Scientia Iranica E (2020) 27(5), 2621–2634

Sharif University of Technology
Scientia Iranica
Transactions E: Industrial Engineering
http://scientiairanica.sharif.edu

A novel selective clustering framework for appropriate labeling of clusters based on K-means algorithm

F. Moslehi, A. Haeri*, and M.R. Gholamian

School of Industrial Engineering, Iran University of Science & Technology, Tehran, Iran.

Received 26 May 2018; received in revised form 21 October 2018; accepted 16 February 2019

KEYWORDS
Machine learning;
Data mining;
Clustering;
K-means algorithm;
Labeling of the clusters.

Abstract. Data mining is a powerful new technology to extract hidden information from data warehouses. Data mining analyzes data from different perspectives and finds useful patterns and knowledge from large volumes of raw data. Clustering is one of the main methods of data mining. K-means algorithm is one of the most common clustering algorithms due to its efficiency and ease of use. One of the challenges of clustering is to identify the appropriate label for each cluster. The selection of a label is done so as to provide a proper description of cluster records. In some cases, choosing an appropriate label is not easy due to the results and structure of each cluster. The aim of this study is to present an algorithm based on the K-means clustering in order to facilitate the allocation of labels to each cluster.

1. Introduction

The data analysis technique is a widespread, interesting research domain in the pattern recognition research community due to the rapid development of computer science and technologies. One of the most important phases in Data Mining (DM) is cluster analysis. In order to cluster, some multi-objective algorithms can be used to automatically partition data [1]. It should be noted that DM consists of a set of computational techniques applied to discover knowledge, hidden patterns, and rules obtained from data in various sciences [2]. DM is concerned with finding the hidden relationships present in business data in order to allow businesses to make predictions for future use [3]. Clustering is a method of splitting a set of records into clusters in a way that records of the same cluster are more analogous to each other than records in a different cluster according to some defined criteria. The K-means algorithm aims to categorize some points into some groups based on a distance scale [4–6]. Due to the lack of a labeled clustering method, it is more difficult to implement than supervised DM methods [7]. It is well known that K-means clustering [8] has become a very common method for splitting the high-dimensional datasets with numerical features.

The K-means type clustering algorithms [8,9] are widely used for real-world applications such as marketing research [10] and DM because of their capability and ability to work with the numerical and categorical datasets [11].

One of the challenges of clustering is the determination of an appropriate label for each cluster. The selection of a label has been done to provide a proper description of cluster records. In some cases, the selection of an appropriate label cannot be easily accomplished due to the results and structure of each cluster. This paper outlines an approach based on the K-means clustering algorithm to facilitate the

*. *Corresponding author. Tel.: +98 21 73225019; Fax: +98 21 73225098*
E-mail addresses: moslehi_fateme@ind.iust.ac.ir (F. Moslehi); ahaeri@iust.ac.ir. (A. Haeri); Gholamian@iust.ac.ir (M.R. Gholamian)

doi: 10.24200/sci.2019.51110.2010

allocation of labels to each cluster. Further, in many DM issues, the dataset contains a large number of fields that specify important fields, and extracting a subset of the fields is very important. By considering the proposed approach, the important and influential variables of the dataset can be identified and the subset of the required fields can be selected. This research attempts to answer the following questions:

- Which variables are important and effective in the dataset clustering process?
- How does the process of formation and cluster change in choosing a subset of the main dataset and adding the step of fields and records?

The K-means clustering algorithm is one of the most widely used clustering techniques. Running K-means algorithm is required to determine the number of clusters and the initial states [12]. Some k points as initial centroids are generated by this method. At the next step, every point is allocated to the cluster with the closest centroid [13,14]. After that, each cluster's centroid is updated. Some data points may shift from one cluster to another. Again, the new centroids are calculated and the data points are allocated to proper clusters. Assigning and updating the centroid continue until the convergence criteria are met, i.e., no point changes clusters, or equivalently until the centroids remain the same. In this algorithm, a usual method to compute the distance between data points and centroids is the Euclidean distance [15].

In the literature, there are many pieces of research on expanding the K-means algorithm; for instance, San et al. [16] proposed a K-means-like algorithm for clustering categorical data. The clustering performance of the algorithm is indicated by two standard datasets, and the obtained results demonstrated that the new algorithm achieved more accurate results than the K-modes algorithm. Yuan et al. [13] introduced the initial centroids algorithm. In the proposed model, the initial centroids were regularly computed. Huang et al. [11] introduced W-K-means as a novel K-means type algorithm. Variable's weights were computed, and the data points were assigned based on these weights. These weights were defined according to Variable's variance. Fahim et al. [17] proposed an enhanced approach based on the K-means algorithm. In this algorithm, the allocation of data points to the appropriate clusters was done with lower time complexity. Ahmed [18] proposed an efficient enhanced K-means clustering algorithm. This algorithm was designed to cluster both the mixed numerical and categorical features. Arai and Barakbah [19] proposed a hierarchical approach for determining better initial cluster centers for the K-means clustering algorithm. In Laszlo and Mukherjee [20], a novel algorithm was presented based on genetic algorithm for determining initial centroids. Genetic algorithm was used to swap neighboring centers for the K-means algorithm. Zalik [21] applied a new K-means algorithm without pre-determining the exact number of clusters. Kao et al. [22] introduced a hybrid clustering technique. The proposed algorithm combined the K-means algorithm, Nelder-Mead simplex search, and Particle Swarm Optimization (PSO). Zhang and Xia [23] presented the initial centroids algorithm based on the K-means to avoid selecting initial centroid randomly. Nazeer and Sebastian [15] introduced an enhanced K-means algorithm to select initial centroids and allocate data points to suitable clusters simultaneously to reduce time complexity. The proposed algorithm performed more accurately and efficiently than the standard K-means algorithm. A novel approach was used by Yedla et al. [24] to determine better initial centroids and provide an efficient way to assign the data points to proper clusters with lower time complexity. Any additional input like threshold values is not considered in this method. The proposed method increased the accuracy of the clustering results. Niknama et al. [25] proposed a different hybrid algorithm named Hybrid K-MICA based on imperialist competitive algorithm and the K-means clustering algorithm. The results showed that the proposed hybrid K-MICA could be considered as an effective technique order to cluster and assign N points to K clusters. In [26], a novel hybrid method was designed for clustering data. This algorithm is a combination of two powerful optimization algorithms: K-means and firefly algorithms. In order to find optimal cluster centers, firefly algorithm was applied and, then, the centers were refined by the K-means algorithm. Celebi et al. [27] reviewed and compared K-means initialization algorithms according to their computational efficiency. Eight popular linear time initialization methodologies on a large and diverse collection of real and synthetic datasets were investigated by different performance indexes. Tzortzis and Likas [28] introduced MinMax K-means algorithm to solve the initialization problem of the K-means algorithm by determining the weights of the clusters. These weights were assigned to each cluster based on their variance. Weights and cluster assignments were treated together in an iterative process. Duan et al. [1] presented a different multilayer data clustering framework based on feature selection and modified the K-means algorithm. They attempted to reduce the dimension of the dataset by selecting an envoy feature subset as a result of the clustering process. Partial distance strategy was used in the proposed algorithm. Guérin et al. [29] proposed a novel clustering algorithm called gap-ratio K-means based on the K-means algorithm. This algorithm can be used for high-dimensional spaces. For this purpose, weights were determined for each dimension of the

feature space before running the K-means algorithm. Lin [30] proposed a new algorithm to enhance the performance of K-means clustering through the linear transformation and the random perturbation of the kernel matrix. Nagwani and Sharaff [31] applied the K-means clustering algorithm and text mining technique for separating and identifying spam and non-spam SMS messages. Chen et al. [32] applied a novel algorithm called ordered K-means clustering algorithm to improve the performance of K-means clustering. They considered K-means clustering as a multi-criteria problem and used PROMETHEE to calculate the closeness of points. Gan and Ng [33] proposed a novel algorithm to provide outlier detection by extending K-means clustering. The KMOR algorithm is able to cluster the data and remove outlier points concurrently.

This paper is organized into five sections. Section 2 outlines the proposed methodology. Sections 3 and 4 present the experimental results and discussions, respectively. Section 5 concludes the paper and provides some insights into future trends.

2. The proposed model

In the field of DM, clustering has many applications. Clustering is an important process in engineering and other fields of scientific research [34]. It is a process in which the records are categorized into separate groups so that the records in each cluster can be highly similar to each other and have the greatest difference from other data groups or clusters [35–38]. The clustering techniques focus on identifying the groups of similar records and naming the records according to the cluster to which they belong. This process takes place without the prior knowledge of the clusters and their features. Clustering is an unsupervised learning task that aims at decomposing a given set of objects into subgroups or clusters based on similarity [39]. One of the purposes of clustering is the better recognition of the dataset.

The major challenge of the clustering process to determine an appropriate label for each cluster. In some cases, due to the particular structure type of the clusters, choosing an appropriate label is not easily possible. In this research, an approach based on the K-means clustering algorithm is presented in order to facilitate the proper allocation of the labels to each cluster. Through step-by-step clustering of the dataset and examining the process of forming clusters, the activity of labeling will be done based on the subset if the allocation of the labels to the clusters is done appropriately. In fact, the main idea of the proposed method is that it is not possible in some cases to determine the label of the entire dataset and, then, it is possible to select a part of the dataset that represents the entire dataset and determines the label of a cluster based on the cluster formed in this part of the dataset. Moreover, in this process, important variables are identified in the clustering process and these variables are expressed as the mentionable indicators in the labeling. The labels are determined according to the effective fields. In conclusion, the principal advantage of this method is the possibility of monitoring, observing, and analyzing the clustering changes, compared to the one-time clustering of datasets. By analyzing these changes, the effective features of the clustering process and the dataset are determined and, also, the appropriate labels of each cluster are selected.

2.1. The proposed algorithm

The main idea of the proposed method is to provide a step-by-step clustering rather than one-time clustering of the dataset. In order to observe and analyze the formation of clusters at each stage, changes need to be made. To this end, the algorithm receives a dataset as input and a small subset of the original dataset is selected. Therefore, in accordance with the dimensions of the dataset, a number of records and fields are determined by the user at each step. After selecting an initial subset of the original set, K-means clustering algorithms run on the selected dataset. At this point, the first clusters will be formed. Then, the clustering process will be repeated at each step until the completion of the dataset fields and records. At each step, a number of fields and records will be added to the initial subset and the selected subset will be wider. After adding fields and records at any time, the K-means clustering algorithm will run on the selected dataset. Then, the clustering results are recorded and compared with those of the previous steps. Changes made to the structure of the clusters and records belonging to each cluster will be investigated. The first step is to select an initial dataset from the entire original dataset. First, the records for this primary subset should be determined. To do so, the clustering with a large number of clusters as selected by the user is implemented on the entire dataset. After that, a record is selected randomly from each cluster. At this point, the first records are selected. In the following, the features of the initial subset must be determined. Then, the variance of the fields on the selected records of the previous step is calculated, and the fields with the highest variance are selected. Thus, the initial dataset is formed. At the next step, the clustering with the desired number is applied to the initial dataset. Then, the records and fields are added to the initial dataset step by step. At each step, the clustering with the K-means algorithm is implemented on the selected dataset, and the clustering results are recorded and compared with those of previous steps. The process of forming clusters is monitored and the required analysis will be performed based on the changes made to the structure of the clusters and the records belonging to

each cluster and an appropriate label for each cluster will be selected. The way to select the records is based on the fact that a large number of clusters are carried out with respect to the dimensions of the dataset, and a record is randomly selected from each cluster. This clustering aims to select the records characterized by the most differences. How the features are chosen is dependent on the variance of the variables. The aim of this step is also to select the features that have made significant differences among the records. Determining the number of fields and records at each step will be selected by the user according to the dimensions of the dataset. The proposed algorithm proceeds following the steps outlined below:

- **Step 1:** *Select the records of the initial subset.* During the first phase, the clustering algorithm with a large number of clusters will be run on the whole dataset with respect to the dimensions of the dataset. Then, a record will be randomly selected from each cluster and, consequently, the records of the initial subset will be determined. The purpose of this step is to select the records characterized by the most differences;
- **Step 2:** *Select the features of the initial subset.* At this step, the features of the first subset are determined. Thus, at first, the variance of all dataset variables is computed for the selected records of the previous stage and the features with the highest variance are selected. The number of these variables is determined by the user's choice and is determined according to the dimensions of the dataset;
- **Step 3:** *Implement the K-means clustering algorithm on the initial subset.* After forming the initial selected subset, the K-means clustering algorithm will be implemented on it and the formed clusters will be investigated;
- **Step 4:** *Add some of the records randomly to the selected dataset.* The initial dataset will be extended gradually. To do so, some records will be selected randomly from the main dataset and added to the selected subset. The number of records will be determined by the user;
- **Step 5:** *Calculate the variance of the variables and select the fields.* At this step, the variance of all the dataset variables on the records of the selected subset will be calculated. Some of the fields with the highest variance will be selected which form the selected dataset of this step;
- **Step 6:** *Implement the K-means clustering algorithm on the selected dataset.* Again, the K-means clustering algorithm will be implemented on a new dataset this time, and its results will be examined.

The fourth to sixth steps will be repeated until all the fields and records are added to the selected dataset. After completing the steps, the clustering results recorded at each step are compared and the process of making changes to the structure of the clusters will be analyzed. In the meantime, if the addition of a variable to the selected subset causes changes in the structure of the clusters, that variable will be selected as an important and effective variable of the dataset. Moreover, if it is possible to determine an appropriate label for clusters in a step, the subset related to that step will be selected as representative of the entire dataset for determining the cluster label. The logic of each step is explained in Table 1. The pseudocode of the proposed algorithm is described in Figure 1.

3. Experimental results

3.1. The algorithm implementation by the hybrid addition of the fields and records step by step

This paper has investigated several experiments to demonstrate the effectiveness of the proposed algorithm. In this section, a dataset containing 385 samples and 50 fields was clustered using the proposed approach. This dataset contains statistics on various banking services of electronic payment instruments including ATM, Pin Pad, Mobile banking, POS, Phone banking, and Internet banking in Iran in 2015. The statistics provided are ordered by month and the

Table 1. The logic of each step.

Step	Logic
Step 1	The purpose of this step is to select the records with the most differences
Step 2	The aim of this step is to select the features that have created the most difference among the records
Step 3	The initial subset is clustered to determine its structure
Step 4	Random records are selected to try different ways to reach the final data
Step 5	The field is added to allow a comparison of the built subsets
Step 6	The consecutive clustering provides the ability to observe the trend of changes

```
Begin
    Select the initial data set from the entire raw data set
        do {
                Cluster with the desired number on the dataset
                Select a record from each cluster randomly
                Compute the variance of the variables in the selected set of records
                Select some fields with the highest degree of variance
            }
        Run the K- means clustering algorithm on the initial data set
        While (all the fields and records are added to the selected data set)
            do {
                Add some of the records randomly to the selected dataset
                Compute the variance of the variables
                Choose the fields with the highest variance
                Run the K-means algorithm on the initial data set
            }
End
```

Figure 1. The pseudocode of the proposed algorithm.

registered service provider to the bank. Analyzing bank databases for analyzing customer behavior is difficult since bank databases are multi-dimensional, composed of monthly account records and daily transaction records [39].

Table 2 lists the fields in this dataset. Since the variables are of different values, the field values were normalized so that the effects of all the fields in the analysis could be the same.

At each step, a subset of the original dataset is selected, and the process of forming the clusters and allocating a label to each cluster is examined. For this purpose, clustering is initially performed on the entire dataset with 50 clusters, and from each cluster, one record is randomly selected as the record of the initial selected subset. To select the dataset fields, the variance of all variables is calculated in the 50 selected records of the previous stage, and 7 fields with the highest variance are selected as the subset of fields of the first step. Thus, the selected subset of the first stage is formed. At the next steps, each time, a number of records are added randomly to the subset of the previous records and the variance of all the variables is calculated in the subset of the records. Some fields with more variance are added to the selected dataset. The mean values of all fields related to each instrument are recorded as the result in the tables. Table 3 shows the performance and the rank of each bank in the amount and the number of electronic payment transactions in 2015. The cluster label is also based on the same statistics published on the website (www.shaparak.ir), which shows their performance in the area of electronic payment. According to the published statistics, the performance of the banks is classified into four categories of privileged, great, fair, and poor.

At first, the standard K-means clustering algorithm was implemented with 4 clusters on the entire original dataset. Given the results and the clusters formed, it is not possible to determine an appropriate label for Clusters 2 and 4. The results of implementing the traditional K-means clustering algorithm are presented in Table 4, Figures 2, and 3.

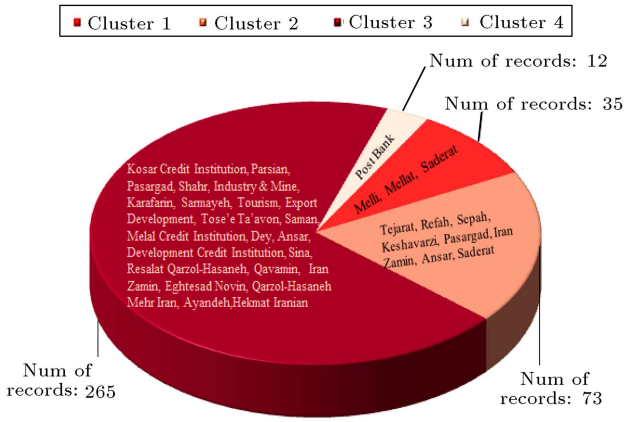

Figure 2. Distribution of banks in each cluster in general clustering.

Description of clusters:

Cluster 1: Long-standing banks with privileged performance in the e-payment. Records related to the three banks of Melli, Mellat, and Saderat that have a long working history and are premier in electronic payment have been located in this cluster;

Cluster 2: Most of the records of this cluster have been made by old banks which play a great role in electronic payments, but records of Ansar, Iran Zamin, and Pasargad banks with less experience than the other banks of the cluster and a poorer performance in the electronic payments are also located in

Table 2. Description of the variables in the bank's dataset.

Variable number	Variable name	Variable number	Variable name
1	Bank	26	Number of intra-bank transfer via Pin Pad
2	Month	27	Amount of transfer/withdrawal via Pin Pad
3	Amount of withdrawals from ATM	28	Number of transfer/withdrawal via Pin Pad
4	Number of withdrawals from ATM	29	Average amount of transfer/withdrawal via Pin Pad
5	Average amount of withdrawals from ATM	30	Amount of transfer/deposit via Pin Pad
6	Number of account balance via ATM	31	Number of transfer/deposit via Pin Pad
7	Number of account check via ATM	32	Average amount of transfer/deposit via Pin Pad
8	Number of intra-bank transfer via ATM	33	Amount of transfer via Shetabi Card with Pin Pad
9	Amount of transfer/withdrawal via ATM	34	Number of transfer via Shetabi Card with Pin Pad
10	Number of Transfer/withdrawal via ATM	35	Average amount of transfer via Shetabi Card with Pin Pad
11	Average amount of transfer/withdrawal via ATM	36	Amount of buying via mobile banking
12	Amount of transfer/deposit via ATM	37	Number of account balance via mobile banking
13	Number of transfer/deposit via ATM	38	Amount of bills pay via mobile banking
14	Average amount of transfer/deposit via ATM	39	Number of bills pay via mobile banking
15	Amount of transfer via Shetabi Card with ATM	40	Average amount of bills pay via Mobile banking
16	Number of Ttansfer via Shetabi Card with ATM	41	Number of account balance via phone banking
17	Average amount of transfer via Shetabi Card with ATM	42	Amount of bills pay via phone banking
18	Amount of bills pay via ATM	43	Number of bills pay via phone banking
19	Number of bills pay via ATM	44	Average amount of bills pay via phone banking
20	Average amount of bills pay via ATM	45	Amount of buying via POS
21	Amount of withdrawals from Pin Pad	46	Amount of buying via internet banking
22	Number of withdrawals from Pin Pad	47	Number of account check via internet banking
23	Average amount of withdrawals from Pin Pad	48	Amount of transfer/deposit via internet banking
24	Number of account balance via Pin Pad	49	Number of transfer/deposit via internet banking
25	Number of account check via Pin Pad	50	Average amount of transfer/deposit via Internet banking

this cluster. There are 12 records, i.e., 16.90% of all the records in this cluster that belong to the three banks of Ansar, Iran Zamin, and Pasargad. With the combination of the banks in this cluster, it is not possible to determine a proper label for this cluster;

Cluster 3: The private and young banks with poor performance in the e-payment. Records related to the private and young banks belong to this cluster. The performance of these banks in the area of electronic payment was poorer than those of the other banks;

Cluster 4: This cluster only includes Post Bank records. No appropriate label is specified for this cluster.

At the next step, the proposed algorithm was implemented on a given dataset. These steps are taken to implement the algorithm and the obtained results are mentioned in the following.

Step 1: Herein, the initial dataset is selected. To do so, the initial dataset comprised 50 records and 10 fields. The variance of all variables was calculated in the selected records, and 10 variables of highest variance were selected. The fields including

Table 3. Bank ranking in e-payment (www.shaparak.ir).

Rank	Total amount of e-payment transactions	Total number of e-payment transactions
1	Mellat	Mellat
2	Bank Melli Iran	Bank Melli Iran
3	Bank Saderat Iran	Bank Saderat Iran
4	Tejarat	Parsian
5	Keshavarzi	Keshavarzi
6	Sepah	Tejarat
7	Refah	Refah
8	Pasargad	Sepah
9	Parsian	Saman
10	Saman	Pasargad
11	Maskan	Eghtesad Novin
12	Ansar	Ansar
13	Qavamin	Maskan
14	Eghtesad Novin	Resalat Qarzol-Hasaneh
15	Resalat Qarzol-Hasaneh	Post Bank
16	Ayandeh	Shahr
17	Shahr	Qavamin
18	Sina	Sina
19	Qarzol-Hasaneh Mehr Iran	Ayandeh
20	Tose'e Ta'avon	Qarzol-Hasaneh Mehr Iran
21	Post Bank	Tose'e Ta'avon
22	Iran Zamin	Hekmat Iranian
23	Tourism	Iran Zamin
24	Dey	Dey
25	Sarmayeh	Tourism
26	Karafarin	Sarmayeh
27	Hekmat Iranian	Karafarin
28	Industry & Mine	Industry & Mine
29	Export Development Bank of Iran	Export Development Bank of Iran
30	Central Bank of Iran	Central Bank of Iran
31	Melal Credit Institution	Kosar Credit Institution
32	Kosar Credit Institution	Melal Credit Institution
33	Development Credit Institution	Development Credit Institution

Table 4. Results of clustering the total bank's dataset by traditional K-means algorithm.

Cluster	ATM	Pin pad	Mobile banking	Phone banking	POS	Internet banking
1	0.503	0.398	0.115	0.195	0.576	0.394
2	0.232	0.175	0.055	0.067	0.162	0.157
3	0.108	0.102	0.022	0.040	0.033	0.116
4	0.115	0.313	0.032	0.026	0.025	0.096

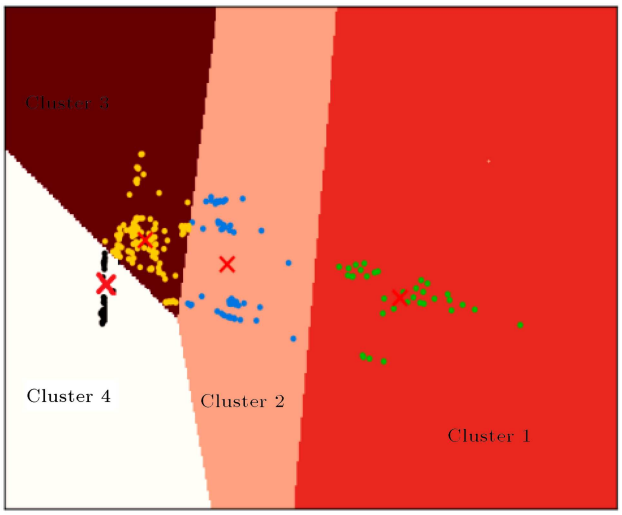

Figure 3. Results of clustering the total bank's dataset.

"Number of withdrawals from ATM", "Number of account balance via ATM", "Number of account check via ATM", "Amount of transfer /deposit via ATM"," Amount of transfer via Shetabi Card with ATM"," Number of bills pay via ATM", "Number of intra-bank transfer via Pin Pad", "Number of transfer/withdrawal via Pin Pad", "Number of transfer/deposit via Pin Pad", and "Number of transfer via Shetabi Card with Pin Pad" were defined as 10 fields of the initial dataset. Then, the K-means clustering algorithm with 4 clusters was implemented on this dataset. Results of this clustering are shown in Table 5, Figures 4 and 5.

Description of the clusters: In comparison with the clustering of the entire dataset, the clustering of the subset including 10 fields and 50 records managed to create a difference between the structures of Clusters 1 and 4. As shown in the Table 4, the post bank's records located in the 4th cluster in the previous stage are part of Cluster 3 at this step. In addition, the records of Melli, Mellat, and Saderat banks are divided into two clusters of 1 and 4. According to the results, choosing this subset to determine the label of clusters is not suitable.

Step 2: 60 random records and 7 fields of "Average amount of withdrawals from Pin Pad",

Table 5. Results of clustering the 10 fields and 50 records.

Cluster	ATM	Pin pad
1	0.771	0.752
2	0.377	0.211
3	0.074	0.030
4	0.678	0.470

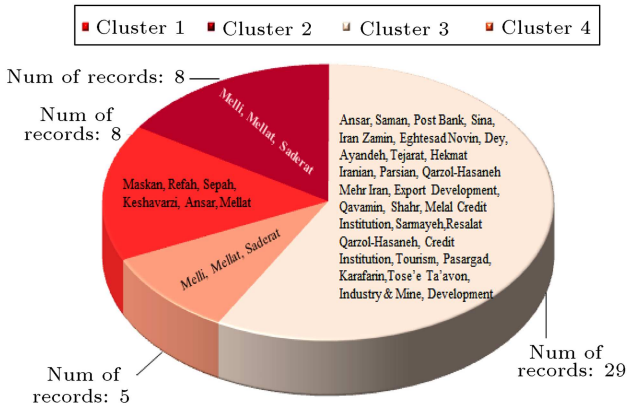

Figure 4. Distribution of banks in each cluster in 10 fields and 50 records clustering.

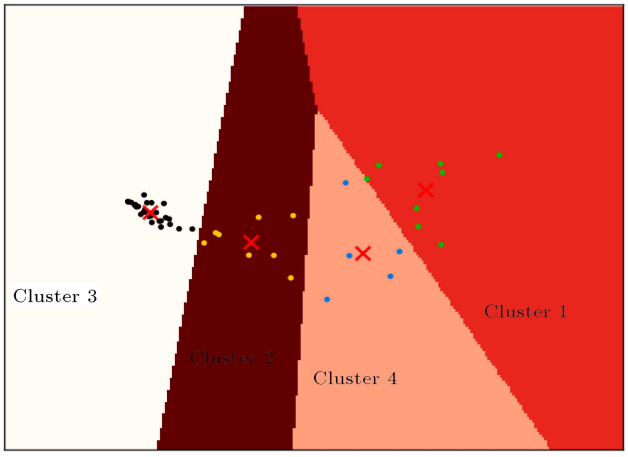

Figure 5. Results of clustering 10 fields and 50 records.

"Number of transfer/withdrawal via ATM", "Number of transfer/deposit via ATM", "Amount of transfer/withdrawal via ATM", "Amount of transfer/withdrawal via Pin Pad", "Amount of transfer/deposit via Pin Pad", and "Amount of transfer via Shetabi Card with Pin Pad" with the highest variance values were added to the initial selected subset. The K-means clustering algorithm was implemented with 4 clusters on the dataset including 17 fields and 110 records. The results of this clustering are shown in Table 6, Figures 6, and 7.

Given the results and the formation of the

Table 6. Results of clustering the 17 fields and 110 records.

Cluster	ATM	Pin pad
1	0.697	0.635
2	0.044	0.028
3	0.089	0.109
4	0.352	0.253

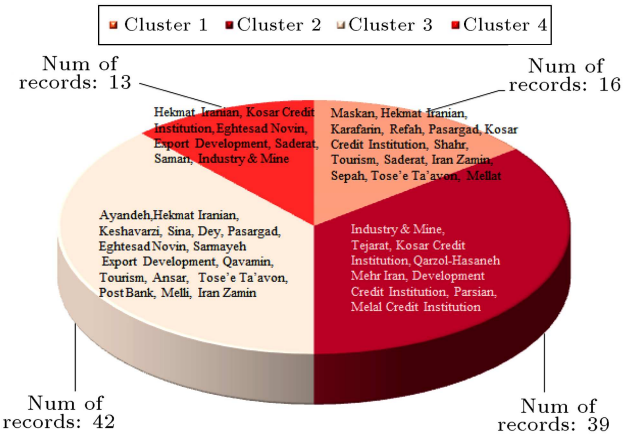

Figure 6. Distribution of banks in each cluster in 17 fields and 110 records clustering.

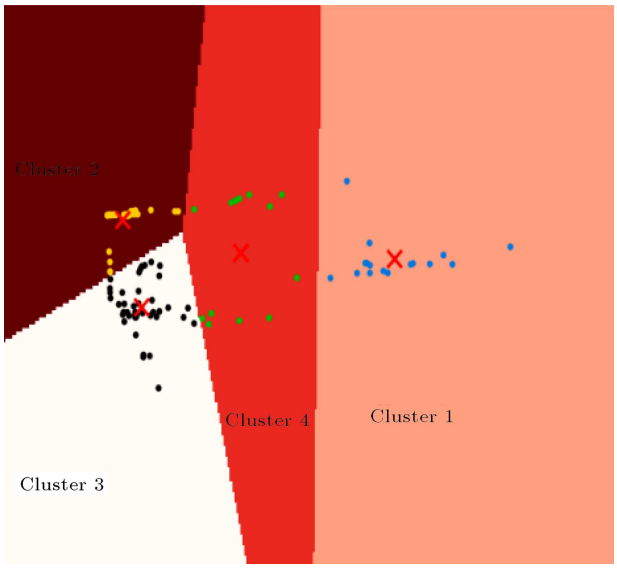

Figure 7. Results of clustering 17 fields and 110 records.

clusters, it is not possible to allocate an appropriate label to clusters.

Step 3: 80 random records and the 13 fields of "Amount of withdrawals from ATM", "Average amount of withdrawals from ATM", "Number of transfer via Shetabi Card with ATM", "Amount of bills pay via ATM", "Number of intra-bank transfer via ATM", "Amount of buying via POS", "Amount of buying via internet banking", "Amount of transfer/deposit via internet banking", "Number of withdrawals from Pin Pad", "Number of account balance via phone banking", "Number of bills pay via phone banking", "Number of account check via Pin Pad", and "Number of account balance via Pin Pad" having the highest variance values were added to the initially selected subset. After extracting the dataset in this stage including 30 fields and 190 records, the K-means clustering algorithm was run with 4 clusters. The results of this clustering are presented in Table 7, Figures 8, and 9.

Description of the clusters

Cluster 1: Long-standing banks with privileged performance in e-payment. Records of the three powerful and old banks including Melli, Mellat, and Saderat are located in this cluster;

Cluster 2: Long-standing banks with great performance in the e-payment. In this cluster, there are records of the old banks with great performance in the area of electronic payment. In the implementation process of the algorithm, only one record, which is equivalent to 5% of the total records of the cluster, is related to Ansar Bank. The combination of this cluster in this step is done in such a way that a proper label can be given to it;

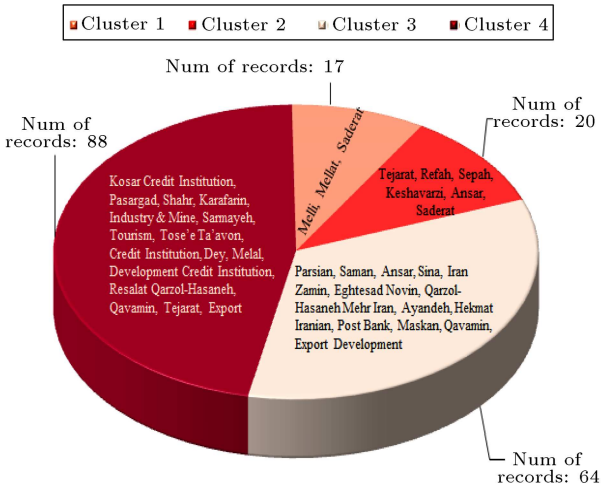

Figure 8. Distribution of banks in each cluster in 30 fields and 190 records clustering.

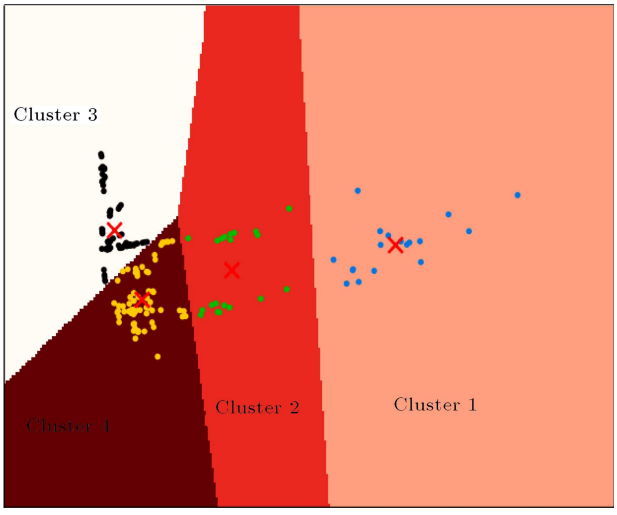

Figure 9. Results of clustering 30 fields and 190 records.

Table 7. Results of clustering 30 fields and 190 records.

Cluster	ATM	Pin pad	Phone banking	POS	Internet banking
1	0.572	0.473	0.2246	0.593	0.473
2	0.271	0.203	0.0476	0.212	0.128
3	0.078	0.112	0.0627	0.049	0.039
4	0.059	0.027	0.0003	0.027	0.021

Cluster 3: Private and young banks with fair performance in the e-payment. Records of the private and young banks with fair performance in the field of electronic payments are located in this cluster. The average number and amount of electronic payment transactions with different instruments in this cluster is higher than that in Cluster 4;

Cluster 4: Private and young banks with poor performance in the e-payment. This cluster is composed of the records of private and young banks with poor performance in the field of electronic payments. The e-banking services of this cluster have received less popularity than the other banks.

Step 4: At this step, seven fields with the highest variance and 60 records were randomly added to the selected dataset of the previous step. The eight added fields include "Average amount of bills pay via phone banking", "Amount of buying via mobile banking", "Average amount of bills pay via mobile banking", "Amount of bills pay via phone banking", "Amount of withdrawals from Pin Pad", "Number of account check via Internet banking", and "Number of transfer/deposit". Then, the K-means clustering algorithm was implemented. The selected dataset of this step includes 37 fields and 250 records, which are divided into 4 clusters. The results of this clustering are described in Table 8, Figures 10, and 11.

Description of the clusters:

By clustering a subset of 37 fields and 250 records, the results similar to the clustering of the raw dataset were obtained. As shown in Figure 10, the composition of Clusters 2 and 4 is not possible for proper labeling.

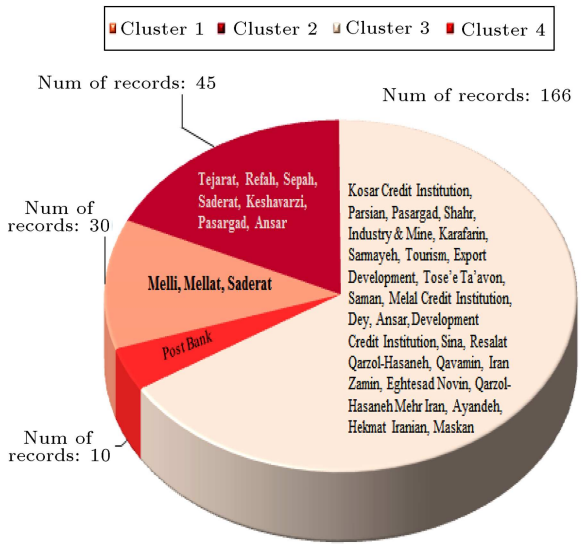

Figure 10. Distribution of banks in each cluster in 37 fields and 250 records clustering.

3.1.1. Selecting an appropriate subset for specifying the clusters' labels

According to the results of the algorithm implementation at each step, the subset of Step 3 containing 30 fields and 190 records is selected as a subset that can be used for determining the label of clusters. The records and the fields of this subset can be extracted from the main dataset and, accordingly, clustering results can be obtained and the labels of the clusters can be specified.

Table 8. Results of clustering 37 fields and 250 records.

Cluster	ATM	Pin pad	Mobile banking	Phone banking	POS	Internet banking
1	0.503	0.421	0.185	0.206	0.583	0.393
2	0.244	0.162	0.078	0.058	0.175	0.103
3	0.062	0.049	0.049	0.048	0.033	0.023
4	0.085	0.311	0.062	0.023	0.024	0.016

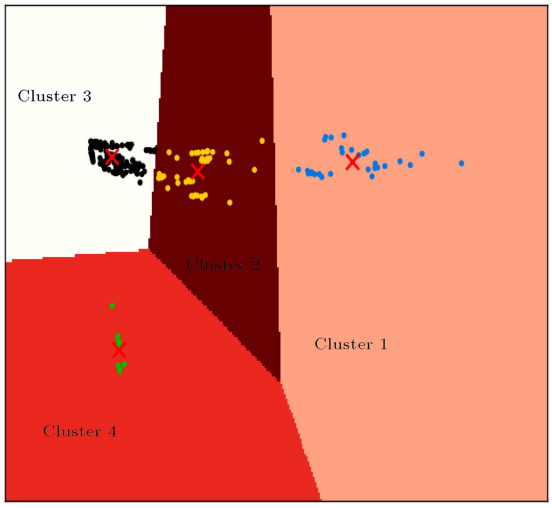

Figure 11. Results of clustering 37 fields and 250 records.

4. Discussion

The main purpose of this study is to propose a method for identifying the important and influential variables of the dataset. In the implementation phase, the algorithm is initially implemented on the selected subsets of the records and fields and, then, the fields and records of the original dataset are added to the selected subset step by step, and the process of clusters formation is monitored. The proposed algorithm was implemented on a dataset consisting of 385 records and 50 fields. This dataset contains the statistics on various electronic banking services including ATM, Pin Pad, Phone banking, Internet banking, Mobile banking, and POS in Iran in 2015. These statistics are recorded and ordered according to the month and the service provider to the bank. First, the K-means clustering algorithm was implemented with 4 clusters on the dataset. According to the clustering results, it was not possible to allocate an appropriate label to the clusters. The proposed algorithm was implemented on the initial dataset containing 10 fields and 50 selected records from the main dataset. Records and fields were gradually added to the selected dataset and the K-means clustering algorithm was implemented on the dataset. According to the clustering results of the dataset with 30 fields and 190 records, determining the proper labeling of the clusters was possible and this subset can be selected and extracted as part of the main dataset, and the cluster labels can be determined based on this subset. The results of this implementation are described in Table 9.

Table 9. Results of the algorithm run by the gradual addition of the fields and records.

Step	Fields' number	Number of samples	Labels
Step 1	4, 6, 7, 12, 15, 19, 26, 28, 31, 34	50	Cluster 1: The long history banks with privileged performance in the e-payment. Cluster 2: — Cluster 3: The private and young banks with poor performance in the e-payment. Cluster 4: —
Step 2	4, 6, 7, 12, 15, 19, 26, 28, 31, 34, 9, 10, 13, 23, 27, 30, 33	110	Cluster 1: — Cluster 2: — Cluster 3: — Cluster 4: —
Step 3	4, 6, 7, 12, 15, 19, 26, 28, 31, 34, 9, 10, 13, 23, 27, 30, 33, 3, 5, 8, 16, 18, 22, 24, 25, 41, 43, 46, 45, 48	190	Cluster 1: The long history banks with privileged performance in the e-payment. Cluster 2: The long history banks with great performance in the e-payment. Cluster 3: The private and young banks with fair performance in the e-payment. Cluster 4: The private and young banks with poor performance in the e-payment.
Step 4	4, 6, 7, 12, 15, 19, 26, 28, 31, 34, 9, 10, 13, 23, 27, 30, 33, 3, 5, 8, 16, 18, 22, 24, 25, 41, 43, 46, 45, 48, 21, 36, 40, 42, 44, 47, 49	250	Cluster 1: The long-standing banks with privileged performance in the e-payment. Cluster 2: — Cluster 3: The private and young banks with poor performance in the e-payment. Cluster 4: —

5. Conclusion

Clustering techniques used for accurate analysis of high-dimensional data are receiving growing interest every day. Different algorithms are presented for data clustering. The K-means algorithm is one of the most powerful and most widely used clustering algorithms that has been applied by data science experts in many data mining projects. Clustering is an unsupervised method and does not require prior knowledge of the data. It can, therefore, be used as a way to help deepen the understanding of the dataset. A serious challenge of the clustering technique is to identify the proper cluster label. In this study, an algorithm based on the K-means clustering algorithm was presented. This approach could help cluster labeling in clustering and select the influential features of the dataset. In cases where the implementation of the clustering algorithm on the entire dataset cannot determine the proper labels for the clusters, the application of the proposed algorithm makes it possible to select and extract a subset of the main dataset so that the proper labeling for the clusters becomes possible. In general, this algorithm will be beneficial as a reliable and effective procedure for solving the appropriate labeling of cluster problems.

References

1. Duan, G., Hu, W., and Zhang, Z. "A novel multilayer data clustering framework based on feature selection and modified K-means algorithm", *International Journal of Signal Processing, Image Processing and Pattern Recognition*, **9**(4), pp. 81–90 (2016). http://dx.doi.org/10.14257/ijsip.2016.9.4.08

2. Haeri, A. and Tavakkoli-Moghaddam, R. "Developing a hybrid data mining approach based on multi-objective particle swarm optimization for solving a traveling salesman problem", *Journal of Business Economics and Management*, **13**(5), pp. 951–967 (2012).

3. Moslehi, F., Haeri, A., and Moini, A. "Analyzing and investigating the use of electronic payment tools in Iran using data mining techniques", *Journal of AI and Data Mining*, **6**(2), pp. 417–437 (2018). DOI: 10.22044/jadm.2017.5352.1643

4. Amezquita-Sanchez, J.P. and Adeli, H. "Feature extraction and classification techniques for health monitoring of structures", *Scientia Iranica*, **22**(6), pp. 1931–1940 (2015).

5. Cheng, T., Li, P., Zhu, S., and Torrieri, D. "M-cluster and X-ray: Two methods for multi jammer localization in wireless sensor networks", *Integrated Computer-Aided Engineering*, **21**(1), pp. 19–34 (2014).

6. Goncalves, N., Nikkilä, J., and Vigario, R. "Selfsupervised MRI tissue segmentation by discriminative clustering", *International Journal of Neural Systems*, **24**(1), 1450004 (2014).

7. Saxena, A., Prasad, M., Gupta, A., et al. "A review of clustering techniques and developments", *Neurocomputing*, **267**, pp. 664–681 (2017).

8. MacQueen, J.B. "Some methods for classification and analysis of multivariate observations", *Proc. 5th Symp. Mathematical Statistics and Probability*, Berkelely, CA, **1**, pp. 281–297 (1967).

9. Huang, Z. "Extensions to the k-means algorithms for clustering large data sets with categorical values", *Data Min Knowl Disc*, **2**, pp. 283–304 (1998).

10. Green, P.E., Kim, J., and Carmone, F.J. "A preliminary study of optimal variable weighting in k-means clustering", *Journal of Classification*, **7**(2), pp. 271–285 (1990).

11. Huang, J.Z., Ng, M.K., Rong, H., et al. "Automated variable weighting in k-means type clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(5), pp. 657–668 (2005). https://doi.org/10.1109/TPAMI.2005.95

12. He, Z. "Evolutionary K-means with pair-wise constraints", *Soft Computing*, **20**(1), pp. 287–301 (2016).

13. Yuan, F., Meng, Z.H., Zhang, H.X., and Dong, C.R. "A new algorithm to get the initial centroids", *Proceedings of the 3rd International Conference on Machine Learning and Cybernetics*, pp. 26–29 (2004).

14. Zhang, C.H. and Xia, S.H. "K-means clustering algorithm with improved initial center", In *Second International Workshop on Knowledge Discovery and Data Mining (WKDD)*, pp. 790–792 (2009).

15. Nazeer, K.A.A. and Sebastian, M.P. "Improving the accuracy and efficiency of the k-means clustering algorithm", In *Proceedings of the World Congress on Engineering*, **1**, pp. 1–5 (2009).

16. San, O.M., Huynh, V.N., and Nakamori, Y. "An alternative extension of the k-means algorithm for clustering categorical data", *International Journal of Applied Mathematics and Computer Science*, **14**, pp. 241–247 (2004).

17. Fahim, A.M., Salem, A.M., Torkey, F.A., et al. "An efficient enhanced k-means clustering algorithm", *Journal of Zhejiang University-Science*, **7**(10), pp. 1626–1633 (2006).

18. Ahmad, A. "k-mean clustering algorithm for mixed numeric and categorical data", *Data & Knowledge Engineering*, **63**, pp. 503–527 (2007). https://doi.org/10.1016/j.datak.2007.03.016

19. Arai, K. and Barakbah, A.R. "Hierarchical K-means: an algorithm for centroids initialization for K-means",

Reports of the Faculty of Science and Engineering, **36**, pp. 25–31 (2007).

20. Laszlo, M. and Mukherjee, S.A. "Genetic algorithm that exchanges neighboring centers for k-means clustering", *Pattern Recognit. Lett.*, **28**(16), pp. 2359–2366 (2007).
21. Zalik, K.R. "An efficient k-means clustering algorithm", *Pattern Recognit. Lett.*, **29**, pp. 1385–1391 (2008).
22. Kao, Y.T., Zahara, E., and Kao, I.W. "A hybridized approach to data clustering", *Expert Syst. Appl.*, **34**(3), pp. 1754–1762 (2008).
23. Zhang, C.H. and Xia, S.H. "K-means clustering algorithm with improved initial center", In *Second International Workshop on Knowledge Discovery and Data Mining (WKDD)*, pp. 790–792 (2009).
24. Yedla, M., Pathakota, S.R., and Srinivasa, T.M. "Enhancing K-means clustering algorithm with improved initial center", *International Journal of Computer Science and Information Technologies*, **1**(2), pp. 121–125 (2010).
25. Niknama, T., Fard, E.T., Pourjafarian, N., et al. "An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering". *Eng. Appl. Artif. Intel.*, **24**(2), pp. 306–317 (2011). https://doi.org/10.1016/j. engappai.2010.10.001
26. Hassanzadeh, T. and Meybodi, M.R. "A new hybrid approach for data clustering using firefly algorithm and K-means", *The 16th CSI International Symposium on Artificial Intelligence and Signal Processing*, AISP, 007–011 (2012). https://doi.org/10.1109/AISP.2012.6313708
27. Celebi, M.E., Kingravi, H.A., and Vela, P.A. "A comparative study of efficient initialization methods for the k-means clustering algorithm", *Expert Syst Appl.*, **40**(1), pp. 200–210 (2013). https://doi.org/10.1016/j.eswa.2012.07.021
28. Tzortzis, G. and Likas, A. "The MinMax k-means clustering algorithm". *Pattern Recognit*, **47**(7), pp. 2505–2516 (2014).
29. Guérin, J., Gibaru, O., Thiery, S., and Nyiri, E. "Clustering for different scales of measurement-the gap-ratio weighted K-means algorithm", *arXiv Preprint arXiv*, pp. 1703.07625 (2017).
30. Lin, K.P. "Privacy-preserving kernel k-means clustering outsourcing with random transformation", *Knowledge and Information Systems*, **49**(3), pp. 1–24 (2016).
31. Nagwani, N.K. and Sharaff, A. "SMS spam filtering and thread identification using bi-level text classification and clustering techniques", *Journal of Information Science*, **43**(1), pp. 75–87 (2017).
32. Chen, L., Xu, Z., Wang, H., and Liu, S. "An ordered clustering algorithm based on K-means and the PROMETHEE method", *International Journal of Machine Learning and Cybernetics*, **9**(6), pp. 917–926 (2018).
33. Gan, G. and Ng, M.K.P. "k-means clustering with outlier removal", *Pattern Recognition Letters*, **90**, pp. 8–14 (2017).
34. Yaghini, M. and Ghazanfari, N. "Tabu-KM: a hybrid clustering algorithm based on tabu search approach", *International Journal of Industrial Engineering & Production Research*, **21**(2) pp. 71–79 (2010).
35. Han, J., Kamber, M., and Tung, A.KH., *Spatial Clustering Methods in Data Mining: A Survey*, London: Taylor & Francis (2001).
36. Jain, A.K., Murty, M.N., and Flynn, P.J. "Data clustering: A review", *ACM Computing Surveys 1999*, **31**, pp. 264–323 (1999).
37. Maimon, O.Z. and Rokach, L., *Data Mining and Knowledge Discovery Handbook*, New York, Springer (2005).
38. Fazel, Z.M. and Zarinbal, M. "Image segmentation: Type-2 fuzzy possibilistic C-mean clustering approach", *Journal of Industrial Engineering & Production Research*, **23**(4) pp. 245–251 (2012).
39. Farajian, M.A. and Mohammadi, S. "Mining the banking customer behavior using clustering and association rules methods", *Journal of Industrial Engineering & Production Research*, **21**(4), pp. 239–245 (2010).

Biographies

Fateme Moslehi received her BSc degree in Computer Engineering from University of Shahid Rajaee, Iran in 2014. She pursued MSc degree in 2015 and received it in Information Technology Engineering from Iran University of Science and Technology in 2017. Her research interests are data mining techniques like rule mining and clustering and soft computing methods, such as genetic algorithm and particle swarm optimization.

Abdorrahman Haeri is an Assistant Professor of Industrial Engineering at Iran University of Science and Technology, Iran. He received his PhD from the Department of Industrial Engineering, College of Engineering, and University of Tehran, Iran in 2013. He holds an MSc in Industrial Engineering, Sharif University of Technology, Iran in 2008. His main areas of teaching and research interests include data envelopment analysis, network design, and data mining. He has published several papers in international conferences and academic journals including Safety Science, Scientometrics, etc.

Mohammad Reza Gholamian is an Associate Professor in School of Industrial Engineering at the Iran University of Science and Technology (IUST), Tehran, Iran. He received his MS degree in Industrial Engineering from Isfahan University of Technology (IUT), Isfahan in 1998 and obtained PhD in Industrial

Engineering from Amirkabir University of Technology (AUT), Tehran in 2005. Presently, he is a faculty member of Systems Engineering Group in School of Industrial Engineering and is actively engaged in conducting Academic, Research and Development Programs in the field of Industrial Engineering. He has contributed more than 172 research papers to many national and international journals and conferences. Besides, he has published 5 books by reputed publishers. His research interests are in the areas of inventory models, supply chain network design, and multi-criteria decision making.