Reducing and Clustering high Dimensional Data through Principal Component Analysis

R.Indhumathi
H.O.D in M.C.A Department,
Vivekanandha College of Arts and Sciences for Women,
Elayampalayam,
Tiruchengode – 637 205.
Tamil Nadu, India.

Dr.S.Sathiyabama
Research Supervisor,
Professor in M.C.A Department,
K.S.Rangasamy College of Technology, Tiruchengode,
Tamil Nadu, India.

ABSTRACT
High dimensional data is phenomenon in real-world data mining applications. Developing effective clustering methods for high dimensional dataset is a challenging problem due to the curse of dimensionality. Usually k-means clustering algorithm is used but it results in time consuming, computationally expensive and the quality of the resulting clusters depends on the selection of initial centroid and the dimension of the data. The accuracy of the resultant value perhaps not up to the level of expectation when the dimension of the dataset is high because we cannot say that the dataset chosen are free from noisy and flawless. Hence to improve the efficiency and accuracy of mining task on high dimensional data, the data must be pre-processed by an efficient dimensionality reduction method. This paper proposes a method in which the high dimensional data is reduced through Principal Component Analysis and then bisecting k-means clustering is performed on the reduced data where there is no initialization of the centroids.

Keywords
Keywords K-means, Dimensionality Reduction, Principal Component Analysis.

1. INTRODUCTION
Cluster analysis is one of the major data analysis methods which is widely used for many practical applications in emerging areas like bioinformatics. The purpose of clustering is to group together data points, which are close to one another. Many algorithms have been developed for clustering. Existing clustering algorithms face difficulty in handling multidimensional data. The inherent scarcity of the points makes multidimensional data a challenge for data analysis.

The most common problem is the rapid degeneration of performance with increasing dimensions because the approaches are originally designed for low dimensional data. There are many approaches to address high dimensionality problem. Simplest approach is the dimension reduction techniques. In these methods, dimension reduction is carried out as a pre-processing step. The standard k-means algorithm generates extremely imbalanced clusters in high dimensional spaces. The dependency of the k-means performance on the initialization of the centres is a major problem. If the initial partitions are not chosen carefully, 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.

Ideally the centroids are chosen to minimize the total "error," where the error for each point is given by a function that measures the discrepancy between a point and its cluster centroid, e.g., the squared distance. Note that a measure of cluster "goodness" is the error contributed by that cluster. For squared error and Euclidean distance, it can be shown that a gradient descent approach to minimizing the squared error yields the following basic K-means algorithm. However k-Means algorithm has the drawbacks of very time consuming and computationally expensive.

Several attempts were made by researchers for improving the performance of the k-means clustering algorithm. Typically the dimensionality reduction is accomplished by applying techniques from linear algebra or statistics such as Principal Component Analysis. This paper proposes a new approach to reduce the dimension of the data and find cluster using Bisecting K-Means which is better than K-Means where initial centroids is not required.

2. METHODOLOGY

2.1. Principal Component Analysis
The central idea of PCA is to reduce the dimensionality of the data set consisting of a large number of variables .In this paper, PCA is used to reduce the dimension of the data. This is achieved by transforming to a new set of variables (Principal Components) which are correlated and which are ordered so that the first few retain the most of the variant present in all original variables. A mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*, which are the linear combinations of the original variables. To reduce the dimensionality (number of variables) of the dataset but retains most of the original variability in the data.

PCA performs a rotation of the data that maximizes the variance in the new axes. It projects high dimensional data into a low dimensional sub-space (visualized in 2-3 dims).Often captures much of the total data variation in a few dimensions (< 5). Exact solutions require a fully determined system (matrix with full rank) .i.e. A "square" matrix with independent rows.

Principal Component can be defined as a linear combination of optimally weighted observed 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 maximum amount of total variance in the observed variables. The second component extracted will account for a maximal account of variance in the data set that was not accounted for by the first 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.

2.2. Bisecting K-Means Algorithm

Bisecting K-means is the extension of the basic 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 is run k-1 times to get k clusters. This algorithm performs better than regular K means because bisecting K-means produces almost uniform sized clusters. While in regular K-means there can be notable difference between sizes of the clusters. As small cluster tends to have high intra cluster similarity, large clusters have very low intra cluster similarity and overall intra cluster similarity decreases.

Basic Bisecting K-Means algorithm:

1. Pick a cluster to split.
2. Find 2 sub_cluster using the basic k-means algorithm.
3. Repeat step2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 & 3 until the desired numbers of clusters is reached.

The text should be in two 8.45 cm (3.33") columns with a .83 cm (.33") gutter.

3. PROPOSED METHOD

As k-means clustering algorithm often does not work well for high dimension, to improve the efficiency, we proposed to apply PCA on the original data set, to obtain a reduced dataset containing possibly uncorrelated variables. Then the reduced data set will be applied to bisecting k-means clustering algorithm to determine precise number of cluster which overcomes the problem of initialization of centroid to make the algorithm more effective and efficient to determine precise number of cluster.

PCA is a transformation, which transforms the dataset to a new coordinate system such that the greatest variance by any projection of the dataset comes to lie on the first coordinate, this is the principal component [6]. PCA is computed by calculating the covariance matrix of the *n*-dimensional dataset. Covariance is defined as the amount by which dimensions of a dataset vary from the mean with respect to each other. Eigenvectors are found for this covariance matrix. These Eigenvectors describe the patterns and characteristics present in the dataset.

The steps involved in this proposed algorithm are as follows.

Input: $X= \{ x_1,x_2,\ldots\ldots x_n\}$

// set of n-data-points.

K – Number of desired cluster.

Output: Reduced data set

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

1. Obtain the input matrix table.
2. Subtract the mean from the dataset in all the n-dimensions.
3. Calculate the covariance matrix of this mean-subtracted dataset.
4. Calculate the eigenvalues and eigenvectors of the covariance matrix
5. Forming a feature vector by selecting the eigenvector with the largest eigenvalues.
6. Deriving the new data set.

Phase- 2: Apply Bisecting K-means algorithm for newly derived dataset.

1. Reduce the D dimension of the N data using Principal Component Analysis (PCA) and prepare another N data with d dimensions (d<D).

2. The Principal components are ordered by the amount of variance.

3. Choose the first principal component as the principal axis for partitioning and sort it in ascending order.

4. Divide the Set into k subsets where k is the number of clusters.

4. EXPERIMENTAL RESULTS

We evaluated the proposed algorithm on the data sets from UCI machine learning repository [9]. We compared clustering results achieved by the k-means, PCA+k-means with random initialization and initial centers derived by the proposed algorithm.

Table 1. Dataset description

Data Sets	#Samples	#Dimensions	#Number of clusters(k)
Iris	1	4	3
Wine	1	1	3
Glass	2	9	6
ImgSeg	2	1	7

The above data sets are used for testing the accuracy and efficiency of the proposed method for the value of k, given in Table 1.

Table 2. Principal component analysis of iris dataset

Compon	eigenvalue	Accumulation(%)
1	4.224	92.4
2	0.242	97.7
3	0.078	99.4
4	0.023	100.

In this work the number of principal components can be decided by a contribution degree of total variance. Table 2 shows the results obtained by a principal component analysis of the Iris data. This shows that three principal components explained about 99.48% of all data. Therefore, there is hardly any loss of information along a dimension reduction.

The results of the experiment for Iris data set is tabulated in Table 3.

Table 3. Performance comparison on iris data

Algorithm	Initial Centroid	Accuracy (%)
k-means	Random Selection	78.7
k-means+ PCA	Random Selection	85.97
Proposed Method	By split	90.55

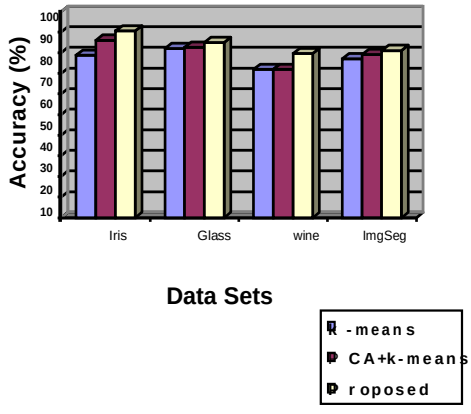

Figure 1. Accuracy on three data sets: Iris, Glass, Wine and ImgSeg

Results presented in Figure 1 demonstrate that the proposed method provides better cluster accuracy than the existing methods. The clustering results of random initial center are the average results over 7 runs since each run gives different results. It shows the proposed algorithm performs much better than the random initialization algorithm.

Table 4 Performance Comparison Based On Time Taken

Dataset	K-Means (Time Taken)(s)	Proposed (Time Taken)(s)
Iris	0.078	0.065
Glass	0.0158	0.0122
Wine	0.0167	0.0131
Imgseg	0.0145	0.0112

In figure 2, we compare the CPU time (seconds) of the proposed method with the existing methods. The execution time of proposed algorithm was much less than the average execution time of k-means when used random initialization.

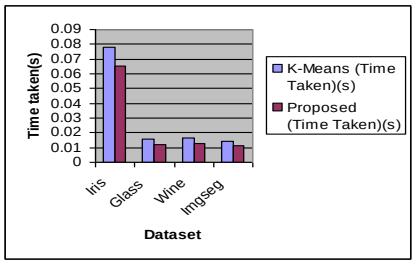

Figure 2: Comparison of time taken

The experimental datasets show the effectiveness of our approach. This may be due to the initial cluster centers generated by proposed algorithm are quite closed to the optimum solution and it also discover clusters in the low dimensional space to overcome the curse of dimensionality.

5. CONCLUSION

In this paper a new approach has been proposed which combines the dimensionality reduction through PCA and a bisecting K-Means algorithm which uses the basic K-Means algorithm. The main objective of applying PCA on original data before clustering is to obtain accurate results. But the clustering results depend on the initialization of centroid. This experiment shows a substantial improvement in running time and accuracy of the clustering results by reducing the dimension and initial centroid selection using PCA. Though it produces better result than the K-Means but when number of document increases, the intra-cluster similarity decreases. A method or algorithm can be taken for further research.

6. ACKNOWLEDGMENTS

I am very grateful to the editors. Also I thank my guide Dr.S.Sathiyabama for many useful and constructive comments and suggestions which helped me significantly improve my work.

7. REFERENCES

[1] Pang-Ning Tang, Michal Steinbach and Vipin Kumar, " Introduction to Data Mining", Pearson Education,Third edition, 2009.

[2] Chris Ding and Xiaofeng He, "K-Means Clustering via Principal Component Analysis",In proceedings of the 21stInternational Conference on Machine Learning, Banff, Canada, 2004

[3] Sandro Saitta, Combining PCA and K-means March 26, 2007 by Filed under: PCA, k-means

[4] Chris Ding and Xiaofeng He ,K-means Clustering via Principal Component Analysis: Proceedings of the twenty-first international conference on Machine learning**,** Page: 29 ,Year of Publication: 2004

[5] Zhang Z., Zhang J. and Xue H.2008.Improved K-means clustering algorithm Proceedings of the congress on Image and signal Processing, Vol.5,n0.5,pp.162-172

[6] Principal component analysis From Wikipedia, the free encyclope

[7] I.T. Jolliffe. Principal Component Analysis. Springer, 2nd edition 2002, ISBN 978-0-387-95442-4.

[8] Rajashree Dash,Debahuti Mishra,Amiya Kumar Rath,Milu Acharya ,A hybridized K- means clustering approach for high dimensional dataset, ,Inertnatioanl Journal of Engineering Science and Technology, Vol.2, No.2, 2010, pp.59-66.

[9] Merz C and Murphy P, UCI Repository of Machine Learning Databases.

[10] A Deterministic Method for Initializing K- Means Clustering, Ting Su,Jennifer Dy, Proceedings of the 16th IEEE International Conference on Tools with Artifical Intelligence, pp.784-786.

[11] Valarrnathie 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.

[12] Sergio M. Savaresi and Daniel L. Boley, On the performance of Bisecting K-Means and PDDP.

[13] N.Tajunisha and V.Saravanan,"An increased performance of clustering high dimensional data using Priniciapl Component Analysis, 2010 First International Conference on Integrated Intelligent Computing" DOI 10.11.09

[14] A k-Means-Based Projected Clustering Algorithm,Yufen Sun,Gang Liy and Kun Xu, 2010 Third International Joint Conference on Computational Science and Optimization, DOI 10.11.09