Journal of Theoretical and Applied Information Technology
15th October 2018. Vol.96. No 19

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

DETERMINATION OF INITIAL CENTROID IN K-MEANS USING PCA FACTOR SCORES

[1]BELABED IMANE, [2]TALIBI ALAOUI MOHAMMED, [3]TALIBI ALAOUI YOUSSEF

[1]Faculty of science, Department of Mathematics and Computer Science, Mohammed First University, Morocco

E-mail : [1]belabedimane@gmail.com, [2]talibialaouim@yahoo.fr

ABSTRACT

Clustering is considered as the task of dividing a dataset, such that elements within each subset are similar between them and are dissimilar to elements belonging to other subsets. One of the most commonly and widely used is K-Means clustering because of its simplicity and performance. The initial centroids for clustering are generated randomly before clustering. If the dataset used is large, then the performance of K-Means will be reduced and also the time complexity will be increased. To overcome this problem, this paper focuses on determining the initial cluster centroids for K-Means. For this purpose, PCA factor score initialization is used in this paper. The experimental results show that the proposed technique provides better clustering and also decreases the time complexity.

Keywords: *K-Means, Principal Component Analysis, PCA factor score, centroid initialization*

1. INTRODUCTION

Data analysis methods are essential for analyzing the ever-growing massive quantity of high dimensional data. On one end, cluster analysis attempts to pass through data quickly to gain first order knowledge by partitioning data points into disjoint clusters, such that data points belonging to the same cluster are similar while data points belonging to different clusters are dissimilar. One of the most popular and efficient clustering algorithms is the K-Means algorithm which uses prototypes (centroids) to represent clusters by optimizing the squared error function.

On the other hand, high dimensional data are often transformed into lower dimensional data via the principal component analysis (PCA) where coherent patterns can be detected more clearly. Such unsupervised dimension reduction is used in very broad areas such as meteorology, image processing, genomic analysis, and information retrieval.

Thus, due to increase in both the volume and the variety of data, it is necessary that there should be advancements in methodology to automatically understand process and summarize the data.

The clustering techniques are used widely to deal with this problem. For K-Means clustering, initial parameters such as number of clusters and initial centroids are needed to be provided [9, 10]. When a large dataset [14] is used in clustering, K-Means will misclassify some data and also the time complexity will be greater. To overcome this problem, the initial centroids mentioned should be effective. For this purpose Principal Component Analysis (PCA) is employed [11].

In this work we use Principal Component Analysis (PCA) for K-Means algorithm to solve the main problem in K-Means by choosing the observations with high PCA factor scores to be the initial centroids for K-Means.

2. RELATED WORK

In order to improve the quality of clustering, several attempts have been made to select the initial centroids carefully, so that the K-Means algorithm should not converge at local optima.

A method was proposed in Rauf et al. [1], to calculate the initial centroids which reduces the number of iterations and improves the elapsed time. The algorithm works in two phases. In the first phase, the cluster size is fixed and the output is initial clusters. In the second phase, the cluster sizes vary and the output is finalized clusters.

A systematic method for finding the initial centroids was presented in Abdul et al. [5]. The centroids obtained by this method are consistent with the distribution of data. Hence it produces

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

clusters with better accuracy compared to the original K-Means algorithm.

The initial starting centroids algorithm based on K-Means was proposed by Agha et al. [7]. The algorithm uses a guided random technique. The experimental results show that the algorithm outperformed the traditional random initialization and improved the quality of clustering by a large margin, especially in complex datasets.

A comparative analysis of various clustering methods, with an emphasis on their computational efficiency, was presented in Celebi et al. [8]. Eight commonly used linear time-complexity initialization methods were compared on a large and diverse collection of datasets using various performance criteria. Experimental results were presented using non-parametric statistical tests. It was concluded that popular initialization methods often perform poorly.

A method for finding the initial centroids was also proposed by Yedla [15]. The method works well in the cases where the input data is uniformly distributed. In the cases where the input data is non-uniform, however, this method does not produce good clustering results. This is where most of the data items in each group lie towards the boundary of the group.

A selection method for initial cluster centroids in K-Means clustering was presented in Aldahdooh et al. [13]. It provides a detailed performance assessment of the proposed initialization method over many datasets with different dimensions. The experimental results show that the proposed initialization method is more effective and converges to more accurate clustering results than those of the random initialization method.

The cluster center initialization algorithm for the iterative clustering algorithm was presented in Khan et al. [11]. The algorithm was based on very similar objects form the core of clusters and their cluster membership remains the same. However, the outliers are more susceptible to change in cluster membership. Hence, these similar patterns (which form the core of clusters) aid in finding initial cluster centers. It is also observed that individual attributes provide information in computing initial cluster centers. The CCIA (Cluster Center Initialization Algorithm) generates clusters which may be more than the number of desired clusters. Similar clusters are merged using the density based multiscale data condensation method to get the desired number of clusters. The centre of these clusters has been used as initial clusters for the K-Means clustering algorithm.

The density-based method for initializing the K-Means clustering algorithm was presented in Zhang et al. [12]. The algorithm uses the density of various areas in data space to choose the initial clustering centers. The density is estimated by using a space-partitioning data structure called K-d tree. Several real-world datasets have been used to test the algorithm and to compare with some other algorithms. It is concluded that their algorithm is more effective for discovering clusters than other well-known algorithms.

Centroid K-Means initialization using PCA eigen values was proposed by Muskatim [28]. The method determines the number of clusters in the K-Means clustering by using covariance matrix on PCA. The clusters will be formed as many as existing attributes; it is raised from the amount of matrix covariance PCA. This study has several disadvantages, such as the number of clusters formed was influenced by the number of attributes, because attribute is a form of eigenvector value in PCA. If in K-Means the number of clusters can be formed as many as desired (number of clusters < number of data) then in PCA K-Means they can only form a maximum number of clusters as many as the number of attributes.

3. METHODOLOGY

3.1 K-Means Clustering

The K-Means algorithm takes two input parameters: the dataset of n, objects, and k, the number of clusters to be created. The algorithm partitions the dataset of n objects into k clusters. Cluster similarity is measured by taking the Euclidean distance between objects. In this way, K-Means finds spherical or ball shaped clusters. The mean value of the objects in a cluster can be viewed as the cluster's center of gravity.

The algorithm works in two phases: in the first phase, k initial centroids are selected randomly, one for each cluster; in the second phase, each object of the given input dataset is associated with the cluster which has the nearest centroid. Euclidean distance is commonly used as a measure to determine the distance between the objects and the centroids. When all the objects from the input dataset are assigned to some clusters, the first iteration is completed, and an early grouping is done. At this point, the algorithm starts a new iteration and recalculates the new centroids, as the inclusion of new data may lead to a change in the cluster

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

centroids. The k centroids may change their position in a step by step manner. Eventually, a situation will be reached where the centroids do not move anymore or the data objects do not change their cluster.

Formally, the K-Means clustering algorithm follows the following steps:
Step 1: Choose a number of desired clusters, k.

Step 2: Choose k starting points to be used as initial estimates of the cluster centroids. These are the initial starting values.

Step 3: Examine each point in the dataset and assign it to the cluster whose centroid is nearest to it.

Step 4: When each point is assigned to a cluster, recalculate the new k centroids.

Step 5: Repeat steps 3 and 4 until no point changes its cluster assignment, or until a maximum number of passes through the dataset is performed.

The K-Means is simple and easy to implement and can be used for processing large datasets, but the algorithm has several limitations:

- It is computationally expensive, time complexity being O (nkl), where n is the total number of objects in the dataset, k is the required number of clusters and l is the number of iterations.
- The quality of final clusters heavily depends on the selection of initial centroids. It means the results may be different for multiple runs of the algorithm for the same input data.

3.2 Principal Component Analysis

Principal Component Analysis (PCA) is a dimension reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information held in the large sets.

The first principal component accounts for as much variability in the data as possible, and each succeeding component accounts for as much remaining variability as possible.

3.2.1 Principal component

Technically, a principal component (PC) can be defined as a linear combination of optimally weighted observed variables which maximize the variance of the linear combination and which have zero covariance with the previous PCs. The first component extracted in a principal component analysis accounts for a maximal amount of total variance in the observed variables. The second component extracted will account for a maximal amount of variance in the dataset that was not accounted for by the first component and it will be uncorrelated with the first component. The remaining components that are extracted in the analysis display the same two characteristics: each component accounts for a maximal amount of variance in the observed variables that was not accounted for by the preceding components; and each is uncorrelated with all of the preceding components.

When the principal component analysis completes, the resulting components will display varying degrees of correlation with the observed variables, but are completely uncorrelated with one another. PCs are calculated using the Eigen value decomposition of a data covariance matrix/ correlation matrix or singular value decomposition of a data matrix, usually after mean-centering the data for each attribute.

3.2.2 Using factor score

Factor scores are composite variables that provide information about an individual's placement on the factor(s).

The extraction method used in this work aims to maximize validity by producing factor scores that are highly correlated with a given factor and to obtain unbiased estimates of the true factor scores.

In this paper, regression factor score extraction is used; it predicts the location of each individual on the factor. Under this process, the computed factor scores are standardized to a mean of zero; however, the standard deviation of the distribution of factor scores (by factor) will be 1.

3.4 Proposed Methodology

An approach to systematically selecting the initial centroids has been proposed here. The centroids are determined following a systematic method, so that different runs of the algorithm on the same dataset produce the same and good quality results.

In fact, the high dimensional data are often transformed into lower dimensional data via the principal component analysis (PCA) where coherent patterns can be detected more clearly. It is also common that PCA is used to project data to a lower dimensional subspace, using the factors that are more meaningful. Consequently, a subject who

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

has high scores on saturated items explains more the factor, consequently this subject is chosen to be the initial centroid for K-Means.

Algorithm: K-Means with PCA factor score initialization.

Input: k: number of clusters, D: a dataset with n objects.

Output: k clusters.

1. Method of choosing initial centroids (Figure 1):

1-1 Apply PCA algorithm.

1-2 Save PCA factor score as a variable using regression method.

1-3 for each factor starting with the one which has the maximum variance.

1-3-1 Choose the object with the high factorial score in ultimate value as initial centroid.

Repeat 1-3 and 1-3-1 until arrived to the number of desired cluster k.

2. Re-assign each object to the cluster to which it is most similar, based on the mean value of the objects in the cluster.

3. Update the cluster means.

4. Repeat until no change.

3.5 Validity Indexes for Clustering

The real usefulness of a cluster partition is hard to measure and depends on its specific application. However, a cluster should meet several common criteria. The criteria relevant to our clustering task are:

- Compactness: The groupings must be homogeneous within and heterogeneous between each other;
- Differentiable: The groupings must be distinguishable conceptually, and respond differently to different potential programs;
- Substantial: The groupings are large enough for a particular program to benefit from;
- Stable: The groupings should be stable over time to be worth designing dedicated programs for;
- Actionable: Actions / programs can be designed to work effectively on each group.

These criteria are qualitative and more or less subjective. To quantify these criteria, various validity indexes are commonly used.

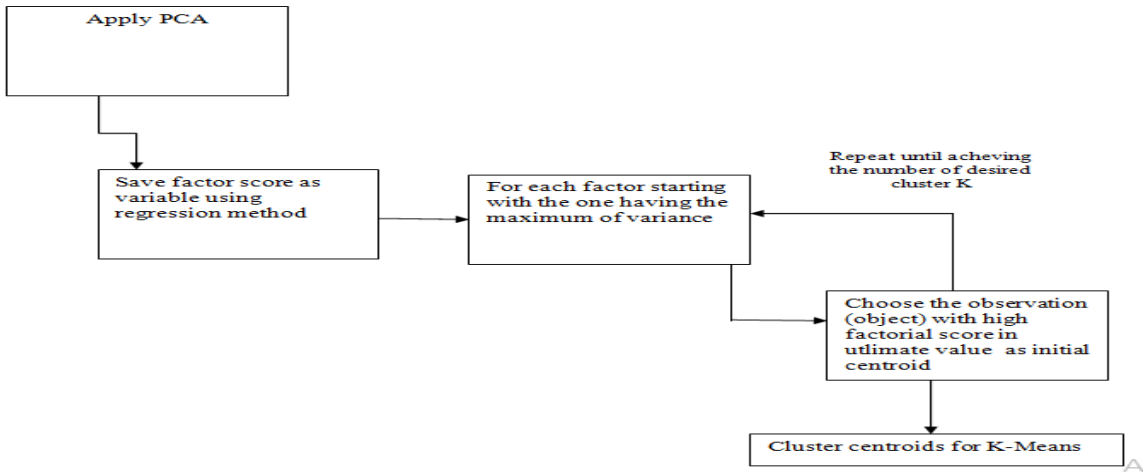

Figure 1: The Proposed methodology initialization for K-Means.

In the following section, the methods used to find the optimal number of clusters is illustrated, such as the Dunn index, the Silhouette index and the Davies-Bouldin index (DBI).

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

These methods are used to see if an algorithm produces better clustering data compared to other clustering methods.

3.5.1 Silhouette index

Silhouette analysis can be used to study the separation distance between the resulting clusters. The Silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters, such as number of clusters, visually. This measure has a range of [-1, 1].

Silhouette coefficients [16] [17] (as these values are referred to) near to +1 indicate that the sample is far away from the neighboring clusters. A value of zero indicates that the sample is very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

3.5.2 Dunn index

The value of the Dunn index (DI) [18] is expected to be large if clusters of the dataset are well separated. If the dataset has compact and well-separated clusters, the distance between the clusters is expected to be large and the diameter of the clusters is expected to be smaller. The clusters are compact and well separated by maximizing the inter-cluster distance while minimizing the intra-cluster distance. Large value results from the Dunn index indicate compact and well-separated clusters.

3.5.3 Davies-Bouldin index

The Davies Bouldin index DBI measures the average similarity between each cluster and its most similar one. A lower value from the DB Index indicates that clusters are highly compact and well separated, which reflects better clustering. The goal of this index is to achieve minimum within-cluster variance and maximum between-cluster separations. [19]

4. RESULTS AND ANALYSIS

The experiment is conducted on datasets from the web for evaluating the proposed clustering initialization of K-Means. The dataset contains 17 variables and 846 observations.

Table 1: Datasets

	V1	V2	V3	V4	V5	V6...	V17
ind1	58	105	183	51	6	265..	183
ind2	58	106	180	51	6	261..	182
ind3	57	109	194	56	6	260..	183
ind4	57	102	181	52	6	257..	184
ind 5	57	106	177	51	5	256..	181
ind6 . .	57 . .	106 . .	172 . .	50 . .	6 . .	255.. . .	183 . .
ind 846	53	101	238	72	4	238	28

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

The data cleaning process is done by performing normalization with Min-Max normalization. The purpose of data normalization is to get the same weight from all of the data attributes without variation or to ensure the result from weighting does not consist of more dominant attributes or is considered more important than the others [20]. Min-max normalization performs a linear transformation on the data, by using a minimum value and a maximum value. Min-max normalization maintains the relationship between the values of the original data [21].

The experiment is devised in two parts: In the first, the tests are done with 6 variables and 200 observations; in the second, in order to evaluate the impact of size effect on the proposed method, the number of variables and the observations is doubled.

According to the original purpose of how to find, the best algorithm for grouping data based on the result of cluster validity. The proposed method is tested with k=3, k=4 and k=5, and also evaluated by seeing the iterations made for finding the final centroids and by calculating the cluster validity indexes.

The first part is illustrated below (6 variables and 200 observations).

4.1 The Result with 6 Variables and 200 Observations

4.1.1 Time complexity

Table 2: Resulted Iteration

The iterations made for convergence	k=3	k=4	k=5
Random centroids	8	7	14
Proposed method	6	6	8

For each number of cluster k, the experiment is conducted 10 times for different sets of values of the initial centroids, which are selected randomly.

In each experiment, the iteration is computed and the average iteration of all experiments is taken. In this step, determining initial centroids with the proposed method and has no considerable gain in terms of complexity (Table 2).

4.1.2 Validity indexes result

Table 3: Silhouette index

	K=3	K=4	K=5
Random centroids	Average=0,4 Mininum=-0,019 Maximum=0,60	Average=0,41 Mininum =-0,03 Maximum=0,65	Average=0,33 Mininum =-0,034 Maximum=0,63
Proposed method	Average=0,4 Mininum =0,098 Maximum=0,61	Average=0,41 Mininum =0,003 Maximum=0,64	Average=0,38 Mininum =-0,047 Maximum=0,66

For the Silhouette index (Table 3) the number of optimal clusters opted for is 4 for both methods. We can deduce that the Silhouette index, in this step, is not significant since the average is less than 0.5 and there are negative values with the exception of k=4 using the proposed method

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

Table 4: Davies-Bouldin index

	K=3	K=4	K=5
Random centroids	1,03	0,93	0,97
Proposed method	1,01	0,94	0,95

For Davies-Bouldin index (Table 4) the number of optimal clusters opted is 4 for both methods, since the lower value is the best.

Table 5: Dunn Index

	K=3	K=4	K=5
Random centroids	0,096	0,073	0,090
Proposed method	0,096	0,097	0,097

Concerning the proposed initialization method (Table 5), the same value was obtained for k=4 and k=5 using Dunn index (the high value is the better). In this case we have opted for the common value with the other indices which is 4. On the other hand, the proposed value for random K-Means is 3.

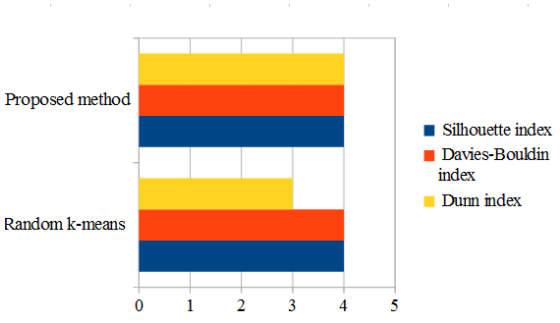

Figure 2: The number of clusters suggested by the validity indexes

From the figure above (Figure 2), in terms of evaluating the cluster validity indexes, we notice that with the proposed method, both the DB Index and the Dunn index specified that the number of optimal clusters should be 4. On the other hand, with random K-Means initialization there is a divergence, when the DB Index suggests that the number of optimal clusters should be 4, the Dunn index suggests that the number of optimal clusters should be 3.

4.2 The Result With 17 Variables And 846 Observation

4.2.1 Time complexity

Table 6: Resulted Iteration

K=3, K=4, K=5	The iterations made for convergence
Random centroids	15 iterations
Proposed method	6 iterations

Also in this step, the experiment is conducted 10 times for different sets of values of the initial centroids, which are selected randomly (Table 6). In each experiment, the iteration was computed and the average iteration of all experiments was taken.

In this step, the number of variables and observations was doubled in order to evaluate the efficiency of the proposed method with a large dataset.

The determination of initial centroids with the PCA factor score method gives it benefits, since the number of iterations made for convergence decreased from 15 iterations to 6 iterations, which is considerable given the volume of data.

4.2.2 Validity indexes result

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

Table 7: Silhouette Index

	K=3	K=4	K=5
Random centroids	Average=0,54 Mininum=-0,04 Maximum=0,71	Average=0,50 Mininum=-0,04 Maximum=0,68	Average=0,5 Mininum=-0,043 Maximum=0,76
Proposed method	Average=0,50 Mininum=0,018 Maximum=0,75	Average=0,51 mininum=0,006 Maximum=0,76	Average=0,5 Mininum=-0,05 Maximum=0,73

For the Silhouette index (Table 7) the number of optimal clusters opted for is 3 for random K-Means and 4 for the proposed method. We deduce also from the above table, that with random K-Means there are negative values of Silhouette index which indicates that there are objects assigned to the wrong cluster, on the other hand we find that with our method there are no negative values with k=3 and k=4.

Table 8: Davies-Bouldin index

	K=3	K=4	K=5
Random centroids	0,54	0,87	0,73
Proposed method	0,79	0,73	1,37

Concerning the Davies-Bouldin index (Table 8) the number of optimal clusters opted is 3 using random K-Means and 4 with the PCA factor score initialization.

Table 9: Dunn index

	K=3	K=4	K=5
Random centroids	0,03	0,03	0,04
Proposed method	0,03	0,04	0,04

Concerning the proposed initialization method (Table 9), the same value was obtained for k=4 and k=5. In this case, we have opted for the common value with the other indices, which is 4. On the other hand, the proposed value for random K-Means is 5.

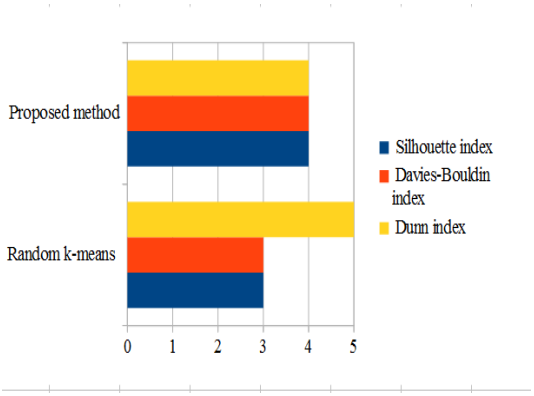

Figure 3: The number of clusters suggested by the validity indexes

From Figure 3, we can deduce that with our method, the Davies-Bouldin index, the Dunn index and the Silhouette index specified that the number of optimal clusters should be 4. However, with random K-Means there is a divergence; the Davies-Bouldin index suggests that the number of optimal clusters should be 3 and the Dunn index suggests that the number of optimal clusters should be 5.

From all results specified above, we can conclude that, in spite of doubling the number of variables and observations, there is no change in terms of the cluster validity index with our method. The entire index suggests that the number of

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

clusters should be 4. Moreover, the number of iterations is always lower than random K-Means.

5 LIMITATIONS OF THIS METHOD

The proposed method of initialization of K-Means centroids is easy to apply and it proves to be an efficient method to determine the initial centroids, which can be used in the K-Means clustering algorithm. Besides solving the problem of non-unique results, our proposed method can also be applicable to large datasets, since the PCA is used in the first step of the method, so it allows the reduction of the data dimension.

Another major advantage of our algorithm over the original K-Means algorithm is that once the initial centroids are systematically determined, the number of iteration required reaching the convergence criteria is reduced largely.

The K-Means algorithm can be applied on numerical data only. But in day to day life, scenarios with a combination of both numerical and categorical data values are encountered, which is the principal limitation of the proposed work.

6 DISCUSSIONS

The main reference of this study was a research by Mustakim about centroid k-Means clustering optimization using eigenvector principal component analysis [28]. In this study to determine the initial centroid in K-Means, the method determines the number of clusters by using covariance matrix on PCA. The clusters which will be formed are as many as existing attributes; it is raised from the amount of matrix covariance PCA. However, eigen vector PCA affects the formation of clusters in K-Means, so PCA K-Means can only form clusters as many as attributes used in the clustering process. However in this article, the proposed method overcomes this limitations and allows forming as many as desired number of clusters, this with a good values of validity indexes.

7 CONCLUSION AND PERSPECTIVES

Since the results of the K-Means clustering method are highly dependent on the selection of initial centroids, there should be a systematic method to determine the initial centroids, which make the K-Means algorithm, converge in global optima and unique clustering result in less iteration. An overview of the existing methods along with new methods to select the initial centroid points for the K-Means algorithm has been proposed in the paper to overcome the deficiency of the traditional K-Means clustering algorithm. The proposed method uses a systematic and efficient way to find initial centroid points based on PCA factor scores initizalization, which produce better clustering in less iteration compared to the traditional K-Means algorithm.

As a perspective, the future work would be carried out in the direction of making the initialization method proposed in this paper applicable to mixed data, that is to say numerical and categorical data.

REFERENCES:

[1] S.A. Rauf, S. Mahfooz, S. Khusro, H. Javed, Enhanced k-mean clustering algorithm to reduce number of iterations and time complexity. Middle East J. Sci. Res. 12(7), 959–963, 2012

[2] Rao, S.G. 2015. “Performance Validation of the Modified KMeans Clustering Algorithm Clusters Data”. International Journal of Scientific & Engineering Research. 6(10). pp. 726-730.

[3] Baarsch, J. and Celebi, M.E. 2012. “Investigation of Internal Validity Measures for K-Means Clustering”. Proceedings of the International Multi Conference of Engineers and Computer Scientiss 2012. 1 March. pp. 16.

[5] K.A.A. Nazeer, M.P. Sebastian, Improving the accuracy and efficiency of the k-Means clustering algorithm, in Proceedings of the World Congress on Engineering, vol. 1, ISBN: 978-988- 17012-5-1, 2009

[6] K. Arai, A.R. Barakbah, Hierarchical k-means: an algorithm for centroids initialization for k-means. Rep. Fac. Sci. Eng. Saga Univ. 36(1), 25–31, 2007

[7]Md.E. Agha, W.M. Ashour, Efficient and fast initialization algorithm for k-means clustering. Int. J. Intell. Syst. Appl. 1, 21–31, 2012

[8] M.E. Celebi, H.A. Kingravi, P.A. Vela, A comparative study of efficient initialization methods for the k-means clustering algorithm. J. Exp. Syst. Appl. 40(1), 200–210, 2013

[9] Donghai Guan, Weiwei Yuan, Young-Koo Lee, Andrey Gavrilov and Sungyoung Lee, "Combining Multi-layer Perceptron and K-Means for Data Clustering with Background Knowledge",Advanced Intelligent Computing Theories and Applications, Springer-Verlag, Vol. 2, Pp. 1220-1226, 2007.

[10] Chen Zhang and Shixiong Xia, "K-means Clustering Algorithm with Improved Initial Center", Second International Workshop on Knowledge Discovery and Data Mining, Pp. 790-792, 2009.

[11] S.S. Khan, A. Ahmad, Cluster center initialization algorithm for k-means clustering. Elesvier J. Pattern Recogn. Lett. 25, 1293–1302, 2004

[12] X. Zhang, Q. Shen, H. Gao, Z. Zhao, S. Ci, A density-based method for initializing the k-means clustering algorithm, in Proceedings of International Conference on Network and Computational Intelligence (ICNCI 2012), IPCSIT, vol. 46, pp. 46–53, 2012

[13] R.T. Aldahdooh, W. Ashour, DIMK means: distance based initialization method for k-means clustering algorithm. Int. J. Intell. Syst. Appl. 5(2), 41–51, 2013

[14] Ismail M. and Kamal M.: Multidimensional data clustering utilization hybrid search strategies,Pattern Recognition Vol. 22(1),PP. 75-89, 1989.

[15] M. Yedla, S. Rao Pathakota, T.M. Srinivasa, Enhancing k-means clustering algorithm with improved initial center. IJCSIT 1(2), 121–125, 2010

[16] Dunn, J.C., ―A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters,‖ J. Cybernetics, vol. 3, 1973, (pg. 32-57).

[17] Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J., Understanding of Internal Clustering Validation Measures‖, Proceeding ICDM '10 Proceedings of the 2010 IEEE International Conference on Data Mining ISBN: 978-0-7695-4256-0 (Pg. 911-916).

[18] Davies, D.L. and Bouldin, D.W., ―A Cluster Separation Measure‖, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224–227.

[19] Kovács, F., Legány, C. and Babos, A., ― Cluster Validity Measurement Techniques‖, AIKED'06 Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (2006), ISBN: 111-2222-33-9 (pg. 388-393).

[20] Xu, Q., Ding, C., Liu, J., and Luo, B. 2015. "PCA-guided search for *K*-means". Pattern Recognition Letters 54. Pp.50–55.

[21] Jain, Y.K., and Bhandare, S.K. 2011. "Min Max Normalization Based Data Perturbation Method for Privacy Protection". International Journal of Computer and Communication Technology. 2(8).

[22] Celebi and Emre, M. 2012. "Deterministic Initialization of the K-Means Algorithm Using Hierarchical Clustering". International Journal of Pattern Recognition and Artifcial Intelligence 26(7). pp. 55-61.

[23] Patel, V.R., and Mehta, R.G. 2011. "Impact of Outlier Removal and Normalization Approach in Modified K-Means Clustering Algorithm". International Journal of Computer Science Issues (IJCSI). 8(5).

[24] Vijay, K., and Selvakumar, K. 2015. "Brain FMRI Clustering Using Interaction KMeansAlgorithm with PCA". International Conference on Communication and Signal Processing (ICCSP). 4(1). pp. 909-913.

[25] S.A. Rauf, S. Mahfooz, S. Khusro, H. Javed, Enhanced k-meanclustering algorithm to reduce number of iterations and timecomplexity. Middle East J. Sci. Res. 12(7), 959–963, 2012.

[26] C.S. Li, Cluster center initialization methof for k-means algorithm over data sets with two clusters, in Proceedings of International Conference on Advances in Engineering, pp. 324-328, 2011

[27] Grice, J. W. (2001). Computing and evaluating factor scores. Psychological Methods, 6(4), 430-450, 2001.

[28] Muskati, centroid K-Means clustering optimization using eigenvector principal component analysis. Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645, 15th August 2017. Vol.95. No.15.

[29] Rendón, E.. Abundez, I., Arizmendi, A., and Quiroz, E.M. 2011. "Internal versus External cluster validation indexes". International Journal of Computers and Communications. 1(5). pp. 27-34.

[30] Sharma, S., Gupta, P., Parnami, P. 2015. "An Approach for Parallel K-means based on Dunn's Index". International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE). 5(5). pp. 1665-1668.