IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 7, July 2016
ISSN (Online) 2348 – 7968 | Impact Factor (2015) - 4.332
www.ijiset.com

# Agglomerative clustering on vertically partitioned data

**R.Senkamalavalli**
Research Scholar,
Department of Computer Science and Engg.,
SCSVMV University, Enathur,
Kanchipuram 631 561
sengu_cool@yahoo.com

**Dr.T.Bhuvaneswari**
Assistant Professor,
Department of Computer Science and Applications,
Queen Mary's College (autonomous),
Mylapore, Chennai – 600 004
t_bhuvaneswari@yahoo.com

***Abstract* — Mining distributed databases is emerging as a fundamental computational problem. A common approach for mining distributed databases is to move all of the data from each database to a central site and a single model is built. Privacy concerns in many application domains prevents sharing of data, which limits data mining technology to identify patterns and trends from large amount of data. Traditional data mining algorithms have been developed within a centralized model. However, distributed knowledge discovery has been proposed by many researchers as a solution to privacy preserving data mining techniques. By vertically partitioned data, each site contains some attributes of the entities in the environment. In this paper, we present a method for Agglomerative clustering algorithm in situations where different sites contain different attributes for a common set of entities for vertically partitioned data. Using association rules data are partitioned into vertically.**

***Keywords - Data mining; agglomerative clustering; distributed data; association rule.***

## I. INTRODUCTION

Data Mining is the technique used by analysts to find out the hidden and unknown pattern from the collection of data. Although the organizations gather large volumes of data, it is of no use if "knowledge" or "beneficial information" cannot be inferred from it. Unlike the statistical methods the data mining techniques extracts interesting information. The operations like classification, clustering, association rule mining, etc. are used for data mining purposes.

The term data distribution means the manner in which the data has been stored at the sites (DB servers). Primarily there are two types of data distribution i) Centralized Data and ii) Partitioned Data. In a centralized data environment all data is stored at single site. While in distributed environment all data is distributed among different sites. Distributed data can further be divided in i) Horizontally and ii) Vertically distributed environments (Fig.1). In horizontal distribution the different sites stores the same attributes for different sets of records. In vertical distribution the sites stores different attributes for the same set of records.

By vertically partitioned, we mean that each site contains some elements of a transaction. Using the traditional online example, one site may contain book purchases, while another has electronic purchases. Using a key such as credit card number and date, we can join these to identify relationships between purchases of books and electronic goods. However, this discloses the individual purchases at each site, possibly violating consumer privacy agreements.

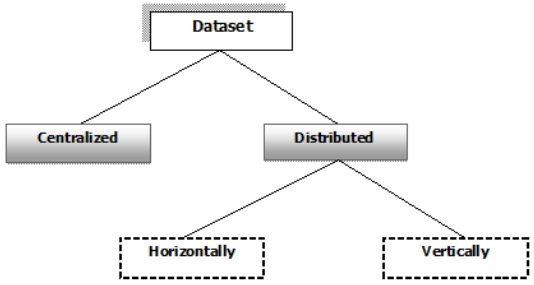


Fig. 1. Classification of dataset

Clustering is the method by which like records are grouped together. Usually this is done to give the end user a high level view of what is going on in the database. Clustering is sometimes used to mean segmentation. Technically it can be defined as the task of grouping a set of objects in such a manner that objects in same group (cluster) are more similar to each other than to those in other groups (clusters). It is main task carried out for machine

learning, pattern recognition, information retrieval, etc. Clustering can be partitioned in i) Hierarchical ii) Partition Based iii) Density Based Clustering (Fig.2).

The hierarchy of clusters is usually viewed as a tree where the smallest clusters merge together to create the next highest level of clusters and those at that level merge together to create the next highest level of clusters.

This hierarchy of clusters is created through the algorithm that builds the clusters. There are two main types of hierarchical clustering algorithms:

- Agglomerative - Agglomerative clustering techniques start with as many clusters as there are records where each cluster contains just one record. The clusters that are nearest each other are merged together to form the next largest cluster. This merging is continued until a hierarchy of clusters is built with just a single cluster containing all the records at the top of the hierarchy (Fig.3).

- Divisive - Divisive clustering techniques take the opposite approach from agglomerative techniques. These techniques start with all the records in one cluster and then try to split that cluster into smaller pieces and then in turn to try to split those smaller pieces.

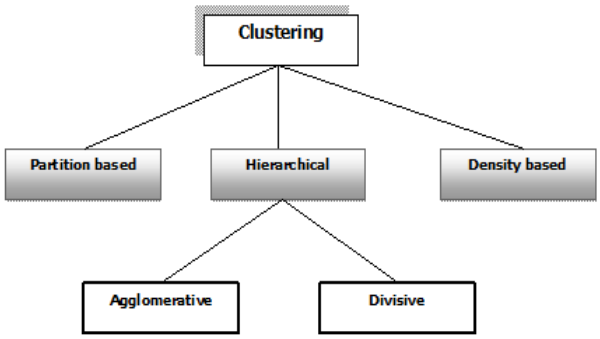


Fig. 2. Categories of Clustering

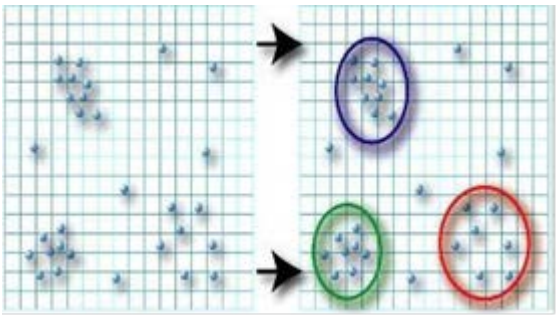

Fig. 3: Concept of Clustering

## II. RELATED WORK

Vaidya et.al [1] presents the k-means technique to preserve privacy of vertically partitioned data. Hwanjo Yu et.al [2] suggests an algorithm for privacy preservation for Support Vector Machines (SVM) based classification using local and global models. Local models are relevant to each participating party that is not disclosed to others while generating global model jointly. Global model remains the same for every party which is then used for classifying new data objects. Liu et.al [3] represents two protocols for privacy preserving clustering to work upon horizontally and vertically partitioned data separately. Inan et.al [4] suggest methods for constructing dissimilarity matrices of objects from different sites in privacy preserving manner. Krishna Prasad et.al [5], mentioned a procedure for securely running BIRCH algorithm over arbitrarily partitioned database. Secure protocols are mentioned in it for distance metrics and procedure is suggested for using these metrics in securely computing clusters. Pinkas [6] represents various cryptographic techniques for privacy preserving. Vaidya [7] presents various techniques of privacy preserving for different procedures of data mining. An algorithm is suggested for privacy preserving association rules. A subroutine in this work suggests procedure for securely finding the closest cluster in k-means clustering for privacy preservation. Nishant [8] suggests scaling transformation on centralized data to preserve privacy for clustering. K-means clustering [9, 10] is a simple technique to group items into k clusters. The basic idea behind k-means clustering is as follows: Each item is placed in its closest cluster, and the cluster centers are then adjusted based on the data placement. This repeats until the positions stabilize.

## III. AGGLOMERATIVE CLUSTERING ON VERTICALLY PARTITIONED DATA

The proposed work is about vertically portioned data mining using clustering technique. In this system, we consider the heterogeneous database scenario considered a vertical partitioning of the database between two parties A and B (Fig.4). The association rule mining problem can be formally stated as follows:

Let $I = i_1, i_2, \ldots i_p$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, called its TID (Transaction Identifier). We say that a transaction T contains X, a set of some items in I, if $X \subseteq T$. An association rule is an implication of the form, $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \phi$. The rule $X \Rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y . The rule $X \Rightarrow Y$ has support s in the transaction set D if s% of transactions in D contain $X \cup Y$.

In this clustering of the Databases will be done so that the responsibility of finding a frequent n item set can be distributed over clusters which will increase the response time as well as decrease the number of messages need be passed and thus avoiding the bottleneck around central site.

The vertically partitioned data are clustered into n clusters using agglomerative clustering algorithm. Then these clusters are separated into n databases. Finally, this information can be stored in external data base for further usage.

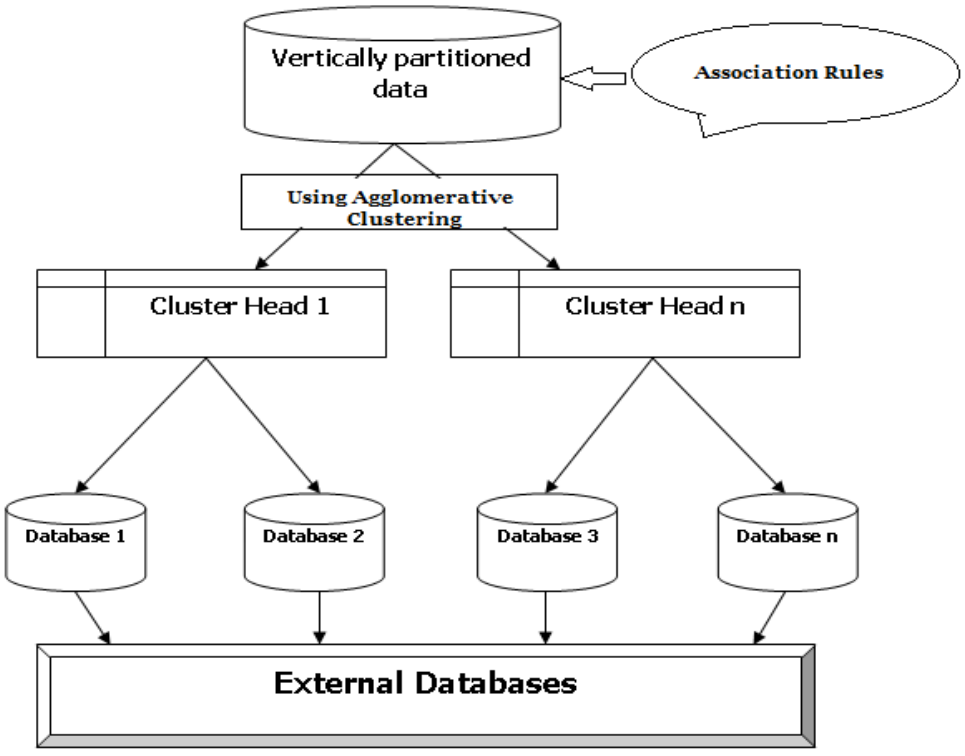


Fig. 4. Proposed system architecture

## IV. AGGLOMERATIVE CLUSTERING ALGORITHM

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc.

A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

### *A. Algorithmic steps*

Let X = {x1, x2, x3, ..., xn} be the set of data points.
Step 1:
Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.

Step 2:
Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.

Step 3:
Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)].

Step 4:
Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).

Step 5:
If all the data points are in one cluster then stop, else repeat from step 2.

## V. FINDINGS AND RESULTS

The work was implemented in one way clustering and efficiency of data obtained was not accurate. This proposed method can be implemented in two way clustering technique which will give better results.

## CONCLUSION

This proposed architecture has been implemented in future for further development using the data mining Toolbox under WEGA Software.

Clustering is an unsupervised learning technique, and as such, there is no absolutely correct answer. For this reason and depending on the particular application of the clustering, fewer or greater numbers of clusters may be desired. The future enhancement of this is to add global caching as caching can be used since data in warehouse tend to change a little over time. Techniques of clustering the databases can be debated upon, since more efficient the division of sites, more efficient will be association rules.

## REFERENCES

[1] J. Vaidya and C. Clifton, “Privacy preserving K-means clustering over vertically partitioned data”, Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining Washington, DC, pp. 206-215, august 24-27 2003.
[2] H. Yu, J.Vaidya and X. Jiang, “Privacy preserving SVM classification on vertically partitioned data ", Adv in Knowledge Disc Data Min. vol. 3918, pp. 647-656, 2006
[3] J. Liu, J. Luo, J. Z. Huang and L. Xiong, "Privacy preserving distributed DBSCAN clustering", PAIS 2012, vol 6, pp.177-185.
[4] A. Inan, S. Kaya, Y. Saygin, E. Savas, A. Hintoglu and A. Levi , "Privacy preserving clustering on horizontally partitioned data", PDM, vol 63, pp. 646-666, 2006.
[5] P. Krishna Prasad and C. Pandu Rangan, "Privacy preserving BIRCH algorithm for clustering over arbitrarily partitioned databases", pp. 146-157, August 2007.
[6] B. Pinkas, "Cryptographic techniques for privacy preserving data mining". Interntional Journal of Applied Cryptography (IJACT) 3(1): 21-45 ,2013.
[7] J. S. Vaidya. “A thesis on privacy preserving data mining over Vertically Partitioned Data". (Unpublished)
[8] P. Khatri Nishant, G.Preeti and P. Tusal, “Privacy preserving clustering on centralized data through scaling transfermation”.International journal of computer engineering& technology (IJCET).vol4,Issue3,pp 449-454,2013.
[9] R. Duda and P. E. Hart. Pattern classification and scene analysis. Wiley, New York.1973.
[10] K. Fukunaga. Introduction to statistical pattern recognition. Academic Press, San Diego, CA, 1990.
[11] S. Paul, “An optimized distributed association rule mining algorithm in parallel and distributed data mining With Xml data for improved response time”, Int J Comp Sci Info Technol. vol. 2, pp. 2, April 2010.
[12] R. J. Gil-Garcia, J. M. Badia-Contelles, and A. Pons-Porrata. Extended star clustering algorithm. Lecture Notes Comp Sci. vol. 2905, pp.480-487, 2003.
[13] G. Lance and W.Williams. A general theory of classificatory sorting strategies. 1:Hierarchical systems. Comp J, vol. 9, pp.373-380, 1967.

[14] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. KDD vol. 99, pp. 16-22, 1999.
[15] A. Pons-Porrata, R. Berlanga-Llavori, and J. Ruiz-Shulcloper. On-line event and topic detection by using the compact sets clustering algorithm. J Intelli Fuzzy Sys, vol. 3-4, pp. 185-194, 2002.

[16] K. Wagstaff and C. Cardie, "Clustering with instance-level constraints", Proceedings of the 17th International Conference on Machine Learning (ICML 2000), Stanford, CA, pp. 1103-1110, 2000.
[17] I. Davidson and S. S. Ravi, "Clustering with constraints and the k-Means algorithm", the 5th SIAM Data Mining Conf. 2005.
[18] I. Davidson and S. S. Ravi, "Hierarchical clustering with constraints: Theory and practice", the 9th European Principles and Practice of KDD, PKDD 2005.
[19] I. Davidson and S. S. Ravi, "Intractability and clustering with constraints", Proceedings of the 24th international conference on Machine learning, 2007.