A Novel Method for Reduction of Error Rates in K-Means Clustering Algorithm

Yousef Farhang, Siti Mariyam Hj. Shamsuddin

Soft Computing Research Group, Faculty of Computing, Universiti Teknologi Malaysia, Skudai, 81310, Johor, Malaysia
E-mail addresses: fyousef2@live.utm.my (Y. Farhang), mariyam@utm.my (S.M. Hj. Shamsuddin).

Abstract: This paper investigated K-means Algorithm, a well known clustering algorithm. K-means clustering algorithms have some shortfalls and defects, and one defect is reviewed in this study. One of the disadvantages of k-means clustering algorithms is that they can produce clusters that do not always include all the correct components. It is due to the presence of the error rate during the clustering process. The purpose of this research was to decrease error rates in the k-means clustering algorithm and to reduce iteration of running this algorithm. A novel method is proposed to calculate the distance between cluster members and cluster center. To evaluate the algorithm proposed in this study, seven well-known data sets consisting of Balance, Blood, Breast, Glass, Iris, Pima and Wine data sets were used. This investigation revealed that the performance of K-means algorithms was increased and resulted in valid clusters and that it reduced error rates, run time and iteration.

[Yousef Farhang, Siti Mariyam Hj. Shamsuddin. **A Novel Method for Reduction of Error Rates in K-Means Clustering Algorithm.** *Life Sci J* 2014;11(8):135-154]. (ISSN:1097-8135). http://www.lifesciencesite.com. 19

Keywords: K-means; clustering algorithm; error rate; iteration; reduction; stable

1. Introduction

Clustering is an important technique used in many fields such as knowledge discovery and information retrieval. It helps researchers find related information more quickly (1). As a result, researchers are kept up to date with new findings in their fields. Clustering is the process of grouping or dividing a set of objects into subsets (called clusters) so that the objects that are similar to one another are placed within the same cluster and dissimilar objects are placed in other clusters (2). In other words, an object is similar to at least one other object in the same cluster and dissimilar to objects in other clusters in terms of predefined distance or similarity measure (3). Currently, clustering as a tool for classification, pattern analysis, information extraction and decision making, has attracted the tendency of numerous investigators. Numerous techniques and approaches have been introduced in the literature. Each of these methods includes a certain measure, and has its own disadvantages and advantages. In general, there is no comprehensive technique and measure for optimal clustering of any kind of data (4).

In this study, a new understanding of the clustering algorithm was expressed. The most prominent, the most commonly used and the most popular clustering algorithm is the K-means algorithm, and it is used in this study. Among clustering algorithms, the K-means clustering algorithm can be used in many fields, including image and audio data compression, pre-process system modeling with radial basis function networks and task decomposition of heterogeneous neural network structure. One problem of clustering algorithms is that the clustering results are not always stable. In repeating the clustering algorithm several times, correct answers may be found in some trials but in others it may not find the correct answers due to instability. The clustering algorithm should be constant and stable which is reviewed in this survey. This problem and gap as mentioned in the fourth part are related to the summary of a section of a Jain article (5-6).

Cormack (1971) first proposed that clusters should be internally integrative and externally segregated, suggesting a certain degree of uniformity within clusters and heterogeneity between clusters (7). So, many investigators tried to operationalize this description by minimizing within-group disparity (8-11).

Following these efforts at maximizing within group uniformity, Sebestyen (1962) and MacQueen (1967) separately developed the *K*-means technique as a strategy that tries to discover optimal partitions (12-13). Based on this significant advancement, *K*-means has become very popular, earning a place in a variety of textbooks on multivariate techniques (14-16), cluster analysis (17), pattern recognition (18), and statistical learning (19-20). There are many surveys in *K*-means clustering algorithm field, yet this algorithm has still not been completely improved. In this paper, we reduced the error rate of the clustering algorithm and increased the stability of this algorithm.

K-means clustering algorithm has a number of disadvantages and problems, and one problem was reviewed in this study. This paper is organized as follows. Section 2 and 3 review the literature about clustering algorithms and K-means clustering algorithm. Section 3 describes the proposed method and research methodology used in this study. Section 5 explains the experiment conducted as a part of this study in the K-means clustering algorithm and improved K-means clustering algorithm, and the results are evaluated in Section 6. Finally, conclusions are drawn and discussed in Section 7.

2. Related Works

In this section, the brief literature of the clustering algorithms is examined in which different researchers have previously expressed and improved these algorithms. Forgy's technique (21) randomly allocates each point to one of the K clusters homogeneously. The centers are then given with the centroids of these primary clusters. This technique has not basis of theoretical as, for example, random clusters have not homogeneity of internal (22). Jancey's technique (23) allocates to each center a combinatorial point randomly generated within the space of data. However, as the data set fills the space, a number of these centers may be too distant from any of the points (22), which might lead to the formation of unfilled clusters (24).

MacQueen (1967) suggested two different techniques (12). The first technique is the default choice in the Quick Cluster method of IBM SPSS Statistics (25), which obtains the first K points in X as the centers. An obvious disadvantage of this technique is its sensitivity into data ordering. The second technique selects the centers randomly from the data points. The foundation behind this technique is that random choice is likely to result in the selection of points from dense regions; points are suitable applicants to be centers.

Ball and Hall's technique (26) obtains the centre of X, as the first center. It then crosses the points in optional order and obtains a point as a center if it is at least T units apart from the formerly selected centers until K centers are taken. The aim of the distance threshold T is to make sure that the seed points are well parted. The Simple Cluster Seeking technique (27) is the same as Ball and Hall's technique with the distinction that the first point in X is obtained as the first center. This technique is applied in the FASTCLUS method of SAS (24, 28).

Maximin technique (29) selects the first center $\mathbf{c}_1$ randomly and the i-th ($i \in \{2,3, . . . ,K\}$) center $\mathbf{c}_i$ is selected to be the point that has the most minimum distance to the formerly chosen centers, that is $\mathbf{c}_1,\mathbf{c}_2, . . . ,\mathbf{c}_{i-1}$. This technique was originally expanded as an approximation to the K-center clustering problem. It should be referred that, motivated with a vector quantization request, Katsavounidis et al.'s variant (29) obtains the point with the greatest Euclidean standard as the first center.

Al-Daoud's density technique (30) first regularly partitions the data space into M decomposed hyper-cubes. It then randomly chooses K N_m/N points as of hypercube m ($m \in \{1,2,\ldots,M\}$) to take a total of K centers where N_m is the points number in hypercube m. Bradley and Fayyad's technique (31) begins by randomly partitioning the data set into J subsets. These subsets are clustered by k-means initialized through MacQueen's second technique producing J sets of intermediate centers, each with K points. These center sets are united into a superset that is then clustered through k-means J times, each time initialized by a diverse center set. Members of the center set that give the least SSE are then taken as the final centers. Pizzuti (32) advanced Al-Daoud's density-based technique using a solution grid method. This technique begins through 2D hypercube and iteratively divides these as the number of points they accept increases.

The k-means++ technique (33) interpolates between maximin technique and MacQueen's second technique. It selects the first center randomly and the *i*-th ($i \in \{2, 3, . . . ,K \}$) center is selected to be $\mathbf{x} \in X$, where md(x) denotes the distance of minimum from a point $\mathbf{x}$ to the previously chosen centers.

The PCA-Part technique (34) applies a divisive hierarchical system based on PCA (Principal Component Analysis). In this method, starting from a first cluster that contains the all data set, the technique iteratively chooses the cluster with the greatest SSE and divides it into two sub-clusters by a hyper-plane that it passes with the center of cluster and is orthogonal to the way of the basic eigenvector of the covariance matrix. This method is repeated until K clusters are taken. The centers are then given through the centers of these clusters.

Lu et al.'s technique (35) applies a two phase pyramidal method. The attributes of each point are first encoded as integers. These points of integer are considered to be at stage 0 of the pyramid. In the phase of bottom-up, starting from stage 0, adjacent data points at stage k ($k \in \{0,1, . . . \}$) are averaged to take weighted points at stage k + 1 until at least 20 K points are taken. Onoda technique (36) first computes K Independent Components (ICs) (37) of *X* and then selects the *i*-th ($i \in \{1, 2, . . . ,K\}$) center as the point that has the least cosine distance (24).

3. K-means Clustering Algorithm

The aim of data clustering, also known as cluster analysis, is to discover the normal grouping of a set of points, objects or patterns. The Merriam-Webster dictionary defines cluster analysis as ''a statistical classification method for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics." The goal is to develop a clustering algorithm that will find the normal groupings in the data of unlabeled objects (5). Cluster analysis or clustering is a method of assigning a set of data objects into clusters where all the objects in a cluster are considered to be similar based on common features. Clustering is an unsupervised learning-based technique for statistical data analysis used in many fields including data mining, pattern recognition, image analysis, and bioinformatics (38-40). Selecting clusters of optimally is an NP-hard problem (41). Clustering algorithms include many algorithms, and *K*-means algorithm is the most popular. *K*-means algorithm is a rather simple but well-known algorithm for grouping objects (6). This algorithm is so well known and has widely applied that researchers consider it the equivalent of clustering algorithms.

The term "K-Means" was first used by James MacQueen in 1967, though the idea originates with Hugo Steinhaus in 1956 (12, 42). A standard algorithm was first proposed by Stuart Lloyd in 1982 as a technique for pulse-code modulation, though it was not published until 1982 (43). The classical K-means clustering algorithm aims to detect a set C of K clusters C_j with cluster mean $\mathbf{c}_j$ to reduce the sum of squared errors. Number of clustering C is a very important parameter (44). This is typically described as follows:

$$E = \sum_{j=1}^{K} \sum_{x_i \in c_j} \|cj - xi\|^2 \qquad (1)$$

E is sum of the square error (SSE) of objects with cluster means for K cluster. Also, ||…|| is a distance metric between a data point x_i and a cluster mean c_j. For instance, the Euclidean distance is defined as:

$$\|x - y\| = \sqrt{\sum_{i=1}^{V} |x_i - y_i|^2} \qquad (2)$$

The mean of cluster C_i is defined by the following vector:

$$c_j = \frac{1}{|C_j|} \sum_{i \in c_j} x_i \qquad (3)$$

The K-Means algorithm is as follows (45):

Algorithm1. The K-means clustering algorithm

```
i.     Assign initial values for cluster means $c_1$ to $c_k$
ii.    Repeat
iii.       for i=1 to n do
iv.            Assign each data point $x_i$ to cluster $C_j$ where $||c_j - x_i||$ is the minimum
v.         end for
vi.        for j=1 to K do
vii.           Recalculate cluster mean $c_j$ of cluster $C_j$
viii.      end for
ix.    until convergence
x.     return C
```

The K-means algorithm is a greedy algorithm, which can only converge to a local minimum, even though recent study has exposed the enormous possibility that K-means could converge to the overall optimum when clusters are well detached (46-47). K-means begins with a primary partition with K clusters and allocates patterns to clusters so as to decrease the squared error. The major stages of standard K-means algorithm are as follows (48):

a. Choose an initial partition with K clusters; repeat stages b and c until membership of cluster stabilizes.

b. Make a new partition through assigning each pattern to its closest cluster center.

c. Calculate new cluster centers.

The Figure 1 expresses an illustration of the standard K-means algorithm on a dataset of two-dimensional with three clusters. Figure 1, sets out a design for a K-means clustering algorithm. (a) Two-dimensional input data with three clusters; (b) three seed points chosen as cluster centers and initial assignment of the data points to clusters; (c) & (d) intermediate iterations updating cluster label and the centers; (e) final clustering obtained by K-means clustering algorithm at convergence [24].

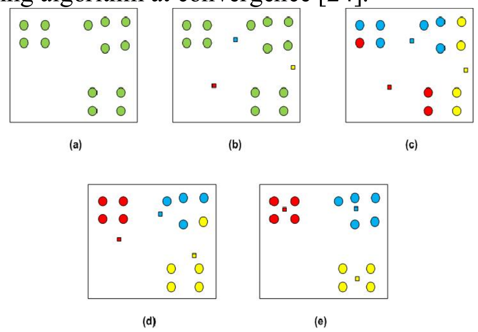

Figure1. K-means clustering algorithm for 3 clusters

Clustering algorithms have many applications, but there are problems with this algorithm. One problem is that the clustering algorithm and K-mean algorithm are not always constant. The clustering algorithm may generate correct answers several times in some trials but in some other trials it may not find the correct answers due to instability. In order for the clustering algorithm to stabilize, it must reduce the number of errors and the number of iteration steps in the algorithm. Therefore, the proposed method tries to reduce the error rate and iteration in the K-means algorithm.

4. Proposed Method

In the method used in this study, numbers were taken from outside of the cluster center. Data sets were added and the k-means clustering algorithm was implemented. In this study is calculated the distance in the clustering algorithm to determine the best method. When the data distance was calculated correctly, cluster errors in the algorithm were reduced. The principal goal of the research methodology used in this study was error reduction in the k-means algorithm. In this section, the initialization of the proposed method is first checked. Then, the improved K-means clustering algorithm is expressed. Last, the problem formulation and proportion is expressed.

4.1 Initialization

In this study, initial value is randomly selected, after the data set was applied as cluster centers are selected randomly in the initial stage. All members of the dataset attributes must be an integer. A set of datasets was generated using MATLAB for testing the effect of various parameters and size of the problem on the time taken through the algorithm. By selecting the required number of cluster centers randomly in the domain [1, number of rows], which chosen randomly is a normal distribution. First, the dataset is applied to MATLAB. If the dataset format is more usable in MATLAB, it must be converted to the format used. Text format for the datasets have been used in this study. Second, the number of rows in the dataset is determined and then the number of clusters is selected as random numbers from 1 to the number of rows. For example, if the number of clusters is three and number of rows in the dataset is 150, three random numbers from 1 to 150 will be selected. So, selected attributes of these rows are initial cluster centers.

4.2 Proposed Algorithm- Reduction of Error Rates in K-Means Algorithm (RER-K-Means)

In the proposed algorithm (Reduction of Error Rates in K-Means algorithm or RER-K-Means algorithm), Equation 5 is used to calculate the distance between the members of dataset and cluster centers. Another difference between RER-K-means algorithm and K-means algorithm is that comparing and finding the minimum distance used a better method, which it is described more in the next section. Actually, Equation 5 is a compatibility function that would calculate and minimize the intra cluster distance. The equation has K clusters of N data vectors classified according to the distance from each cluster center; it is located at one of the clusters. In this equation, the total aggregate of Euclidean distance of all the data vectors from cluster centers that they own is calculated and added to each other.

Therefore, by determining the optimal, centers can easily be clustered and the answer is that one that is best clustered. Using the equation, the number of clustering errors is reduced and it is close to being stable. It follows that the main objective of this equation is to ensure that minimum distances between the centers of the clusters are optimized, till, K-means clustering algorithm is to be improved. This study calculated the distance between the center of the cluster and the cluster members using one of the best ways to calculate distance, which is the Euclidean function in MATLAB. Also, calculations of the distance between centers of the cluster and the cluster members are eliminated as additional unnecessary operations have a negative impact on the calculation. Accordingly, the proposed algorithm will be clustered; cluster members will be assigned to data sets, reducing the error rate and stabilizing clustering algorithms.

4.3 Problem Formulation and Proportion

The RER-*K*-means clustering algorithm is further described in this section. The programming code was written using MATLAB software. A coding program was used to reduce the complexity of the algorithm, and the best method for clustering data was calculated in the *K*-mean algorithm. In the following algorithm, the RER-*K*-means clustering algorithm that was implemented in MATLAB is shown. In algorithm 2, the RER-K-means clustering algorithm is described, which the Euclidean method was used to calculate the distance between clusters. The RER-K-means clustering algorithm (Reduction of Error Rates in K-means clustering algorithm) has eleven stages, which are described below.

Algorithm2. The RER-K-means clustering algorithm
Input: Dataset as Number, Number of Classes, Number of Attributes, Number of Instances Number of Iteration: 50 Times Number of Running: 20 Times Output: Clusters, No. True, No. Errors, Intra Cluster Distance, Iteration

Step1: At first, the target dataset is applied to the MATLAB software. The dataset must have the clustering conditions.

Step2: In this step, the number of rows of the dataset is found followed by selecting desired numbers of rows randomly as cluster centers. The selected attributes of the random rows are assumed to be initial cluster centers.

Step3: Specifying the number of iterations, it is considered 50 steps for all datasets in this study. All main processes were placed into this loop. This is known as the named outer loop.

Step4: A loop is created for the first to the last dataset in which all the main instructions can be placed. This loop is the inter loop.

Step5: At this stage, the distances of cluster centers which have been previously considered from all members of the dataset are calculated. To calculate the distance, the coordinates of the cluster center in one array and attributes of a row as dataset in another array are placed, and then the distance between these two arrays is calculated using the following formula. This operation is carried out for all cluster centers in one step.

Step6: In this step, the distances of all cluster centers from one of the datasets are calculated separately and the minimum distance is taken into consideration. Now, members of datasets are placed in the cluster with the minimum distance.

Step7: In this step, some variables are defined to represent summation of distances between cluster center and its members. The number of define variables should be equal to the number of clusters. For instance, if there are 3 clusters, three variables s1, s2 and s3 are defined in which s_i is summation of distances among ith cluster center to its member. (i=1, 2, 3)

Step8: This step is the end of inter loop. It means that steps 4 to 7 are run until the ending condition of inter loop.

Step9: Variable S which is intra cluster distance is defined as summation of s_1, s_2, s_3 and so on. From converging of S it is deducted that algorithm has stabilized. Generally, S should be tried to minimize as far as possible.

Step10: The means of any cluster should be determined separately. Then, at the end of any step, the determined means are considered as cluster centers for the next step.

Step11: This step is the end of outer loop. It means steps 3 to 8 are run until the ending condition of outer loop.

Assume that matrix A represents a dataset as follows:

$$A = \begin{bmatrix} A(1,1) & A(1,2) \dots A(1,n) \\ A(2,1) & A(2,2) \dots A(2,n) \\ A(3,1) & A(3,2) \dots A(3,n) \\ \dots \dots \dots \dots & \dots \dots \dots \dots \dots \\ A(m,1) & A(m,2) \dots A(m,n) \end{bmatrix} \qquad (4)$$

In which, each row shows one member of data sets. Supposing that jth row is cluster center, the following formula can be used to determine intra cluster distance:

$$D(j) = \sum_{i=1}^{m} \sqrt{\sum_{k=1}^{n} [A(i,k) - A(j,k)]^2} \qquad (5)$$

4.4 The formula for calculating error rates

The equation shown below was used to calculate the error rate. It was required to calculate two measures; the number of error patterns and the total number of patterns:

$$\text{Error rate } (\%) = \frac{\textit{Numbr of Error patterns}}{\textit{Total number of patterns}} * 100 \qquad (6)$$

Equation 6 was used to find the error rate in the improved K-means clustering algorithm and the K-means clustering algorithm in all data sets of this study. In next section, it will be seen that the RER-K-means algorithm reduced the error rate and iteration. In this algorithm, additional operations that have a negative effect on the calculation must be avoided. In all the data sets, the K-means clustering and the RER-K-means algorithms implementation were similar and only the data set name and data set coordinates were changed by the algorithms.

5. Experimental Results

The clustering results are compared with k-means and improved k-means algorithm. These are implemented with the number of clusters as equal to the number of classes. Meanwhile, the number of data sets selected to solve the problem in the next section can be fully expressed. To check the results, two important criterions are used to error rates and iteration of running.

5.1 Experimental Data

Experiments have been performed on seven data sets which consist of Balance, Blood, Breast, Glass, Iris, Pima and Wine that were selected from standard data set UCI. Each of them is described in the following:

Balance Scales (Balance): Balance Scale data set is composed of 625 instances, 4 attributes and 3 classes. Each example is classified as having the balance scale tip to the right, tip to the left, or balanced. Balance dataset contains 46.08% of class L, 7.84% of class B and 46.08% of class R.

Blood Transfusion Service Center (Blood): This data set adopted the donor database of Blood Transfusion Service Center in Hsin-Chu City in Taiwan. Blood data set has 748 samples which are 748 donors selected at random from the donor database. This data set has 5 attributes which include R (Recency - months since last donation), F (Frequency - total number of donations), M (Monetary - total blood donated in c.c.), T (Time - months since first donation), and a binary variable representing whether donor donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood). The dataset contained 76% no (0) and 24% yes (1).

Breast Cancer Wisconsin, Original (Breast): The Breast Cancer Wisconsin dataset has 699 instances of cytological analysis of fine needle aspiration of breast tumors. In this data set each instance contains 10 attributes that are computed from a digitized image of a fine needle aspiration of a breast mass. Attributes of this data set include radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The dataset contains 241 (34.48%) malignant instances and 458 (65.52%) benign instances (49).

Glass Identification (Glass): This data set has 214 samples and seven classes. Every sample in this data set has 9 attributes. Seven kinds of glass are in the data sets including building windows float, building windows non-float, vehicle windows float, vehicle windows non-float, containers, tableware and headlamps.

Iris: This data set is based on Iris flowers recognition with three different classes each consisting of 50 samples. Every sample has four attributes. It presents 150 instances containing width and length measures of the sepals and petals of three species of the flower Iris: 'Setosa', 'Versicolor' and 'Virginical'. With 4 attributes and 3 classes, each containing 50 objects, the aim is to cluster similar species based on their measurements (49-50).

Pima: This data set is allocated to recognize diabetic patients. A total of 768 samples are classified into two groups consisting of 500 and 268 samples, respectively. Every sample in this data set has 8 attributes (50).

Wine: The Wine dataset has 178 instances and 13 attributes, which correspond to the results of chemical analyses performed with three types of wines produced in the same region of Italy, but from different cultivations. Attributes include alcohol content, acidity, alkalinity, color intensity, among others. The dataset has 59 instances of the first class, 71 instances of the second class and 48 instances of the third (49).

The databases used were obtained from the UCI data warehouse (51). Additional information about number of instances, number of attributes and clusters contained in each dataset are offered in Table 1.

The relevant datasets are implemented in the clustering algorithm and proposed algorithm and compared in depth. In this study, two measures are used to compare the names of the error rates and number of iterations. In the next section, these two measures will be discussed above data sets.

Table1. The data sets used in the experiments.

No.	Name of Dataset	Number of Instances	Number of Attributes	Number of Classes	Size of Dataset
1	Balance Scales (Balance)	624	4	3	Medium
2	Blood Transfusion Service Center (Blood)	748	5	2	Large
3	Breast Cancer Wisconsin, Original (Breast)	699	10	2	Medium
4	Glass Identification (Glass)	214	9	7	Small
5	Iris	150	4	3	Small
6	Pima Indians Diabetes (Pima)	768	8	2	Large
7	Wine	178	13	3	Small

5.2 Results of Error Rate and Number of Iteration in the Proposed Algorithm

In this section, results concerning the number of errors of the proposed K-means clustering algorithm on the data sets are reviewed. In the previous section, it was noted that in this study, seven data sets have been selected to analyze the proposed K-means algorithm. These data sets are standard and are selected from UCI data sets. The proposed K-means algorithm is applied to the respective data sets to determine the results; the number of errors and the graphs and charts can then be fully expressed.

In this study, our main objective is to improve the K-means clustering algorithms. To verify the improved algorithm, the improved clustering algorithm will be tested on the some data sets to answer the question of whether this algorithm is improved or not. Thus, the seven data sets (Balance, Blood, Breast, Glass, Iris, Pima and Wine) are implemented separately in the MATLAB software of the proposed algorithm and the results are discussed in this study.

i. Balance

The proposed k-means algorithm was first applied to the Balance dataset. The specifications of this data set are described in the previous section, but will be mentioned briefly here. The Balance data set has 625 instances and 3 classes. In the MATLAB software, code programming proposed K-means algorithm is implemented and Balance data set is loaded. The proposed algorithm clustered Balance data set is shown in Figure 2. In this figure, a scattering diagram of improved K-means clustering algorithm on the Balance data set is shown. This diagram highlights the 625 members and the distance among members in this data set.

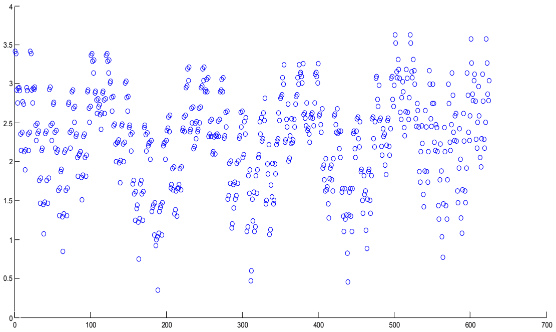

Figure 2. Scatter diagram of Balance data set clustered with improved K-means algorithm

In Figure 3, the clustering of three clusters in Balance data set with improved K-means clustering algorithm is displayed. This data set has three clusters that are not regular which means that the first, second and third clusters are merged, and that the Balance data set is such that the clusters are not completely separated.

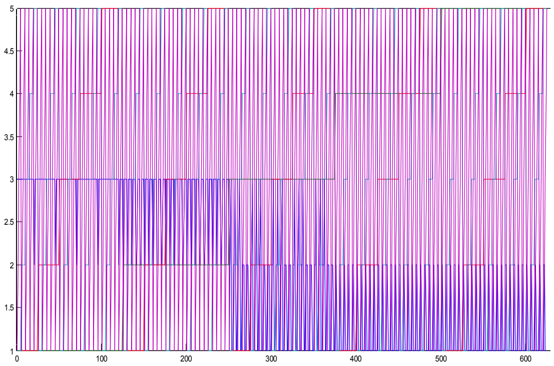

Figure 3. Display clustering the Balance data set with improved K-means algorithm

In Table 2, the complete results of the K-means clustering algorithm and the improved K-means clustering algorithm are fully described in the Balance data set. Each algorithm was run twenty times and each time, the algorithm was run 50 times to achieve stability. In general, the algorithm is executed 1000 times for each data set. In this table, five factors (number of true, number of errors, intra cluster distance, iteration and error rate) are compared with the improved K-means clustering algorithm and K-means clustering algorithm on the Balance data set. In all factors, the proposed algorithm is much better than previous algorithms.

Table 2**.** The Number of true, Number of errors, Intra cluster distance, iterations and error rate for the 20 runs of K-means and the improved K-means on Balance data set

	Improved K-means algorithm					**K-means algorithm**				
Running	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**
1	341	284	1425.84	7	45.44	129	496	1429.72	13	79.36
2	326	299	1423.92	12	47.84	290	335	1438.73	10	53.6
3	294	331	1423.85	9	52.96	370	255	1431.43	28	40.8
4	325	300	1425.84	11	48	246	379	1426.48	8	60.64
5	443	182	1426.63	17	29.12	304	321	1436.08	13	51.36
6	348	277	1425.70	6	44.32	335	290	1433.10	22	46.4
7	298	327	1426.00	7	52.32	347	278	1432.60	13	44.48
8	420	205	1426.07	7	32.8	89	536	1426.20	17	85.76
9	323	302	1425.93	5	48.32	107	518	1427.73	19	82.88
10	479	146	1425.92	16	23.36	330	295	1426.42	24	47.2
11	340	285	1425.93	14	45.6	296	329	1435.24	7	52.64
12	301	324	1426.27	5	51.84	379	246	1434.87	6	39.36
13	373	252	1424.19	9	40.32	370	255	1433.90	7	40.8
14	199	426	1425.84	15	68.16	263	362	1428.74	11	57.92
15	303	322	1431.23	3	51.52	179	446	1436.75	7	71.36
16	318	307	1426.42	12	49.12	243	382	1427.70	8	61.12
17	277	348	1423.85	8	55.68	342	283	1439.08	6	45.28
18	325	300	1425.73	9	48	370	255	1433.90	22	40.8
19	341	284	1425.81	6	45.44	323	302	1434.04	10	48.32
20	428	197	1426.47	9	31.52	190	435	1438.57	11	69.6

The Table 3 are examined summarizes the results in the Table 2. In Table 3, the proposed algorithm is evaluated against the previous algorithm using four major criteria (worst, best, mean and standard deviation) in the Balance data set. In this table, it can be seen that the proposed algorithm reduces the number of errors and the intra-cluster distance. It means the proposed algorithm provided better clustering. The most important criteria in the table are that the standard deviation of the five factors in the proposed algorithm is smaller than the previous algorithm, hence closer to stability. Whenever the standard deviation is small, distance has less variation in the results, including that the algorithm is stable.

Table 3. Worst, best, mean and standard deviation of intra-cluster distance, Numbers of True, Numbers of errors, iterations and error rate for the 20 runs of K-means, and the improved K-means on Balance data set

	Improved K-means algorithm					**K-means algorithm**				
Criteria	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**
Average	340	284	1425	9.3	45.5	275	349	1432	13.1	55.9
Std. Dev.	63	63	1.3	3.9	10.2	91	91	4.2	6.65	14.7
Best	479	146	1423	3	23.3	347	278	1426	6	44.4
Worst	199	426	1430	17	68.1	89	536	1439	24	85.7

ii. Blood

Secondly, the proposed k-means algorithm was applied to the Blood dataset. The specifications of this data set are described in the previous section, but will be mentioned here briefly. The Blood data set has 748 instances and 2 classes. In the MATLAB software, code programming proposed K-means algorithm is implemented and Blood data set is loaded. The proposed algorithm clustered Blood data set is shown in Figure 4. In this figure, the scattering diagram shows improved K-means clustering algorithm on the Blood data set. This diagram indicates the 748 members and the distance among members in this data set. In the diagram it is shown that a small number of members are scattered, mostly in the one level.

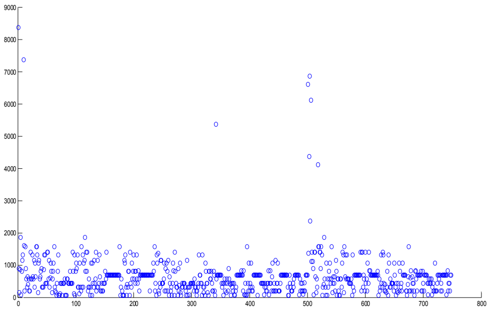

Figure 4. Scatter diagram of Blood data set clustered with improved K-means algorithm

Figure 5 display two clusters in Blood data set clustered with improved K-means clustering algorithm. In this data set are two clusters that are not regular. This means that the first and second clusters are merged; the Blood data set is such that the first and second clusters are not completely separated.

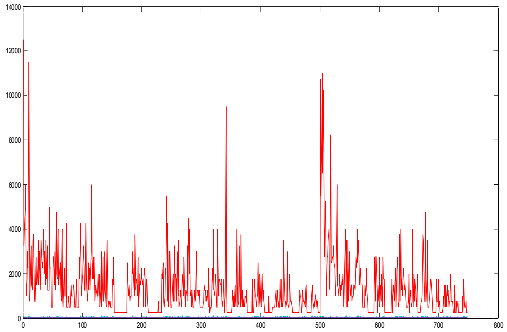

Figure **5.** Display of Blood data set clustering with improved K-means algorithm

In Table 4, the complete results of the K-means clustering algorithm and the improved K-means clustering algorithm are fully described for the Blood data set. Each algorithm was run twenty times and each time, the algorithm was repeated 50 times to achieve stability. In general, the algorithm is executed 1000 times for each data set. In this table, five factors (number of true, number of errors, intra cluster distance, iteration and error rate) are compared with the improved K-means clustering algorithm and K-means clustering algorithm on the Blood data set. In all factors, the proposed algorithm is much better than previous algorithms.

Table 4. The Number of true, Number of errors, Intra cluster distance, iterations and error rate for the 20 runs of K-means and the improved K-means on Blood data set

	Improved K-means algorithm					**K-means algorithm**				
Running	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**
1	543	205	469637	3	27.4	541	207	469637	8	27.6
2	541	207	469637	6	27.6	201	547	508640	5	73.1
3	541	207	469637	6	27.6	541	207	469637	9	27.6
4	541	207	469637	7	27.6	207	541	469637	12	72.3
5	207	541	469637	7	72.3	541	207	726188	4	27.6
6	541	207	469637	7	27.6	541	207	469637	9	27.6
7	570	178	482775	2	23.7	207	541	469637	12	72.3
8	541	207	469637	5	27.6	541	207	609979	7	27.6
9	541	207	469637	4	27.6	541	207	469637	9	27.6
10	541	207	469637	4	27.6	207	541	482775	4	72.3
11	540	208	469637	8	27.8	541	207	469637	7	27.6
12	541	207	469637	3	27.6	541	207	469637	9	27.6
13	541	207	469637	5	27.6	539	209	469637	7	27.9
14	541	207	469637	3	27.6	207	541	469637	11	72.3
15	541	207	469637	8	27.6	541	207	469637	8	27.6
16	541	207	469637	8	27.6	539	209	532933	5	27.9
17	541	207	469637	5	27.6	205	543	469637	11	72.5
18	541	207	469637	5	27.6	541	207	469637	9	27.6
19	570	178	469637	7	23.7	541	207	482775	5	27.6
20	541	207	469637	8	27.6	541	207	469637	8	27.6

Table 5 examines a summary of the results from Table 4. In Table 5, the proposed algorithm and the previous algorithm are evaluated using four major criteria (worst, best, mean and standard deviation) in the Blood data set. In this table, it can be seen that the proposed algorithm reduces the number of errors and the intra-cluster distance. It means that the proposed algorithm offers better clustering. The most important criteria in the table is that the standard deviation of the five factors proposed algorithm was smaller than previous algorithms, indicating proximity to stability. Whenever the standard deviation is small, there is less distance variation in the results which implies the algorithm is stable.

Table 5. Worst, best, mean and standard deviation of intra-cluster distance, Numbers of True, Numbers of errors, iterations and error rate for the 20 runs of K-means, and the improved K-means on Blood data set

	Improved K-means algorithm					K-means algorithm				
Criteria	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Errors rate (%)**
Average	527	220	470294	5.5	29.5	440	307	495910	7.9	41.1
Std. Dev.	75	75	2937	1.9	10.1	157	157	63900	2.4	21.0
Best	570	178	469637	2	23.7	541	207	469637	4	27.6
Worst	207	541	482775	8	72.3	201	547	726188	12	73.1

iii. Breast

In this section, the proposed k-means algorithm was applied to the Breast dataset. The specifications of this dataset are described in the previous section but will, however, be mentioned briefly here. The Breast data set has 699 instances and 2 classes. In the MATLAB software, code programming the proposed K-means algorithm is implemented and Breast data set is loaded. The proposed algorithm clustered Breast data set is shown in Figure 6. In this figure, the scattering diagram is shows an improved K-means clustering algorithm on the Breast data set. This diagram indicates the 699 members and the distance among members in this data set.

In Figure 7, two clusters of Breast data set clustered with improved K-means clustering algorithm are displayed. In this data set are two clusters that are not regular. This indicates that the first and second clusters are merged; the Breast data set is such that the first and second clusters are not separated.

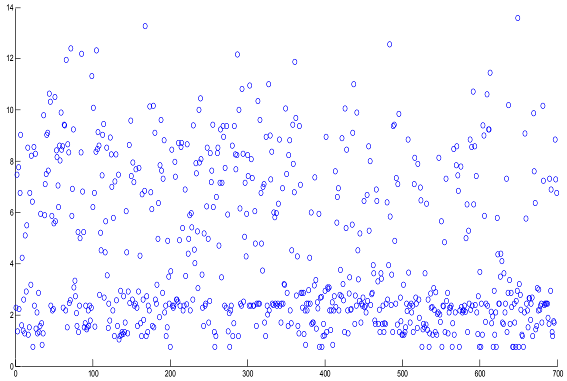

Figure 6. Scatter diagram of Breast data set clustered with improved K-means algorithm

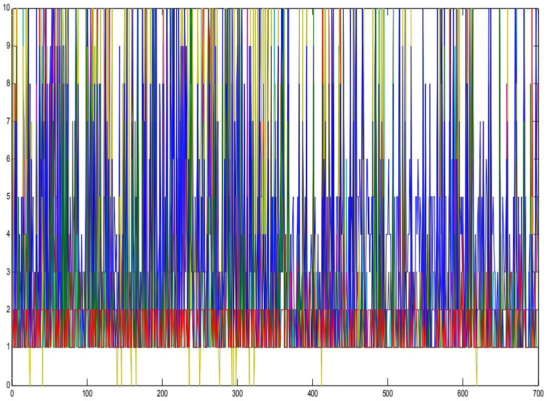

Figure 7. Display clustering the Breast data set with improved K-means algorithm

In Table 6, the complete results of the K-means clustering algorithm and the improved K-means clustering algorithm are fully described in the Breast data set. Each algorithm was run twenty times and each time, the algorithm was repeated 50 times to achieve stability. In general, the algorithm is executed 1000 times for each data set. In this table, five factors (number of true, number of errors, intra cluster distance, iteration and error rate) are compared to the improved K-means clustering algorithm and K-means clustering algorithm on the Breast data set. In all factors, the proposed algorithm is much better than previous algorithms.

In Table 7, the proposed algorithm and the previous algorithm are evaluated using four major criteria (worst, best, mean and standard deviation) in the Breast data set. In this table, it can be seen that the proposed algorithm reduces the number of errors and the intra-cluster distance. This indicates that the proposed algorithm offers better clustering. The most important criteria in the table are the standard deviation of the five factors proposed algorithm is smaller than previous algorithm, which implies stability. Whenever the standard deviation is small, there is less distance variation in the results which implies the algorithm is stable.

Table 6. The Number of true, Number of errors, Intra cluster distance, iterations and error rate for the 20 runs of K-means and the improved K-means on Breast data set

	Improved K-means algorithm					K-means algorithm				
Running	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**
1	670	29	3056	6	4.1	670	29	3056	7	4.1
2	670	29	3056	5	4.1	670	29	3056	7	4.1
3	670	29	3056	6	4.1	670	29	3056	6	4.1
4	670	29	3056	4	4.1	665	34	3056	6	4.8
5	670	29	3056	4	4.1	670	29	3056	9	4.1
6	670	29	3056	7	4.1	670	29	3056	7	4.1
7	670	29	3056	5	4.1	670	29	3056	6	4.1
8	670	29	3056	5	4.1	670	29	3056	6	4.1
9	670	29	3056	5	4.1	670	29	3056	6	4.1
10	670	29	3056	4	4.1	670	29	3056	7	4.1
11	670	29	3056	4	4.1	660	39	3056	5	5.5
12	670	29	3056	6	4.1	670	29	3056	6	4.1
13	670	29	3056	6	4.1	670	29	3056	6	4.1
14	670	29	3056	6	4.1	670	29	3056	6	4.1
15	670	29	3056	6	4.1	670	29	3056	8	4.1
16	670	29	3056	5	4.1	670	29	3056	7	4.1
17	670	29	3056	6	4.1	670	29	3056	6	4.1
18	670	29	3056	5	4.1	670	29	3056	6	4.1
19	670	29	3056	4	4.1	670	29	3056	9	4.1
20	670	29	3056	6	4.1	670	29	3056	6	4.1

Table 7. Worst, best, mean and standard deviation of intra-cluster distance, Numbers of True, Numbers of errors, iterations and error rate for the 20 runs of K-means, and the improved K-means on Breast data set

	Improved K-means algorithm					K-means algorithm				
Criteria	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Errors rate (%)**
Average	670	29	3056	5.2	4.14	669	29.7	3059	6.6	4.25
Std. Dev.	0	0	1.39	0.9	0	2.4	2.4	8.96	1	0.35
Best	670	29	3056	4	4.14	670	29	3056	5	4.14
Worst	670	27.4	3056	7	3.93	660	39	3069	9	5.57

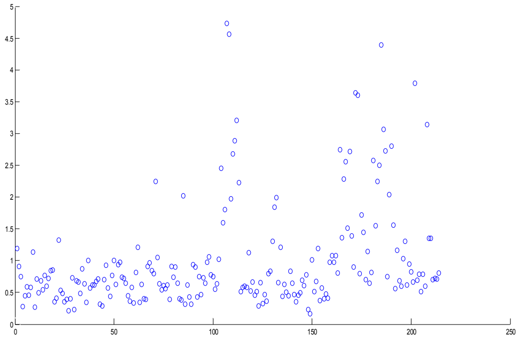

Figure 8. Scatter diagram of Glass data set clustered with improved K-means algorithm

iv. *Glass*

In this section, the proposed k-means algorithm was applied to the Glass dataset. The specifications of this data set are described in the previous section, but will be mentioned briefly here. The Glass data set has 214 instances and 7 classes. In the MATLAB software, code programming the proposed K-means algorithm is implemented and Glass data set is loaded. The proposed algorithm clustered Glass data set is shown in Figure 8. In this figure, the scattering diagram shows improved K-means clustering algorithm on the Glass data set. This diagram indicates the 214 members and the distance among members in this data set.

In Figure 9, clustering of seven clusters in Glass data set with improved K-means clustering algorithm is displayed. This data set has seven regular clusters, indicating that the clusters are not merged.

In Table 8, the complete results of the K-means clustering algorithm and the improved K-means clustering algorithm are fully described in the Glass data set. Each algorithm was run twenty times and each time, the run was repeated 50 times to achieve stability. In general, the algorithm is executed 1000 times for each data set. In this table, five factors (number of true, number of errors, intra cluster distance, iteration and error rate) are compared to the improved K-means clustering algorithm and K-means clustering algorithm on the Glass data set. In all factors, the proposed algorithm is much better than previous algorithms.

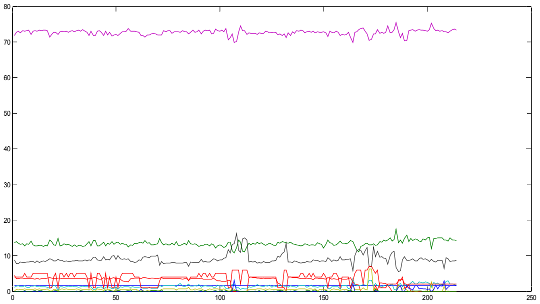

Figure 9. Display clustering the Glass data set with improved K-means algorithm

Table 9 are examined summarizes the results in the Table 8. In Table 9, the proposed algorithm and the previous algorithm are evaluated using four major criteria (worst, best, mean and standard deviation) in the Glass data set. In this table, it can be seen that the proposed algorithm reduces the number of errors and the intra-cluster distance, which means that the proposed algorithm offer better clustering. The most important criteria in the table is the standard deviation of the five factors proposed algorithm is smaller than the previous algorithm, indicating that the algorithm is close to stable. Whenever the standard deviation is small, there is less distance variation in the results which implies the algorithm is stable.

Table 8. The Number of true, Number of errors, Intra cluster distance, iterations and error rate for the 20 runs of K-means and the improved K-means on Glass data set

	Improved K-means algorithm					**K-means algorithm**				
Running	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**
1	148	66	212	8	30.8	112	102	205	10	47.6
2	118	96	203	11	44.8	119	95	206	24	44.3
3	122	92	210	13	42.9	86	128	250	10	59.8
4	121	93	206	16	43.4	122	92	255	15	42.9
5	122	92	206	14	42.9	111	103	214	10	48.1
6	112	102	207	8	47.6	100	114	215	15	53.2
7	119	95	203	14	44.3	121	93	213	13	43.4
8	122	92	203	9	42.9	120	94	211	15	43.9
9	121	93	204	8	43.5	143	71	224	12	33.1
10	124	90	205	12	42.5	120	94	210	19	43.9
11	148	66	208	16	30.4	110	104	208	10	48.5
12	122	92	213	15	42.9	101	113	213	18	52.8
13	122	92	212	9	42.9	137	77	216	11	35.9
14	124	90	208	8	42.5	109	105	208	15	49.0
15	117	97	209	8	45.2	120	94	242	10	43.9
16	150	64	209	9	29.0	121	93	209	13	43.4
17	122	92	206	8	42.9	98	116	221	13	54.2
18	122	92	205	10	42.9	118	96	208	17	44.8
19	140	74	203	9	34.5	121	93	212	12	43.4
20	122	92	208	12	42.9	110	104	219	12	48.5

Table9. Worst, best, mean and standard deviation of intra-cluster distance, Numbers of True, Numbers of errors, iterations and error rate for the 20 runs of K-means, and the improved K-means on Glass data set.

	Improved K-means algorithm					**K-means algorithm**				
Criteria	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Errors rate (%)**
Average	125	88	207	10.8	41.1	114	99	218	13.7	46.28
Std. Dev.	11	11	3.1	2.9	5.15	12	12	14.3	3.672	6.051
Best	150	64	203	8	29.9	143	71	206	10	33.17
Worst	112	102	213	16	47.6	86	128	255	24	59.81

v. Iris

In this section, the proposed k-means algorithm was applied to the Iris dataset. The specifications of this data set are described in the previous section, but will be mentioned briefly here. The Iris data set has 150 instances and 3 classes. This data set is used in many research studies and is very popular and so will be used in this study. In the MATLAB software, code programming the proposed K-means algorithm is implemented and Iris data set is loaded. In Figure 10, the scattering diagram is shown of improved K-means clustering algorithm on the Iris data set. This diagram indicates the 150 members and the distance among members in this data set.

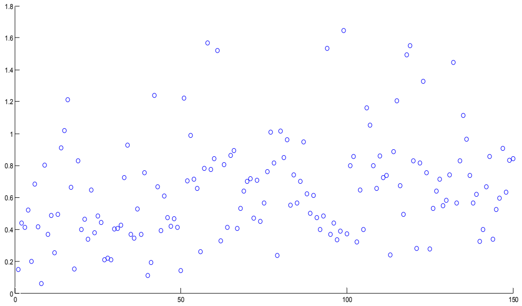

Figure 10. Scatter diagram of Iris data set clustered with improved K-means algorithm

In Figure 11, clustering of three clusters in Iris data set with improved K-means clustering algorithm is displayed. This data set has three clusters (first cluster, the first to fifty members; second cluster, members fifty one to one hundred; third cluster, members one hundred one to one hundred fifty). It can be seen that, in the first cluster, there are no errors after clustering, which means an error rate for the first cluster of zero. In the second cluster, there is a small error rate, but in the third cluster, the error rate is higher than both previous clusters. In general, the graphs display the clustering in the Iris data set using improved K-mean clustering algorithm.

In Table 10, the complete results of the K-means clustering algorithm and the improved K-means clustering algorithm are fully described in the Iris data set. Each algorithm was run twenty times and each time, the run was repeated 50 times to achieve stability. In general, the algorithm is executed 1000 times for each data set. In this table, five factors (number of true, number of errors, intra cluster distance, iteration and error rate) are compared to the improved K-means clustering algorithm and K-means clustering algorithm on the Iris data set. In all factors, the proposed algorithm is much better than previous algorithms.

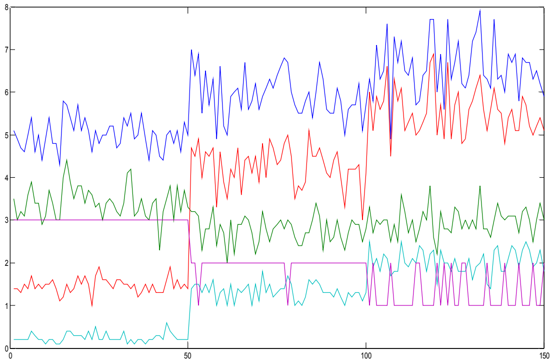

Figure 11. Display clustering the Iris data set with improved K-means algorithm

Table10. The Number of true, Number of errors, Intra cluster distance, iterations and error rate for the 20 runs of k-means and the improved k-means on Iris data set

	Improved K-means algorithm					K-means algorithm				
Running	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**
1	134	16	97.3	7	10.6	125	25	122	8	16.6
2	133	17	97.3	8	11.3	126	24	122	4	16
3	133	17	97.3	4	11.3	125	25	97.3	4	16.6
4	134	16	97.3	6	10.6	133	17	97.3	7	11.3
5	134	16	97.3	8	10.6	126	24	103	12	16
6	134	16	97.3	3	10.6	133	17	97.3	7	11.3
7	134	16	97.3	8	10.6	123	27	97.3	6	18
8	136	14	97.3	4	9.3	133	17	97.3	5	11.3
9	134	16	97.3	9	10.6	133	17	97.3	5	11.3
10	133	17	97.3	8	11.3	126	24	110	7	16
11	133	17	97.3	9	11.3	133	17	97.3	15	11.3
12	134	16	97.3	6	10.6	133	17	97.3	4	11.3
13	134	16	97.3	3	10.6	126	24	97.3	9	16
14	133	17	97.3	8	11.3	133	17	97.3	12	11.3
15	134	16	97.3	4	10.6	133	17	97.3	6	11.3
16	133	17	97.3	5	11.3	125	25	122	4	16.6
17	133	17	97.3	6	11.3	133	17	112	11	11.3
18	133	17	97.3	8	1	133	17	97.3	8	11.3
19	134	16	97.3	3	10.6	123	27	97.3	13	18
20	133	17	97.3	5	11.3	133	17	97.3	9	11.3

In Table 11, the proposed algorithm and the previous algorithm are evaluated using four major criteria (worst, best, mean and standard deviation) in the Iris data set. In this table, it can be seen that the proposed algorithm reduces the number of errors and the intra-cluster distance. It means that the proposed algorithm is better at clustering. The most important criteria in the table is the standard deviation of the five factors proposed algorithm is smaller than the previous algorithm, thus closer to being stable. Whenever the standard deviation is small, there is less distance variation in the results which implies the algorithm is stable.

Table11. Worst, best, mean and standard deviation of intra-cluster distance, Numbers of True, Numbers of errors, iterations and error rate for the 20 runs of K-means, and the improved K-means on Iris data set

	Improved K-means algorithm					**K-means algorithm**				
Criteria	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Errors rate (%)**
Average	133	16.3	97.3	6.1	10.9	129	20.6	102	7.8	13.7
Std. Dev.	0.74	0.7	0.01	2.10	0.4	4.1	4.15	9.55	3.31	2.7
Best	136	14	97.3	3	9.3	133	17	97.3	4	11.3
Worst	133	17	97.3	9	11.3	123	27	122	15	18

vi. Pima

In this section, the proposed k-means algorithm was applied to the Pima dataset. The specifications of this data set are described in the previous section, but will be mentioned briefly here. The Pima data set has 768 instances and 2 classes. In the MATLAB software, code programming the proposed K-means algorithm is implemented and Pima data set is loaded. The proposed algorithm clustered Pima data set is shown in Figure 12. In this figure, the scattering diagram shows improved K-means clustering algorithm on the Pima data set. This diagram indicates the 768 members and the distance among members in this data set. In the diagram, it is shown that a small number of members are scattered.

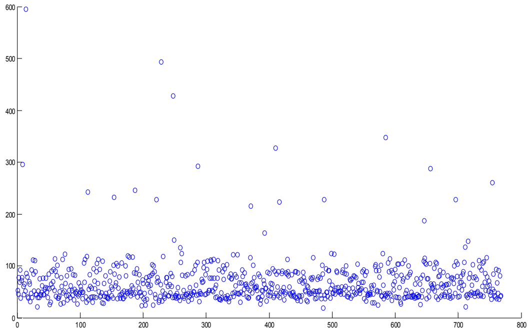

Figure 12. Scatter diagram of Pima data set clustered with improved K-means algorithm

In Figure 13, clustering of two clusters of Pima data set with improved K-means clustering algorithm is displayed. In this data set are two clusters they are not regular, which means that the first and second clusters are merged; the Pima data set is such that the first and second clusters are not separated.

In Table 12, the complete results of the K-means clustering algorithm and the improved K-means clustering algorithm are fully described in the Pima data set. Each algorithm was run twenty times and each time, the algorithm was run 50 times to achieve stability. In general, the algorithm is executed 1000 times for each data set. In this table, five factors (number of true, number of errors, intra cluster distance, iteration and error rate) are compared the improved K-means clustering algorithm and K-means clustering algorithm on the Pima data set. In all factors, the proposed algorithm is much better than previous algorithms.

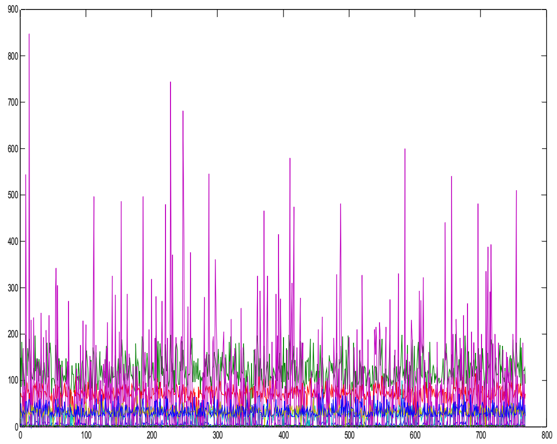

Figure13**.** Display clustering the Pima data set with improved K-means algorithm

Table 13 are examined summarizes the results in the Table 12. In Table 13, the proposed algorithm is evaluated with the previous algorithm using four major criteria (worst, best, mean and standard deviation) in the Pima data set. In this table, it can be seen that the proposed algorithm reduces the number of errors and the intra-cluster distance, which means that the proposed algorithm offer better clustering. The most important criteria in the table is the standard deviation of the five factors proposed algorithm is smaller than the previous algorithm, and is closer to being stable. Whenever the standard deviation is small, there is less distance variation in the results which implies and the algorithm is stable.

Table 12. The Number of true, Number of errors, Intra cluster distance, iterations and error rate for the 20 runs of K-means and the improved K-means on Pima data set

	Improved K-means algorithm					K-means algorithm				
Running	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**
1	507	261	52072	15	33.9	507	261	52072	16	33.9
2	507	261	52072	13	33.9	507	261	52216	14	33.9
3	507	261	52072	16	33.9	507	261	52072	16	33.9
4	507	261	52072	8	33.9	507	261	52072	18	33.9
5	507	261	52072	14	33.9	507	261	52167	10	33.9
6	507	261	52072	8	33.9	507	261	52216	12	33.9
7	507	261	52072	12	33.9	507	261	52072	14	33.9
8	507	261	52072	15	33.9	500	268	52072	15	34.8
9	507	261	52072	10	33.9	507	261	52072	11	33.9
10	507	261	52072	14	33.9	507	261	52072	12	33.9
11	507	261	52072	13	33.9	489	279	52110	15	36.3
12	507	261	52072	14	33.9	507	261	52072	12	33.9
13	507	261	52072	14	33.9	507	261	52072	14	33.9
14	507	261	52072	16	33.9	507	261	52072	14	33.9
15	507	261	52072	15	33.9	507	261	52290	18	33.9
16	507	261	52072	15	33.9	507	261	52072	22	33.9
17	507	261	52072	11	33.9	507	261	52479	16	33.9
18	507	261	52072	9	33.9	507	261	52072	17	33.9
19	507	261	52072	11	33.9	507	261	52072	14	33.9
20	507	261	52072	12	33.9	507	261	52072	13	33.9

Table13**.** Worst, best, mean and standard deviation of intra-cluster distance, Numbers of True, Numbers of errors, iterations and error rate for the 20 runs of K-means, and the improved K-means on Pima data set.

	Improved K-means algorithm					**K-means algorithm**				
Criteria	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Errors rate (%)**
Average	507	261	52070	12.7	33.9	505	262	52124	14.6	34.1
Std. Dev.	0	0	4.8	2.5	0	4.2	4.2	105	2.7	0.5
Best	507	261	52072	8	33.9	507	261	52072	10	33.9
Worst	507	261	52072	16	33.9	489	279	52479	22	36.3

vii. Wine

In the final step, the proposed k-means algorithm was applied to the Iris dataset. Although the specifications of this dataset are described in the previous section, they will be mentioned briefly here. The Wine data set has 178 instances and 3 classes. Because this data set similar to Iris data set is used in many research studies and is very popular, it is also used in this study. In the MATLAB software, code programming the proposed K-means algorithm is implemented and Wine data set is loaded. Then, the proposed algorithm clustered Wine data set into three clusters as shown in Figure 14.

In Figure 15, the scattering diagram is shown of improved K-means clustering algorithm on the Wine data set. This diagram indicates the 178 members and the distance among members in this data set.

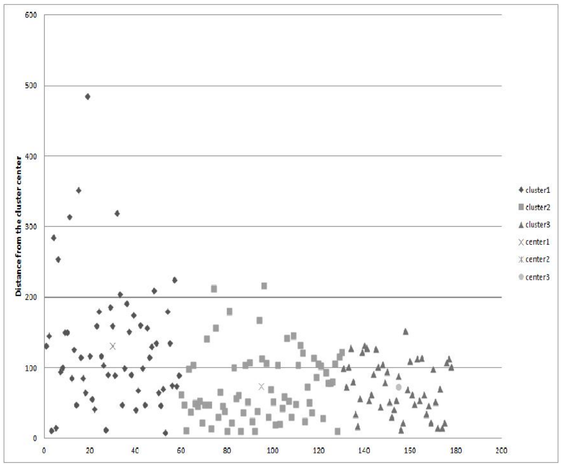

Figure14. Wine data set clustered into three clusters with improved k-means algorithm

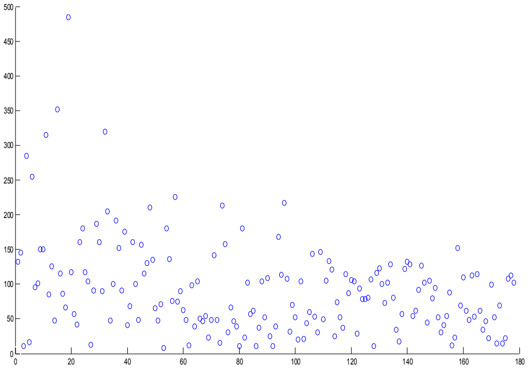

Figure 15. Scatter diagram of Wine data set clustered with improved K-means algorithm

In Figure 16, clustering of three clusters in Wine data set with improved K-means clustering algorithm is displayed. In this data set there are three clusters (first cluster includes members one to fifty nine; the second cluster, members sixty to one hundred thirty; and in the third cluster, members one hundred thirty one to one hundred seventy eight). It can be seen that in the first cluster, after clustering, there is a small error rate, but in the second and third cluster, there are high error rates. In general, the graphs display the clustering in the Wine data set using the improved K-mean clustering algorithm.

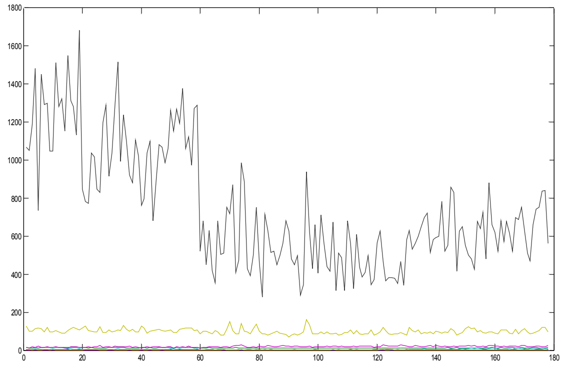

Figure 16. Display of clustering the Wine data set with improved K-means algorithm

In Table 14, the complete results of the K-means clustering algorithm and the improved K-means clustering algorithm are fully described in the Wine data set. Each algorithm was run twenty times and each time, the run was repeated 50 times to achieve stability. In general, the algorithm is executed 1000 times for each data set. In this table, five factors (number of true, number of errors, intra cluster distance, iteration and error rate) are compared to the improved K-means clustering algorithm and K-means clustering algorithm on the Wine data set. In all factors, the proposed algorithm is much better than previous algorithms.

Table14. The Number of true, Number of errors, Intra cluster distance, iterations and error rate for the 20 runs of K-means and the improved K-means on Wine data set

	Improved K-means algorithm					**K-means algorithm**				
Running	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**	**No. True**	**No. Errors**	**Intra cluster distance**	**Iteration**	**Error rate (%)**
1	125	53	16555	6	29.7	125	53	16555	7	29.7
2	125	53	16555	5	29.7	125	53	16555	8	29.7
3	125	53	16555	8	29.7	110	68	17292	8	38.2
4	125	53	16555	7	29.7	125	53	16555	8	29.7
5	125	53	16555	6	29.7	125	53	16555	7	29.7
6	125	53	16555	8	29.7	125	53	16555	11	29.7
7	125	53	16555	4	29.7	110	68	17390	12	38.2
8	132	46	16662	9	25.8	110	68	18467	10	38.2
9	125	53	16555	6	29.7	125	53	16555	9	29.7
10	125	53	16555	7	29.7	125	53	16555	9	29.7
11	125	53	16555	6	29.7	125	53	16555	8	29.7
12	132	46	16555	7	25.8	125	53	16555	7	29.7
13	125	53	16555	7	29.7	125	53	16555	10	29.7
14	125	53	16555	7	29.7	125	53	16904	7	29.7
15	125	53	16555	8	29.7	125	53	16555	9	29.7
16	125	53	16555	6	29.7	125	53	16555	9	29.7
17	125	53	16555	5	29.7	125	53	16555	8	29.7
18	125	53	16555	6	29.7	125	53	17474	9	29.7
19	125	53	16555	7	29.7	125	53	16911	7	29.7
20	125	53	16555	7	29.7	125	53	16555	8	29.7

Table 15 are examined summarizes the results in the Table 14. In Table 15, the proposed algorithm along with the previous algorithm is evaluated using four major criteria (worst, best, mean and standard deviation) in the Wine data set. In this table, it can be seen that the proposed algorithm reduces the number of errors and the intra-cluster distance. It means that the proposed algorithm has better clustering. The most important criteria in the table are that the standard deviation of the five factors in the proposed algorithm is smaller than the previous algorithm, hence closer to being stable. Whenever the standard deviation is small, there is less distance variation in the results which implies and the algorithm is stable.

Table 15. The Worst, best, mean and standard deviation of intra-cluster distance, Numbers of True, Numbers of errors, iterations and error rate for the 20 runs of K-means, and the improved K-means on Wine data set

Criteria	Improved K-means algorithm					K-means algorithm				
	No. True	No. Errors	Intra cluster distance	Iteration	Error rate (%)	No. True	No. Errors	Intra cluster distance	Iteration	Errors rate (%)
Average	2.1	2.1	23.9	1.1	1.2	122	55.2	16811	8.5	31
Std. Dev.	132	46	16555	4	25.8	5.4	5.4	496	1.3	3
Best	125	53	16662	9	29.7	125	53	16555	7	29.7
Worst	2.1	2.1	23.9	1.1	1.2	110	68	18467	12	38.2

In this section were expressed the experimental results obtained from the study. Seven data sets were used to test the proposed algorithm that was then compared with K-means clustering algorithms. In this part, data sets were completely described, and the proposed algorithm was then evaluated on the data sets with four criteria (average, standard deviation, best and worst) and five factors (Numbers of True, Numbers of errors, intra-cluster distance, iterations and error rate). Improved K-means clustering algorithm has better performance than the K-mean clustering algorithm in the all factors and criteria. In the next section the proposed algorithm and K-means clustering algorithm are compared in more detail.

6. Evaluation of Result

In this section the results of the experiments conducted in section 4 are compared to the results of the improved K-means clustering algorithm and the K-means clustering algorithm. One important factor for the clustering algorithm is intra cluster distance that will be reviewed first.

Figures 17, 18, 19, 20, 21, 22 and 23 are the analyzed intra cluster distance in the K-means algorithm with the improved K-means algorithm on the seven data sets (Balance, Blood, Breast, Glass, Iris, Pima and Wine). For better comparison, both algorithms are run 20 times on all data sets. All diagrams that can be seen in the intra cluster distance of the proposed clustering algorithm have been improved and in all cases the intra cluster distance is reduced. Also, the intra cluster distance is constant during program execution, indicating the algorithm is stable. It can be seen that the proposed algorithm 20 times the intra cluster distance, meaning that the algorithm is stable. One of the main problems in the K-means clustering algorithm is stability, which the proposed algorithm has almost solved.

Table 16 shows the average intra cluster distance in the above figures. It can be seen from the above figures that the clustering algorithm's behavior is not consistent; the answer is sometimes right and sometimes not. To solve this problem, the proposed algorithm was used. The following table shows a comparison of the average intra cluster distance in all seven data sets and that the proposed algorithm has reduced the average intra cluster distance.

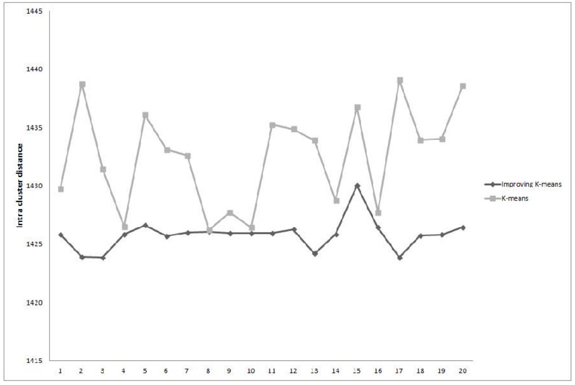

Figure 17. Diagram comparison of intra cluster distance between improved K-means algorithm and K-means algorithm in the Balance data set

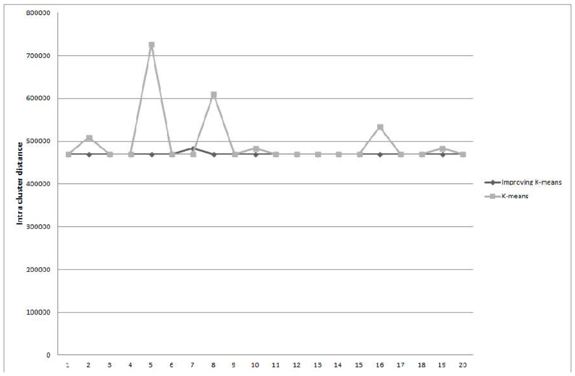

Figure 18. Diagram comparison of intra cluster distance between improved K-means algorithm and K-means algorithm in the Blood dataset

Table 17 shows the standard deviation intra cluster distance in the above figures for 20 algorithm runs. The standard deviation is small and close to zero, which is very good, indicating that the algorithm is stable. In the following table it can be seen that the proposed algorithm has reduced standard deviation which means the proposed algorithm is very near to being stable. As a result, the improved K-means clustering algorithm finds correct and stable answers most of the time.

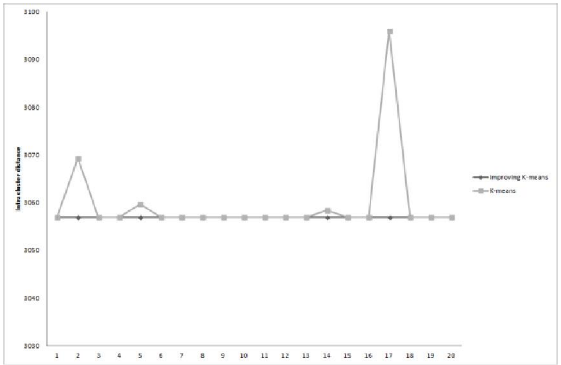

Figure 19. Diagram comparison of intra cluster distance between improved K-means algorithm and K-means algorithm in the Breast dataset

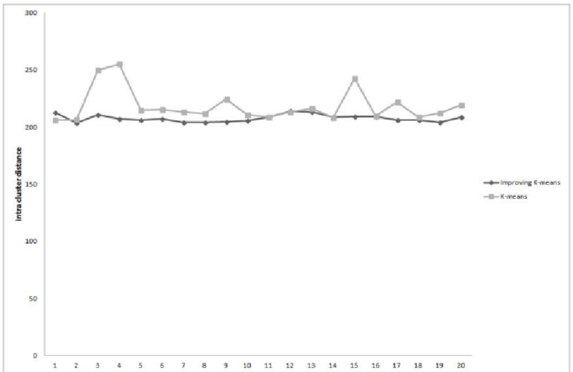

Figure 20. Diagram comparison of intra cluster distance between improved K-means algorithm and K-means algorithm in the Glass dataset

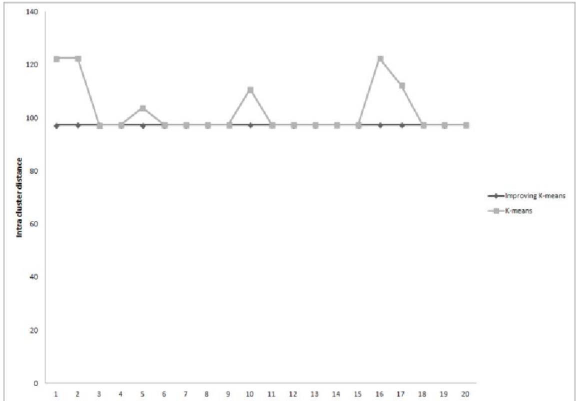

Figure 21. Diagram comparison of intra cluster distance between improved K-means algorithm and K-means algorithm in the Iris data set

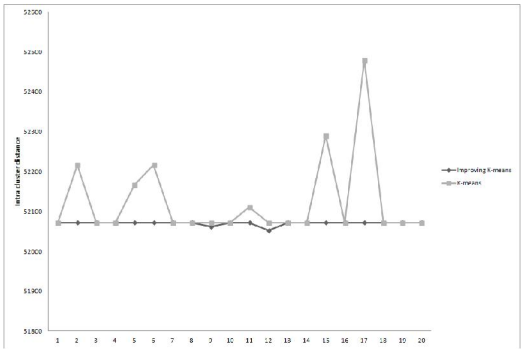

Figure 22. Diagram comparison of intra cluster distance between improved K-means algorithm and K-means algorithm in the Pima data set

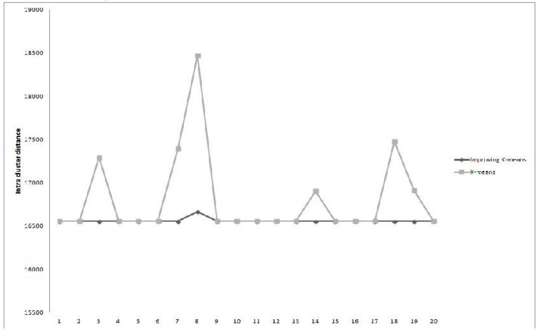

Figure 23. Diagram comparison of intra cluster distance between improved K-means algorithm and K-means algorithm in the Wine dataset

Table 16. Mean of intra cluster distance for the 20 running of K-means algorithm, and the improved K-means algorithm on seven data sets (Balance, Blood, Breast, Glass, Iris, Pima and Wine)

Intra cluster distance (Average)		
Data set	**Improved K-means algorithm**	**K-means algorithm**
Balance	1425.81	1432.56
Blood	47029.49	495911
Breast	3056.96	3059.74
Glass	207.6	218.37
Iris	97.33	102.86
Pima	52070.74	52124.6
Wine	16561.03	16811.03

Table 17. Standard deviation of intra cluster distance for the 20 runs of K-means algorithm, and the improved K-means algorithm on seven data sets (Balance, Blood, Breast, Glass, Iris, Pima and Wine)

Intra cluster distance (Std)		
Data set	**Improved K-means algorithm**	**K-means algorithm**
Balance	1.32	4.29
Blood	2937.76	63900.08
Breast	1.39	8.96
Glass	3.16	14.32
Iris	0.01	9.55
Pima	4.89	105.29
Wine	23.92	496.55

Table 18 displays the average error rate in the above figures over 20 algorithm runs. This table shows that the error rate of the improved K-means algorithm is lower rather than the K-means clustering algorithm. This means that the proposed algorithm is able to improve clustering and also reduce the number of errors. Another factor is the stability of the algorithm is reduction of in the algorithm that it can be seen which the proposed algorithm.

Table 18. Mean of error rate for the 20 runs of K-means algorithm, and the improved K-means algorithm on seven data sets (Balance, Blood, Breast, Glass, Iris, Pima and Wine).

Error Rate (Average)		
Data set	**Improved K-means algorithm**	**K-means algorithm**
Balance	45.58	55.98
Blood	29.51	41.14
Breast	4.14	4.25
Glass	41.16	46.28
Iris	10.9	13.73
Pima	33.98	34.14
Wine	29.38	31.03

Figure 24 shows a comparison of the average error rate between the improved K-means clustering algorithm and the K-means clustering algorithm in the seven data sets over 20 runs. In this figure it can be seen that the error rate in the proposed algorithm is less than K-means clustering algorithm. In Figure 32, Blood and Balance data sets have most reduction in the error rate. In general, there is a reduction in error rates in the all data sets.

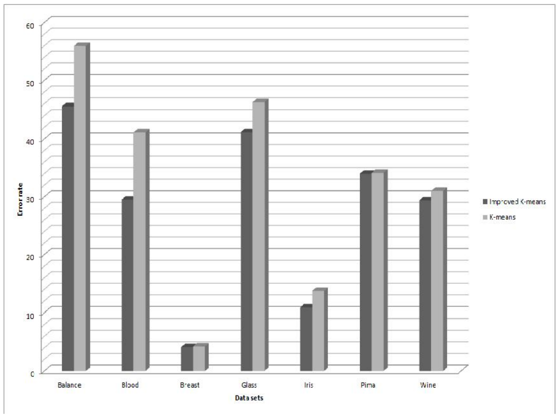

Figure 24. Comparison diagram of the error rates for seven datasets using the improved K-means algorithm and the K-means algorithm

In this section, the results of the experiments obtained in section 4 are discussed and evaluated. The results were discussed with the three criteria, intra cluster distance (average), intra cluster distance (standard deviation) and error rate (average) for seven data sets. In the three comparison criteria between the improved K-means clustering algorithms and K-means clustering algorithm it can be seen that the improved K-means clustering algorithm is the best in each case. In general, the proposed algorithm reduces the error rate and intra cluster distance, and it will lead the clustering algorithm to become stable.

7. Conclusion

This paper focuses on a disadvantage of the K-means clustering algorithm, which is that the clustering algorithm has a high error rate. It also referred to one of the main problems in the K-means algorithm which is that the K-means clustering algorithm is not always stable. In this study, an algorithm was proposed to solve this problem in order to reduce the error rate in K-means clustering algorithms and to stabilize the algorithm. In this paper, examining the improved K-means clustering algorithm with the K-means clustering algorithm involved the consideration of five factors (average, standard deviation, best and worst) and four criteria (numbers of true, numbers of errors, intra-cluster distance, iterations and error rate). For comparing the improved K-means algorithm and K-means algorithm seven data sets were used (Balance, Blood, Breast, Glass, Iris, Pima and Wine) and the proposed algorithm shows better performance in all these data sets. In summary, the proposed algorithm has better efficiency than the K-means clustering algorithm in the all measures used in this study, the intra cluster distance and error rate was reduced in the proposed algorithm and the improved algorithm is closer to being stable.

In this study a method is proposed to solve one of the main problems of K-means clustering algorithm, which is that the algorithm is not always consistent. In this survey, one of the best ways to calculate the distance it to use the Euclidean distance calculation in the MATLAB software to calculate the members distance from the cluster center. Also, it should be noted that when calculating the Euclidean distance, additional operations that have a negative effect on the calculation must be avoided. Thus, the purpose of this paper is to improve the calculation of the members distance from the center of the cluster, which this will help to stabilize algorithm clustering and reduce the error rate.

Future work related to this paper can be done as a continuation as other problems of clustering algorithms can be studied using these data sets. Also, the proposed algorithm in this paper can be examined on other data sets and clustering algorithms and the obtained results compared. Finally, other criteria can be studied with the proposed algorithm on the new data set and data sets in this article.

Acknowledgment:

This publication was supported by The Ministry of Higher Education (MOHE) under Long Term Research Grant Scheme, R.J130000.7828.4L805-MULTI-VARIATE PATTERN ANALYSIS FOR FORENSIC IDENTIFICATION AND VERIFICATION (Research university grant). The authors would like to thank Universiti Teknologi Malaysia (UTM), Research Management Centre (RMC) for the support in R & D, Soft Computing Research Group (SCRG) for the inspiration in making this study a success.

Corresponding Author:

Prof. Dr. Siti Mariyam Hj. Shamsuddin
Soft Computing Research Group
Faculty of Computing, Universiti Teknologi Malaysia, Skudai, 81310, Johor, Malaysia
Tel.: +60 123710679, Fax: +607-5538793
E-mail: mariyam@utm.my

References

1. Leuski A, editor. Evaluating document clustering for interactive information retrieval. Proceedings of the tenth international conference on Information and knowledge management; 2001: ACM.
2. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: A survey. Knowledge and Data Engineering, IEEE Transactions on. 2004;16(11):1370-86.
3. Kogan J, Nicholas CK, Teboulle M. Grouping multidimensional data: Springer; 2006.
4. Bayat F, Mosabbeb EA, Jalali AA, Bayat F. A non-parametric heuristic algorithm for convex and non-convex data clustering based on equipotential surfaces. Expert Systems with Applications. 2010;37(4):3318-25.
5. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 2010;31(8):651-66.
6. Ju C, Xu C. A New Collaborative Recommendation Approach Based on Users Clustering Using Artificial Bee Colony Algorithm. The Scientific World Journal. 2013;2013.
7. Cormack RM. A review of classification. Journal of the Royal Statistical Society Series A (General). 1971:321-67.
8. Cox DR. Note on grouping. Journal of the American Statistical Association. 1957;52(280):543-7.
9. Engelman L, Hartigan JA. Percentage points of a test for clusters. Journal of the American Statistical Association. 1969;64(328):1647-8.
10. Fisher WD. On grouping for maximum homogeneity. Journal of the American Statistical Association. 1958;53(284):789-98.
11. Thorndike RL. Who belongs in the family? Psychometrika. 1953;18(4):267-76.
12. MacQueen J, editor. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability; 1967: California, USA.
13. Sebestyen GS. Decision-making processes in pattern recognition (ACM monograph series). 1962.
14. Johnson RA, Wichern DW. Applied multivariate statistical analysis: Prentice hall Upper Saddle River, NJ; 2002.
15. Lattin JM, Carroll JD, Green PE. Analyzing multivariate data: Thomson Brooks/Cole Pacific Grove, CA; 2003.
16. Timm NH. Applied multivariate analysis: Springer; 2002.
17. Gordon A. Classification. 1999. Chapman&Hall, CRC, Boca Raton, FL. 1999.
18. Duda RO, Hart PE, Stork DG. Unsupervised learning and clustering. Pattern classification. 2001:571.
19. Hastie T, Tibshirani R, Friedman JJH. The elements of statistical learning: Springer New York; 2001.
20. Steinley D. K‐means clustering: A half‐century synthesis. British Journal of Mathematical and Statistical Psychology. 2006;59(1):1-34.
21. Forgy EW. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics. 1965;21:768-9.
22. Anderberg MR. Cluster analysis for applications. DTIC Document, 1973.
23. Jancey R. Multidimensional group analysis. Australian Journal of Botany. 1966;14(1):127-30.
24. Emre Celebi M, Kingravi HA, Vela PA. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications. 2012.
25. Norusis MJ. IBM SPSS statistics 19 statistical procedures companion: Prentice Hall; 2012.
26. Ball GH, Hall DJ. A clustering technique for summarizing multivariate data. Behavioral science. 1967;12(2):153-5.
27. Tou JT, Gonzalez RC. Pattern recognition principles. 1974.
28. Institute S, Publishing PS. SAS/STAT 9.2 User's Guide: The Glimmix Procedure (Book Excerpt): SAS Institute; 2008.
29. Katsavounidis I, Jay Kuo C-C, Zhang Z. A new initialization technique for generalized Lloyd iteration. Signal Processing Letters, IEEE. 1994;1(10):144-6.
30. Al-Daoud MdB, Roberts SA. New methods for the initialisation of clusters. Pattern Recognition Letters. 1996;17(5):451-5.
31. Bradley PS, Fayyad UM, editors. Refining Initial Points for K-Means Clustering. ICML; 1998.

32. Pizzuti C, Talia D, Vonella G. A divisive initialisation method for clustering algorithms. Principles of Data Mining and Knowledge Discovery: Springer; 1999. p. 484-91.
33. Arthur D, Vassilvitskii S, editors. k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms; 2007: Society for Industrial and Applied Mathematics.
34. Su T, Dy JG. In search of deterministic methods for initializing K-means and Gaussian mixture clustering. Intelligent Data Analysis. 2007;11(4):319-38.
35. Lu J, Tang J, Tang Z, Yang J-Y. Hierarchical initialization approach for K-Means clustering. Pattern Recognition Letters. 2008;29(6):787-95.
36. Onoda T, Sakai M, Yamada S. Careful Seeding Method based on Independent Components Analysis for k-means Clustering. Journal of Emerging Technologies in Web Intelligence. 2012;4(1):51-9.
37. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. Neural Networks, IEEE Transactions on. 1999;10(3):626-34.
38. Barani G, Poovendhiran N. Impact of Industrial Policy on Small Scale Industries: A Cluster Analysis. Life Science Journal. 2013;10(3s).
39. Chang D, Zhao Y, Zheng C, Zhang X. A genetic clustering algorithm using a message-based similarity measure. Expert Systems with Applications. 2012;39(2):2194-202.
40. Chellatamilan T, Suresh R. Concept Based Query Expansion and Cluster Based Feature Selection for Information Retrieval. Life Science Journal. 2013;10(7s).
41. Wang T, Hung WN. Reliable Node Clustering for Mobile Ad Hoc Networks. Journal of Applied Mathematics. 2013;2013.
42. Steinhaus H. Sur la division des corp materiels en parties. Bull Acad Polon Sci. 1956;1:801-4.
43. Lloyd S. Least squares quantization in PCM. Information Theory, IEEE Transactions on. 1982;28(2):129-37.
44. Huang H, Tang Q, Liu Z. Adaptive Correction Forecasting Approach for Urban Traffic Flow Based on Fuzzy-Mean Clustering and Advanced Neural Network. Journal of Applied Mathematics. 2013;2013.
45. Chau M, Cheng R, Kao B, editors. Uncertain data mining: a new research direction. Proceedings of the Workshop on the Sciences of the Artificial, Hualien, Taiwan; 2005.
46. Jiawei H, Kamber M. Data mining: concepts and techniques. San Francisco, CA, itd: Morgan Kaufmann. 2001;5.
47. Meilă M, editor. The uniqueness of a good optimum for k-means. Proceedings of the 23rd international conference on Machine learning; 2006: ACM.
48. Jain AK, Dubes RC. Algorithms for clustering data: Prentice-Hall, Inc.; 1988.
49. Gorgônio FL, Costa JAF. PartSOM: A Framework for Distributed Data Clustering Using SOM and K-Means.
50. Yazdani D, Golyari S, Meybodi MR, editors. A new hybrid approach for data clustering. Telecommunications (IST), 2010 5th International Symposium on; 2010: IEEE.
51. Asuncion A, Newman D. UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2007).

4/26/2014