Improving the Performance of K-Means Clustering For High Dimensional Data Set

P.Prabhu

Assistant Professor in Information Technology
DDE, Alagappa University
Karaikudi, Tamilnadu, India

N.Anbazhagan

Associate Professor
Department of Mathematics
Algappa University,
Karaikudi,Tamilnadu,India

***Abstract t*— Clustering high dimensional data is the cluster analysis of data with anywhere from a few dozen to many thousands of dimensions. Multiple dimensions are hard to think in, impossible to visualize, and, due to the exponential growth of the number of possible values with each dimension, impossible to enumerate. Hence to improve the efficiency and accuracy of mining task on high dimensional data, the data must be preprocessed by efficient dimensionality reduction methods such as Principal Component Analysis (PCA).Cluster analysis in high-dimensional data as the process of fast identification and efficient description of clusters. The clusters have to be of high quality with regard to a suitably chosen homogeneity measure. K-means is a well known partitioning based clustering technique that attempts to find a user specified number of clusters represented by their centroids. There is a difficulty in comparing quality of the clusters produced Different initial partitions can result in different final clusters. Hence in this paper we proposed to use the Principal component Analysis method to reduce the data set from high dimensional to low dimensional. The new method is used to find the initial centroids to make the algorithm more effective and efficient. By comparing the result of original and proposed method, it was found that the results obtained from proposed method are more accurate.**

Keywords : Clustering , k-means; principal component analysis; dimension reduction; initial centroid

I. INTRODUCTION

Data mining is the process of exploration and analysis, by automatic or semi automatic means, of large quantities of data in order to discover meaningful patterns and rules [8]. Data mining can be performed on various types of database and information repositories, but the kind of patterns to be found are specified by various data mining functionalities such as association, correlation analysis, classification, prediction, cluster analysis etc.,

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering[12]. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A good clustering method will produce high quality of clusters with high intra-cluster similarity and low inter-cluster similarity.

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 developed for low dimensional data, often do not work well for high dimensional data and the result may not be accurate most of the time due to outliers. 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. 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.

With high dimensional data, many features are redundant or irrelevant. The redundant features are of no help for clustering. There are many approaches to address this problem. The Simplest approach is dimension reduction technique such as Principal component analysis (PCA) as a preprocessing.

Dimension reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction [1]. As dimensionality increases, query performance in the index structures degrades. Dimensionality reduction algorithms are the only known solution that supports scalable object retrieval and satisfies precision of query results [13]. Feature transforms the data in the high-dimensional space to a space of fewer dimensions [9].The data transformation may be linear, as in principal component analysis (PCA), but any nonlinear dimensionality reduction techniques also exist [14]. To improve the efficiency the noisy and outlier data may be removed we have to reduce the no. of variables in the original data set.

II. METHODOLOGIES

A. Principal Component analysis

Principal Component Analysis (PCA) is an exploratory tool designed by Karl Pearson in 1901[1] .The algorithm was introduce to psychologist in 1933 by H.Hotelling(1). We know that implementing PCA is the equivalent of applying Singular Value Decomposition (SVD) on the covariance matrix of a data set [3,12].PCA is a classical technique; the central idea of PCA is to reduce the dimensionality of the data set consisting of a large number of variables[1]. This is achieved by transforming to a new set of variables (Principal Components) which are uncorrelated and, which are ordered so that the first few retain the most of the variants present in all of the original variables [10]. It 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. The main advantage of PCA is that once we have found these patterns in the data and we can compress the data i.e., by reducing the number of dimensions without much loss of information. It has become an essential one in high dimensional data in order to determine the number of clusters and provides a statistical framework to model the cluster structure.

B. Principal Component

Technically, a principal component (PC) 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. After finding principal components reduced dataset is applied to k-means clustering. 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.

C. The K-Means clustering Algorithm

K-means is the simplest and most popular classical clustering method that is easy to implement. It is algorithm to classify or to group your objects based on attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. It is also called centroid method. The K-means method uses the Euclidean

distance measure, which appears to work well with compact clusters. The psedocode for the k-means clustering algorithm is listed below as algorithm 1[4].

Algorithm 1

Input:

X = {d1, d2,.....,dn} //set of *n* data items.
k // Number of desired clusters

Output:

A set of *k* clusters.

Steps:

1. Arbitrarily choose *k* data-items from X as initial centroids;
2. Repeat
   a) Assign each item *d*i to the cluster which has the closest centroid;
   b) Calculate new mean for each cluster;

Until convergence criteria is met.

The most widely used convergence criteria for the k-means algorithm is minimizing the SSE. The k-means algorithm always converges to a local minimum. The computational complexity of the algorithm is O(nkl), where n is the total number of objects in the dataset, k is the required number of clusters and l is the number of iterations. The time complexity for the high dimensional data set is O(nmkl) where m is the number of dimensions. The major drawback of this algorithm is that it produces different clusters for different sets of values of the initial centroids[4].

III. PROPOSED METHOD

In the modified method discussed in this paper, we used Principal Component Analysis to reduce the dataset and initial cluster centers are derived by median of k partitioned reduced data [6]. The reduced projected data is applied to k-means clustering. Algorithm 2 describes the proposed method.

Algorithm 2 : Proposed Method

Input: *X* = {*d1, d2,……..,dn*} // set of n data items.
Output: A set of *k* clusters

// finding reduced projected dataset Y using Principal Component Analysis

Steps

1. Organize the dataset in a matrix A.
2. Normalize the data set using Z-score. V'= (V-mean (A))/std (A).
3. Calculate the Singular Value Decomposition (SVD) of the data matrix. X =UDV
4. Calculate the coefficients and variance using the diagonal elements of D.
5. Sort variances in decreasing order.
6. Choose the first 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.

// Finding the Initial Centroids

1. For a data set with dimensionality, d, compute the variance of data in each dimension(column).
2. Find the column with maximum variance and call it as max and sort it in any order.
3. Partition the data points into K subsets, where K is the desired number of clusters.
4. Find the median of each subset.

5. Use the corresponding data points (vectors) for each median to initialize the cluster centers.

// Apply K-Means Clustering With Reduced Datasets Y and initial centroid .

Steps:

1. Arbitrarily choose *k* data-items from X as initial centroids;
2. Repeat
   a) Assign each item *d*i to the cluster which has the closest centroid;
   b) Calculate new mean for each cluster;
   Until convergence criteria is met.

IV. EXPERIMENTAL SETUP

In all experiments we use MATLAB software as a powerful tool to compute clusters and windows Vista with Intel Core 2Due CPU 2.8 GHZ with RAM 2.0GB. We evaluated the proposed algorithm on multivariate wine data set taken from UCI machine learning repository (6) for testing the efficiency of the proposed algorithm. The data is the result of a chemical analysis of wines grown in a region in Italy but derived from three different cultivars. There are three classes. The dataset consists of 178 examples each with 13 continuous attributes. The data set contains distribution 59 examples of class 1, 71 examples for class 2 and 48 examples for class 3.

Table 1. The Variances, Variances in Percentages, and Cumulative Variances in Percentages Corresponding to PCs

Principal Component No	Variance	Variance in %	Cumulative Variance
1	4.7059	36.19923	36.20
2	2.4970	19.20769	55.41
3	1.4461	11.12385	66.53
4	0.9190	7.069231	73.60
5	0.8532	6.563077	80.16
6	0.6417	4.936154	85.10
7	0.5510	4.238462	89.34
8	0.3485	2.680769	92.02
9	0.2889	2.222308	94.24
10	0.2509	1.930000	96.17
11	0.2258	1.736923	97.91
12	0.1688	1.298462	99.21
13	0.1034	0.795385	100.00

V. EXPERIMENTAL RESULTS

We compared clustering results achieved by the k-means, PCA+k-means with random initialization and initial centers derived by the proposed algorithm. In existing methods the initial centroid for standard k-means algorithm is selected randomly. The experiment is conducted several times for wine dataset of values of the initial centroids, which are selected randomly. In each experiment, the accuracy was computed and taken the average accuracy of all experiments. In proposed method , PCA is applied to find reduced dataset. The reduced data set is partitioned into k sets and median of each set is obtained and used as initial cluster centers and then assigned each data points to its nearest cluster centroid. Table 1 shows variance, percentage of variance and cumulative variance obtained using PCA.

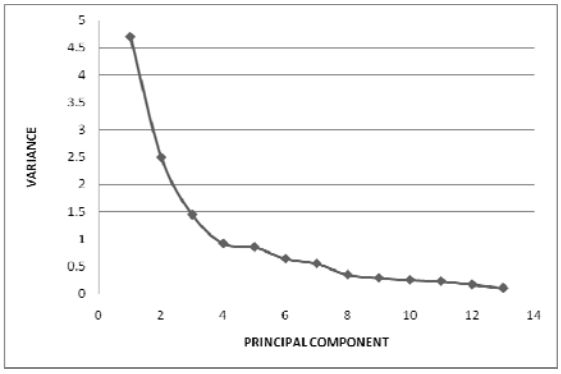

Figure 1. Variance of Principal Component

The Figure1 shows the results obtained by a principal component analysis of the wine dataset. It shows that the first three principal components are having more variance than the others.

The following figure 2 shows the cluster results for wine dataset using the proposed method with k=3 clusters.

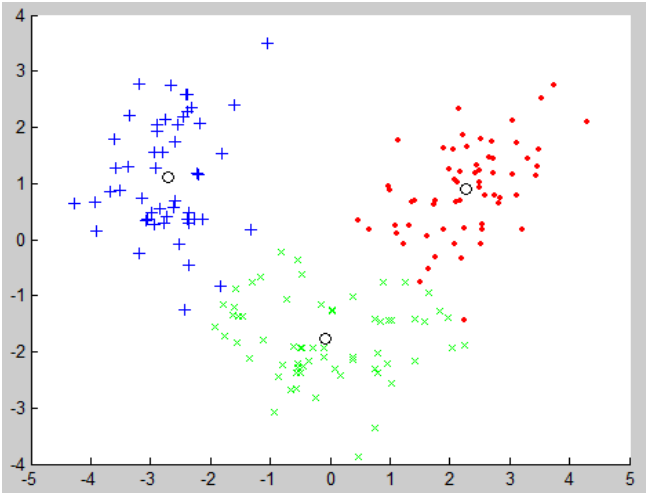

Figure 2 clusters of size k =3 of wine data set using Proposed Method.

The following table 2 show the performance comparison on wine data set with different methods with k = 3 number of clusters.

Table 2. Performance comparison on wine data set

No.of Clusters	Method	Run	Initial Centroid	Accuracy (%)
Wine dataset k=3	k-means	10	Random	87.64
	Kmeans + PCA	10	Random	88.76
	Proposed	1	Proposed	92.13

Results presented in Figure 3 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 10 runs since each run gives different results. It shows the proposed algorithm performs much better than the random initialization algorithm.

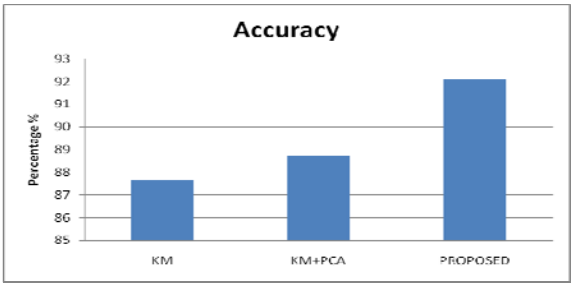

Figure 3. Accuracy on Wine dataset

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.

CONCLUSION

In this paper a new approach has been proposed which combines the dimensionality reduction through PCA and new initialization method is used to improve the performance of clustering. 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 improvement in accuracy of the clustering results by reducing the dimension and improved initial centroid by partitioning projected dataset and finding the median as centroids. Several experiments may be conducted to test the performance with different dimensions with different initial centroid methods for better accuracy as future work.

REFERENCES

[1] Chris Ding and Xiaofeng He, "K-Means Clustering via Principal Component Analysis", In proceedings of the 21st InternationalConference on Machine Learning, Banff, Canada, 2004.
[2] Hotelling, H., Analysis of a complex of statistical variable into principal components; J. Educ. Psych., vol. 24, 417-441 (1933).
[3] Jiawei Han, Micheline Kamber, "Data Mining concepts and Techniques", Morgan Kaufmann Publishers,San Fracisco,CA,USA.
[4] K.A.Abdul Nazeer,M.P.Sebastian "Improving the Accuracy and Eficiency of the k-means clustering algorithm" Proceeding sot he world congress on Engineering vol I WCE2009,July 1-3,2009,London,U.K 2009.
[5] Kohei Arai* and Ali Ridho Barakbah Hierarchical K-means: an algorithm for centroids initialization for K-means,Reports of the Faculty of Science and Engineering,Saga University, Vol. 36, No.1, 2007
[6] Lindsay I Smith A tutorial on Principal Components Analysis; February 26, 2002
[7] Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available:ftp://ftp.ics.uci.edu/pub/machine-Learning-databases
[8] Michael J.A.Berry Gordon Linoff,Mastering Data Mining" John wiley & sons ptd,Ltd,Singapore 2001.
[9] Moth'd Belal. Al-Daoud ,A New Algorithm for Cluster Initialization, World Academy of Science, Engineering and Technology. (2005)
[10] Pang-Ning Tang, Michal Steinbach and Vipin Kumar, " Introduction to Data Mining", Pearson Education, Third edition, 2009.
[11] Rajashree Dash, Debahuti Mishra, Amiya Kumar Rath, Milu Acharya: A hybridized k-means clustering approach for high dimensional dataset, International Journal of Engineering, Science Technology, Vol. 2, No. 2,pp. 59-66 2009.
[12] Shlens, J.; A tutorial on Principal Component Analysis; (2003)
[13] Xu R. and Wunsch D, Survey of clustering algorithms, IEEE Trans. Neural Networks, ,Vol. 16, No. 3, pp. 645-678 2005.
[14] Yeung Ka Yee and Ruzzo Walter L., 2000. An empirical study on principal component analysis for clustering gene expressionData",Tech. Report, University of Washington.

AUTHORS PROFILE

P.PRABHU received the Bachelor's Degree in B.Sc Computer Science from Madurai Kamaraj University in 1990, Master's Degree in Computer Applications from Bharathiar University in 1993, and M.Phil degree in Computer Science from Bharthidasan University in 2005.He is working as Assistant Professor in Information Technology, Directorate of Distance Education, Alagappa University, Karaikudi, Tamilnadu, India. He has published articles in National and International Journals. He has presented papers in National and International Conferences. His research area is Data Mining and Computer Networks.

N. Anbazhagan is currently Associate Professor of Mathematics in Alagappa University, Karaikudi, India. He received his M. Phil and Ph.D in Mathematics from Madurai Kamaraj University, Madurai, India and M.Sc in Mathematics from Cardamom Planters Association College, Bodinayakanur, India. He has received Young Scientist Award (2004) from DST, New Delhi, India, Young Scientist Fellowship (2005) from TNSCST, Chennai, India and Career Award for Young Teachers(2005) from AICTE, India. He has successfully completed one research project, funded by DST, India. His research interests include Stochastic modeling, Optimization Techniques, Inventory and Queueing Systems. He has published the research articles in several journals, including Stochastic analysis and applications, APJOR and ORiON.