Performance analysis of k-means with different initialization methods for high dimensional data

Tajunisha[1] and Saravanan[2]

[1]Department of Computer Science, Sri Ramakrishna College of Arts and Science (W), Coimbatore, Tamilnadu, India
tajkani@gmail.com
[2]Department of Computer Application, Dr. N.G.P. Institute of Technology, Coimbatore, Tamilnadu, India
tvsaran@hotmail.com

ABSTRACT :

Developing effective clustering method for high dimensional dataset is a challenging problem due to the curse of dimensionality. Among all the partition based clustering algorithms, k-means is one of the most well known methods to partition a dataset into groups of patterns. However, the k-means method converges to one of many local minima. And it is known that, the final result depends on the initial starting points (means). Many methods have been proposed to improve the performance of k-means algorithm. In this paper, we have analyzed the performance of our proposed method with the existing works. In our proposed method, we have used Principal Component Analysis (PCA) for dimension reduction and to find the initial centroid for k-means. Next we have used heuristics approach to reduce the number of distance calculation to assign the data point to cluster. By comparing the results on iris data set, it was found that the results obtained by the proposed method are more effective than the existing method.

KEYWORDS:

k-means, principal component analysis, dimension reduction, initial centroid.

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 functionality like class 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 to one another and different from the objects in other groups. A good clustering method will produce high quality clusters with high intra-cluster

DOI : 10.5121/ijaia.2010.1404

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.

Clustering is a common unsupervised learning technique used to discover the natural groups of similar objects represented by vectors of measurements, in multidimensional spaces. A clustering algorithm typically considers all features of the data in an attempt to learn as much as possible about the objects. However, with high dimensional data, many features are redundant or irrelevant. The redundant features are of no help for clustering; even worse, the irrelevant features may hurt the clustering results by hiding clusters in noises. There are many approaches to address this problem. The simplest approach is dimension reduction techniques including principal component analysis (PCA) and random projection. In these methods, dimension reduction is carried out as a preprocessing step.

The standard k-means algorithm [5, 8] for cluster analysis developed for low dimensional data, often do not work well for high dimensional data and the results may not be accurate most of the time due to noise. Different methods have been proposed [1] by combining PCA with k-means for high dimensional data. But the accuracy of the k-means clusters heavily depending on the random choice of initial centroids.

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. To combat 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 being very time consuming and computationally expensive.

In this work, initial centers are determined using PCA and k-means method is modified by using heuristic approach to assign the data-point to cluster. This paper compares the results of existing methods with the proposed method on iris dataset.

2. K-MEANS CLUSTERING ALGORITHM

K-means is a prototype-based, simple partitional clustering technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids. A cluster centroid is typically the mean of the points in the cluster. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice. The algorithm consist of two phases: the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to nearest centroid. The k-means algorithm works as follows:

a) Select initial centroid of the k clusters. Repeat steps b through c until the cluster membership stabilized.

b) Generate a new partition by assigning each data to its closest cluster centoid.

c) Compute new cluster centroid for each cluster.

The most widely used convergence criteria for the k-means algorithm is minimizing the SSE.

$$SSE = \sum_{j=1}^{k} \sum_{x_i \in c_j} \left\| x_i - \mu_j \right\|^2 \quad \text{Where } \mu_j = \frac{1}{n_j} \sum_{x_i \in c_j} x_i$$

Denotes the mean of cluster c_j and n_j denotes the no. of instances in c_j.

The k-means algorithm always converges to a local minimum. The particular local minimum found depends on the starting cluster centroids. The k-means algorithm updates cluster centroids till local minimum is found. Before the k-means algorithm converges, distance and

centroid calculations are done while loops are executed a number of times, say l, where the positive integer l is known as the number of k-means iterations. The precise value of l varies depending on the initial starting cluster centroids even on the same dataset. So 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.

3. PRINCIPAL COMPONENT ANALYSIS

In general, handling high dimensional data using clustering techniques obviously a difficult task in terms of higher number of variables involved. In order to improve the efficiency, the noisy and outlier data may be removed and minimize the execution time and we have to reduce the no. of variables in the original data set. Principle Component Analysis is a common technique for finding patterns in high dimensional data [6].

The central idea of PCA is to reduce the dimensionality of the data set consisting of a large number of variables. 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 steps involved in PCA are

Step1: Obtain the input matrix Table

Step2: Subtract the mean

Step3: Calculate the covariance matrix

Step4: Calculate the eigenvectors and eigenvalues of the covariance matrix

Step5: Choosing components and forming a feature vector

Step6: deriving the new data set.

The eigenvectors with the highest eigenvalue is the principal component of the data set. In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. To reduce the dimensions, the first d (no. of principal components) eigenvectors are selected. The final data has only d dimensions.

The main objective of applying PCA on original data before clustering is to obtain accurate results so that the researchers can do analysis in better way. Secondly, minimize the running time of a system because time taken to process the data is a significant one. Normally it takes more time when the number of attributes of a data set is large and sometimes this dataset not supported by all the clustering techniques hence the number of attributes are directly proportional to processing time. 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 uncorrelated and, which are ordered so that the first few retain the most of the variant present in all of the original variables. In our work, The first Principal Component is selected to find the initial centroid for the clustering process.

4. EXISTING METHODS

There is no commonly accepted or standard “best” way to determine either the no. of clusters or the initial starting point values. The resulting set of clusters, both their number and their centroids, depends on the specified choice of initial starting point values. Two simple approaches to cluster initialization are 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 and set, which is closest to optimal, is chosen. However, testing different initial sets are considered impracticable criteria, especially for large number of clusters. Therefore different methods have

been proposed to find the initial centroid given in literature [11]. In this paper, we have presented the related works presented on iris dataset.

Fahim A M et al. [2] proposed an efficient method for assigning data points to clusters. The original k-means algorithm is computationally very expensive because each iteration computes the distances between data points and all the centroids. Fahim's approach makes use of two distance functions for this purpose- one similar to k-means algorithm and another one based on a heuristics to reduce the number of distance calculations. But this method presumes that the initial centroids are determined randomly, as in the case of the original k-means algorithm. Hence there is no guarantee for the accuracy of the final clusters. In 2009, Fahim A M et al. [3] 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.

Nazeer et al. (2009) proposed [10] an enhanced k-means to improve the accuracy and efficiency of the k-means clustering algorithm. In this algorithm two methods are used, one method for finding the initial centroids and another method for an efficient way of assigning data points to appropriate clusters. This algorithm to a fine the initial centroid is described as follows: initially, 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. To assign the data points to cluster centroid, they have used heuristics approach.

Madhu Yedla et al. (2010) proposed [7] method to find the better initial centroids with reduced time complexity. For assigning the data points they followed the paper [2][10]. In this method first check the given dataset contain the negative value attributes or not. Then transform all the data points in the dataset to the positive space by subtracting the each data point attribute with the minimum attribute value in the given data set. Next, calculate the distance for each data point from the origin. Sort the datapoints accordance with the distances. And partition the sorted datapoint into k equal sets and the middle point in each set is taken as initial centroid. In this paper they have used heuristics approach to assign the datapoints to initial centroid.

But all the above methods do not work well for high dimensional data sets. In our proposed work [12], the new approach was proposed to reduce the dimension and to find the initial centroid using PCA. To improve the efficiency of our method we have used heuristics approach to reduce the number of distance calculation in the standard k-means algorithm. In this paper we have compared and analyzed the results of proposed method with the existing methods on iris datasets.

5. PROPOSED METHOD

The proposed method that performs data partitioning with Principal component. It partitions the given data set into k sets. The median of each set can be used as good initial cluster centers and then assign each data points to its nearest cluster centroid. The Proposed model is illustrated in Figure 1.

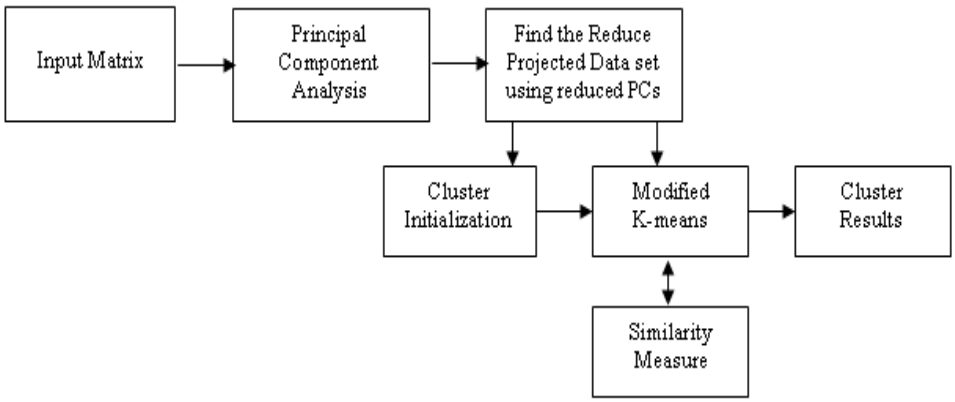

Figure.1 Proposed Model

In this method Following are the steps of the proposed algorithm.

Algorithm 1: The proposed method
Steps:

1. Reduce the dimension of the data into d dimension and determine the initial centroid of the clusters by using Algorithm 2.
2. Assign each data point to the appropriate clusters by using Algorithm 3.

In the above said algorithm the data dimensions are reduced and the initial centroids are determined systematically so as to produce clusters with better accuracy.

Algorithm 2: Dimension reduction and finding the initial centroid using PCA.
Steps:

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.
5. Find the median of each subset.
6. Use the corresponding data points for each median to initialize the cluster centers.

The initial centroids of the clusters are given as input to Algorithm 3. It starts by forming the initial clusters based on the relative distance of each data-point from the initial centroids. The Euclidean distance is used for determining the closeness of each data point to the cluster centroids. For each data-point, the cluster to which it is assigned and its distance from the centroid of the nearest cluster are noted. For each cluster, the centroids are recalculated by taking the mean of the values of its data-points. The procedure is almost similar to the original k-means algorithm except that the initial centroids are computed systematically.

The next stage is an iterative process which makes use of a heuristic method to improve the efficiency. During the iteration, the data-points may get redistributed to different clusters. The method involves keeping track of the distance between each data-point and the centroid of its present nearest cluster. At the beginning of the iteration, the distance of each data-point from the new centroid of its present nearest cluster is determined. If this distance is less than or equal to the previous nearest distance, that is an indication that the data point stays in that cluster itself and there is no need to compute its distance from other centroids. This result in the saving of time

required to compute the distances to k-1 cluster centroids. On the other hand, if the new centroid of the present nearest cluster is more distant from the data-point than its previous centroid, there is a chance for the data-point getting included in another nearer cluster. In that case, it is required to determine the distance of the data-point from all the cluster centroids. This method improves the efficiency by reducing the number of computations.

Algorithm 3: Assigning data-points to clusters

Steps:

1. Compute the distance of each data-point x_i (1<=i<=n) to all the centroids c_j(1<=j<=k) using Euclidean distance formula..
2. For each data object xi, find the closest centroid cj and assign x_i to the cluster with nearest centroid c_j and store them in array Cluster[ ] and the Dist[ ] separately.
   Set Cluster[i] = j, j is the label of nearest cluster.
   Set Dist[i]= d(xi, cj), d(xi, cj) is the nearest Euclidean distance to the closest center.
3. For each cluster j (1<=j<=k), recalculate the centroids;
4. Repeat
5. for each data-point
   5.1 Compute its distance from the centroid of the present nearest cluster
   5.2 If this distance is less than or equal to the previous nearest distance, the data-point stays in the cluster
   Else
   For every centroid c_j
   Compute the distance of each data object to all the centre
   Assign the data-point x_i to the cluster with nearest centroid c_j
6. For each cluster j (1<=j<=k), recalculate the centroids;

Until the convergence criteria is met.

This algorithm requires two data structure Cluster[ ] and Dist[ ] to keep the some information in each iteration which is used in the next iteration. Array cluster[ ] is used for keep the label if the closest centre while data structure Dist[ ] stores the Euclidean distance of data object to the closest centre. The information in data structure allows this function to reduce the number of distance calculation required to assign each data object to the nearest cluster, and this method makes the improved k-means algorithm faster than the standard k-means algorithm.

6. EXPERIMENTAL RESULTS

We evaluated the proposed algorithm on iris 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 given in Table 4. Table 2 shows the results obtained by paper [10]. Table 3 shows the results obtained by paper [7]. The initial centroid for standard k-means algorithm is selected randomly. The experiment is conducted 7 times for different sets of values of the initial centroids, which are selected randomly. In each experiment, the accuracy and time was computed and taken the average accuracy and time of all experiments.

TABLE 1. PRINCIPAL COMPONENT ANALYSIS OF IRIS DATASET

Component	eigenvalue	Accumulation(%)
1	4.2248	92.46
2	0.2422	97.76
3	0.0785	99.48
4	0.0237	100.00

In the proposed work the number of principal components can be decided by a contribution degree about total variance. Table 1 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.

Table 2. Performance comparison on iris data sets by paper[10]

No. of Cluster	Algorithm	Run	Accuracy	Time Taken(ms)
K=3	k-means	7	78.7	70.7
	Enhanced method	1	88.6	67

Table 3. Performance comparison on iris data sets by paper[7]

No. of Cluster	Algorithm	Run	Accuracy	Time Taken(sec)
K=3	k-means	7	63.14	0.096
	Improved method	1	88.66	0.086

Table 4 Performance comparison on iris data sets by proposed method

No. of Cluster	Algorithm	Run	Accuracy	Time Taken(sec)
K=3	k-means	7	78.7	0.13
	k-means + PCA	7	85.97	0.12
	Proposed Method	1	90.55	0.08

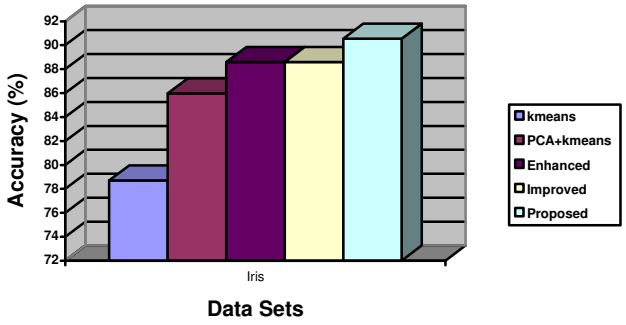

Figure 2. Performance comparison on iris data sets

Results presented in Figure 2 demonstrate that the proposed method provides better cluster accuracy than the existing methods. It shows the proposed algorithm performs much better than

the random initialization algorithm and other author's initialization method. The experimental dataset 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.

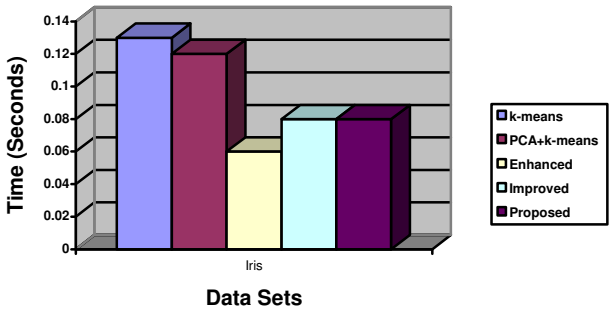

Figure 3. Execution time results on data sets: Iris, Glass, Wine and ImgSeg

In figure 3, 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. Our proposed method provides higher accuracy than the other author's method and takes moreover equal time.

7. CONCLUSION

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. Our proposed method finding the initial centroid and cluster the data in low dimensional space. In this paper, we have analyzed the performance of our proposed method with the existing works. By comparing the results on iris data set, it was found that the results obtained by the proposed method are more accurate and efficient compared to the existing method. In our future work, we will apply our proposed method to microarray cancer datasets.

REFERENCES

[1] 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.

[2] Fahim A.M,Salem A.M, Torkey A and Ramadan M.A (2006) : An Efficient enchanced k-means clustering algorithm,Journal of Zhejiang University,10(7): 1626-1633,2006.

[3] 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.

[4] Ismail M. and Kamal M. (1989): Multidimensional data clustering utilization hybrid search strategies,Pattern Recognition Vol. 22(1),PP. 75-89.

[5] Jiawei Han M.K (2006): Data mining Concepts and Techniques, morgan Kaufmann publishers, An imprint of Elsevier.

[6] Jolliffe I.T. (2002): Principal Component Analysis, Springer, Second edition.

[7] 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.

[8] Margaret H.Dunham (2006): Data Mining-Introductory and Advanced Concepts, Pearson Education.

[9] Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available: ftp://ftp.ics.uci.edu/pub/machine-Learning-databases

[10] 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.

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

[12] Tajunisha N., Saravanan V.,"An increased performance of clustering high dimensional data using Principal Component Analysis",Proceedings of the IEEE first international conference on integrated intelligent computing pp 17-21,(2010).