Int. J. Nonlinear Anal. Appl. 13 (2022) 2, 173–184
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.23544.2555

Clustering algorithm for electronic services customers: A case study of the Banking Industry

Safanaz Heidari[a], Reza Radfar[a,*], Mahmood Alborzi[a], Mohammad Ali Afshar Kazemi[a], Ali Rajabzadeh Ghatari[b]

[a]*Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran*

[b]*Department of Management, Tarbiat Modares University, Tehran, Iran*

(Communicated by Madjid Eshaghi Gordji)

Abstract

Today, recognizing and retaining customers is one of the major challenges of customer-oriented organizations, especially in the field of banking, which has attracted the attention of many researchers. With the increasing growth of customers and the use of electronic devices that have led to the production of large volumes of data, customer behavior analysis can be considered as a competitive factor for them. In this paper, considering the varied density and data growth that leads to computational overhead, a combined approach is used of the RFM method, density-based clustering algorithm and Map-Reduce(which is an efficient and low-cost framework for running synchronous algorithms) are used. The results show that the proposed algorithm is more accurate than VDMR-DBSCAN. Also, the output of the algorithm is in the form of 5 clusters, the results of which can help managers in identifying valuable customers, This method leads to increased revenue and reduce unnecessary costs that occur due to lack of recognition and incorrect segmentation of customers.

Keywords: DBSCAN, clustering algorithm, RFM
2020 MSC: 91G45

1 Introduction

Today, recognizing and retaining customers is one of the major challenges of customer-oriented organizations, especially in the field of banking, which has attracted the attention of many researchers. With the increasing growth of customers and the use of electronic devices that have led to the production of large volumes of data, customer behavior analysis can be considered as a competitive factor for them. By using Big data technology, banks will be able to continuously use this information to monitor the transactional behaviors of their customers. Also, will be able to provide them with more diverse services. Big data refers to a set of data and related operations that have a very high volume and from a commercial point of view, refers to new methods of data utilization that by analyzing large volumes of data, processing rapid changes in data production and movement, makes it cost-effective to identify and segregation structured and unstructured data and to measure the accuracy of the data. Big Data came into being when data sets became so large, bulky, and complex that traditional tools for processing them became inefficient. Therefore, by collecting large amounts of data from various sources, they use Big data for business decisions and

*Corresponding author
Email addresses: safanazheidari@gmail.com (Safanaz Heidari), Radfar@gmail.com (Reza Radfar), Mahmood_alborzi@yahoo.com (Mahmood Alborzi), Dr.mafshar@gmail.com (Mohammad Ali Afshar Kazemi), alirajabzadeh@gmail.com (Ali Rajabzadeh Ghatari)

Received: May 2021 *Accepted:* July 2021

quick identification of behaviors, which is a good opportunity to use the banking network data and analyze customer behavior from this sector. Some banking industry experts predict a sevenfold increase in the amount of data available in the coming years, so mass data technology is a solution to exploit and use this amount of information. Big data technology represents a new way for banks to interact and use their data. As a result, banks need to change their patterns to design, develop, deploy and support big data solutions. A wave of technology has emerged in providing the flexibility and scalability needed for these changes. New methods for storing data such as databases (NoSQL) can be used in this case. Data distribution and computing software (such as Hadoop) has reached a level of maturity that can deliver the expected performance of a modern platform, while this amount of data has never been used before. Exactly when banks need to re-evaluate technologies, the methods of implementing and deploying big data must also change. Agile development methodologies have been developed to enable rapid, repetitive, and incremental deployment and deployment that can be used to rapidly access data in a way that is properly measured understood and analyzed. Today, the components of a comprehensive big data framework are available and ready to use, and it seems that the time has come for banks to enter this technology. Analyzing data and the Internet has made it much easier now to monitor and evaluate the progress of banks than in the past, which is made possible by access to a wealth of personal information from customers; But now, with big data technology, banks will be able to use this information continuously to monitor their customers' transactional behaviors at the time of occurrence (and almost immediately), and this will help banks to provide better services and resources. These real-time services increase their overall profitability. As the number of bank customers increases, so does the need to provide services that are affected by their requests and needs. However, the responsibility to protect the funds and personal information of customers is one of the most important issues for banks. Some of the most important applications of big data in the banking industry are: change in service level, fraud detection and prevention, development of advanced reporting based on analysis, customer segmentation, marketing and customer relationship management, anti-money laundering, product personalization To the customer is risk management, inspection and monitoring. Data mining is the process of discovering patterns and processes that are regular and hidden in large, distributed data. Different data mining methods are used depending on what kind of knowledge is considered in the data mining process. Clustering is a data mining technique that uses an unsupervised learning approach to find distributions and patterns of unlabeled data sets. The purpose of clustering is to maximize similarities within groups and minimize similarities between groups by dividing the data points into a number of nodes in which the members within the group have the most similarity and the two groups are completely separate. Customer clustering is the process of identifying customers who have specific characteristics and consumption habits in the public market. Clustering divides customers into homogeneous clusters that have similar needs and characteristics of customers within each cluster. Therefore, by recognizing customers, personalized products and services can be provided according to their needs. Today, one of the most common models for segmenting and identifying customers based on value is the RFM model proposed by Hughes in 1994, which is based on three variables: Recency, Frequency and Monetary value in extracting the behavioral characteristics of customers. Is used and affects the probability of future purchases of customers [8]. By Using the RFM model, marketing managers can effectively target valuable customers and then develop marketing strategies based on their values [20]. Various types of clustering methods such as hierarchical clustering, network clustering, ..., and density-based clustering have been proposed by researchers. This article focuses on the density-based clustering method. In this method, which was proposed by Martin et al. In 1996, the definition of a cluster based on two parameters is the minimum number of points and radius that clusters are defined as dense areas of the data set. Objects in low-density areas separate the clusters from each other, which can be noise or border points. This method connects points that are within a certain range (in a neighbourhood radius). The advantage of this method over other clustering methods such as K-means is that it is not sensitive to the shape of the data and can also detect irregular shapes in the data [27, 26]. With the rapid growth and development of information technology, data is generated at high production rates in various fields such as business, email, disease, etc. This data is provided to users in structured, semi-structured, and non-structured forms. New technologies are needed to store and extract useful information from this volume of data called Big data. Although various algorithms have been proposed in this regard, each has looked at the problem from one aspect. Considering the inefficiency and inefficiency of existing algorithms due to running on one machine and other challenges associated with Big data, In this article, we have tried to look at the problem from several dimensions and offer a solution.1- First, due to the distribution of data in various databases, it is necessary to design a data warehouse, which traditional data warehouses do not meet this volume of data, and the Hadoop-based database, which is Hive, has been used.2- After preparing the data, RFM model variables have been extracted. 3-The issue is the correct data partitioning that the PRBP algorithm has been used [10]. 4- The DBSCAN density-based clustering algorithm is inefficient in many cases because our goal is to provide a variable density algorithm. For this purpose, the local density of each point has been used to separate clusters of different densities. 5- To solve the problem of a machine not responding to a large amount of data, Map-Reduce has been used to process data in several machines. The rest of the article is organized into several sections: In the second section, the literature and research background are discussed. The methodology is

introduced in the third part and the fourth part evaluates the performance of the proposed algorithm and compares it with another algorithm and the final part is the conclusion.

2 Research Background

Today, almost everything is done electronically. People exchange information through the Internet, buy, and sell through the Internet. E-commerce vendors use the Internet to market goods and services, improve revenue and brand awareness. The provision of electronic services in the banking sector has created the need for continuous identification of consumers and analysis of their behavior. Analyzing customer behavior in the banking sector is quite complex because the databases are multidimensional and consist of monthly and daily transaction records of a large number of customers [18]. With the advent of big data analysis, there will be more informed and data-driven strategies for identifying and communicating with customers. Therefore, those e-service providers can obtain and analyze massive data from the exchange of electronic information, to gain a better understanding of consumer behavior. By identifying customer value based on information about the use of electronic services over a period, banks are developing different marketing strategies to retain the most valuable customers.

3 RFM Model

Recency, Frequency, and Monetary Value (RFM) is an effective customer segmentation model that distinguishes important customers from large volumes of data. It is also a behavioral analysis that can be used for market segmentation [11, 29]. Hughes describes that the main asset of the RFM method is, on the one hand, to obtain customer behavioral analysis in order to group them into homogeneous clusters, and, on the other hand, to develop a marketing plan tailored to each specific segment of the market. RFM analysis improves market segmentation by examining time (R), frequency (F), and money spent (M value) on a particular product or service, Young says Customers who have recently bought over and over again and spent the most money are likely to react to future promotions [36]. Once customers' RFM model scores are determined, customers can be grouped into segments and then their profitability analyzed. To know the RFM value of the customer [12] using the RFM value, the symbol " ↑ " is a value above average and the symbol " ↓ " is a value below average. This means that more value is better for the company and the lower value is not good for the company. But for R, the symbol" ↓ " means below average and for the company is optimal, and the symbol " ↑ " means above average and its value is not good for the company. The cluster belonging to the symbol R " ↓ " F " ↑ " M " ↑ " is named with the loyal customer, the symbol R " ↑ " F " ↓ " M " ↓ " is named as the lost customer, the symbol R " ↓ " F " ↓ " M " ↓ " to the new customer and the symbol R " ↓ " F " ↑ " M " ↓ " It is named after the future customer.

4 Big Data

The term Big data refers to datasets that other traditional database management systems no longer respond to this amount of data. Bulk data is datasets that are larger than the capabilities of software tools and storage systems. Which are defined by three characteristics of high volume, high variety, and high production speed, but the value characteristic is also sometimes mentioned [21]. IBM definition of Big Data: This definition includes four features for big data, three of which are taken from the 3V definition of Gartner Institute and IBM itself adds the fourth dimension. This added dimension is called correctness. Sometimes the opposite is used, is uncertainty. These four dimensions together describe IBM's 4V model.

5 Apache Hadoop

Apache Hadoop is a Java-based open-source software framework meant for distributed processing of very large datasets across thousands of distributed nodes. A Hadoop cluster divides data into small parts and distributes them across the nodes. Doug Cutting and Mike Cafarella created the Hadoop framework in [4]. Apache Hadoop is developed to scale up from the single server to in cluster of multiple machines, each of these offering its own (local) computation and storage capabilities [3]. Structurally, Hadoop is a software infrastructure for the parallel processing of big data sets in large clusters of computers. The inherent property of Hadoop is the partitioning and parallel processing of mass data sets. Hadoop is based on Map-Reduce programming which is suitable for any kind of data.

6 Map-Reduce

Map-Reduce is a framework for implementing distributed and parallel algorithms in datasets [17]. This framework was introduced by Google in 2004 to support distributed processes on a distributed datasheet across clusters of computers. The model follows the rule of split and overcome. Thus, dividing the input data sets into separate pieces that are processed in parallel to the mapping phase. Then the sorting operations of the mapping outputs are performed by the framework and used as inputs for the reduction phase. These operations are carried out in three phases: the mapping phase, the sorting phase, and the reduction phase [15]. The main idea of Map-Reduce is to divide the data into fixed-size chunks which are processed in parallel which take advantage of that. Also, it is designed to avoid computer node failure issues (fault tolerance) [24, 14].

7 DBSCAN

The density-based method in clustering is one of the most popular clustering methods in which data in the data set is split based on density, and high-density points are separated from the low-density points based on the threshold. The density-based method is the basis of density-based clustering algorithms [13]. The advantage of this method over other clustering methods such as K-Means is that it is not sensitive to the shape of the data and can also detect irregular shapes in the data. In contrast to its advantages, this algorithm does not support a variety of densities. Other algorithms are presented to improve this imperfection.

8 Experimental background

In recent years, the integration of the RFM method and data mining analysis techniques has been proposed for various areas such as customer identification and customer profitability analysis. Tavakoli and his colleagues in an article entitled Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining, addressed the challenge of not paying attention to customer behavior in the RFM model and changing it. They presented the $R + FM$ model and segmented their statistical population based on the k-means model and compared the output with the RFM model running on the same data. They developed strategies for each sector and were able to increase the number of purchases and their financial value in the shopping cart [34]. Hussein used the DBSCAN and K-means algorithms for clustering, and the results of implementing these two algorithms show that each of them can be used to split the customer. Unlike K-means DBSCAN is not bound to data form. They showed that the DBSCAN algorithm is more suitable for clustering [19]. Sohrabi et al. in [32] using the RFM model and its combination with the value of customer life and finally its use in the banking industry categorized and ranked customers in 8 sections. After categorizing and determining the rank of each of the categories, they stated the appropriate strategies that the bank should use in dealing with each of the categories [32]. By using the RFM model and neural network, Afsar et al. categorized and ranked the bank's credit customers and divided the customers into 10 clusters, and granted facilities to the customers based on the points of each section [1]. Sohrabi et al. in [31], an article aimed at assisting marketing and sales managers in the pharmaceutical industry, by identifying and analyzing different segments of customers and providing proposals tailored to each segment, to maintain and increase their purchases using data mining, have suggested. They clustered and analyzed the pharmacy based on the variables of Recency, Frequency, monetary value, and purchase time in the RFML model. As a result of this segmentation, three categories of pharmacies named: low-cost and low-profit pharmacies, with average and loyal and profitable purchases and profits in terms of sales process were identified and based on this segmentation, related analyzes It is provided [31]. Maleki et al. In 2016 to segment, the customers of the Qom Telecommunication Internet sector, after determining the values of RFM model indices, determined the weight of each index based on the hierarchical analysis process and then divided the customers into two clusters. . The results provide the basis for analyzing customer characteristics and their prioritization [25]. Gharib et al. [16] proposed a hybrid data-mining model using association rules and clustering. They identified customer behavior patterns and based on them, they placed customers in 4 behavioral groups. These results can help senior managers to adopt appropriate strategies to improve behavioral patterns [16]. In [23], Khodabandehloo and Rahman also introduced a combination of data mining approaches to customer segmentation in order to achieve the desired segmentation results. According to their proposed method, customers were classified into four groups, and to better understand customers, each section was classified into ascending and descending sections [23]. Yousefizad and Sorayai [37], in a study based on the RFM model clustered customers and the results showed that the K-means method is a better method for customer clustering. After clustering and forming a customer pyramid, they were placed in 5 groups and introduced services and facilities suitable for each group [37]. In another study, Baradaran and Farokhi introduced the developed RFMC model, and the results of this study show that the accuracy of the model proposed by them is

higher than %5.5 the RFM model. In addition to analyzing the customer behavior of each cluster, a model based on the forward neural network has been developed to predict the number of customer clusters based on their behavioral and demographic characteristics [6]. In 2020, Digiuristik et al. presented a combined approach of the RFM method, the K-means clustering algorithm, and the SVM. They concluded that the proposed algorithm to identify valuable customers is very effective that can lead to increased revenue in the organization [22]. In [5], Ballstar et al. presented a segmentation of customers on the cashback website. The segmentation is based on the activity and business role of the customers in the social network of the site. This study shows how the role of the customer in the cashback website social network determines the customer behavior and activity on the website. . The proposed segment describes the customer in terms of profitability and age. These findings explain customer behavior in e-commerce and the value of using personal retention strategies for each cluster rather than general or customer purchasing strategies. This article describes how customers move between clusters and allows professionals to increase customer loyalty and long-term profitability [5]. Namvar et al. in [28] introduced a new method of customer segmentation that includes clustering in two phases. They showed that combining demographic information and two-stage clustering leads to relatively better clustering [28]. In [2], Aghlis et al. used RFM to analyze the customer database of Greek banks and, using the concept of customer value pyramid, categorized customers into five categories of customers based on K-means and two-step clustering algorithms. Top, main customers, regular customers, retail customers, and inactive customers and evaluated customer profitability in increasing banking services [9].

9 Methodology

Due to the nature of the research, the research method used is based on design science. The emphasis of design science is based on the design of new styles and the transformation of existing styles. In this model, all designs begin with the knowledge of a problem. On the other hand, design science research, sometimes known as "improvement research," focuses on problem solving or improving the nature or performance of an activity. According to the general design cycle model, the suggestions that are offered to create a solution are derived from existing knowledge or theories of the problem area, and then an attempt to implement a structure according to the proposed solution has been done. In the following, according to the implicit or explicit functional characteristics of the solution created, the partial or completely successful implementation of that proposal will be evaluated. In this cycle, the development and evaluation of proposals are often repeated. The last part of this model is the conclusion that represents the endpoint of a unique and special design project [30]. Design science research should produce appropriate products in the form of a structure, a model, a method, or a sample. The output of our design science research is in the form of a new algorithm for density-based clustering using mapping-reduction in cloud data warehouses. In the present study, while introducing customer performance variables from the perspective of bank managers, an algorithm for clustering customers of the bank's electronic services has been proposed so that they can be clustered based on customer value and provide facilities appropriate to each cluster. In order to evaluate the proposed algorithm, one of the country's banks that provides modern banking services has selected. Customer transaction data has collected whether through mobile banking, ATMs, or system software. So that all the data done through credit card transactions were examined. The services provided in this bank include balance announcement, balance transfer, withdrawal of funds, reports related to checks, payment of facilities, etc. All services have been provided through mobile phones, websites, ATMs, etc. Done. The main data set used for this purpose has been obtained from the bank and the performance data period is related to transactions made using the bank's credit cards for the last quarter of 2020. This dataset contains 1352406 credit card users whose user identity is completely confidential and will not be disclosed. The resulting database contains information about the bank's customers, gender, place of residence (region), age, average monthly income, as well as data related to transactions such as the date and value of transactions, based on which the RFM attributes are calculated. The RFM model is one of the most popular models in calculating customer value and clustering based on it. The strength of this model is that it extracts customer characteristics with fewer criteria using clustering. The purpose of this study is to provide a method for customer clustering to identify customers with a focus on customer value, which is based on the improved RFM and DBSCAN model. According to the proposed algorithm presented in 3 layers, in the first layer, data are collected from different sources and placed in the data warehouse. At this stage, the data is also cleared, which means that incomplete and poor quality data is removed or replaced because each record must contain content for all three variables. The challenge in this article is to query and analyze big datasets for which SQL is not suitable. Hive can be a good alternative to SQL. Apache Hadoop is an open-source data warehouse used to query and analyze massive datasets - structured and semi-structured data - in distributed environments. Apache Hive, located at the top of Hadoop, is used to receive and convert various types of data. In this optimization, using hive, we have converted and loaded the data warehouse into a star schema. At this point, the existing records for each client must be converted to a format that can be used in the RFM model. Therefore, the research variables

were extracted based on the RFM model, in which three indicators include the Recency (R) of the time interval from the last customer interaction with the bank, the number of times (F) includes the number of times the customer's financial interaction with the bank during a certain period. In addition, financial value (M) is the amount of money that a customer has spent over a period to complete a transaction, expressed as the amount of balance a person has at the time of account turnover. Because the collected data for the model indicators are not of the same type and scale, the data should be normalized. To normalize the data, we use Formula 1, so that x is the original value, x_{max} is the maximum value, x_{min} is the minimum value, and x ' is the normalized value.

1. $x^{'} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$

In the second layer, the clustering process is performed on each node independently. Each mapper reads the data as a (key, value), which key=null and value=partition. Sooner than starting the clustering process, as regards the prime purpose of this research, is to create an appropriate algorithm for clustering varied density data, local density for point x is calculated according to the following functions (Eqs. 2,3), which is better than counting the points in the neighbourhood radius (EPS).

2. d (x,y)= $\sqrt[2]{\sum_{i=1}^{n}(xi - yi)}$
3. Local-density= $\sum_{i=1}^{K} d(x?xi)$

The important point to note in the calculation of local density is the exact determination of the parameter K, which has been selected in several test steps for a suitable amount of K. Local densities (LD_i) obtained are arranged in the local-density list in ascending order using merge sorting algorithm. The points belonging to the same cluster have close values of LD_i, which can be calculated for the adjacent points of p_i and p_j in the local density list; Eq.4 calculate the density difference between the two points.

4. $LDVar_{(p_i,p_j)} = \frac{(LD_{p_j} - LD_{p_i})}{(LD_{p_i})}$

After calculating the density difference, we set LDVarlist to determine the points that are located on the same level as the cluster. The values in LDVarlist which are greater than a threshold λ (calculated by Eq. 5 are separated out and put into separate LDlevel (density level set).

5. $\lambda = Ex(LDVarlist) + w.SD(LDVarlist)$

Ex: mathematical expectation;
SD: standard deviation;
W: tuning coefficient (for multi-density datasets w=2.5 is a suitable value [35] .
By specifying the set of density levels, levels that have the same density and the density difference between the two levels is less than 0.2 (Eq. 6) are merged [35]. In this step, the value of ε is calculated for each level (Eq. 7) and stored in the Eps List.

6. $\text{DenGrade}(DLS_i, DLS_j) = \frac{\text{mean local density}(DLS_j}{\text{mean local density}(DLS_i} - 1$
7. $\varepsilon_i = \max \text{local density}(DLS_i)\sqrt{\frac{\text{median local density}(DLS_i}{\text{mean local density}(DLS_i}}$

By calculating LDlevel list values, noise and boundary points should be wiped out of the list; the EPS values for each level is set, assuming that each object must know its own EPS radius. The largest EPS in each level is considered as Max-EPS and stored in EPSList. By specifying the minpts=k and EPS_i parameters, we call the DBSCAN algorithm for each level. The KD-tree spatial index will be used to obtain optimal query data in the dataset before the start of the algorithm. Eventually, last clusters of varied density are procured and remained points are determined as noise points. The clustering outcomes after implementing the DBSCAN algorithm are divided into three groups of boundary regions ,local regions and unvisited points. Before sending the output of map phase to the reduces phase, the operation of the combination in each mapping takes place separately to merge between the chunks of a mapper, and in the shuffle phase, a combination of the mappers is done. The clustering outcomes of the local region and unvisited points are stored in the local disk and boundary region are sent to the reduce phase.

Local region: Output $(qt\_index, partitionindex + qt\_cluster\_id)$.

Unvisited points: Output $(qt\_index, partitionindex + qt.Eps - value)$.

Boundary region: Output $(qt\_index, partition\_index + qt_cluster\_id + qt.iscore_point + qt.Eps - value + density)$.

The reduce phase, receives the pairs of clusters from adjacent partitioning and Identifies data points with the same *qt_index* of adjacent partitions that have the ability to merge in the merge phase. The output of this phase is a list of clusters that can integrate with each other. Points with the same *qt_index* are executed at the same reducer. In fact, this phase decides if the two clusters that share the boundary points merged or not. Two clusters are merged; the boundary point is the core point in one of the clusters, second provided that Eps values difference is equal to or less than θ. In this way, we prevent the integration of clusters by different densities. The value of θ is not fixed, depending upon the quality of clusters. If the clusters are near each other and not suitable for merging, the boundary point which is part of both the clusters should be assigned to one of the clusters. The point is earmarked to the cluster with the least difference of Eps value and kdist values. In the reducing phase, the output of the mapping phase, which is a list of merging clusters, are merged together. The output of this phase is a list of clusters that could be merged together. In the last layer, the clusters are merged, after which; the clusters are sorted in descending order, relabelled by the first cluster in the sorted list and the clusters in the local disk are relabelled. The remaining points are not marked as noise; they are rather marked as unvisited points. In this study, first, all datasets are called and tables are stored in Hive as a star schema. Then, based on the purpose of the research, the required variables are extracted and converted, and stored in HDFS. In this research, based on the RFM model, customer clustering is performed using the proposed algorithm. For clustering analysis based on the RFM model, the initial data set needs pre-processing. First, the appropriate variables are selected from all the variables. . For recency, the "Transaction Date" and "Customer Number" attributes are used to specify the duration (day) of the customer who recently made the transaction. For frequency, the "Customer Number", "Transaction Date" and "Operation Code" attributes are used to determine the repetition of a customer's purchase over a specified period. For monetary, the feature "Customer number", "Account balance" is used to determine the total amount traded (account entry and exit) by a customer. After specifying the variables, the data was cleared, so the data Missing values are identified and replaced with the average value of the data. Due to the differences in the index units of the RFM model, these values must be normalized based on the same unit. For this purpose, the data were normalized by the max-min method. The higher the value of the two indicators M and F, the more favorable the customer situation, and conversely, the lower the value of the R index, the better the customer situation. After normalization, we used the PRBP algorithm to divide the data, which were divided into 3 sections with minimum boundary points. According to the number of sections obtained, we used 2 slave nodes and a master node. At this stage, in each node, the local density of each point was calculated separately and based on the local density, the value of ε was determined and the DBSCAN algorithm was implemented independently in each node. The output of the results was observed after the mapping and reduction steps in both cluster and noise points, where the noise points can be labelled unobserved to be re-examined in the steps of adding new data. At this stage, 5 clusters were identified, which analysis and evaluation of each cluster are reviewed below.

10 Experimental setup

The main goal of this article is to cluster customers using the RFM model and improved the DBSCAN algorithm to find loyal and profitable customers to achieve the highest profitability among competitors. Proper interpretation of each cluster is essential in creating customer-centric business intelligence. Considering that for each customer, each of the normal parameters R, F, M can be above the normal average or below the normal average, they can be combined in 8 cases. Table 1 shows the 8 states obtained from the combination of the three parameters. Clusters show the

Table 1: Different combination modes in the RFM model

	Modes
1	R↑ F↑ M↑
2	R↓ F↓ M↓
3	R↑ F↓ M↓
4	R↑ F↑ M↓
5	R↑ F↓ M↑
6	R↓ F↑ M↓
7	R↓ F↓ M↑
8	R↓ F↑ M↑

behavioral differences of each customer well. People in a cluster are closest to each other in terms of characteristics.

5 clusters are obtained based on the proposed algorithm, which is interpreted in the continuation of each cluster. The first cluster (R↓ F↑ M↑): The customers of the first cluster, which includes the fewest customers, are the customers with the highest transaction frequency and financial volume, and the time interval between transactions is very short. In other words, customers whom all have an average of high financial value index and in the transaction freshness index have an average lower than the average of all customers and also the average frequency of their transaction index is better than other clusters. From the point of view of banking experts, these customers are considered as special and loyal banking customers, who are often the owners of stocks and large investment companies, and maintaining this group is very vital and valuable for the bank. The second cluster (R↓ F↓ M↑): It includes customers who have not had much time since their last visit to the bank and their average transaction is low, which is good for the bank and also have done good financial transactions that can be profitable for the company, but They have a low average transaction frequency, which can be converted into loyal customers by offering new offers and facilities to these people so that these people can also be placed in the first cluster. The number of people who are in this cluster is more than the first cluster, so the attention of bank managers to these customers is very important because there are customers who are profitable for the company and maintaining these customers will cost less for the company. The third cluster(R↑ F↑ M↑): The customers of this cluster are customers who have more financial transaction frequency, in other words, they have a higher average than the total and also the financial value index of this group is higher than the average. But it has been a long time since they went to the bank that these people should be among the valuable customers by providing facilities and the bank should try to turn them from potential to actual. The fourth cluster(R↑ F↓ M↑): There are customers who have a higher account balance than the average, but have not done any financial transactions for a long time. In other words, a long time has passed since their last transaction and the number of their transactions is low. Since these customers have higher account balances, they can be valuable to the bank, but these people are not considered loyal customers in terms of loyalty to the bank, so the bank should try to maintain and increase the loyalty by providing more interaction and better services to customers. And make it possible to increase the transaction and reduce the time interval between transactions. Fifth cluster(R↓ F↓ M↓): It includes customers who have low purchase frequency and their financial index is also low and not long after their last transaction. In fact, new customers are the best opportunity for the bank so that bank managers can turn them into loyal and valuable customers from the beginning. A number of customers were not included in any of the above 5 clusters, which are considered noise points, and these points were not considered in this study. Which can be re-examined in future research using the incremental algorithm. According to Table 2, the highest number of customers are in the third cluster and the lowest number is in the first cluster. In this table, the number of customers in each cluster is specified separately. Percentages also represent the average of each parameter that is measured relative to the total average. These values are obtained after Max-Min normalization. In parameters F and M, values above the average are desirable for the bank, and in the case of parameter R, values below the average are desirable. . Therefore, as explained in the interpretation of clusters, managers and banking experts can behave differently towards the customers of each cluster based on the output obtained from the algorithm. Which can lead to improved bank performance.

Table 2: Algorithm output in step 3

R	F	M	Number customers(total: 1352406)	cluster
0.38	0.67	0.87	11259	1
0.37	0.43	0.74	160665	2
0.63	0.57	0.62	792176	3
0.70	0.45	0.60	211202	4
0.93	0.66	0.32	123589	5
0.59	0.56	0.36		totall average

To compare the MR-VDBSCAN algorithm with the VDMR-DBSCAN algorithm[7] used from 3 criteria Rand, Jaccard and Fowlkes-Mallows, which are among the clustering evaluation criteria. Three criteria calculate the degree of similarity of the clusters obtained from the evaluated algorithms with the labelled clusters. The results are shown in tables **??**. The best parameters are set for the evaluated algorithms by performing various tests on the data set. These criteria calculate the similarity between the clusters obtained from the evaluation algorithms. The results presented in the tables show that the MR-VDBSCAN algorithm has a higher similarity index than the VDMR-DBSCAN algorithm. Table 3 shows the Jacquard criteria for the best results compared to VDMR-DBSCAN; This means that the clusters generated by this algorithm are more like labelled clusters.

Table 3: Jaccard test results

	dataset
VDMR-DBSCAN	0.942
MRVDBSCAN	0.975

The Fowlkes–Mallows Index is the geometric mean of precision and recall. This measure is based on the pairwise approach to calculate TP, TN, FP, and FN. The value of the Fowlkes–Mallows Index is between 0 and 1, and a high value means better accuracy. As shown in Table 4 we found that the proposed algorithm with .95 for the dataset has the highest similarity.

Table 4: Fowlkes–Mallows Index test result

	dataset
VDMR-DBSCAN	0.882
MRVDBSCAN	0.953

Also, Rand-measure—the similarity measure—was used to evaluate the effectiveness of the proposed algorithm. The value of Rand-measure is within [0,1] and the near number to 1 indicates that the two partitions are similar. To make the competition more fairly we choose the same values for parameters in all algorithms because the parameter setting can influence both the clustering results and the execution time. The proposed algorithm has a high similarity; it can, therefore, be said that it has superb precision and the output of the criteria is verifying the subject, too, as shown in Table 5. This table shows the similarity of the two clustering algorithms in the data set. We see that the MR-VDBSCAN algorithm can provide better performance than other algorithms. In particular, the similarity of the proposed algorithm is .87 for the data set.

Table 5: Rand-measure test results

	dataset
VDMR-DBSCAN	0.794
MRVDBSCAN	0.872

F-measure was used to evaluate the accuracy of the final clusters based on precision and recall. F-measure is commonly used in evaluating the effectiveness of clustering algorithms. F-measure for quality of clustering algorithm is given by the following formula; where N is the total number of the data points and ci is a candidate cluster [33]. Figure 1 shows that the proposed algorithm has higher accuracy compared to VDMR-DBSCAN.

$$f = \sum_{i}^{l} \frac{|c_i|}{N} * F(i,j)$$

11 Conclusion

Given the growing banking industry and fierce competition between banks, identifying customers is essential to making a profit and is the key to their success. Each customer creates a different value for the organization. Customer segmentation can help bank managers identify the value of each customer as well as identify appropriate strategies for each. In this paper, the RFM method was used to identify and cluster customers. The customers of the surveyed bank were selected for evaluation in the last quarter of the year. Based on the proposed algorithm, customer data were

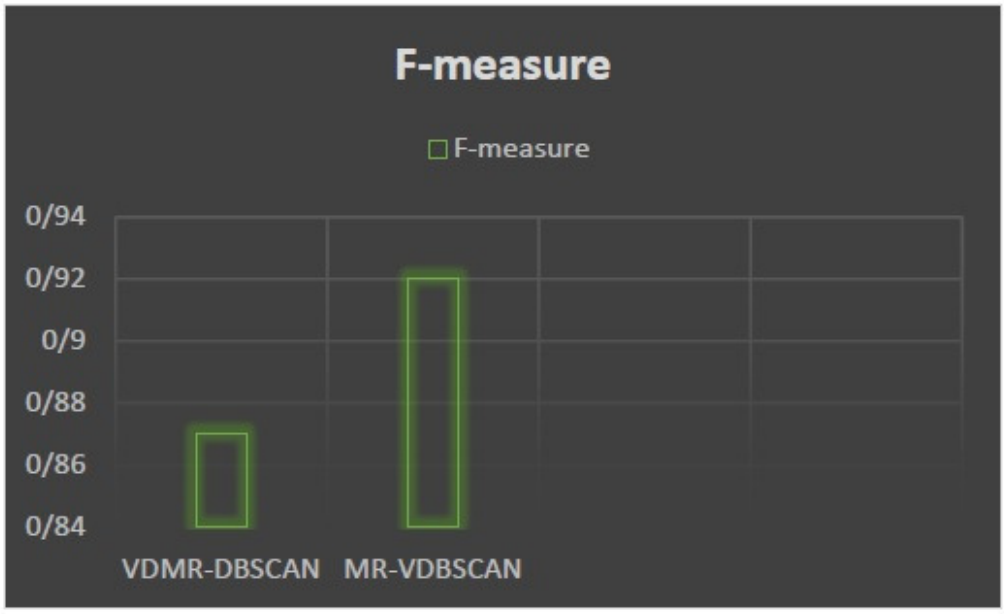

Figure 1: F-measure evaluation

analyzed and divided into 5 clusters. The lowest number of customers was related to the first cluster and the highest was related to the third cluster. Of course, some customers did not fall into any clusters. Because the Euclidean distance and density were used for clustering, some of the data were not at a suitable distance for cluster formation. Customer clustering can help decision-makers identify market segments and develop marketing and sales strategies for customer satisfaction. Also in this paper, a comparison with the VDMR-DBSCAN algorithm is made, which shows that the proposed algorithm has higher accuracy in data clustering than the comparison.

References

[1] A. Afsar, R. Hoshdar and R. Minaie, *Customer credit clustering to provide tailored facilities*, Iran. J. Manag. Res. **14** (2012), no. 4, 1–24.

[2] V. Aggelis and D. Christodoulakis, *Customer clustering using rfm analysis*, Proc. 9th WSEAS Int. Conf. Comput. 2005, p. 2.

[3] I. Akbar, T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani and S.A. Khan, *The rise of "big data" on cloud computing:review and open research issues*, Inf. Syst. **47** (2015), 98–115.

[4] N. Aktar, M.V. Ahmad and S. Khané, *Clustering on big data using Hadoop*, Int. Conf. Comput. Intell. Commun. Networks (CICN), IEEE, 2015, p. 789–795.

[5] M.T. Ballestar, P.G. Carles and J. Sainz, *Customer segmentation in e-commerce: Applications to the cashback business model*, J. Bus. Res. **88** (2018), 407–414.

[6] V. Baradaran and Z. Farokhi, *Customer segmentation in the banking industry using the developed RFMC model*, Brand Manag. Quart. **1** (2015), no. 2, 135–154.

[7] S. Bhardwaj and S.K. Dash, *VDMR-DBSCAN:Varied density mapreduce DBSCAN*, Int. Conf. Big Data Anal. Springer, Cham, 2015, p. 134-150.

[8] C. Cheng and Y. Chen, *Classifying the segmentation of customer value via RFM model and RS theory*, Expert Syst. Appl. **36** (2009), no. 3, 4176–4184.

[9] D. Christodoulakis and V. Aggelis, *Customer clustering using RFM analysis*, Expert Syst. Appl. **36** (2009), 2678–2685.

[10] B.-R. Dai and I.-C. Lin, *Efficient map/reduce-based DBSCAN algorithm with optimized data partition*, Fifth Int. Conf. Cloud Comput. IEEE, 2012, p. 59–66.

[11] R.A. Daoud, A. Amine, B. Bouikhalene and R. Lbibb, *Customer segmentation model in e-commerece using clustering techniques and LRFM model: The cese of online stores in Morocco*, Int. J. Comput. Inf. Enng. **9** (2015), no. 8, 2000–2010.

[12] A. Dursun and M. Caber, *Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis*, Tourism Manag. Perspect. **18** (2016), 153–160.

[13] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, *A density-based algorithm for discovering clusters in large spatial databases with noise*, kdd, **96** (1996), no. 34, 226–231.

[14] F. Eugen, L. Ramakrishnan and C. Morin, *Performance and energy efficiency of big data applications in cloud environments: A Hadoop case study*, J. Parall. Distrib. Comput. **79** (2015), 80–89.

[15] X. Fu, S. Hu and Y. Wang, *Research of parallel DBSCAN clustering algorithm based on MapReduce*, Int. J. Database Theory Appl. **7** (2014), no. 3, 41–48.

[16] A. Gharib, A. Toloie and K. Heidarzadeh, *Providing a combined model of data mining using association rules and clustering to identify dominant patterns of customer behavior (Case study- Ansar Bank)*, Manag. Futurology **30** (2019), no. 3, 189–201.

[17] Y He, H Tan, W Luo, S Feng and J Fan *MR-DBSCAN: A scalable MapReduce-based DBSCAN algorithm for heavily skewed data*, Front. Comput. Sci. **8** (2014), 83–99.

[18] N.-C. Hsieh, *An integrated data mining and behavioral scoring model for analyzing bank*, Expert Syst. Appl. 27 (2004), no. 4, 623–633.

[19] A.S. Hussain, *Customer segmentation using centroid based and density based clustering algorithms*, 3rd Int. Conf. Electric. Inf. Commun. Technol. IEEE, 2017, p. 1–6.

[20] S. Irvin, *Using lifetime value analysing for selecting new customers*, Credit World **83** (1994), no. 3, 37–40.

[21] A.J. Ishwarappa, *A brief introduction on big data 5Vs Characteristics and Hadoop technology*, Procedia Comput. Sci. **48** (2015), 319–324.

[22] S. Jafari, A. Farajzadeh and S. Morad, V. Djurisic, L. Kascelan, S. Rogic and B. Melovic, *Bank CRM optimization using predictive classification based on the support vector machine method*, Appl. Artif. Intell. **34** (2020), no. 12, 941–955.

[23] S. Khodabandelou and M. Zivari Rhman, *Providing a new approach for segmenting customers based on their purchasing behavior change over time in electronic business*, J. Inf. Technol. Manag. **9** (2017), no. 2, 277–300.

[24] C.W. Lu, C.M. Hsieh, C. H. Chang and C. Yang, *An improvement to data service in cloud computing with content sensitive transaction analysis and adaptation*, IEEE 37th Ann. Comput. Software Appl. Conf. Workshops, 2013, p. 463–468.

[25] M.M. Maleki, A. Zarei and Z. Hajiloo, *Identifying and segmenting key customers for prioritizing them based on lifetime value using RFM model (Case study: Internet customer of Qom Telecommunications Company)*, Iran. Bus. Manag. **8** (2016), no. 2, 461–478.

[26] N. Maitry and D. Vaghela, *Survey on different density based algorithms on spatial dataset*, Int. J. Adv. Res. Comput. Sci. Manag. Stud. **2** (2014), no. 2, 362–366.

[27] A.K. Nafees and R.T. Abdul, *An overview of various improvements of DBSCAN algorithm in clustring spatial databases*, IJRCCE. **5** (2016), no. 2, 360–363.

[28] M. Namvar, S. Khakabimamghani and M.R. Gholamian, *A two phase clustering method for intelligent customer segmentation*, Int. Conf. Intell. Syst. Modell. Simul. IEEE, 2010, p. 215–219.

[29] S. Qadaki Moghaddam, N. Abdolvand and S. Rajaee Harandi, *A RFMV modeland customer segmantation based on variety of products*, Inf. Syst. Telecommun. **5** (2017), no. 3, 155–161.

[30] A. Rajabzadeh Ghatari, S. Nikghadam Hojati and M. Faridi Masule, *An introduction to design science and meta-analysis*, Neghah Danesh, Tehran, 2015.

[31] B. Sohrabi, R.V. Iman and N. Nikaien, *Segmentation of pharmaceutical industry customers based on RFML model*, Bus. Manag. **8** (2016), no. 4, 861–884.

[32] B. Sohrabi, A. Khanlari and N. Ajorlu, *A model for determining the life cycle value of customers in the banking industry*, Manag. Res. Iran. **10** (2011), no. 1, 224–239.

[33] J. Song, C. Guo, Z. Wang, Y. Zhang, G. Yu and J.M. Pierson, *Haolap: A Hadoop based OLAP system for big data*, J. Syst. Software **102** (2015), 167–181.

[34] M. Tavakoli, M. Molavi, V. Masoumi, M. Mobini, S. Etemad and R. Rahmani, *Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: A case study*, IEEE 15th Int. Conf. e-Bus. Engin. 2018, p. 119–126.

[35] Z. Xiong, R. Chen, Y. Zhang and X. Zhang, *Multi-density DBSCAN algorithm based on density levels partitioning*, J. Inf. Comput. Sci. **9** (2012), no. 10, 2739–2749.

[36] A.X. Yang, *How to develope new approaches to RFM segmentation*, J. Target. Measur. Anal. Market. **13** (2004), no. 1, 50–60.

[37] A. Yusefizad and A. Sorayaie, *Customer review and clustering, based on the RFM model and design a model to provide services to key customers*, J. Execut. Manag. **10** (2001), no. 20, 175–198.