IMPROVING THE EFFICIENCY OF CLUSTERING BY USING AN ENHANCED CLUSTERING METHODOLOGY

Bikram Keshari Mishra[1], Nihar Ranjan Nayak[2], Amiya Kumar Rath[3], Sagarika Swain[4]
[1&2]Deptt. of Computer Sc. & Engg., Silicon Inst. Of Technology, Bhubaneswar, India
[3]Deptt. of Computer Sc. & Engg., DRIEMS, Cuttack, India
[4]Deptt. of Computer Sc. & Engg., Koustav Inst. Of Self Domain, Bhubaneswar, India

ABSTRACT

Clustering in data analysis means data with similar features are grouped together within a particular valid cluster. Each cluster consists of data that are more similar among themselves and dissimilar to data of other clusters. Clustering can be viewed as an unsupervised learning concept from machine learning perspective. In this paper, we have proposed an Enhanced Clustering Methodology to obtain better clustering quality with much reduced complexity. We have evaluated the performances of the classical K-Means approach of data clustering, it's modified Global K-Means, an Efficient K-Means and the proposed Enhanced K-Means method. The accuracy of all these algorithms were examined taking several data sets from UCI [21] repository of machine learning databases. Their clustering efficiency has been compared in conjunction with two typical cluster validity indices, namely the Davies-Bouldin Index and the Dunn's Index for different number of clusters, and our experimental results demonstrated that the quality of clustering by proposed method is much proficient than the other mentioned K-Means based algorithms when larger data sets with more number of attributes are taken into consideration. Apart from this it has been found that, the computational time for clustering determined by the proposed algorithm is much lower than the other discussed methods.

Keywords: *Cluster analysis, Cluster validity indices, K-Means, Global K-means, Efficient K-Means, and Enhanced K-Means clustering.*

I. INTRODUCTION

Retrieving information faster from a group has always been an important issue. Several approaches have been developed for this purpose, one of them is data clustering. Therefore much attention is now paid to invent new fast and improved clustering algorithms. The main goal of clustering is that, the objects present in a group will be much similar to one another and different from the objects present in other groups.

In order to elaborate the concept of clustering, let us take a simple example of the library management system. In a library several books concerning to a large variety of topics are available. They are more or less kept in form of clusters. The books that have some kind of similarities among them are placed in one cluster i.e. books relating to computer architecture are kept in one shelf and books on algorithm analysis are kept in another shelf, and so on and finally, the shelves are labelled with their relative names. Now, when a user searches for a book of specific kind, he would only have to proceed to that particular shelf and check for the book instead of searching in the entire library.

The definition of what constitutes a cluster is always not well defined, and in most applications clusters are not well separated from each other hence, most clustering techniques represent a result as a classification of the data into non-overlapping groups. Clustering is often confused with

classification, but there are some differences between the two. In classification, the objects are assigned to some already pre-defined class, whereas in clustering the classes are to be defined.
Learning valuable information from huge volume of data makes the clustering techniques widely applicable in several domains including artificial intelligence, data compression, data mining and knowledge discovery, information retrieval, pattern recognition and pattern classification , and so on.
In this paper, we have compared the basic K-Means based methods for data clustering and have introduced a new clustering algorithm which is far more effective than the classical K-Means [19] algorithm, it's modified Global K-Means [12] algorithm and the revised Efficient K-Means [13] algorithm. We have implemented these algorithms on various data sets with varied sizes. The major downsides of K-Means, Global K-Means and Efficient K-Means are discussed and in order to curtail such difficulties and improve the clustering quality and efficiency, we have proposed a simple model known as Enhanced Clustering Algorithm. We have checked the check the quality of clustering results by using Dunn's separation index (DI) [8] and Davies-Bouldin's index (DBI) [7] respectively on the given algorithms and have also recorded the amount of computational time taken by each of them.
This paper is organized as follows: In Section II we briefly present the basic idea of cluster validity measures and two widely used validity indices such as DI and DBI used for determining the quality of results obtained from clustering. Section III presents the efficient and productive works done by several researchers in this relevant area. Different K-Means based data clustering methods are their effectiveness and downsides are mentioned in Section IV and our proposed enhanced data clustering method is mentioned in Section V. Simulation and experimental results are shown in Section VI. Finally, Section VII concludes the paper.

II. CLUSTERING VALIDATION

Clustering analysis is a task of assigning a set of related objects into their respective cluster. There is no specific recommended algorithm for clustering analysis, it can be achieved by various algorithms that differ significantly in the way in which they group the similar objects efficiently into their desired cluster. In fact, if cluster analysis is to make a significant contribution to any relevant application area, then much attention must be paid to cluster a validity issue which is normally concerned with determining the optimal number of clusters and checking the quality of clustering results. Evaluation of clustering results is generally referred to as cluster validation. Cluster validity issue by and large concerned with determining the optimal number of clusters and checking the fineness of clustering results. Many different indices of cluster validity have been already proposed. In this section, we discuss briefly the Dunn's separation Index and Davies-Bouldin's Index which we have used in our proposed clustering algorithm for examining the soundness of clusters.

2.1. Dunn's Index

The main goal of Dunn's index (DI) measure [8] is to maximize the inter-cluster distances and minimize the intra-cluster distances. Dunn's index is defined as:-

$$DI(c) = \min_{i \in c} \left\{ \min_{j \in c, j \neq i} \left\{ \frac{\delta(A_i, A_j)}{\max_{k \in c} \{\Delta(A_k)\}} \right\} \right\} \qquad \text{.....} \quad (1)$$

where,

$$\delta(A_i, A_j) = \min \left\{ d(\underline{x}_i, \underline{x}_j) \middle| \underline{x}_i \in A_i, \underline{x}_j \in A_j \right\}$$

$$\Delta(A_k) = \max \left\{ d(\underline{x}_i, \underline{x}_j) \middle| \underline{x}_i, \underline{x}_j \in A_i \right\}$$

d is a distance function, and A_j is the set whose elements are the data points assigned to the ith cluster. The number of cluster that maximizes DI is taken as the optimal number of the clusters.

2.2 Davies-Bouldin's Index

Another measure, the Davies-Bouldin's index (DBI) [7] is a function of the ratio of the sum of within-cluster distribution to between-cluster separation.

The within i^{th} cluster distribution is defined as:-

$$S_{i,q} = \left(\frac{1}{|A_i|} \sum_{\underline{x} \in A_i} \left\| \underline{x} - \underline{v}_i \right\|_2^q \right)^{1/q} \qquad \text{.....} \quad \textbf{(2)}$$

The between i^{th} and j^{th} separation is given by:-

$$d_{ij,t} = \left\{ \sum_{s=1}^{p} \left| v_{si} - v_{sj} \right|^t \right\}^{1/t} = \left\| \underline{v}_i - \underline{v}_j \right\|_t \qquad \text{.....} \quad \textbf{(3)}$$

where, $\underline{v}_i$ is the ith cluster centre, and $(q, t) \geq 1$, and both q & t are integers and can be selected independently of each other. $|A_i|$ is the number of elements in A_i.

Next, from equation **(2)** and **(3)** we define $R_{i,qt}$ as:-

$$R_{i,qt} = \max_{j \in c, j \neq i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\} \qquad \text{.....} \quad \textbf{(4)}$$

Finally, Davies-Bouldin's index is given by:-

$$DB(c) = \frac{1}{c} \sum_{i=1}^{c} R_{i,qt} \qquad \text{.....} \quad \textbf{(5)}$$

The objective is to minimize the DBI for achieving proper clustering.

The appropriate clustering algorithm and parameter settings heavily depend on the input data set taken into consideration. An ideal cluster can be said to be a set of data points that is more isolated and compact from other data points.

III. RELATED WORKS

K-Means clustering as proposed by [1] can be used for the feature set obtained using the histogram refinement method which is based on the concept of coherency and incoherency. A non-metric distance measure for similarity estimation based on the characteristic of differences [2] is presented and implemented on K-Means clustering algorithm. The performance of this kind of distance and the Euclidean and Manhattan distances were then compared. A new line symmetry based classifier (LSC) [3] deals with pattern classification problems. LSC is well-suited for classifying data sets having symmetrical classes, irrespective of any convexity, overlap and size. A modified version of the K-Means algorithm for data clustering [4] adopts a novel non-metric distance measure based on the idea of point symmetry. This kind of distance can be applied in data clustering and human face detection. The shortcomings of the standard K-Means clustering algorithm can be found in the literature [5] in which a simple and efficient way for assigning data points to clusters is proposed. Their improved algorithm reduces the execution time of K-Means algorithm to a great extends. A system for analyzing students results based on cluster analysis and using standard statistical algorithms to arrange their scores data according to the level of their performance is described in [6]. An incremental approach to K-means clustering method that adds one cluster centre at a time through a search procedure is given in Global K-Means algorithm proposed by A. Likas et al. in [12]. A simple and efficient implementation of K-Means clustering algorithm called the filtering algorithm given by [13] shows that the algorithm runs faster as the separation between clusters increases. The various types of clustering algorithms along with their applications in some benchmark data sets were surveyed in

[14]. Here, several proximity measures, cluster validation and various tightly related topics were discussed. A new generalized version of the conventional K-Means clustering algorithm which performs correct clustering without pre-assigning the exact cluster number can be found in [15]. Based on the definition of nearest neighbour pair C.S. Li et al. in [16] proposed a new cluster centre initialization method for K-Means algorithm. In iterative clustering algorithms, selection of initial cluster centres is extremely important as it has a direct impact on the formation of final clusters. An algorithm to compute the initial cluster centres for K-Means algorithm was given by M. Erisoglu et al. in [17] and their newly proposed method has good performance to obtain the initial cluster centres converges to better clustering results and almost all clusters have some data in it.

IV. DIFFERENT DATA CLUSTERING TECHNIQUES

In this section, we have briefly focused on the different approaches of K-Means based clustering algorithms widely used for clustering datasets of diverse characteristics, occurrences and number of classes. The different kind of clustering methods are as follows:

4.1. K-Means Clustering Algorithm

The K-Means Clustering algorithm is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. In 1967, Mac Queen [19] firstly proposed the K-Means algorithm. During every pass of the algorithm, each data is assigned to the nearest partition based upon some similarity parameter (such as Euclidean distance measure). After the completion of every successive pass, a data may switch partitions, thereby altering the values of the original partitions.
Various steps of the standard K-Means clustering algorithm is as follows: -

(1) The number of clusters is first initialized and accordingly the initial cluster centres are randomly selected.
(2) A new partition is then generated by assigning each data to the cluster that has the closest centroid.
(3) When all objects have been assigned, the positions of the K centroids are recalculated.
(4) Steps 2 and 3 are repeated until the centroids no longer move any cluster.

The main objective of K-Means is the minimization of an objective function that determines the closeness between the data and the cluster centers, and is calculated as follows:

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} \left\| d(X_i, C_j) \right\| \qquad \text{.....} \quad \textbf{(6)}$$

where, $\left\| d(X_i, C_j) \right\|$ is the distance between the data X_i and the cluster centre C_j.

But the major drawback is that, the result of K-Means strongly depends on the initial selection of centroids and it is very difficult to compare the quality of the clusters produced as very far data from the centroid may pull the centroid away from the real one.

4.2. Global K-Means Clustering Algorithm

The Global K-means algorithm was first proposed by [12], is an incremental approach to classical K-means clustering method that adds one cluster centre at a time through a search procedure consisting of N (where N is the size of the data set) executions of the K-means algorithm from a suitably determined initial position. The main purpose of Global K-Means method is that, instead of finding all cluster centres at once, it proceeds in an incremental fashion by adding one cluster centre at a time. Using Global K-Means, for solving a K-cluster problem, this algorithm begins by solving a one-cluster problem first. In this case, a cluster is formed with the centre being the centroid of all the data present in the dataset. After knowing the first cluster centre, the next step is to add a new second cluster at its optimal position. This is done by running the K-Means algorithm for N number of times with the first centre being already obtained in the one-cluster problem and the second cluster's starting position will be each individual data X_i in the data set where, $1 \leq i \leq N$. The final solution to this two-cluster problem will be the best solution obtained from the N-execution of K-Means algorithm. The above procedure is repeated, and one cluster centre is incrementally added at a time. Hence, the

solution of a K-clustering problem is obtained from the solution of a (K-1) clustering problem, once the newly determined centre is placed at an appropriate optimal position within the data set. In this paper, we have considered several parameters in choosing the best solution, after executing K-Means algorithm for N times.

4.2.1. Pseudo-code:

We now discuss the informal high-level description using the structural conventions of programming language of Global K-means algorithm. The pseudo-code for solving a K-clustering problem with N number of data in the data set is as follows:

```
for i = 1 to K
{
    if the value of i is 1, then a cluster is formed with the centre C_i being the centroid of all
       the data present in the given data set.
    else
        for j = 1 to N
        {
            Run K-Means algorithm with initial values of
                { { C_i , C_i-1 , C_i-2 , … …, C_i-N-2 }, {X_j} }
        }
}
```

Experimental results indicate that the performance of this method is excellent in grouping similar objects in their respective groups. But, the only downside of this method is that, the computational time can be rather too long.

4.3. Efficient K-Means Clustering Algorithm

D. Napeleon and P. Ganga Laxmi [18] proposed a method for making the K-Means algorithm more effective and efficient, so as to get better clustering with reduced complexity using uniform distribution data points. The basic idea of this Efficient K-Means clustering algorithm is outlined as follows:

Initially, the distances between each data and all other data in the data set are computed. Then, the closest pair of data is found from the whole data set and is kept in an intermediate set D_1. These two data are then removed from the whole data set D. The next step is to determine the data which is closest to the set D_1. After finding that, it is added to D_1 and deleted from D. This whole procedure is repeated until the number of elements in the set D_1 reaches a threshold value (0.75 * (N / K)) where, N is the number of data items and K is the number of desired clusters.

Similarly, following the above procedure, we again select another intermediate data set D_2. Repeat this until K such sets of data are obtained. Finally, by averaging all the vectors in each intermediate data set, we obtain the initial centroids.

V. PROPOSED CLUSTERING METHOD

In this proposed method of clustering algorithm, we have slightly modified our algorithm and have used it more effectively which works much proficiently than the already discussed three standard clustering techniques. The algorithm is outlined as follows:-

5.1. An Enhanced Clustering Methodology

The algorithm operates in two phases:

Phase I: Selection of K cluster centres :

(1) Take initial cluster centre c_1 randomly as any data from the input dataset I_p .

(2) Choose the next cluster center $c_i = p'$ with a probability given by:

$$\left(\frac{D(p')^2}{\sum_{p \in I_p} D(p)^2} \right)$$

where, D(p) denotes the shortest distance from any data point p to the already chosen nearest cluster centre [20].

(3) Repeat step (2) until K numbers of cluster centres is chosen.

Phase II: The clustering of dataset :

(4) Compute the Euclidean distance between each data present in the input dataset and all K cluster centres.

(5) Assign each data to its corresponding nearest cluster centre c_i.

(6) (a) Declare two matrices C[ ][ ] and D[ ][ ].

(b) Store the cluster number in which the data p is assigned to in step (5) in the matrix C[ ][ ].

(c) Store the distance of data p to its nearest cluster in matrix D[ ][ ].

(7) Recalculate the cluster centre for each cluster c_i .

(8) Repeat step (9) until convergence is reached.

(9) For each data p, compute the distance to its nearest cluster centre.

(a) If this calculated distance is less than or equal to the previous stored distance in matrix D[ ][ ], then the data remains in the initial cluster.

(b) Else

Find the distance of each data to all the cluster centres and assign the data to that cluster which is nearest to its centre.

(10) End of clustering.

VI. EXPERIMENTAL RESULTS

We examined the performance of the above described algorithms on a number of benchmark data sets taken from the UCI [21] repository of machine learning databases. To assess the efficiency of our method, we compared the results obtained by general K-Means algorithm, Global K-Means algorithm as well as the Efficient K-Means algorithm against the clustering results returned by our proposed Enhanced Clustering algorithm on different data sets varying in their size and characteristics. The initial number of clusters is given by the user during the execution of the program.

The performances of the discussed algorithms as well as our proposed Enhanced Clustering algorithm are measured in terms of two standard validity measures namely Dunn's index (DI) [8] and Davies-Bouldin's index (DBI) [7]. The validity measures tests the quality of clustering by making a comparison between the results obtained from clustering with the information given for that data set. Table 2 gives a comparative analysis of various clustering algorithms by considering DI and DBI measures on various sized datasets.

The graphical representation of the performance analysis of K-Means, Global K-Means, Efficient K-Means and proposed clustering algorithm is shown in Figure 1(a) and 1(b) respectively.

All these algorithms were implemented in MATLAB 7.8.0 on Intel Core 2 Duo system. The processing time of all the above specified data clustering algorithms on various datasets were recorded as can be seen from Table 3.

Table 1 shows some characteristics of the data sets used in this paper.

Table 1: Characteristics of some data sets from UCI repository.

Data set	Number of Attributes	Number of Classes	Number of Records
Iris	4	3	150
Wine	13	3	178

Glass	11	2	214
Abalone	8	3	4177
Mushroom	22	2	8124

Table 2: Comparison of K-Means, Global K-Means, Efficient K-Means and proposed Enhanced Clustering algorithm by considering Dunn's and Davies-Bouldin's indices on different sized data sets.

Algorithms / Data Set	K-Means		Global K-Mean		Efficient K-Mean		Proposed Enhanced Method	
	DI	**DBI**	**DI**	**DBI**	**DI**	**DBI**	**DI**	**DBI**
Iris (k=3)	0.0233	0.6791	0.0168	0.7731	0.0823	0.6791	0.0887	0.6225
Glass (k=2)	0.0415	1.4461	0.0282	1.7242	0.0914	1.4351	0.0938	0.8552
Wine (k=3)	0.0227	0.6880	0.0154	1.4869	0.0477	0.6894	0.0508	0.6458
Abalone (k=3)	0.0019	0.8493	0.0013	1.0627	0.0046	0.8966	0.0057	0.8252
Mushroom (k=2)	0.0349	1.7455	0.0276	1.9884	0.0652	1.7962	0.0729	1.4706

Table 3: Clustering running times of K-Means, Global K-Means, Efficient K-Means and proposed Enhanced Clustering algorithm on different data sets.

Algorithm	**Iris** (k=3)	**Glass** (k=2)	**Wine** (k=3)	**Abalone** (k=3)	**Mushroom** (k=2)
K-Means	0.1406	0.0938	0.2313	4.3750	28.2969
Global K-Means	34.3750	42.6528	38.8603	446.3177	722.0824
Efficient K-Means	0.5000	0.8125	0.4375	18.0156	66.3425
Proposed Enhanced Method	0.0781	0.0625	0.1813	2.2656	12.4063

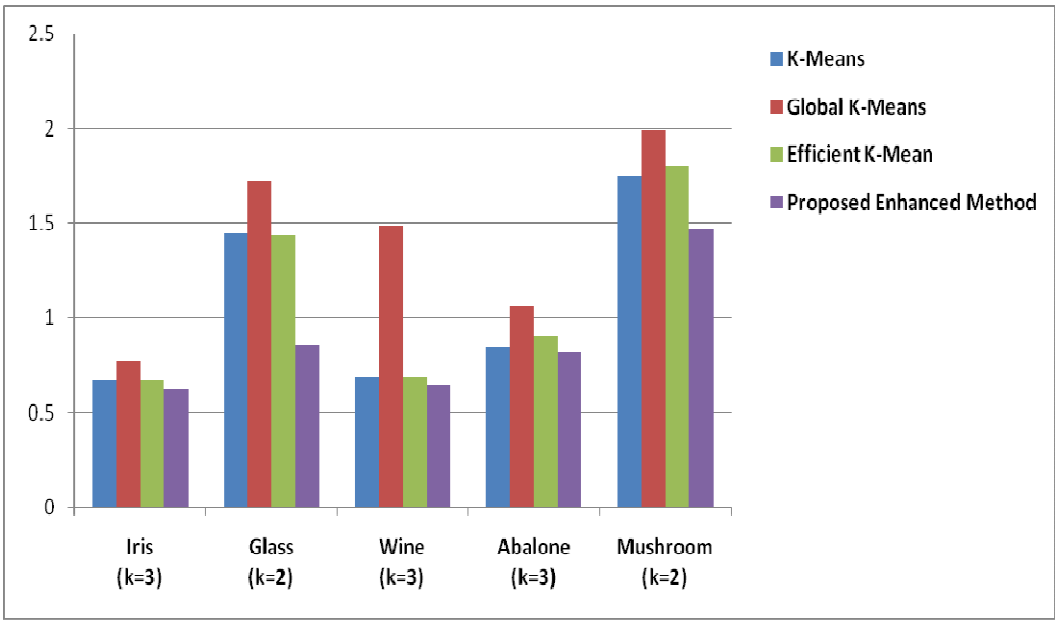

Figure 1(a): Performance analysis of K-Means, Global K-Means, Efficient K-Means and proposed Enhanced Clustering algorithm on several data sets based on Davies-Bouldin's index.

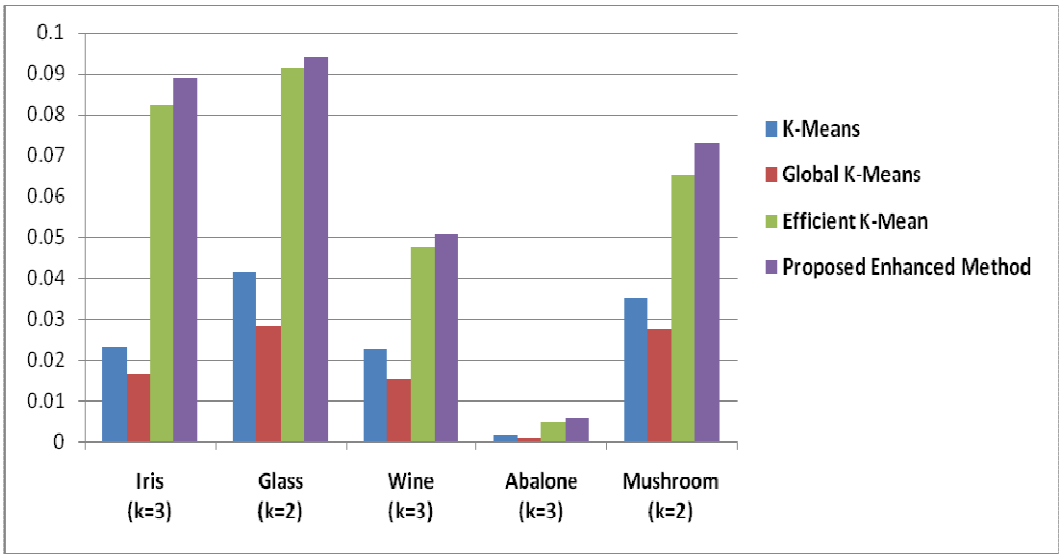

Figure 1(b): Performance analysis of K-Means, Global K-Means, Efficient K-Means and proposed Enhanced Clustering algorithm on several data sets based on Dunn's index.

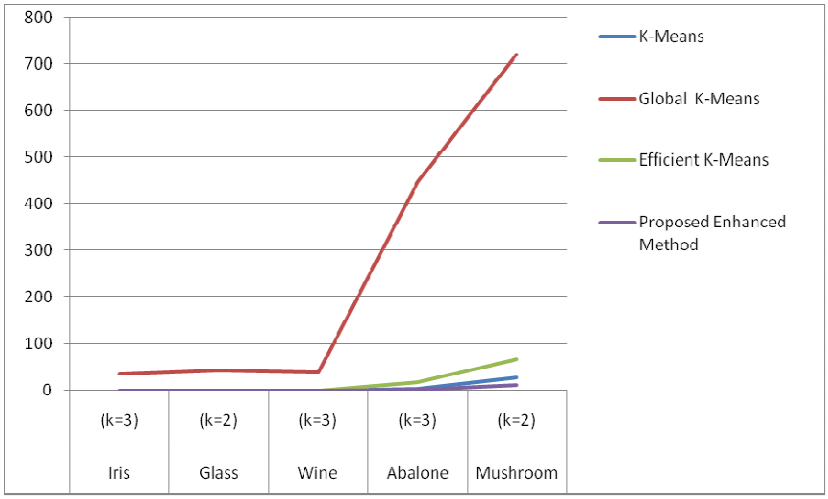

Figure 2: Line chart showing the performance comparison of computational time of K-Means, Global K-Means, Efficient K-Means and proposed Enhanced Clustering algorithm on several data sets.

VII. CONCLUSIONS

In this paper, we have examined a few varieties of imperative clustering algorithms – the customary K-Means algorithm, its modified Global K-Means version, its competent Efficient K-Means algorithm and our projected Enhanced Clustering methodology. It can be seen from the experimental result that, K-Means algorithm can do a pretty good job in clustering data sets in any K numbers of clusters. However, the algorithm strongly depends on the initial selection of centroids and it is very difficult to compare the quality of the clusters produced, as very far data from the centroid may pull the centroid away from the real one. The Global K-means algorithm is an incremental approach to classical K-means clustering method and to a great extent effective in situations where the correct clustering is required but, the main problem with this method is its much stretched execution time. The Efficient K-Means is a revised version of the usual K-Means and works very efficiently and gives suitable result only for uniform distribution of data points. In order to curtail such difficulties and improve the clustering quality and efficiency especially on varied data sets, we have proposed a simple model known as Enhanced Clustering technique. Considering both the DI and DBI parameters for cluster validation on various sized data sets, the results obtained by our proposed algorithm produces better quality of clustering as compared to the discussed K-Means-based algorithms. In addition to this, the

clustering computational time of the proposed algorithm is much lower than the other discussed techniques.

REFERENCES

[1] P. Jeyanthi and V. Kumar, "Image Classification by K-means Clustering", "Advances in Computational Sciences and Technology",vol.3,pp.1–8, 2010.

[2] Z. Li, J. Yuan, H. Yang and Ke Zhang, "K-mean Algorithm with a Distance Based on the Characteristic of Differences", "IEEE International conference on Wireless communications, Networking and mobile computing", pp. 1-4, Oct.2008.

[3] S. Saha, S. Bandyopadhyay and C.Singh, "A New Line Symmetry Distance Based Pattern Classifier", "International joint conference on Neural networks as part of 2008 IEEE WCCI", pp.1426-1433, 2008.

[4] M. C. Su and C. H. Chou, .A modified version of the K-Means algorithm with a distance based on cluster symmetry, IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 674.680, 2001.

[5] Shi Na, L. Xumin, G. Yong, "Research on K-Means clustering algorithm - An Improved K-Means Clustering Algorithm", "IEEE Third International Symposium on Intelligent Information Technology and Security Informatics", pp.63-67, Apr.2010.

[6] O. J. Oyelade, Oladipupo and I. C. Obagbuwa, "Application of K-Means Clustering algorithm for prediction of Students Academic Performance", "International Journal of Computer Science and Information Security(IJCSIS)", pp. 292-295, Vol. 7, No. 1, Feb.2010.

[7] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure", "IEEE Trans. Pattern Analysis and Machine Intelligence",vol.1, pp.224-227, 1979.

[8] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", J. Cybernetics, vol. 3, pp. 32- 57, 1973.

[9] L. Hubert. and J. Schultz., "Quadratic assignment as a general data-analysis strategy", "British Journal of Mathematical and Statistical Psychology", vol.29, pp.190-241, 1976.

[10] P. Rousseeuw, J. Silhouettes, "A graphical aid to the interpretation and validation of cluster analysis", "Journal of Computational and Applied Mathematics", vol.20, pp.53-65, 1987.

[11] J. C. Bezdek, "Numerical Taxonomy with Fuzzy Sets", "J. Math. Biol.", vol.1, pp.57-71, 1974.

[12] A.Likas, N. Vlassis and J. Verbeek, "The global k-means clustering algorithm", "Pattern Recognition Society, Published by Elsevier Science Ltd.", vol.36, issue 2, pp. 451-461, Feb.2003.

[13] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko and A. Wu, "An Efficient K-Means Clustering Algorithm: Analysis and Implementation", "IEEE Transactions on Pattern analysis and Machine Intelligence", vol. 24,no.7, 2002

[14] R. Xu and D. Wunsch, "Survey of Clustering Algorithms", "IEEE Transactions on Neural networks", vol. 16, no. 3, May 2005.

[15] Y.M. Cheung, "A New Generalized K-Means Clustering Algorithm", "Pattern Recognition Letters, Elsevier",vol.24,issue15, 2883–2893, Nov.2003.

[16] C. S. Li, "Cluster Centre Initialization Method for K-means Algorithm Over Data Sets with Two Clusters", "2011 International Conference on Advances in Engineering, Elsevier", pp. 324-328, vol.24, 2011.

[17] M. Erisoglu, N. Calis and S. Sakallioglu, "A new algorithm for initial cluster centres in K-Means algorithm", "Published in Pattern Recognition Letters", vol. 32, issue 14, Oct.2011.

[18] D. Napoleon and P. G. Laxmi, "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points", "IEEE Trendz in Information science and computing", pp.42-45, Feb.2011.

[19] J. Mac Queen, "Some methods for classification and analysis of multivariate observations", "Fifth Berkeley Symposium on Mathematics, Statistics and Probability", pp.281-297, University of California Press, 1967.

[20] David Arthur and Sergei Vassilvitskii : K-Means++: The advantages of careful seeding, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027—1035, 2007.

[21]C.Merz and P.Murphy, UCI Repository of Machine Learning Databases, Available: fttp://ftp.ics.uci.edu/pub/machine-learning-databases.

AUTHOR'S BIOGRAPHIES

Bikram Keshari Mishra is currently pursuing his Ph.D in Computer Science and Engineering from CMJ University, Meghalaya in the field of Data Mining. He has completed his M-Tech in Computer Science and Engineering from Biju Patnaik University of Technology, Odisha, India. He is currently working as a Senior Asst. Professor in the Department of Computer Science & Engineering in Silicon Institute of Technology for the last five years. His research interest includes data mining and image processing.

Nihar Ranjan Nayak is currently doing his Ph.D in Computer Science and Engineering from CMJ University, Meghalaya in the field of Image Processing. He holds a M-Tech degree in Computer Science from Utkal University, Odisha, India. He is currently working as a Senior Asst. Professor in the Department of Information Technology in Silicon Institute of Technology since 2006. His research interest includes data mining and image processing.

Amiya Kumar Rath was awarded Ph.D in Computer Science in the year 2005 from Utkal University for the work in the field of Embedded system. Presently working with Dhaneswar Rath Institute of Engineering & Management Studies (DRIEMS) as Professor of Computer Science & Engg. cum Principal of Degree Wing. Contributed more than 30 research level papers to many national and International journals and conferences. Having research interests include Embedded System, Ad-hoc Network, Sensor Network, Power Minimization, Bi-clustering, Evolutionary Computation and Data Mining.

Sagarika Swain holds a M-Tech degree in Computer Science and Engineering from Biju Patnaik University of Technology, Odisha, India. Her area of interest includes data mining, image processing and fuzzy logic.