Study on clustering security – in Distributed Databases

P. Ganesh #[1], M.Sumathi#[2], D. Prabakar #[3], Dr. T. Kalaikumaran#[4]

[1#] *Assistant Professor, Dept of CSE, SNS College of Technology-India*
E-Mail ID: ganeshdino@gmail.com
[2]*# Assistant Professor, Dept of CSE, SNS College of Technology-India*
E-Mail ID: sumathitvasagam@gmail.com
[3]*# Assistant Professor, Dept of CSE, SNS College of Technology-India.*
E-Mail ID: prabakaralam@gmail.com
[4#] *Professor & Head, Dept of CSE, SNS College of Technology-India.*
E-Mail ID: proftkalaikumaran@gmail.com

***Abstract* - Currently, a growing number of companies have strived to obtain a competitive advantage through participation in corporative organizations. However, no company wants to share information about their customer and transact business with other companies and even competitors, because it is needed to maintain commercial confidentiality and due to local legislation matters. Hence, a large number of studies in this research area, called privacy preserving data mining – where security and confidentiality of data must be maintained. A comprehensive review of these studies is presented below.**

***Key Words* - *Data* Sets, Cluster Ensemble, Data Structures.**

I. INTRODUCTION

Clustering is the process of discovering groups within high-dimensional databases, based on similarities, with a minimal knowledge of their structure. Traditional clustering algorithms perform it over centralized databases, however, recent applications require datasets distributed among several sites[1] . The gathering of all distributed databases in a central unit, followed by algorithm application, it is important to take into consideration some issues, namely: The possibility of existence of similar data with different names and formats.

(i) Differences in data structures.
(ii) Conflicts between one and another database.

On the other hand, integration of several database in a single location is not suggested, when it is composed of very large databases. Due to all of these problems related to database integration, research for algorithms. That perform data mining in a distributed way is demanding the methods with the ability to process clustering securely that has motivated the development of algorithms to analyze each database separately and to combine the partial results to obtain a final result. This chapter presents a wide review on privacy-preserving data clustering. Also discuss issues related to the utilization of classification and clustering ensembles. And some techniques of information merging used in literature to combine results that come from multiple clustering processes are analyzed[3][7] . Then, are discussed about security and privacy-preserving in distributed data clustering, and also present an alternative approach to this problem based on the partSOM architecture and discuss about the confidentiality of the information that is analyzed through application of this approach in geographically distributed database cluster analysis.

II. RELATED WORK

A. Data partitioning methods

There are two distinct situations that demand the need for effecting cluster analysis in a distributed way. Certain current applications hold databases so large, that it is not possible to keep them integrally in the main memory, even using robust machines. There are three approaches to solve this problem:

(i) Storing data in a secondary memory and clustering data subsets separately. Partial results are kept and, in a posterior stage, are gathered to cluster the whole set.
(ii) Using an incremental clustering algorithm, in which every element is individually brought to the main memory and associated to one of the existing clusters or allocated in a new cluster. The results are kept and the element is discarded, in order to grant space to the other one.
(iii) Using parallel implementation, in which several algorithms work simultaneously on stored data, increasing efficacy.

In cases in which the data set is unified and needs to be divided in subsets, due to its size, two approaches are normally used: horizontal and vertical partitioning (Figure 1). The first approach is more used and consists in horizontally splitting database, creating homogeneous data subsets, so that each algorithm operates on different records considering, however, the same set of attributes. Another approach is vertically dividing the database, creating heterogeneous data subsets; in this case, each algorithm operates on the same records, dealing, however, with different attributes [1][4].

In cases in which the data set is already partitioned, as in applications which possess distributed databases, besides the two mentioned approaches, it is still possible meet situations in which data is simultaneously disperse in both forms, denominated arbitrary data partitioning which is a generalization of the previous approaches.Both horizontal and vertical database partitioning are common in several areas of research, mainly in environments with distributed systems and/or databases, to which commercial application belongs.

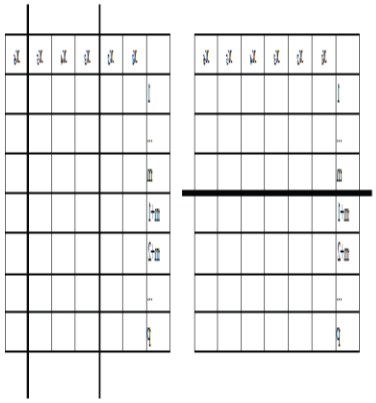

Figure1. Horizontal and vertical partitioning

When applied to distributed databases, vertical partitioning offers two great advantages which may influence system performance.

(i) The frequency of queries necessary to access different data fragments may be reduced, once that it is possible to obtain necessary information with a smaller number of SQL queries.
(ii) The amount of recovered and transferred unnecessary information in a traditional query to memory may also be reduced.

B. Cluster Ensemble

Cluster Ensemble is defined as a combination of two or more solutions come from application of different algorithms or variations of a same algorithm on a dataset, or even, on subsets thereof. There are four approaches for classifying system development [7][8], which may be extended to cluster ensemble development:

(i) Application of several instances of a same algorithm on the same database, changing the initialization parameters of the algorithm and combining its results.
(ii) Application of different clustering algorithms on a same database, intending to analyze which algorithm obtains the best data clustering.
(iii) Application of several instances of a same clustering algorithm on subsets of slightly

different samples, obtained with or without reposition.

(iv) Application of several instances of a same clustering algorithm on different subset of attributes.

In spite of all reported, both multiple classifier systems and cluster ensembles have been more and more used. Some neural classifying model denominated concurrent self organizing maps (CSOM), which is composed of a collection of small SOM networks. CSOM model present some conceptual differences from tradition SOM model – the major is in the training algorithm, which is supervised.

In tests performed with CSOM model, we consider three applications in which this model presents fair results. Face recognition, speech recognition and multi-spectral satellite images [2][8].A huge self-organizing map is divided into several parts with the same size and distributed among the machines of the cluster. A SOM cluster training architecture and methodology distributed along a computational grid, in which it is considered the ideal number of maps in the ensemble, the impact of the different kinds of data used in the training and the most appropriate period for weight updating. The authors performed a series of experiments and obtained important conclusions which can be extended to other SOM network parallel training algorithms:

If the latency time of the ensemble members the periodical weight adjusts and the synchrony time of the maps are very short, in comparison to the computational time of each training stage, the utilization of a SOM ensemble brigs about good results, regarding training time and accuracy.

(i) In the performed tests, the ideal number of maps in an ensemble was between 5 and 10 networks.

(ii) The choice of the several utilized parameters in the training (learning and decrement rate) and the frequency which calculations of the map average are also factors of great importance in reducing mean square error.

(iii) SOM ensemble presents quite superior results as the dimension of the data set increases.

Cluster ensemble application on different attribute subsets has been analyzed mainly in image segmentation. Picture SOM or PicSOM in a hierarchical architecture in which several algorithms and methods can be applied jointly for image recovering based on contents. Originally, PicSOM utilizes multiple instances of TS-SOM algorithm – which is composed by structured trees of self-organizing maps, hierarchically organized. Each TS-SOM is trained with a different set of characteristics, such as colour, texture or form. PicSOM architecture is an example of SOM network combination, whose result is a solid system for image recovery based on content similarity.

C. Combining ensemble results

A problem which is inherent to cluster ensembles is partial results combining. There are three most common approaches to solve this problem under different points of view.

(i) The first approach consists of analyzing similitude among different partitions produced through utilization of similarity metrics among partitions.

(ii) The second uses hyper-graphs to represent relationship among the objects and applies hyper-graph partitioning algorithms on them to find the clusters.

(iii) In the third approach the elements of input set are labeled and, then, labels are combined to present a final result, normally through some voting system.

D. Security and privacy preserving data mining

Data security and privacy-preserving are among the primal factors which motivate creation and maintenance of distributed database.

It is known that combining several sources of data during a KDD process increases analysis process, even though it jeopardizes security and privacy-preserving of data involved in the process .Wherefore, data mining algorithms which operate in distributed way must take into consideration not only the way data is distributed among the units, in order to avoid unnecessary

transferences, rather they must also ensure that transferred data is protected against occasional attempts of undue appropriation attempts.

A potentially Security restrictions hinder sharing information from customers among partner companies in several countries and create a series of problems related to privacy-preserving, preventing companies from adopting this strategy [7][9].

Privacy-preserving cluster analysis rises as a solution to this problem, permitting that the parties to cooperate among them in knowledge extraction, preventing obligation of each of them of revealing their individual data to the others. This approach concentrates its efforts in algorithms which assure privacy and security to data involved in the process, mainly in applications in which security has fundamental importance, for instance, in medical and commercial applications.

E. The partSOM architecture clustering process

This section presents a cluster ensemble methodology for privacy preserving clustering in distributed databases, using traditional and well known algorithms, such as self-organizing maps and K-means. The proposed methodology combines a clustering architecture, the partSOM architecture with principles of vector quantization, building a cluster ensemble model that can be used to cluster analysis in distributed environments composed by a set of partner companies involved in this process, avoiding jeopardizing the privacy of their customers [8].

The main idea of this process is focused on omission of real information about customers, changing a set of real individuals for one (or more) representative (and fictional) individual with similar statistical characteristics of the real individuals. This strategy, based on vector quantization principles, enables that a group of individuals with similar characteristics to be able to be represented by a single individual (vector) corresponding to that group.

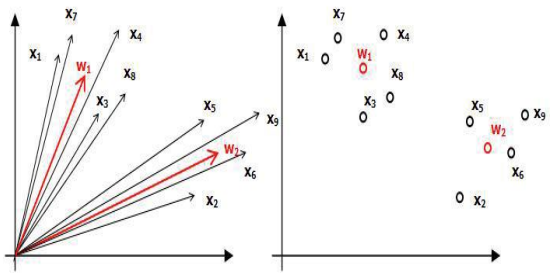

Figure 2. Example of a vector quantization process in a bidimensional plan

As illustrated in Figure 2, the vectors {**x**1, **x**3, **x**4, **x**7, **x**8} can be represented by **w**1 vector and {**x**2,**x**5, **x**6, **x**9} can be represented by **w**2 vector. This strategy is used to reduce the amount of space required to store or transmit a dataset and has been widely used by clustering tasks and data compression of signals, particularly voice and image. The partSOM architecture presents a strategy to carry out cluster analysis over distributed databases using self-organizing maps and K-means algorithms. This process is separated in two stages:

i) Data are analyzed locally, in each distributed unit.
ii) In a second stage, a central unit receives partial results and combines them into an overall result.

The partSOM algorithm, embedded in partSOM architecture, consists of six steps and is presented as it follows. An overview of the complete architecture is showed in Figure 3.

(i) A traditional clustering algorithm is applied in each local unit, obtaining a reference vector, known as the codebook, from each local data subset.
(ii) Each input data is compared with codebook issues and the index corresponding to the most similar vector present in the codebook is stored in an index vector. So, a data index vector is created based on representative objects instead of original objects.
(iii) Each remote unit sends the codebook and the index vector to the central unit, which will conduct the unification of all partial results.

(iv) The central unit is responsible for receiving index vectors and codebooks from each local unit and combining partial results and building a whole database. In this process, index vector issues are substituted by the similar issues in the codebook.
(v) The clustering algorithm is applied on the whole database obtained in previous step, to identify existing clusters in the collective database.
(vi) A segmentation algorithm is applied on results obtained after the final cluster process,in order to improve the quality of the visualization results.

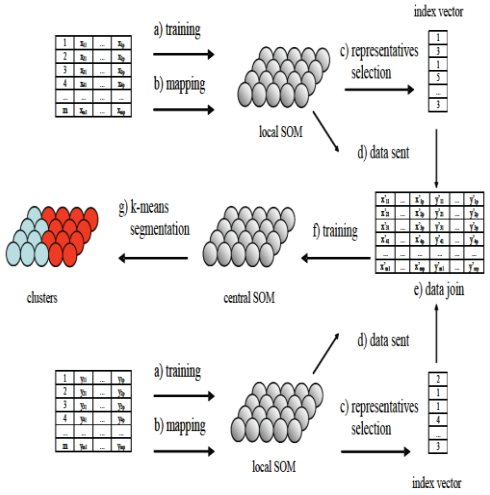

Figure 3. An overview of the partSOM architecture with SOM and K-means algorithms.

F. Some contributions to partSOM clustering process

This section presents some contributions to increase security and privacy preserving in a Clustering process using the partSOM architecture.

G. The pre-processing stage

In real world applications, raw data usually are named *dirty data*, because they can contain errors, missing values, redundant information or are incomplete and inconsistent. So, most of data mining process needs a pre-processing stage that objectives to carry out tasks such as data cleaning, data integration and transformation, data reduction, although this important step is sometimes neglected in data mining process.

Conventionality, a relational database is a set of two-dimensional tables interrelated by one or more attributes. Each table is a two-dimensional structure of rows and columns where each row represents a record from the database and each column represents an attribute associated with that record. Figure 4 suggests a sample of a typical table in a database.After pre-processing stage, data are usually arranged in single table known as data matrix, which must satisfy the requirements of the chosen algorithm. The data matrix **D** is formed by a set of *n* vectors, where each vector represents an element of the input set. Each vector has *p* components, which correspond to the set of attributes that identify it [3][6]. A data matrix example, related to the previous presented table in Figure 4, is shown in table 1.In this example, some attributes were removed, others were transformed and the whole dataset was normalized. As discussed in literature, this stage contributes to privacy and security maintenance of data and information stored in database, because real data are replaced by a set of representatives with same statistical distribution of original data. Thus, since only codebook and index vector are sent to the central unit and no real information is transferred, the security is maintained.

H. The pruning algorithm

In terms of partSOM architecture, the most suitable algorithm during the initial codification stage in the local units is the self-organizing maps. In this case, the codebook may contain a few entries with little or no representation in the input set, known as dead neurons. These elements occur with some frequency in clustering processes using the SOM, what has been cited in the literature.

In terms of K-means algorithm, codebook elements with little representation may correspond to outliers or noise in the input data and, eventually, these elements can be discarded from representatives set without great impairment to the maintenance of the statistical distribution of data. So, in both cases, it is possible to include a pruning algorithm in a stage before the transfer of data to the central unit, to reduce the size of the codebook and avoiding moving items that are not used (or are not relevant) in data reconstruction [4].

The procedure for reducing the codebook is performed by a pruning algorithm (Figure 4), which will be detailed below.

The pruning algorithm receives the input dataset **X** = {**x**1, **x**2, ..., **x**N}, the trained codebook **W** = {**w**1,**w**2, ..., **w**k}, the set of representatives **R** and an integer value T, which corresponds to the representation threshold required for each element. Then, the algorithm searches for elements whose representation is less than or equal to the threshold and eliminates them from the codebook. Finally, the representative choice algorithm is called again to reselect the representatives of each input dataset.

#	Name	Sex	Age	Wage	Civil State	Children	…	State
1	A.Araújo	M	39	2.300,00	Married	3	..	RN
2	Q.Queiroz	F	82	1.350,80	Widowed	2	..	PB
3	W.Wang	M	21	720.50	Single	1	..	CE
4	E.Eudes	F	18	1.420,00	Single	0	..	SP
5	S.Silva	M	16	450,00	Single	0	..	RN
6	G.Gomes	M	42	32.827,52	Married	2	..	DF
7	K.Key	F	38	410.50	Divorced	1	..	SE
	…	…	…	…	…	…	..	…
N	M .Mendes	M	21	3.500,00	Married	4	..	BA

Table 1. Sample of a typical table in a database

Importantly, the pruning algorithm is an optional step, whose objective is to reduce the amount of data transferred between the remote units and central unit. In the particular case in which the threshold value is zero, only the inactive neurons are eliminated without any change in the outcome.

Pruning Algorithm

Input: input dataset (X); original *codebook* (W); representatives set (R); threshold (T)
Output: modified *codebook* (W'); modified representatives set (R')
procedure pruning(X,W,R,T)
for each wj €□W
cont = 0
for each ri €□R
if (R[i] = j)
cont = cont + 1
endif
endfor
if (cont <= T)
W' = remove(W,j)
endif
endfor
R' = choose_representative(X,W')
return(W',R')

Figure 4. The pruning procedure algorithm

I. The covariance matrix

The first step in partSOM architecture uses a vector quantization process to effect a compression in the input data and thus reduce the amount of data transferred to the central unit. As in any process of data compression, there are losses associated with vector quantization and, possibly some of the information existing in the input data is discarded during the first stage of the algorithm [5] .To minimize the losses occurring in the process of vector quantization is the use of additional statistical information contained in the original sample, so that the reconstructed data are as similar as possible to the input data. The covariance matrix of a set of data allows extracting the variance and correlation between the samples, and an efficient solution to create random samples containing the same statistical characteristics of the original sample.Thus, if the covariance matrices of each cluster are drawn in remote units and sent along with the codebooks, so that each centroid can carry information about the variance of the data that it represents, and this information could be used to generate samples with a statistical distribution even more similar to the original dataset, helping to reduce losses associated with the process of vector quantization.

III. Conclusion

This chapter discussed the utilization of cluster ensembles in data clustering and classification tasks. Matters related to existence of geographically distributed databases and mechanisms used for data partitioning were analyzed. It was also presented a wide review on algorithms and strategies used in data mining, mainly in clustering tasks.

Following, matters related to distributed data clustering security and privacy were addressed. Eventually, some information fusion techniques used to combine results come from multiple clustering solutions were cited in reviewed works. The partSOM architecture was presented as a proposal for performing cluster analysis on geographically distributed databases, such as discussed in previous works.

However, this study focused specifically on issues related to security and privacy preserving in distributed databases clustering. The main contribution of this work was a bibliographic review about the theme and a discussion about some techniques that can be used in a privacy preserving distributed databases clustering process, including:

i) A data pre-processing stage, which objectives to remove all information that could be used to identify an individual.
ii) A pruning algorithm to reduce the amount of data transferred between the local and central units.
iii) The use of a covariance matrix from each local data unit to reduce losses during the process of vector quantization.

Future research directions will be focused on extent the partSOM architecture, including use of others privacy-preserving strategies. Furthermore, it is necessary to apply and to evaluate this model in real world applications.

ACKNOWLEDGMENT

Our Sincere thanks to our most respective Professor and Dean of the Department of Computer science and Engineering Dr.S.Karthik, Our beloved professor and Head of the Department Dr.T.KalaiKumaran and the people who are mainly motivate to prepare and publish this article in esteemed journal.

REFERENCES

[1] Abhijith Kashyap, Vagelis Hristidis, Michalis Petropoulos, and Sotiria Tavoulari,"Effective Navigation of Query Results Based on Concept Hierarchies", ieee transactions on knowledge and data engineering, vol. 23, no. 4, april 2011

[2] Agrawal, S. Chaudhuri, G. Das and A. Gionis: Automated Ranking of Database Query Results. In Proceedings of First Biennial Conference on Innovative Data Systems Research (CIDR),2003.

[3] Bast, A. Chitea, F. Suchanek, and I. Weber, "ESTER: Efficient Search on Text, Entities, and Relations," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2007.

[4] Demidova, X. Zhou, and W. Nejdl, "IQP: Incremental Query Construction, a Probabilistic Approach," Proc. 26th IEEE Int'l Conf.Data Eng. (ICDE), 2010.

[5] Demidova, X. Zhou, G. Zenz, and W. Nejdl, "SUITS: Faceted User Interface for Constructing Structured Queries from Keywords," Proc. 14th Int'l Conf. Database Systems for Advanced Applications (DASFAA), 2009.

[6] Elena Demidova, Xuan Zhou, and Wolfgang Nejdl, Member, IEEE Computer Society "A Probabilistic Scheme for Keyword-Based Incremental Query Construction" ieee transactions on knowledge and data engineering, vol. 24, no. 3, march 2012

[7] Gideon Zenz , Xuan Zhou , Enrico Minack , Wolf Siberski "From Keywords to Semantic Queries Incremental Query construction on semantic web", Elsevier 18 July 2009.

[8] He, H. Wang, J. Yang, and P.S. Yu, "BLINKS: Ranked Keyword Searches on Graphs," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2007.

[9] Hristidis and Y. Papakonstantinou, "DISCOVER: Keyword Search in Relational Databases," Proc. Int'l Conf. Very Large Data Bases (VLDB), 2002.