INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

AN IMPROVED CLUSTERING ALGORITHM FOR BIG DATA ANALYSIS

T. Kamalavalli[1], S. Vinothini[2], R. Kuppuchamy[3]

[1,2]Assistant Professor,kamali.ajayan@gmail.com, vinothi@psnacet.edu.in
[3] Associate Professor, rkuppuchamy@psnacet.edu.in
Author Correspondence: Department of MCA, PSNA College of Engineering and Technology,
Dindigul- 624622, Tamilnadu

Abstract: - Big Data concerns large-volume, complex, growing data sets with multiple, autonomous Sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Clustering is a well known technique in identifying intrinsic structures and find out useful information from large amount of data. In this paper, 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*:* Big Data, Clustering, Principal Component Analysis, K-Means

1. Introduction

Big Data concerns large-volume, complex, growing data sets with multiple, autonomous Sources [4]. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Every day, 2.5 quintillion bytes of data are created and 90% of the data in the world today were produced within the past two years (IBM 2012). Our capability for data generation has never been so powerful and enormous ever since the invention of the Information Technology in the early 19th century. Existing methods are incapable of handling this Big Data. As a result, the unprecedented data volumes require an effective data analysis algorithms and prediction platform to achieve fast-response and real-time classification and clustering.[5]

Clustering is the fundamental operation in data mining, which groups similar objects into classes or clusters. 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 [1]. Clustering has been used in many application domains, including biology, medicine, anthropology and economics. Clustering applications include plant and animal classification, disease classification, image processing, pattern recognition and document retrieval. One of the first domain in which clustering was used was biological taxonomy. Recent uses include examining web log data to detect the usage patterns [2].

Many clustering algorithms exist in the literature. The major clustering methods can be classified into the following categories [1].

Partitioned Methods
Hierarchical Methods
Density Based Methods
Grid Based Methods
Model Based Methods

Among these, Partition based methods are simple and mostly investigated by many researchers to improve its performance. This paper investigates partition algorithm especially K-means algorithm. A partitioning method initially creates k partitions, called clusters, from given set of n data objects. The parameter K indicates number of partitions to construct. It then uses an iterative relocation technique that attempts to improve the partitioning by moving the objects from one group to another. The general criterion of good partitioning is that objects in the same clusters are “close” or related to each other, whereas objects of different clusters are “far apart” or very different.

K-Means Algorithm

The K-Means clustering algorithm is proposed by Mac Queen in 1967 which is a partition-based cluster analysis method. It has been widely used in cluster analysis because of its higher efficiency and scalability and converges fast when dealing with large data sets. This algorithm partitions the data into K clusters (C1, C2,….CK), represented by their centers or means. The center of each cluster is calculated as the mean of all the instances belonging to that cluster.
Figure 1 presents the pseudo-code of the K-means algorithm. The algorithm starts with an initial set of cluster centers, chosen at random or according to some heuristic procedure. In each iteration, each instance is assigned to its nearest cluster center according to the Euclidean distance between the two. Then the cluster centers are re-calculated.

```
Input: S (instance set), K (number of cluster)
Output: clusters
1: Initialize K cluster centers.
2: while termination condition is not satisfied do
3: Assign instances to the closest cluster center.
4: Update cluster centers based on the assignment.
5: end while
```

Figure 1: K-Means Algorithm

The center of each cluster is calculated as the mean of all the instances belonging to that cluster:

Where Nk is the number of instances belonging to cluster k and μk is the mean of the cluster k.
A number of convergence conditions are possible. For example, the search may stop when the partitioning error is not reduced by the relocation of the centers. This indicates that the present partition is locally optimal. Other stopping criteria can be used also such as exceeding a pre-defined number of iterations.

Principal Component analysis is a pre-processing 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[6] 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. [7,8] 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 sub-cluster 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 kmeans 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.[9] 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[10] 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.

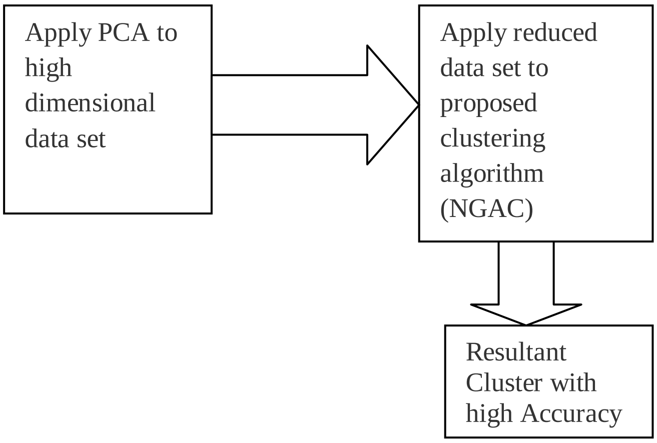
Apply PCA to high dimensional data set
Apply reduced data set to proposed clustering algorithm (NGAC)
Resultant Cluster with high Accuracy

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 discard set.
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.

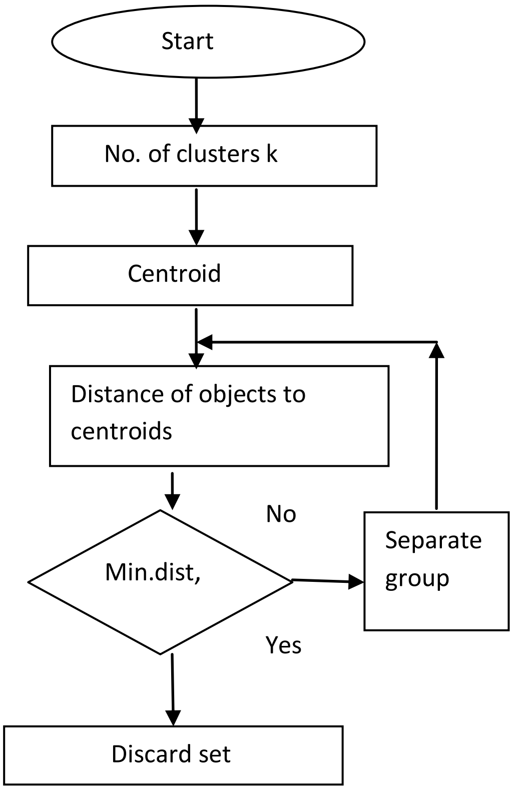

Figure 2: Block diagram of nearest group around stable centroid Algorithm (NGAC)

3. RESULT AND DISCUSSION

Evaluation and interpretation are two vital operators of the output. Evaluation typically plays the role of measuring the results. It can also be one of the operators for the data mining algorithm, such as the sum of squared errors which was used by the selection operator of the genetic algorithm for the clustering problem [12].

To solve the data mining problems that attempt to classify the input data, two of the major goals are: (1) cohesion—the distance between each data and the centroid (mean) of its cluster should be as small as possible, and (2) coupling—the distance between data which belong to different clusters should be as large as possible. In most studies of data clustering or classification problems, the sum of squared errors (SSE), which was used to measure the cohesion of the data mining results, can be defined as

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} D(x_{ij} - c_i),$$

$$c_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij},$$

where k is the number of clusters which is typically given by the user; nini the number of data in the ith cluster; xijxij the jth datum in the ith cluster; cici is the mean of the ith cluster; and n=∑ki=1nin=∑i=1kni is the number of data. The most commonly used distance measure for the data mining problem is the Euclidean distance, which is defined as

$$D(p_i, p_j) = \left(\sum_{l=1}^{d} |p_{il}, p_{jl}|^2 \right)^{1/2},$$

where pi and pj are the positions of two different data. For solving different data mining problems, the distance measurement D(pi,pj)D(pi,pj) can be the Manhattan distance, the Minkowski distance, or even the cosine similarity [13] between two different documents.

Accuracy (ACC) is another well-known measurement [14] which is defined as

$$\mathrm{ACC} = \frac{\text{Number of cases correctly classified}}{\text{Total number of test cases}}.$$

A useful graphical user interface is another way to provide the meaningful information to an user. As explained by Shneiderman in [15], we need “overview first, zoom and filter, then retrieve the details on demand”. The useful graphical user interface [38, 41] also makes it easier for the user to comprehend the meaning of the results when the number of dimensions is higher than three. How to display the results of data mining will affect the user’s perspective to make the decision. For instance, data mining can help us find “type A influenza” at a particular region, but without the time series and flu virus infected information of patients, the government could not recognize what situation (pandemic or controlled) we are facing now so as to make appropriate responses to that. For this reason, a better solution to merge the information from different sources and mining algorithm results will be useful to let the user make the right decision.

4. CONCLUSION

In this paper, we have proposed a new approach for data clustering in large volume of data. From the system perspective, the KDD process is used as the framework for these studies and is summarized into three parts: input, analysis, and output. From the perspective of big data analytics framework and platform, the discussions are focused on the performance-oriented and results-oriented issues. 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.

REFERENCES:

[1] J. Han and M. Kamber. Data Mining: Concepts and Techniques, 2nd Ed. Morgan Kaufman. 2006.

[2] M. H. Dunham. Data Mining: Introductory and Advanced Topics. Pearson Education. 2006.

[3] Bughin et al. 2010, J Bughin, M Chui, J Manyika, Clouds, big data, and smart assets: Ten tech-enabled business trends to watch, McKinSey Quarterly, 2010

[4] IBM 2012, what is big data: Bring big data to the enterprise, http://www-01.ibm.com/software/data/bigdata/, IBM.

[5] Labrinidis and Jagadish 2012, A. Labrinidis and H. Jagadish, Challenges and Opportunities with Big Data, In Proc. of the VLDB Endowment, 5(12):2032-2033, 2012

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

[7] H. Greg and E. Charles, "Alternatives to the kmeans algorithm for better clusterings." CIKM '02 Proceedings of the eleventh international conference on CIKM, 2002.

[8] Charles Elkan. Using the triangle inequality to accelerate k-means. In 20th International Conference on Machine Learning (ICML-2003), Washington, DC, 2003.

[9] 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

[10] 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

[11] JuntaoWang ,Xiaolong Su , "An improved KMeans clustering algorithm", Communication Software and Networks (ICCSN)- 2011, Page(s):44 – 46.

[12] Krishna K, Murty MN. Genetic kk-means algorithm. IEEE Trans Syst Man Cyber Part B Cyber. 1999; 29(3):433–9.

[13] Liu B. Web data mining: exploring hyperlinks, contents, and usage data. Berlin, Heidelberg: Springer-Verlag; 2007.

[14] d'Aquin M, Jay N. Interpreting data mining results with linked data for learning analytics: motivation, case study and directions. In: Proceedings of the International Conference on Learning Analytics and Knowledge, pp 155–164

[15] Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings of the IEEE Symposium on Visual Languages, 1996, pp 336–343