A Review: RFM Approach on Different Data Mining Techniques

Article · September 2013

CITATION
1

READS
296

1 author:

Ankit Kharwar
Uka Tarsadia University
17 PUBLICATIONS **11** CITATIONS

SEE PROFILE

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 10, October 2013)

A Review: RFM Approach on Different Data Mining Techniques

Chandni Naik[1], Prof. Ankit Kharwar[2], Niyanta Desai[3]

[1, 3]*Student of M.Tech Computer Engineering in Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli*
[2]*Assistant Professor, Department of Computer Engineering, Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli*

***Abstract*-- Data mining is a well-known technique, which can be used to extract hidden information about customers' behaviors. It is used to improve the customer relationship management processes by various Organizations. Previous researches in constraint based pattern mining emphasis only on the concept of frequency.But the changes in the environment may occur frequently, so the frequently occuring pattern in past may not happen again in the future.To deal with these issues, in this paper the concept of recency, frequency, and monetary is consider. Based on the RFM value, customers can be clustered into different groups and the group information is very useful in market decision making.In this paper we have explained about sequential pattern mining using RFM, clustering using RFM, classification using RFM and association rule mining using RFM for customer segmentation, customer behavior prediction and product recommendation.**

***Keywords*-- Association rule mining, Classification, Clustering, RFM, Sequential pattern mining.**

I. INTRODUCTION

RFM stands for Recency, Frequency and Monetary value. RFM analysis is marketing technique used for analyzing customer behavior.

Recency: How recently a customer has purchased? [1]
Frequency: How often the customer purchases? [1]
Monetory: How much the customer spends? [1]

To enhance customer segmentation this techniques is dividing customers into various groups for future personalization accommodations and to recognize customers who are more liable to respond to promotions. [1]

Previous approaches emphasis only on the concept of frequency, which means, if a pattern is not frequent, it is abstracted from further consideration. In real life, however, Changes in the environment may occur frequently and patterns discovered may also change over time. Thus, the recent behavior of the users' is not necessarily the identical as the past ones and it may be possible that the frequently occuring pattern in past may not happen again in the future. [2] Furthermore, a pattern which has a low value is not important to the retailer. (e.g., prices or profits).Without considering the monetary value, retailers may be flooded by a large number of low-value patterns. [3]

So to deal with these issues, the concept of recency, Frequency, and monetary is used.

Combination of RFM analysis and data mining techniques give fruitful information for current and new customers. Compared to another cluster analysis method *clustering* based on RFM attributes provides more behavioral knowledge of customers' actual marketing levels. By using a customer demographic variable and RFM attributes *Classification* rules are discovered which provides useful information for managers to predict future customer behavior such as how recently the customer will probably purchase, how often the customer will purchase, and what will the value of his/her purchases. To meet the customer needs and provide a better recommendation the *Association rule mining* based on RFM attributes is used which analyzes the relationships of product properties and customers' contributions/ loyalties. [1]

The paper is organized as: section 2 of the paper gives RFM on sequential pattern mining. Clustering using RFM is explained in section 3. Classification using RFM is explained in section 4.Association rule mining using RFM is explained in section 5.We conclude the paper in section 6.

II. SEQUENTIAL PATTERN MINING USING RFM

Sequential pattern mining is the mining of frequently occurring ordered events or subsequences as patterns. For example *"Customers who buy a Canon digital camera are likely to buy an HP color printer within a month"*. [4]

A customer's data-sequence A can be represented by <(a1, t1, q1), (a2, t2, q2)... (a_n, t_n, q_n)>, where (a_j, t_j, q_j) indicates that item aj was purchased at time tj with quantity qj, $1 \leq j \leq n$, and $tj-1 \leq tj$ for $2 \leq j \leq n$. If items occur at the same time in the data-sequence, they are ordered alphabetically. [5] Several definations are given below.

Subsequence: A sequence<(ab)>is contained in data-sequence A =<(b, 1, 10), (c, 3, 5), (a, 5, 40), (b, 5, 20), (d, 7, 30), (a, 8, 20), (e, 8,10)>, because both items a and b occur in A at time 5. The sequence <(ab)(ae)> is a subsequence of A since itemsets (ab) and (ae) are contained in A at time 5 and time 8, respectively, and $t_{(ab)} < t_{(ae)}$. [5]

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 10, October 2013)

Compactness constraint: Assume that ms_length = 25, ms_length is user-specified maximum span length ,a sequence B =<(ab)(e)>is called a c-subsequence or compact subsequence of sid10 since (1) itemsets (ab) and (e) are contained in sid10 at time 40 and time 60, (2) $t_{(ab)} < t_{(e)}$, and (3) $t_{(e)} - t_{(ab)} \leq$ ms length = 25. [5]

Sid	Sequence
10	<(a,10,5),(c,30,4),(a,40,2),(b,40,1),(d,60,3),(e,6 0,1),(c,80,6)>
20	<(d,30,2),(a,50,15),(b,5 0,4),(e,50,3),(d,90,6)>
30	<(a,30,3),(b,30,4),(e,45,2),(c,55,8),(a,60,6),(b,60,1),(a,85,3),(e,85,7)>
40	<(b,20,2),(d,35,3),(b,40,1),(c,40,12),(e,70,5),(a,100,130),(d,100,50)>
50	<(c,70,160),(b,85,7),(d,85,6),(a,90,90),(d,100,10) >

Fig. 1. A Sequence database [5]

Frequency subsequence: B =< (ab) (e)> is a c-subsequence of sid10 and sid30. Besides, B occurrences two times in sid30, where the first occurrence is during the 30–45 time interval and the second is during 60–85. Thus, Fscore (B, sid30) = 2, Fscore (B, sid10) = 1, and finally the total frequency score of B is equal to 3. [5]

Recency subsequence: Assume the current timestamp $t_{current}$ = 110 and decay speed δ = 0.1, since c-subsequence B=< (ab) (e)> has two occurrences in sid30, and the most recent one occurs during time interval 60–85. A c-subsequence B's Rscore in sid30, Rscore (B, sid30) = $(1-0.1)^{110-85}$ =0.0717898. Rscore (B, sid10) = $(1-0.1)^{110-60}$=0.0051538. Hence, $TRscore_{DB}$ (B) = 0.0717898 + 0.0051538 = 0.0769436. [5]

Table I
A List Of Item Unit Prices [5]

Items	Price
A	10
B	150
C	25
D	45
E	80

Monetary subsequence: B = (ab) (e) occurs once in sid10 and twice in sid30. For sid10 and sid30, Mscore (B, sid10) = 2×10 + 1×150 + 1×80 = 250 and Mscore (B, sid30) = (3×10 + 4×150 + 2×80) + (6×10 + 1×150 + 7×80) = 1560.The $TMscore_{DB}$ (B) = 1560 + 250 = 1810. [5]

Sequence Monetary value (SM): The compact prefixes of B=< (a)> are [sid10:20:40:40] < (a, 10, 5), (c, 30, 4)> .The sequence monetary value of compact projection is=5×10 + 4×25=150. [5]

Total Sequence Monetary value (TSM): The total sequence monetary value of compact postfix B is the sum of all the sequence monetary values of B in SDB. [5]

The RFM-PostfixSpan Algorithm

The RFMPostfixSpan algorithm is developed by modifying the well-known PrefixSpan algorithm [6], which recursively partitions a sequence database into a number of projected databases and retrieves the RFM-SPs by exploring only the local frequent patterns in each projected database. RFM-PostfixSpan partitions SDB from the postfix to efficiently retrieve the recency score of a pattern. [5]

Step 1: Find RF-SP pattern

Scan SDB once and count the $TRscore_{SDB}$, $TFscore_{SDB}$ and $TMscore_{SDB}$ for each item to find the complete set of 1-RF-SPs and 1-RFM-SPs [5]. The complete set found which includes <a>: (0.549867, 8, 2540, 7160); <b>: (0.080289, 7, 3000, 5895); <c>: (0.061061, 5, 4750, 1810); <d>:(0.896466, 7, 3600, 9480); <e>: (0.094583, 5, 1440, 3060), where the notation "<pattern>: ($TRscore_{SDB}$, $TFscore_{SDB}$, $TMscore_{SDB}$, TSM_{SDB}). The pattern is RF-SP if $TRscore_{SDB} \geq$ Rminsup, $TFscore_{SDB} \geq$ Fminsup, $TMscore_{SDB}$ +TSM≥ Mminsup and The pattern is RFM-SP if $TRscore_{SDB} \geq$ Rminsup, $TFscore_{SDB} \geq$ Fminsup, $TMscore_{SDB} \geq$ Mminsup. [5]

Step 2: Divide and Search

Make the projected database of remaining postfix and calcualte the SM and TSM value of each postfix. [5]

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 10, October 2013)

Step3: Find subsets of sequential patterns

We start to find 2-RF-SPs and 2-RFM-SPs with postfix < (a)>, For the first compact projection [sid10:20:40:40]: <(a, 10, 5), (c, 30, 4)>, the Rscore, Fscore and Mscore of the two patterns,<(c)(a)> And <(a)(a)>, will be calculated based on recency,frequency,montary formula.Like wise all pattern of all five compact projection will be calculated.

Finally total $TRscore_{SDB}$, $TFscore_{SDB}$, $TMscore_{SDB}$ will calculated and constraint will check. The algorithm then proceeds to build the <(a)(a)>projected database and find 3-RF-SPs and 3-RFM-SPs with postfix <(a)(a)> by calling RFMPostfixSpan(<(a)(a)>, 2, $SDB_{|<(a)(a)>}$). Continuing in this way yields the complete set of RF-SPs and RFM-SPs in SDB are found. [5]

Postfix	TSM	Sid	Mscore of postfix	Start time	End time	Projected(prefix)database	SM
<(a)>	7160	10	20	40	40	*<(a,10,5),(c,30,4)>*	150
		20	150	50	50	*<(d, 30,2)>*	90
		30	30	85	85	*<(e,45,2),(c,55,8),(a,60,6),(b,60,1)>*	1200
			60	60	60	*<(a,30,3),(b,30,4),(e,45,2),(c,55,8)>*	
		40	1300	100	100	*<(e,70,5)>*	400
		50	900	90	90	*<(c,70,160), (b,85, 7),(d,85,6)>*	5320

Fig. 2 < (a)> - projected databases [5]

III. CLUSTERING USING RFM

The segmentation commences with recency, then frequency, and finally monetary value. It begins with sorting customers based on recency, most recent purchasers at the top. The customers are then split into five equal groups, and given the top 20% a recency score of 5, the next 20% a score of 4 and so on. Customers are then sorted and scored for frequency same as recency. This process is then undertaken for monetary also. Finally, all customers are ranked by combining R, F, and M values. [1] To group the customer with homogeneous RFM value we used K-Means++ [7] clustering algorithm because it is more advantages in terms of runtime and clustering quality. Instead of generating randomly, K-Means++ has a particular way of choosing centers.It first determines the initial center points by calculating their squared distance from the most proximate center which is already chosen. Because of new method for calculating an intial center points, KMeans++always finds better clusters than K-Means.K-Means++ algorithm is a much faster than K-Means because required number of iterations is determines by initialization procedure. [1] Here we are taking the dataset of sports store.

Table II
The Customer Segments Generated By K-Means++ Clustering Based On Rfm Values [1]

Cluster	Recency		Frequency		Monitory		RFM Pattern	Customer Type
	Day	R	#	F	TL	M		
C1	65.4	4.57	6.28	4.89	485.1	4.79	R↑F↑M↑	Best
C2	83.5	4.32	1.52	3.44	146.8	3.42	R↑F↑M↑	Valuable
C3	75.1	4.44	1.18	3.05	70.1	1.49	R↑F↑M↓	Shopper
C4	202.4	2.86	1.01	2.02	69.5	1.47	R↑F↓M↓	FirstTime
C5	247.8	2.22	4.27	4.51	387.4	4.67	R↓F↑M↑	Churn
C6	325.8	1.38	2.26	3.76	137.5	2.94	R↓F↑M↓	Frequent
C7	290.1	1.86	1.00	1.41	138.1	3.33	R↓F↓M↑	Spenders
C8	339.1	1.35	1.00	1.00	69.5	1.53	R↓F↓M↓	Uncertain
Overall		2.85		3.01		2.95		

From above figure we can see that the frequency and monitory value of Customers in C5 has high value but his recency value is low. So it indicates that something gose wrong with these customers, and therefore, it is required to contact with these customers by sending an e-mail, and give some offers on product.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 10, October 2013)

The clusters which contain RFM values with at least two upper arrows (↑) can be chosen as target ones.We require some action on customer of that cluster.After clustering, to evaluate the clustering results standard deviation and sum of square error (SSE) metrics were applied. The final result will be chosen based on SSE, so the cluster which has a minimum SSE will be selected as final result. [1]

IV. Classification Using RFM

Customer segments which are revealed by clustering algorithm is used by the classification algorithm as input and discover a classification rule by considering customers demographic variables such as their ages, genders, occupations, marital statuses and R-F-M attributes. C4.5 [8] decision tree algorithm is used for discovery of a classification rule. C4.5 algorithm works in a following way, first divide-and-conquer strategy used to create an initial tree and then it prunes the tree to avoid over fitting problem. Calculation of overall entropy and information gain of all the attributes takes place. The attribute with the highest information gain is opted to make the decision. [1]

V. Association Rule Mining Using RFM

To extract recommendation rules association rule mining was applied after the classification, mainly, it extracts the frequent purchase patterns from each cluster of the customers. A frequent pattern which is extracted by association rule mining is representing the common purchasing behavior of customers with analogous RFM values and demographic Variables.FP-Growth [4] algorithm is used for an association rule mining. After customers were classified by demographic variables, Feature attributes generate a recommendation list which is determined using a classification rule inducer.Here we are taking the example of sports store in which we consider those rules which has 40% confidence and 2% support. For example, if a customer in segment C3 (R↑F↑M↓) buys a soccer ball, then marketers should recommend backpack and water bottles products. However, if a customer in segment C4 (R↑F↓M↓) buys a soccer ball, then marketers should recommend of-kick product. So according to the RFM value of the customer marketer should recommend a product to the customer. [1]

VI. Conclusion

In this paper we consider the concept of RFM analysis of all data mining techniques like sequential pattern mining, clustering, classification and association rule mining.

We consider the concept of recency, frequency and monitory because changes in the environment may occur frequently, so it is possible that the patterns that appear frequently in the past may not appear again in the future. To efficiently discover complete set of RFM-SP pattern we go for pattern growth method because it significantly reduce the runtime and also retain more meaningful result for user. For clustering we used k-means++ algorithm because it has new method for choosing center so it is advantageous in term of cluster quality and runtime. Discovery of classification rules we consider customers demographic variables and R-F-M attributes because it will be helpful to target a customer profile more clearly. Discovery of the association rule FP-Growth algorithm is considered because it will eliminate the stage of candidate generation and scan the database two times only so it will decrease the runtime.

REFERENCES

[1] Derya Birant, Data Mining Using Rfm Analysis,Knowledge-Oriented Applications In Data Mining, Prof.Kimito Funatsu (Ed.),Isbn:978-953-307-154-1,In Tech, 2011.

[2] Chetna Chand, Amit Thakkar, And Amit Ganatra, "Target Oriented Sequential Pattern Mining Using Recency And Monetary Constraints," International Journal Of Computer Applications, Vol. 45, No. 10, May 2012.

[3] Mi-Hao Kuo A, Shin-Yi Wub, Kwei Tang Yen-Liang Chen A, "Discovering Recency, Frequency, And Monetary (Rfm) Sequential Patterns From Customers' Purchasing Data," Electronic Commerce Research And Applications, Pp. 241-251, March 2009.

[4] Micheline Kamber Jiawei Han, Data Mining:Concept And Techniques, 2nd Ed.: Morgan Kaufmann, 2006.

[5] Tony Cheng-Kui Huangb, Yu-Hua Kaoa Ya-Han Hua, "Knowledge Discovery Of Weighted Rfm Sequential Patterns From Customer Sequence Databases," The Journal Of Systems And Software, 779-788 2013.

[6] Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto,Qiming Chen, Umeshwar Dayal, Mei-Chun Hsu Jian Pei, "Prefixspan: Mining Sequential Patterns Efficiently By Prefix-Projected Pattern Growth," In 17th International Conference On Data Engineering, 2001, Pp. 215-224.

[7] Sergei Vassilvitskii David Arthur, "K-Means++: The Advantages Of Careful Seeding," In Proceedings Of Acm-Siam Symposium On Discrete Algorithms, New Orleans, 2007, Pp. 1027-1035.

[8] John Ross Quinlan, C4.5:Programs For Machine Learning.: Morgan Kaufmann, 1993.

[9] Fan Wu,Tzu-Wei Yeh Ya-Han Hu, "Considering Rfm-Values Of Frequent Patterns In Transactional Databases," In Software Engineering And Data Mining, Chengdu , 2010, Pp. 422 - 427.