6

Data Mining Using RFM Analysis

Derya Birant
Dokuz Eylul University
Turkey

1. Introduction

RFM stands for Recency, Frequency and Monetary value. RFM analysis is a marketing technique used for analyzing customer behavior such as how recently a customer has purchased (recency), how often the customer purchases (frequency), and how much the customer spends (monetary). It is a useful method to improve customer segmentation by dividing customers into various groups for future personalization services and to identify customers who are more likely to respond to promotions.

In recent years, data mining applications based on RFM concepts have also been proposed for different areas such as for the computer security (Kim et al., 2010), for automobile industry (Chan, 2008) and for the electronics industry (Chiu et al., 2009). Research cases of data mining with RFM variables include different data mining techniques such as neural network and decision tree (Olson et al., 2009), rough set theory (Cheng & Chen, 2009), self organizing map (Li et al., 2008), CHAID (McCarty and Hastak, 2007), genetic algorithm (Chan, 2008) and sequential pattern mining (Chen et al., 2009; Liu et al., 2009).

Integration of RFM analysis and data mining techniques provides useful information for current and new customers. *Clustering* based on RFM attributes provides more behavioral knowledge of customers' actual marketing levels than other cluster analyses. *Classification* rules discovered from customer demographic variables and RFM variables provides useful knowledge 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. *Association rule mining* based on RFM measures analyzes the relationships of product properties and customers' contributions / loyalties to provide a better recommendation to satisfy customers' needs.

This chapter presents incorporating RFM analysis into data mining techniques to provide market intelligence. It proposes a new three-step approach which uses RFM analysis in data mining tasks, including clustering, classification and association rule mining, to provide market intelligence and to assist market managers in developing better marketing strategies. In our model, (i) once clustering task is used to find customer segments with similar RFM values, (ii) then, using customer segments and customer demographic variables, classification rules are discovered to predict future customer behaviors, (iii) finally; association rule mining is carried out for product recommendation. The proposed model depends on the sentence "the best predictor of future customer behavior is past customer behavior". (Swearingen, 2009)

The purpose of this study is to provide better product recommendations than simple recommendations, by considering several parameters together: customer's segment, the

current RFM values of the customer, potential future customer behavior and products frequently purchased together. To the best of our knowledge, this chapter is the first in applying the RFM criterion in three data mining tasks, applied one after another, using customer demographic data, customer transaction data, and product properties. Experiments, which were carried out using the datasets collected by a sports store in Turkey through its e-commerce website, empirically demonstrate the benefits of using our model in direct marketing.

The rest of the chapter is organized as follows. Section 2 introduces the basics of RFM analysis and explains the recency, frequency and monetary concepts in detail. Section 3 reviews the literature and describes how data mining and RFM analysis are combined in the previous studies. Section 4 presents our proposed model and describes its architecture in detail. Section 5 demonstrates how the proposed model can be used to analyze a real world data, as a case study, including data preprocessing, RFM analysis, customer segmentation, customer behavior prediction and product recommendation. Finally, Section 6 concludes the chapter.

2. RFM analysis

The concept of RFM was introduced by Bult and Wansbeek (1995) and has proven very effective (Blattberg et al., 2008) when applied to marketing databases. RFM analysis depends on Recency (R), Frequency (F), and Monetary (M) measures which are three important purchase-related variables that influence the future purchase possibilities of the customers.

Recency refers to the interval between the time, that the latest consuming behavior happens, and present. Many direct marketers believe that most-recent purchasers are more likely to purchase again than less-recent purchasers. *Frequency* is the number of transactions that a customer has made within a certain period. This measure is used based on the assumption that customers with more purchases are more likely to buy products than customers with fewer purchases. *Monetary* refers to the cumulative total of money spent by a particular customer.

In order to demonstrate RFM analysis, an example dataset (customer transaction data) is given in Table 1. Table 2 shows the steps of RFM analysis, which involves scaling customers based on each RFM factor separately. The segmentation starts with recency, then frequency, and finally monetary value. It begins with sorting customers based on recency, i.e. period since last purchase, in order of lowest to highest (most recent purchasers at the top). The customers are then split into quintiles (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 – from the most to least frequent, coding the top 20% as 5, and the less frequent quintiles as 4, 3, 2, and 1. This process is then undertaken for monetary as well. Finally, all customers are ranked by concatenating R, F, and M values. This example shows that RFM analysis can be useful even if database is small of only 15 transactions whereas it would be more powerful when the database grows.

RFM analysis assigns value-scores to each customer on the basis of her past behavior. Using the quintile system explained above, at the most, 125 different scores (5x5x5) can be assigned. These cells differ in size from one another. A customer's score can range from 555 being the highest, to 111 being the lowest. The best customers are in quintile 5 for each factor (555) that have purchased most recently, most frequently and have spent the most money.

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

Table 1. An example dataset: customer transactions

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

Table 2. Customer quintiles and RFM values of customers

RFM provides a simple framework for quantifying customer behavior. For example, it is possible to infer from Table 2 that customer with id 9, which has RFM score 155, has made a high number of purchases with high monetary values but not for a long time. Something might have gone wrong with this customer, for example, he/she has most likely defected to a competitor's products and services or has found an alternate source and that is why his/her recency score is low. At this situation, marketers can contact with this customer and get feedbacks about how to do it better because he/she is one of the valuable customers according to his frequency and monetary values. Moreover, it is possible to plan a customer reactivation program and send him/her an extreme promotion in an effort to get his/her

attention. While customers with score 155 need a reminder, 551's need to be upsold, and 515's need a sticky recurring relationship. For example, if the RFM score of a customer is identified as 515, marketers can prepare a special customer packet that includes a thank-you letter, a list of company benefits, and an incentive to make another purchase from the online store within the next 30 days.

Several studies have discussed the different versions of RFM analysis. For example, in Weighted RFM (WRFM) version, each R,F,M value is multiplied by a weight value, w_R, w_F and w_M according to its relative importance to make intuitive judgments about ranking ordering. Another version, Timely RFM (TRFM) was proposed to deal with the product periodicity i.e. to analyze different product demands in different times. RFD (Recency, Frequency, Duration) version was proposed for the web site visitors to consider the duration i.e. how long someone spends on a website. RML (Recency, Monetary and Loyalty) is an adaptation of RFM, for annual transaction environments. Loyalty is typically a normalized form of Frequency in an annual period. RFR (Recency, Frequency, Reach) was proposed for social graph, i.e. Recency - last post, Frequency - total number of posts, Reach - networks, friends. FRAT (Frequency, Recency, Amount and Type of goods) is an extended version of RFM. It induces an improvement of the segmentation by way of taking into account the categories of bought products, for example, 0 - no buy, 1 - buy a compact car, 2 - buy an economy car, 3- buy a midsize car, 4 - buy a luxury car, where the order is defined in increasing order of size.

3. Data mining + RFM

3.1 Clustering using RFM

In recent years, several researchers have considered RFM variables in developing clustering models. For example, Hosseini et al. (2010) combined weighted RFM model into K-Means algorithm to improve Customer Relationship Management (CRM) for enterprises. Wu et al. (2009) applied RFM model and K-Means method in the value analysis of the customer database of an outfitter in Taiwan to establish strong relationship and eventually consolidate customer loyalty for high profitable long-term customers. Chuang and Shen (2008) first assessed the weights of R, F, M in order to know their relative importance by Analytical Hierarchy Process method, then evaluated Customer Lifetime Values (CLV) by clustering analysis and finally, sorted customers by self-organizing map method to recognize high value customer groups.

Differently from the previous Clustering+RFM studies, this chapter proposes using K-Means++ (Arthur & Vassilvitskii, 2007) algorithm to find customer segments with similar RFM values. We propose K-Means++ algorithm instead of other clustering algorithms such as K-Means, self-organizing map because of its advantages in terms of runtime and clustering quality.

K-Means++ was proposed as a specific way of choosing centers for the K-Means algorithm, instead of generating randomly. It determines the initial center points by calculating their squared distance from the closest center already chosen. Through new seeding method, K-Means++ consistently finds better clusters than K-Means and yields a much faster because the initialization procedure that ultimately determines the number of iterations to run before stopping. For example, on a small dataset, K-Means++ terminates almost twice as fast while achieving potential function values about 20% better, on the larger dataset, it is obtained up to 70% faster and the potential value is better by factors of 10 to 1000. (Arthur &

Vassilvitskii, 2007) For these reasons, we propose K-Means++ algorithm in this chapter, instead of K-Means or other clustering algorithms.
K-Means++ is a partitioning cluster algorithm by grouping *n* vectors based on attributes into *k* partitions, where k < n, according to some measure. The name comes from the fact that *k* clusters are determined and the centre of a cluster is the *mean* of all vectors within this cluster. The algorithm starts with determining *k* appropriate initial centroids, then assigns vectors to the nearest centroid using Euclidean distance and re-computes the new centroids as means of the assigned data vectors. This process is repeated over and over again until vectors no longer changed clusters between iterations.

3.2 Classification using RFM

Recently, integration of classification techniques and RFM was studied by Olson et al. (2009) to analyze customers' response possibilities to a specific product promotion. They compared three data mining techniques: logistic regression, decision trees and neural networks, and discussed the relative tradeoffs among these data mining algorithms in the context of customer segmentation. Cheng and Chen (2009) also combined RFM attributes and rough set theory (the LEM2 algorithm) to mine classification rules that help enterprises finding out the characteristics of customers in order to strengthen CRM. Furthermore, in order to evaluate the accuracy rate of the generated classification rules, they compared their approach with different three methods: Decision Tree, Artificial Neural Networks and Naive Bayes. According to the empirical results, their procedure outperforms the other methods listed in terms of accuracy rate. Ha (2007) used decision tree technique to track changes in RFM values of customers over time, to discover classification rules related to transition paths and thus to predict the next customers' RFM values from the current customers' RFM values.
Differently from the previous Classification+RFM studies, we apply a classification algorithm using customer segments discovered by a clustering algorithm and propose the discovery of classification rules by considering customers' demographic variables such as their ages, genders, occupations, and marital statuses.

3.3 Association rule mining using RFM

In data mining, association rules are descriptive patterns of the form X→Y, where X is termed the left-hand-side, and is the conditional part of an association rule; meanwhile, Y is called the right-hand-side, and is the consequent part. Association rule mining (ARM) is a task for discovering the hidden, interesting association rules, between items in the database, having support ≥ minsup threshold. The support of an association rule indicates how frequently that rule occurs in the data. Higher support corresponds to a stronger correlation between the items in the database.
Several studies applied ARM using RFM variables to analyze customer behaviors. For example, Chen et al. (2005) recorded all customer behavior patterns (emerging patterns, added patterns, perished patterns and unexpected patterns) generated by ARM for tracking changes in customer behaviors at different time snapshots. Liu and Shih (2005) proposed an approach depending on the idea is that if customers have had similar behavior, then they are very likely also to have similar RFM values. They firstly applied two hybrid methods (Weighted RFM-based method and the preference-based Collaborative Filtering method), and then extracted frequent patterns to represent the common behavior of customers with

similar purchases. Niyagas et al. (2006) used association rule mining technique and marketing techniques (RFM analysis) together to analyze historical data of e-banking usages from a commercial bank in Thailand. They applied Apriori algorithm to detect the relationships within the features of e-banking services.

Sequential Pattern Mining (SPM) is the extended version of the ARM. While ARM does not consider the order of transactions, SPM extracts frequent sequences while maintaining their order. SPM is more complicated than ARM because not only the frequent itemsets but also the temporal relationships must be found. Recently, SPM and RFM model were studied together. Chen et al. (2009) developed a novel algorithm for generating all RFM sequential patterns from customers' purchasing data. Liu et al. (2009) proposed a novel hybrid recommendation method that combines the segmentation-based sequential rule method with the segmentation-based K-Nearest Neighbors-Collaborative Filtering (KNN-CF) method. In their proposed method, sequential rules are extracted using customers' RFM values from the purchase sequences in the database.

Differently from the previous ARM+RFM and SPM+RFM studies, this chapter proposes the application of ARM after clustering and classification tasks to provide better product recommendations to customers i.e. according to their segments, RFM values and demographic variables.

4. Integrated approach

This section presents a new three-step approach which uses RFM analysis in data mining tasks. In our approach, (i) once *clustering* task is used to find customer segments with similar RFM values, (ii) then, *classification* rules are discovered using demographic variables (age, gender, education level etc.) and RFM values of customer segments to predict future customer behaviors, (iii) finally; *association rule mining* is carried out for product recommendation.

The proposed model can assist managers in developing better marketing strategies that fully utilize the knowledge resulting from data mining and RFM analysis. It is useful for predicting customer behaviors according to their demographic variables, because not all customers have purchased identical amounts, some have ordered more often, and some have ordered more recently. In addition, it provides better product recommendations than simple recommendations, by considering several parameters together: customer's segment, the current RFM values of the customer, potential future customer behavior and products frequently purchased together.

Figure 1 shows the IPO (Input, Process and Output) diagram of the proposed model. The model consists of five major parts: data preprocessing, RFM analysis, customer segmentation, prediction, and product recommendation with their evaluation processes. Each part of the approach is applied one after another. The output of each part becomes the input of the next part(s). The detail processes of each part are expressed as follows.

Step 1. Data Preprocessing

Data preprocessing step is needed to make knowledge discovery easier and correctly. Data preparation operations such as reduction in number of attributes, outlier detection, normalization, discretization, concept hierarch generation significantly improve the model; in fact a further increasing the prediction accuracy and saving in elapsed time.

In this step, the following operations should be made:

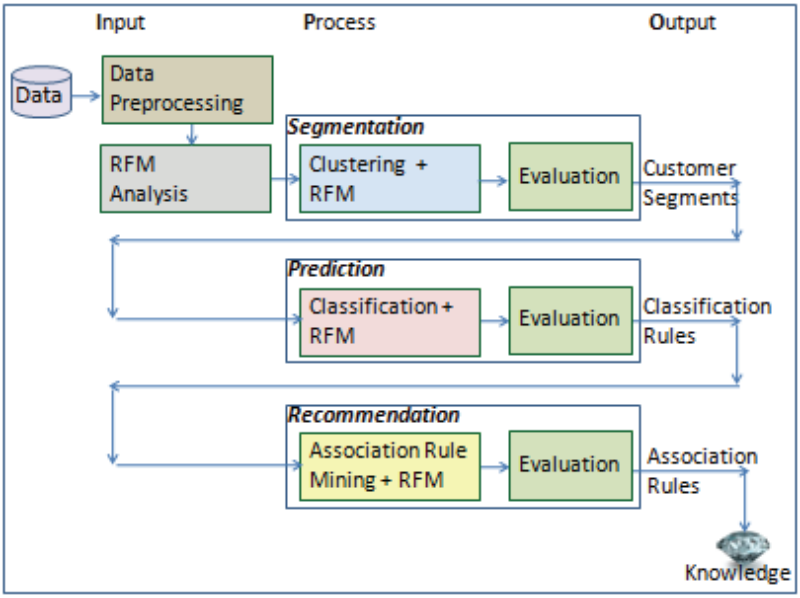

Fig. 1. IPO (Input, Process, Output) diagram of the proposed model

Dimensionality Reduction: Unnecessary attributes should be deleted, such as attributes that have only a few values (the others are null) or have only single value.

1.1 *Filling*: Missing values should be filled in using an appropriate approach.

1.2 *Handling*: Outliers and inaccurate values should be handled and removed from the dataset.

1.3 *Transformation*: Data should be transformed into an appropriate format.

1.4 *Discretization*: Before association rule mining task, continuous attributes should be encoded by discretizing the original values into a small number of value ranges. Because they have nearly a different value for every case; with such a high cardinality they provide little meaning to the association rule mining process. One common example of this phenomenon is the attribute that stores age values. The age attribute can be grouped into four ranges such as child (0-12), teenager (13-19), adult (20-59) and senior (60+).

1.5 *Concept Hierarchy Generation*: This method can be used to replace low level concepts (such as cities Istanbul, Ankara, or Izmir) by higher level concepts (such as states Marmara, Central Anatolia or Aegean).

Step 2. RFM Analysis

In this step, RFM analysis is applied by defining the scaling of R–F–M attributes. This process is divided into four parts introduced in the following:

2.1 Sort the data of three R–F–M attributes by descending or ascending order.

2.2 Partition the three R–F–M attributes respectively into 5 equal parts and each part is equal to 20% of all. The five parts are assigned 5, 4, 3, 2 and 1 score that refer to the customer contributions. The '5' refers to the most customer contribution, while '1' refers to the least contribution to revenue.

2.3 Repeat the previous sub-processes (2.1 and 2.2) for each R-F-M attribute individually. There are total 125 (5 x 5 x 5) combinations since each attribute in R–F–M attributes has 5 scaling (5, 4, 3, 2 and 1).

Step 3. Customer Segmentation

This step divides customers into numerous groups with similar RFM values, and assigns each customer to an appropriate segment. RFM analysis is used to evaluate customer loyalty, and thus identify the target customers with high RFM values by clustering analysis. The main advantage of this process is to be able to adopt different marketing strategies for different customer segments. Moreover, clustering customers into different groups improves the quality of recommendation, helps decision-makers identify market segments more clearly and therefore develop more effective strategies.

The detail process of this stage is expressed into two sub-steps.

3.1 *Clustering*: According to R–F–M attributes for each customer, data is partitioned into k clusters using the K-Means++ algorithm. (Arthur & Vassilvitskii, 2007) We propose K-Means++ algorithm instead of other clustering algorithms such as K-Means, SOM because of its advantages explained in Section 3.1.

Let D be a dataset expressed in terms of p attributes from the set $A = \{A_1, A_2,...,A_p\}$, and $A_r \in A$, which contains the intervals since last transactions, $A_f \in A$, which contains the number of transactions within a certain period, and $A_m \in A$, which contains the amount of money spent within a certain period. Each tuple $t \in D$ has p tuples $t = (CustomerID, r_i, f_i, m_i, \ldots)$, where $r_i \in \text{Range}(A_r)$ is a value in the range of the attribute A_r, $f_i \in \text{Range}(A_f)$ is a value in the range of the attribute A_f, $m_i \in \text{Range}(A_m)$ is a value in the range of the attribute A_m. Dataset D expressed as $D=<(1,r_1, f_1,m_1,\ldots), (2,r_2, f_2,m_2,\ldots),\ldots>$ is partitioned into k clusters $C = (C_1, C_2,..., C_k)$.

3.2 *Evaluation of Clustering Results*: The purpose of this step is to evaluate the quality of the clusters, to ensure compact clusters with little deviation from the cluster centroids and while to ensure larger separation between different clusters. Different methods can be used for evaluating the efficiency of data segmentation such as Standard Deviation (σ) defined in Eq. 1, Sum of Squared Error (SSE) defined in Eq.2.

$$\sigma=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-c)^2} \tag{1}$$

where x_i (i=1,2,..*N)* is an element in the cluster with N objects and c is the center of the cluster.

$$SSE=\sum_{i=1}^{K}\sum_{x\in c_i} dist(c_i,x)^2 \tag{2}$$

where k is the number of clusters and c_i is the center of i^{th} cluster.

Step 4. Prediction

In this step, classification rules are discovered using demographic variables (age, gender, education level etc.) and RFM values of customer segments to predict

future customer behaviors. For example, if *age* = *teenager* and *gender* = *male* and *state* = *Aegean* then $R\uparrow F\uparrow M\downarrow$, where the sign ↑ denotes that the value is greater than an average and sign ↓ denotes that the value is smaller than an average.

The rationale of this step is that if customers have similar demographic values, then they are very likely also to have similar RFM values. In fiercely competitive environments, discovering classification rules using customer demographic values is important for helping decision makers to target customer profiles more clearly. Additionally, the effect of classification rules on recommendations should be investigated to make more effective marketing strategies.

The detail process of this stage is expressed into two sub-steps.

4.1 *Classification:* Using customer demographic variables and R–F–M attributes, classification rules are discovered by C4.5 Decision Tree (Quinlan, 1993) algorithm. In data analysis techniques, the capabilities of C4.5 for classifying large datasets have already been confirmed in many studies.

C4.5 algorithm first grows an initial tree using the divide-and-conquer strategy and then prunes the tree to avoid overfitting problem. It calculates overall entropy and information gains of all attributes. The attribute with the highest information gain is chosen to make the decision. So, at each node of tree, C4.5 chooses one attribute that most effectively splits the training data into subsets with the best cut point, according to the entropy and information gain.

Let D be a dataset expressed in terms of p attributes from the set $A = \{A_1, A_2,...,A_p\}$, and k classes from the set $C = (C_1, C_2,..., C_k\}$. Thus each sample $d \in D$ has $p+1$ tuples $d = <V_1, V_2,.., V_p; C_j>$, where $V_i \in \text{Range}(A_i)$ is a value in the range of the attribute $A_i \in A$ and $C_j \in C$. A decision tree is constructed using C4.5 algorithm that selects an attribute A_i and a subset of its values V_i to branch on.

4.2 *Evaluation of Classification Accuracy*: Commonly used validation techniques for classification are simple validation, cross validation, n-fold cross validation, and bootstrap method. In our model, we propose n-fold cross validation technique because it matters less how the data gets divided. In this technique, dataset is divided into n subsets and the method is repeated n times. Each time, one of the n subsets is used as the test set and the other $n-1$ subsets are put together to form a training set. Then the average error across all n trials is computed.

Step 5. Product Recommendation

The core concept of this work is to extract recommendation rules from each customer group by considering classification rules and using FP-Growth Algorithm (Han et al., 2000). So, the purpose of this step is to identify the associations between customer segments, customer profiles and product items purchased together. By applying such an algorithm, it is possible to recommend products with associated rankings, which results in better customer satisfaction and cross selling.

The detail process of this stage is expressed into two sub-steps.

5.1 *ARM*: FP-Growth (Frequent Pattern Growth) is one of the Association Rule Mining (ARM) algorithms. Among the other ARM algorithms such as Apriori, Eclat, Mafia, it extracts the rules very fast from data by constructing a prefix tree and traversing this tree to generate rules. The algorithm scans the database two times only. Because of these reasons, FP-Growth algorithm is preferred in this study.

FP-Growth starts with compressing the database into a frequent-pattern tree (FP-Tree). During this process, it also constructs a header table which lists all frequent 1-itemsets to improve the performance of the tree traversal. Each item in the header table consists of two fields: item name and head of node link, which points to its first occurrence in the tree. After constructing FP-Tree and header table, the algorithm starts to mine the FP-tree by considering the items from the bottom of the header table and by recursively building conditional FP-Trees.

5.2 *Evaluation of Association Rules*: ARM algorithms use support and confidence thresholds and usually produce a large number of association rules which may not be interesting. An association rule is valid if it satisfies some evaluation measures. Evaluation process is needed to handle a measure in order to evaluate its interestingness.

In our approach, we propose to evaluate interestingness of mined rules and to express the relevance of rules with two descriptive criteria: Lift and Loevinger. These two criteria are defined on itemsets X, Y and rule $R: X \rightarrow Y$ as follows:

$$\text{Lift}(R)=\frac{P(XY)}{P(X)P(Y)} \tag{3}$$

$$\text{Loevinger}(R)=1-\frac{P(X)P(\neg Y)}{P(X \neg Y)} \tag{4}$$

Lift criterion represents the probability scale coefficient of having Y when X occurs. Loevinger criterion normalizes the centered confidence of a rule according to the probability of not satisfying its consequent part Y. In general, greater Lift and Loevinge values indicate stronger associations.

5. Case study

This section presents a case study which demonstrates how our proposed model was applied on the real-world data collected by a sports store. All steps of proposed model using a real world data is expressed in detail.

5.1 Data preprocessing

Dataset used in this case study was provided by a sports store in Turkey and collected through its e-commerce website within two years period. The complete dataset included 1584 different product demands in 54 sub-groups and 6149 purchase orders of 2666 individual customers. The purchase orders included many columns such as transaction id, product id, customer id, ordering date, quantity, ordering amount (price), sales type, discount and whether or not promotion was involved. While customer table included demographic variables such as age, gender, marital status, education level and geographic region; product table included attributes such as barcode, brand, color, category, sub-category, usage type and season.

Data preprocessing step handles outliers, fills missing values and makes dimensionally reduction, transformation, concept hierarchy generation, normalization and discretization. From the sport dataset, unnecessary attributes like e-mail addresses, telephone number

were obviously inappropriate to be used in data mining and were discarded. Continuous attributes were encoded by discretizing the original values into a small number of value ranges. For example, the age attribute was grouped into four ranges: child (0-12), teenager (13-19), adult (20-59) and senior (60+); the number of children attribute was replaced with four groups: 0, 1, 2 and 3+. In addition, gender attribute was encoded as *m* and *f* instead of *male* and *female*. Furthermore, concept hierarchy generation method was used to replace low level concepts (city) by higher level concepts (state). Recency attribute was constructed by calculating time interval between the last transaction date and present for each customer. Frequency attribute was constructed by finding the number of transactions that each customer has made within the certain period. Monetary attribute was constructed by calculating the cumulative total of money spent by each customer. Table 3 shows the partial data from customers, products and orders tables.

Customers

CID	Age	Sex	State	Education	Marital S.	Child	Year	...
5	Teenager	M	Aegean	Middle	NeverM	0	4	…
8	Adult	M	Marmara	HighSchool	Married	0	3	…
19	Adult	F	BlackSea	HighSchool	Married	3+	4	…
…	…	…	…	…	…	…	…	…

Products

PID	PName	Price	Brand	Group	Type	Color	Sex	...
100	NK DRI FIT PO	42	Nike	TShirt	Running	NK10	Male	…
106	PM AIKI JR	81	Puma	Sneaker	Soccer	PM03	Child	…
110	AD MALV OH	125	Adidas	Jersey	Soccer	AD05	Male	…
…	…	…	…	…	…	…	…	…

Orders

TID	PID	CID	Date	Quantity	Discount	Total	Type	...
T1	106	19	2008.12.2	1	0	81	SS	…
T2	100	8	2008.12.2	1	0	42	YS	…
T3	110	5	2008.12.3	1	0	125	SS	…
…	…	…	…	…	…	…	…	…

Table 3. An example data from customers, products and orders tables

5.2 RFM model

All customers were ranked by considering their recency, frequency and monetary values and they were represented by R-F-M codes. Table 4 shows example R-F-M values of some customers after RFM analysis. For example, it is possible to infer from the first row in Table 4 that customer with id 5 has R-F-M values 4-3-4 respectively. This customer has made a high number of purchases with high monetary values, not long ago.

Figure 2 shows the distribution of the number of customers with respect to their RFM values. The distribution of RFM values varies within the limits of 0 - 4.6%. At the most, the customers have the RFM value 555 (125 customers), followed by RFM value 113 (108 customers), and next, 107 customers have the RFM value 321. Some RFM values such as 121, 125, 231, 311 etc. were not assigned to any customer.

CID	Recency (Day)	Frequency (Number)	Monetary (TL)	R	F	M	RFM
5	95	4	237	4	3	4	434
8	269	10	790	5	5	5	555
19	321	1	81	1	1	2	112
…	…	…	…	…	…	…	…

Table 4. Example R-F-M values of some customers after RFM analysis

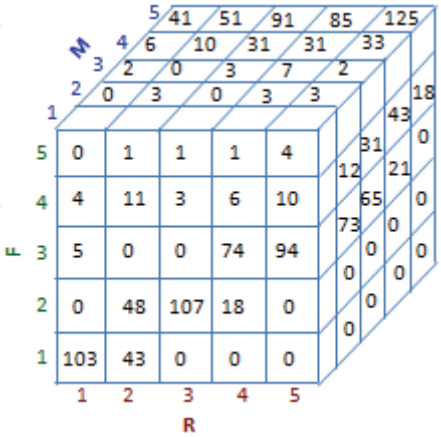

Fig. 2. RFM distribution: 125 possible RFM values and the number of customers

5.3 Customer segmentation

K-Means++ clustering was employed to group customers with similar RFM values. Customers were segmented into eight target markets in terms of the period since the last transaction (recency), purchase frequency and total purchase expenditure (monetary). The *k* parameter was set to 8, since eight (2x2x2) possible combinations of inputs (RFM) can be obtained by assigning ↑ or ↓, according to the average to R,F,M values of a cluster being less than or greater than the overall average. If the average R (F, M) value of a cluster exceeded the overall average R (F, M), then an upward arrow ↑ was included, otherwise and downward arrow ↓ was included. For example, R↑F↓M↓ represents that the average recency value of a customer segment is greater than overall average, while frequency and monetary average values are smaller than overall averages. These eight customer groups include best customers (most valuable), valuable customers, shoppers, first-time customers, churn customers, frequent customers, spenders, and uncertain customers (least valuable).

Table 5 presents the result, listing eight clusters, each with the corresponding number of customers, their average actual and scaled R, F and M values. The last row also shows the overall average for all customers. The last two columns of Table 5 show the RFM pattern for each cluster and corresponding customer type. While cluster C5 contains the maximum number of customers (425 customers, 16%), C6 includes the minimum, only 135 customers (5%).

Customer segment C1 contains the most valuable customers, because it consists of customers who have recently made regular purchases, and also have higher average

purchase frequency and purchase expenditure. It is followed by cluster C2, and next cluster C3. Cluster C4 (R↑F↓M↓) may include first-time customers, who have recently visited the company, with higher recency and lower purchase frequency and monetary expenditure. Customers in C5 have made a high number of purchases with high monetary values but not for a long time. Something might have gone wrong with these customers, and therefore, it seems to be an indicator of churn likelihood. It is needed to contact with these customers i.e. sending an e-mail, and to plan a customer reactivation program i.e. promotion suggestion. Cluster 8 is concluded to be the least valuable for the business, because customers coded as 111, 112, 121 are generally the least likely to buy again.

Cluster	Size	Recency (Avg.)		Frequency (Avg.)		Monetary (Avg.)		RFM Pattern	Customer Type
		Day	R	#	F	TL	M		
C1	309	65.4	4.57	6.28	4.89	485.1	4.79	R↑F↑M↑	Best
C2	392	83.5	4.32	1.52	3.44	146.8	3.42	R↑F↑M↑	Valuable
C3	415	75.1	4.44	1.18	3.05	70.1	1.49	R↑F↑M↓	Shopper
C4	300	202.4	2.86	1.01	2.02	69.5	1.47	R↑F↓M↓	FirstTime
C5	425	247.8	2.22	4.27	4.51	387.4	4.67	R↓F↑M↑	Churn
C6	135	325.8	1.38	2.26	3.76	137.5	2.94	R↓F↑M↓	Frequent
C7	381	290.1	1.86	1.00	1.41	138.1	3.33	R↓F↓M↑	Spenders
C8	309	339.1	1.35	1.00	1.00	69.5	1.53	R↓F↓M↓	Uncertain
Overall	2666		2.85		3.01		2.95		

Table 5. The customer segments generated by K-Means++ clustering based on RFM values

The clusters that have RFM values with at least two upper arrow (↑) can be selected as target ones, all customers who belong to these clusters become candidates for conducting suitable marketing strategies, which attract the most attention.

After customer segmentation, standard deviation and SSE metrics were used to evaluate clustering results. All clusters had a lower standard deviation and SSE values. The result, as shown in Figure 3, confirmed that these eight clusters were significantly distinguished by recency, frequency, and monetary. Standard deviation values ranges from 0.67 being the highest, to 0.33 being the lowest. In the experiments, K-Means++ algorithm was run 10 times with different initial center values and the clustering result with minimum SSE was selected as final result.

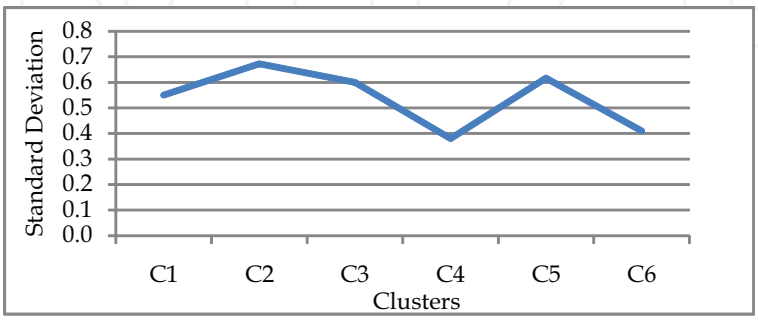

Fig. 3. Standard deviations of clusters (customer segments)

5.4. Customer behavior prediction

A customer segment is not as enough to identify, and then to predict customer's behavior. Many direct marketers believe that the RFM variables of customers are generally associated with customer profiling. For example, customers with profiles *age* = *teenager* and *gender* = *female* and *state* = *Aegean* can generally have $R\uparrow F\uparrow M\downarrow$ pattern, while customers with profiles *age* = *senior* and *gender* = *male* and *state* = *EasternAnatolia* can generally have $R\downarrow F\uparrow M\downarrow$ pattern. For this reason, in this step, classification rules were discovered using demographic variables (age, gender, education level etc.) and RFM values of customer segments.

Figure 4 shows a part of classification rules, found in the case study, that identify customer profiles and the associated RFM values. For example, rule 1 shows that customer profile with (State=Aegean, EducationLevel=Bachelors, MaritalStatus=Married, Gender=M) is highly related to R↑F↑M↑ pattern. Similarly, classification rule 5 represents that a customer profile (State=EasternAnatolia, Gender=F) is dominant or most strongly associated with R↑F↓M↓ pattern.

Rule 1: **if** *State*=Aegean **and** *EducationLevel*=Bachelors **and** *MaritalStatus*=Married **and** *Gender*=M **then** R↑F↑M↑
Rule 2: **if** *State*=Aegean **and** *MaritalStatus*=Married **and** *Gender*=M **and** *Age*=Adult **then** R↑F↑M↑
Rule 3: **if** *State*=Marmara **and** *Membership*=3 **and** *EducationLevel*=HighSchool **and** *Children*=0 **then** R↓F↑M↑
Rule 4: **if** *State*=CentralAnatolia **and** *Age*=Teenager **and** *MaritalStatus*=NeverMarried **and** *Membership*=3 **then** R↑F↑M↓
Rule 5: **if** *State*=SouthEasternAnatolia **and** *Children*=3+ **and** *Gender*=M **then** R↓F↑M↓
Rule 6: **if** *State*=Mediterranean **and** *EducationLevel*=Middle **then** R↓F↓M↑
Rule 7: **if** *State*=EasternAnatolia **and** *Gender*=F **then** R↑F↓M↓
Rule 8: **if** *State*=BlackSea **and** *Children*=3+ **and** *Gender*=F **then** R↓F↓M↓

Fig. 4. A part of classification rules found in the case study

In our experiments, classification accuracy was observed by using 5-fold cross validation technique. The highest classification accuracy 81% is obtained when different values were given to parameters (confidence factor, minimum number of objects, number of folds etc.) as inputs.

5.5 Product recommendation

In the proposed approach, after generating classification rules, association rule mining was applied to extract recommendation rules, namely, frequent purchase patterns from each group of customers. The extracted frequent purchase patterns represent the common purchasing behavior of customers with similar RFM values and with similar demographic variables. For example, not all women age 45-54 have the same tendency to purchase a product; so we should also consider their RFM values, customer segments and the other products frequently purchased together with that product.

After customers were classified by demographic variables, the recommendation list was generated by feature attributes determined using a classification rule inducer. Parameters were set up to identify association rules that had at least 40% confidence and 2% support imposed on the FP-Growth association rule algorithm. Figure 5 shows a part of association rules, found in the case study. 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. Other rules (Rule 7 and Rule 8) denote that marketers should recommend two different products (Reebok Sneakers or Converse Shoes) to customers according to their different RFM values.

Rule 1: {C1, Adidas soccer jersey (man), Adidas soccer jersey (woman)} → {Adidas soccer jersey (child)}
Rule 2: {M>3, Adidas Sneaker (child)} → {Adidas Socks, Adidas Equipment Bag}
Rule 3: {C3, Adidas Soccer ball} → {Adidas Backpack (unisex), Adidas Water Bottles}
Rule 4: {C4, Adidas Soccer ball} → {Nike of-kick}
Rule 5: {C5, Converse Sneaker (woman), Puma Sneaker (man)} → {Nike Cap (unisex)}
Rule 6: {C6, Adidas T-Shirt (male)} → {Adidas Short (male), Adidas Training Bag}
Rule 7: {R<=3, F<=3, M>3} → {Reebok Sneakers}
Rule 8: {R<=3, F<=3, M<=3} → {Converse Shoes}

Fig. 5. A part of association rule set on support 2% and confidence 40% for each customer segment

In the evaluation process, association rules were reduced by more than 50% to the set of potentially interesting and valuable rules. For example, the number of association rules related to C4 customer segment was reduced from 67 to 42. These reduction percentages also give weight to the need of taking into consideration the information brought by the confirmation property.
In the proposed approach, it is possible to predict the customer segment of a new customer from classification rules, according to her/his profile, and then a recommendation list can be generated according to his/her predicted segment.

6. Conclusion

This chapter proposes a novel three-step approach which uses RFM analysis in three data mining tasks: clustering, classification and association rule mining, applied one after another. Firstly, customer segments with similar RFM values are identified to be able to adopt different marketing strategies for different customer segments. Secondly, classification rules are discovered using demographic variables (age, gender, education level etc.) and RFM values of customer segments to predict future customer behaviors and to target customer profiles more clearly. Thirdly, association rules are discovered to identify the associations between customer segments, customer profiles and product items purchased, and therefore to recommend products with associated rankings, which results in better customer satisfaction and cross selling.

This chapter presents incorporating RFM analysis into data mining techniques to provide market intelligence. It aims to bring attention of data miners and marketers to the importance and advantages of using RFM analysis in data mining. In order to evaluate the proposed model and empirically demonstrate the benefits of using this model in direct marketing, a case study was carried out using the datasets collected within two years period by a sports store in Turkey through its e-commerce website. According to experimental study results, proposed approach provides better product recommendations than simple recommendations, by considering several parameters together: customer's segment, the current RFM values of the customer, potential future customer behavior and products frequently purchased together.

Future research can focus in the followings: First, the proposed approach can be tested for different versions of RFM such as Weighted RFM (WRFM), Timely RFM (TRFM), FRAT (Frequency, Recency, Amount and Type of goods). As the number of additional variables increases, the number of cells will geometrically increase. For example, if we add two types of product parameter, the number of FRAT cells becomes $2 \times 5 \times 5 \times 5 = 500$. Thus, it is unrealistic to estimate RFM model with more than two additional variables. Second, the effectiveness of the proposed approach can be evaluated for different application domains such as for the web site visitors (RFD), for annual transaction environments (RML), and for social graphs (RFR).

7. References

Arthur, D. & Vassilvitskii, S. (2007). K-Means++ The advantages of careful seeding, *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027-1035, ISBN:978-0898716245, New Orleans, January 2007, Society for Industrial and Applied Mathematics, USA.

Blattberg, R.C.; Kim, B-D. & Neslin, S.A. (2008). *Database Marketing: Analyzing and Managing Customers*, Chapter 12, pp. 323-337, Springer, ISBN: 978-0387725789, New York, USA.

Bult, J. R. & Wansbeek, T. (1995). Optimal selection for direct mail, *Marketing Science*, Vol. 14, No. 4, (Fall 1995) 378-394, ISSN:0732-2399.

Chan, C.C.H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer, *Expert Systems with Applications*, Vol. 34, No. 4, (May 2008) 2754-2762, ISSN:0957-4174.

Chen, M.; Chiu, A. & Chang, H. (2005). Mining changes in customer behavior in retail marketing, *Expert Systems with Applications*, Vol. 28, No. 4, (May 2005) 773-781, ISSN:0957-4174.

Chen, Y-L.; Kuo, M-H.; Wu, S-Y. & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data, *Electronic Commerce Research and Applications*, Vol. 8, No. 5, (October 2009) 241-251, ISSN: 1567-4223.

Cheng, C-H. & Chen, Y-S. (2009). Classifying the segmentation of customer value via RFM model and RS theory, *Expert Systems with Applications*, Vol. 36, No. 3, (April 2009) 4176-4184, ISSN: 0957-4174.

Chiu, C-Y.; Kuo, I-T. & Chen, P-C. (2009). A market segmentation system for consumer electronics industry using particle swarm optimization and honey bee mating optimization, *Global Perspective for Competitive Enterprise, Economy and Ecology*, Springer London, pp. 681- 689.

Chuang, H. & Shen, C. (2008). A study on the applications of data mining techniques to enhance customer lifetime value – based on the department store industry, *Proceedings of the 7th International Conference on Machine Learning and Cybernetics*, pp. 168-173, ISBN: 978-1424420964, Kunming, China, July 2008, IEEE.

Ha, S.H. (2007). Applying knowledge engineering techniques to customer analysis in the service industry, *Advanced Engineering Informatics*, Vol. 21, No. 3, (July 2007) 293–301, ISSN:1474-0346.

Han, J.; Pei, H.& Yin. Y. (2000). Mining Frequent Patterns without Candidate Generation. *Proceedings of Conference on the Management of Data (SIGMOD'00)*, pp. 1-12, ISBN:1581132174, Dallas, Texas, United States, May 2000, ACM New York, NY, USA.

Hosseini, S.M.; Maleki, A. & Gholamian, M.R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty, *Expert Systems with Applications*, Vol. 37, No. 7, (July 2010) 5259–5264, ISSN:0957-4174.

Kim, H. K.; Im, K. H. & Park, S. C. (2010). DSS for computer security incident response applying CBR and collaborative response, *Expert Systems with Applications*, Vol. 37, No. 1, (January 2010) 852-870, ISSN:0957-4174.

Li, S-T.; Shue, L-Y. & Lee, S-F. (2008). Business intelligence approach to supporting strategy-making of ISP service management, *Expert Systems with Applications* ,Vol. 35, No. 3, (October 2008) 739–754, ISSN:0957-4174.

Liu, D-R. & Shih, Y-Y. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value, *Information & Management*, Vol. 42, No. 3, (March 2005) 387-400, ISSN:0378-7206.

Liu, D-R.; Lai, C-H. & Lee, W-J. (2009). A hybrid of sequential rules and collaborative filtering for product recommendation, *Information Sciences*, Vol. 179, No. 20, (September 2009) 3505-3519, ISSN:0020-0255.

McCarty, J. A. & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression, *Journal of Business Research*, Vol. 60, No. 6, (June 2007) 656-662, ISSN:0148-2963.

Niyagas, W.; Srivihok, A. & Kitisin, S. (2006). Clustering e-banking customer using data mining and marketing segmentation, *ECTI Transaction CIT*, Vol. 2, No. 1, (2006) 63-69.

Olson, D.L.; Cao, Q.; Gu, C. & Lee, D. (2009). Comparison of customer response models, *Service Business*, Vol. 3, No. 2, (June 2009) 117-130, ISSN: 1862-8516.

Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning*, Morgan Kaufmann Publishers. 302 pages.

Swearingen, C. (2009). *101 Powerful Marketing Strategies for Growing Your Business Now!*, SmallBiz Marketing Services, pp. 24-27.

Wu, H-H.; Chang, E-C. & Lo, C-F. (2009). Applying RFM model and K-Means method in customer value analysis of an outfitter, *Global Perspective for Competitive Enterprise, Economy and Ecology*, ISSN: 1865-5440, Part 12, pp. 665-672, ISBN:978-1848827615, Springer London.

Knowledge-Oriented Applications in Data Mining
Edited by Prof. Kimito Funatsu

ISBN 978-953-307-154-1
Hard cover, 442 pages
Publisher InTech
Published online 21, January, 2011
Published in print edition January, 2011

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by 'Data Mining' address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821