MultiCraft

International Journal of Engineering, Science and Technology
Vol. 2, No. 2, 2010, pp. 59-66

INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND TECHNOLOGY

www.ijest-ng.com

A hybridized K-means clustering approach for high dimensional dataset

Rajashree Dash[1], Debahuti Mishra[*], Amiya Kumar Rath[2], Milu Acharya[3]

[1,*,3] *Institute of Technical Education and Research, Bhubaneswar, Orissa, INDIA*
[2] *Director, College of Engineering Bhubaneswar, Orissa, INDIA*
[*] *Corresponding Authors' e-mails: debahuti@iter.ac.in, rajashree_dash@yahoo.co.in ,amiyamiya@rediffmail.com, milu_acharya@yahoo.com*

Abstract

Due to incredible growth of high dimensional dataset, conventional data base querying methods are inadequate to extract useful information, so researchers nowadays is forced to develop new techniques to meet the raised requirements. Such large expression data gives rise to a number of new computational challenges not only due to the increase in number of data objects but also due to the increase in number of features/attributes. Hence, to improve the efficiency and accuracy of mining task on high dimensional data, the data must be preprocessed by an efficient dimensionality reduction method. Recently cluster analysis is a popularly used data analysis method in number of areas. K-means is a well known partitioning based clustering technique that attempts to find a user specified number of clusters represented by their centroids. But its output is quite sensitive to initial positions of cluster centers. Again, the number of distance calculations increases exponentially with the increase of the dimensionality of the data. Hence, in this paper we proposed to use the Principal Component Analysis (PCA) method as a first phase for K-means clustering which will simplify the analysis and visualization of multi dimensional data set. Here also, we have proposed a new method to find the initial centroids to make the algorithm more effective and efficient. By comparing the result of original and new approach, it was found that the results obtained are more accurate, easy to understand and above all the time taken to process the data was substantially reduced.

Keywords: Cluster analysis, K-means Algorithm, Dimensionality Reduction, Principal Component Analysis, Hybridized K-means algorithm

1. Introduction

Data mining is a convenient way of extracting patterns, which represents knowledge implicitly stored in large data sets and focuses on issues relating to their feasibility, usefulness, effectiveness and scalability. It can be viewed as an essential step in the process of knowledge discovery. Data are normally preprocessed through data cleaning, data integration, data selection, and data transformation and prepared for the mining task. Data mining can be performed on various types of databases and information repositories, but the kind of patterns to be found are specified by various data mining functionalities like class/concept description, association, correlation analysis, classification, prediction, cluster analysis etc.

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 (or related) to one another and different from (or unrelated to) the objects in other groups. A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns. A good survey on clustering methods is found in Xu *et al*. (2005).

K-means is a commonly used partitioning based clustering technique that tries to find a user specified number of clusters (k), which are represented by their centroids, by minimizing the square error function. Although K-means is simple and can be used for a wide variety of data types, it is quite sensitive to initial positions of cluster centers. There are two simple approaches to cluster center initialization i.e. either to select the initial values randomly, or to choose the first k samples of the data points. As an alternative, different sets of initial values are chosen (out of the data points) and the set, which is closest to optimal, is chosen. However, testing different initial sets is considered impracticable criteria, especially for large number of clusters Ismail *et al* (1989). Therefore, different methods have been proposed in literature by Pena *et al.* (1999). Again the computational complexity

of original K-means algorithm is very high, especially for large data sets. In addition the number of distance calculations increases exponentially with the increase of the dimensionality of the data. When the dimensionality increases usually, only a small number of dimensions are relevant to certain clusters, but data in the irrelevant dimensions may produce much noise and mask the real clusters to be discovered. Moreover when dimensionality increases, data usually become increasingly sparse, due to which data points located at different dimensions can be considered as all equally distanced and the distance measure, which, essentially for cluster analysis, becomes meaningless. Hence, attribute reduction or dimensionality reduction is an essential data-preprocessing task for cluster analysis of datasets having a large no. of features/attributes.

Dimensionality reduction specified by Maaten *et al*. (2007) and Davy *et al*. (2007) is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality that corresponds to the intrinsic dimensionality of the data. It falls into two categories i.e. Feature Selection (FS) and Feature Reduction (FR). Feature Selection algorithm aims at finding out a subset of the most representative features according to some objective function in discrete space. The algorithms of FS are always greedy. Thus, they sometimes cannot even find the optimal solution in the discrete space. Feature Extraction/ Feature Reduction algorithms aim to extract features by projecting the original high-dimensional data into a lower-dimensional space through algebraic transformations. It finds the optimal solution of a problem in a continuous space, but the computational complexity is more comparative to feature selection algorithm. Various types of feature reduction methods have been developed. PCA is a commonly used feature reduction method in terms of minimizing the reconstruction error.

Traditional K-means algorithm for cluster analysis developed for low dimensional data, often do not work well for high dimensional data like microarray gene expression data and the results may not be accurate most of the time due to noise and outliers associated with original data. Also the computational complexity increases rapidly as the dimension increases. Hence, to improve the efficiency, we proposed a method to apply PCA on original data set, so that the correlated variables exist in the original dataset would be transformed to possibly uncorrelated variables, which are reduced in size. Before applying PCA the dataset needs to be normalized, so that any attribute with larger domain will not dominate attributes with smaller domain. The resulting reduced data set obtained from the application of PCA will be applied to a K-means clustering algorithm. Here also we have proposed a new method to find the initial centroids to make the algorithm more effective and efficient. 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.

Section 1 of the paper deals with the introductory concepts of clustering, K-means clustering, its limitations, the need of dimensionality reduction for clustering and the goal of the paper. Some recent related works and other preliminaries on K-means algorithm, dimensionality reduction methods and some concepts of PCA have been discussed in section 2. Section 3 describes our new proposed algorithm for K-means clustering. Section 4 describes our approach in various steps with experimental activities and corresponding result discussion followed by conclusion in Section 5.

2. Related Work

Several attempts were made by researchers to improve the effectiveness and efficiency of the K-means algorithm. Yuan *et al*. (2004) proposed a systematic method for finding the initial centroids. However, Yuan's method does not suggest any improvement to the time complexity of the K-means algorithm. Belal *et al*. (2005) proposed a new method for cluster initialization based on finding a set of medians extracted from a dimension with maximum variance. Zoubi *et al*. (2008) proposed a new strategy to accelerate K-means clustering by avoiding unnecessary distance calculations through the partial distance logic. Fahim *et al*. (2009) proposed a method to select a good initial solution by partitioning dataset into blocks and applying K-means to each block. But here the time complexity is slightly more. . Though the above algorithms can help finding good initial centers for some extent, they are quite complex and some use the K-means algorithm as part of their algorithms, which still need to use the random method for cluster center initialization. Deelers *et al*. (2007) has proposed an enhancing K-means algorithm based on the data partitioning algorithm used for color quantization. The algorithm performs data partitioning along the data axis with the highest variance. Nazeer *et al*. (2009) proposed an enhanced K-means algorithm, which combines a systematic method for finding initial centroids and an efficient way for assigning data points to cluster. This method ensures the entire process of clustering in O (n^2) time without sacrificing the accuracy of clusters. Similarly Xu *et al*. (2009) specify a novel initialization scheme to select initial cluster centers based on reverse nearest neighbor search. But all the above methods do not work well for high dimensional data sets. Yeung *et al*. (2000) presented an empirical study on principal component analysis for clustering gene expression data. But here the initial centroids are chosen randomly. Chao *et al*. (2005) also proposed a method for dimension reduction for microarray data analysis using Locally Linear Embedding.

2.1 K-means Clustering Algorithm

The K-means algorithm is one of the partitioning based, nonhierarchical clustering methods. Given a set of numeric objects $\boldsymbol{X}$ and an integer number k, the K-means algorithm searches for a partition of $\boldsymbol{X}$ into k clusters that minimizes the within groups sum of squared errors. The K-means algorithm starts by initializing the k cluster centers. The input data points are then allocated to one of the existing clusters according to the square of the Euclidean distance from the clusters, choosing the closest. The mean (centroid) of each cluster is then computed so as to update the cluster center. This update occurs as a result of the change in the

membership of each cluster. The processes of re-assigning the input vectors and the update of the cluster centers is repeated until no more change in the value of any of the cluster centers.

The steps of the K-means algorithm are written below:

1. Initialization: choose randomly *K* input vectors (data points) to initialize the clusters.
2. Nearest-neighbor search: for each input vector, find the cluster center that is closest, and assign that input vector to the corresponding cluster.
3. Mean update: update the cluster centers in each cluster using the mean (centroid) of the input vectors assigned to that cluster.
4. Stopping rule: repeat steps 2 and 3 until no more change in the value of the means.

2.2 Principal Component Analysis (PCA)

Principal Component Analysis by Valarmathie *et al*. (2009) and Yan *et al*. (2006) is an unsupervised Feature Reduction method for projecting high dimensional data into a new lower dimensional representation of the data that describes as much of the variance in the data as possible with minimum reconstruction error. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. It transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. Hence, PCA is a statistical technique for determining key variables in a high dimensional data set that explain the differences in the observations and can be used to simplify the analysis and visualization of high dimensional data set, without much loss of information.

2.3 Principal Component (PC)

Technically, a principal component 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 data set 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 is uncorrelated with all of the preceding components. When the principal component analysis will complete, 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. Covariance matrix is preferred when the variances of variables are very high compared to correlation. It would be better to choose the type of correlation when the variables are of different types. Similarly the SVD method is used for numerical accuracy.

2.4 Elimination Methods of Unnecessary PCs

The transformation of the dataset to the new principal component axis produces the number of PCs equivalent to the no. of original variables. But for many datasets, the 1st several PCs explain the most of the variances, so the rest can be eliminated with minimal loss of information. The various criteria used to determine how many PCs should be retained for the interpretation is as follows:

- Using Scree Diagram plots the variances in percentage corresponding to the PCs, which will automatically eliminate the PCs with very low variances.
- Fixing a threshold value of variance, so that PCs having variance more than the given threshold value will be retained rejecting others.
- Eliminate PCs whose Eigen values are smaller than a fraction of the mean Eigen value.

3. Proposed Hybridized K-means Clustering Algorithm

As original K-means clustering algorithm often does not work well for high dimension, hence, to improve the efficiency, we proposed to apply PCA on original data set, to obtain a reduced dataset containing possibly uncorrelated variables. Then the resulting reduced data set will be applied to the K-means clustering algorithm to determine the precise no. of clusters. As quality of the final clusters heavily depends on the selection of the initial centroids, here we proposed a new method to choose such data objects as initial centroids whose squared Euclidian distance is maximum among all the data objects, to make the algorithm more effective and efficient.

The proposed model is illustrated in Figure 1.

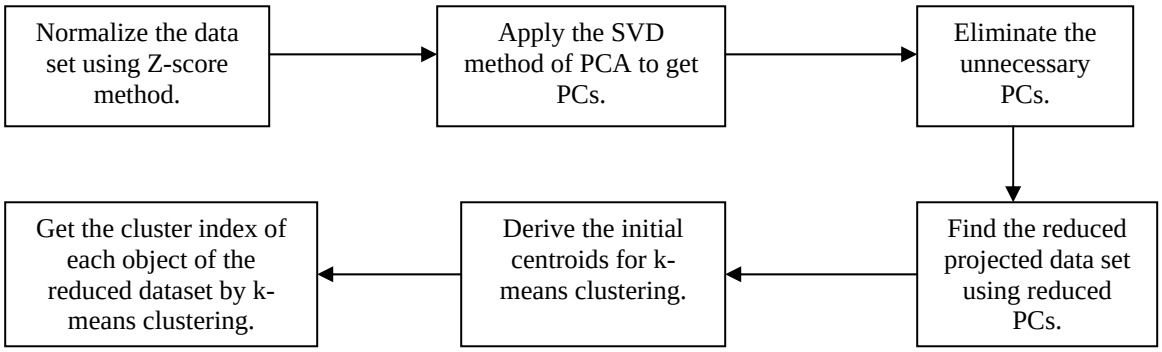

Figure 1: A Hybridized model for K-means Clustering

The steps of the hybridized k-means clustering algorithm are as follows.

Input: $X = \{d1, d2, \ldots\ldots, dn\}$ // set of n data items.
K // Number of desired clusters.
An array Cen [ ] having size k initially being empty.

Output: A set of k clusters

// Phase-1: Apply PCA to reduce the dimension of the data set

1. Organize the dataset in a matrix X.
2. Normalize the data set using Z-score.
3. Calculate the singular value decomposition of the data matrix. $X = UDV^T$
4. Calculate the variance using the diagonal elements of D.
5. Sort variances in decreasing order.
6. Choose the p principal components from V with largest variances.
7. Form the transformation matrix W consisting of those p PCs.
8. Find the reduced projected dataset Y in a new coordinate axis by applying W to X.

//Phase-2: Find the initial centroids

9. Set m=1.
10. Compute the distance between each data points in the set Y.
11. Choose the two data points y_i and y_j such that distance (y_i, y_j) is maximum.
12. $Cen[m] = y_i$; $Cen[m+1] = y_j$; $m=m+2$;
13. Remove the two objects y_i , y_j from Y.
14. While ($m <= k$)
    1. Find the distance of each object in Y to $Cen[i]$, for i = 1 to m-1.
    2. Find the average of all the distances to the centroid for each object in Y.
    3. Choose the data object y_o having maximum average distance from previous centroids.
    4. $Cen[m] = y_o$; $m = m+1$;
    5. Remove the object y_o from Y.

// Phase-3: Apply the K-means clustering with the initial centroids given in array Cen.

15. For each data point, in set Y, find the nearest cluster center from list Cen that is closest and assign that data point to the corresponding cluster.
16. Update the cluster centers in each cluster using the mean of the data points, which are assigned to that cluster.
17. Repeat the steps 15 and 16 until there are no more changes in the values of the centroids.

4. Experimental Activities and Result Discussion

Initially, we evaluated the proposed algorithm on a synthetic dataset with 15 data objects having 10 attributes as shown in table 1. Then three datasets, Pima Indian Diabetes data set, Breast Cancer data set and SPECTF Heart data set, taken from the UCI

machine learning repository are used for testing the accuracy and efficiency of the hybridized algorithm. Here the Sum of Squared Error (SSE), representing distances between data points and their cluster centers have used to measure the clustering quality. Among two solutions for a given dataset, the smaller the value of SSE and higher the accuracy, the better the solution.

Step 1: Normalizing the original data set
Using the Normalization process, the initial data values are scaled so as to fall within a small-specified range. An attribute value V of an attribute A is normalized to V' using Z-Score as follows:
$V' = (V\text{-mean}(A))/\text{std}(A)$
It performs two things i.e. data centering, which reduces the square mean error of approximating the input data and data scaling, which standardizes the variables to have unit variance before the analysis takes place. This normalization prevents certain features to dominate the analysis because of their large numerical values.

Table 1: The original data matrix X with 15 data objects having 10 attribute values.

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
Data1	1	5	1	1	1	2	1	3	1	1
Data2	2	5	4	4	5	7	10	3	2	1
Data3	3	3	1	1	1	2	2	3	1	1
Data4	4	6	8	8	1	3	4	3	7	1
Data5	5	4	1	1	3	2	1	3	1	1
Data6	6	8	10	10	8	7	10	9	7	1
Data7	7	1	1	1	1	2	10	3	1	1
Data8	8	2	1	2	1	2	1	3	1	1
Data9	9	2	1	1	1	2	1	1	1	5
Data10	10	4	2	1	1	2	1	2	1	1
Data11	11	1	1	1	1	1	1	3	1	1
Data12	12	2	1	1	1	2	1	2	1	1
Data13	13	2	1	1	1	2	1	2	1	1
Data14	14	5	3	3	3	2	3	4	4	1
Data15	15	1	1	1	1	2	3	3	1	1

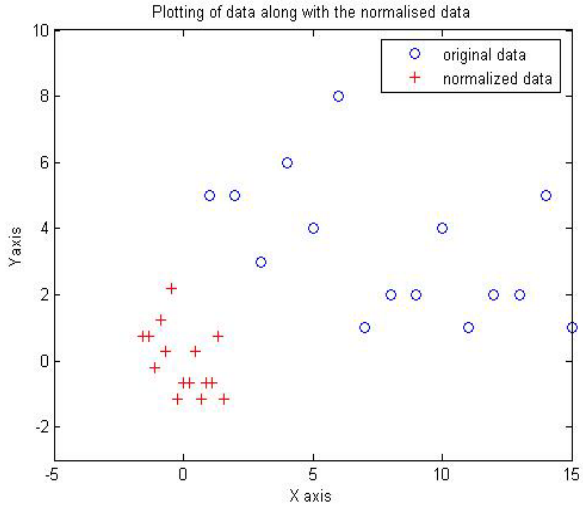

Figure 2: Plotting of data along with the normalized data.

Step 2: Calculating the PCs using Singular Value Decomposition of the normalized data matrix
Applying the steps given in phase 1 of the new proposed algorithm, the no. of PCs obtained is same with the no. of original variables. To eliminate the weaker components from this PC set we have calculated the corresponding variance, percentage of variance and cumulative variances in percentage, which is shown in Table 2. Then we have considered the PCs having variances less than the mean variance, ignoring the others. The reduced PCs are shown in Table 3.

Step 3: Finding the reduced data set using the reduced PCs
The transformation matrix with reduced PCs is formed and this transformation matrix is applied to the normalized data set to produce the new reduced projected dataset, which can be used for further data analysis. The reduced data set is shown in Table 4. We have also applied the PCA on three biological dataset and the reduced no. of attributes obtained for each dataset is shown in Figure 3.

Step 4: Comparison of efficiency and accuracy of the original k-means clustering and proposed algorithm.
The clustering results shown in Figure 4 and 5 by applying the standard k-means clustering to the original synthetic dataset and the proposed method to the reduced dataset are approximately same, but the time taken for clustering will be reduced due to less number of attributes. Again we compared the clustering results obtained by the k-means algorithm using random initial centers and initial centers derived by the proposed algorithm over 4 datasets with original dimension and with reduced dimension based on the sum of squared error distances (SSE), which is shown in Figure 6 and 7. The clustering results of k-means using random initial centers are the average results over 10 runs since each run gives different results. The SSE value obtained and the time taken in ms for 4 datasets with original k-means and new proposed algorithm is given in Table 5.

Table 2: The variances, variances in percentage, cumulative variances in percentage corresponding to the PCs .

	Variances	**Variances in %**	**Cumulative variances in%**
pc1	6.210578	62.10578	62.10578
pc2	1.054022	10.54022	72.646
pc3	1.016014	10.16014	82.80614
pc4	0.86546	8.654603	91.46075
pc5	0.455458	4.554576	96.01532
pc6	0.24649	2.464901	98.48022
pc7	0.108508	1.085079	99.5653
pc8	0.030248	0.302483	99.86779
pc9	0.010731	0.10731	99.9751
pc10	0.00249	0.024904	100

Table 3: Reduced PCs having variance greater than the mean variance.

Pc1	**Pc2**	**Pc3**
0.161104	-0.70692	0.222666
-0.34575	0.06143	0.104589
-0.37811	-0.12852	0.225285
-0.37785	-0.12565	0.21828
-0.34923	0.061232	-0.09679
-0.34712	0.275694	-0.1149
-0.28668	0.21192	-0.29648
-0.34062	-0.24545	-0.14089
-0.34594	-0.26379	0.318984
0.091787	0.457921	0.780392

Table 4: The reduced dataset containing 3 attributes

	v1	**v2**	**v3**
data1	0.536985	1.045863	-0.56448
data2	-2.76212	1.914677	-1.15055
data3	0.858597	0.73036	-0.64746
data4	-2.78991	-0.43292	1.310152
data5	0.502757	0.444459	-0.51029
data6	-7.19228	-0.525	0.259885
data7	0.691562	0.513194	-1.21072
data8	1.149898	-0.19297	-0.28827
data9	2.060598	1.744708	2.866126
data10	1.08491	-0.31268	-0.00678
data11	1.749453	-0.8052	-0.20218
data12	1.620462	-0.6419	-0.08667
data13	1.656486	-0.79997	-0.03688
data14	-0.70786	-1.51675	0.500656
data15	1.540447	-1.16586	-0.23254

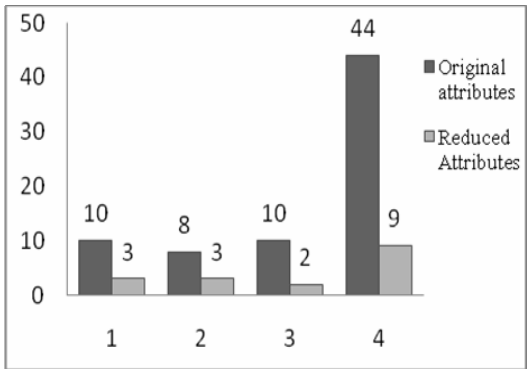

Figure 3: Plotting of original and reduced no. of attributes for Synthetic, Pima Indian Diabetes, Breast Cancer and SPECTF Heart datasets.

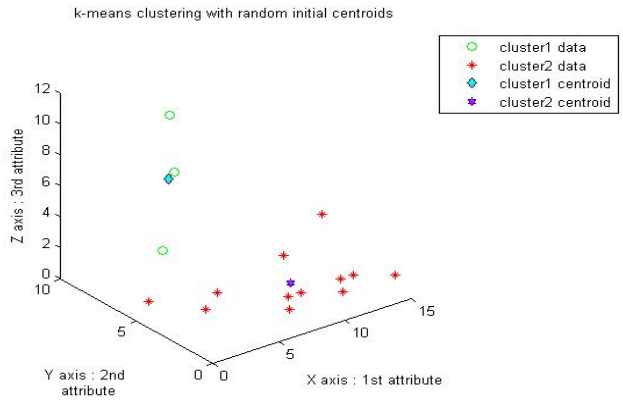

Figure 4: Clustering with original dataset by standard K-means algorithm.

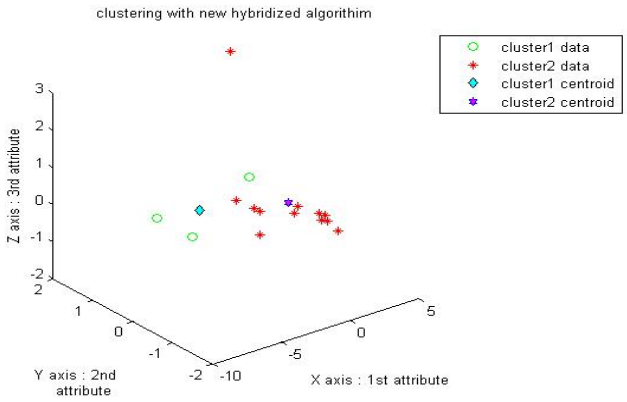

Figure 5: Clustering with reduced dataset by proposed algorithm.

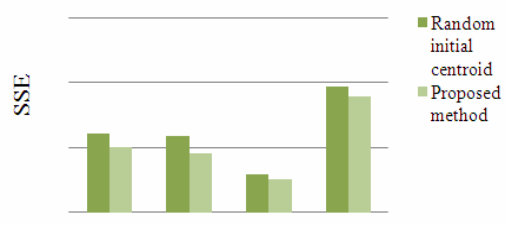

Figure 6: SSE results on synthetic, PID, Breast cancer, SPECTF heart datasets with original dimension.

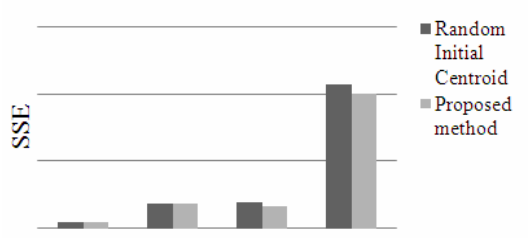

Figure 7: SSE results on synthetic, PID, Breast cancer, SPECTF heart datasets with reduced dimension.

Table 5: SSE values obtained and time taken in ms with original k-means and new proposed algorithm

Dataset	No of Instances	Original K-means Algorithm		Proposed Algorithm	
		SSE	Time taken(ms)	SSE	Time taken(ms)
Synthetic	15	608.446	78	47.8	65
Pima Indian Diabetes	50	59255	158	182.39	122
Breast Cancer	80	29253	167	165.94	131
SPECTF Heart	40	97075	145	996.8	112

The above results show that the new algorithm provides better SSE values for all the cases. Hence, in this regard it increases the efficiency of the original k-means algorithm. The accuracy of clustering determined by comparing the clusters obtained by the experiments with the available clusters for three data sets in UCI data set is shown in Figure 8. In all the cases the proposed algorithm provides better accuracy compared to the original k-means algorithm.

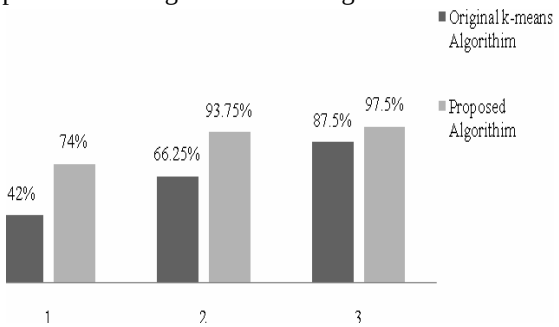

Figure 8: Clustering accuracy of PID, Breast cancer, SPECTF Heart datasets.

5. Conclusion

In this paper a hybridized K-means algorithm has been proposed which combines the steps of dimensionality reduction through PCA, a novel initialization approach of cluster centers and the steps of assigning data points to appropriate clusters. Using the proposed algorithm a given data set was partitioned in to k clusters 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 experimental results show that the proposed algorithm provides better efficiency and accuracy comparison to original k-means algorithm with reduced time. Though the proposed method gave better quality results in all cases, over random initialization methods, still there is a limitation associated with this, i.e. the number of clusters (k) is required to be given as input. Again the method to find the initial centroids may not be reliable for vary large dataset. Evolving some statistical methods to compute the value of k, depending on the data distribution is suggested for future research. Methods for refining the computation of initial centroids are worth investigating.

References

Belal M. and Daoud A., 2005. A new algorithm for cluster initialization, *World Academy of Science, Engineering and Technology*, Vol. 4, pp. 74-76.

Chao Shi and Chen Lihui, 2005. Feature dimension reduction for microarray data analysis using locally linear embedding, *3rd Asia Pacific Bioinformatics Conference*, pp. 211-217.

Davy Michael and Luz Saturnino, 2007. Dimensionality reduction for active learning with nearest neighbour classifier in text categorization problems, *Sixth International Conference on Machine Learning and Applications*, pp. 292-297.

Deelers S. and Auwatanamongkol S., 2007. Enhancing K-means algorithm with initial cluster centers derived from data partitioning along the data axis with the highest variance, *International Journal of Computer Science*, Vol. 2, No. 4, pp. 247-252.

Fahim A. M., Salem A. M., Torkey F. A., Saake G. and Ramadan M. A., 2009. An efficient k-means with good initial starting points, *Georgian Electronic Scientific Journal: Computer Science and Telecommunications*, Vol. 2, No. 19, pp. 47-57.

Ismail M. and Kamel M., 1989. Multidimensional data clustering utilization hybrid search strategies, *Pattern Recognition*, Vol. 22, No. 1, pp.75-89.

Maaten L.J.P., Postma E.O. and Herik H.J. van den, 2007. Dimensionality reduction: A comparative review", *Tech. rep. University of Maastricht*.

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.

Pena J. M., Lozano J. A. and Larranaga P., 1999. An empirical comparison of four initialization methods for the k-means algorithm, *Pattern Recognition Letters*, Vol. 20, No. 10, pp. 1027-1040.

Valarmathie P., Srinath M. and Dinakaran K., 2009. An increased performance of clustering high dimensional data through dimensionality reduction technique, *Journal of Theoretical and Applied Information Technology*, Vol. 13, pp. 271-273.

Xu R. and Wunsch D., 2005. Survey of clustering algorithms, *IEEE Trans. Neural Networks*, Vol. 16, No. 3, pp. 645-678.

Xu Junling, Xu Baowen, Zhang Weifeng, Zhang Wei and Hou Jun, 2009. Stable initialization scheme for K-means clustering, *Wuhan University Journal of National Sciences*, Vol. 14, No. 1, pp. 24-28.

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.

Yeung Ka Yee and Ruzzo Walter L., 2000. An empirical study on principal component analysis for clustering gene expression Data",*Tech. Report, University of Washington.*

Yuan F., Meng Z. H, Zhang H. X and Dong C. R, 2004. A new algorithm to get the initial centroids, *Proc. of the 3rd International Conference on Machine Learning and Cybernetics*, pp. 1191–1193.

Zhang Z., Zhang J. and Xue H., 2008. Improved K-means clustering algorithm, *Proceedings of the Congress on Image and Signal Processing*, Vol. 5, No. 5, pp. 162-172.

Zoubi M. B. Al., Hudaib A., Huneiti A. and Hammo B., 2008. New efficient strategy to accelerate k-means clustering algorithm", *American Journal of Applied Sciences*, Vol. 5, No. 9, pp. 1247-1250.

Biographical notes

Rajashree Dash has completed her B.Tech in Computer Sc. & Engineering from KIIT University. Now she is perusing her M.Tech in Computer Sc. & Engg at Institute of Technical Education & Research (ITER) under Siksh `O` Anusandhan University, Bhubaneswar. Her research areas include Data mining, Computer Graphics etc.

Debahuti Mishra is an Assistant Professor and research scholar in the department of Computer Sc. & Engg, Institute of Technical Education & Research (ITER) under Siksh `O` Anusandhan University, Bhubaneswar. She received her Masters degree from KIIT University, Bhubaneswar. Her research areas include Data mining, Bio-informatics Software Engineering, Soft computing . She is an author of a book Aotumata Theory and Computation by Sun India Publication (2008).

Dr.Amiya Kumar Rath obtained Ph.D in Computer Science in the year 2005 from Utkal University for the work in the field of Embedded system. Presently working with College of Engineering Bhubaneswar (CEB) as Professor of Computer Science & Engg. Cum Director (A&R) and is actively engaged in conducting Academic, Research and development programs in the field of Computer Science and IT Engg. Contributed more than 30 research level papers to many national and International journals. and conferences Besides this, published 4 books by reputed publishers. Having research interests include Embedded System, Adhoc Network,Sensor Network ,Power Minimization, Biclustering, Evolutionary Computation and Data Mining.

Dr. Milu Acharya obtained her Ph.D at Utkal University. She is a Professor in Department of Computer Applications at Institute of Technical Education and Research (ITER), Bhubaneswar. She has contributed more than 20 research level papers to many national and International journals and conferences Besides this, published 3 books by reputed publishers. Her research interests include Biclustering,Data Mining , Evaluation of Integrals of analytic Functions , Numerical Analysis , Complex Analysis , Simulation and Decision Theory.

Received February 2010
Accepted March 2010
Final acceptance in revised form April 2010