A Nearest Grouping Around The Stable Centroid Technique For Addressing Scalability Issues In Clustering High Dimensional Data

T. Kamalavalli[1], S. Vinothini[2] and Dr. P. Jaganathan[3]

[1]Associate Professor in PSNA College of Engg.& Tech, India.

[2]Associate Professor in PSNA College of Engg.& Tech, India.

[3] Professor in PSNA College of Engg.& Tech, India.

Abstract

K-means algorithm is the most popular and partition based clustering algorithm. It is computationally expensive and the quality of resulting clusters heavily depends on the selection of initial centroid and the dimension of the data [10]. Several methods have been proposed in the literature for improving performance of the k-means clustering algorithm. In this paper, we propose a Nearest Group Around the Centroids method to make the clustering more effective and efficient by using PCA. The performance is compared with k-means algorithm and the results obtained are more effective, easy to understand and above all, the time taken to process the data is substantially reduced

Keywords: k-means, principal component analysis, dimension reduction.

1. INTRODUCTION

Data mining is the process of discovering non-trivial, previously unknown and potentially useful information from large volumes of data[2]. Data mining is an important method for extracting valuable information from all sizes of databases: large and small. Data miners need fast response time when building their models so they can construct the most effective model. Additionally, most algorithms rely on a set of initial parameters that are hard to be capture and tuned. Most existing data mining techniques operate in a batch mode where it is necessary to have all of the relevant data at once.

Clustering of very large high dimensional data sets is an important problem[3]. There are a number of different clustering algorithms that are applicable to very large data sets, and a few that address high dimensional data.

Scalability remains a significance issue for large scale datasets[1]. Data mining applications place the following two primary requirements on clustering algorithms: Scalability to large dataset[4] and non-presumption of any canonical data properties like convexity. Many clustering algorithms generate accurate clusters on small data sets with limited dimensions; because those algorithms were initially developed for applications where accuracy was more important than speed. Algorithmic complexity of most of these algorithms is of the order of O(n) or O(n2). In order for these algorithms to work effectively in a database application with large data sets, they must be made scalable. Scanning through millions of records more than once is an expensive option. Besides computational cost, it is not possible to fit such large amount of data in memory so as to make all those comparisons. It is important to find an algorithm with a complexity of less than O(n2) and ideally O(n) that could work on a limited buffer at a given time. However, the scalability of data mining techniques is very important due to the rapid growth in the amount the data.

Principal Component analysis is a preprocessing stage of data mining and machine learning, dimension reduction not only decreases computational complexity, but also significantly improves the accuracy of the learned models from large data sets. PCA[9] is a classical multivariate data analysis method that is useful in linear feature extraction. Without class

labels it can compress the most information in the original data space into a few new features, i.e., principal components. 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 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.

2. RELATED WORK

Several algorithms for the improvement of scalability in clustering have been reported in the literature as given below. Ayaz et al. [5,6] proposed that scalability in k-means can be achieved by identifying compressible and discardable sets of objects in clusters. An object is discardable if its membership for a given cluster is confirmed i.e it lies at the center of that cluster. An object is compressible if it belongs to tight subcluster within a cluster. A separate data structure must be maintained that could keep statistics and clustering features about discardable and compressible objects identified in each iteration. By employing a buffer where data objects are saved in compressed form and this effectively decreases the number of scans of entire dataset.

For improving the performance and efficiency of k-means clustering, various and numerous methods have been proposed [11]. A hybridized K-Means clustering approach for high dimensional data set was proposed by Dash, et al.[7] and in his paper he used PCA for dimensional reduction and for finding the initial centroids a new method is employed that is by finding the mean of all the data sets divided in to k different sets in ascending order. This approach stumble, when time complexity is taken into account and it may eliminates some of the features which are also important for explicit extraction of information.
For improving the performance of K-Means clustering M Yedla et al[8] proposed an enhanced K-Means algorithm with improved initial center by the distance from the origin. The approach seems to solve the initialization problem but does not give any guarantee regarding the performance of the algorithm in terms of Time complexity and other matters.

3. PROPOSED METHOD

K-means Clustering is an important algorithm for identifying the structure in data. K-means is the simplest clustering algorithm. This algorithm uses predefined number of clusters as input. The original algorithm is based on random selection of cluster centers and iteratively improving the results. First, the need for number of clusters in advance, is difficult since the underlying structure is not known. Second selection of cluster centers randomly in local optima. In this work taking all the nearest records around the cluster center (centroid) are examined whether they are closer to cluster center or not(based on proper constraint). If it, group the records around the centroid. If not examine with other centroids to identify its group. In this approach, no need to move the centroid means no need to calculate the distance vector each time. The proposed solution, 'Nearest Grouping Around the stable Centroid' is tested on both row store and column store databases.

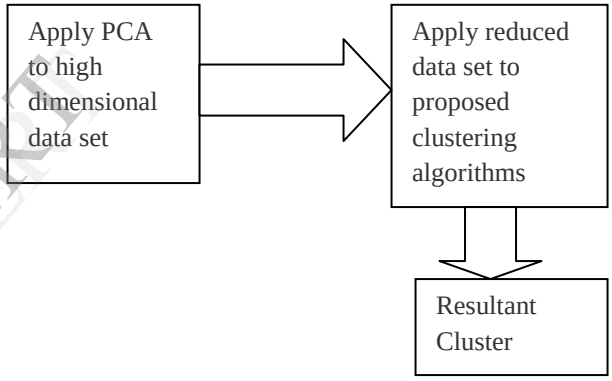

Figure :1 Working model of NGAC

The proposed algorithm for Nearest Grouping Around the stable Centroid is

D={d1,d2,d3,….dn} //set of n data points

K -number of desired clusters.

Discard set // Points that are unlikely to change membership.

Steps:

1. Initialize k centroids from D
2. Group the data points around the centroid[cluster]
3. Examine the distance of each data points with the centroids for k clusters

4. For each k cluster, data points are unlikely to change membership are removed from the cluster and are placed in discardset.
5. For each k cluster remaining data points in the cluster are examined with the centroids of k-1 cluster
6. If the data set is exhausted, then finish. Otherwise repeat from step 3

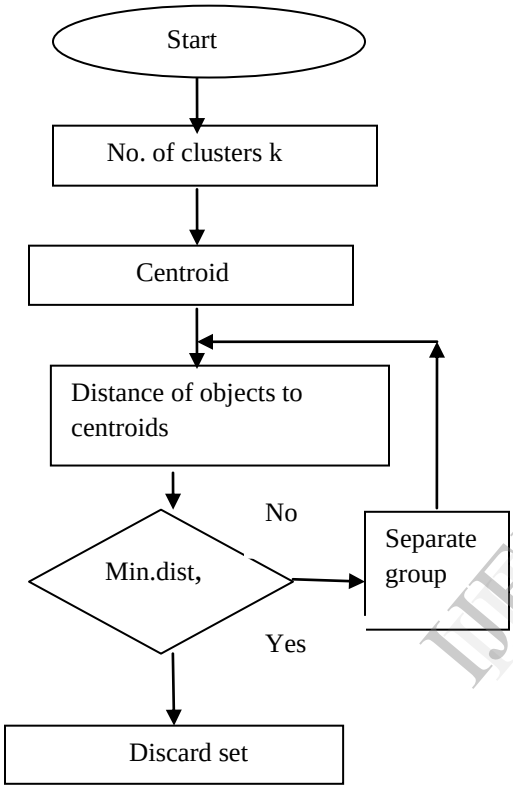

Figure 2: Block diagram of Nearest group around stable centroid Algorithm (NGAC)

4. EXPERIMENTAL RESULTS

The proposed algorithm is developed in java language and the high dimensional data set is preprocessed by Principal Component Analysis (PCA) using Weka. The reduced data set is given to the proposed algorithm for better clustering in terms of time. The scalability in terms of time is measured for various data sets like the breast cancer dataset, iris dataset are obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. The experiment is repeated for a number of times and the results are shown in table 1 and figure 3.

Table 1: Performance Analysis

S.No	Data Set	Time Taken(Sec)	
		K-Means with PCM	NGAC with PCM
1`	Iris	0.03	0.01
2	Breast Cancer	0.04	0.02
3	Parkinsons	0.04	0.03

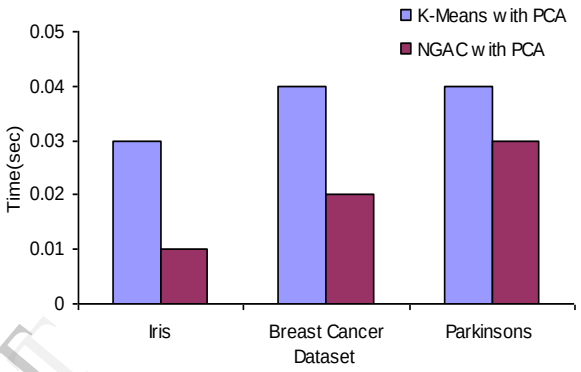

Figure 3: Performance analysis for various dataset

5. CONCLUSION

In this paper, we have proposed a new approach for data clustering. This approach reduces the overhead of finding the mean of cluster center each time in k-Means. The proposed method improves the scalability by means of having a fixed cluster center. This approach ensure that the total mechanism of clustering in time without loss of correctness of clusters. Now the problem of determining the number of clusters before hand is still worth working for and also certain area of improvement for dimension reduction and outlier elimination is still to be explored and unveiled.

6. REFERENCES

[1] Huajing Li,Zaiquing Nie, Wang-Chien Lee,"Scalable community Discovery on Textual Data with Relations.[http://www.ics.uci.edu/~mlearn/MLRepository.html] Irvine, CA: University of California, Department of Information and Computer Science.

[2] J. Han and M. Kamber (2001), "Data Mining: Concepts and Techniques" San Francisco, Morgan Kauffm ANN Publishers.
[3] F. Fredrik L. James E. Charles. Scalability of clustering algorithms revisited. ACM SIGKDD 2000,
[4] Tian Zhang , Raghu Ramakrishnan , Miron Livny., "Birch: An efficient data clustering method for very large databases"1996.
[5] H. Greg and E. Charles,"Alternatives to the k-means algorithm for better clusterings." CIKM '02 Proceedings of the eleventh international conference on CIKM, 2002.
[6] Charles Elkan. Using the triangle inequality to accelerate k-means. In 20th International Conferenceon Machine Learning (ICML-2003), Washington, DC, 2003.
[7] Dash et.al , "A Hybridized k-Means Clustering Algorithm for High Dimensional Dataset", International Journal of Engineering, vol. 2, No. 2, pp.59-66,2010.
[8] M Yedla et al[6]. "Enhancing K means algorithm with improved initial center", (IJCSIT) International Journal of Computer Science logies, Vol. 1 (2) , pp- 121-125,2010.
[9] Jolliffe I.T. (2002): Principal Component Analysis, Springer, Second edition
[10]Nazeer, Kumar, Sebastian, "Enhancing the K-means Clustering Algorithm by Using a O(n logn) Heuristic Method for Finding Better Initial Centroids ", Emerging Applications of Information Technology (EAIT), 2011, Page(s): 261 – 264.
[11] JuntaoWang ,Xiaolong Su , "An improved K-Means clustering algorithm", Communication Software and Networks (ICCSN)- 2011, Page(s): 44 – 46.