The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/0368-492X.htm

Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis

Peiman Alipour Sarvari and Alp Ustundag
Istanbul Technical University, Istanbul, Turkey, and
Hidayet Takci
Cumhuriyet University, Sivas, Turkey

Abstract

Purpose – The purpose of this paper is to determine the best approach to customer segmentation and to extrapolate associated rules for this based on recency, frequency and monetary (RFM) considerations as well as demographic factors. In this study, the impacts of RFM and demographic attributes have been challenged in order to enrich factors that lend comprehension to customer segmentation. Different types of scenario were designed, performed and evaluated meticulously under uniform test conditions. The data for this study were extracted from the database of a global pizza restaurant chain in Turkey. This paper summarizes the findings of the study and also provides evidence of its empirical implications to improve the performance of customer segmentation as well as achieving extracted rule perfection via effective model factors and variations. Accordingly, marketing and service processes will work more effectively and efficiently for customers and society. The implication of this study is that it explains a clear concept for interaction between producers and consumers.

Design/methodology/approach – Customer relationship management, which aims to manage record and evaluate customer interactions, is generally regarded as a vital tool for companies that wish to be successful in the rapidly changing global market. The prediction of customer behaviors is a strategically important and difficult issue because of the high variance and wide range of customer orders and preferences. So to have an effective tool for extracting rules based on customer purchasing behavior, considering tangible and intangible criteria is highly important. To overcome the challenges imposed by the multifaceted nature of this problem, the authors utilized artificial intelligence methods, including k-means clustering, Apriori association rule mining (ARM) and neural networks. The main idea was that customer clusters are better enhanced when segmentation processes are based on RFM analysis accompanied by demographic data. Weighted RFM (WRFM) and unweighted RFM values/scores were applied with and without demographic factors and utilized to compose different types and numbers of clusters. The Apriori algorithm was used to extract rules of association. The performance analyses of scenarios have been conducted based on these extracted rules. The number of rules, elapsed time and prediction accuracy were used to evaluate the different scenarios. The results of evaluations were compared with the outputs of another available technique.

Findings – The results showed that having an appropriate segmentation approach is vital if there are to be strong association rules. Also, it has been determined from the results that the weights of RFM attributes affect rule association performance positively. Moreover, to capture more accurate customer segments, a combination of RFM and demographic attributes is recommended for clustering. The results' analyses indicate the undeniable importance of demographic data merged with WRFM. Above all, this challenge introduced the best possible sequence of factors for an analysis of clustering and ARM based on RFM and demographic data.

Originality/value – The work compared k-means and Kohonen clustering methods in its segmentation phase to prove the superiority of adopted segmentation techniques. In addition, this study indicated that customer segments containing WRFM scores and demographic data in the same

Kybernetes
Vol. 45 No. 7, 2016
pp. 1129-1157

0368-492X
DOI 10.1108/K-07-2015-0180

clusters brought about stronger and more accurate association rules for the understanding of customer behavior. These so-called achievements were compared with the results of classical approaches in order to support the credibility of the proposed methodology. Based on previous works, classical methods for customer segmentation have overlooked any combination of demographic data with WRFM during clustering before proceeding to their rule extraction stages.

Keywords Customer segmentation, Performance evaluation, Association rule algorithm, Demographic variables, RFM analysis, Self-organizing map (SOM)

Paper type Research paper

1. Introduction

In the 1990s, in the business domain the concept of customer relationship management (CRM) gradually emerged, which prevailed from its very first years, gaining prominence as a legitimate area of scholarly inquiry and stimulating the interest of the global business and research community (Soltani and Navimipour, 2016). CRM is the operational model by which enterprises understand and influence customer behavior via interaction in order to obtain new customers, keep old customers and increase customer loyalty and thereby improve profits (Chung and Chen, 2016). Faced with increasing complexity and competition, today's firms need to develop innovative activities to understand customers' needs and improve customer satisfaction and retention (Razieh *et al.*, 2012). The main objective of CRM is to create long-lasting and profitable relationships with customers. The increased digitization of transactions has resulted in a boost to the information about customers stored in large transactional databases (Khajvand *et al.*, 2011). Furthermore, it is the strongest and the most efficient method for maintaining and creating relationships with customers (Soltani and Navimipour, 2016).

Clustering is the process of collecting a set of physical or abstract objects into groups of similar objects (Hosseini and Mohammadzadeh, 2016). A prominent application of database marketing is the clustering of customers for direct marketing, by which analysts try to find homogeneous groups of customers with respect to their response behavior using so-called data mining (DM) tools (Ambler *et al.*, 2002). In this regard, some authors have proposed soft clustering methods, which yield more promising results than hard clustering ones as well as greater clustering quality within segments than are derived from the finite mixture model. Wu and Cho compared customer distributions for hard clustering and soft clustering. In soft clustering, each customer has a mixed-membership score associated with each latent class. They emphasized: micro-segments for shopping behavior (buying frequency and money spent), micro-segments for online customer satisfaction and demographic characteristics of customers based on experimental results (Coussement *et al.*, 2014).

In recent years, database marketing techniques have evolved from simple recency, frequency and monetary (RFM) models (models involving the recency and frequency of customer purchases, and their payments in monetary terms) into statistical techniques such as chi-square automatic interaction detection (CHAID) and logistic regression. More recently, models using artificial neural networks (ANNs) have been advocated in many enterprises to serve for CRM by analyzing valuable customer information (C.H. Cheng and Chen, 2009; Xu and Chu, 2015). Their ease of use and quick implementation are the reasons that marketers continue to employ RFM models despite the development of advanced DM techniques. Also, they are easily understood by

managers and decision makers. That having been said, RFM models have some blind spots. Primarily, their lack of both exploratory and forecasting abilities are negative qualities in a predictive model. Second, they presuppose a continuous mode of behavior for every customer in order to target them for promotional activity. In other words, the technique does not take into account the impact of individuals' life stages or life cycle transitions on the likelihood of their responses. Finally, if only the most attractive RFM segments are used as the primary targeting method, there is a risk that other profitable segments might be neglected.

For this study, we aimed to propose different algorithms using customers' RFM analysis outputs and their demographic data for different scenario forms. To extract true product-based (process-based) rules from customer transactions, customer segmentation needs to be done using correct target clusters. Up to now analysis based on RFM customer segmentation has been considered solely on RFM indices. In other words, in order to formulate a comprehensive customer clustering model using RFM analysis, weighting factors and other attributes, such as demographic variables, should not be ignored. It is contended that a performance analysis of all the scenario types will support the idea that a unification of customer transaction attributes is necessary to derive true cluster-based rules.

The rest of this study is organized as follows: Section 2 gives an overview of related works. Section 3 explains model factors, data characteristics and descriptions. Section 4 describes methodologies and the adopted algorithms. Section 5 covers empirical analysis and shows how to select the best-fitted algorithms, as well as offering a practical challenge to the methodology, and performance analysis. Then, an interpretation of the results is provided. Section 6 challenges some benchmarks by comparing their methods' performances with that of the proposed approach. In Sections 7 and 8, one can find a discussion of the work's implications and the conclusion of the paper.

2. Literature review

CRM has attracted a lot of attention, and many businesses that are end-users of information technology solutions have spent considerable sums investing in implementing CRM systems. These systems have been integrated, to a greater or lesser extent, with their operational and business processes. However, what should be borne in mind is that CRM is a basic, commonsense idea that can be put into practice with nothing more than a spreadsheet and a modest database (Nettleton, 2014). One of the most important instruments for business to absorb and retain customers is by deploying a CRM technique. With such a technique, an association can implement customer-based strategies to improve relationships with customers that result in an increase of customers' loyalty and satisfaction and in their better retention. Also, it is useful for improving customer relationships and cooperation with them, by integrating customers' values and their requirements into business processes. Without a doubt, analyzing data from customer transactions is a very important aspect of CRM. Shortening sales cycles and developing incrementally closer relationships with customers are other benefits of good CRM. In their literature review paper, E.W.T. Ngai *et al.* (2009) mentioned that CRM is categorized into four dimensions which are customer identification, customer attraction, customer retention and customer development. What is more, DM is divided into seven functions (i.e. association, classification, clustering, forecasting, regression, sequence discovery and visualization) (Ngai *et al.*, 2009).

To maximize the performance of CRM by an enterprise, it needs to have a defined, segmented group of its audience of consumers and customers. Segmentation was first

introduced in the marketing literature by W. Smith (1956). Later, Claycamp and William (1968) mentioned segmentation as an alternative to product differentiation strategy. The main idea of segmentation or clustering is to group similar customers together.

Fallis (2013) mentioned that there are many clustering algorithms in the literature. It is difficult to provide a crisp categorization of clustering methods because these categories may overlap such that a method may have features from several categories and, in general, the major fundamental clustering methods can be classified as partitioning methods (like k-means clustering algorithms), hierarchical methods (such as agglomerative and divisive hierarchical clustering algorithms), density-based methods (density-based clustering based on connected regions with high density and clustering based on density distribution functions are the most popular methods) and grid-based methods (like STING: the statistical information grid) (Fallis, 2013). As another effective clustering algorithm, the self-organizing map (SOM) of Kohonen, which is well-known for its ability to map an input space with an neural network (NN) (Demartines and Blayo, 1992), is very popular for clustering customers nowadays.

A segment can be described as a set of customers who have similar characteristics based on demography, behaviors, values and so on. The selection of segmentation techniques has become more important due to the fact that developments in information and communication technologies, especially database management and DM systems have changed ways of marketing.

The vast availability of data, and the inefficient performance of traditional statistical techniques (or statistics-oriented segmentation tools) when handling such voluminous amounts of data, have stimulated researchers to find more effective segmentation tools in order to discover more useful information about their markets and customers. Thus, knowledge discovery (KD) and DM have been seen as a solution to this problem. Disciplines such as machine learning, statistics, artificial intelligence (soft and hard computing techniques), expert systems, and data and knowledge management technologies have been incorporated within KD and DM, making use of their theories and algorithms (Hiziroglu, 2013). Teichert (2008) tried to segment airline customers. They used some demographic and monetary factors and segmented their customers into business and leisure categories. They could not generalize the method and concluded that their findings were doubtful due to the lack of customers' frequencies and loyalty factors that were observed. Chung and Chen (2016) used purchase importance and demographic factors to segment their customers at the point of a mandatory service encounter. Despite service marketing literature indicating that loyal customers are more likely to participate in service coproduction than are new customers, they ignored some important factors (frequency and recency) and this limited the application of their work. The above-mentioned articles examined non-RFM-based segmentation.

RFM analysis originated in the practice of direct marketing by catalog sales companies in the 1960s (Blattberg *et al.*, 2008). RFM methodology is very effective for customer segmentation. According to Kahan (1998), RFM is easy to use and can generally be implemented very quickly (Kahan, 1998). Furthermore, it is a method that managers and decision makers can understand (Marcus, 1998). Hu and Yeh considered recorded transactions without collecting customer information. Therefore, they defined the RFM pattern and developed a novel algorithm to discover complete sets of RFM patterns to approximate sets of customers. Instead of evaluating the values of patterns from a customer's point of view (i.e. those that not only occurred frequently, but involved a recent purchase and a higher percentage of revenue), their study measured pattern ratings directly by considering RFM factors, using the RFM scores of frequent

patterns (Hu and Yeh, 2014). A study by Ambler *et al.* (2002) showed that RFM was the second most common method used by direct marketers, after cross-tabulation. By the definition of Robert S. Michael, a cross-tabulation is a joint frequency distribution of cases based on two or more categorical variables. Displaying a distribution of cases by their values on two or more variables is known as contingency table analysis and is one of the more commonly used analytic methods in the social sciences. One of the limitations of cross-tabulation is that it is time-consuming because it uses the bulk of values, whereas RFM analysis can group the bulk of values into classes of scores.

Coussement *et al.* (2014) investigated the influence of problems with data accuracy – an important dimension of data quality – using RFM analysis for customer segmentation for two real-life direct marketing data sets. In spite of the availability of more statistically sophisticated methods, J.A. McCarty and M. Hastak (2007) investigated RFM, CHAID and logistic regression as analytical methods for direct marketing segmentation, using two different data sets (McCarty and Hastak, 2007). This work indicated that RFM may provide results similar to CHAID and logistic regression. Overall, the study concluded that RFM can perform at an acceptable level in many database marketing situations when a direct marketer is limited to using basic transaction variables. Meanwhile, this paper addressed the broader issue that RFM may focus too much attention on transaction information and ignore individual difference information (e.g. values, motivations, lifestyles) which is very important for assessing individual preferences. Chu Chai Henry Chan (2008) presented an approach that combined customer-targeting and customer segmentation for campaign strategies (Chan, 2008). Yen-Liang Chen *et al.* included two recency and monetary examples and proposed the RFM pattern for sequential pattern mining (SPM). Although several researchers developing DM methods have considered RFM variables, this paper was the first to apply RFM in SPM. The RFM patterns, as traditional sequential patterns, could be applied in various e-commerce applications, such as cross-selling, product recommendation, personalized marketing, e-catalog design and product bundle design (Chen *et al.*, 2009). This investigation identifies customer behavior using an RFM model and then it uses a customer lifetime value (LTV) model to evaluate proposed segmented customers. Additionally, this work proposes using a generic algorithm (GA) to select more appropriate customers for each campaign strategy. With the aid of DM tools, they constructed a new customer segmentation method based on RFM, demographic and LTV data (Namvar *et al.*, 2009). Their new customer clustering technique consists of two consecutive phases. First, with k-means segmentation, customers are segmented into different groups with respect to their RFM. Second, using demographic data, each cluster is partitioned again into new clusters. They adopted DM methods by combining SOMs and k-means techniques to apply an RFM model for a hair salon in Taiwan to segment customers and develop marketing strategies and they suggested that investigating customer demographic characteristics might suggest different marketing implications. Coussement *et al.* were interested in the performance of segmentation. They investigated data quality's influence on clustering processes and segments' quality using values analysis, logistic regression and decision trees (DTs) (Coussement *et al.*, 2014).

Even though customer clustering based on RFM analysis has been studied widely for several papers, none have tried to show how RFM analysis factors and demographic data may change customer segments' precision. To the best of our knowledge, all works that proposed RFM analysis for customer segmentation have clustered customers ignoring weighting factors and demographic data while segmenting.

3. Data characteristics

Customers have a variety of dissimilarities according to their characteristics. In consumer and industrial marketing literature, several segmentation variables can be found, such as geographic, demographic, firmographic, behavioral, decision-making-process-related variables, purchasing behavior, situation factors, personality, lifestyle, psychographics and so on (Hiziroglu, 2013).

3.1 Demographic data

Demographics are determinate statistics for a data set of a customer population. Moreover, demographics are used to discern the study of measurable subsets. Marketing and public opinion polling are other areas of study that use demographic data. Generally speaking, ethnicity, gender, age, employment status, etc., are considered as being demographic data.

The description of time-oriented demographic changes in a population is incurred from demographic trends. For instance, the average spend rate of a population may increase or decrease over time. Both dispensations and trends of values within a demographic variable are of interest (Power and Elliott, 2006; Ryder, 1965).

3.2 RFM data

RFM is a model that differentiates considered customers from a mass of data by three attributes: interval of customer consumption, frequency and payment value. The detailed definitions of RFM are as follows:

(1) The recency of the last purchase that is shown by R refers to the duration of time between the last purchase time and the time of a survey. Here the desired state is shorter in duration time so that R is bigger.

(2) The frequency of purchases, which is shown by F, refers to the number of transactions in a specific time cycle. In a desirable state, a bigger value for F means that there has been a high repetition of purchases.

(3) The monetary value of purchases, which is shown by M, refers to the money consumption value expended by customers during a certain time-period. The desired state for this is for there to be much money, so M is bigger. Although an RFM model is a good method for differentiating important customers in large amounts of data by the three variables, there are two studies that have different opinions with respect to the three variables of the RFM model. Hughes (1994) considered that the three variables were of equal importance and, therefore, the weights of the three variables were identical. On the other hand, Stone (1995) indicated that the three variables were different in importance based on the characteristics of an industry. Thus, the weights of the three variables were not equal (C.-H. Cheng and Chen, 2009).

One can replace M with duration, which can be used to analyze customer behavior, for example, the value of the time spent by radio audiences on a certain channel.

Most businesses will keep transaction records for their customers. All that is needed is a table with customer names, their dates of purchase and their purchase values. One methodology is to assign a scale of 1 to 10, where 10 is the maximum value and to stipulate a formula whereby the data suits the scale.

4. Proposed methodology

In recent years, DM has not only had great popularity in research areas, but also in commercialization. Nowadays, by utilizing DM tools for assisting CRM, some techniques, which include DTs, ANNs, GAs, association rule mining (ARM), etc., are usually used in fields such as engineering, science, finance and business to solve customer-related problems. Generally, no DM tool for CRM is perfect, because there are some uncertainties inherent in their use. For example, with DTs, too many instances lead to large DTs that decrease classification accuracy rates. ANNs have long training times, especially in large data sets, and it is a trial-and-error process. GAs converge slowly, have large computation times and are less stable. ARM can generate rules of scopes that are huge and so may be redundant (C.-H. Cheng and Chen, 2009).

One of the major challenges for any firm of marketers is to identify the business target for their products or services. All customers have specific profiles and preferences as business targets. To develop a marketing strategy and marketing plan, these profiles have to be established and used. In this regard, we first clustered the community into the form of grouped customers, and then, to extract proper decision support rules, used the features of a typical cluster. We were going to investigate different ARM-based segmentation approaches. We used these clusters for rule mining. In order to obtain the most effective rules with less redundancy, various combinations of model scenarios have been designed, and their performances have been evaluated by machine learning algorithms.

We used the abilities of the above-mentioned tools to develop a logical and reliable ARM-based CRM procedure. The individual steps of the proposed methodology for this effort are shown in Figure 1. The principal methodology for customer-transaction-based ARM contains five phases.

4.1 Data preprocessing phase

The data preprocessing phase is the first stage for the preparation of raw data before its transformation into refined and usable data. At this phase the operations of filling, handling, transformation and discretization are performed on raw data. This refining step includes operations such as attributes numbers regulation, outlier detection, normalization, discretization and concept hierarchy generation. It affects prediction accuracy and elapsed time duration directly and positively. The following procedures are undertaken in this regard:

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

Filling: missing values should be filled in using an appropriate method. In this case, we replaced values with specified expressions depending on their condition, type and nullity using MATLAB and IBM SPSS.

Handling: outliers and inaccurate values should be handled and removed from the data set.

Transformation: data should be transformed into an appropriate format.

Discretization: before an ARM task, continuous attributes should be encoded by discretizing the original values into a small number of value ranges. Because they have a different value for nearly every case, with such a high cardinality they provide little information of meaning to the ARM process (Birant, 2003).

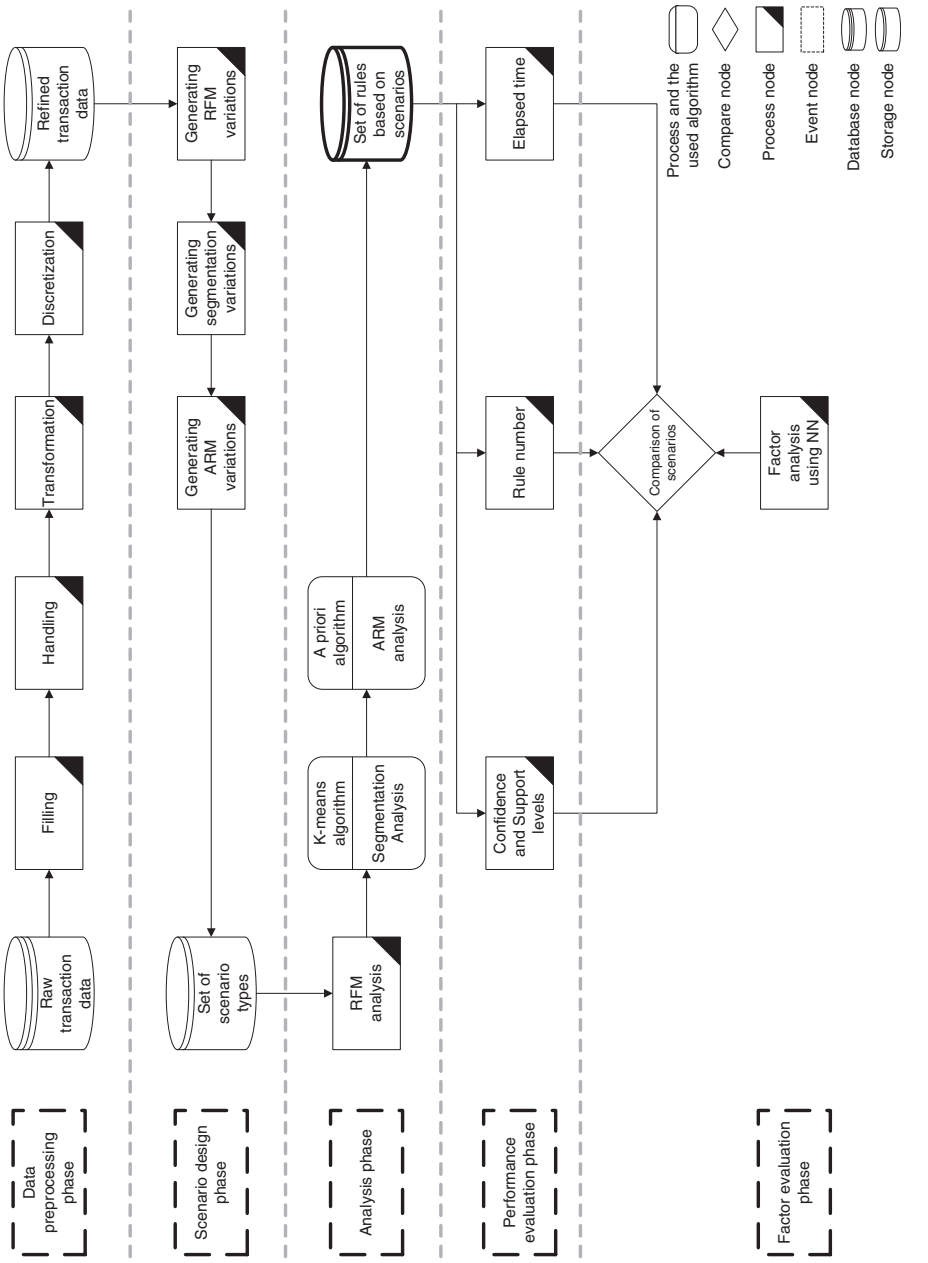

Figure 1.
Proposed methodology and operation steps flowchart

4.2 Scenario design phase

The output of this phase was refined transaction data. After this, it became our main database for upcoming analysis. Our data contained specific factors and the main idea of this work was based on the factors analysis flows. So to capture the best and most accurate streams, in the second phase of this methodology, we tried to develop whole possible scenarios analysis. Every individual scenario is a combination of sequences of factors and their different specifications. In this phase, different combinations of RFM analysis, segmentation and ARM have been designed and stored as the set of analysis scenarios.

Different attributes and data types in each transaction led us to provide comprehensive scenario types which covered all possible configurations through basic DM applications. RFM analyses have seven different outputs; RFM values and RFM scores (recency score, frequency score and monetary score) as well as an averaged RFM score (the weighted mean of RFM scores). Meanwhile, RFM values are classed in five score categories from 1 to 5. In addition, there are different combinations of importance weighting for weighted RFM (WRFM) analysis which were used for this test. Clusters may contain WRFM values solely or WRFM values with demographic data. Table I depicts all 42 possible scenario types induced from regular customer transaction data. For instance, in scenario 1, RFM values with equal weights ($W_R = W_F = W_M$) were considered at the segmentation step and demographic factors, which had been used for various analyses during the ARM step. As another example, for scenario 22, at the segmentation step, RFM scores were used alongside demographic factors. It should be noted that, in this scenario, the weights of monetary and frequency were equal and higher than the weight of recency ($W_M = W_F > W_R$). In other words, for scenarios 15-28 and 37-42, demographic data had been used together with RFM outputs during the clustering process, and in other scenarios we used demographic data only at the ARM step. Scenarios 1, 8 and 30 were the most common approaches in the literature which used RFM factors only covering $W_R = W_F = W_M$ during the clustering phase (Coussement *et al.*, 2014).

4.3 Analysis phase

At the analysis phase, each scenario was run and analyzed. The results of these analyses were stored as the set of rules based on scenarios. This phase contained the performance of three fixed steps which were RFM analysis, segmentation and ARM analysis. However, each analysis was run under different scenarios.

4.3.1 RFM analysis. RFM analysis is a marketing technique used for analyzing customer behavior, such as when a customer has purchased lately (recency), how often the customer purchases (frequency) and what is the purchase amount the customer has spent (monetary). It is a useful method for improving customer segmentation by dividing customers into various groups for future personalization services and for identifying customers who are more likely to have a tendency to respond to promotions. RFM analysis depends on recency (R), frequency (F), and monetary (M) measures which are three important purchasing-related variables influenced by future purchasing possibilities. Related calculations are shown in the following equation:

$$R(C_i) = \frac{R_i - R_{Min}}{R_{Max} - R_{Min}} \tag{1}$$

Scenario	RFM analysis phase RFM weights W_R: weight of recency; W_F: weight of frequency; W_M: weight of monetary	Segmentation phase Segmentation factors RFM outputs RFM score	RFM value	RFM average	Demographic factors	Association rules mining phase Association rule mining With demographics	Without demographics
1	$W_R = W_F = W_M$		✓			✓	
2	$W_R > W_F = W_M$		✓			✓	
3	$W_F > W_R = W_M$		✓			✓	
4	$W_M > W_R = W_F$		✓			✓	
5	$W_R = W_F > W_M$		✓			✓	
6	$W_M = W_R > W_F$		✓			✓	
7	$W_M = W_F > W_R$		✓			✓	
8	$W_R = W_F = W_M$	✓				✓	
9	$W_R > W_F = W_M$	✓				✓	
10	$W_F > W_R = W_M$	✓				✓	
11	$W_M > W_R = W_F$	✓				✓	
12	$W_R = W_F > W_M$	✓				✓	
13	$W_M = W_R > W_F$	✓				✓	
14	$W_M = W_F > W_R$	✓				✓	
15	$W_R = W_F = W_M$		✓		✓		✓
16	$W_R > W_F = W_M$		✓		✓		✓
17	$W_F > W_R = W_M$		✓		✓		✓
18	$W_M > W_R = W_F$		✓		✓		✓
19	$W_R = W_F > W_M$		✓		✓		✓
20	$W_M = W_R > W_F$		✓		✓		✓
21	$W_M = W_F > W_R$		✓		✓		✓
22	$W_M = W_F > W_R$	✓			✓		✓
23	$W_R = W_F = W_M$	✓			✓		✓
24	$W_R > W_F = W_M$	✓			✓		✓
25	$W_F > W_R = W_M$	✓			✓		✓
26	$W_M > W_R = W_F$	✓			✓		✓
27	$W_R = W_F > W_M$	✓			✓		✓
28	$W_M = W_R > W_F$	✓			✓		✓
29	$W_M = W_F > W_R$			✓		✓	
30	$W_R = W_F = W_M$			✓		✓	
31	$W_R > W_F = W_M$			✓		✓	
32	$W_F > W_R = W_M$			✓		✓	
33	$W_M > W_R = W_F$			✓		✓	
34	$W_R = W_F > W_M$			✓		✓	
35	$W_M = W_R > W_F$			✓		✓	
36	$W_M = W_F > W_R$			✓	✓		✓
37	$W_R = W_F = W_M$			✓	✓		✓
38	$W_R > W_F = W_M$			✓	✓		✓
39	$W_F > W_R = W_M$			✓	✓		✓
40	$W_M > W_R = W_F$			✓	✓		✓
41	$W_R = W_F > W_M$			✓	✓		✓
42	$W_M = W_R > W_F$			✓	✓		✓

Table I.
All 42 possible scenario types induced from regular customer transaction data

Note: ✓, the attribute is selected

$$F(C_i) = \frac{F_i - F_{Min}}{F_{Max} - F_{Min}} \quad (2)$$

$$M(C_i) = \frac{M_i - M_{Min}}{M_{Max} - M_{Min}} \quad (3)$$

in which $R(C_i)$, recency value of i customer; R_i, duration passed since i customer last purchased; R_{Min}, minimum duration passed since a customer last purchased; R_{Max}, maximum duration passed since a customer last purchased; $F(C_i)$, frequency value of i customer; F_i, frequency of the i customer purchases; F_{Min}, minimum purchase frequency of each customer; F_{Max}, maximum purchase frequency of each customer; $M(C_i)$, monetary value of i customer; M_i, monetary value of i customer; M_{Min}, minimum monetary value of customer purchase; and M_{Max}, maximum monetary value of customer purchase.

4.3.2 Segmentation. Purchase transactions and the tracking of the purchasing behavior of customers was segmented during the customer segmentation phase. Customer segmentation is a vital DM methodology used in CRM. There are plenty of approaches to clustering and algorithms in the extant literature. We have introduced and compared two of them to select the best algorithm below.

K-means clustering algorithm. The k-means clustering method accepts D as an input data set, a number, K, of segments that are to be formed and a function $dist(X_i, X_k)$ that indicates a kind of homogeneity between each couple of inspections.

Suppose C_h, $h = 1, 2, \ldots, K$ is a cluster whose centroid is defined as the point z_h the coordinates of which are equal to the mean of each feature for the observations belonging to the cluster. That is:

$$Z_{hj} = \frac{\sum_{X_i \in C_h} x_{ij}}{card\{C_h\}} \quad (4)$$

SOM. The SOM is an excellent tool used at the exploratory phase of DM. It projects input space on prototypes of a low-dimensional regular grid that can be utilized effectively to visualize and explore properties of the data (Vesanto and Alhoniemi, 2000). Like most ANNs, SOMs operate in two modes: training and mapping. "Training" builds the map using input examples (a competitive process, also called vector quantization), while "mapping" automatically classifies a new input vector. An SOM consists of components called nodes or neurons. Associated with each node is a weight vector of the same dimension as the input data vectors and a position in the map space. The usual arrangement of nodes is a two-dimensional regular spacing in a hexagonal or rectangular grid. The SOM describes a mapping from a higher dimensional input space to a lower dimensional map space. The procedure for placing a vector from data space onto the map is to find the node with the vector of closest (smallest distance metric) weight to the data space vector (Ultsch and Siemon, 1990). The Kohonen algorithm is the most famous clustering tool based on SOM. As has been shown in Section 5.2, clustering quality and runtime were the most compelling cognitive reasons that convinced us to use the k-means segmentation algorithm as it portions n observations into k clusters when compared with the Kohonen algorithm (SOM). The k-means algorithm segments observations with the nearest means into the same cluster.

4.3.3 Association rule mining and algorithms. ARM is a famous method for extracting dealings between features and factors in large-scale data sets. It is trusted to extract reliable rules and principles using different dimensions of interests. Many algorithms for generating association rules have been presented over time. Three of the most popular algorithms are presented as follows.

FP-growth algorithm. Frequent pattern mining plays an essential role for mining associations (Agrawal *et al.*, 1993, 1996; Mannila *et al.*, 1994; Nichol *et al.*, 2008), correlations (Brin *et al.*, 1997), causality (Silverstein *et al.*, 1998), sequential patterns (Agrawal and Srikant, 1994), episodes (Mannila *et al.*, 1997), multidimensional patterns (Yusof and Kraak, 2015), max patterns (Bayardo, 1998), partial periodicity (Han *et al.*, 1999), emerging patterns (Yu *et al.*, 2011) and many other important DM tasks.

Let $I = \{a_1, a_2, \ldots, a_m\}$ be a set of items, and a transaction database $DB = \langle T_1, T_2, \ldots, T_n \rangle$, where $T_i (i \in [1, ..., n])$ is a transaction which contains a set of items in I. The support (or occurrence frequency) of a pattern A, where A is a set of items, is the number of transactions containing A in DB. A pattern A is frequent if A's support is no less than a predefined minimum support threshold, ξ. Given a transaction database DB and a minimum support threshold ξ, the problem of finding the complete set of frequent patterns is called the frequent pattern mining problem.

Éclat algorithm. Éclat stands for "equivalence class transformation" and is a depth-first search algorithm that uses the set intersection. It is a naturally elegant algorithm suitable for both sequential as well as parallel execution with locality-enhancing properties. It was first introduced by Zaki *et al.* (1997) in a series of papers written in 1997.

Apriori. Apriori is a seminal algorithm for finding frequent item sets using candidate generation. It is characterized as a level-wise complete search algorithm using the anti-monotonicity of item sets. Thus, "if an item set is not frequent, any of its superset is never frequent." By convention, Apriori assumes that items within a transaction or item set are sorted in lexicographic order (Wu *et al.*, 2008).

The Apriori algorithm was proposed by Agarwal and Srikant in 1994. Apriori is designed to operate on databases containing transactions (e.g. collections of items bought by customers, or details of a website's frequentation). Each transaction is seen as a set of items (an item set). Given a threshold C, the Apriori algorithm identifies the item sets which are subsets of at least C transactions in the database. Apriori uses a "bottom-up" approach, where frequent subsets are extended one item at a time (a step is known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. It generates candidate item sets of length K from item sets of length K−1. Then it prunes the candidates which have an infrequent subpattern. The candidate set contains all frequent K-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

The pseudocode for the algorithm is given below for a transaction database T and a support threshold ofε. Usual set theoretic notation is employed, though it should be noted that T is a multiset. C_k is the candidate set for level k. At each step, the algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma. Count (c) accesses a field of the data structure that represents candidate set c, which is initially assumed to be zero. Many details are omitted below, usually the most important part of the implementation

is the data structure used for storing the candidate sets, and counting their frequencies (Agrawal and Srikant, 1994).

Apriori(T, ε)
$L_1 \leftarrow$ {large 1−item sets}
$k \leftarrow 2$
while $L_{k-1} \neq \varnothing$
$C_k \leftarrow \{a \cup \{b\} | a \in L_{k-1} \wedge b \notin a\} - \{c | \{s | s \subseteq c \wedge |s| = k-1\} \nsubseteq L_{k-1}\}$
for transactions $t \in T$
$C_t \leftarrow \{c | c \in C_k \wedge c \subseteq t\}$
for *candidates* $c \in C_t$
$count[c] \leftarrow count[c] + 1$
$L_k \leftarrow \{c | c \in C_k \wedge count[c] \geqslant \in\}$
$k \leftarrow k + 1$
return $\underset{k}{\cup} L_k$

Based on an analysis of Section 5.3, we implanted all three association rule algorithms in our data set and after analysis comparison, the Apriori algorithm came out as the strongest rule generator compared with other algorithms.

4.4 Performance evaluation phase

At the end of the analysis phase, we obtained different rules from different scenarios with different analysis flows and streams. It was clear that we needed to distinguish the best scenario type and factor stream. So, using a proper performance evaluation technique to assess all the scenario types was inevitable. Meanwhile, we needed to compare the results of our proposed methodology with other available performance analysis techniques. During the performance evaluation phase, rules number, confidence and support values were used for factor evaluation. This phase tried to select the best scenario from a large number of scenario types.

Based on previous sections, to have the most comprehensive and accurate rules, we needed to have enriched and accurate clustering factors for distinguishing better clustering results. The evaluation of predictive accuracy is part of the development process for a model and it is based on specific numerical indicators. It was possible to define reasonable measures of quality and significance for clustering methods. At the end of the ARM step, we had a set of rules and their specifications per scenario type. These specifications were very important criteria for the performance evaluation of a scenario. These criteria are introduced below.

4.4.1 Number of rules. The most important factor of ARM is the number of extracting rules. A bigger rule number for a certain level of preregulated confidence and support showed the accuracy of rule extraction process. In the same condition for all scenario types, the number of extracted rules was measured to define the description of rules in detail.

4.4.2 Elapsed time. One of the most important factors for the performance analysis of a technique or stream was the time section in which a problem was worked out by its solver. Elapsed time is the difference between the beginning time and the finishing time of a process. For this experiment, we calculated and compared the elapsed time of every scenario type.

4.4.3 Confidence and support levels. Confidence and support are two association rule evaluation measures. Let two item sets be $L \subset O$ and $H \subset O$ such that $L \cap H = \varnothing$ and a

data set of T. An association rule is a stochastic concept denoted by L⇒H and it means that if T covers L, then T also covers H with a probability of p, which is termed the confidence of the rule in D, defined as:

$$p = conf\{L \Rightarrow H\} = \frac{f(L \cup H)}{f(L)} \tag{5}$$

The rule L⇒H is said to have a support s in D if the ratio of transactions covering both L and H is equal to s, that is, if:

$$s = \sup\{L \Rightarrow H\} = \frac{f(L \cup H)}{m} \tag{6}$$

The support of a rule expresses the ratio of transactions that contain both the antecedent and the consequent of the rule.

4.5 Comparison and evaluating clusters using NN phase

In order to verify the accuracy of our performance analysis, we have to evaluate performances of the scenarios by using the separate phases of factor evaluation. This phase, which is based on the logic of NN, is an attempt to evaluate the enrichment levels of different factors that are basic elements of extracted rules. This enables us to compare our results. The details of every phase is examined next.

NNs are simple models of the way that the nervous system operates. Their basic units are neurons, which are typically organized into layers. An NN, sometimes called a multilayer perceptron, is basically a simplified model of the way the human brain processes information. It works by simulating a large number of interconnected simple processing units that resemble abstract versions of neurons. The processing units are arranged in layers. There are typically three parts in an NN: an input layer, with units representing the input fields; one or more hidden layers; and an output layer, with a unit or units representing the output field(s). The units are connected with varying connection strengths (or weights). Input data are presented to the first layer, and values are propagated from each neuron to every neuron in the next layer. Eventually, a result is delivered from the output layer. The network learns by examining individual records, generating a prediction for each record and making adjustments to the weights whenever it makes an incorrect prediction. This process is repeated many times, and the network continues to improve its predictions until one or more of the stopping criteria have been met.

To be able to compare the results of analysis and verify the best-selected stream and scenario, we needed to use NNs for factor analysis. The role of NN analysis is to predict the sale rate of every product, based on different types of factor shapes. So at the end of a prediction analysis we should have the importance degree of a certain factor. In other words, when clustering a certain data set by different combinations of attributes, using NNs can help to find the most important factors to predict the rate at which a certain product is preferred by the customers of a mentioned data set. For this step, we tried to find the most important features and attributes that were affecting certain factors. Using this approach led us to measure the effects of different clusters that had been established from different scenarios. If the importance degrees of clusters are high, we can be sure that our clusters are enriched sufficiently; otherwise, the clustering method or attribute combinations that produced the cluster results would have to have been in a wrong sequence.

5. Empirical analysis and results

A web-based extraction of transactions from the database of a global pizza restaurant chain has been presented. The transaction data structure was formed of customer identifications, product purchases and their demographic data. For this empirical analysis, we tried to extract the most accurate and reliable rules and relationships between different data attributes as possible. These rules were useful for better CRM and product recommendations.

5.1 Data structure

First of all, we analyzed the structure of a real-life data set, which was gathered from a global pizza restaurant chain in Turkey. It had more than two and a half million different transactions for customers with a customer identification (ID) number. The data set's other attributes were; dates of orders, city where the customer lived, sex/gender category, age category (obtained from date of birth attribute through data preprocessing), payment or expenditure of the customer per order, and the other attributes showed products. The data set had seven demographic and 49 product attributes. Figure 2 illustrates some of these.

As shown in Table II, all the transactions for a certain customer were combined, calculated and analyzed. As the results show, each customer was identified by an ID and the other attributes presented were, respectively, recency value, frequency value, monetary value, recency score, frequency score, monetary score and averaged RFM score. Thanks to this analysis, the RFM information could be used as a type of extra data attribute for processing.

5.2 Selection of the best segmentation algorithm

At this stage we performed segmentation on the data set, using both k-means and Kohonen (SOM) algorithms, to select the best clustering algorithm for our methodology. The segmentation analysis results for all 2,553,470 transactions, using both K-means and Kohonen algorithms, have been illustrated in Figure 3. Nine clusters were conducted using the Kohonen algorithm and the elapsed time for this analysis was 20 minutes and 25 seconds. By comparison, the k-means algorithm resulted in five clusters (that were precalculated using the Dunn index) that were produced in just 25 seconds. Glancing at Figure 3, it is clear that a cluster result created by Kohonen contains just three records and it showed that the number of clusters had to be adjusted. Obviously, clustering quality and the runtime of the k-means algorithm was more reliable. Based on these factors, adopting the k-means algorithm promised to enhance the superiority of our methodology.

At the clustering stage, the output and quality of the clusters resulting from running the k-means algorithm largely depended on selecting the proper number of clusters. For selecting the optimal number of clusters, there are many indicators such as the Davies-Bouldin index, the silhouette width or the Dunn index. In this work we used the Dunn index. Dunn's index measures compactness (the maximum distance between the data points of clusters) and cluster separation (the minimum distance between clusters). This measurement serves to find the right number of clusters in a data set, where the maximum value of the index represents the right partitioning given the index (Pakhira *et al.*, 2004).

The aim of all these indices is to have meaningful clusters where the data objects within the same cluster are similar to one another and dissimilar to the objects in other clusters. In this research, we evaluated the optimum number as $K = 5$, based on the Dunn Index, for use as the number of clusters in the k-means algorithm.

1	Customer ID	Date of Order	CITY	SEX/Gender	AGE_Category	Date of Birth	Payment	Beverage 1	Beverage 2	Beverage 3	Pizza 1	Pizza 2	Pizza 3
2	154596	2014-03-15 20:56:00	1	1	1.00	1995-01-11	$139.00	null	null	1	null	null	null
3	888917	2014-03-15 20:56:00	1	2	1.00	1983-04-20	$118.00	null	null	null	null	null	null
4	153834	2014-03-15 20:55:00	1	2	1.00	1973-11-29	$409.00	1	null	null	null	1	null
5	546771	2014-03-15 20:55:00	3	2	1.00	1968-03-16	$299.00	null	null	null	null	null	null
6	546771	2014-03-15 20:55:00	3	2	1.00	1968-03-16	$11.00	null	1	null	null	null	null
7	591676	2014-03-15 20:55:00	4	2	2.00	1980-07-04	$222.00	null	1	1	1	3	2
8	591676	2014-03-15 20:55:00	4	2	2.00	1980-07-04	$22.00	null	null	null	null	null	null
9	591676	2014-03-15 20:55:00	4	2	2.00	1980-07-04	$341.00	null	null	null	null	null	null
10	591676	2014-03-15 20:55:00	4	2	2.00	1980-07-04	$59.00	null	null	null	null	null	null
11	922450	2014-03-15 20:55:00	5	2	2.00	1996-05-03	$232.00	null	null	null	null	null	null
12	922450	2014-03-15 20:55:00	5	2	2.00	1996-05-03	$163.00	null	1	null	null	null	null
13	964654	2014-03-15 20:55:00	1	1	2.00	1996-03-16	$264.00	null	1	null	null	1	null
14	964654	2014-03-15 20:55:00	1	1	2.00	1996-03-16	$177.00	null	null	null	null	null	null
15	964654	2014-03-15 20:55:00	1	1	1.00	1996-03-16	$2,426.00	10	12	22	11	45	33
16	874917	2014-03-15 20:55:00	6	2	1.00	1986-02-20	$139.00	null	null	null	null	null	null
17	503203	2014-03-15 20:55:00	2	2	1.00	1994-07-25	$139.00	null	null	null	null	null	null
18	503203	2014-03-15 20:55:00	2	2	1.00	1994-07-25	$39.00	null	null	null	null	1	null
19	273409	2014-03-15 20:55:00	1	1	1.00	1980-09-03	$165.00	null	null	null	null	null	null
20	273409	2014-03-15 20:55:00	1	1	1.00	1980-09-03	$2.00	null	null	null	null	null	null
21	874925	2014-03-15 20:55:00	7	2	1.00	1995-11-01	$163.00	null	1	null	null	null	null
22	874925	2014-03-15 20:55:00	7	2	1.00	1995-11-01	$323.00	12	2	1	3	3	2
23	89169	2014-03-15 20:55:00	1	1	2.00	1989-01-19	$278.00	null	null	null	null	null	null
24	438286	2014-03-15 20:55:00	1	1	2.00	1986-10-16	$139.00	null	null	null	null	null	null
25	552856	2014-03-15 20:55:00	18	2	1.00	1980-06-02	$182.00	null	1	null	null	null	null
26	552856	2014-03-15 20:55:00	18	2	1.00	1980-06-02	$182.00	null	1	null	null	null	null

Figure 2.
A partial capture of a data set including attributes and transactions

ID	Recency	Frequency	Monetary	Recency score	Frequency score	Monetary score	Average RFM score
247968	300	3	669	1	3	4	1,420.0
926746	300	3	4,092	1	3	5	1,620.0
821702	300	3	751	1	3	4	1,420.0
716158	300	1	139	1	1	1	420.0
260212	300	4	4,980	1	4	5	1,820.0
812921	300	1	139	1	1	1	420.0
964319	300	1	139	1	1	1	420.0
24127	300	3	409	1	3	4	1,420.0
649539	300	1	165	1	1	2	620.0
833556	300	3	4,682	1	3	5	1,620.0
945488	300	3	328	1	3	3	1,220.0
39396	300	4	288	1	4	2	1,220.0

Table II.
RFM calculation results for each customer

5.3 Selection of the best ARM algorithm

Apriori (SPSS Clementine v12), Eclat and FP-growth (version 0.99.491 – © 2009-2015 RStudio, Inc.) ARM algorithms were applied to the data set and the results of analysis are illustrated in Table III hereafter. A number of rules, the minimum support percentage, the minimum confidence percentage and elapsed time were comparison criteria. Clearly the performance (minimum level of support and confidence and also shorter elapsed time) of Apriori was more advantageous compared with the other two algorithms. On the other hand, the higher number of rules expressed more detailed association rules.

At the association rule step, we intended to use Apriori with a fixed confidence level of up to 80 and support of up to 20 for all scenario types.

5.4 Performing scenario types

All 42 scenarios in Table I, embedded in the stream are shown in Figure 4, Each scenario type was modeled by SPSS Clementine v12. In all the model streams we used RFM analysis, k-means and Apriori algorithm tools. Based on each scenario type, the tools options were regulated. In this test, we used five beans for RFM with different combinations of the weights one and/or two based on each scenario. Figure 4 illustrates a general model stream from IBM SPSS Modeler.

5.5 Rule evaluation

After performing all the scenario types, we recorded their results as a number of rules, cluster numbers and times elapsed. At that point, it was the turn of the NN to signify all the important attributes which might take the role of the most powerful factors by forecasting the five most-sold products (products 53, 65, 55, 11 and 39). Figure 5 illustrates the stream of the NN-based evaluation. The variable importance percentage shows that using certain scenarios enriched factors better than other scenarios did. For example, as shown in Table VI, for scenario number 23, k-means clusters were more important for forecasting sales amounts of product VAR00065.

5.6 Results of analysis

A real-life customer transaction data set was evaluated by a proposed algorithm to achieve the best customer purchase behavior pattern and the extraction of

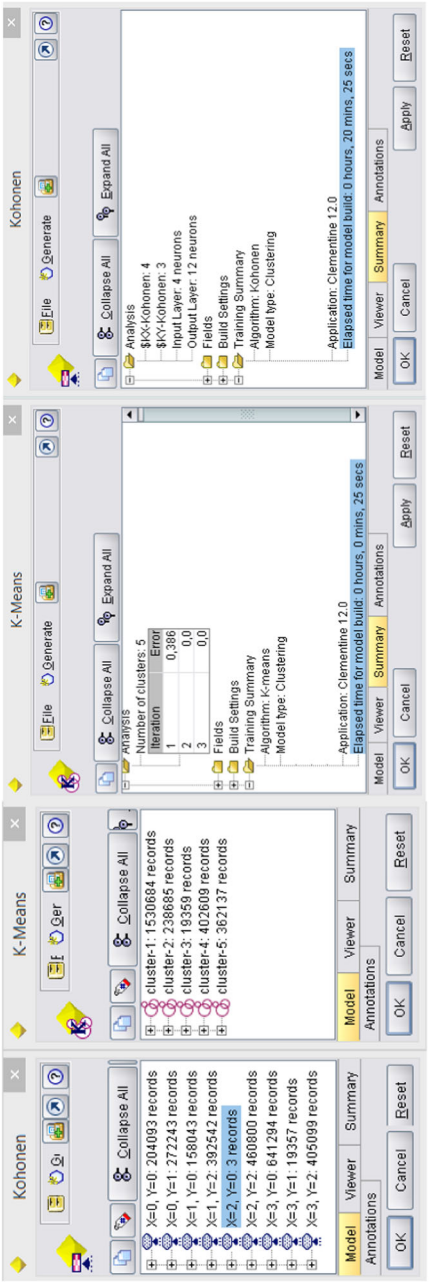

Figure 3.
Comparison of the k-means and Kohonen clustering algorithms targeting the number of clusters and elapsed time

association rules. For this, different series of streams and a variety of attributes were examined as scenario types. All 42-scenario types were applied, executed and evaluated by an Intel Dual Core i7 microprocessor computer and all the possible results are outlined in Table IV. Customer clusters, support and confidence levels were prearranged as the same for all scenario types, however, the number of clusters were reduced to three or two in some scenarios. This was due to the disability of the embedded algorithms in certain scenario types that made it impossible to see the dissimilarities when using the k-means clustering algorithm.

In different scenario types, the number of rules varied between 329 and 2,491 and a higher rule number was desired. Also, the elapsed times ranged from seven to 223 minutes and shorter elapsed times were desired. According to our evaluations since scenario 25 had the most detailed rules with the most appropriate cluster numbers and a shorter elapsed time, it proved to be the best option according to our proposed methodology (i.e. the togetherness of demographic and RFM factors while clustering, and causes more accurate association rules).

A comparison step was completed by using an NN to examine clusters' content effectiveness levels to predict the top-five products sold. All effectiveness levels for the five products that were highly recommended by the rule association algorithm have been recorded. All these recorded effectiveness levels were averaged and are illustrated in Table V. Scenario types 23, 24 and 25 had better effectiveness level averages using NN. The average of scenario type 25 was explicitly the best. As a result, scenario 25 was the best scenario after a comparison of Tables IV and VI. Scenario (stream) 25 emphasized an algorithm by clustering customers not only with their RFM data but with a composition of RFM scores and demographic data. Furthermore, this scenario type showed that monetary was the most important factor for clustering among the RFM analysis results.

In the results of ARM, our best scenario (no. 25) extracted 2,491 rules. These rules were in the shape of antecedents and consequents. The antecedents showed the different factors and purchase behaviors of customers. We had five different clusters composed of RFM analysis outputs and demographic data. (The contents of clusters are shown in Table V.) The consequents indicated those products that had been purchased by customers with those antecedents. In Table V we have shown the partial results of rules as an example, especially for the most purchased products.

6. Benchmarking

To improve the quality of the work, the 25th scenario of our proposed methodology, which concluded the study as the best scenario type, was compared with similar works that have been conducted lately. Different algorithms from similar works were applied to our data set. The results of the proposed methodology and the results of the benchmarks are summarized in Table VII.

	Number of rules	Minimum support %	Minimum confidence %	Elapsed time (Min.)
Apriori	613	71	80.42	14
Eclat algorithm	439	42	75	32
FP-growth algorithm	541	22	70	15

Table III. Comparison of Apriori, Eclat and FP-growth rule mining algorithms

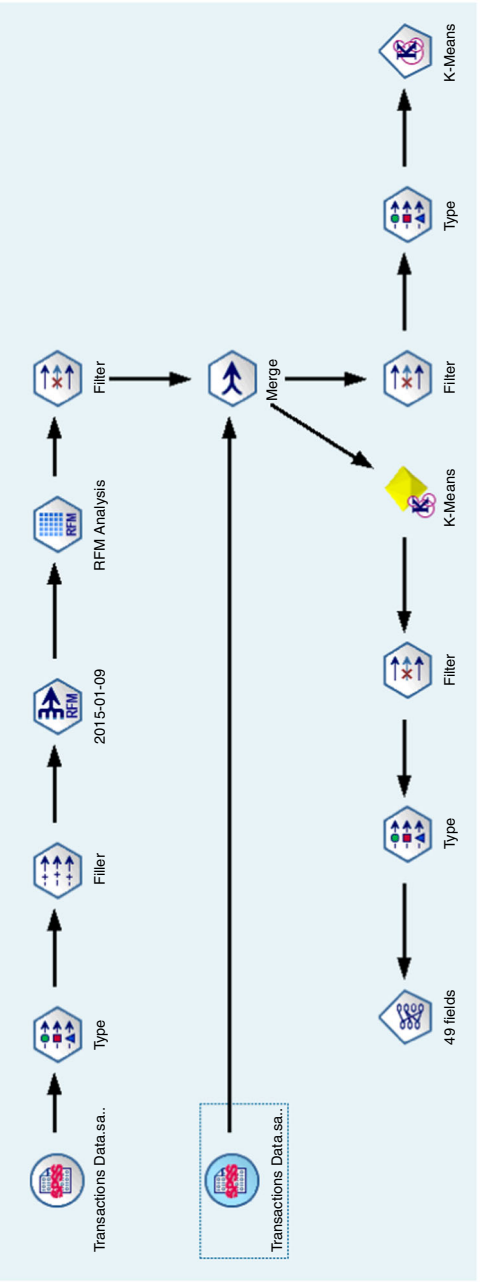

Figure 4.
A general model stream including different combinations of inputs and outputs

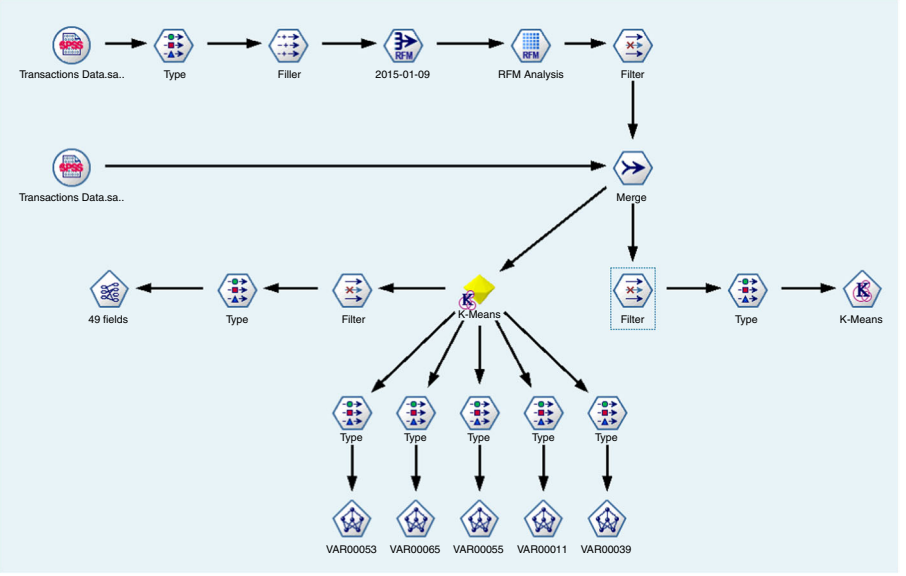

Figure 5.
Model stream including NN prediction flow

Scenario (1-14)	Cluster number	Rule number	Time elapsed (min.)	Scenario (15-28)	Cluster number	Rule number	Time elapsed (min.)	Scenario (29-42)	Cluster number	Rule number	Time elapsed (min.)
1	5	647	127	15	5	517	10	29	2	522	22
2	5	763	119	16	5	517	11	30	5	329	24
3	5	547	113	17	5	517	14	31	5	1,653	27
4	5	547	15	18	5	517	13	32	5	1,747	25
5	5	1,647	220	19	5	517	14	33	5	1,062	25
6	5	1,647	223	20	5	517	15	34	5	1,060	24
7	5	1,647	112	21	5	517	15	35	3	1,056	23
8	5	1,555	35	22	5	517	7	36	4	517	22
9	5	1,555	34	*23**	*5*	*1,517*	*9*	37	4	517	24
10	5	1,555	35	*24**	*5*	*1,517*	*8*	38	4	517	22
11	5	1,555	31	*25***	*5*	*2,491*	*9*	39	4	517	26
12	5	1,555	33	26	5	1,175	7	40	4	517	26
13	5	1,555	33	27	5	613	8	41	4	611	19
14	5	1,555	32	28	5	613	10	42	4	423	18

Notes: *Better scenario; **the best scenario

Table IV.
Performance results of different scenarios

Dursun and Caber (2016) focussed on profiling profitable hotel customers by RFM analysis in their work (Dursun and Caber, 2016). They used equiponderance RFM scores ($W_R = W_F = W_M$). We used their approach which was the same as that which we used for the segmentation phase of our eighth scenario. Ghodousi *et al.* (2016) gathered an information set from a municipal district in a city and analyzed it to prioritize urban needs as they were assessed for citizen satisfaction (Ghodousi *et al.*, 2016). In total, 43 citizen needs were identified and categorized using k-means clustering and the fuzzy clustering method based on equally weighted ($W_R = W_F = W_M$) RFM values. Margianti *et al.* (2016) used clustering by the affinity propagation and RFM model on 1,000

customers' data in their research. They used the distance method of affinity propagation algorithm with two Euclidean distances and the Manhattan distance techniques to facilitate companies' analysis of customer transaction data (Margianti *et al.*, 2016; Nazari and Ersoy, 1992).

It is clear that the elapsed time (nine minutes), the clustering factor's effectiveness level (0.284) and the number of rules (2,491) for our proposed method, made it undeniably better than other approaches. Figure 6 presents the results for clustering performances graphically, adopting the algorithms of benchmarks for our data set using the MATLAB program.

7. Implications

Modern marketing methods use multiple criteria to group clients, targeted campaigns and publicity messages for products and services (Nettleton, 2014). Obtaining information about customer characteristics, and analyzing the previously recorded data for generating association rules and CRM strategies, has become vital for companies due to global competition. DM analysis enables managers to discover the potential demands of targeted customers by the observation of their past behaviors. The current study emphasizes clustering, which is a very useful DM tool for reducing the complexity of very large transactional data sets. On this road, clustering's accuracy is another important factor that influences the effectiveness of association rules directly. To reach an acceptable level of cluster accuracy, RFM analysis, in particular, has added extra factors to customer clustering analysis. As customer clustering's effectiveness levels are very important, one may need to have more attributes to make the resulting clusters as enriched as possible. Demographic data from customers are easy to use and important factors that are simply accessible for companies. Interestingly, the demographic data from customers are improving clusters' effectiveness levels and the quality of extracted rules as these data are integrated with RFM score factors.

In general, companies have used these association rules to make offers and campaigns tailored for their current or potential customers (which is known as "gamification"). Irrelevant, wrong offers or messages sent too often to customers will decrease the impression made by campaigns, causing financial and customer losses. By using the proposed methodology for customer clustering and with consequent ARM, relevant and true messages can be delivered to target customers with greater confidence and more precision.

Table V. Deriving significant rules to identify certain product consumers from scenario no. 25, which is the best scenario (from among 2,491 samples)

Rule	Antecedent	Consequent (product)	Support %	Confidence %
1	Cluster = 1, beverage 2 = yes, pizza 3 = yes	Var00053 = yes	89.8	99.9
2	Cluster = 2, beverage 23 = yes, pizza 11 = yes	Var00065 = yes	83.0	98.7
3	Cluster = 5, beverage 11 = yes, pizza 2 = yes	Var00055 = yes	82.11	97.1
4	Cluster = 5, beverage 7 = yes, pizza 8 = yes	Var00011 = yes	69.21	99.2
5	Cluster = 4, beverage 7 = yes, pizza 2 = yes	Var00039 = yes	70.43	80.42
Cluster 1 = (R score = 4, F score = 1, M score = 1, City = 5, Sex Cat. = 2, Age Cat. = 2)				
Cluster 2 = (R score = 5, F score = 2, M score = 2, City = 62, Sex Cat. = 1, Age Cat. = 3)				
Cluster 5 = (R score = 5, F score = 2, M score = 2, City = 12, Sex Cat. = 1, Age Cat. = 3)				
Cluster 5 = (R score = 5, F score = 2, M score = 2, City = 1, Sex Cat. = 2, Age Cat. = 1)				
Cluster 4 = (R score = 5, F score = 2, M score = 2, City = 43, Sex Cat. = 2, Age Cat. = 3)				

Scenario	Effectiveness level for product 0065	Effectiveness level for product 0055	Effectiveness level for product 0011	Effectiveness level for product 0039	Effectiveness level for product 0053	Averages of effectiveness levels
1	0.0447	0.0465	0.0597	0.1123	0.0903	0.0707
2	0.0601	0.047	0.0632	0.0489	0.0865	0.06114
3	0.016	0.1319	0.0449	0.0772	0.0357	0.06114
4	0.00402	0.0218	0.0218	0.0202	0.0176	0.017084
5	0.0132	0.0906	0.0494	0.00137	0.00121	0.031156
6	0.0097	0.0188	0.0315	0.032	0.018	0.022
7	0.0449	0.0215	0.0284	0.0307	0.02344	0.029788
8	0.0056	0.0307	0.035	0.0615	0.0519	0.03694
9	0.00899	0.0466	0.0186	0.0349	0.0447	0.030758
10	0.00148	0.0187	0.0373	0.0429	0.0467	0.029416
11	0.000817	0.0245	0.0337	0.0381	0.036	0.0266234
12	0.00192	0.016	0.0278	0.035	0.0504	0.026224
13	0.024	0.028	0.0265	0.0468	0.0477	0.0346
14	0.00374	0.0522	0.0448	0.03	0.0393	0.034008
15	0.0266	0.0144	0.0301	0.0146	0.0242	0.02198
16	0.0144	0.0286	0.021	0.0232	0.0186	0.02116
17	0.0184	0.0264	0.0267	0.0126	0.0165	0.02012
18	0.0186	0.0192	0.0292	0.0206	0.0353	0.02458
19	0.062	0.098	0.0133	0.0302	0.0248	0.04566
20	0.0915	0.026	0.0233	0.0336	0.0254	0.03996
21	0.0122	0.0107	0.0281	0.0242	0.0218	0.0194
22	0.0306	0.0379	0.0165	0.0274	0.0253	0.02754
23*	*0.408*	*0.0319*	*0.0129*	*0.0221*	*0.195*	*0.13398*
24*	*0.143*	*0.0306*	*0.0145*	*0.0307*	*0.424*	*0.12856*
25**	*0.148*	*0.0452*	*0.177*	*0.0294*	*0.284*	*0.13672*
26	0.003	0.0218	0.00771	0.0122	0.0239	0.013722
27	0.00554	0.0299	0.00137	0.034	0.0242	0.019002
28	0.00162	0.0331	0.0197	0.0184	0.0289	0.020344
29	0.021	0.00119	0.0528	0.0062	0.0124	0.018718
30	0.00158	0.0232	0.0478	0.0467	0.0588	0.035616
31	0.036	0.0389	0.0379	0.0242	0.135	0.0544
32	0.0413	0.113	0.0326	0.0194	0.1058	0.06242
33	0.0116	0.0222	0.0307	0.0218	0.061	0.02946
34	0.0197	0.0228	0.0254	0.0298	0.0147	0.02248
35	0.0681	0.0117	0.017	0.0218	0.115	0.04672
36	0.0633	0.0225	0.0319	0.00842	0.0219	0.029604
37	0.00172	0.0281	0.0145	0.0319	0.02	0.019244
38	0.0494	0.0115	0.0347	0.035	0.0154	0.0292
39	0.0136	0.0262	0.011	0.0228	0.014	0.01752
40	0.0201	0.0137	0.0312	0.016	0.0148	0.01916
41	0.0246	0.016	0.01	0.0248	0.0171	0.0185
42	0.0213	0.0173	0.0197	0.0226	0.0183	0.01984

Table VI.
Analysis results of clusters' enrichment levels (effectiveness levels) using NN for the five most-sold products

8. Conclusion

Businesses are always trying to obtain trustable and convenient methods of association rule extraction using customer transaction data to capture a better quality of CRM. They are driven to try to recognize the needs of their societies and their customers more precisely. So, to distinguish the needs of the mass of their customers and create some meaningful interaction between producers and consumers, they have

	Segmentation approach	Clustering tool	Number of clusters	Number of rules	Effectiveness level for product Var.53	Elapsed time
Dursun and Caber (2016)	(RFM scores) $W_R = W_F = W_M$	K-means	5	1,555	0.0519	35
Ghodousi *et al.* (2016)	(RFM values) $W_R = W_F = W_M$	K-means	5	517	0.0465	127
		Fuzzy K-means	16+	1,240	0.195	215
Margianti *et al.* (2016)	(RFM values) $W_R = W_F = W_M$	Affinity propagation – Euclidean distance	4	1,804	0.0097	9.8
		Affinity propagation – Manhattan distance	5	1,625	0.0307	7.7
Proposed scenario	RFM score +demographic data	K-means	5	2,491	0.284	9

Table VII. Summarized benchmarking results

made efforts to understand customers by their behaviors and transactions. To achieve this, they have to sort through a large volume of their customers' data to identify where to find competitive advantages. In this vein, we have sought to examine the best kind of customer segmentation approach that is based on RFM and demographic attributes. Challenging the impacts of RFM and demographic attributes for enriching customer segments is a result of the clustering of factors, and the extraction of more reliable association rules was the main purpose of this study.

In this study we used transactions data from a global pizza restaurants chain to evaluate extracted rules performance. As part of this evaluation, the impacts of demographic attributes for enriching customer segmentation factors have been challenged. Different types of scenario have been designed, performed and evaluated meticulously under uniform test conditions. An effective methodology based on a comprehensive scenario design stage was inaugurated. The results were evaluated and compared using efficient machine learning techniques. All the test results were recorded and summarized. This study indicates that the weights of RFM attributes affect rule association performance positively. Moreover, to gain more accurate customer segments, a combination of WRFM and demographic attributes is recommended. Accordingly, the proposed methodology resulted in the best outcomes and scores, so the importance of WRFM and demographic data in the clusters has been proved. Although we have used limited numbers of demographic data as attributes, the results' analysis indicates the undeniable importance of demographic factors merged with RFM data. Using more accurate and stronger rules, marketing managers can prepare more targeted and effective campaigns based on accurate customer segments. For future studies, using an expanded scope of attributes, we will use the firmographic and behavioral data of customers in the phase of input enrichment. Moreover, in the solving phase, considering novel machine learning concepts will play a greater role in combining different clustering and rule extraction techniques. Thus, the present methodology will be enhanced by using metaheuristic clustering and association rule algorithms that can improve training speeds, elapsed times and the ability of DM tools to extract more powerful and more comprehensive rules.

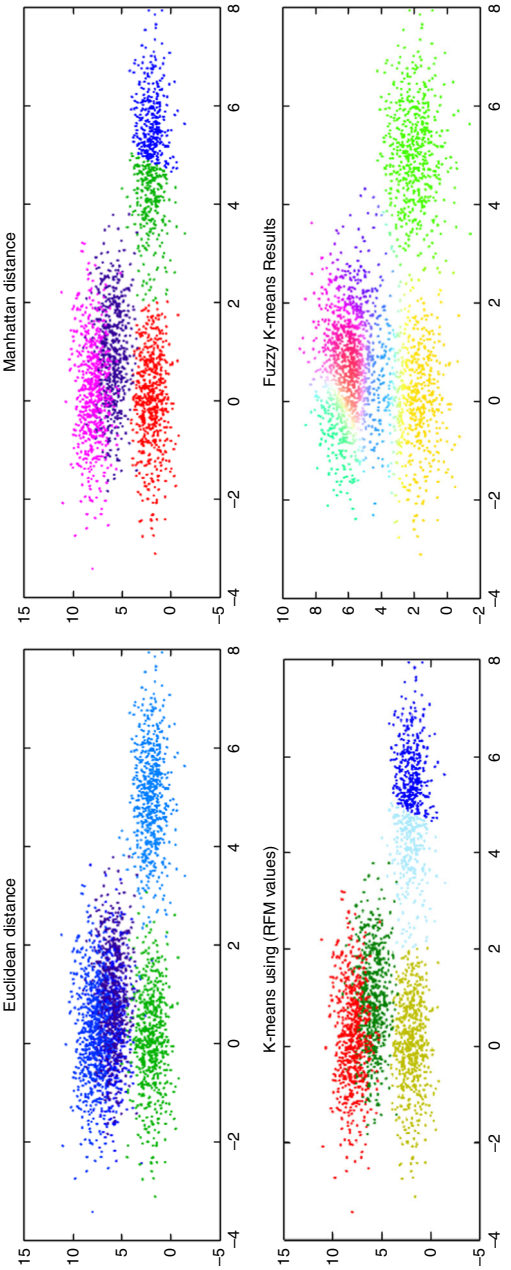

Figure 6.
Graphical output of clustering using benchmark techniques

References

Agrawal, R. and Srikant, R. (1994), "Fast algorithms for mining association rules in large databases", *in VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases, Morgan Kaufmann*, pp. 487-499.

Agrawal, R., Imielinski, T. and Swami, A. (1993), "Mining association rules between sets of items in large databases", *ACM SIGMOD Record*, Vol. 22, May, pp. 207-216, available at: http://doi.org/10.1145/170036.170072.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. (1996), "Fast discovery of association rules", in Usama, R.U., Fayyad, M., Piatetsky-Shapiro, G. and Smyth, P. (Eds), *Advances Knowledge Discovery and Data Mining*, Vol. 12, AAAI, pp. 307-328, available at: www.cs.helsinki.fi/hannu.toivonen/pubs/advances.pdf

Ambler, T., Bhattacharya, C.B., Edell, J., Keller, K.L., Lemon, K.N. and Mittal, V. (2002), "Relating brand and customer perspectives on marketing management", *Journal of Service Research*, Vol. 5 No. 1, pp. 13-25, doi: 10.1177/1094670502005001003.

Bayardo, R.J. (1998), "Efficiently mining long patterns from databases", *ACM SIGMOD Record*, Vol. 27 No. 2, pp. 85-93, doi: 10.1145/276305.276313.

Birant, D. (2003), "Knowledge-oriented applications in data mining", in Funatsu, D. (Ed.), *Knowledge-Oriented Applications in Data Mining*, InTech Published, Shanghai, pp. 91-108.

Blattberg, R.C., Kim, B.-D. and Neslin, S.A. (2008), *Database Marketing: Analyzing and Managing Customers*, Springer, New York.

Brin, S., Motwani, R. and Silverstein, C. (1997), "Beyond market baskets: generalizing association rules to correlations", *ACM SIGMOD Record*, Vol. 26 No. 2, pp. 265-276, doi: 10.1145/253262.253327.

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, pp. 2754-2762, doi: 10.1016/j.eswa.2007.05.043.

Chen, Y.-L., Kuo, M.-H., Wu, S.-Y. and Tang, K. (2009), "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data", *Electronic Commerce Research and Applications*, Vol. 8 No. 5, pp. 241-251, doi: 10.1016/j.elerap.2009.03.002.

Cheng, C.H. and Chen, Y.S. (2009), "Classifying the segmentation of customer value via RFM model and RS theory", *Expert Systems with Applications*, Vol. 36 No. 3/Part 1, pp. 4176-4184, doi: 10.1016/j.eswa.2008.04.003.

Chung, Y.-C. and Chen, S.-J. (2016), "Study on customer relationship management activities in Taiwan tourism factories", *Total Quality Management & Business Excellence*, Vol. 27 Nos 5-6, pp. 581-594, doi: 10.1080/14783363.2015.1019341.

Claycamp, H.J. and William, F.M. (1968), "A theory of market segmentation", *Journal of Marketing Research*, Vol. 5, pp. 388-394.

Coussement, K., Van den Bossche, F.A.M. and De Bock, K.W. (2014), "Data accuracy's impact on segmentation performance: benchmarking RFM analysis, logistic regression, and decision trees", *Journal of Business Research*, Vol. 67 No. 1, pp. 2751-2758, doi: 10.1016/j.jbusres.2012.09.024.

Demartines, P. and Blayo, F. (1992), "Kohonen self-organizing maps: is the normalization necessary?", *Complex Systems*, Vol. 6 No. 2, pp. 105-123.

Dursun, A. and Caber, M. (2016), "Using data mining techniques for profiling profitable hotel customers: an application of RFM analysis", *Tourism Management Perspectives*, Vol. 18, May, pp. 153-160, available at: http://doi.org/10.1016/j.tmp.2016.03.001

Fallis, A. (2013), *Data Mining Concepts and Techniques. Journal of Chemical Information and Modeling*, Vol. 53, Elsevier, pp. 1689-1699, available at: http://doi.org/10.1017/CBO9781107415324.004

Ghodousi, M., Alesheikh, A.A. and Saeidian, B. (2016), "Analyzing public participant data to evaluate citizen satisfaction and to prioritize their needs via K-means, FCM, and ICA", *Cities*, Vol. 55 No. 1, pp. 70-81, available at: http://doi.org/10.1016/j.cities.2016.03.015

Han, J.H.J., Dong, G.D.G. and Yin, Y.Y.Y. (1999), "Efficient mining of partial periodic patterns in time series database", *Proceedings 15th International Conference on Data Engineering (Cat. No. 99CB36337)*, pp. 1-10, doi: 10.1109/ICDE.1999.754913.

Hiziroglu, A. (2013), "Soft computing applications in customer segmentation: state-of-art review and critique", *Expert Systems with Applications*, Vol. 40 No. 16, pp. 6491-6507, doi: 10.1016/j.eswa.2013.05.052.

Hosseini, Z.Z. and Mohammadzadeh, M. (2016), "Knowledge discovery from patients' behavior via clustering-classification algorithms based on weighted eRFM and CLV model: an empirical study in public health care services", *Iranian Journal of Pharmaceutical Research*, Vol. 15 No. 1, pp. 355-367.

Hu, Y.H. and Yeh, T.W. (2014), "Discovering valuable frequent patterns based on RFM analysis without customer identification information", *Knowledge-Based Systems*, Vol. 61 No. 3, pp. 76-88, available at: http://doi.org/10.1016/j.knosys.2014.02.009

Hughes, A.M. (1994), *Strategic Database Marketing*, Probus Publishing Company, Chicago.

Kahan, R. (1998), "Using database marketing techniques to enhance your one-to-one marketing initiatives", *Journal of Consumer Marketing*, Vol. 15 No. 5, pp. 491-493, doi: 10.1108/07363769810235965.

Khajvand, M., Zolfaghar, K., Ashoori, S. and Alizadeh, S. (2011), "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study", *Procedia Computer Science*, Vol. 3 No. 1, pp. 57-63, available at: http://doi.org/10.1016/j.procs.2010.12.011

McCarty, J.A. and Hastak, M. (2007), "Segmentation approaches in data-mining: a comparison of RFM, CHAID, and logistic regression", *Journal of Business Research*, Vol. 60 No. 6, pp. 656-662, doi: 10.1016/j.jbusres.2006.06.015.

Mannila, H., Toivonen, H. and Verkamo, A.I. (1994), "Efficient algorithms for discovering association rules", *AAAI Workshop on Knowledge Discovery in Databases KDD94*, Vol. 118 Nos 1-4, pp. 181-192, available at: http://ukpmc.ac.uk/abstract/CIT/21659.

Mannila, H., Toivonen, H. and Verkamo, A.I. (1997), "Discovery of frequent episodes in event sequences", *Data Mining and Knowledge Discovery*, Vol. 1 No. 3, pp. 259-290, doi: 10.1023/A:1009748302351.

Marcus, C. (1998), "A practical yet meaningful approach to customer segmentation", *Journal of Consumer Marketing*, Vol. 15 No. 5, pp. 494-504, doi: 10.1108/07363769810235974.

Margianti, E.S., Refianti, R., Mutiara, A.B., Nuzulina, K., Technology, I. and Technology, I. (2016), "Affinity propagation and RFM-model for CRM's data analysis", *Journal of Theoretical and Applied Information Technology*, Vol. 84 No. 2, pp. 272-282.

Namvar, M., Gholamian, M.R. and Khakabi, S. (2009), "Electronic business model selection based on firm's ıntellectual capital", *The 8th International Conference on E-Business*, pp. 132-140.

Nazari, J. and Ersoy, O.K. (1992), "Electrical and computer engineering implementation of back-propagation neural networks with MatLab, ECE Technical Reports, Paper 275, available at: http://docs.lib.purdue.edu/ecetr/275.

Nettleton, D. (2014), "Chapter 13 – CRM – customer relationship management and analysis", in Dierna, A. (Ed.), *Commercial Data Mining*, pp. 195-208, Morgan Kaufmann, Waltham, MA, available at: http://doi.org/10.1016/B978-0-12-416602-8.00013-3

Ngai, E.W.T., Xiu, L. and Chau, D.C.K. (2009), "Application of data mining techniques in customer relationship management: a literature review and classification", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 2592-2602, doi: 10.1016/j.eswa.2008.02.021.

Nichol, M.B., Knight, T.K., Dow, T., Wygant, G., Borok, G., Hauch, O. and O'Connor, R. (2008), "Fast algorithms for mining association rules", *The Annals of Pharmacotherapy*, Vol. 42 No. 1, pp. 62-70, doi: 10.1.1.40.7506.

Pakhira, M.K., Bandyopadhyay, S. and Maulik, U. (2004), "Validity index for crisp and fuzzy clusters", *Pattern Recognit*, Vol. 37, pp. 487-501.

Power, C. and Elliott, J. (2006), "Cohort profile: 1958 British birth cohort (national child development study)", *International Journal of Epidemiology*, Vol. 35 No. 1, pp. 34-41, doi: 10.1093/ije/dyi183.

Razieh, Q., Baqeri-Dehnavi, M., Minaei-Bidgoli, B. and Amooee, G. (2012), "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, pp. 172-181, available at: www.tjmcs.com

Ryder, N.B. (1965), "The cohort as a concept in the study of social change", *American Sociological Reviewl Review*, Vol. 30 No. 6, pp. 843-861, available at: http://doi.org/10.2307/2090964

Silverstein, C., Ullman, J., Brin, S. and Motwani, R. (1998), "Scalable techniques for mining causal structures", *Data Mining and Knowledge Discovery*, Vol. 24 No. 146, pp. 594-605.

Smith, W.R. (1956), "Product differentiation and market segmentation as an alternative marketing strategy", *Journal of Marketing*, Vol. 21, pp. 3-8.

Soltani, Z. and Navimipour, N.J. (2016), "Customer relationship management mechanisms: a systematic review of the state of the art literature and recommendations for future research", *Computers in Human Behavior*, Vol. 61 No. 1, pp. 667-688, available at: http://doi.org/10.1016/j.chb.2016.03.008

Stone, B. (1995), *Successful Direct Marketing Methods*, NTC Business Books, Lincolnwood, IL, pp. 37-57.

Teichert, S. (2008), "Impact of ammonium permeases mepA, mepB, and mepC on nitrogen-regulated secondary metabolism in Fusarium fujikuroi", *Eukaryot Cell*, Vol. 7 No. 2, pp. 187-201.

Ultsch, A. and Siemon, H.P. (1990), "Kohonen's self organizing feature maps for exploratory data analysis", *Proceedings of the International Neural Network Conference (INNC-90)*, pp. 305-308.

Vesanto, J. and Alhoniemi, E. (2000), "Clustering of the self-organizing map", *IEEE Transactions on Neural Networks*, Vol. 11 No. 3, pp. 586-600, doi: 10.1109/72.846731.

Wu, X., Kumar, V., Ross, Q.J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008), "Top 10 algorithms in data mining", *Knowledge and Information Systems*, Vol. 14 No. 1, pp. 1-37, available at: http://doi.org/10.1007/s10115-007-0114-2

Xu, L. and Chu, H. (2015), "Big Data analytics toward intelligent mobile service provisions of customer relationship management in e-commerce.

Yu, H.-H., Chen, C.-H. and Tseng, V.S. (2011), "Mining emerging patterns from time series data with time gap constraint", *International Journal of Innovative Computing, Information and Control*, Vol. 7 No. 9, pp. 5515-5528.

Yusof, N. and Kraak, M. (2015), "Mining sequential pattern of multi-dimensional wind profiles", *GeoComputation*, Dallas, Texas, pp. 395-400.

Zaki, M.J., Parthasarathy, S., Ogihara, M. and Wei, L. (1997), "New algorithms for fast discovery of association rules", *3rd International Conference on Knowledge Discovery and Data Mining*, pp. 283-286, available at: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.5143

Further reading

Smith, W. (1995), "Product differentiation and market segmentation as alternative", *Marketing Management*, Vol. 4 No. 3, pp. 63-65.

About the authors

Peiman Alipour Sarvari is a PhD Candidate researching Industrial Engineering at the Istanbul Technical University. He graduated with a MS from the Gazi University in 2013. His current fields of interest include data mining, supply chain management and logistics.

Alp Ustundag was born in 1977. He graduated from the Industrial Engineering Department of the Istanbul Technical University (ITU) in 2000. He received his MBA Degree from the Boğaziçi University in 2002 and his Doctoral Degree from the ITU in 2008. He conducted research studies at the Logistics Department at the Dortmund University, Germany, in 2007. He became an Associate Professor at the ITU in 2011. He is currently the Head of the Radio-frequency Identification (RFID) Research and Test Laboratory at the ITU. He has conducted substantial research and consulting projects in reengineering, logistics and supply chain management for major Turkish companies. His current research interests include RFID, supply chain and logistics management, innovation and technology management, risk management, IT/IS systems, soft computing and optimization. He has published many papers in international journals and presented various studies at national and international conferences. Alp Ustundag is the corresponding author and can be contacted at: ustundaga@itu.edu.tr

Hidayet Takci is an Academician at the Cumhuriyet University, Sivas, Turkey. He has worked in fields such as data mining, text mining, machine learning and security. Hitherto, he has also lectured for courses in data mining and applications, text mining, neural networks, etc. He has contributed to many papers and projects in the field of data mining, text mining, information retrieval and web security.