Combining RFM Model and Clustering Techniques for Customer Value Analysis of a Company selling online

Rachid AIT DAOUD
Department of physics
Sultan Moulay Slimane University
Beni Mellal, Morocco
r.aitdaoud@usms.ma

Abdellah AMINE
Department of Mathematics and Applications Mathematics
Sultan Moulay Slimane University
Beni Mellal, Morocco
a.amine@usms.ma

Belaid BOUIKHALENE
Department of Computer Science
Sultan Moulay Slimane University
Beni Mellal, Morocco
b.bouikhalene@usms.ma

Rachid LBIBB
Department of physics
Sultan Moulay Slimane University
Beni Mellal, Morocco
rachid.lbibb@gmail.com

Abstract—A case study of applying RFM (recency, frequency, and monetary) model and clustering techniques in the sector of electronic commerce with a view to evaluating customers' values is presented. Self-organizing maps method (SOM) is first used to determine the best number of clusters and then K-means method is applied to classify 730 customers into eight clusters when R, F, and M are the segmenting variables, and then developing effective marketing strategies for each cluster. The average values of RFM are computed for each cluster and the overall customers. The values of RFM variables for each cluster greater than those of the overall average are identified. The results show that the cluster 7 is the most important cluster because the average values of R, F and M are higher than the overall average value. In summary, the purpose of this case study is customer segmentation using RFM model and clustering algorithms (SOM and K-means) to specify loyal and profitable customers for achieving maximum benefit and a win-win situation.

Keywords—RFM model, Customer value, loyalty, Cluster analysis, Self-Organizing Maps method (SOM), K-means algorithm.

I. INTRODUCTION

A real case study for an online selling company in Morocco is employed by combining RFM (recency, frequency and monetary) model and data mining techniques (cluster analysis) to achieve better market segmentation and improve customer satisfaction. Data mining techniques such as Self Organizing Maps and K-means are used in this study to divide all customers into an appropriate number of clusters. On the other hand, the customers are segmented into similar clusters according to their RFM values. Therefore, the characteristics of each cluster are examined in order to determine and retain profitable and loyal customers and then develop effective marketing strategy for each cluster of customers. The transactional data consist of 730 customers who have purchased the website of the company from November 2013 to January 2015. The profile for each customer includes the customer identifier, gender, birth date, city, shopping frequency, date of first transaction, date of last purchase and the total spending at the online store.

The transaction recorded for each customer must be transformed to a usable format for the RFM model in this study. Therefore, customer values of different clusters can be measured by the use of clustering techniques and RFM model.

The remainder of the paper is as follows: Section 2 provides the literature review on RFM (recency, frequency and monetary) model. In Section 3, Cluster analysis is depicted. A case study of applying RFM model and clustering analysis is summarized in Section 4 and last section presents a brief conclusion.

II. LITERATURE REVIEW OF RFM MODEL

Recency, frequency and monetary (RFM model) is an effective method of segmenting and it is likewise a behavioral analysis that can be employed for market segmentation [1, 2].Hughes [1] described that the main asset of the RFM method is, on the one hand, to obtain customers' behavioral analysis in order to group them into homogeneous clusters, and on the other hand, to develop a marketing plan tailored to each specific market segment. RFM analysis improves the market segmentation by examining the when (recency), how often (frequency), and the money spent (monetary) in a particular item or service [3]. Yang [3] has summarized that customers who had bought most recently, most frequently, and had spent the most money would be much more likely to react to the future promotions.

978-1-5090-0478-2/15/$31.00 ©2015 IEEE

The advantage of RFM model resides in its relevance as long as it operates on several variables which are all observable and objective. They are all available at the order's past for each customer. These variables are classified according to three independent criteria, namely recency, frequency and monetary. Recency is the time interval between the last purchase and a present time reference; a lower value corresponds to a higher probability that a customer will make a repeat purchase. Frequency is the number of transactions that a customer has made in a particular time period and monetary means the amount of money spent in this specified time period [4]. The traditional approach to adopt RFM model is to sort the customers' data via each variable of RFM and then divide them into five equal quintiles [5, 6].The process of segmentation begins with sorting all customers based on recency, then frequency and monetary. For recency, the customer database is sorted in an ascending order (most recent purchasers at the top). Customers are then sorted for frequency and monetary in a descending order (most frequently and had spent the most money were at the top).The customers are then split into quintiles (five equal groups), and given the top 20%segment is assigned as a value of 5, the next 20% segment is coded as a value of 4, and so on. Therefore, all customers are represented by one of 125 RFM cells, namely, 555, 554, 553, . . . , 111.

Customers who have the most score are profitable. In this study, we adopt another approach proposed by chang et al. (2010) [7], it consistsof using the original data rather than the coded number. The definitions are as follows: recency is the time interval between the first day of study period and the last purchase; frequency is the number of transactions that a customer has made in a particular time period and monetary means the amount of money spent in this specified time period.

III. Review of cluster analysis

Data mining techniques have been widely employed in different domains. As the transactions of an organization become much larger in size, data mining techniques, particularly the clustering technique, can be applied to divide all customers into several clusters based on some similarities in these customers [8]. Clustering techniques are used to identify a set of groups that both minimize within-group variation and maximize between-group variation according to a distance or dissimilarity function [9].

The SOM (Self-Organizing Map) is an unsupervised neural network methodology, which need only the input, is used to clustering for problem solving [10] and market screening [11]. The network is formed by an unsupervised competitive learning algorithm, which can detect for itself (which means that no human intervention is needed during the learning process) patterns, strong features, and correlation in the large input data and code them in the output. The patterns of SOM in a high-dimensional input space are originally very complicated. When projected on a graphical map display, its structure, after clustering, turns out to be not only understandable but more transparent as well[12].

K-means method is used for grouping n vectors based on attributes into k partitions, where $k < n$, depending to some measures. The name comes from the fact that k clusters are identified and the center of a cluster is the mean of all vectors within this cluster. The algorithm starts with choosing k random initial centroids, then assigns vectors to the nearest centroid using Euclidean distance and recalculates the new centroids as means of the assigned data vectors. This process is repeated many times until vectors no longer altered clusters between iterations [13]. The K-means method is arguably a non-hierarchical method.

However, SOM has a few disadvantages. For example, with the result generated by SOM technique, it is difficult to detect clustering boundaries, a fact which limits their application to automatic knowledge [10]. Furthermore, in the k-means technique, the number of clusters and the initial starting point are randomly selected, which means that the algorithm has to turn several times to identify strong forms, because the final result depends on the initial starting points (different initial k objects may produce different clustering results). Due to the weakness of SOM and k-means method, the integration of these methods becomes desirable. Punj and Steward (1983) [14] took this view, adopting a two-staged clustering method by integrating the hierarchical method into the non-hierarchical. Kuo and al [15] have pointed out that it is preferable to use iterative partitioning methods instead of the hierarchical methods if the initial centroid and number of clusters are provided. If the information is provided, the iterative method consistently finds better clusters and higher accuracy than the hierarchical methods and yields faster results because the initialization procedure that ultimately determines the number of iteration is already executed. Therefore, self-organizing maps is used to determine the number of clusters for K-Means method.

IV. A case study

In this section the proposed model to determine loyal and profitable customers is described. Fig. 1 shows the required steps for the proposed model.

Dataset used in this case study was provided by a company selling online in Morocco and collected through its e-commerce website. The customer purchase database consists of 730 customers who purchased directly from the company website from November 2013 to January 2015. The profile for each customer includes the customer identifier, gender, birth date, city, shopping frequency, date of first transaction, date of last purchase and the total expense. Specifically, the notations of male and female are m and f, respectively.

The definition of the RFM model used in this study is shown in Table I.

The descriptive statistics for R, F and M are presented in Table II.

The maximum and minimum recencys' values in terms of days are calculated to be 438 and 1, the larger value demonstrates that the customer has shopped recently. In what concerns the frequency, the maximum and minimum values are calculated to be 19 and 1, respectively. For monetary, the

highest and lowest total amount spent are MAD 13731.60 and MAD 69, respectively (calculated in Moroccan Dirham).

Therefore, it is necessary to further decompose the dataset for more details. Table III present the characteristics of R, F and M for different genders, where the percentage of male customers is twice as high as that of females (67,9% versus 32,1%).

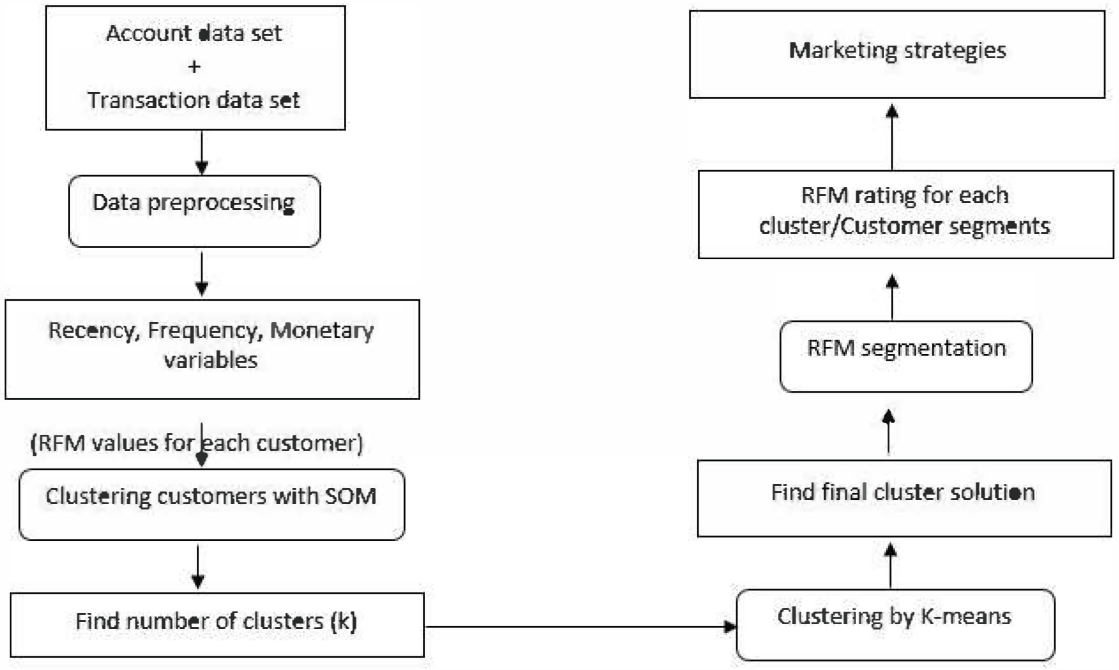

Fig. 1. Framework for customer segmentation based on RFM model and clustering techniques

TABLE I. AN EXAMPLE DATASET FROM CUSTOMERS, PRODUCTS AND TRANSACTIONS TABLES

Customer table

CID	Gender	Birth_date	Marital_S	address	City	…
00025	M	17/09/1983	1	Lot 03, Bloc…	Casa	…
00107	M	03/04/1990	0	Avenue Hass...	Rabat	…
00115	f	11/08/1985	1	N 19, apt03…	Casa	…
…	…	…	…	…	…	…

Product table

PID	Brand	Category	Subcategory	Description	Price	…
A00140	17/09/1983	6	64	VibroActio..	209.00	…
I00077	03/04/1990	3	32	CovrePhone..	100.00	…
A00125	11/08/1985	6	61	SlimSport..	79.99	…
…	…	…	…	…	…	…

Transaction table

TID	CID	PID	Date	Qte	Total	…
T00044	00107	A00140	15/09/2014	1	209.00	…
T00045	00025	I00077	07/09/2014	2	200.00	…
T00046	00107	A00125	01/10/2014	1	79.99	…
…	…	…	…	…	…	…

TABLE II. EXAMPLE R-F-M VALUES OF SOME CUSTOMERS AFTER DATA PREPROCESSING

CID	Recency	Frequency	Monetary
00025	311	7	1600.99
00107	335	11	3052.74
00115	42	2	225.00
…	…	…	…

TABLE III. DATA FORM WITH FOUR ATTRIBUTES.

Attribute name	Data content
Gender	Gender of the customers
Recent transaction time (R)	Refers to the number of days between the first day of study period and the last day of purchase
Frequency (F)	Refers to the total number of purchases from November 2013 to January 2015
Monetary (M)	Refers to the total amount spent by customers from November 2013 to January 2015. (Moroccan Dirhams or MAD)

TABLE IV. THE DESCRIPTIONS OF RECENCY, FREQUENCY AND MONETARY.

Variables	Max	Min	Average	Standard deviation
R	438	1	296,81	113,93
F	19	1	10,16	6,17
M	13731,60	69,00	3947,18	3405,75

TABLE V. CHARACTERISTICS OF R, F AND M FOR DIFFERENT GENDERS.

Description of "Gender"					
Gender=f			Gender=m		
Number of customers	[32,1 %] 234		Number of customers	[67,9 %] 496	
Variables	Group	Overral	Variables	Group	Overral
Average of R	242,53	287,90	Average of R	309,31	287,90
Average of F	6,99	10,16	Average of F	11,65	10,16
Average of M	2772,98	3986,63	Average of M	4559,20	3986,63

According to the proposed model, we use the Self-organizing maps and k-means methods of Tanagra to perform cluster analysis. First step: By applying self-organizing maps (SOM) method to cluster 730 customers, we found that the number eight is the best number of clustering based on the characteristics of recency, frequency and monetary. The result of this method is shown in Fig. 2. Later the number eight of cluster (k) generated by SOM can be used as a parameter for the second step. In this step, K-means method is applied to find the final solution. Therefore, Table IV summarizes the information of these eight clusters in terms of R, F, and M by K-means method when the number of clusters is set to eight.

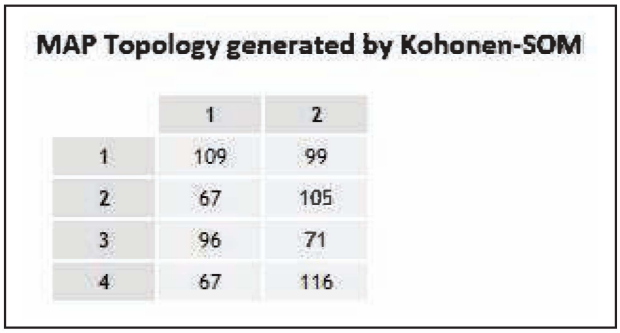

Fig. 2. Eight Clusters generated by SOM technique

By observing Table IV, most of the customers are grouped in clusters 6 and 7. More importantly, Cluster 7 is the most important cluster because the average values of L, R, F and M are higher than the overall average value, which might indicate that these customers purchase recently and frequently with a high money spent. It is worth investing lots of resources to sustain good relationships with these customers from the online store viewpoints. In addition to Cluster 7, customers in cluster 4 have high R and M values, which might indicate that they shopped the store recently with high money spent. So, they could be the customers with profit potential in the near future. Cluster 4 is called big spender customers.

For Clusters 1 and 5, customers have the lowest values of R, F and M compared to other clusters. In terms of recency, these customers have recently joined the online store, and in terms of frequency and monetary, these customers are not those who spend a lot of money, and even not those who often purchase. This cluster is called uncertain new customers. In addition to Clusters 1 and 5, customers in the Cluster 6 are new customers; the only difference between the two is that customers in Cluster 6 have recently shopped at the merchant site. They could become more important customers if the store could increase their frequency and monetary.

Cluster 2 has the characteristics of high F and M values but low R value. The low R value indicates that these customers have not shopped the store recently. Thus, the marketing strategy is to invite them to come back. Finally, Cluster 3 and 8, the values of R and F are above the average values. They have the characteristics of high recency, high frequency, it

can be said that these two clusters are loyal, but they do not have the right profile to become profitable customers, because their contribution to the company is still low. The marketing strategy is to encourage these customers to spend much more money.

TABLE VI. DESCRIPTIVE STATISTICS OF EIGHT CLUSTERS BASED ON K-MEANS METHOD.

Cluster	Number of customers	Male	Female	Average of R	Averageof F	Average of M	pattern
Cluster n1	89	50	39	243,68539	4,764045	1560,1137	***
Cluster n2	96	70	26	175,41667	13,104167	3987, 0856	*FM
Cluster n3	79	67	12	376, 37975	17,341772	10691, 82	RF*
Cluster n4	44	30	14	374,86364	5,386364	8797, 2724	R*M
Cluster n5	98	38	60	67,897959	3,173469	923,72276	***
Cluster n6	119	71	48	376, 2605	4,97479	1302, 9202	R**
Cluster n7	107	87	20	365, 90654	17,233645	5890, 529	RFM
Cluster n8	98	83	15	355,41837	14,061224	2867,7194	RF*
Total	**730**	**497**	**235**	**287.90**	**10.16**	**3 986.63**	

V. CONCLUSION

The main purpose of this paper is customer segmentation and measuring their loyalty by combining RFM model and data mining techniques (Clustering analysis). First, behavioral variables, recency, frequency and monetary were obtained using RFM model, then customers were segmented by applying two different methods, in the first self-organizing maps is applied to determine that eight might be the best number of clusters in this study. Later, K-means method is used to classify 730 customers into eight clusters in accordance with R, F, and M variables.

The results have demonstrated that group 7 is the most important cluster, as the average values of R, F, and M are superior to the overall average value. In summary, with the applications of RFM model and clustering techniques, the online store can easily identify high-value and profit potential customers and then design different marketing strategies to maximize its profits for different types of clusters.

VI. ACKNOWLEDGMENT

We would like to thank the online store specialized in electronics, fashion, home appliances, and children's items in Morocco for providing us with the data.

REFERENCES

[1] Hughes AM. Boosting response with RFM. Marketing Tools 1996;5:4-7.

[2] Marakas GM. Decision Support Systems in the 21st Century, Second Edition. Prentice Hall, Upper Saddle River, NJ, 2003.

[3] Yang AX. How to develop new approaches to RFM segmentation. Journal of Targeting, Measurement and Analysis for Marketing 2004;13(1):50-60.

[4] Wang CH. Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques. Expert Syst. Appl 2010., 37: 8395-8400.

[5] O'Connor, G. C., & O'Keefe, B. Viewing the web as a marketplace: the case of small companies. Decision Support Systems, 21(3), 1997, pp. 171–183.

[6] J. T. Wei, S. Y. Lin, and H. H. Wu, "A review of the application RFM model," African Journal of Business Management, vol. 4, no. 19, 2010 pp. 4199–4206.

[7] Chang, E. C., Huang, H. C., & Wu, H. H. Using K-means method and spectral clustering technique in an outfitter's value analysis. Quality & Quantity, 44(4), 2010, pp 807–815.

[8] Huang, S., Chang, E. C., & Wu, H. H. A case study of applying data mining techniques in an outfitter's customer value analysis. Expert Systems with Applications, 36, 2009, pp 5909–5915.

[9] Jo-Ting Wei, Shih-Yen Lin, Chih-Chien Weng, Hsin-Hung Wu.. Customer relationship management in the hairdressing industry: An application of data mining techniques. Expert Systems with Applications, 40, 2013, pp 7513–7518.

[10] Wang, S. Cluster analysis using a validated self-organizing method: Cases of problem identification. Intelligent Systems in Accounting, Finance and Management, 10(2), 2001, pp 127–138.

[11] Fish, K. E., & Ruby, P. An artificial intelligence foreign market screening method for small businesses. International Journal of Entrepreneurship, 13, 2009, pp 65–81.

[12] Churilov, L., Bagirov, A., Schwartz, D., Smith, K., & Michael, Data mining with combined use of optimization techniques and self-organizing maps for improving risk grouping rules: Application to prostate cancer patients. Journal of Management Information Systems, 21(4), 2005, pp 85–100.

[13] Derya Birant. Data Mining Using RFM Analysis, Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-154-1, InTech, DOI: 10.5772/13683, 2011. Available from: http://www.intechopen.com/books/knowledge-oriented-applications-in-data-mining/data-mining-using-rfm-analysis.

[14] Punj, G., & Steward, D. W. Cluster analysis in marketing research: review and suggestions for applications. Journal of Marketing Research, 20, 1983, pp 134-148.

[15] Kuo, R. J. Ho, L. M. Hu, C. M. Integration of self-organizing feature map and K-means algorithm for market segmentation, Computers & Operations Research, Vol.29, No.11, 2002, pp. 1475-1493.