A COMPARATIVE ANALYSIS BETWEEN K-MEAN AND Y-MEANS ALGORITHMS IN FISHER'S IRIS DATA SETS.

V.Leela#1, K.Sakthi priya*2,R.Manikandan#3
#1M.tech VLSI Design, Department of Computing, SASTRA university,Thanjavur-613401,India.
Email:leelaeclipse@gmail.com
*2 M.tech VLSI Design, Department of Computing, SASTRA University,Thanjavur-613401,India.
Email:sakthieceb@gmail.com
#3Senior Assistant Professor, Department of ICT,SASTRA University,Thanjavur-613401,India.
Email:manikandan75@core.sastra.edu

***ABSTRACT*:** Cluster analysis plays a vital role in various fields in order to group similar data from the available database. There are various clustering algorithm available in order to cluster the data but the entire algorithm are not suitable for all process .This paper mainly address with the comparative performance analysis of partition based k-mean and y-mean algorithm in Iris flower datasets. The experimental results of iris data set show that the Y-Means algorithm yields the best results in clustering and time complexity compared with k-Mean algorithm in little iteration time.

Keywords - K-Mean Algorithm, Y-Means Algorithm, Cluster Analysis**.**

1. INTRODUCTION

Cluster analysis groups the given data objects based on only information found in the data and describes the objects and their relationships. The objective is that objects within a group be similar to one another and different from the objects in other groups. The data objects which have the maximum similarity within a group and the greater the difference between the groups are, the better or more distinct the clustering. Clustering is an effective technique for exploratory data analysis, and has found applications in a wide variety of areas.

In this paper, we mainly review two algorithms k-means and y-mean algorithm. MostExisting methods of clustering can be categorized into three: partitioning, hierarchical, and grid-based and model-based methods. The k-Means and y-means are examples of partitional methods. The y-mean and k-mean are compared in the data sets of iris flower to cluster the three species of iris flower and the results are obtained in Matlab.

2. METHODOLOGY

Clustering is one of the most widely performed analyses on gene expression data. Every clustering algorithm is based on the index of similarity or dissimilarity between data points. Each cluster is a collection of data objects that are similar to one another are placed within the same cluster but are dissimilar to objects in other clusters.The iris data sets are taken from three different species inorder to classify each species with common data sets. The clustering process for each algorithm differs from in order to classify the similar groups.

2.1. THE K-MEANS ALGORITHM

K-Means is one of the simplest unsupervised learning algorithms used to partition the given data objects in clustering. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The main procedure is to initialize k centroids, one for each cluster groups. These centroids have to be selected carefully since their placement will always affect the end result. Finally, the k-mean algorithm aims at minimizing an objective function in the data objects, in this case a squared error function. The objective function

$$\mathbf{M}=\sum_{j=1}^{k}\sum_{i=1}^{k}\left\|\mathbf{x}_{a}^{(j)}-\mathbf{c}_{b}\right\|^{2}$$

Where ||xa(j)-yb||2 is a chosen distance measure between a data point xa_j and cluster centre cb, is an is an indicator of the distance of the n data points from their respective cluster centers.

So, the better choice is to place them as far as possible. The algorithm is composed of following steps,

(1) Place k points into the space represented by the objects that are being clustered. The k- points represent initial group of centroids.

(2) Assign each object the group that has the closest centroid.

(3) After the assignment of centroids to each objects, recalculate the positions of the k centroids.

(4) Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups of the objects into groups from which the metric to be minimized can be calculated.

The algorithm is drastically sensitive to the initial randomly selected cluster centers. The k-Means algorithm can be run iteratively to minimize this effect.

K-means include:

-simplicity and applicability for a wide variety of data types. It is also quite efficient, even after multiple iterations are often performed.

-It provide best result when the cluster is intensive and the distinction between clusters is obvious.

-It is efficient and scalable when the data set is large.

Weaknesses of K-means include:

-It depends on initial centroid and the final number of clusters and undergoes degeneracy.

-The algorithm is not apposite for nonconvex shape clusters nor cluster sizes that are highly variable.

-A sensitivity to noise points, marginal points and isolated points.

2.2. THE Y-MEAN CLUSTERING

Y-means is based on the K-means algorithm. The main difference between the two is Y-means' ability to autonomously decide the number of clusters based on the statistical nature of the data. This makes the number of final clusters that the algorithm produces a self-defined number rather than a user-defined constant as in the case of K-means. This overcomes one of the main drawbacks of K-means since a user-defined k cannot guarantee a suitable partition of a dataset with an unknown distribution, a random value of initial k usually results in poor clustering.

Y-means can find out an appropriate value of final k (centroids), which is independent of the initial k experiments by using a sequence of splitting, deleting and merging the clusters, even without the knowledge of the distribution of data. To eliminate the effect of dominating features due to the feature-range differences, the dataset is first normalized. Next, the standard K-means algorithm is run over the training data. Due to the fact that the final number of clusters is independent of initial k. Moreover, the selection of the k initial centroids is again independent of the final results. The standard K-means algorithm uses Euclidian distance as a distance function. The Y-means algorithm uses the following function to identify a single outlier per iteration for each cluster. . Let Obc(Bj, Cl) be an Boolean outlier detection function:

$$O_{bc}(B_x, C_y) = \begin{cases} true & pj \text{ is an outlinear to } c_l \\ false & otherwise \end{cases}$$

$\forall x, \forall y: j \in [1, m], l \in [1, n]$

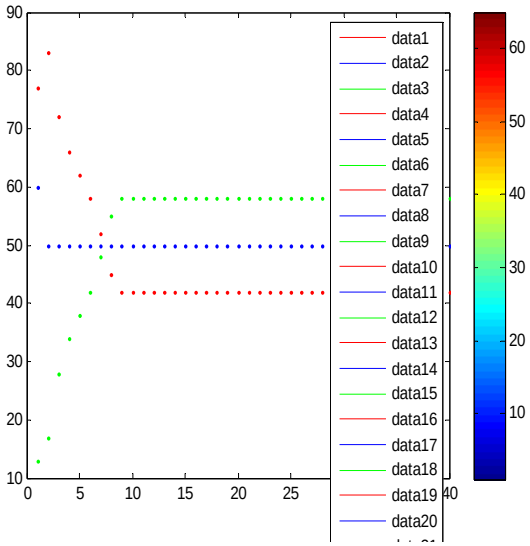
The Y-means algorithm iteratively identifies outliers and converts them to new centroids.

3. EXPERIMENTAL RESULT

Experimental work is done through MATLAB programming language. An important step in most clustering process is to select a distance measure, which determine the similarity between each data objects from calculation. This will manipulate the shape of the clusters, as some data objects will be close to one another according to one distance and farther away according to another. They are distinction whether the clustering uses symmetric or asymmetric distances.

The iris flower have various species ,in that three species namely Iris setosa, Iris virginica and Iris versicolor sre taken for clustering based on the available data sets provided by Fisher's Iris data set. The fifty data sets classify the length and width of sepals, petals of three species commonly. One of the clusters contains Iris setosa and the other cluster contains both Iris virginica and Iris versicolor and is not separable without the species information.so this experimental results proves the efficiency by clustering all the three species with time complexity. The k-mean algorithm classifies the species with user defined iteration along with dependency of initial centroids and final number of clusters. .

The output obtained from the clustering are shown below by K-mean clustering are shown in fig-1 and fig-2



shows the iteration when N=40.

Fig-1: K-MEAN CLUSTERING

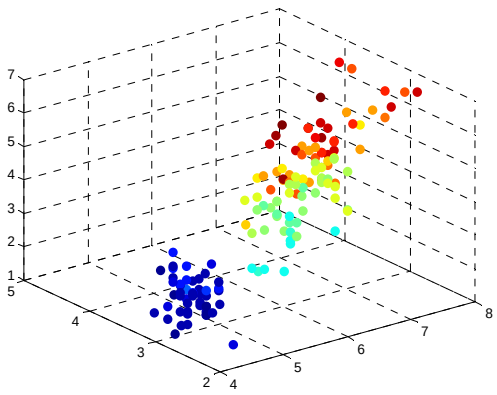


Fig 2: K-MEAN CLUSTERING WHEN N=40

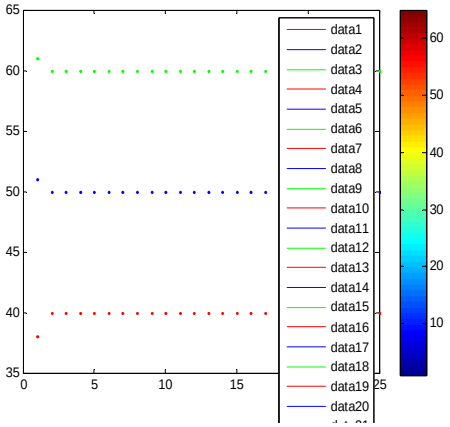


Fig 3: Y-MEAN CLUSTERING

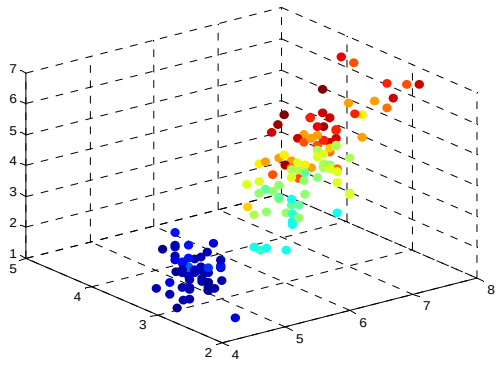


Fig 4: Y-MEAN CLUSTERING WHEN N=25

The Y-mean clustering algorithm overcomes the drawbacks of k-mean clustering by working on trained set of normalized input data. Then undergoes the process of splitting, merging with the deletion of empty clusters to avoid degeneracy. The output obtained by Y-mean cluster the data set in few iteration N=25 when compared with k-mean is shown in figure 3 and 4.The red ,blue and green color indicates the three species of Iris flowers in it.

TABLE 1 K-MEAN RESULTS

Cluster		1	2	3	4	5	Total	Distance
1	size	58	69	27	25	21	218	37
2	size	39	41	35	47	38	219	8
3	size	46	33	43	40	38	196	9
4	size	38	41	44	39	38	221	0
5	size	51	38	43	30	38	221	10

TABLE 2 Y-MEAN RESULTS

Cluster		1	2	3	4	5	Total	Distance
1	size	39	43	39	49	30	172	0
2	size	62	28	46	29	46	156	8
3	size	37	44	31	36	42	165	9
4	size	52	45	28	37	38	162	0
5	size	33	33	38	42	52	163	10

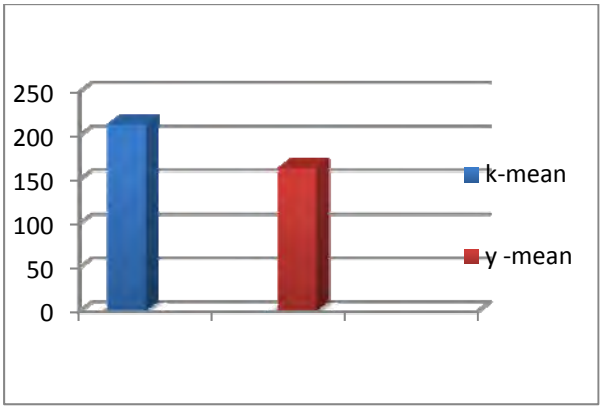


Fig 5: AVERAGE RUN TIME

The results of both the algorithms are analyzed based on the number of data points and the computational time of each algorithm. The performance of the partition algorithm is analyzed by experimental results in iris data sets. The number of data points is clustered by the algorithm as per the distribution of arbitrary shapes of the data points. Time complexity analysis is a part of computational complexity theory that is used to describe an algorithms use of computational resources; in this case, the best case and the worst maximum and minimum time taken by the Y-Means algorithm is 172 and 156 respectively. Like, from table 2, 221 and 196 are the maximum and minimum time taken by the K-mean algorithm. The performance of the algorithms have been analyzed for several iterations by considering different data points (for which the results are not shown) as input (300 data points, 400 data points etc.) and the number of clusters are 10 and 15 (for which also the results are not shown), the obtained results are found to be highly adequate. Figure 5 shows that the graph of the average results of the distribution of data points. The average execution time is taken from the tables 1 and 2. It is easy to identify from the figure 5 that there is a difference between the times of the algorithms. Here, it is found that the average execution time of the Y-Means algorithm is very less by comparing the K-Means algorithm.

4. CONCLUSIONS

This paper presents comparative analysis of a two unsupervised clustering algorithm namely y-mean and k-mean for data classification. Experimental results show that Y-means has a very good performance compared K-means. Furthermore, we also analyzed the overall performance of our algorithm by using three different species of iris datasets as initial input for clustering. The outcomes of this experiment provide the best clustering of the data set for Y-means than k-mean in few iteration steps and improved run time. Our future work will mostly concentrate on various ways to improve not only the performance of the algorithm but also on the accuracy and efficiency of the clustering.

References

[1] Chan, P.K., Mahoney, M.V., Arshad, and M.H.*: Managing cyber threats: Issues, approaches, and challenges. In: Learning Rules and Clusters for Anomaly Detection in Network Traffic*, ch. 3, pp. 81–99. Springer, Heidelberg (2005).

[2] Cortes, C., Vapnik, V.: *Support-vector networks. Machine Learning 20(3), 273–297 (1995) 3. Cover, T., Hart, P.G.: Nearest neighbor pattern classification*. IEEE Transactions on Information Theory IT-13(1), 21–27 (1967).

[3] Maria Camila N. Barioni, Humberto L.Razente, Agma J. M. Traina, Caetano Traina Jr, *An efficient approach to scale up k-medoid based algorithms in large databases, 2006.*

[4] Marta V. Modenesi, Myrian C. A. Costa, Alexandre G. Evsukoff,, and Nelson F.F.Ebecken, *Parallel Fuzzy C-Means Cluster Analysis, High Performance Computing for Computational Science - VECPAR 2006.*

[5] Kaufman, L. and P.J. Rousseeuw,*"Finding Groups in Data: an Introduction to Cluster Analysis"*, John Wiley and Sons, 1990

[6] Manikandan .R, *Improving Efficiency of textual static web content mining using clustering techniques* ,Journal of Theoretical and Applied Information Technology ,Vol.33,No.2,2011