monetary value of the purchases in a particular time interval . Higher the monetary value more is the value of M.

The Table.1 below shows a dataset of customer transactions that are analyzed using RFM analysis with values them being initially in days , number and amount respectively [4].

TABLE 1
AN EXAMPLE OF DATASET : CUSTOMER TRANSACTION [4]

CustomerID	Recency (Day)	Frequency (Number)	Monetary (TL)
1	3	6	540
2	6	10	940
3	45	1	30
4	21	2	64
5	14	4	169
6	32	2	55
7	5	3	130
8	50	1	950
9	33	15	2430
10	10	5	190
11	5	8	840
12	1	9	1410
13	24	3	54
14	17	2	44
15	4	1	32

In the RFM analysis customer segmentation is done by first sorting customers based on their Recency value that is the most recent will be at the top ,then with the Frequency value with the most frequent at the top and finally the monetary valu with the highest monetary value at the top . The customers are therefore split into five quintiles with the top 20% having the score 5 the next 20% having score 4 and so on. The process is repeated for all the three criterias and finally the values are merged to get every individual customers rank . The Table 2 given below depicts the customers are assigned to different quintiles and how their values can be merged to get customer values [4].

TABLE 2
CUSTOMER QUINTILES AND RFM VALUES OF CUSTOMERS [4]

CID	Rec.	R	CID	Freq.	F	CID	Mon.	M	CID	RFM
12	1	5	9	15	5	9	2430	5	1	544
1	3	5	2	10	5	12	1410	5	2	454
15	4	5	12	9	5	8	950	5	3	111
7	5	4	11	8	4	2	940	4	4	222
11	5	4	1	6	4	11	840	4	5	333
2	6	4	10	5	4	1	540	4	6	222
10	10	3	5	4	3	10	190	3	7	433
5	14	3	7	3	3	5	169	3	8	115
14	17	3	13	3	3	7	130	3	9	155
4	21	2	14	2	2	4	64	2	10	343
13	24	2	4	2	2	6	55	2	11	444
6	32	2	6	2	2	13	54	2	12	555
9	33	1	15	1	1	14	44	1	13	232
3	45	1	3	1	1	15	32	1	14	321
8	50	1	8	1	1	3	30	1	15	511

1.4 Clustering

Clustering is the process of grouping similar objects.There we use various clustering methods like partitioning, hierarchical clustering , density based , grid based etc . The main motive of clustering here is to group together customers with similar buying patterns than those who are different from them . The resulting clusters should have minimum dissimilarity within the cluster and maximum dissimilarity with other clusters.

1.5 Classification

Classification algorithms are used to derive rules from the clustered results obtained in the clustering phase . The rules are useful for identifying each and every customer from his buying patterns.There are various techniques for classification like decision tree technique , neural network etc.

2 METHOD PROPOSED IN DIFFERENT PAPERS

The entire process of RFM Analysis is depicted in fig 1 . The Sections below describe them in detail.

2.1 Step1 : Data Preprocessing

In this step the data is preprocessed to remove missing , incorrect values , normalization , discretization of attribute values are also done here then removal of unwanted attributes , concept hierarchy generation like converting city value to state is done here in [1][2][3][4].

2.2 Step2 : RFM Analysis

Approach 1 : It is carried out using following steps [1][3][4]:

- Arrange the three attributes in either ascending or descending order
- Club the RFM attributes into 5 equal parts where each part is equal to 20% of all. The score 5 is assigned to most contributing one then 4 to next highest contributing one and so on till 1 as shown in Table 2.
- This process is repeated with all three criterias RFM to get the final RFM score.

Approach2 [2][6]: Weighted RFM is used in recent techniques where rather than assigning equal weights to all R,F and M values of 1:1:1 different weights are assigned to them. This is useful in situations where the organizations focus more on one parameter over the other.This is generally a problem of multiple criteria decision making ie. to decide among multiple options and is solved using Fuzzy analytical hierarchical processing(FAPH).This process gets you the weights of criterias R,F and M which are multiplies to their respective values to finally get the customer loyalty value.

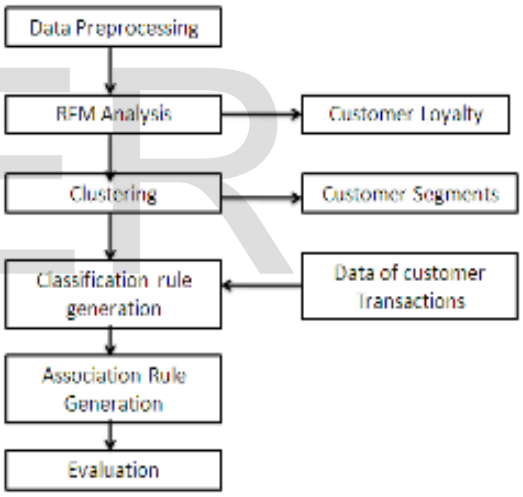

Fig.1 Basic Steps in Datamining using RFM Analysis

2.3 Step3 : Clustering

Clustering algorithm used different approaches are different techniques are as follows [1][3]:

Approach 1 : Here clustering of RFM values is done using Kmeans algorithm. The algorithm is as follows :-
a) Partition the items into initial clusters randomly.
b) Now calculate the centroid of each cluster.
c) Compute the distance of each point from the centroids obtained in previous step and reassign the item to the centroid which is closest to it.
d) Repeat b) and c) steps till the cluster results obtained is the same.

Approach 2 : The results from Kmeans algorithm is largely de-

pendent on the initial centre selection so the results are accurate only if the centre selection is correct. Kmeans++ suggests an improved way of selecting centres by calculating the initial centres for the clusters [4]. It calculates initial centres by calculating their squared distance from the closest centre already chosen. Therefore one cen get consistent result using it.

Approach 3 : Rather than going for centroid selection directly one's result will be more accurate if one can know how many clusters are actually formed after clustering is actually done and then perform kmeans to actually get the clusters [3] . It will be helpful in removing outliers if any from the clusters . So a two step way for clustering ie by performing Hierarchical Agglomerative Clustering first to determine the optimal number of clusters and then apllying Kmeans algorithm will give better results .

2.4 Step4 : Classification

Approach 1 : The rules of classification are discovered by using the clustering results obtained in the previous step. In one approach it is carried out using C4.5 decision tree algorithm where in entropy and gain of each attribute is calculated and the one with the highest gain becomes the root node and the others become its child [4]. The process is repeated till the time all attributes value pairs have been considered and leaf nodes are reached.

Approach 2 : It just uses another variation of decision tree ie. C5 algorithm to determine the classification rules[3] . The rules returned have great use of demographic variables like age,gender etc like c4.5 algorithm to get the final rules.

Approach 3 : Since decision trees have large number of instances to be handled it becomes difficult to build and achieve the desired accuracy. So another technique of generating rules is proposed by using LEM2 algorithm [1] to classify data with vague and imprecise values. It can handle data of the form –

TABLE 3
SAMPLE RESULT AFTER CLUSTERING

Cid	Recency	Frequency	Monetary	loyalty
10	High	Very High	High	High
11	Low	Very High	Medium	Medium
12	Very Low	Very High	Low	Low
13	Very Low	Very High	Low	VeryLow

As shown in Table 3. above Cid 12 and 13 have the same values for RFM but their loyalty values are different LEM2 has the capability to handle such values unlike decision tree which cannot handle those.

Approach 4 : In weighted RFM there is a step to rank each cluster obtained after clustering to know and decide the strategic plan for each cluster.One of the proposed techniques in paper [2] is TOPSIS ie. Technique Of Order Preference by Similarity To Ideal Solution .

2.5 Step5 : Generating Association Rules [4]

Association rules are generated by using Frequent pattern tree or by using Apriori algorithm by setting some support and checking if a particular pattern is appearing equal to or more than that value if yes then its saved as an association rule .

2.6 Step6 : Evaluation Of Results [1][4]

It is generally carried out by comparing accuracy of results from different classification techniques like Naive Bayes , Decision tree, Neural Network etc on the same set of data. Also one can use different number of classes on the output to see if the accuracy varies by how much amount in case of clustering by using Kmeans.

Some other techniques propose use of n fold cross validation technique to evaluate the results of classification and criteria called Lift & Loevinger to evaluate the association rules accuracy.

3 INFERENCES :

TABLE 4
INFERENCES OBTAINED FROM THE STUDY

Criteria for comparison	Data Mining Using RFM Analysis	Classifying the segmentation of customer value via RFM model and RS theory	Developing a model for measuring customer's loyalty and value with RFM technique and clustering algorithms	A Hybrid Model of Data Mining and MCDM Methods for Estimating Customer Lifetime Value
RFM Analysis	Defining the scaling of RFM attribute for every customer	Same procedure	Same procedure	Same procedure
RFM Normalization	Not present	Not present	Present using min-max algorithm	Present using min-max algorithm
Clustering Algorithm	K-means++	K-means	Two Step	K-means
RFM weight ratio	1:1:1	1:1:1	1:1:1	Here it is done using AHP
Classification Rule generation	C4.5	Rough set theory	C5 algorithm	TOPSIS ie. Technique for Order of Preference by Similarity to Ideal Solution method of MCDM (Multiple criteria decision making)
Clustering efficiency	It is better than K-means because the initial centroid selection is done using squared distance	It is greatly dependent on the initial centroid selection and hence not as good as K-means++	It uses Hierarchical Agglomerative clustering first then K-means to reduce the error caused due to initial centroid selection	Same as paper 1 .

The Table 4.above depicts the inferences obtained from various approaches studied in RFM Analysis . Here each individual row represents the criteria on which the techniques are compared and the columns heads give the names of the

different papers suggesting the different strategies.

The preferred spelling of the word "acknowledgment" in American English is without an "e" after the "g." Use the singular heading even if you have many acknowledgments. Avoid expressions such as "One of us (S.B.A.) would like to thank ... ." Instead, write "F. A. Author thanks ... ." Sponsor and financial support acknowledgments are included in the acknowledgment section. For example: This work was supported in part by the US Department of Commerce under Grant BS123456 (sponsor and financial support acknowledgment goes here). Researchers that contributed information or assistance to the article should also be acknowledged in this section.

4 Conclusion

The use of RFM analysis helps improving customer relationship even better than by directly going for data mining since it incorporates customer demographic variables as well in getting the results . The use of Weighted RFM gives a further enhancement to the technique of RFM analysis by providing organizations to decide their own priority of one factor over the other . Thus in the competing world of today RFM analysis helps organizations to better attain their goals of profit and customer relationship.

5 References

[1] Ching-Hsue Cheng, You-Shyang Chen,"Classifying the segmentation of customer value via RFM model and RS theory", Expert Systems with Applications 36 (2009) 4176–4184, Department of Information Management, National Yunlin University of Science and Technology, 123, Section 3, University Road, Touliu, Yunlin 640, Taiwan,2009 .

[2] Amir Hossein Azadnia, Pezhman Ghadimi, Mohammad Molani- Aghdam, '"A Hybrid Model of Data Mining and MCDM Methods for Estimating Customer Lifetime Value" , Proceedings of the 41st International Conference on Computers & Industrial Engineering, Department of Engineering, Ayatollah Amoli branch, Islamic Azad University, Amol, Iran ,Department of Manufacturing & Industrial Engineering, Universiti Teknologi Malaysia, Skudai, Malaysia .

[3] Razieh qiasi, Malihe baqeri-Dehnavi, Behrooz Minaei-Bidgoli, Golriz Amooee , "Developing a model for measuring customer's loyalty and value with RFM technique and clustering algorithms" , The Journal of Mathematics and Computer Science Vol. 4 No.2 (2012) 172 – 181 , Department of Information Technology, University of Qom, Qom, Iran, raziehghiasi@gmail.com, Department of Information Technology, University of Qom, Qom, Iran, Programming_bagheri@yahoo.com, Department of Computer Engineering, University of Science and Technology, Tehran, Iran, minaeibi@cse.mcu.ed, Department of Information Technology, University of Qom, Qom, Iran, 2012 .

[4] Derya Birant , "Data minig using RFM analysis" , Knowledge-Oriented Applications in Data Mining InTech , University Campus STeP Ri Slavka Krautzeka 83/A51000 Rijeka, Croatia ,2011 .

[5] Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai, "A Two-Step Method for Clustering Mixed categroical and Numeric Data", Tamkang Journal of Science and Engineering, Vol. 13, No. 1, pp. 11_19 (2010), Department of Computer Science and Information Engineering, National Changhua University of Education, Changhua, Taiwan 500, R.O.C.,2010.

[6] Babak Daneshvar Rouvendegh, Turan Erman Erkan , "Selection of academic Staff using fuzzy analytical processing: FAPH pilot study",Technical Gazette 19, 4(2012),923-929.

Book referred -

[7] Data Mining: Concepts and Techniques , Second Edition , Jiawei Han University of Illinois at Urbana-Champaign Micheline Kamber .

[8] Tutorial referred –Dr. Rainer Haas Dr. Oliver Meixner Institute of Marketing & Innovation University of Natural Resources and Applied Life Sciences, Vienna, 'An Illustrated Guide to the Analytical Hierarchical Process' , University of Natural Resources and Applied Life Sciences, Vienna http://www.boku.ac.at/mi

- *Divya D. Nimbalkar is currently pursuing masters degree program in Computerr Engineering inNMIMS University, India, PH-9975672424 E-mail: ddndivya@mail.com*
- *Asst Prof. Paulami Shah is currently teaching inComputer Engineering Dept , MPSTME, NMIMS University, India, PH-9833968946. E-mail:paulami.shah@nmims.edu*