A review of data mining methods in RFM-based customer segmentation

E Ernawati[1,*], S S K Baharin[2] and F Kasmin[2]

[1] Department of Informatics, Universitas Atma Jaya Yogyakarta, Indonesia
[2] Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia

*ernawati@uajy.ac.id

Abstract. Data mining (DM) is the process of extracting knowledge from data. Knowledge from customer behaviour segmentation is useful for companies in setting the target market and developing a marketing strategy. Recency Frequency Monetary (RFM) model is the most behaviour segmentation used. Many customer-segmentation studies in various application areas use the RFM model that collaborates with DM. With many methods in DM, the selection of appropriate methods can reveal useful hidden patterns in customer segments. This paper aims to analyse DM methods that collaborate with the RFM model and synthesize them to propose a customer segmentation framework. This study uses a comprehensive literature review published in 2015-2020. The most widely used methods are clustering and visualization from seven DM methods analysed. Due to the increased visualization function and the need for customers' geo-demographic data to be considered in the analysis, this study presents a new framework for using DM methods with the RFM based segmentation in the Geographic Information Systems (GIS) environment. This framework helps analysts utilize DM methods to uncover and understand customer characteristics, so companies can set the target market and develop a marketing strategy to increase their competitive advantage.

1. Introduction

Customers are individuals or businesses that purchase products or services from a company. Marketing strategies are applied to increase customer satisfaction and loyalty. However, with so many clients, not all plans are suitable for everyone. Customer segmentation (CS) can be chosen to divide customers into smaller homogeneous groups so that the marketing strategy can target each group individually. CS helps the company target valuable customers and develop marketing activities.

The most popular methods for segmenting customers are data mining (DM) and customer behaviour analysis [1]. The RFM (Recency Frequency Monetary) model is the most widely applied approach to customer behaviour analysis [1–3]. By adopting IT, businesses gather a large amount of data, so generating meaningful information from data is essential. DM methods can identify the significant hidden trends and the relationships inside the data [4]. CS-based on DM has the superiority of improving accuracy over traditional classification [5]. Each DM method has its utility according to its nature, so understanding its function in CS helps researchers and practitioners choose the right one.

Previous literature reviews on RFM model applications [6] and the use of RFM analysis concepts in DM methods [7] were conducted. The survey on DM methods in market segmentation [8] and review of classification and clustering methods in CS [9] are also available. Gončarovs [10] classified data analytics techniques in customer relationship management into seven DM techniques. However, there is no literature review specifically addressing the collaboration of DM methods and RFM-based models in CS. This study aims to survey and organize references that use RFM-based models and DM methods for CS into seven DM methods [10]. Then propose a framework for using DM methods in RFM-based CS.

2. Methodology

This study's literature search comes from three databases: Scopus, Web of Science (WoS), and Emerald. Search keywords used are: 'customer segmentation' OR 'market segmentation' AND 'data mining' AND 'RFM' within 2015-2020. Two hundred and ninety-seven articles from the journals and conference proceedings were compiled. This study's first step is to eliminate duplication then screen the papers’ abstract for CS-related research using RFM-based models and DM methods. Subsequently, reviewing the full-text articles using the following selection criteria: 1). English paper, 2). CS shall use RFM or modified-RFM model and DM methods, 3). Segmentation is based on actual empirical data or a real case study in a particular application area. Based on the inclusion criteria, this study involved 44 articles to be reviewed and classified.

3. Results and discussion

3.1. RFM-based model customer segmentation

CS separates customers into smaller, homogeneous, distinct, and specific customer groups based on customer features. First introduced in 1956 by Smith, CS is now a key and popular marketing activity. Many companies use it to gain a deeper understanding of their customers' characteristics and needs [11]. The CS study has been applied to many application areas such as retail [2,11–13], e-business [14–18], wholesale [19], Small and Medium Enterprise (SMEs) [20], as well as in the field of financial service [21–26], health care service [27,28], IT [29–32], and others. The majority of applications were applied to commercial organizations, except [33] in non-commercial institutions.

Variables used for grouping are an essential aspect of CS [29]. Various types of segmentation exist, such as geographic, demographic, behavioural, and psychographic. RFM analysis is a common behaviour segmentation to explain customer purchase behaviour [34] based on transaction data. Recency (the novelty of customer relations with the company), frequency (the number of purchases made by customers), and monetary (the amount of money paid by customers in a specific period) are three variables in the original RFM model. The valued customers have the highest frequency and monetary value and the lowest recency [1].

The RFM scoring process for quantifying the behaviour of customers uses the quintile method. The first quintile with the highest values (the smallest for recency) marked as 5. The next quintile marked as 4, and so on. Finally, all customers are presented by 555, 554, 553, ...,112, 111. The most valuable customer group is 555, whereas the worst customer group is 111. Other approaches using the actual values of each RFM factor from customer data transactions. Typically, this approach continues with the normalization process to obtain a specific score. Some researchers used hard coding by discretizing the exact value of each RFM-based model factor following the predetermined range categories.

To segment the customers, the majority of studies used clustering (88%); the rest used CLV (Customer Lifetime Value), such as using RFM ranking scores [30] and calculating CLV values based on weighted RFM [35]. Grouping based on clustering has more accuracy than the CLV method [27]. The CLV value is a single RFM score counted by multiplying each RFM value and its weight. The weight determination for each variable depends on the factor’s importance in the application [4,11]. In addition to using equal weight [11,25,33], some researchers used other methods such as the Analytical Hierarchy Process (AHP) [1,14,19,20,23,27,36,37], Fuzzy AHP [31,38], Fuzzy Analytical Network Process (F-ANP) [26], and Entropy method [2,14].

From 44 references reviewed, 57% of them used the original RFM model as a segmentation variable. This model widely used because it uses fewer segmentation variables [19], making it simple, easy to implement, and easy to understand by managers and decision-makers [11,39]. In addition to the original RFM, various modifications to the RFM model used as segmentation variables. Change of the RFM model was done by redefining one or more of the RFM variables or adding new variables [27], sometimes excluding some variables [11]. The change is adapted to the application context, the nature of the product/service, or the contribution of consumers to the company's value [11,38,40]. In the insurance sector, modifications to the RFM model carried out by adding factors related to risk [24], diversity of types of insurance purchased [25], time, number, and amount of expenses of an insurance claim [26]. In e-commerce CS, the RFM model's modification is done by entering users' satisfaction with e-commerce websites and customer activities in e-commerce [14]. The specific characteristics of e-commerce users, such as the frequency of browsing, the number of categories browse, the positive feedback rate, the forwarding rate indicator [16] also included. In non-profit institutions, the monetary factor of the RFM model is not a critical variable to consider, so it can be replaced by other variables that are more important to describe the user's characteristics. For the segmentation of library users in universities, the M (Monetary) factor in the RFM model is replaced by C (College), which indicates the student's college [33]. Despite its advantages, the RFM model does not pay attention to the customer's personal and demographic information [1]. Yoseph and Heikkila [12], Beheshtian-Ardakani et al. [18], Akhondzadeh-Noughabi and Albadvi [39] added demographic data (such as age, gender, nationality), Ardakani et al. [18] also added geographic data (city). The demographic factor is an essential segmentation variable merged with the RFM model [39].

3.2. Data mining methods in RFM-based model customer segmentation

DM is a process for discovering meaningful patterns, trends, and knowledge from large data sets. In this study, the DM methods used together with RFM-based CS are classified into seven groups, i.e., clustering, association, sequence discovery, classification, forecasting, regression, and visualization.

Clustering is the process of partitioning observations or cases into clusters of similar objects. Objects in a cluster are like one another, while objects in other clusters are dissimilar. In CS, clustering acts as an executor to group customers into clusters based on their characteristics. In this study, K-means is the most widely used clustering algorithm. K-means is popular in the application areas because it is simple to understand, interpret, and apply [2]. Some K-means modifications to address its shortcomings, such as determining the number of clusters and initial cluster centres, are also used, among others: K-means++ [23], Bisecting K-means [15], Improve K-means [41]. For handling data uncertainty, Fuzzy C-means is used [3,23,24,31,37]. The combination of SOM (Self-Organizing Maps) and K-means is often used. SOM is used to determine the optimum number of clusters (k), whereas K-means conduct the segmentation [4,17,32,42]. The optimum number of clusters is a crucial consideration in clustering. A validity index has been used to determine the optimum number of clusters. The most widely used validation indices are the Silhouette, the Davies Bouldin index, and the Dunn index. Based on observations, the number of segments chosen was between 2-10 groups, whereas 4-5 segments were the most applied. The number of segments applied is not too large because it will be difficult for the marketing analyst to interpret them and design marketing strategies for selected customer segments.

Association Rule Mining (ARM) is a set of algorithms used to uncover rules for discovering interesting associations and correlations between items that exist together. In CS, ARM generates rules that describe frequently occurring patterns in a particular customer segment and discover the relationship between items so that customer group characteristics are better known. Apriori is a widely used algorithm for extracting association rules [18,30,35,39]. The sequential discovery extracts interesting subsequence from data sequences. Customers' dynamic behaviour can be detected by analysing sequential patterns in customer behaviour from segmented clusters and observing customer segments' changes. Using the Generalized Sequential Pattern (GSP) algorithm, dominant patterns of customer shifts between different groups can provide a good insight into customer changes between segments [29].

Classification is the process of finding a model for classifying data into a predetermined class based on certain criteria. In the segmentation of customers, the classification is used to predict future customer behaviour in segmented customer groups. A decision tree is a classification structure commonly used in customer segments to generate rules [43]. Some of the Decision Tree classification algorithms used are C4.5 [15], CART [24], J48 [28], CHAID [27], C5.0 [18]. The customer segments obtained from the clustering results become leaf nodes of the decision tree, while the internal nodes represent a test based on the segmentation variables used for grouping. Besides prediction, the classification model is also used for insight and profiling. The rules extracted from the decision tree can be used to predict future customers' behaviour.

In CS, forecasting is used to predict future segmented customer behaviour. Forecasting is the process of making future value predictions based on past and present data. A model is constructed with the model parameters are estimated from historical data to forecast the value of a target variable. Regression analysis is a method for estimating a continuous dependent variable's value based on one or more independent variables. In the segmentation of customers, regression models are used to determine the behaviour of valuable customer segments.

Data visualization is usually used in conjunction with other DM techniques to understand the patterns found better, leading to further insight into its characteristics. CS results can be displayed in various 2D or 3D visual forms, such as scatter diagrams, heat maps, histograms, and tree structures, SOM map. Principal Component Analysis (PCA) and Uniform Manifold Approach and Projection (UMAP) methods for dimensionality reduction are also used for visualization [15].

Based on this study, the DM methods used in conjunction with the RFM-based CS are clustering (39 articles), visualization (25 articles), association (9 articles), classification (6 articles), forecasting (3 articles), regression (2 articles) and sequence discovery (1 article). Clustering is the most used in CS as it suits the clustering function to group customers into several clusters based on their behavioural similarities. Another widely used technique is visualization. It is in line with Gončarovs [10], who reported that among the seven DM techniques, clustering is the most commonly used for data analytics to support decision making, followed by visualization and classification. This fact indicates an increased visualization function to understand better patterns or associations found in customer segments.

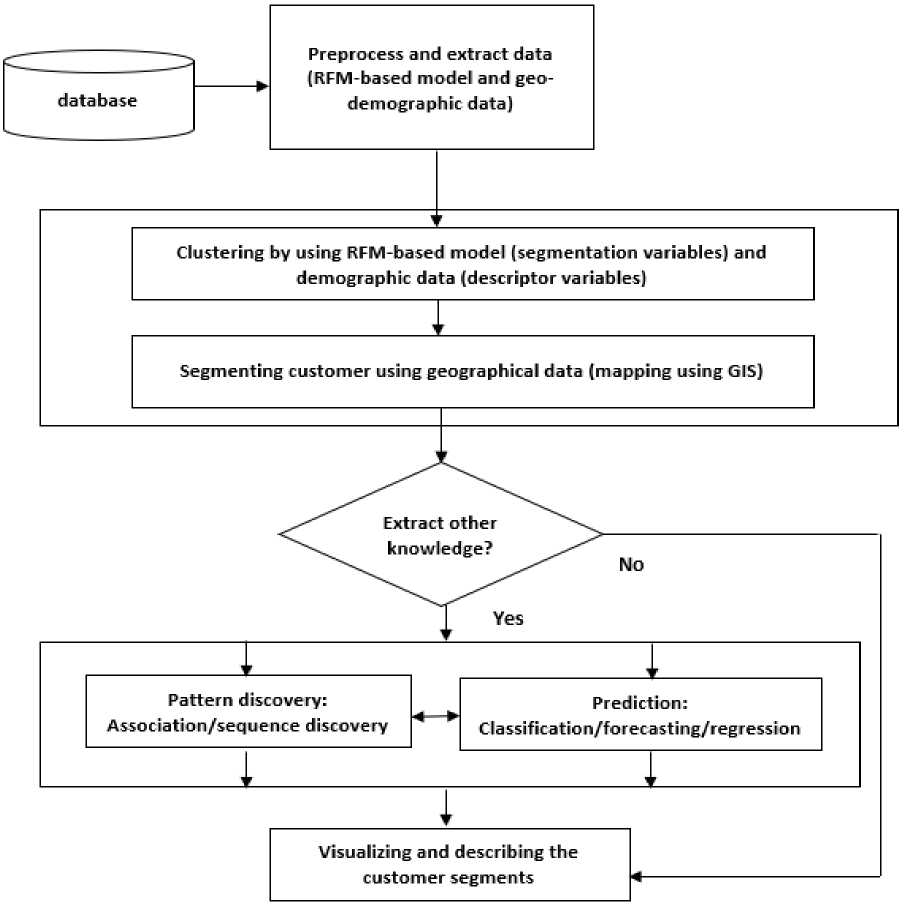
database
Preprocess and extract data (RFM-based model and geo-demographic data)
Clustering by using RFM-based model (segmentation variables) and demographic data (descriptor variables)
Segmenting customer using geographical data (mapping using GIS)
Extract other knowledge?
No
Yes
Pattern discovery: Association/sequence discovery
Prediction: Classification/forecasting/regression
Visualizing and describing the customer segments

Figure 1. The framework of CS using the RFM-based model collaborated with DM methods.

Figure 1 shows the proposed framework for using DM methods that collaborated with the RFM based CS. Clustering's function to classify customers, the descriptive DM (association or sequence discovery), is useful to discover the pattern or relationship between items. At the same time, predictive DM (classification, forecasting, or regression) predicts customer behaviour and visualization to understand the customer segment's characteristics better. This framework uses GIS to visualize customers' locations and uses demographic data as descriptor variables. This framework helps the analysts utilize DM methods to understand better and uncover customer behaviour, so companies can develop appropriate marketing strategies to increase their competitive advantage.

4. Conclusion

This study has reviewed and classified articles that use DM and RFM-based models in CS. Clustering and visualization are DM methods that are most used together with RFM-based models. The role of each DM method in CS has been presented. Each DM method has its function, which complements each other to state the customers' behaviour and predict the customers' behaviour based on past and present data. The framework for using the collaboration of DM and RFM-based CS incorporate in GIS circumstance has been proposed.

References

[1] Moghaddam S Q, Abdolvand N and Harandi S R 2017 A RFMV Model and Customer Segmentation Based on Variety of Products *J. Inf. Syst. Telecommun.* **5** 155–61
[2] Haghighatnia S, Abdolvand N and Rajaee Harandi S 2017 Evaluating discounts as a dimension of customer behavior analysis *J. Mark. Commun.*

[3] Maulina N R, Surjandari I and Rus A M M 2019 Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering *2019 16th International Conference on Service Systems and Service Management (ICSSSM)* (IEEE)

[4] Dursun A and Caber M 2016 Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis *Tour. Manag. Perspect.* **18** 153–60

[5] Lu Z, Peiyi W, Ping C, Xianglong L, Baoqun Z and Longfei M 2019 Customer Segmentation Algorithm Based on Data Mining for Electric Vehicles *2019 IEEE 4th Int. Conf. Cloud Comput. Big Data Anal.* 77–83

[6] Wei J, Lin S and Wu H 2010 A review of the application of RFM model *African J. Bus. Manag.* **4** 4199–206

[7] Naik C, Kharwar P A and Desai N 2013 A Review : RFM Approach on Different Data Mining Techniques *Int. J. Emerg. Technol. Adv. Eng.* **3** 725–8

[8] Dutta S, Bhattacharya S and Guin K K 2015 Data Mining in Market Segmentation: A Literature Review and Suggestions *Das K., Deep K., Pant M., Bansal J., Nagar A. Proc. Fourth Int. Conf. Soft Comput. Probl. Solving. Adv. Intell. Syst. Comput. vol 335. Springer, New Delhi*

[9] Asiabi T P K and Tavoli R 2015 A Review of Different Data Mining Techniques in Customer Segmentation *J. Adv. Comput. Res.* **6** 51–63

[10] Gončarovs P 2017 Data Analytics in CRM Processes: A Literature Review *Inf. Technol. Manag. Sci.* **20** 103–8

[11] Peker S, Kocyigit A and Eren P E 2017 LRFMP model for customer segmentation in the grocery retail industry: a case study *Mark. Intell. Plan.* **35**

[12] Yoseph F and Heikkila M 2019 Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method *Proc. - Int. Conf. Mach. Learn. Data Eng. iCMLDE 2018* 77–82

[13] Dogan O, Aycin E and Bulut Z A 2018 Customer Segmentation By Using RFM Model and Clustering Methods : a Case Study in Retail Indutry *Int. J. Contemp. Econ. Adm. Sci.* **8** 1–19

[14] He X and Li C 2016 The Research and Application of Customer Segmentation on E-commerce Websites *2016 6th International Conference on Digital Home* pp 203–9

[15] Pondel M and Korczak J 2018 Collective clustering of marketing data-recommendation system upsaily *Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, pp. 801-810*

[16] Li X and Li C 2018 The research on customer classification of B2C platform based on k-means algorithm *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference* (IEEE) pp 1871–5

[17] Daoud R A, Bouikhalene B, Amine A and Lbibb R 2015 Combining RFM Model and Clustering Techniques for Customer Value Analysis of a Company selling online *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech*

[18] Beheshtian-Ardakani A, Fathian M and Gholamian M 2018 A novel model for product bundling and direct marketing in e-commerce based on market segmentation *Decis. Sci. Lett.* **7** 39–54

[19] Rezaeinia S M and Rahmani R 2016 Recommender System Based on Customer Segmentation (RSCS) *Kybernetes* **45**

[20] Marisa F, Ahmad S S S, Yusof Z I M, Fachrudin and Aziz T M A 2019 Segmentation Model of Customer Lifetime Value in Small and Medium Enterprise (SMEs) using K-Means Clustering and LRFM model *Int. J. Integr. Eng.* **11** 169–80

[21] Sheikh A, Ghanbarpour T and Gholamiangonabadi D 2019 A Case Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting *J. Business-to-bus. Mark.* **26** 197–207

[22] Aryuni M, Didik Madyatmadja E and Miranda E 2018 Customer Segmentation in XYZ Bank

Using K-Means and K-Medoids Clustering *2018 International Conference on Information Management and Technology*

[23] Patel Y S, Agrawal D and Josyula L S 2016 The RFM-based Ubiquitous Framework for Secure and Efficient Banking *2016 1st International Conference on Innovation and Challenges in Cyber Security*

[24] Moeini M and Alizadeh S H 2016 Proposing a New Model for Determining the Customer Value Using RFM Model and Its Developments (Case Study on the Alborz Insurance Company) *J. Eng. Appl. Sci. 11 828-836* **11** 828–36

[25] Hamdi K and Zamiri A 2016 Identifying and Segmenting Customers of Pasargad Insurance Company Through RFM Model (RFM) *Int. Bus. Manag. 10(18)4209-4214* **10**

[26] Ravasan A Z and Mansouri T 2015 A Fuzzy ANP Based Weighted RFM Model for Customer Segmentation in Auto Insurance Sector *Int. J. Inf. Syst. Serv. Sect. 7(2), 71-86, April. 2015*

[27] 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 *Iran. J. Pharm. Res.* **15** 355–67

[28] Tarokh M J and EsmaeiliGookeh M 2019 Modeling patient's value using a stochastic approach: An empirical study in the medical industry *Comput. Methods Programs Biomed.* **176** 51–9

[29] Akhondzadeh-Noughabi E and Albadvi A 2015 Mining the dominant patterns of customer shifts between segments by using top-k and distinguishing sequential rules *Manag. Decis.* **53** 1976–2003

[30] Chen Q, Zhang M and Zhao X 2017 Analysing customer behaviour in mobile app usage *Ind. Manag. Data Syst.* **117** 425–38

[31] Safari F, Safari N and Montazer G A 2016 Customer lifetime value determination based on RFM model *Mark. Intell. Plan.* **34** 446–61

[32] Panuš J, Jonášová H, Kantorová K, Doležalová M and Hořačková K 2016 Customer segmentation utilization for differentiated approach *The International Conference on Information and Digital Technologies 2016* pp 227–33

[33] Weng C H 2016 Knowledge discovery of digital library subscription by RFC itemsets *Electron. Libr.* **34** 772–88

[34] Tsiptsis K and Chorianopoulos A 2009 *Data Mining Techniques in CRM: Inside Customer Segmentation* (John Wiley & Sons)

[35] Wong E and Wei Y 2018 Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction model *Int. J. Retail Distrib. Manag.* **46** 406–20

[36] Dachyar M, Esperanca F M and Nurcahyo R 2019 Loyalty Improvement of Indonesian Local Brand Fashion Customer Based on Customer Lifetime Value (CLV) Segmentation *IOP Conference Series: Materials Science and Engineering* vol 598

[37] Monalisa S, Nadya P and Novita R 2019 Analysis for Customer Lifetime Value Categorization with RFM model *Procedia Comput. Sci.* **161** 834–40

[38] Güçdemir H and Selim H 2015 Integrating multi-criteria decision making and clustering for business customer segmentation *Ind. Manag. Data Syst.* **115** 1022–40

[39] Sarvari P A, Ustundag A and Takci H 2016 "Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis *Kybernetes* **45**

[40] Singh A, Rana A and Ranjan J 2015 Proposed analytical customer centric model for an automobile industry *Int. J. Data Mining, Model. Manag.* **7** 314–30

[41] Li H, Yang X, Xia Y, Zheng L, Yang G and Lv P 2018 K-LRFMD: Method of Customer Value Segmentation in Shared Transportation Filed Based on Improved K-means Algorithm *J. Phys. Conf. Ser. 1060 012012*

[42] Wei J T, Lin S-Y, Yang Y-Z and Wu H-H 2019 The application of data mining and RFM model in market segmentation of a veterinary hospital *J. Stat. Manag. Syst.*

[43] Singh A and Rumantir G W 2015 Two-tiered Clustering Classification Experiments for Market Segmentation of EFTPOS Retailers *Australas. J. Inf. Syst.* **19** S117–32