Journal of Theoretical and Applied Information Technology
10th March 2014. Vol. 61 No.1

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

CLUSTERING OF HIGH DIMENSIONAL DATASET USING K-MAM (MAX-AVG-MIN) METHOD WITH PRINCIPAL COMPONENT ANALYSIS A HYBRID APPROACH

S.DHANABAL[#1] , DR.S.CHANDRAMATHI[*2]

[#]Assistant Professor, Jansons Institute of Technology, Coimbatore, Tamilnadu, INDIA

[*]Dean – Electrical Science, Sri Krishna Institute of Technology, Coimbatore, Tamilnadu, INDIA

EMAIL : [1]vhdhanabal@gmail.com, [2]schandrarajan@yahoo.com

ABSTRACT

Clustering the high-dimensional data set is one of the main issues in clustering analysis. Reducing the data from high-dimensional to a meaningful representation of low dimensional will increase the efficiency of clustering algorithms. The performance of K-means clustering algorithm is poor for high dimensions. Hence, to improve the efficiency, in this paper, we apply the efficient dimension reduction technique, Principal Component Analysis (PCA), to obtain possible uncorrelated variables, called Principal Components (PCs), from the original dataset. Then the reduced set is used to find the initial centroid using the k-means initialization method, k-MAM (Maximum-Average-Minimum) and then it is applied to K-Means clustering algorithm. The results are compared to k-means with PCA. The final results show that the k-MAM with PCA outperforms well in terms of accuracy and number of iterations compared to k-means, k-MAM and k-means with PCA for high dimensional data set.

KEYWORDS: *Clustering-Dimensionality Reduction - Principal Component Analysis - K-Means Algorithm- Initialization Method- K-MAM.*

1. INTRODUCTION

Dimension reduction is the process of transforming the high-dimensional data set into a meaningful representation of lower dimensionality that corresponds to the intrinsic dimensionality of the data whereas Clustering is the process of grouping the data which are close together. Kriegel et. al identified [1] four problems that has to be overcome for clustering the high-dimensional data set (a) The Curse of dimensionality is that, multiple dimensions are hard to visualize and complete enumeration of all subspaces become intractable with increasing dimensionality (b) The concept of distance becomes less precise as the number of dimensions grow, since the distance between any two points in a given dataset converges. The discrimination of the nearest and the farthest point in particular becomes meaningless. (c) Local Relevance Problem is the one in which different clusters might be found in different subspaces , so a global filtering of attributes is not sufficient (d) Given a large number of attributes , it is likely that some attributes are correlated. Hence, clusters might exist in arbitrarily oriented affine subspaces.

The most popular, unsupervised, partition clustering algorithm is k-means. It is used to find 'k' clusters which minimize SSE (Sum-Squared-error). The performance of k-means algorithm fully depends on the initial seed. If the initial partitions are not chosen carefully, then the computation will run the chance of converging to a local minimum rather than the global minimum solution. The initialization step is therefore very important. To overcome this problem it might be a good idea to run the algorithm several times with different initializations. If the results converge to the same partition then it is likely that a global minimum has been reached. This, however, has the drawback of time consuming and computationally expensive. Various initialization techniques for k-means are proposed by many researchers at various point of time. All methods are working towards achieving better local minimum than global minimum. We are proposing an algorithm called "k-means MAM" which can converge very

Journal of Theoretical and Applied Information Technology
10th March 2014. Vol. 61 No.1

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

fast with better accuracy. Our main objective is proposing a framework to combine the k-Max-Avg-Min (MAM) initialization method of clustering with principal component analysis (PCA) dimension reduction method to overcome aforesaid difficulties and improving efficiency and accuracy in K-Means algorithm to apply in high dimensional datasets

This paper is organized as follows: Section 2 presents an overview of clustering algorithms, K-Means clustering and drawbacks of k-means clustering algorithm. Section 3 presents the works related to the initialization of k-means along with PCA. The proposed K-MAM and the algorithm of PCA with K-MAM method is explained in Section 4. Section 5 deals with the Letter Recognition Dataset description. Experimental setup and the result analysis are given in section 6. Section 7 deals with conclusion and future work.

2 Clustering

Cluster analysis is one of the major data analysis methods widely used for many practical applications in emerging areas. Clustering is the process of finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups. A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity [2]. The quality of a clustering result depends on the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns [3].

2.1 K-Means Clustering Algorithm

K-means is one of the popular clustering algorithms which converge very fast [4]. The algorithm begins with random initial centroids and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centroids until a convergence criterion is met after certain number of iterations. The K-means algorithm is popular because it is easy to implement, and its time complexity is O (n), where 'n' is the number of patterns. Since there is at most k^n possible clustering, the process will always terminate. The basic algorithm works as follows:

Algorithm 1: K-means Algorithm

Step 1: Arbitrarily choose K centre locations (C_1,..., C_K).

Step 2: Assign each X_i to its nearest cluster centre C_i.

Step 3: Update each cluster centre C_i as the mean of all X_i that have been assigned as

closest to it.

Step 4: Calculate D = $\sum_{i=0}^{n} \min_{i=1..k}$ d(Xi,Ci)]

Step 5: if the value of D has converged, then return (C_1, ..., C_K); else go to Step 2.

In real, the algorithm requires very few iteration than any other clustering algorithms. Despite being used in a wide area of applications, the K-Means algorithm is not exempt of drawbacks. Some of these drawbacks have been extensively reported in the literature. The most important are listed below:

1. It does not provide approximation guarantee.[5]
2. It converges to the local optimum solutions. [6]
3. The results obtained from this algorithm are strongly dependent on its initial points.[6]
4. Number of clusters need to be known in advance.[7]

The main advantage of this approach stems from the fact that this framework is able to obtain better clustering with reduced complexity and also provides better accuracy and efficiency for high dimensional datasets.

3 RELATED WORKS

Dimension reduction comprises of Feature Selection and Feature Extraction . Increase in dimensionality, degrades the performance of the query in the index structures. Dimensionality reduction algorithms are the only known solution that supports scalable object retrieval and satisfies precision of query results. Feature transforms the data in the high-dimensional space to a space of fewer dimensions. The data transformation is linear as in principal component analysis (PCA).

Adnan Alrabea et.al [8] uses Principal Component Analysis (PCA) for generating the first principal component for initializing the centroid for k-means clustering. Initially, the principal components in the dataset are gathered using PCA. From the obtained components, the first principal component is used for initializing the cluster centroid. It has a better accuracy. Different methods have been proposed by combining PCA

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

with k-means for high dimensional data set. But the accuracy of the k-means clusters heavily depending on the random choice of initial centroids.

Fahim A M et al. [9] proposed an efficient method for assigning data points to clusters. In this, the distance between data points and all the centroids are computed for each iteration. The distance function is calculated based on Euclidean Distance and another one based on a heuristic to reduce the number of distance calculations. The main drawback of this method is that the initial centroids are determined randomly which in turn increases complexity.

Madhu Yedla et al. [10] proposed an initialization method to find the better initial centroids with reduced time complexity. This method starts with checking the given dataset whether it contains the negative value attributes or not. Then, it transform all the data points in the dataset to the positive space by subtracting each data point attribute with the minimum attribute value in the given data set. After that, it calculate the distance for each data point from the origin and sort it based on the distance. Finally, partition the sorted dataset into k equal sets and the middle point in each set is taken as initial centroid. They used heuristics approach to assign the data points to initial centroid.

K. A. Abdul Nazeer, M. P. Sebastian [11] proposed an enhanced k-means initialization. Their method iinitially, compute the distances between each data point and all other data points in the set of data points. Then find out the closest pair of data points and form a set A1 consisting of these two data points, and delete them from the data point set D. Then determine the data point which is closest to the set A1, add it to A1 and delete it from D. Repeat this procedure until the number of elements in the set A1 reaches a threshold. At that point go back to the second step and form another data-point set A2. Repeat this till 'k' such sets of data points are obtained. Finally the initial centroids are obtained by averaging all the vectors in each data-point set.

Tajunisha et al.[12] proposed the performance analysis of various initialization technique of k-means on high dimensional data . M.Emre Celebi et al.[13] compared eight linear time complexity initialization methods and found that Forgy, Macqueen, and maximin methods often perform poorly. R.Indhumathi et al [14] proposed a hybrid approach by bisecting k-means algorithm. It starts with one large cluster of all the data points and divides the whole dataset into two clusters. K-means algorithms run multiple times to find a split that produce maximum intra cluster similarity. Then the cluster with largest size is picked to split further. This cluster can be chosen based upon minimum intra cluster similarity also. This algorithm runs for k-1 times to get k clusters. This algorithm performs better than regular K means because bisecting K-means produces almost uniform sized clusters. But the running time complexity is high as it spends more time to find a split.

4 K-MAM INITIALIZATION TECHNIQUE

In most of the initialization methods, it is observed that initial centroids are chosen randomly or by using a systematic approach. In such a way that the initial seeds may either fall in a dense area or it will give more outliers as a result. This is because points are chosen as maximum or somewhere else between maximum and minimum in most of the cases. In K-means algorithm, some data points which are far away from the centroid may be discarded due to high SSE value even though the point belongs to that centroid. To overcome this problem, our proposed method[17] can consider the points at the extreme ends and then finds the centroid for k-means. This can be implemented in three phases: (i) First, apply PCA to reduce the dimensionality of the given dataset (ii) Apply k-MAM method to find the initialization and (ii) using the initial centroids, apply normal k-means algorithm for finding the clusters. The two phases are depicted in the figure 1.

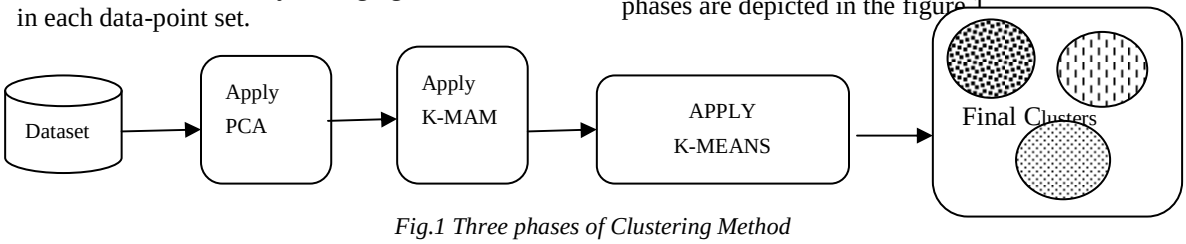

Fig.1 Three phases of Clustering Method

Journal of Theoretical and Applied Information Technology
10th March 2014. Vol. 61 No.1

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

Initially, choose the first data point (D_1) as the initial seed and calculate the distance to all other points. Almost all the initialization methods, explained above, choose the initial seed randomly or using some mathematical formulation. If the initial seeds are chosen randomly, again there is a chance of having more outliers. So, take the first data point P_1 as the initial seed and then calculate the distance from the initial seed to all other data points using Euclidean distance formula (1)

$$D(x, y) = [\sum_{i=1}^{n} (x_i - y_i)^2]^{1/2} \quad \text{-------- (1)}$$

where D(X,Y) is the distance between the point x_i ε X and y_i ε Y belongs to the data set D.

Once the distance is calculated, choose the maximum distance and mark it as P_2. Then, from P_2, again find the distance using equation (1) to all other data points and choose the minimum distance, say P_3. This is chosen because this point is the nearest point from P_2. This will give the minimum point on one end. Then, from P_3, again find the distance using equation (1) to all other data points and choose the maximum distance, say P_4. This is chosen because this point is the farthest point from P_3.This will give the maximum point on the other end. Then, calculate the mean, say P_5, by adding P_3 and P_4 and divided by number of rows in the dataset. The mean point may be chosen approximately nearest to the midpoint. Then, add these centroids to the centroid set. If k=2, then choose P_3 and P_4 as centroids whereas if the number of clusters, say k=3, then choose P_3, P_4 and P_5 as centroids and proceed as normal k-means. If k>3, then iteratively choose centroids by calculating the average (i)between minimum and average distances, (ii) between average and maximum distances inclusive of minimum (P_3) and maximum (P_4) centroids. In all the calculations, we kept the maximum, average and minimum which is chosen at first. Proceed in this way until 'k' cluster centroids are found.The convergence criterion for the k-means algorithm is minimizing the Sum-Squared-Errors (SSE).The formula to calculate SSE value is given below (2).

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist(C_i, X)2 \quad \text{---------- (2)}$$

where *dist* is the standard Euclidean distance between the centroid C_i and X in the Euclidean distance. The overall implementation is given in the algorithm 1,2 and algorithm 3. The main advantage of this algorithm is to find the centroids which are spread between minimum and maximum point of the given data set.

Algorithm 1: PCA for dimension reduction

Principal component analysis (PCA) [18] in multivariate statistics is widely adopted as an effective unsupervised dimension reduction method and is extended in many different directions. The main justification of dimension reduction is that PCA uses singular value decomposition (SVD) which gives the best low rank approximation to original data in L2 norm. The algorithm for PCA is given below.

Let $x_1, x_2, ..., x_M$ are *N* vectors

Step 1: Calculate $\bar{\upsilon} = 1/m \sum_{i=1}^{M} xi$

Step 2 : Subtract the mean : $\Phi_i = x_i - \bar{\upsilon}$

Step 3 : Form the matrix $A = [\Phi_1 \Phi_2 .... \Phi_m]$ where A is N x M matrix then compute

$$C = 1/M \sum_{n=1}^{M} \Phi_n \Phi_n^T = A.A^T$$

where C is the Covariance matrix.

Step 4: Compute the eigen values of C $=\mu_1 > \mu_2 > .... > \mu_N$.

Step 5: compute the eigenvectors of *C*: $u_1, u_2, \ldots, u_N$

- Since *C* is symmetric, *u*1, *u*2, . . . , *uN* form a basis, (i.e., any vector *x* or actually $(x - \bar{\upsilon})$, can be written as a linear combination of the eigenvectors):

$$x - \bar{\upsilon} = b_1 u_1 + b_2 u_2 + \ldots + b_N u_N = \sum_{i=0}^{N} b_i u_i$$

Step 6: (**Dimensionality Reduction Step**) keep only the terms corresponding to the *K* largest eigen values:

$$x - \bar{\upsilon} = \sum_{i=0}^{k} b_i u_i \quad \text{where } K << N$$

Step 7: return Reduced Dataset rd.

Journal of Theoretical and Applied Information Technology
10th March 2014. Vol. 61 No.1

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

Algorithm 2: Cluster K-MAM (Dataset rd, Cluster k, Centroid C{})

Step 1:

Select the first data instance P_1 as the initial seed from the dataset *d*.

Step 2:

Calculate the distance D_1 from P_1 to all the data instances and select the maximum distance, say P_2.

Step 3:

Calculate the distance D_2 from P_2 to all the data instances and select the minimum distance, say P_3.

Step 4:

Calculate the distance D_3 from P_3 to all the data instances and select the maximum distance, say P_4. Sort the distances such that P_3 is the first instance and P_4 is the last instance.

Step 4:

If the number of clusters, say *k*, is 2, then add P_3 and P_4 to the centroid set *C*. If the number of clusters, *k*=3, then add P_5 to the centroid set C where $P_5 = (P_{3+} P_4)/n$. Here 'n' is the number of rows in the given data set.

Step 4:

If the number of clusters is greater than 3, then calculate mean between P_{i-1} and P_{i+1} and add it to the centroid set C where i= 4 to n.

Step 5:

If the number of centroids is equal to the number of clusters required then stop else calculate mean between P_i and P_{i+1} and add it to the centroid set *C*

Step 6:

Repeat step 4 and 5 until '*k*' cluster centroids are found.

Step 7:

Return the centroid set *C* to *k*-means program and stop the process.

The modified k-means algorithm which takes the centroid from *k-MAM* is given below:

Algorithm 3: Modified K-means

Step 1:

Let C = K-MAM(Dataset D, Cluster K, Centroid C{})

Step 2:

Assign each X_i to its nearest cluster centre C_i.

Step 3:

Update each cluster centre C_i as the mean of all X_i that have been assigned as closest to it.

Step 4:

Calculate *dist* = $\sum_{i=n}^{n} \min_{i=1..k}$ d(Xi,Ci)]

Step 5:

If the value of *dist* has converged, then stop; else go to Step 2

Step 6:

Stop the process.

5. DATASET DESCRIPTION

The dataset used in this experiment is Letter recognition dataset which is taken from UCI machine repository dataset [15]. The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15. For the better analysis, we have chosen 2291 instances with 17 attributes to identify the letters A, B and C.

6. EXPERIMENTAL SETUP

In this section, the performance of the proposed method is compared based on five criteria, one on effectiveness and three on efficiency. The effectiveness criterion is SSE and efficiency criteria are number of iterations, CPU time, misclassified percentage and accuracy. They are defined below:

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

1. **Sum-Squared Errors (SSE)**: This is the Objective function of k-means algorithm. The convergence of the dataset is either based on the number of iterations reached or SSE value is greater than the threshold value.

2. Number of Iterations: K-means requires number of iterations until reaching convergence when it is initialized by the centroid.

3. CPU Time: This is the total time taken by the CPU for the initialization and clustering phases.

4. Misclassified Percentage: This is the percentage of total number of instances that is wrongly classified by the total number of instances.

5. Accuracy: This is the percentage of total number of instances that is perfectly classified by the total number of instances. Accuracy discovers the one-to-one relationship between clusters and classes and measures the extent to which each cluster contains data points from the corresponding class and it has been used as performance measures for clustering analysis. Accuracy can be described

$$\text{Accuracy} = \text{Max} \sum_{C_k L_m} T(C_k, L_m))/n$$

where *n* is the number of data points, *Ck* denotes the *k*-the cluster, and *Lm* is the *m*-the class. $T(C_k,L_m)$ s the number of data points that belong to class *m* are assigned to cluster *k*. Accuracy is then computed as the maximum sum of for all pairs of clusters and classes, and these pairs have no overlaps.

The experiments are carried out on a PC with an Intel Core i3 processor (2.4 GHz) and 4G byte of internal memory running in the Windows 7 Home premium operating system. For finding the principal components, we have used Weka 3.6.4 tool. The implementation of our algorithm is done in DOT NET platform using C# language. The results are plotted in graph using MS-Excel.

6.1 Experimental Results

Letter recognition original dataset is reduced using principal component analysis reduction method. The ultimate goal of the work is to reduce the SSE value and thereby increasing the accuracy and decreasing the execution time. As the number of obtained principal component is same with the number of original value, the weaker components are eliminated from this PC set. For this, we have calculated the corresponding variance, its proportion and cumulative variances. Then the ranking is calculated based on the cumulative variance which is shown in Table 1. We have eliminated PCs whose Eigen values are smaller than a fraction of the mean Eigen value. The Eigen value, proportion, cumulative variance and its ranking are shown in Table1. For the analysis purpose, we clustered the letter recognition dataset into 5, 6,7 and 8 clusters. The experiments are conducted starting with minimum of five principal components. Principal components whose Eigen value less than 0.2 are ignored. So, we have selected top 5 PCs.For analysis, the transformation matrix with reduced PCs is applied to the normalized data set to produce the new reduced projected dataset.

From the original dataset, we applied k-MAM initialization technique to find the centroids. Figure 1 shows the result of running k-means and k-MAM initialization method on the letter dataset without using dimension reduction. The result shows that the misclassified % of k-means is 1.22 % higher than the k-MAM initialization ie., k-MAM has a higher accuracy of 1.22% than k-means. The number of iterations and the CPU time is also low by 7 and 14 ms respectively.

Journal of Theoretical and Applied Information Technology
10th March 2014. Vol. 61 No.1

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

Table 1. The Eigen Value, Proportion, Cumulative Variance And Ranking Of Reduced Pcs.

Sl.No	EigenValue	Variation%	Cumulative value	Ranking
1	4.29539	26.846	26.846	0.7315
2	2.62544	16.409	43.255	0.5674
3	1.72108	10.757	54.012	0.4599
4	1.3691	8.557	62.569	0.3743
5	1.05136	6.571	69.14	0.3086
6	0.98007	6.125	75.265	0.2473
7	0.88927	5.558	80.823	0.1918
8	0.62586	3.912	84.735	0.1527
9	0.59547	3.722	88.457	0.1154
10	0.49189	3.074	91.531	0.0847
11	0.42648	2.666	94.197	0.058
12	0.2662	1.664	95.861	0.0414

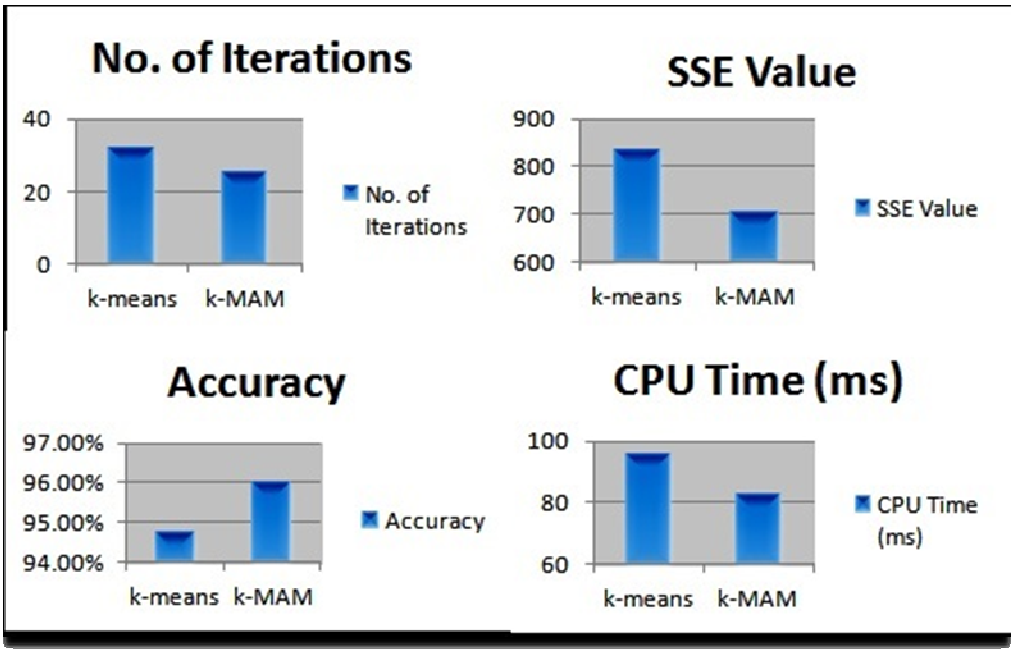

Fig.1. Comparison Of K-Means And K-MAM Without Dimension Reduction

Criteria	Reduced No. of Iterations	SSE Value	Accuracy%	Reduced CPU Time (ms)
Increased Efficiency	7	133.03	1.22	13.5

Journal of Theoretical and Applied Information Technology
10th March 2014. Vol. 61 No.1

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

Table 2. Comparison Of Increased Efficiency Of K-MAM Over K-Means Without Dimension Reduction.

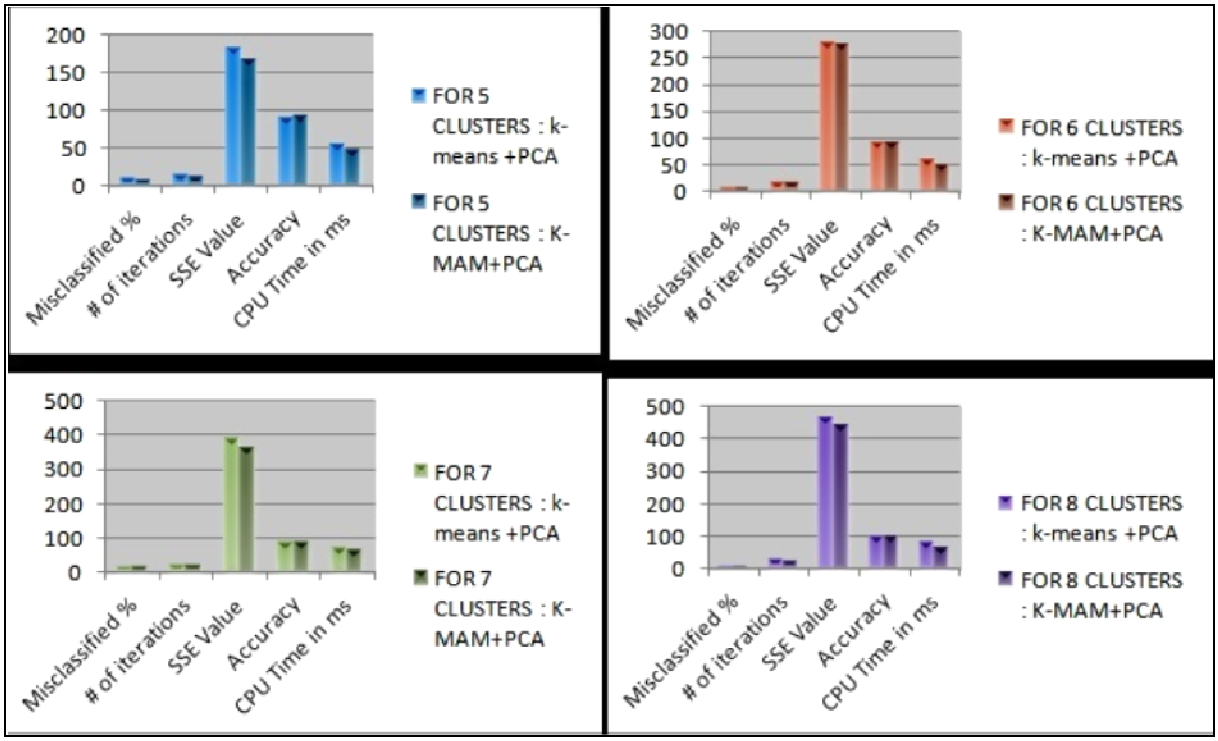

Fig.2 Comparison Analysis Of K-Means + PCA Over K-MAM +PCA With Dimension Reduction On Letter Recognition Dataset. A) 5 Clusters B) 6 Clusters C) 7 Clusters D) 8 Clusters.

Figure 2 shows the comparison results of k-means with PCA and k-MAM with PCA for the selected clusters 5, 6, 7 and 8. K-means with PCA value specified in the table are taken based on the mean of 10 runs. For the 5 Clusters, the accuracy is improved by 2.28 %, # of iterations are minimized by 2 and the execution time is considerably reduced by 7.7 ms when considering k-MAM with PCA over k-means with PCA. For the 6 Clusters, the accuracy, misclassified % and the # of iterations are all same but the CPU time is little bit reduced to 6.2 ms. This is because k-MAM with PCA converges very fast compared to the k-means with PCA. For the 7 Clusters, the # of iterations are same but the accuracy is increased by 1.37 and its CPU time is reduced by 8.3 ms. SSE value is also considerably reduced by 97.46. In the case of 8 Clusters, the accuracy is tremendously good and its misclassified % is only 2.28% compared to k-means with PCA. Also the CPU time is reduced by 13.99 ms. SSE values of k-MAM with PCA for all the Clusters are very less compared to the SSE values of k-means with PCA. This is because the sum of the total clustering errors for all clusters was reduced as much as possible while inter distances between clusters are maintained to be as large as possible.

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

Table 3 Increased Efficiency Of K-MAM + PCA Over K-Means + PCA With Dimension Reduction On Letter Recognition Dataset

Criteria	Increased Efficiency of k-MAM +PCA over K-means + PCA (Selected Clusters)			
	5	6	7	8
Misclassified %	2.71	0	1.37	2.28
# of Iterations	2	3	3	5
SSE value	15.04	3.05	27.46	24.08
Accuracy	+2.71	0	+1.37	+2.28
Reduced CPU time in ms	7.7	9.9	8.3	13.99

7. CONCLUSION AND FUTURE WORK

In this paper a dimensionality reduction through PCA, is applied to *k-means* algorithm and k-MAM initialization method. Using Dimension reduction of principal component analysis, original Letter recognition dataset is transformed to a reduced data set. Then, it was partitioned in to k clusters using k-means and k-MAM initialization algorithm, in such a way that the sum of the total clustering errors for all clusters was reduced as much as possible while inter distances between clusters are maintained to be as large as possible. The proposed a new algorithm, k-MAM with PCA is applied to initialize the clusters and then centroids are applied to k-means algorithm. The final results show that the k-MAM with PCA outperforms well in terms of accuracy and number of iterations compared to the k-means, k-MAM and k-means with PCA for high dimensional data.

In future, rough set theory can be applied as a dimension reduction technique to apply k-means and k-MAM.

REFERENCES

[1] Kriegel,Hans-Peter;Kroger,Peer:Zimek, Arthur , " Clustering High Dimesional Data: A survey on subspace clustering, pattern based clustering and correlation clustering" ACM Transactions on Knowledge Discovery from Data (New York,NY:ACM)3(1):1-58,2009

[2] Yan Jun, Zhang Benyu, Liu Ning, Yan Shuicheng, Cheng Qiansheng, Fan Weiguo, Yang Qiang, Xi Wensi, and Chen Zheng,2006. Effective and efficient dimensionality reduction for large- scale and streaming data preprocessing, *IEEE transactions on Knowledge and Data Engineering*, Vol. 18, No. 3, pp. 320-333

[3] Arthur, D. and Vassilvitskii, S. How slow is the k-means method?, pp. 144–153,2006

[4] MCQUEEN, J. "Some methods for classification and analysis of multivariate observations", 1967

[5] E.W. Forgy, Cluster analysis of multivariate data: efficiency versus interpretability of classifications, Biometrics 21 (3) 768–769,1965.

[6] S.Z. Selim, M.A. Ismail, K-means type algorithms: a generalized convergence theorem and characterization of local optimality, IEEE Trans. Pattern Anal. Mach. Intell. 6 81–87,1984.

[7] Fathian, M. & Amiri, B. A Honeybee-mating Approach for Cluster Analysis. *Advance Manufacture Tech*. 1(38), 809–821, 2008.

[8] Adnan Alrabea, A. V. Senthilkumar, Hasan Al-Shalabi, and Ahmad Bader, Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with PCA, *Journal of Advances in Computer Networks, Vol. 1, No. 2, June 2013*

[9] Chris Ding and Xiaofeng He (2004) : k-means Clustering via Principal component Analysis, In Proceedings of the 21st international conference on Machine Learning, Banff, Canada.

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

[10] Fahim A.M., Salem A.M., Torkey F.A., Saake,G and Ranadan M.A., " An Efficient k-means with good initial starting points", Georgian Electronic Scientific Journal, Computer Science and Telecommunication vol 2, No19., PP -47 -57 ,2009.

[11] Madhu Yedla et al. (2010) : "Enhancing K-means clustering algorithm with improved initial centers",International Journal of Computer Science and information Technologies. Vol.1(2), 2010, 121-125.

[12] Nazeer K. A., Abdul and Sebastian M.P. (2009): Improving the accuracy and efficiency of the k-means clustering algorithm, Proceedings of the World Congress on Engineering,Vol. 1, pp. 308-312.

[13] Tanjunisha and Saravan, Performance analysis of K-means with different initialization for high dimensional data" , International journal of Artificial Intelligence and application vol1 no.4, October 2010.

[14] Emre Celebi, Hassan A. Kingravi, Patricio A. Vela, A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm ",Expert Systems with Applications, 40(1): 200– 210, 2013.

[15] R.Indhumathi, Dr.S.Sathiyabama Reducing and Clustering high Dimensional Data through Principal Component Analysis, *International Journal of Computer Applications (0975 – 8887)*, Volume 11– No.8, December 2010

[16] C.L. Blake, C.J. Merz, UCI repository of machine learning databases. Available from: <http:// ww.ics.uci.edu/-mlearn/

[17] S. Dhanabal and S. Chandramathi, 2013. An Efficient K-Means Initialization Using Minimum Average- Maximum (MAM) Method. Asian Journal of Information Technology, 12: 77-82.

[18] I.T. Jolli_e. Principal Component Analysis. Springer Verlag, 1986.