International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Distributed Data Clustering : A Comparative Analysis

V. Maria Antoniate Martin*[1], Dr. K. David[2], B. Merlinsuganthi[3]
*[1]Research Scholar, Department of Computer Science, Research and Development Centre, Bharathiar University, Coimbatore, Tamil Nadu, India
[2]Assistant Professor, Department of Computer Science, The Rajah's College, Pudukkottai, Tamil Nadu, India
[3]Student, Department of Information Technology, St. Joseph's College, Trichy, Tamil Nadu, India

ABSTRACT

Distributed computing plays an important role in the Data Mining process. Cluster analysis is one of the most common techniques in data mining. Clustering is a task of grouping a set of objects in such a way that objects is in the same group. Data mining is a function that assigns items in a collection to target categories or classes. There are many different techniques and algorithms are available for distributed data clustering. Cluster analysis itself is not one specific algorithm, but the general task to be solved. Many researchers have proposed clustering algorithms, which work efficiently in the distributed mining. This paper compares the performance of distributed clustering algorithms namely, Distributed k-means algorithm and partition algorithm. In this research paper we have to discuss, the comparative analysis of some of these distributed clustering.

Keywords : Distributed cluster, Centroid, k-Means, k-Medoid, CLARA.

I. INTRODUCTION

This Cluster is an unsupervised learning process. Clustering is a major task in data analysis and data mining applications. It is the assignment of combination a set of objects so that objects in the identical group are more related to each other than to those in other groups. Cluster is an ordered list of data, which have the familiar characteristics. Cluster analysis is a descriptive data-mining task where the goal is to group similar objects in the same cluster and dissimilar objects in different clusters. Applications of clustering include clustering customers for the purpose of market segmentation and grouping similar documents together in response to a search engine request.

Cluster analysis can be done by finding similarities between data according to the characteristics found .It groups data objects on information found in the data that describes the objects and their relationships. Clustering algorithms can be applied to a wide range of problems, including customer segmentation, image segmentation, information retrieval and scientific data analysis. There are different clustering algorithms; each has its own characteristics. Data clustering techniques are developed to partition an initial set of observation data into collections with small in- group distances and big out-group distances.

The main objective of distributed clustering algorithms is to cluster the distributed datasets without necessarily downloading all the data to the single site. It assumes that the objects to be clustered reside on different sites. This process is carried out on two different levels: local level and global level. On the local level, all sites carry out clustering process independently from each other. After having completed the clustering, a local model or local representative is.

The K-means algorithm is one of the most commonly used clustering algorithms. The "K" in its name refers to the fact that the algorithm looks for a fixed number of clusters, which are defined in terms of proximity data points to each other. It is determined, which should reflect an optimum trade-off between complexity and accuracy in is the process of group in objects in toclusters such that objects within a cluster have High similarity in comparison to one another, but are very dissimilar to objects in other clusters.

II. CLUSTER

There is no clustering algorithm performing best for all datasets. Each dataset requires both expertise and insight to choose a single best clustering algorithm, and it depends on the nature of application and patterns to be executed. In data, mining of data can be done using two learning approaches, supervised and unsupervised learning. A well-known limitation of data clustering algorithms, such as the k-means algorithm, is that the number of clusters has to be specified beforehand, based for example, on subjective evaluations or a priori analysis.

The aim partition-based algorithm is to decompose clusters. Where the numbers of the resulting cluster is predefined by the user. The algorithm uses an iterative method and based on a distance measure updates the cluster of each object. The numbers of clusters has to be determined in advance and spherical shapes can be determined as cluster. Clustering the process of grouping physical or abstract objects into classes of similar object. Traditional clustering methods require that all data have to be located at the site, where they are analyzed and cannot be applied in the case of multiple distributed datasets, unless all data are transferred in a single location and then clustered. Due to technical, economical or security reasons, it is not always possible to transmit all the data from different local sites to single location and then perform global clustering.

III. PARTITIONING ALGORITHM

Partitioning clustering algorithm split the data points into k division, where each division represent a cluster and k<=n, where n is the number of data points. Partitioning methods are based on the idea that a cluster can be represented by a Centre point. The cluster must exhibit two properties, they are (a) each collection should have at least one object (b) every object should belong to accurately one collection. The main drawback of this algorithm is whenever a point is close to the centre of another cluster; it gives poor outcome due to overlapping of data points .It uses a number of greedy heuristics schemes of iterative optimization.

There are many methods of partitioning clustering: They are K-Means, K-Medoid Method, PAM (Partitioning around Medoid), and CLARA (Clustering Large Applications).

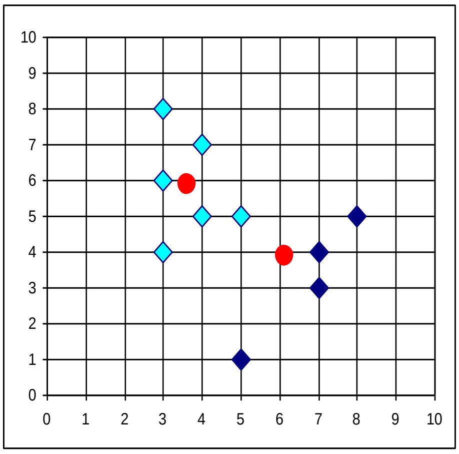
All paragraphs must be indented. All paragraphs must be justified, i.e. both left justified and right justified.

K-Means algorithm

- In this algorithm, a cluster is represented by its centroid, which is a mean (average pt.) of points within a cluster. This works efficiently only with numerical attributes and a single outlier can negatively affect it.
- The k-means algorithm is the most popular clustering tool that is used in scientific and industrial applications. The classical k-means algorithm provides an easy-to-implement approximate solution to. Reasons for popularity of k-means are ease of interpret-tation, of implementation, scalability, speed of convergence, adaptabil- ity to sparse data, and ease of out-of-core implementation.
- Here we have used K-Means clustering algorithm. The input to the algorithm is a data from multiple entries in table of database to be

clustered with same property. The output from the clustering algorithm provides the average distance from cluster members to the center of each cluster.

- The similarity is obtained from the distance from the center of the cluster. In our approach we have used K-Means Cluster based algorithm. K-Means is a widely use clustering algorithm. In that, we use random seeds data and according to that arrange the clusters.



K-Means

IV. K-MEDOIDALGORITHM

Medoid are representative objects of a data set or a cluster with a data set whose average dissimilarity to all the objects in the cluster is minimal. The k-means algorithm is sensitive to outliers since an object with an extremely large value may substantially distort the distribution of data. Instead of taking the mean value of the objects in a cluster as a reference point, the Medoid can be used, which is the most centrally located object in a cluster. K-Medoid method uses representative objects as reference points instead of taking the mean value of the objects in each cluster. While median indicates the value around that, which all data items are evenly distributed around it. The basic idea of this algorithm is to first compute the K representative objects, which are called as Medoid. After finding the set of Medoid, each object of the data set is assigned to the nearest Medoid.

Algorithm: k-Medoid

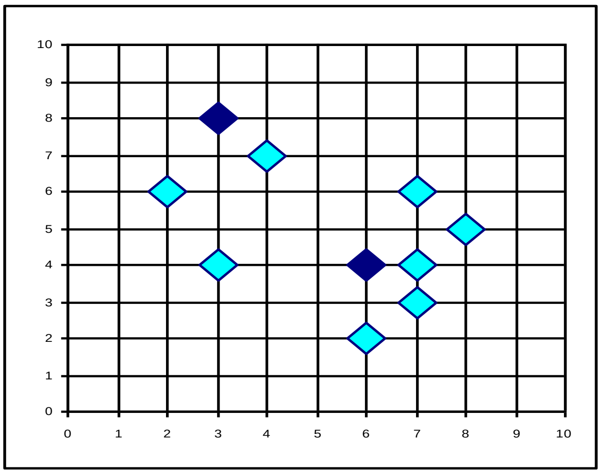
Input: The number of clusters k and a database containing n objects

Output: A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest Medoid.

Method: Arbitrarily choose k objects as the initial Medoid

- ✓ Repeat
- ✓ Assign each remaining object to the cluster with the nearest Medoid
- ✓ Randomly select a non-Medoid object, o random.
- ✓ • Compute the total cost, S of swapping OJ with orandom to form the new set of k Medoid
- ✓ Until no change.

The results are presented here in tabular a graphical form for many experimental runs. Number of clusters chosen by user is



K-Medoid

.

V. CLARANS

To deal with huge data sets, a sampling based technique, called CLARA (Clustering Large Applications) and an improved version which is based on randomized search called CLARANS (Clustering Large Applications based upon Randomized Search) can be used. CLARA cannot find a good clustering if any of the best-sampled Medoid is far from the best k-Medoid.

***Pros and Cons of Partitioning Algorithm**:*

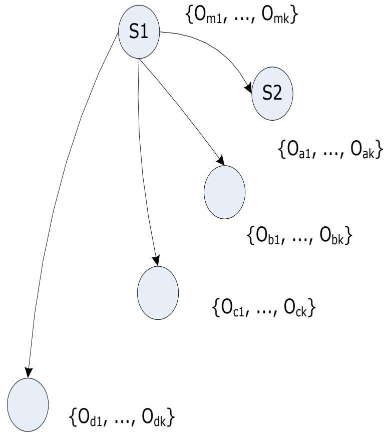
It is simple to understand and implement. It takes less time to execute as compared to other techniques. The drawback of this algorithm is the user has to provide pre-determined value of k and it produces spherical shaped clusters. It cannot handle with noisy data object. CLARA uses 5 samples, each with 40+2k points, each of which are then subjected to PAM, which computes the best Medoid from the sample. A large sample usually works well when all the objects have equal probability of getting selected.

The complexity of computing Medoid from a random sample is O (ks2+k (n-k)), Where s is the size of the sample. A typical k-Medoid partitioning algorithm works effectively for small data sets, but does not scale well for large data sets. To deal with larger data sets, a sampling-based method, called Clara (clustering large applications) can be used. The idea behind CLARA is as follows: Instead of taking the whole set of data into consideration, a small portion of the actual data is chosen as a representative of the data.

Medoid are then chosen from this sample partitioning Around Medoid. If the sample is selected in a fairly random manner, it should closely represent the original data set. The representative objects (Medoid) chosen will likely be similar to those that would have been chosen from the whole data set. Clara draws multiple samples of the data set, applies PAM on each sample, and returns its best clustering as the output. The effectiveness of CLARA depends on the sample size. Notice that PAM searches for the best k-Medoid among a given data set, whereas CLARA searches for the best k- Medoid among the selected sample for the data set.

CLARA cannot find the best clustering if any sampled Medoid is not among the best k- Medoid. A k-Medoid. Type algorithm called CLARANS (Clustering Large Applications based upon RAN (domized Search) was proposed that combines both sampling technique with PAM. However, unlike CLARA, CLARANS does not confine itself to any sample at any given time. While CLARA has a fixed sample with some randomness in each step of the search, CLARANS draws a sample with some randomness in each step of the search.

- ✓ A graph abstraction, $G_{n,\,k}$
- ✓ Each vertex is a collection of k Medoid
- ✓ S1   S2 | = k – 1
- ✓ Each node ask (n-k) neighbors
- ✓ Cost of each node is total dissimilarity of objects to their Medoid
- ✓ PAM searches whole graph
- ✓ CLARA searches sub graphoid.



VI. CONCLUSION

In this paper, comparative study has been performed on the k-means, k-Medoid, CLARA, partitioning algorithm. The comparative study is performed because of accuracy and efficiency parameters. The partition based algorithms work well for finding spherical-shaped cluster small to medium-sized databases. The k-Medoid method is more robust than k-means in the presence of noise and outliers because a Medoid is less influenced by outliers or other extreme values than a mean. But its processing is more costly than the k-means method. The k-Medoid method works effectively for small data sets, but does not well for large data sets. To deal with larger data sets, a sampling-based method, called CLARA can be used. The effectiveness of CLARA depends on the sample size. CLARANS is the most effective portioning method among all. K-means clustering algorithms forms clusters with less time and more accuracy than otherAlgorithms. In terms of time and accuracy, K-means produces better results as compared to other algorithms.

VII. REFERENCES

[1]. Jiawey Han, MichelineKamber," Data Mining Concepts And Techniques" Morgan Kaufmann Publishers, NewDelhi, 2001.

[2]. Chauhan R, Kaur H, Alam M A, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications, (0975 – 8887) November2010Vol.10– No.6. .

[3]. DataClusteringAlgorithmsAvailable https://sites.google.com/site/dataclusteringalgorithms/density-based-clustering-algorithm.3(2).

[4]. K.Kameshwaran and K. Malarvizhi, "Survey on Clustering Techniques in Data Mining", International Journal of Computer Science and Information Technologies (0975-9646), Vol. 5(2), 2014.

[5]. G. Olive, R. Setola, and C. Hadjicostis, "Distributed *k*-Means algorithm," http://arxiv.org/abs/1312.4176.

[6]. http://en.wikipedia.org/wiki/Data clustering.

[7]. http://wwwusers.cs.umn.edu/~han/dmclass/cluster_survey_10_02_00.pdf.

[8]. Sony N, Ganatra A (2012) Categorization of Several Clustering Algorithms from Different Perspective: A Review. IJARCSSE.

[9]. Han J, Kamber M (2001) Data Mining. Kaufmann Publishers, Morgan(2).

[10]. Rao IKR (2003) Data Mining and Clustering Techniques DRTC Workshop on Semantic Web, pp.23-30.

[11]. J. A. Hartigan and M. A. Wong, "A k-means clustering Algorithm," Applied Statistics, 28:100-- 108, 1979.

[12]. Zhong Wei, et al. "Improved K-Means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property" IEEE Transactions on Nanobioscience, Vol.4, No.3. Sep. 2005. 255-265.

[13]. A.K. Jain, M.N. Murty, and P. J. Flynn (1999) "Data Clustering: a review", ACM Computing Surveys (CSUR), Vol.31, Issue 3, 1999.

[14]. Grabmeier, J. and Rudolph, A. (2002) "Techniques of Cluster algorithms in data mining" Data Mining and Knowledge Discovery, 6, 303-360.

[15]. Han, J. and Kamber, M. (2001) "Data Mining: Concepts And Techniques", (Academic Press, San Diego, California, USA).

[16]. Pham, D.T. and Afify, A.A. (2006) "Clustering Techniques and their applications in engineering". Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science,

[17]. A.K. Jain, M.N. Murty, and P. J. Flynn (1999) "Data Clustering: a review", ACM Computing Surveys (CSUR), Vol.31, Issue 3, 1999.

[18]. K. Wagstaff, C. Cardie, S. Rogers, and S. Schrodl, "Constrained k-means clustering with background knowledge," in Proceedings of the International Conference on Machine Learning, pp. 577– 584, 2001.

[19]. Davidson and S. S. Ravi, "Clustering with constraints: feasi- ability issues and the k-means algorithm," in Proceedings of the fifth SIAM International Conference on Data Mining (SDM ’05), vol. 5, pp. 201–211, 2005.

[20]. P. Brucker, "On the complexity of clustering problems," in Optimization and Operations Research, pp. 45–54, Springer, Berlin, Germany, 1978.

VIII. AUTHOR DETAILS

V. Maria Antoniate Martin is a Research Scholar in Computer Science at Bharathiar University, Coimbatore, Tamil Nadu, India. He is also working as an Assistant Professor in Department of Information Technology at St. Joseph's College, Tiruchirappalli, Tamil Nadu, India. He received his Bachelor of Science degree in Computer Science from Bharathidasan University in 2003; He completed his Masters in Science in Computer Science from the same University in 2006. He also completed his Masters in Philosophy in Computer Science from the same University in 2011.He has seven years of teaching experience. He has published seven research articles in reputed International Journals. He is also the co-author of a publication in a National Conference of importance. His area of research is Data Mining.

Dr. K. David is an Assistant Professor in the Department of Computer Science at H.H. the Rajah's College, Pudukkottai, TamilNadu, 622001. He has over fifteen years of teaching experience. He has published scores of papers in peer reviewed journals of National and International repute and is currently guiding seven Ph.D., scholars. His research interests include, UML, OOAD, Knowledge Management, Web Services and Software Engineering.

B.Merlin suganthi is a student of M.Sc. Computer Science, St. Joseph's College, Trichy - 620002.She received her Bachelor of Science degree in Computer Science from Bharathidasan University in 2016.