Volume 3, Issue 10, October 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Distributed K-Means in Data Clustering Space

Anurag Bhardwaj*
Computer Sciences Corporation,
India

Ashutosh Bhardwaj
Infosys Limited,
India

Abstract— This paper discusses distributed clustering based on K-means and coarse-grained algorithm which combines local search advantages of K-means with distributed global optimization merits. This is a variant of K-means. Normalization before clustering scales the data values into a suitable range using various techniques. The algorithm produces an optimum quality of clusters in distributed environment.

Keywords— Distributed Clustering, K-Means, Normalization, Migration.

I. INTRODUCTION

The process grouping a set of physical or abstract objects into classes of similar objects is called clustering. The grouping is accompanied by finding similarities between data according to characteristics found in actual data. The groups are called clusters. Given a database D = { t_1, t_2, t_3, ……., t_n} of tuples and an integer value 'k', the clustering problem is to define a mapping f : D -> {1, …..,k} where each t_i is assigned to one cluster K_j, 1<=j<=k. A cluster, „K_j, contains precisely those tuples mapped to it; that is, K_j = { t_i | f(ti) = K_j} 1<=i<=n, and t_i € D.[4] Clustering is one of the main entity in data mining. The two prominent types of clustering are: Hierarchical clustering and Partitional clustering.

A. *Hierarchical Clustering*

Hierarchical clustering creates a hierarchical decomposition of given set of data objects. Each level in the hierarchy has a separate set of clusters. At the lowest level, each item is in its own cluster.[6] At the highest cluster, all items belong to the same cluster. Hierarchical method can be classified as the following two types of clustering schemes

1. *Agglomerative Clustering*

It is a bottom up approach that starts with each individual item in its own cluster and then iteratively merges clusters until all items belong to one cluster. Different agglomerative algorithms differ in how the clusters are merged at each level.

2. *Divisive Clustering*

It is a top down approach which starts with all the objects in the same cluster. With each successive iteration, a cluster is split up into smaller clusters, until eventually all objects are in their own cluster. The idea is to split up clusters where some elements are not sufficiently close to other elements.

B. *Partitional Clustering*

A partitional clustering creates the cluster in one step as opposed to several steps done in hierarchical clustering. Various types of partitional clustering are: Partitional minimum spanning tree, Squared error clustering algorithm, K-means clustering algorithm, Nearest neighbour algorithm, PAM (Partitioning Around Medoids) algorithm.

II. DISTRIBUTED CLUSTERING

With the increase in amount of data and storage capacity of computers, the field of data mining is growing fast. There exists a set of different techniques related to data mining out of which data clustering is one of the most in demand nowadays.[8] A set of homogeneous data is categorized into distinct clusters based on similarities between them. The following two properties impose a necessity to use distributed algorithms on data clustering to achieve better performance:

- Data is to be distributed over a set of clusters which are far apart.
- High computational cost of clustering algorithms.

K-means clustering aims at partitioning, 'n' observations into 'k' clusters, where each observation belongs to the cluster with nearest mean. The basic principle of K-means method is to select K clustering center randomly, allocating every node to nearest cluster calculated by Euclidean distance, then calculating the average of every reallocated clustering and iterating as new clustering center. Generally, high computational complexity arises due to the evaluative function[10] in clustering algorithms. K-means algorithm uses iterative heuristics search method which decreases the computational complexity and helps in achieving high performance. Distributed algorithms based on K-means are well efficient hence, and the result after clustering is expected to have high precision.[1]

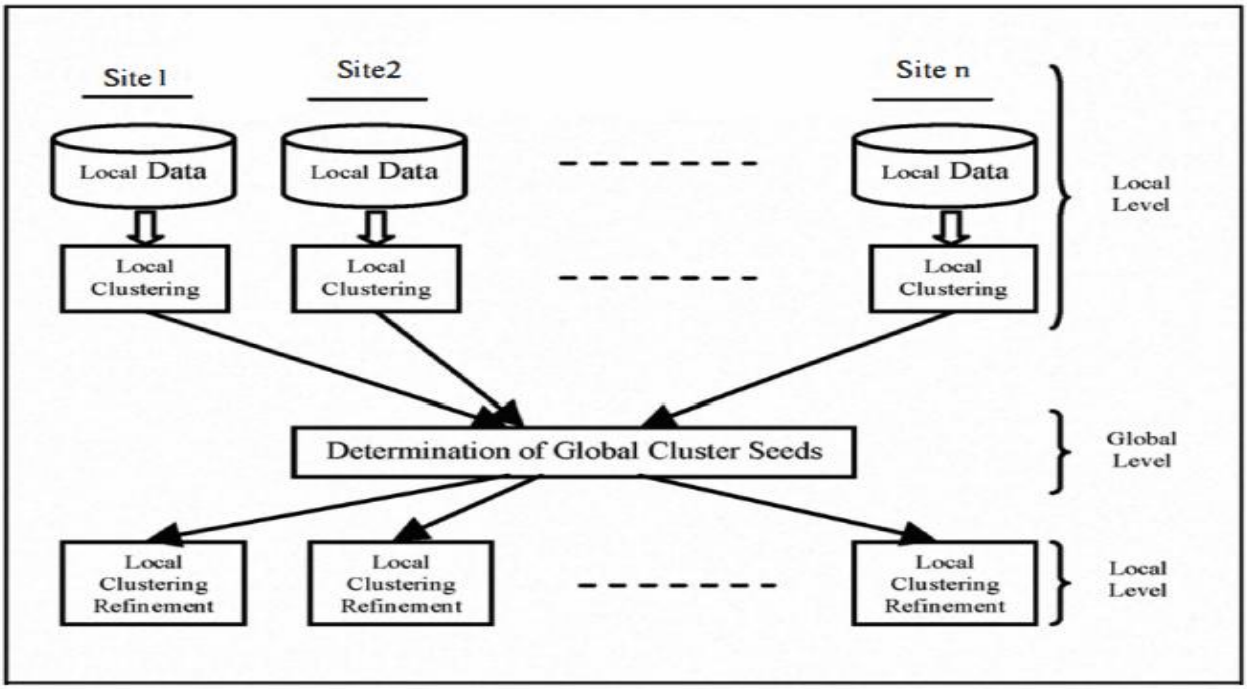

Fig.1 Distributed Clustering.

Here the distributed clustering algorithm based on K-means and Coarse-grained Parallel Genetic Algorithm (CPGA) combines the advantages of both CPGA and K-Means and decreases the network overload in an efficient manner with migration strategy. CPGA is basically related to parallelization of genetic algorithm sometimes referred as Island parallel algorithm whereas during migration, unqualified individuals[5] of the objective population are replaced by optimal individuals of the population.

III. METHOD AND EVALUATION

Clustering issues over distributed environment can be solved with the proposed algorithm effectively. Other issues include enhancement perspective of local data and reducing the overloading of network by adopting proper migration technique. Theoretical and experimental analysis confirms the feasibility of Distributed Clustering Based on K-means and Coarse-grained parallel genetic algorithm (DCBKC). Migration saves a lot of communication time. In island parallel algorithm migration strategy is adjusted by allowing optimal individuals to immigrate to adjacent sub-populations. Every sub-population evolves independently and then selects individuals satisfied optimum fitness function value for migrating m generation of one time. Fitness function is defined as:

$$f(s) = \frac{1}{\sum_{i=1}^{k} \sum_{x \in C_i} d(x, r_i)}$$

where C_i is i^{th} clustering division, r_i is i^{th} clustering center.

The technique of global normalization is applied before clustering of distributed data sets is done. The step by step algorithm to normalize distributed K-means clustering is presented. Initially, first minimum and maximum values are taken out from datasets[3] and transmitted to the center where global minimum and maximum values are taken into account. These two values are then transmitted to local sites to perform global normalization using Min-max technique. Further, the normalized objects are clustered using K-means algorithm. All local centroids are then merged, grouped into similar centroids and a global centroid is obtained. The global centroid is transmitted to local site, where Euclidean distance of each object is computed and assigned to each cluster centroid. The performance of normalization based distributed K-means (NDKM) clustering algorithm is monitored. Quality analysis of it is done by comparing with other normalization techniques.

The Euclidian distances are normalized as it is very sensitive to differences in magnitude or scales. Three alternative approaches exist to normalize the Euclidian distances:

- Performing linear transformation on linear data.

$$v' = \frac{v - \min_a}{(\max_a - \min_a)}$$

- Normalizing the values for an attribute based on mean ($\bar{A}$) and standard deviation (σ_A) of A (Attribute).

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

- Normalizing the distances by moving the decimal point of values of attribute A.

$$v' = \frac{v}{10^j}$$

The impact of normalization in the process of distributed K-means is explored. The experimental analysis performed on Iris dataset of UCI machine learning data repository confirms that normalization produces an optimum quality of clusters in distributed environment. All experiments were implemented in VBA programming backdrop.

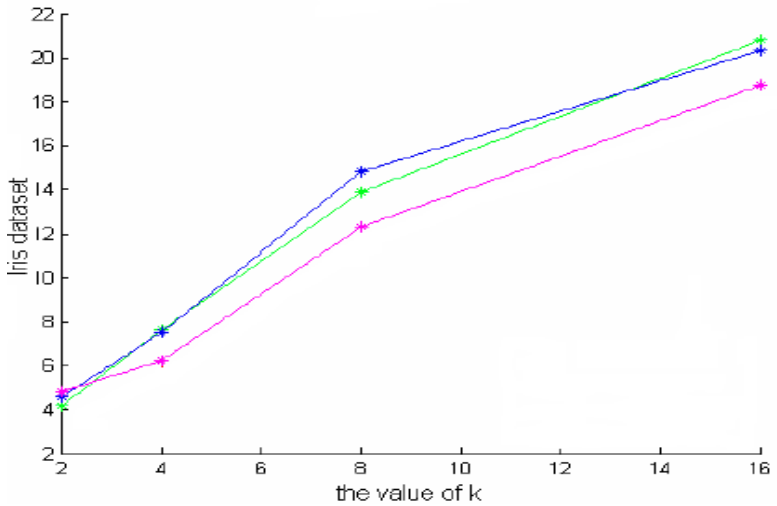

Fig. 2 Plot of normalization with respect to k using Iris Dataset at three different instance.

Distributed clustering algorithm based on K-mean provides cooperation between sites as well as takes care of data security and site independency.[7] One site is unaware of the details of the other. The sites cooperate with each other and high degree of data security is maintained. The proposed model provides independent clustering of the data in the local sites,[2] global clustering in the central site and sending the results to the other sites. Clustering is performed at two levels: local and global.

Steps to perform clustering in a distributed environment:
Step 1: Local clustering (by the algorithm above)
Step 2: Extraction of local properties
Step 3: Determining global model based on the local models
Step 4: Updating the local models
The main parameters for communication with the central site are:

$$\vec{S} = \sum_{i=1}^{N} \vec{X}_i, \vec{M} = \frac{\vec{S}}{N}, \vec{S}S = \sum_{i=1}^{N} \vec{X}_i^2$$

$$\vec{SD} = \left(\frac{\sum_{i=1}^{N} (\vec{X}_i - \vec{M})^2}{N} \right)^{\frac{1}{2}} = \left(\frac{\vec{SS} - 2\vec{M} \times \vec{S} + N \times \vec{M}^2}{N} \right)^{\frac{1}{2}}$$

where, S: Sum, M: Mean, SS: Sum of squares, SD: Standard Deviation.

Each of the sites does clustering locally and then sends it to the central site. Central site then finds the Euclidian distance,„d" between any two cluster centers.

Step 5: Find the property i in which the difference of the two centers

$\Delta D_i = |C_{1i} - C_{2i}|$ is the maximum.

Step 6: If the following equation holds, then the clusters are put in the same neighbourhood:

$$(d-T)\times\frac{\Delta D_i}{d}\le \max(SD_{1i}, SD_{2i})$$

Experimental studies show that the new distributed clustering algorithm is highly efficient with low communication cost and bandwidth required and moreover, offers privacy of local sites making it suitable for networks.

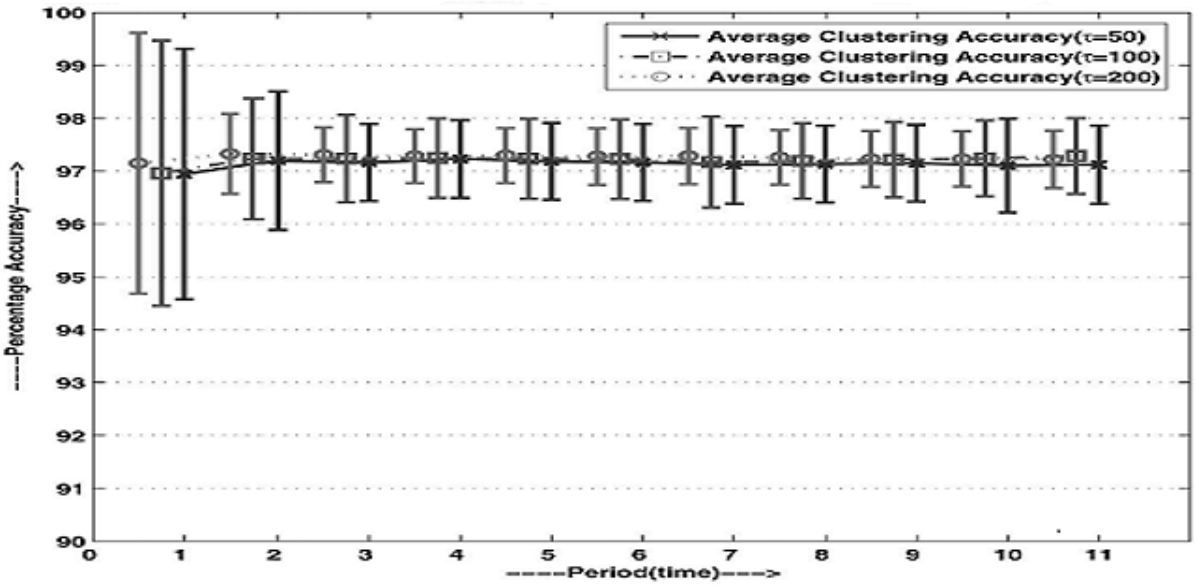

Fig.3 Variation of clustering accuracy over time with stationary data distribution.

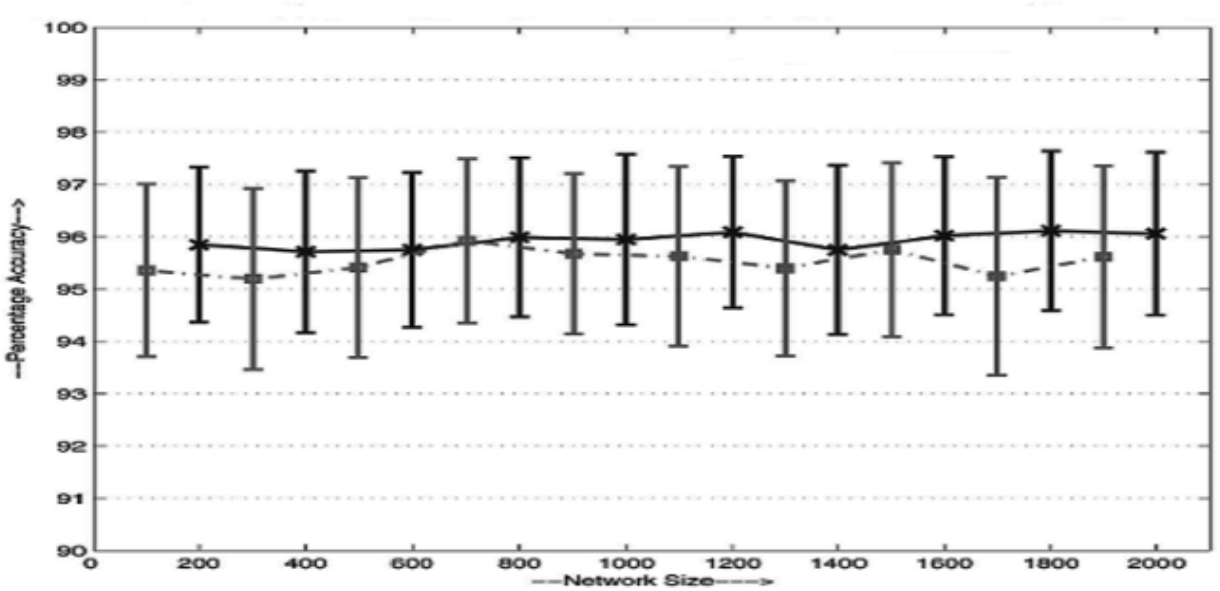

Fig.4 Clustering accuracy with increasing network size for K-means with non-stationary data distribution.

	K-Means + CPGA	NDKM	Probability Tree
Parameters	Euclidean Distance	Euclidean Distance	Standard Deviation
Techniques Involved	Iterative Heuristic Search, Migration	Min-Max Normalization	Fuzzy Logic
Network Load	Decreases	No Effect	Decreases

Security	Low	Low	Very High
Efficiency	High	Higher than DCBKC	High

Table 1: Comparison between DCBKC, NDKM and Probability Tree.

IV. CONCLUSION

A clear picture of distributed clustering based on K-means is presented in this paper with efficient algorithm and an extension with normalization. It is based on K-means and CPGA which combines local search advantages of K-means with distributed global optimization merits of CPGA. The algorithm is efficient and feasible and distinguishes itself from the usual hierarchy methods. Privacy of local sites and data security is possible.[9] Normalization before clustering scales the data values into a suitable range using various techniques. Use of normalization produces an optimum quality of clusters in distributed environment. Future work will involve improving the mechanism of migration taking privacy into account. Convergence speed and efficiency also can be improved for a complicated data mining problem in a large, distributed environment.

REFERENCES

[1] G. Forman and B. Zhang, "Distributed Data Clustering Can Be Efficient and Exact," SIGKDD Explorations, vol. 2, no. 2, pp. 34-38,

[2] Veeraswamy et al.,"An Implementation of Mining Weighted Association Rules without Preassigned Weights" International Journal of Advanced Research in Computer Science and Software Engineering 2(8),August- 2012, pp. 276-283

[3] Januzaje E, Kriegel H, Pfeifle M, Scalable Density Based Distributed Clustering. Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD). Springer, 2004.231 – 244.

[4] David Hand, Heikki Mannila, Padhraic Smyth. Principles of Data Mining. Cambridge: The MIT Press,2001.

[5] Aleksander lazarevic, et. al. "Distributed clustering and local regression for knowledge discovery in multiple spatial database". School of electrical engineering and computer science, Washington state university 2000.

[6] Jiuyong, L., Wong, R.C.-W., Fu, A.W.-C., Jian, P.: Achieving k-Anonymity by Clustering in Attribute Hierarchical Structures. In: Tjoa, A.M., Trujillo, J. (eds.) DaWaK 2006. LNCS, vol. 4081, pp. 405–416. Springer, Heidelberg (2006)

[7] Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J.: UCI Repository of Machine Learning Databases (1998), available at http://www.ics.uci.edu/~mlearn/MLSummary.html

[8] H. Kargupta and K. Sivakumar, "Existential Pleasures of Distributed Data Mining," Data Mining: Next Generation Challenges and Future Directions, AAAI Press, 2004.

[9] I. Sharfman, A. Schuster, and D. Keren, "A Geometric Approach to Monitoring Threshold Functions over Distributed Data Streams," ACM Trans. Database Systems, vol. 32, no. 4, pp. 23:1-23:29, 2007.

[10] E Joshi "Improving Performance of Algorithms in Distributed Computing with Perspective of Green Information Technology"2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 18.

