A novel method for K-Means clustering algorithm

Jinguo Zhao*

School of Computer and Information Science, Hunan Institute of Technology, Hunan, 421002, China

Received 1 March 2014, www.cmnt.lv

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 centre. 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.

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

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 [33]. 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 [26]. 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 [31]. 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 [6].

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 modelling 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 [23].

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. So, many investigators tried to operationalize this description by minimizing within-group disparity [11, 14, 15, 45]. 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. Based on this significant advancement, K-means has become very popular, earning a place in a variety of textbooks on multivariate techniques [28, 32, 46], cluster analysis [17], pattern recognition [12], statistical learning [19, 43]. 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.

***Corresponding author** e-mail: jinguo2014@126.com

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 [16] randomly allocates each point to one of the K clusters homogeneously. The centres 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 [2]. Jancey's technique [25] allocates to each centre a combinatorial point randomly generated within the space of data. However, as the data set fills the space, a number of these centres may be too distant from any of the points [2], which might lead to the formation of unfilled clusters [13].

MacQueen (1967) suggested two different techniques. The first technique is the default choice in the Quick Cluster method of IBM SPSS Statistics [38], which obtains the first K points in X as the centres. An obvious disadvantage of this technique is its sensitivity into data ordering. The second technique selects the centres 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 centres. Ball and Hall's technique [5] obtains the centre of X, as the first centre. It then crosses the points in optional order and obtains a point as a centre if it is at least T units apart from the formerly selected centres until K centres 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 [47] is the same as Ball and Hall's technique with the distinction that the first point in X is obtained as the first centre. This technique is applied in the FASTCLUS method of SAS [13, 22].

Maximin technique [30] selects the first centre c1 randomly and the i-th (i{2,3,...,K}) centre ci is selected to be the point that has the most minimum distance to the formerly chosen centres, that is c1,c2,...,ci-1. This technique was originally expanded as an approximation to the K-centre clustering problem. It should be referred that, motivated with a vector quantization request, Katsavounidis et al.'s variant [30] obtains the point with the greatest Euclidean standard as the first centre.

Al-Daoud's density technique [1] first regularly partitions the data space into M decomposed hyper-cubes. It then randomly chooses K Nm/N points as of hypercube m (m {1,2,…,M}) to take a total of K centres where Nm is the points number in hypercube m. Bradley and Fayyad's technique [7] 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 centres, each with K points. These centre sets are united into a superset that is then clustered through k-means J times, each time initialized by a diverse centre set. Members of the centre set that give the least SSE are then taken as the final centres. Pizzuti [40] 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 [3] interpolates between maximin technique and MacQueen's second technique. It selects the first centre randomly and the i-th (i{2,3,...,K}) centre is selected to be x, where md(x) denotes the distance of minimum from a point x to the previously chosen centres.

The PCA-Part technique [44] 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 centre 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 centres are then given through the centres 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{0,1,...}) are averaged to take weighted points at stage k+1 until at least 20 K points are taken. Onoda technique [39] first computes K Independent Components (ICs) [21] of X and then selects the i-th (i{1,2,...,K}) centre as the point that has the least cosine distance [13].

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 unlabelled objects [23]. 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 [8]. Selecting clusters of optimally is an NP-hard problem [48]. Clustering algorithms include many algorithms, and K-means algorithm is the most popular. -means algorithm is a rather simple but well-known algorithm for grouping objects [29]. 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. A standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-

code modulation, though it was not published until 1982. The classical K-means clustering algorithm aims to detect a set C of K clusters Cj with cluster mean cj to reduce the sum of squared errors. Number of clustering C is a very important parameter [20].

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 [27,37]. K-means begins with a primary partition with K clusters and allocates patterns to clusters so as to decrease the squared error.

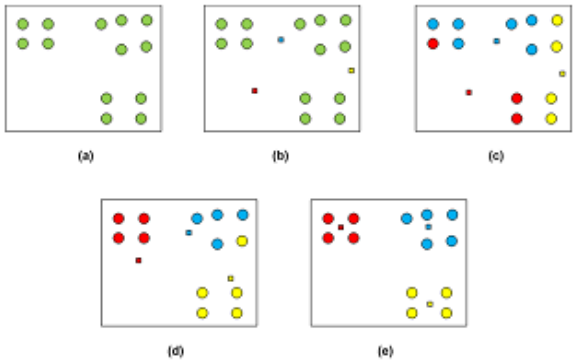

FIGURE 1 K-means clustering algorithm for 3 clusters

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 centres and initial assignment of the data points to clusters; (c) & (d) intermediate iterations updating cluster label and the centres; (e) final clustering obtained by K-means clustering algorithm at convergence [23]. 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 The proposed method

In the method used in this study, numbers were taken from outside of the cluster centre. 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 INITIALIZING THE CENTRE

In this study, initial value is randomly selected, after the data set was applied as cluster centres 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 centres 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 centres.

4.2 PROPOSED ALGORITHM

In the proposed algorithm (Reduction of Error Rates in K-Means algorithm or RER-K-Means algorithm), equation is used to calculate the distance between the members of dataset and cluster centres. 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 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 centre; it is located at one of the clusters. In this equation, the total aggregate of Euclidean distance of all the data vectors from cluster centres that they own is calculated and added to each other.

Therefore, by determining the optimal, centres 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 centres of the clusters are optimized, till, K-means clustering algorithm is to be improved. This study calculated the distance between the centre 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 centres 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

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.

Step 1: At first, the target dataset is applied to the MATLAB software. The dataset must have the clustering conditions.

Step 2: In this step, the number of rows of the dataset is found followed by selecting desired numbers of rows randomly as cluster centres. The selected attributes of the random rows are assumed to be initial cluster centres.

Step 3: 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.

Step 4: 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.

Step 5: At this stage, the distances of cluster centres which have been previously considered from all members of the dataset are calculated. To calculate the distance, the coordinates of the cluster centre 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 centres in one step.

Step 6: In this step, the distances of all cluster centres 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.

Step 7: In this step, some variables are defined to represent summation of distances between cluster centre 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 si is summation of distances among ith cluster centre to its member. (i=1,2,3).

Step 8: This step is the end of inter loop. It means that steps 4 to 7 are run until the ending condition of inter loop.

Step 9: Variable S which is intra cluster distance is defined as summation of s1, s2, s3and so on. From converging of S it is deducted that algorithm has stabilized. Generally, S should be tried to minimized as far as possible.

Step 10: The means of any cluster should be determined separately. Then, at the end of any step, the determined means are considered as cluster centres for the next step.

Step 11: This step is the end of outer loop. It means steps 3 to 8 are run until the ending condition of outer loop.

4.4 THE FORMULA

The steps shown in section 4.3 were used to calculate the error rate. It was required to calculate two measures; the number of error patterns and the total number of patterns, which 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 DATA SET

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 Centre (Blood): This data set adopted the donor database of Blood Transfusion Service Centre 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 [18].

Glass Identification (Glass): This data set has 214 samples and seven classes. Every sample in this data set has 10 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 [18, 49].

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 [49].

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 [18]. The databases used were obtained from the UCI data warehouse [4].

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.

5.2 RESULTS OF ERROR RATE

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 analyse 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. In Figure 2, 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.

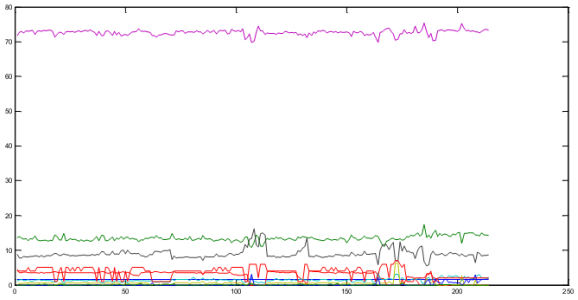

FIGURE 2 Display clustering the Glass data set with improved K-means algorithm

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.

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 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. 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. In Figure 3, 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.

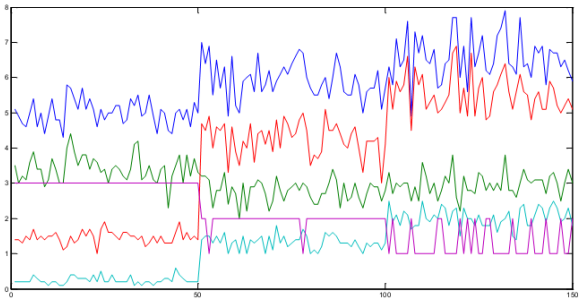

FIGURE 3 Display clustering the Iris data set with improved K-means algorithm

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 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. 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. 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. 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.

5.3 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. 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.

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.

6 Conclusions

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 centre. 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 centre 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.

References

[1] Al-Daoud M d B, Roberts S A 1996 New methods for the initialisation of clusters *Pattern Recognition Letters* **17** 451-5
[2] Anderberg M R 1973 Cluster analysis for applications *DTIC Document*
[3] Arthur D, Vassilvitskii S 2007 K-means++ The advantages of careful seeding *in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* 1027-35
[4] Asuncion A, Newman D 2007 UCI Machine Learning Repository *University of California School of Information and Computer Science Irvine CA* ed 24
[5] Ball G H, Hall D J 1967 A clustering technique for summarizing multivariate data *Behavioral science* **12** 153-5
[6] Bayat F, et al. 2010 A non-parametric heuristic algorithm for convex and non-convex data clustering based on equipotential surfaces *Expert Systems with Applications* **37** 3318-25
[7] Bradley P S, Fayyad U M 1998 Refining Initial Points for K-Means Clustering *in ICML* 91-9
[8] Chang D, et al. 2012 A genetic clustering algorithm using a message-based similarity measure *Expert Systems with Applications* **39** 2194-202
[9] Chau M, et al. 2005 Uncertain data mining: a new research direction *in Proceedings of the Workshop on the Sciences of the Artificial Hualien Taiwan* 199-204
[10] Cormack R M 1971 A review of classification *Journal of the Royal Statistical Society Series A (General)* 321-67
[11] Cox D R 1957 Note on grouping *Journal of the American Statistical Association* **52** 543-7
[12] Duda R O, et al. 2001 Pattern classification *2nd Edition. New York*
[13] Emre C M, et al. 2012 A comparative study of efficient initialization methods for the K-means clustering algorithm *Expert Systems with Applications*
[14] Engelman] L, Hartigan J A 1969 Percentage points of a test for clusters *Journal of the American Statistical Association* **64** 1647-8
[15] Fisher W D 1958 On grouping for maximum homogeneity *Journal of the American Statistical Association* **53** 789-98
[16] Forgy E W 1965 Cluster analysis of multivariate data: efficiency versus interpretability of classifications *Biometrics* **21** 768-9
[17] Gordon A 1999 Classification. 1999 *Chapman&Hall CRC Boca Raton FL*
[18] Gorgônio F L, Costa J A F PartSOM A Framework for Distributed Data Clustering Using SOM and K-Means
[19] Hastie T 2001 The elements of statistical learning *Springer New York*
[20] Huang H, et al. 2013 Adaptive Correction Forecasting Approach for Urban Traffic Flow Based on Fuzzy-Mean Clustering and Advanced Neural Network *Journal of Applied Mathematics*
[21] Hyvarinen A 1999 Fast and robust fixed-point algorithms for independent component analysis *Neural Networks IEEE Transactions on* **10** 626-34
[22] S. Institute and P. S. Publishing, SAS/STAT 9.2 User's Guide The Glimmix Procedure (Book Excerpt) SAS Institute 2008
[23] Jain A K 2010 Data clustering: 50 years beyond K-means *Pattern Recognition Letters* **31** 651-66
[24] Jain A K, Dubes R C 1988 Algorithms for clustering data *Prentice-Hall Inc*
[25] Jancey R 1966 Multidimensional group analysis *Australian Journal of Botany* **14** 127-30
[26] Jiang D, Tang C, Zhang A 2004 *Knowledge and Data Engineering, IEEE Transactions on* **16**(11) 1370-86
[27] Jiawei H, Kamber M 2001 Data mining: concepts and techniques *San Francisco, CA, itd: Morgan Kaufmann* **5**
[28] Johnson R A, Wichern D W 2002 Applied multivariate statistical analysis *Prentice hall Upper Saddle River, NJ* **5**
[29] Ju C, Xu C 2013 A New Collaborative Recommendation Approach Based on Users Clustering Using Artificial Bee Colony Algorithm *The Scientific World Journal*
[30] Katsavounidis I, Kuo J C-C, Zhang Z 1994 *Signal Processing Letters IEEE* **1**(10) 144-6
[31] Kogan J, et al. 2006 Grouping multidimensional data *Springer*
[32] Lattin J M, et al. 2003 Analyzing multivariate data *Thomson Brooks/Cole Pacific Grove, CA*
[33] Leuski A 2001 Evaluating document clustering for interactive information retrieval *in Proceedings of the tenth international conference on Information and knowledge management* 33-40
[34] Lloyd S 1982 *Information Theory, IEEE Transactions on* **28**(2) 129-37
[35] Lu J 2008 Hierarchical initialization approach for K-Means clustering *Pattern Recognition Letters* **29** 787-95
[36] MacQueen J 1967 Some methods for classification and analysis of multivariate observations *in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 14
[37] Meilă M 2006 The uniqueness of a good optimum for K-means *in Proceedings of the 23rd international conference on Machine learning* 625-32
[38] Norusis M J 2012 IBM SPSS statistics 19 statistical procedures companion *Prentice Hall*
[39] Onoda T 2012 Careful Seeding Method based on Independent Components Analysis for K-means Clustering *Journal of Emerging Technologies in Web Intelligence* **4** 51-9
[40] Pizzuti C 19999 A divisive initialisation method for clustering algorithms *in Principles of Data Mining and Knowledge Discovery ed Springer* 484-91
[41] Sebestyen G S 1962 Decision-making processes in pattern recognition (*ACM monograph series*)
[42] Steinhaus H 1956 Sur la division des corp materiels en parties *Bull. Acad. Polon. Sci* **1** 801-4
[43] Steinley D 2006 K-means clustering: A half-century synthesis *British Journal of Mathematical and Statistical Psychology* **59** 1-34
[44] Su T, Dy J G 2007 In search of deterministic methods for initializing K-means and Gaussian mixture clustering *Intelligent Data Analysis* **11** 319-38
[45] Thorndike R L 1953 Who belongs in the family? *Psychometrika* **18** 267-76
[46] Timm N H 2002 Applied multivariate analysis *Springer*
[47] Tou J T, Gonzalez R C 1974 Pattern recognition principles
[48] Wang T, Hung W N 2013 Reliable Node Clustering for Mobile Ad Hoc Networks *Journal of Applied Mathematics*
[49] Yazdani D, et al. 2010 A new hybrid approach for data clustering *in Telecommunications (IST), 2010 5th International Symposium on* 914-9

Author

Jinguo Zhao, born in June, 1965, Shaodong, Hunan Province, China

Current position, grades: associate professor in School of Computer and Information Science, Hunan Institute of Technology, China.
University studies: Database.
Scientific interest: semantic web and database.
Publications: 15 papers.