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

Fateme Moslehi¹, Abdorrahman Haeri²* & Mohammad Reza Gholamian³

1. MSc student of Information Technology Engineering, Iran University of Science & Technology, Tehran, Iran, Email: moslehi_fateme@ind.iust.ac.ir.
2. Assistant Professor of School of Industrial Engineering, Iran University of Science & Technology, Tehran, Iran, Email: ahaeri@iust.ac.ir.
3. Associate Professor of School of Industrial Engineering, Iran University of Science & Technology, Tehran, Iran, Email: Gholamian@iust.ac.ir

*Corresponding author: Phone: +98-21-73225019, Fax: +98-21-73225098.

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 in such a way as to provide a proper description of the cluster records. In some cases, choosing the 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.

Keyword: Machine learning, Data Mining, Clustering, K-means algorithm, labeling of the clusters.

1. Introduction

The data analysis technique is widespread interest research area in pattern recognition research community due to computer science and technologies are developing rapidly. One of the most important phases in data mining is cluster analysis. In order to cluster, some multi-objective algorithms can be used which automatically partition the data [1]. It should be noted that data mining (DM) consists of a set of computational techniques which are 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 different cluster pursuant to some defined criteria. The K-means algorithm aims to categorize some points into some groups based on a distance scale [4, 5, 6]. Due to the lack of a labeled clustering method, it is more difficult to implement than supervised data mining methods [7]. It is well-known that K-means clustering [8] has become a very common method for splitting the high dimensional data sets with numerical features.

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

One of the challenges in clustering is the determination of an appropriate label for each cluster. The selection of a label has been done in such a way as 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 in order to facilitate the allocation of labels to each cluster. Also, in many data mining issues, the data set contains a large number of fields that specify important fields and extracting a subset of the fields is very important.
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 by choosing a subset of the main data set 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 clusters’ centroid are updated. Some data points may shift from one cluster to another. Again, the new centroids are calculated and the data points are allocated to the proper clusters. Assigning and updating the centroid are continued until the convergence criteria are met i.e. no point changes clusters, or equivalently until the centroids remain the same. In this algorithm, usual method to compute the distance between data points and centroids is Euclidean distance [15].

In the literature there are many researches for expanding the K-means algorithm for instance; San et al. [16] have proposed a K-means-like algorithm for clustering categorical data. The clustering performance of the algorithm is indicated by two standard data sets and the obtained results have demonstrated that the new algorithm achieved more accurate results in comparison with the K-modes algorithm. Yuan et al. [13] have introduced the initial centroids algorithm. In the proposed model, the initial centroids were regularly computed. Huang et al. [11] have 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] have proposed an enhanced approach based on the K-means algorithm. In this algorithm, allocating data points to the appropriate clusters was done with less time complexity. Ahmed and Dey [18] have proposed an efficient enhanced the K-means clustering algorithm. This algorithm was designed to cluster both the mixed numeric and categorical features. Arai and Barakbah [19] have proposed the hierarchical approach for determining the better initial cluster centers for the K-means clustering algorithm. In Laszlo and Mukherjee [20], a novel algorithm has been proposed based on the Genetic algorithm for determining initial centroids. Genetic algorithm is used to swap neighboring centers for the K-means algorithm. Zalik [21] has proposed a new K-means algorithm with no need to pre-determine the exact number of clusters. Kao et al. [22] have introduced a hybrid clustering technique. The proposed algorithm combined the K-means algorithm, Nelder-Mead simplex search, and PSO. Zhang and Xia [23] have presented the initial centroids algorithm based on the K-means to avoid selecting initial centroid randomly. Nazeer and Sebastian, [15] have introduced an enhanced the K-means algorithm to select initial centroids and allocating the data points to suitable clusters simultaneous reducing time complexity. The proposed algorithm has performed more accurate and efficient in comparison with the standard K-means algorithm. In [24] a novel approach was used by Yedla et al. to find the better initial centroids and providing an efficient way of assigning the data points to proper clusters with less time complexity. Any additional input like threshold values did not use in this method. The proposed method increased the accuracy of the clustering results. Niknama et al. [26] have 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 can be considered as an effective technique order to cluster and assign N points to K clusters. In [28] a novel hybrid method has been designed for clustering data by Hassanzadeh and Meybodi. This algorithm is a combination of two powerful optimization algorithms: K-means and firefly algorithm. 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. [29] have 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 data sets were investigated by different performance indexes. Tzortzis and Likas, [30] have introduced MinMax K-means algorithm in order to solve the initialization problem of the K-means algorithm by determining weights for the clusters. These weights have been assigned to each cluster based on their variance. Weights and the cluster assignments were treated together, through an iterative process. Duan et al. [1] have proposed a different multilayer data clustering framework based on feature selection and modified the K-means algorithm. They have attempted to reduce the dimension of the data set, by selecting an envoy feature subset as a result of this the clustering process. Partial distance strategy was used in the proposed algorithm. Guérin et al. [31] have 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 [32] has proposed a new algorithm to enhance the performance of K-means clustering by the use of the linear transformation and the random perturbation of the kernel matrix. Nagwani & Sharaff [33] have applied the K-means clustering algorithm and text mining technique for separating and identifying spam and non-spam SMS messages. Chen et al. [34] have proposed a novel algorithm to improve the performance of K-means clustering called ordered K-means clustering algorithm. In this paper they supposed K-means clustering as a multi-criteria problem and used PROMETHEE to calculate the closeness of points. Gan et al. [35] have 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 divided into five sections. The proposed methodology is outlined in Section 2. Section 3 presents the experimental results and discussions is presented in section 4. Section 5 concludes the paper and provides some insights into future trends.

2. The Proposed model

In the field of data mining, clustering has many applications. Clustering is an important process in engineering and other fields of scientific research [36]. It is a process in which the records are categorized into separate groups so that the records in each cluster are most similar to each other and have the greatest difference with other data groups or clusters [37, 38, 39]. 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 [40]. One of the purposes of clustering is the better recognition of the data set.

A major difficulty in the clustering process is the determination of an appropriate label for each cluster. In some cases, due to the structure 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 a step-by-step clustering of the dataset and examining the process of forming clusters, the activity of labeling would 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 in some cases that it is not possible to determine the label for the entire dataset, and then it is possible to select a part of the dataset which represents the entire dataset and determine the label of cluster based on the cluster formed in this part of dataset. Moreover, during this process, the important variables are identified in 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. Those by
analyzing these changes, the effective features of the clustering process and the dataset are determined and also the appropriate labels to 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, the 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 in each step. After selecting an initial subset of the original set, K-means clustering algorithms run on the selected data set. At this point the first clusters will be formed. Then the clustering process will be repeated in each step until the completion of the dataset fields and records. In 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. In each step, the clustering results are recorded and compared with those of the previous steps. Changes made in the structure of the clusters and records belonging to each cluster will be investigated. The first step is to select an initial data set from the entire original data set. First, the records for this primary subset should be determined. To do so, the clustering with a high number and by the user’s selection is implemented on the entire data set. 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 data set is formed. In the next step, the clustering with the desired number is applied to the initial data set. Then the records and fields are added to the initial data set step by step. At each step, the clustering with the K-means algorithm is implemented on the selected data set and the clustering results are recorded and compared with those of previous steps. In each stage, the process of forming clusters is monitored, and the required analysis will be performed based on the changes made 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 high number clusters are carried out with respect to the dimensions of the dataset, and a record is randomly selected from each cluster. The aim of this is to select the records which have the most differences. The way of choosing the features is based on the variance of the variables. The aim of this step is also selecting the features that have created the highest difference among the records. Determining the number of fields and records in 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 the large number of cluster will be run on the whole dataset with respect to the dimensions of the dataset. Then a record will be randomly selected from the each cluster, and consequently the records of the initial subset will be determined. The purpose of this step is to select the records which have the most differences.

- **Step 2:** Select the features of the initial subset.
  In 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 which have 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.

  In this step, the variance of all the data set 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 data set of this step.

The pseudo code of 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 proposed algorithm. In this section a data set 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 registered service provider to the bank. Analyzing bank databases for analyzing customer behavior is difficult since bank databases are multi-dimensional, comprised of monthly account records and daily transaction records [41]. Table 2 lists the fields in this data set. Since the variables have different values, the field values were normalized so that the effects of all the fields in the analysis were the same.

In each step, a subset of the original data set is selected, and the process of forming the clusters and the labels allocated to each cluster are examined. For this purpose, clustering is initially performed on the entire data set with 50 clusters, and from each cluster, one record is randomly selected as the record of the initial selected subset. To select the data set 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. In 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 indicates their performance in the area of electronic payment.

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, Figure 2 and Figure 3.

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

**Step 1:** In this step, the initial dataset is selected. To do so, the initial dataset was formed with 50 records and 10 fields. In this step, the variance of all variables was calculated in the selected records, and the 10 variables with the highest value of variance were selected. The fields “Number of withdrawals from ATM”, “Number of Account Balance 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 /Deposit 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 data set. Then, the K-means clustering algorithm with 4 clusters was implemented on this dataset. Results of this clustering are presented in Table 5, Figure 4 and Figure 5.

Description of clusters:

**Cluster 1:** The long history banks with the privileged performance in the e-payment. Records related to the 3 banks of Melli, Mellat, and Saderat that have had a long 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 the electronic payments, but records of Ansar, Iran Zamin, and Pasargad banks which have less experience than the other banks of the cluster and have had a poorer performance in the electronic payments, are also located in this cluster. There are 12 records, i.e. 16.90% of all the cluster records in this cluster that belong to 3 banks of Ansar, Iran Zamin and Pasargad. With the combination of the banks in this cluster, therefore, it is not possible to determine the 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 are in this cluster. The performance of these banks in the area of electronic payment was poorer than the other banks.

**Cluster 4:** This cluster only includes post bank records. No appropriate label is specified for this cluster.

In comparison with the clustering of the entire dataset, the clustering of the subset including 10 fields and 50 records made a difference in the structure of clusters 1 and 4. As shown in the table above, the post bank’s records that were located in the 4th cluster in the previous stage are put in cluster 3 in this step. Also, the records of the banks Melli, Mellat, and Saderat are divided into two clusters of 1 and 4. According to the results, choosing this subset is not suitable as the subset that can be used to determine the label of clusters.

**Step 2:** In this step, 60 records randomly, and 7 fields of “Average amount of withdrawals from Pin Pad”, “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 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
Given the results and the formation of the clusters, it is not possible to determine the appropriate label for clusters.

**Step 3:** In this step, 80 records randomly 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” having the highest variance were added to the initial selected subset. After the extracting of 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 shown in Table 6, Figure 6 and Figure 7.

**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:** During this step, 7 fields with the highest variance as well as 60 records were randomly added to the selected dataset of the previous step. The eight added fields are “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 and 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, Figure 10 and Figure 11.

According to the results of the algorithm implementation in each step, the subset of step 3 with 30 fields and 190 records is selected as the subset that can be used for determining the clusters’ labels. The records and the fields of this subset can be extracted from the main dataset and accordingly, the clustering results can be gained and the labels of the clusters can be specified.
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 the 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 determine an appropriate label for the clusters. The proposed algorithm was implemented on the initial dataset containing 10 fields and 50 selected records from the main data set. 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.

5. Conclusion
Clustering techniques are attracting increasing interest for accurate analysis of high-dimensional data. Different algorithms are presented for data clustering. The K-means algorithm is one of the most powerful and most widely used clustering algorithms that 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 understanding of the data set. A serious challenge of clustering technique is to identify the proper cluster label. In this study, an algorithm based on the K-means clustering algorithm is presented. This approach could help cluster labeling in clustering and selecting 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, with the help of the proposed algorithm, it would be possible to select and extract a subset of the main data set so that the proper labeling for the clusters becomes possible. In general, this algorithm will be beneficial as a reliable and effective procedure in 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/ijisip.2016.9.4.08

[2] Haeri, A., & 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., 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/fjadm.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] Gon_calves, N., Nikkil □a, J. and Vig_arlo, R. “Selfsupervised mri tissue segmentation by discriminative clustering”, International Journal of Neural Systems, 24(1), 1450004 (16 pages) (2014).

[7] Saxena, A., Prasad, M., Gupta, A., and 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. 5-th Symp. Mathematical Statistics and Probability, Berkeley, CA 1967; 1:281–297.

[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., and 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., Zhangz, 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., 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., and et al. “An efficient enhanced k-means clustering algorithm”, Journal of Zhejiang University-Science, 7(10), pp. 1626-1633 (2006).

[18] Ahmad, A. “k-means 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., 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., 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] Niknam, T., Fard, E.T., Pourjafarian, N., and 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 2012), (Aisp) 007– 011. https://doi.org/10.1109/AISP.2012.6313708
[39] Maimon, O.Z., and Rokach, L. “Data mining and knowledge discovery handbook”, New York: Springer (2005).

[40] Fazel, Z. M., & Zarinbal, M. “Image Segmentation: Type–2 Fuzzy Possibilistic C-Mean Clustering Approach”, Journal of Industrial Engineering & Production Research, 23(4) pp. 245-251(2012).

[41] Farajian, M. A., & 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).

Figure Captions

Figure 1. The pseudo code of proposed algorithm.

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

Figure 3. Results of clustering the total bank's data set.

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

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

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

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

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

Figure 9. Results of clustering the 30 fields and 190 records.
**Figure 10.** Distribution of banks in each cluster in 37 fields and 250 records clustering.

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

---

**Table captions**

**Table 1.** The logic of each step.

**Table 2.** Explain the variables in the bank's data set.

**Table 3.** Bank ranking in e-payment. ([www.shaparak.ir](http://www.shaparak.ir)).

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

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

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

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

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

**Table 9.** The results of the algorithm run by the gradual addition of the fields and records.

---

**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**
Table 1

| Step | Logic |
|------|-------|
| Step 1 | The purpose of this step is to select the records which have the most differences. |
| Step 2 | The aim of this step is to select the features that have created the highest 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 comparison with the built subsets. |
| Step 6 | The consecutive clustering provides the ability to observe the trend of changes. |

Table 2

| Variable Number | Variable Name | Variable Number | Variable Name | Variable Number | Variable Name |
|-----------------|---------------|-----------------|---------------|-----------------|---------------|
| 1               | Bank          | 18              | Amount of Bills Pay via ATM | 35              | Average amount of Transfer via Shetabi Card with Pin Pad |
| 2               | Month         | 19              | Number of Bills Pay via ATM | 36              | Amount of Buying via Mobile banking |
| 3               | Amount of withdrawals from ATM | 20              | Average amount of Bills Pay via ATM | 37              | Number of Account Balance via Mobile banking |
| 4               | Number of withdrawals from ATM | 21              | Amount of withdrawals from Pin Pad | 38              | Amount of Bills Pay via Mobile banking |
| 5               | Average amount of withdrawals from ATM | 22              | Number of withdrawals from Pin Pad | 39              | Number of Bills Pay via Mobile banking |
| 6               | Number of Account Balance via ATM | 23              | Average amount of withdrawals from Pin Pad | 40              | Average amount of Bills Pay via Mobile banking |
| 7               | Number of Account Check via ATM | 24              | Number of Account Balance via Pin Pad | 41              | Number of Account Balance via Phone banking |
| 8               | Number of intra-bank transfer via ATM | 25              | Number of Account Check via Pin Pad | 42              | Amount of Bills Pay via Phone banking |
| 9               | Amount of Transfer /Withdrawal via ATM | 26              | Number of intra-bank transfer via Pin Pad | 43              | Number of Bills Pay via Phone banking |
| 10              | Number of Transfer /Withdrawal via ATM | 27              | Amount of Transfer /Withdrawal via Pin Pad | 44              | Average amount of Bills Pay via Phone banking |
| 11              | Average amount of Transfer /Withdrawal via ATM | 28              | Number of Transfer /Withdrawal via Pin Pad | 45              | Amount of Buying via POS |
| Variable Number | Variable Name | Variable Number | Variable Name | Variable Number | Variable Name |
|-----------------|---------------|-----------------|---------------|-----------------|---------------|
| 12              | Amount of Transfer /Deposit via ATM | 29              | Average amount of Transfer /Withdrawal via Pin Pad | 46              | Amount of Buying via Internet banking |
| 13              | Number of Transfer /Deposit via ATM | 30              | Amount of Transfer /Deposit via Pin Pad | 47              | Number of Account Check via Internet banking |
| 14              | Average amount of Transfer /Deposit via ATM | 31              | Number of Transfer /Deposit via Pin Pad | 48              | Amount of Transfer /Deposit via Internet banking |
| 15              | Amount of Transfer via Shetabi Card with ATM | 32              | Average amount of Transfer /Deposit via Pin Pad | 49              | Number of Transfer /Deposit via Internet banking |
| 16              | Number of Transfer via Shetabi Card with ATM | 33              | Amount of Transfer via Shetabi Card with Pin Pad | 50              | Average amount of Transfer /Deposit via Internet banking |
| 17              | Average amount of Transfer via Shetabi Card with ATM |                | Number of Transfer via Shetabi Card with Pin Pad |                |                |

Table 3

| 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                            |
| Rank | Total amount of e-payment transactions | Total number of e-payment transactions |
|------|----------------------------------------|----------------------------------------|
| 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

| 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            |

Table 5

| Cluster | ATM       | Pin Pad    |
|---------|-----------|------------|
| 1       | 0.771     | 0.752      |
| 2       | 0.377     | 0.211      |
| 3       | 0.074     | 0.030      |
| 4       | 0.678     | 0.470      |

Table 6

| Cluster | ATM       | Pin Pad    |
|---------|-----------|------------|
| 1       | 0.697     | 0.635      |
| 2       | 0.044     | 0.028      |
| 3       | 0.089     | 0.109      |
| 4       | 0.352     | 0.253      |
### Table 7

| 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            |

### Table 8

| Cluster | ATM     | Pin Pad | Mobile Banking | Phone Banking | POS     | Internet Banking |
|---------|---------|---------|---------------|---------------|---------|------------------|
| 1       | 0.562   | 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            |

### Table 9

| Step     | Fields’ number | Number of samples | Labels                                                                 |
|----------|----------------|-------------------|------------------------------------------------------------------------|
| Step One | 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 Two | 4, 6, 7, 12, 15, 19, 26, 28, 31, 34, 9, 10, 13, 23, 27, 30, 33, 3 | 110               | Cluster 1: -                                                             |
|          |                |                   | Cluster 2: -                                                             |
|          |                |                   | Cluster 3: -                                                             |
|          |                |                   | Cluster 4: -                                                             |
| Step Three | 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 Four | 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 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: -                                                             |
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, Energy Sources, Part B: Economics, Planning, and Policy, Journal of Engineering Manufacture: Part B, Int. J. Production Research, Int. J. Advanced Manufacturing and Technology, Int. J. Productivity and Quality Management (IJPQM).

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 M.S. degree in Industrial Engineering from Isfahan University of Technology (IUT), Isfahan in 1998 and obtained Ph.D. in Industrial Engineering from Amirkabir University of Technology (AUT), Tehran in 2005. Presently he is 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 this, 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.