PROPOSAL FOR DEVELOPING AN APPROACH TO CONSTRAINTS BASED MULTI-DIMENSIONAL DATA CLUSTERING AIDED WITH ASSOCIATIVE CLUSTERING USING COMPARATIVE STUDY

B.KRANTHI KIRAN [1], Dr. A VINAYA BABU [2]

[1]*Assistant Professor, Department of Computer Science and Engineering, JNTUHCEJ, Karimnagar(India)*

[2]*Professor, Department of Computer Science and Engineering, JNTUniversity Hyderabad(India)*

ABSTRACT

Our paper outlines a proposal for developing an approach to constraints based multi-dimensional data clustering aided with associative clustering. Our proposed approach evolved from comparative study of associative clustering, gene expression, multi-dimensional large data sets in a distributed environment from state of the art. Our proposed approach will work well in many of the emerging distributed, ubiquitous, possibly privacy-sensitive data mining applications. Clustering on multidimensional data sets using parallel algorithms may degrade performance linearly with increase in data size as well other challenges such as different data types and formats, computational complexity have to be taken care of. We are proposing a scalable parallel algorithm which will cater to these challenges.

Keywords: *Associative clustering, gene expression, multidimensional data set,scalable paraallel algorithm*

I. INTRODUCTION

Data mining evolves as a promising solution in discovering knowledge hidden in databases. Data Mining has been formally defined as "the non-trivial extraction of implicit, previously unknown and potentially useful information from data in databases" [1], [2]. Data mining has been utilized for multiple needs both in the private and public sectors. Precise usage of data mining include market segmentation, fraud detection, direct marketing, interactive marketing, market basket analysis, trend analysis and more [3, 4, 5, 7]. Advances in computing and communication over wired and wireless networks have resulted in many pervasive distributed computing environments. These environments often come with different distributed sources of data and computation. Mining in such environments naturally calls for proper utilization of these distributed resources. However, most off-the-shelf data mining systems are designed to work as a monolithic centralized application. They normally download the relevant data to a centralized location and then perform the data mining operations [1-7]. This centralized approach does not work well in many of the emerging distributed, ubiquitous, possibly privacy-sensitive data mining applications. Distributed Data Mining (DDM) offers an alternate approach to address this problem of mining data using distributed resources [6].

Clustering [16, 26] has been studied extensively for more than forty years in data mining field and across many disciplines due to its broad applications. Clustering is the process of assigning data objects into a set of disjoint groups called clusters so that objects in each cluster are more similar to each other than objects from different clusters. The literature presents with an enormous number of algorithms for efficient clustering of data. These algorithms can be categorized into nearest-neighbor clustering, fuzzy clustering, partitional clustering, hierarchical clustering, artificial neural networks for clustering, statistical clustering algorithms, density-based clustering algorithm and so on. In these methods, hierarchical and partitional clustering algorithms are two primary approaches of increasing interest in research communities. Hierarchical clustering algorithms can

usually find satisfiable clustering results. Although the hierarchical clustering technique is often portrayed as a better quality clustering approach, this technique does not contain any provision for the reallocation of entities, which may have been poorly classified at the early stage. Furthermore, most of the hierarchical algorithms are very computationally intensive and require much memory space [25].

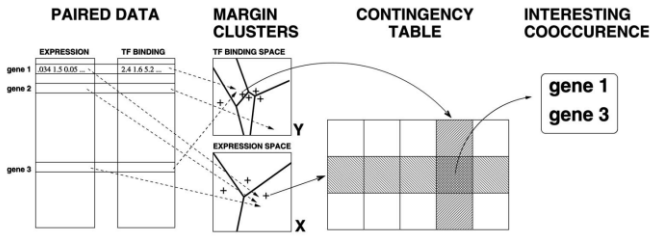

Fig. 1.Clustering extracted from Kaski et al [1]

Recently, large data clustering has been extensively studied in many areas, including statistics, machine learning, pattern recognition, and image processing [13-15]. In the areas, the scalability of clustering methods and the techniques for big data clustering much active research has been devoted. To overcome the problems occurred in large database clustering different methods has been introduced, including initialization by clustering a sample of the data and using an initial crude partitioning of the entire data set [7]. However, the most prominent representatives are partitioning clustering methods such as CLARANS [11]; hierarchical clustering methods such as BIRCH [10]; grid clustering methods such as STING [8] and WAVECLUSTER [9]. Each method has its advantages and shortcomings. They are not suitable for processing very large databases. It is difficult to acquire both high accuracy and efficiency in a clustering algorithm of large data. The two targets always conflict with each other. To process massive data sets, the power of a single computer is not enough. Parallel and distributed clustering is the key technique. It will be highly scalable and low cost to do clustering in a distributed environment.

II. RELATED WORK

Samuel Kaski et al.[23] have proposed method on associative clustering for exploring dependencies between functional genomics datasets. High-throughput genomic measurements, interpreted as co-occurring data samples from multiple sources, open up a fresh problem for machine learning: What is in common in the different data sets, that is, what kind of statistical dependencies are there between the paired samples from the different sets. They introduce a clustering algorithm for exploring the dependencies. Samples within each data set are grouped such that the dependencies between groups of different sets capture as much of pair wise dependencies between the samples as possible. They have formalized this problem in a novel probabilistic way, as optimization of a Bayesfactor. The method is applied to reveal commonalities and exceptions in gene expression between organisms and to suggest regulatory interactions in the form of dependencies between gene expression profiles and regulator binding patterns.

Yao yuhui et al.[24] have described an approach to the analysis of gene expression data using Associative Clustering Neural Network(ACNN). ACNN dynamically evaluates similarity between any two gene samples through the interactions of a group of gene samples. It has feasibility to more robust performance than those similarities evaluated by direct distances. The clustering performance of ACNN has been tested on the Leukemia's data set. The experimental results demonstrate that ACNN

can achieve superior performance in high dimensional data (7129 genes). The performance can be further enhanced when some useful feature selection methodologies are incorporated. The study has shown ACNN can achieve 98.61% accuracy on clustering the Leukemias dataset with correlation analysis.

Inderjit S et al. [12] presented a parallel implementation of the k-means clustering algorithm based on the message passing model. Their algorithm exploits the inherent data-parallelism in the k-means algorithm. They analytically showed that the speedup and the scale up of their algorithm approach the optimal as the number of data points increases. Wen-Yen Chen et al, [17] have investigated representative methods of approximating the dense similarity matrix because of the drawback of spectral clustering in large database. The one of the method was compared by sparsifying the matrix and the other by the Nyström method. Here, they designed a parallel implementation and its scalability was evaluated. By retaining nearest neighbors and investigating its parallelization they pick up the strategy of sparsifying the matrix. Here, both the memory used and computation on distributed computers were parallelized by them. The experimental result on various dataset showed that the designed algorithm can effectively handle large problems.

Eshref Januzaj *et al*¸ [18] have designed a scalable density-based distributed clustering algorithm. Here, a user-defined trade-off between clustering quality and the number of transmitted objects were allowed by the designed clustering algorithm. The procedure included in the designed method were, according to a quality criterion reflecting their suitability to serve as local representatives they ordered all objects located at a local site. The best among these representatives was send to a server site. It was then clustered with a slightly enhanced density-based clustering algorithm. Their experimental result showed that their designed algorithm outperformed in high quality clustering with scalable transmission cost. Josenildo Costa da Silva and Matthias Klusch, [19] have addressed the confidentiality problems in distributed data clustering, notably the interference problem. Here, for distributed data clustering they designed an algorithm which was called as KDEC-S. It was to provide mining results when the confidentiality of original data was preserved. The confidentiality level of KDEC-S method was stated only with the presented confidentiality framework. The basic plan of the designed method was not to reconstruct the original data to the given extent..

Ruoming Jin *et al*, [20] have designed a method called Fast and Exact K-means Clustering (FEKM). Only one or a small number of passes on the entire dataset was required by the designed method and provably produced the same cluster centers as reported by the original k-means algorithm. Here, the cluster centers were adjusted by taking one or more passes over the entire datasets before this the designed algorithm created initial cluster centers by sampling. Also a theoretical analysis was provided by them to show that the cluster centers were equal as the ones computed by the original k-means algorithm. The experimental result of real and synthetic datasets showed that the designed algorithm was performed better compared to K-means. Also, here they described and evaluated a distributed version of FEKM which was called as DFEKM. It was best for analyzing data that was distributed across loosely coupled machines. The DFEKM provided better result than two other possible options for exact clustering on distributed data, which were down-loading all data and running sequential k-means, or running parallel k-means on a loose coupled configuration. If there was a significant load imbalance the designed method outperformed parallel k-means.

To find global clustering patterns Genlin Ji and Xiaohan Ling [21] have designed a distributed clustering method based on ensemble learning. The distributed data sources were analyzed and mined by the designed method. The two stages of the distributed clustering were, firstly doing clustering in local sites and then in global site. The local clustering results were transmitted to server site form an ensemble and combining schemes of ensemble learning used the ensemble to generate global clustering results. The generated global pattern from ensemble was mathematically converted to be a combinatorial

optimization problem. A novel distributed clustering algorithm called DK-means was introduced here as an implementation for the model. The experimental results showed that the DK-means achieved similar results to K-means which clusters centralized data set at a time. It was scalable to data distribution varied in local sites, and also showed validity of the model.

Olivier Beaumont *et al*, [22] have deliberated the resource clustering problem in large scale distributed platforms, such as BOINC and WCG. Here, they planned to remove the single computing resource constraint by executing the task on a set of resources. Their goal was to design a distributed method for a large set of resources which enables to build clusters. They explained about a generic 2-phases method which was based on resource augmentation and whose approximation ratio was 1/3. Also, a distributed version of the above method was designed when the metric space was for a small value of D and the $L\infty$ norm was used to define distances. It took O rounds and messages both in expectation and with high probability, where n was the total number of hosts.

III. PROBLEM STATEMENT

The recent advancement in digital world creates very large data to do their relevant process in various domains. Due to this uncontrollable growth of data, clustering played major role to partition into a small sets to do relevant processes within the small sets. But, again, the additional challenge of cluster identification problem is how to deal with large data since most of algorithms are suitable only for small data. The usual way of handling multi-dimensional data in clustering is to solve clustering problem with parallel algorithm. The important assumption here is that the parallel algorithm can do better in terms of time consumption but the effectiveness should be also satisfactory. It means that the recent cluster methods should be applicable to do with multi-dimensional databases and performance should decrease linearly with data size increase. Also, when developing a large data clustering, the additional challenges like, different data format and data handling should be taken into account without much computational complexity. A few of the existing clustering algorithms of large data either can handle both data types but are not efficient when clustering large data sets or can handle large data sets efficiently but are limited to numeric attributes. So recently, the parallel clustering provided significant contribution in the large data clustering. So a scalable parallel clustering can help in handling the above furnished problems.

In order to address the clustering problem related to multi-dimensional data clustering, a number of techniques has been implemented. To address the multi-dimensional data clustering, we consider the gene expression data. Which is one of the most discussed multi-dimensional data. The proposed approach, mainly concentrate on the constraints based multi-dimensional data clustering. The constraints helps in identifying the right data to be clustered and the knowledge regarding the data also considered as a constraint which improves the accuracy of clustering. The data constraints also helps in specifying the data relevant to the clustering task. The dimensional or level constraints confine the dimension of data to be examined in a database. Earlier Samuel kasaki [23] have proposed a method based on associative clustering. Inspired from their research, we planned in corporate associative clustering method to the constraints based multi-dimensional data. The associative clustering method help to identify relationship between the two clusters based on the constraint values. The method will be implemented using the java programming language. The standard gene expression dataset will be used for the experimental process.

IV.OBJECTIVES OF PROPOSED WORK

Earlier Samuel kasaki [23] have proposed a method based on associative clustering. Inspired from their research, we planned in corporate associative clustering method to the constraints based multi-dimensional data. The associative clustering method help to identify relationship between the two clusters based on the constraint values. The method will be implemented using the java programming language. The standard gene expression dataset will be used for the experimental process.

V. CONCLUSIONS

In this paper, we focused on deriving and proposing a new scalable parallel algorithm for clustering on multidimensional data sets. Clustering on multidimensional data sets using parallel algorithms may degrade performance linearly with increase in data size as well other challenges such as different data types and formats, computational complexity have to be taken care of. We made a comparative study of state of the art for compiling and deriving our proposed scalable algorithm.

REFERENCES

[1] Osmar R. Z., "Introduction to Data Mining", In: Principles of Knowledge Discovery in Databases. CMPUT690, University of Alberta, Canada, 1999.

[2] Kantardzic, Mehmed. "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley and Sons, 2003.

[3] E. Wainright Martin, Carol V. Brown, Daniel W. DeHayes, Jeffrey A. Hoffer and William C. Perkins, "Managing information technology", Pearson Prentice-Hall 2005.

[4] Andrew Kusiak and Matthew Smith, "Data mining in design of products and production systems", in proceedings of Annual Reviews in control, vol. 31, no. 1, pp. 147- 156, 2007.

[5] Mahesh Motwani, J.L. Rana and R.C Jain, "Use of Domain Knowledge for Fast Mining of Association Rules", in Proceedings of the International Multi-Conference of Engineers and Computer Scientists, 2009.

[6]. Souptik Datta Kanishka Bhaduri Chris Giannella Ran Wolff Hillol Kargupta "Distributed Data Mining in Peer-to-Peer Networks", Journal of internet computing, vol.10, no.4, pp.18-26. 2006.

[7] Ron Wehrens and Lutgarde M.C. Buydens, "Model-Based Clustering for Image Segmentation and Large Datasets via Sampling",Journal of Classification, Vol. 21, pp.231-253, 2004.

[8] W. Wang, J. Yang, R. Muntz, STING,"A Statistical Information Grid Approach to Spatial Data Mining", VLDB, 1997.

[9] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: A multi-resolution clustering approach for very large spatial databases", VLDB, pp. 428-439, 1998.

[10] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases", In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp.103-114, 1996.

[11] Ng R. T., Han J.: "Efficient and Effective Clustering Methods for Spatial Data Mining", Proceedings 20th International Conference on Very Large Data Bases, pp.144-155, 1994.

[12] Inderjit S. Dhillon and Dharmendra S. Modha, "A Data-Clustering Algorithm On Distributed Memory Multiprocessors", Proceedings of KDD Workshop High Performance Knowledge Discovery, pp. 245-260, 1999.

[13] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-bsed algorithm for discovering clusters in large spatial databases with noise", In SIGKDD, pp. 226–231, 1996.

[14] R. T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining", In VLDB, 1994.

[15] T. Zhang, R. Ramakrishnan, and M. Livny. Birch, "An efficient data clustering method for very large databases", In SIGMOD, pp. 103–114, 1996.

[16] Jinchao Ji , Wei Pang, Chunguang Zhou, Xiao Han, Zhe Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data", journal of Knowledge-Based Systems, vol. 30, pp. 129-135, 2012.

[17] Chen L, Chen CL, Lu M., "A multiple-kernel fuzzy C-means algorithm for image segmentation", IEEE Transaction on System Man Cybernetics: Part B, vol. 41, no. 5, pp. 1263-74, 2011.

[17] Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin and Edward Y. Chang, "Parallel Spectral Clustering in Distributed Systems", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.33, No.3, pp.568 – 586, 2011.

[18] Eshref Januzaj, Hans-Peter Kriegel and Martin Pfeifle, "Scalable Density-Based Distributed Clustering", Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), pp.231-244, 2004.

[19] Josenildo Costa da Silva and Matthias Klusch, "Inference in Distributed Data Clustering", Engineering Applications of Artificial Intelligence, Vol.19, No.4, pp.363-369, 2005.

[20] Ruoming Jin, Anjan Goswami and Gagan Agrawal, "Fast and Exact Out-of-Core and Distributed K-Means Clustering", Journal of Knowledge and Information System, Vol. 10, No.1, pp. 17-40, 2006.

[21] Genlin Ji and Xiaohan Ling, "Ensemble Learning Based Distributed Clustering", Emerging Technology in Knowledge Discovery and Data Mining, Vol. 4819, pp 312-321, 2007.

[22] Olivier Beaumont, Nicolas Bonichon, Philippe Duchon, Lionel Eyraud-Dubois and Hubert Larcheveque, "A Distributed Algorithm for Resource Clustering in Large Scale Platforms", Principles of Distributed Systems, Vol.5401, pp.564-567, 2008.

[23] Samuel Kaski, Janne Nikkila¨ , Janne Sinkkonen, Leo Lahti, Juha E.A. Knuuttila, and Christophe Roos," Associative Clustering for ExploringDependencies between Functional Genomics Data Sets", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 2, NO. 3,pp: 203-216, 2005.

[24] Yao Yuhui, Chen Lihui, Andrew Goh, Ankey Wong, " CLUSTERING GENE DATA VIA ASSOCIATIVE CLUSTERING NEURAL NETWORK", Proceedings of the 9th International Conference onNeural Information Processing, Vol.5, pp: 2228- 2232, 2002.

[25] Hesam Izakian, Ajith Abraham, Vaclav Snasel, "Fuzzy Clustering Using Hybrid Fuzzy c-means and Fuzzy Particle Swarm Optimization", World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), India, IEEE Press, pp. 1690-1694, 2009.

[26] Swagatam Das, Ajith Abraham, Amit Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm", IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems And Humans, Vol. 38, No. 1, 2008.