Journal of Theoretical and Applied Information Technology
30th September 2020. Vol.98. No 18
© 2005 – ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPROVING CUSTOMER CLUSTERING BY OPTIMAL SELECTION OF CLUSTER CENTROIDS IN K-MEANS AND K-MEDOIDS ALGORITHMS

SHAHLA MOUSAVI[1], FARSAD ZAMANI BOROUJENI[2], SAEED ARYANMEHR[3]

[1] Faculty of Computer Engineering, Mazandaran Branch, Nonprofit University, Sari, Iran

[2, 3] Faculty of Engineering, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran

E-mail: [1]shahla.mousavi@khuisf.ac.ir, [2]f.zamani@khuisf.ac.ir, [3]saeed.aryanmehr@khuisf.ac.ir

ABSTRACT

Clustering technique is one of the most important tools for knowledge discovery, during which the samples are divided into categories whose members are similar to each other. One of the most common and widely-used clustering solutions is partition-based clustering algorithms such as K-Means and K-Medoids which have attracted a lot of attention in the field of customer clustering. However, in these algorithms, the initial cluster centroids are usually randomly selected from the initial samples, making the final result of the clustering undesirable in most cases. In this research, a solution is proposed for the optimal selection of initial cluster centroids in K-Means algorithm. In the proposed method, the initial cluster centroids are selected based on a heuristic method to provide the input for the clustering algorithm. To evaluate the effectiveness of the proposed method, the K-Means, K-Medoids, and improved K-Means algorithms were tested on a real data set obtained from Central Insurance Company in Iran. According to standard evaluation criteria, the proposed method had a greater impact on improving clustering results than the other two methods.

Keywords: *Partition-Based Clustering, Customer Clustering, Selection of Cluster Centroid, K-Means, K-Medoids*

1. INTRODUCTION

Nowadays, data mining plays a key role in helping organizations and companies improve their business decisions. By analyzing customer data through data mining methods, companies can use the collected data to guide, optimize, and automate customer-centric transactions. One of the most important uses of data mining in business is customer relationship management or CRM, in which customer clustering can lead to the identification of groups of customers who have similar behavior in terms of purchasing habits or receiving behavioral services. Clustering is one of several data mining techniques that is used widely in customer information clustering and does not require background knowledge about them. Clustering can be considered as partitioning of data objects into groups of similar objects and such groups are called clusters. In fact, the term "cluster" refers to a subset of data objects that are very similar to each other and less similar to other data objects [1]. Different clustering algorithms can be divided into different categories, including partition-based, hierarchical, network-based, density-based, model-based, and constraint-based algorithms.

A partition-based method divides n data objects into k parts, referred to as clusters [2]. A couple of the most common partition-based algorithms are K-Means and K-Medoids methods that use Euclidean space as their similarity measure. Random selection of first centroid is one of the k-means disadvantages [3]. Thus, selecting initial points with a short distance from each other can lead to termination of the clustering algorithm at local optima. Also, selecting points with a large distance from other points, leads to unbalanced clusters with low density. If we could distribute the initial cluster centroids uniformly among other data points, the quality of clustering could be improved based on quality assessment criteria. Examining the results of applying this technique on K-Means and K-Medoids can be helpful in selecting the appropriate clustering method for different data. It seems that obtaining useful information from customer data using improved clustering methods is useful for improving customer relationship management. Since different algorithms may be appropriate for different data, this study compares the results of applying naïve K-Means, improved K-Means and K-Medoids on the data set of an insurance company. In these methods, the number of clusters must be pre-determined. However, as we do not have a background knowledge on distribution of data in a data set, it is hard to guarantee the optimal selection of number of clusters. Therefore, the number of clusters will be determined based on the parameters determining the quality of clustering. A complete description of performance of these algorithms is presented in the following sections.

2.1 K-Means Algorithm

The K-Means algorithm is one of the well-known unsupervised learning algorithms for clustering problems [4]. This algorithm is often used as a prerequisite step for other algorithms. It uses the partitioning-based clustering in which data sets are divided into a predetermined number of clusters. The main idea in this algorithm is to determine the optimal centroid for each of the clusters. The best choice for the cluster centroid in the K-Means algorithm is to place them as far apart as possible. Then, each record in data set is assigned to the nearest cluster centroid. The K-Means algorithm use a simples methods for selecting initial centroids Such that, a centroid is randomly assigned to each cluster and each data object is assigned to a cluster with the closest cluster centroid. After this initial allocation, the cluster centroid is recalculated and assignment of data objects to the new centroids is repeated until a convergence condition is satisfied. In fact, in each iteration of the algorithm, the cluster centroids are updated and the algorithm continues until no change is occur in the location of cluster centroids. As a result, k clusters are created that display a set of n data objects. The K-Means method uses the average of data points as centroid to represent the clusters, so it is sensitive to outliers. It means that a data object with very high value can disrupt the distribution of data between points and is denoted by EQ (1).

$$E = \sum_{i=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(i)} - c_i \right\|^2 \qquad (1)$$

In a simple type of this method [5], some points are randomly selected first according to the number of clusters required. Then, the data are assigned to one of these clusters according to the degree of closeness (similarity), and accordingly, new clusters are created. By repeating this process, new centroid can be calculated for them in each repetition by averaging the data and the data can be re-assigned to new clusters. This process continues until the data remain unchanged. The function is considered as the objective function:

Where, $\left\| x_i^{(j)} - c_j \right\|^2$ is the criterion of distance between the points and the centroid of J^{th} cluster? The following algorithm is the basic algorithm for this method:

1-First, K points are randomly selected as the centroid of the clusters.
2-Each data sample is assigned to a cluster whose centroid has the smallest distance to that data.
3-After assigning all the data to one of the clusters, a new point is calculated as the centroid for each cluster (mean points belonging to each cluster)
4-Steps 2 and 3 are repeated until the centroids of the clusters remain unchanged [5].

Figure (1) shows the implementation of the K-Means algorithm for partitioning of 300 data objects into 3 clusters. The centroids of the clusters (black dots) remained unchanged after 9 repetitions of the algorithm.

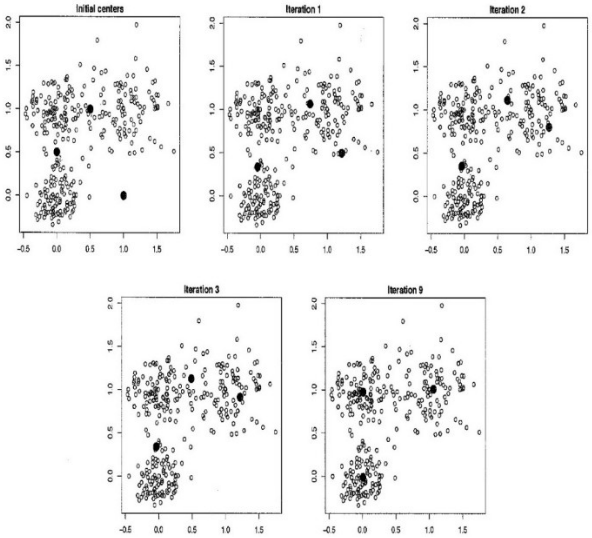

Figure 1: Execution of K-Means algorithm, n = 300 and k = 3

Although the termination of the above algorithm is guaranteed, it does not yield a single solution and it does not always have an optimal solution. After presenting the K-Means algorithm, the researchers identified some of the problems with some types of data sets. This algorithm is very sensitive to outliers, so the presence of an object with large value causes a significant deviation in data distribution. The K-Mediods algorithm [6] reduces this sensitivity by making changes to K-Means. Instead of minimizing the Euclidean squared sum of distance between the points (usually placed as a target function in K-Means), this algorithm minimizes the sum of the differences between the pair of points.

2.2 K-Medoids Algorithm

One of the methods presented in the partition clustering is the K-Medoids algorithm, which has been proposed to improve the K-Means algorithm and solve some of its problems, including sensitivity to outliers. It should be noted that in some data sets that have several nominal (non-numerical) features, it is not possible to select the centroid of the cluster using the mean calculation of the points within the cluster. Therefore, the K-Medoids method solves this problem by using data median instead of the mean as the centroid of the cluster. The median is the most central data object among the points of a cluster. Therefore, K objects are randomly selected as median to represent the clusters, and all data objects are assigned to clusters with the closest median. After processing all the points, a new median is determined for each cluster, which can be a better representative of a cluster, and thus, the whole process will be repeated. In each repetition, the medians change and the algorithm continues until other medians do not change. This algorithm also starts with an initial guess for the central points of the cluster randomly (k points are randomly selected from the points $X_1,...,X_n$). Then, the following steps are repeated:

1- For each $i = 1,...,n$, we identify the closest cluster centroid c_k to Xi and we will have: $C(i) = k$

2- For each $k = 1,...,k$, we set c_k equal to the median of the $k(X*k)$ cluster points, meaning that the point Xi in the cluster k minimizes $\sum_{c(j)=k} \|X_j - X_i\|_2^2$.

The repetition continues until the dispersion within the cluster does not change. In other words:

1- The algorithm clusters each point based on the nearest centroid to it.

2- Each centroid of the cluster is replaced by the median of the points within the cluster.

K-Medoids algorithm generally returns greater value of $\sum_{k=1}^{K} \sum_{C(i)=k} \|X_i - c_k\|_2^2$. This algorithm is more complex computationally compared to K-Means algorithm, since selecting of median is more complicated than selecting of mean. It should be noted that the K-Medoids algorithm has this important feature that the centroids of centroids are part of the cluster points [7]. In the K-Medoids method, one cluster is shown by one of its points. This method is a simple solution because it covers all types of data, and medians have an implicit resistance to outliers, so that out of centroid points do not affect them. When the medians are selected, the clusters are defined as a subset of points close to the centroid of the selected cluster, and the objective function is determined by the mean distance or another criterion of similarity between a point and the centroid of its cluster. Figure 2 illustrates the division of 300 data points into 3 clusters using the K-Means algorithm. The algorithm terminates after 6 repetitions. It should be noted that compared to clustering of this data with the K-Means algorithm, only three data points had different labels.

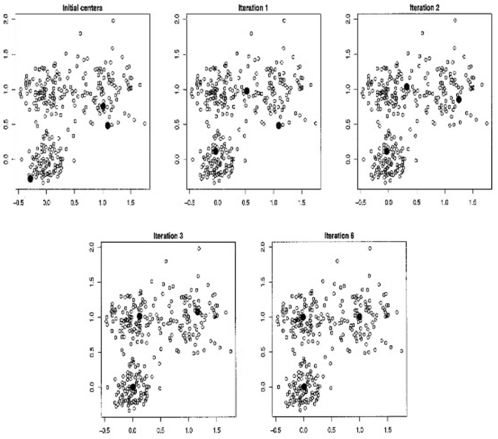

Figure 2: Execution of K-Medoids algorithm, $n = 300$ *and* $k = 3$

2.3 Review Of Literature

Shalini and Chauhan [8] compared two algorithms of K-Means and K-Medoids and examined their strengths and weaknesses. Examining features of the two methods, they concluded that partition-based clustering methods are more suitable for spherical clusters in small and moderate data sets.

Wagstaff and Cradie [9] presented a method in which they can use background knowledge to improve the K-Means algorithm. They argue that the current clustering methods have no way to apply background knowledge, even when such information is available. They showed how the K-Means clustering algorithm can use this information and how it can increase accuracy in clustering. Pham et al [10] also presented a method to determine the number of clusters for different data sets. As one of the shortcomings of the K-Means method is the number of clusters required before execution of algorithm, this method can help overcome this limitation.
Velmurugan and Santhanam argue that the most important advantage of clustering is that important patterns and structures can be obtained directly from a very large set of data without a need for background knowledge. In their research, they analyzed two methods of K-Means and K-Medoids based on their basic structure, and selected the best method based on efficiency in each category. It observed that mean time for normal distribution is longer than the meantime for uniform distribution. The results of several executions of the program showed that the K-Means algorithm was more efficient for small data sets and the K-Medoids algorithm was more efficient for moderate data sets [11]. Zhang and Cheng [12] presented a heuristic method for optimal selection of initial centroids instead of random selection. For this purpose, they used data centroid density to get better results in selecting initial centroids. Since the final result of clustering depends on the selection of the initial clusters, this algorithm can be used to improve the K-Means method and the results can be compared with the basic K-Means algorithm and the K-Medoids algorithm.

2. PROPOSED MODEL

The main idea of the improved K-Means algorithm is to calculate the $SimMatrix[n][n]$ similarity matrix for the vector X with m dimensions. Neighborhood matrices of similarity and density of similarity are also calculated according to the $SimMatrix$. First, candidates of the initial centroids are placed in an array called X'. In the X' array, the point x_i with the highest density of similarity in the defined similarity neighborhood is selected as one k cluster centroids and is recorded in $initC$. By selecting any point, x_i removes all points that meet the limitations of $SimNeighbor(x_i, a)$ from the X' array. In a similar method, k number of initial cluster centroids is determined and recorded in $initC$. Then, the clustering algorithm of the X data set is executed on K centroid of the initial cluster located in $initC$. As explained, our proposed improved algorithm hysterically selects points with the highest similarity density as the initial cluster centroids. The heuristic selection algorithm of k cluster centroids is as follows:
Algorithm input: array X containing n data elements with m dimension, X' array to represent candidate points, similarity neighborhood threshold a, similarity weighting factors l and set of initial clustering centroids $initC = f$.

Algorithm output: K number of clusters that form the C set, provided that: $i^{1}\ j, 0£\ i, j£\ K, C_i ÇC_j = f$.

It can be executed in the following five steps:
Step 1: using the equals (1): the similarity degree of each pair is calculated from the data of the vector set X and stored in the similarity matrix of $SimMatrix[n][n]$.

Step 2: in remaining x_i in the set X', the vector x_i is selected with the highest value of density (x_i).

Step 3: the x_i vector and all vectors in $SimNeighbor(x_i.a)$ are removed from the X' array. The X' array is displayed as:
$\{SimNeighbor(x_i - a)\}- \{x_i\}- X' ® X'$.

Step 4) the selected x_i point is added to the $initC$ set in the second step. Accordingly, $initC$ is displayed as $initC È x_i ® initC$. If the number of $initC$ set members is less than K, the algorithm execution will continue by jumping to the second step, otherwise, the clustering algorithm will be executed with the prepared initial centroids and at the end the clustering evaluation coefficient, obj will be calculated. The data set used in this study is related to information of Tehran Central Insurance customers in 2011 and 2012. This data set retains third-party customer insurance information, personal information and vehicle information, and information about accidents and damages. Clustering of customer is one of the needs of stores and business firms such as Insurance and other centroids dealing with customers. Table 1 presents the features related to the data set and the information related to their type and numerical interval.

Table 1: Attributes of the tested dataset

Feature name	Feature type	Interval
Gender	Nominal	Male-female
Accident damage	Quantitative	Natural
Vehicle age	Quantitative	Natural
Insurance	Quantitative	Natural

The first evaluation criterion used in this study to compare and assess the quality of clustering by the three methods is the obj criterion, which was described in the previous sections. This criterion is obtained by EQ (2). The higher the value of this criterion, the better the clustering quality. Therefore, clustering that has a higher value than the obj criterion has a more desirable result.

Another evaluation criterion is the sum of the sum of means of squared distance to the determined cluster centroid. To obtain this value, after completing the clustering and finding k centroid of clusters, we first calculate the Euclidean distance of each point from the centroid of the cluster to which it belongs. We do this for all points and in all clusters. For each cluster, we calculate the mean of squared distance obtained from the centroid of the cluster, and finally, we calculate the sum of the mean values obtained from all the clusters. This criterion is obtained by EQ (2).

$$MSD = \sum_{K} \frac{\sum_{i=1}^{m^k} \left\| d_i - d^k \right\|^2}{m^k} \qquad (2)$$

Where, k represents the number of clusters, m^k represents the number of k cluster points, and d^k represents the centroid of the k cluster. By calculating this criterion for different clusters, as the obtained value is lower, the clustering will be better.

3. RESULTS

The determined algorithms were implemented by Matlab software and executed on the data set of Tehran Central Insurance customers. Accordingly, customer data were clustered based on quality evaluation criteria using improved K-Means, K-Medoids and K-Means algorithms. To evaluate the proposed method for improving the K-Means algorithm, after implementation, we should compare this method with the basic K-Means algorithm and the K-Medoids algorithm. Randomly selection of initial points in these algorithms will increase the execution time and reduce the quality of clustering. Both techniques work without a supervisor, so having basic information of customers is enough and we do not need to have previous information of customers. With the heuristic selection of initial points of the cluster centroids compared to their random selection, a significant improvement will be achieved in the quality of clustering.

3.1 Comparison of methods based on obj evaluation criterion

In Figure (3) compares the obj evaluation criteria for the determined algorithms. This comparison was made for a number of different clusters. This evaluation criterion measures the quality of clustering based on the similarity of the points belonging to a cluster (density within each cluster) and the difference between the points belonging to different clusters. There is not much difference in the quality of clustering results in the number of less clusters, but with increasing the number of clusters, the improved K-Means algorithm shows better results than other methods. Such a result is expected because when the number of clusters is very small, many points are placed in the same cluster without considering specific and accurate criteria and without the necessary precision. With increasing the number of clusters to a reasonable level, the way of selecting the initial cluster centroids becomes more important, and the heuristic selection of initial clusters compared to randomly selection can significantly improve clustering results. Therefore, as the number of clusters increases, the qualitative difference between the clustering results in the improved K-Means algorithm increases with those of other methods. Also, in this data set, the K-Medoids algorithm and the K-Means basic algorithm follow a relatively similar performance in clustering and are not significantly different in terms of performance. Therefore, due to the optimization of the selection of initial cluster centroids using heuristic methods, as expected, better results were obtained from clustering and the density within the clusters increased.

Also, if we evaluate the Diagram exclusively for each method, we conclude that with increasing the number of clusters at the real and logical level, the obj evaluation criterion also increases at a uniform rate. It means improvement of the quality of clustering in more clusters. This conclusion can be justified by the fact that if the appropriate similarity criterion is selected, the density within each cluster increases with increasing number of clusters, and the difference between the points of different clusters increases accordingly, so the points within a cluster are more similar to each other. For example, if the Euclidean distance criterion is used to measure the similarity of points within a cluster, with increasing the number of clusters the points within each cluster will be less mean distance from each other and the density within each cluster will increase.

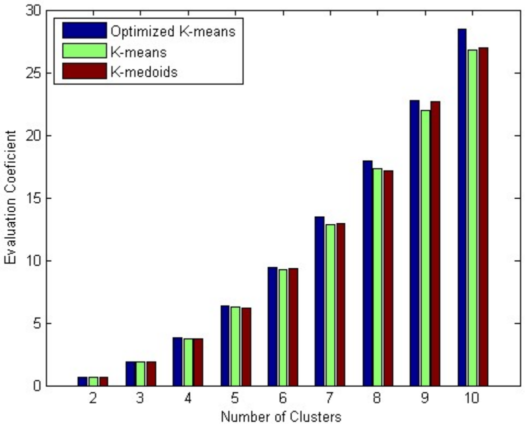

Figure 3: comparison of clustering methods based on obj criteria based on the number of different clusters

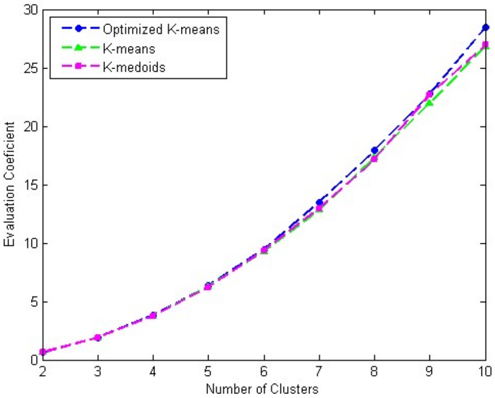

Figure 4: Linear diagram of comparing methods based on obj criteria

3.1 Comparison of methods based on the sum of means of squared distances to the cluster centroid

As described in the previous chapter, when the total sum of means of squared Euclidean distance of the points from the cluster centroids is lower, the clustering result will be better, because it indicates the high density of points within the obtained clusters. Figures 3 and 4 compare the values obtained from the calculation of this criterion after executing three clustering methods on the tested data set. Unlike previous standards, these values have greater differences in lower number of clusters, so that in lower number of clusters, mean of squared distance (MSD) in K-Means method is very different from that of other two methods. Therefore, this method is not suitable for small clusters at all. As the number of clusters increases, the difference in the value MSD in the methods decreases. However, in each time execution of the methods with any number of improved K-Means algorithm clusters, the lowest value of MSD and the best clustering result are obtained. It should be noted that when the distance between the points within each cluster with the cluster centroid is lower (indicating the similarity or closeness of the points within each cluster relative to each other), density within each cluster will be higher. Now, if the initial cluster centroids are selected using a heuristic method and not randomly, the decrease in the total mean distance between the points and each cluster centroid will indicate the appropriate accuracy and quality of clustering. Therefore, increasing the quality of clustering in the improved K-Means method is an expected result. However, in general, as the number of clusters increases, the value of MSD decreases exclusively in each of the three methods, and better clustering is achieved. This result can be justified by reducing the distance between points in each cluster and increasing the number of clusters as a result. With increasing the number of clusters, the K-Medoids algorithm does show good performance compared to the basic K-Means and the improved K-Means. Therefore, it seems that the K-Medoids method is not efficient compared to K-Means methods for clustering the desired data set and similar data.

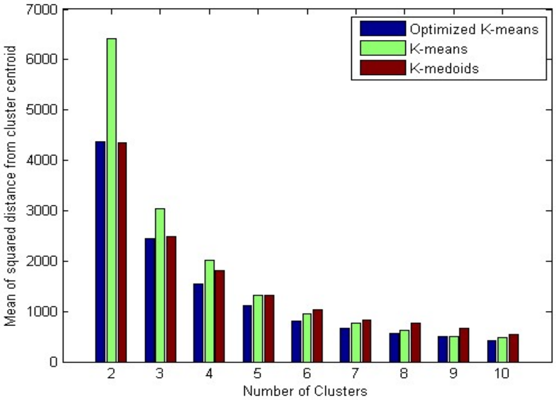

Figure 5: Comparison of methods based on the sum of means of squared distances for the number of different clusters

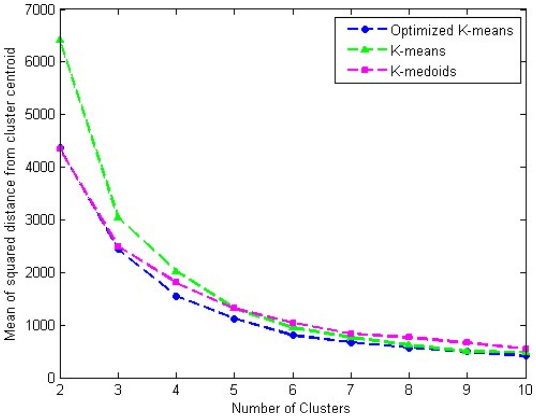

Figure 6: linear diagram for comparing the three methods based on the sum of means of squared distance from the points to cluster centroids

4. DISCUSSION AND CONCLUSION

The diagrams obtained from the implementation of the proposed method show that selecting the right number of clusters can have a great impact on improving the quality of clustering. In fact, if the number of clusters is too large, the concept of clustering disappears, because each data can form a cluster in such cases. However, if the number of clusters is very low compared to the data, the density within the clusters will also decrease, and some clusters may merge and many data may be placed in the same cluster without appropriate similarity criteria. However, the right and optimal selection of initial cluster centroids will have a significant impact on the final quality of clustering. As expected, in the proposed algorithm of improved K-Means, due to the use of the heuristic method for selecting initial cluster centroids, a significant improvement was observed in cluster quality and accuracy based on quality evaluation criteria compared to K-Means and K-Medoids methods. At K-Medoids, each cluster centroid is a member of the data set. The K-Medoids method is more resistant to outliers compared to K-Means, so it usually yields better quality results than K-Means method, but it is more complex in terms of computation. In the K-Means method, the cluster centroid can be placed anywhere in the problem space. One of the features of this method is that the selection of initial cluster centroids is done randomly at first, so the heuristically selection of initial cluster centroids can significantly improve the quality and accuracy of clustering compared to the methods in which the initial cluster centroids are selected randomly. Using this method, the density within each cluster increases, which means that the data inside a cluster are closer to each other in terms of distance and are more similar in terms of criteria and data that belong to different clusters have the lowest similarity and the highest distance.

REFRENCES:

[1] F. Shuweihdi, "Clustering and classification with shape examples," University of Leeds, 2009.

[2] Y. Zhang and E. Cheng, "An optimized method for selection of the initial centers of k-means clustering," in *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, 2013, pp. 149-156: Springer.

[3] S. Poomagal, P. Saranya, and S. Karthik, "A novel method for selecting initial centroids in K-means clustering algorithm," *International Journal of Intelligent Systems Technologies and Applications,* vol. 15, no. 3, pp. 230-239, 2016.

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, vol. 1, no. 14, pp. 281-297: Oakland, CA, USA.

[5] C. Elkan, "Using the triangle inequality to accelerate k-means," in *Proceedings of the 20th international conference on Machine Learning (ICML-03)*, 2003, pp. 147-153.

[6] L. Kaufman, P. Rousseeuw, and Y. Dodge, "Clustering by Means of Medoids in Statistical Data Analysis Based on the," ed: L1 Norm,~ orth-Holland, Amsterdam, 1987.

[7] R. Tibshirani, "Clustering 1: K-means, k-medoids," *R. Tibshirani, Data Mining,* pp. 36-463, 2013.

[8] S. S. Singh and N. Chauhan, "K-means v/s K-medoids: A Comparative Study," in *National Conference on Recent Trends in Engineering & Technology*, 2011, vol. 13.

[9] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *Icml*, 2001, vol. 1, pp. 577-584.

[10] D. T. Pham, S. S. Dimov, and C. D. Nguyen, "Selection of K in K-means clustering," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science,* vol. 219, no. 1, pp. 103-119, 2005.

[11] T. Velmurugan and T. Santhanam, "Performance analysis of k-means and k-medoids clustering algorithms for a randomly

generated data set," in *Proceedings of the International Conference on Systemics, Cybernetics and Informatics*, 2008, pp. 578-583.

[12] C. Zhang and S. Xia, "K-means clustering algorithm with improved initial center," in *2009 Second International Workshop on Knowledge Discovery and Data Mining*, 2009, pp. 790-792: IEEE.