Secured disclosure of data in multiparty clustering

G. Kirubhakar [1] , Dr. C. Venkatesh [2]

[1.] Department of Computer Science and Engineering, Surya Engineering College, Erode, Tamilnadu, India
[1.] ergalaxy81@gmail.com, kirubhakarg@gmail.com
[2.] Dean, Faculty of Engineering,Erode Builder Educational Trust's Group of Institutions, Kangayam, Tamilnadu, India
[2.] dean.foe@ebet.edu.in

Abstract: Data mining can extract important knowledge from large data collections sometimes these collections are distributed among multiple parties. Privacy concerns may prevent the parties from directly sharing the data and some type of information about data. This work presents a distributed privacy-preserving k-clustering. K-means were used for clustering and that will be applied to the data bases that are distributed between many parties. The participants of the protocol learn only the final cluster centers on completion of the protocol. It uses data perturbation techniques for securing the information about data.

[G. Kirubhakar, C. Venkatesh. **Secured disclosure of data in multiparty clustering.** *Life Sci J* 2012;9(4):2580-2585]. (ISSN: 1097-8135). http://www.lifesciencesite.com. 383

Keywords: security, privacy, data dissemination, clustering, multiparty

1. Introduction

With the development of data analysis and processing technique, organizations, industries and governments are increasingly publishing micro data (i.e., data that contain non aggregated information about individuals) for data mining purposes, studying disease outbreaks or economic patterns. While the released datasets provide valuable information to researchers, they also contain sensitive information about individuals whose privacy may be at risk (P. Samarati, 2001).

Now a day's these data bases are distributed among several sites. Data mining techniques on distributed data bases however reveals sensitive information about individuals. Here the concept of securing data mining comes.

Securing distributed data mining allows cooperative computation of data mining algorithms without requiring the participating organizations to reveal their individuals data items to each other (D.Aruna, 2011).

2. Problem Definition

Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from the point of view of privacy preservation. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. However, there are situations where the sharing of data can lead to mutual benefit.

Despite the potential gain, this is often not possible due to the confidentiality issues which arise. It is well documented that the unlimited explosion of new information through the Internet and other media has reached a point where threats against privacy are very common and deserve serious thinking.

Consider a scenario that there are several hospitals involved in a multi-site medical study. Each hospital has its own data set containing patient records. These hospitals would like to conduct data mining over the data sets from all the hospitals with the goal of obtaining more valuable information via mining the joint data set. Due to privacy laws, one hospital cannot disclose their patient records to other hospitals. How can these hospitals achieve their objective? Can privacy and collaborative data mining coexist? In other words, can the collaborative parties somehow conduct data mining computations and obtain the desired results without compromising their data privacy? We show that privacy and collaborative data mining can be achieved at the same time.

Common examples arise in health science, where data may be held by multiple parties: commercial organizations (such as drug companies, or hospitals), government bodies (such as the Food and Drug Administration) and non-government organizations (such as charities). Each organization is bound by regulatory restrictions (for instance privacy legislation), and corporate requirements (for instance on distributing proprietary information that may provide commercial advantage to competitors). In such a case, an independent researcher may not receive access to data at all, while even members of one of these organizations see an incomplete view of the data. However, data from multiple sources may be needed to answer some important questions. A classical example occurs for an organization like the CDC (Center for Disease Control and Prevention), who are mandated with detecting potential health threats, and to do so they require data from a range of sources (insurance companies, hospitals and so on), each of whom may be reluctant to share data.

3. Literature Survey

The problem of secured data mining has found considerable attention in recent years because of the recent concerns on the privacy of underlying data (V.S.Verykios, 2004).

Various secured data mining techniques fall under:

- *K*-Anonymity
- Cryptographic techniques
- Randomized Response techniques
- Data modification

Many recent papers on privacy have focused on the perturbation model and its variants. Methods for inference attacks in the context of the perturbation model have been discussed by Acerkerman.M.S (1999).

A number of papers have also appeared on the *k*-anonymity model recently. Other related works discuss the method of top-down specialization for privacy preservation, and workload-aware methods for anonymization (W.Du, 2004).

Agrawal (2000) develops a new distribution-based data mining algorithm for the classification problem, whereas the techniques in Vaidya and Clifton (2002) and Rizvi and Haritsa (2002) develop methods for privacy-preserving association rule mining.

Another branch of privacy preserving data mining which uses cryptographic techniques was developed (S.Laur, 2006).

Randomized Response technique was first introduced by Warner as a technique to solve a survey problem (H.Polat, 2005).

In condensation approach, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way ensure high privacy protection (V.S. Verykios, 2004).

The goal of this paper is to present technologies to solve security related data mining problems over large data sets at multiple sites or parties with reasonable efficiency.

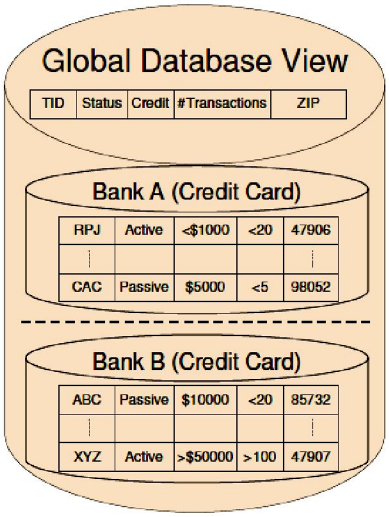
4. Materials and Methods

4.1. Cluster Analysis

Clustering is an important data mining problem. The goal of clustering, in general, is to discover dense and sparse regions in a dataset. Most previous work in clustering focused on centralized data whose inherent geometric properties (V.S. Verykios, 2004) can be exploited to naturally define distance functions between points. Recently, the problem of clustering at distributed sites started receiving interest.

4.2. Proposed System

1) Data partitioning methods: There are two distinct situations that demand the need for effecting cluster analysis in a distributed way. The first occurs when the volume of data to be analyzed is relatively great, which demand a considerable computational effort, which sometimes is even unfeasible, to accomplish this task.



TID	Status	Credit	#Transactions	ZIP

RPJ	Active	<$1000	<20	47906
⋮				⋮
CAC	Passive	$5000	<5	98052

ABC	Passive	$10000	<20	85732
⋮				⋮
XYZ	Active	>$50000	>100	47907

Fig.1. Database partitioning

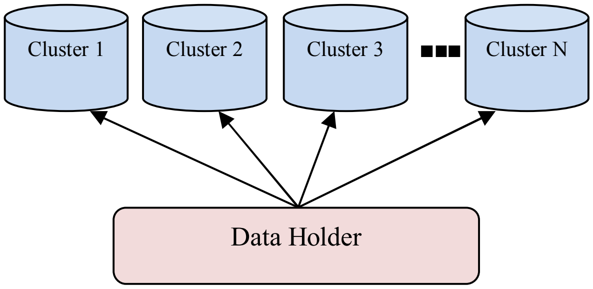


Fig.2. Multiparty clustering

The best alternative, then, is splitting data, cluster them in a distributed way and unify the results. The second occurs when data is naturally distributed among several geographically distributed units and the cost associated to its centralization is very high as in Fig. 1. Certain current applications hold databases so large, that it is not possible to keep them integrally in the main memory, even using robust machines.

Kantardzic (2002) presents three approaches to solve this problem:

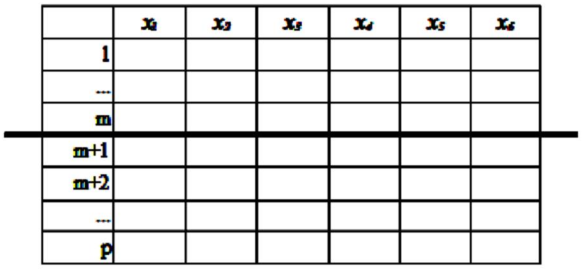
a) Storing data in a secondary memory and clustering data subsets separately. Partial results are kept and, in a posterior stage, are gathered to cluster the whole set;

b) 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 as shown in Fig. 2.;

c) 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 (Fig. 3 and 4).

Fig.3. Horizontal partitioning



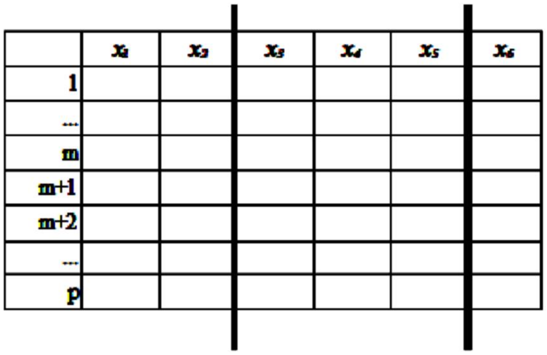


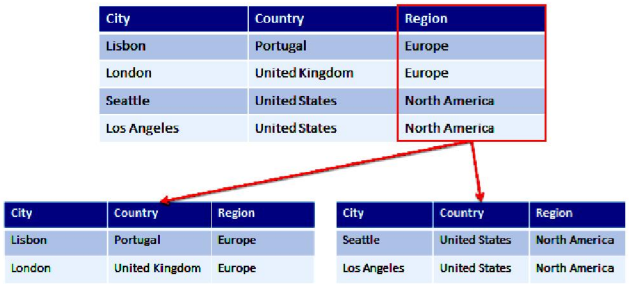
Fig.4. Vertical partitioning

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.

2) General aims of partitioning and placement: Before proceeding, a few definitions of the terminology used are in order. Partitioning (also known as fragmentation) is the fragmentation of a relational table into subsets, called partitions as in Fig. 5. Placement is the assignment of these partitions to physical storage media. The collective term for these is allocation. Note that some workers use the term partitioning to mean allocation. The general aims of data partitioning and placement in database machines are to

1. reduce workload (e.g. data access, communication costs, search space)
2. balance workload
3. speed up the rate of useful work (e.g. frequently accessed objects in main memory)

3) Data perturbation: It can be broadly divided into two sets of techniques – probability distribution, which is not dealt with here, and fixed-data perturbation, designed specifically for numerical/categorical (not statistical) data. Fixed-data perturbation methods usually generate an entirely new database, for secondary use. In their simplest form, only a single attribute is perturbed – but techniques also exist for the transformation of multiple attributes.



City	Country	Region
Lisbon	Portugal	Europe
London	United Kingdom	Europe
Seattle	United States	North America
Los Angeles	United States	North America

City	Country	Region
Lisbon	Portugal	Europe
London	United Kingdom	Europe

City	Country	Region
Seattle	United States	North America
Los Angeles	United States	North America

Fig.5. Partitioned entities

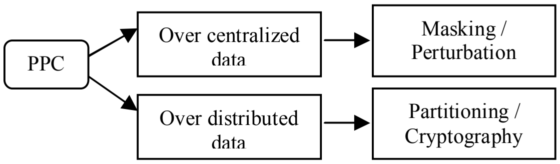


Fig.6. A taxonomy of data mining security solutions

4.3. Security Primitives and Data Perturbation

Taxonomy of data mining security solutions is shown in Fig. 6 and perturbation is one among them which is taken into account in this work.

4.4. Algorithms

The central idea of general k-means clustering (Algorithm 1) is as follows:

```
Algorithm 1: Centralized K-means Clustering
1. Select K points as initial centroids
2. repeat
3.    Form K clusters assigning all points to closest centroid
4. Recompute the centroid of each cluster
```

The central idea of proposed algorithm (Algorithm 2) is similar to the one proposed by (D.Aruna, 2011). The algorithm is as follows: This algorithm is used as a function in k-means clustering algorithm to securely compute the nearest cluster for the given entity i.e. to which cluster should an entity to be assigned. This algorithm is invoked for every single entity in each iteration. Each party has its input data, the distance component corresponding to each of the k clusters i.e. which is equivalent of having a matrix of distances of dimensions r*k. Euclidean distance measure is used to compute the distance between the entity and the cluster centroid.

```
Algorithm 2: Multiparty K-means Clustering
1. DH splits and secures data by random perturbation
2. DH distributes secured data to all parties
3. Each party performs clustering separately
4. Each party returns intermediate centroids
```

Data Holder(DH) applies the random perturbation on the splitted data and publishes to all other parties. Each party selects randomly m number of entities from the dataset as initial starting points and every party computes the local distances of their attributes for k clusters. Since there are many parties and each of them sends the computed intermediate centroid to party DH.

Then compute closest cluster algorithm is described in Algorithm 3 below.

```
Algorithm 3: Merge and Compute closest cluster
1. DH uses centroids collected as initial centroids
2. Performs clustering on the whole data
3. DH computes final centroids
```

DH combines all randomized masked local distances with respect to each entity and compares the masked local distances with k cluster and assigns the entity to the closest cluster. The final output of privacy preserving k-means clustering algorithm is that all the parties will know to which cluster each entity is assigned.

5. Measurements

This experiment is performed in two phases. In the first phase, the data mining task-clustering is performed centralized without securing the sensitive details. In the second phase, the same data mining task-clustering is performed among multiparty in a distributed manner by securing the sensitive attributes.

For Youtube dataset, the data quality of the secured dataset is then compared with the data quality of the original dataset for estimating the effectiveness of secured disclosure in preserving the patterns.

The same experiment performed in many runs at varying entity counts, security levels, party counts and accuracy levels are noted. The attributes of the entities are distributed among parties equally or unequally, which does not show any effect on the algorithms. The proposed algorithm is applied to Youtube user dataset consist of 5000 entities and 4 attributes for each entity. The clustering results in both centralized and multiparty modes are shown below.

The various experimental results are shown for 5000 records in Tables 1-4 and Fig. 7 and 8, which conclude that the results are likely to be fine at entity count more than 1000 having multiparty cluster levels around 5 with perturbation security level around 10%.

6. Results

Data set taken: Youtube data set
Number of attributes: 4
Sensitive attributes: Uploads, Watches
Data mining task examined: Clustering Number of entities: 5000
Number of parties: 5
Security method: Partitioning and Perturbation
Security level: 10%

6.1. Detailed results

6.1.1. Centralized clustering

32 iterations
Elapsed Time - 34.468750

Table 1: Centralized clustering detailed results

Clusters	Centroids	Entities
C1	17, 2453	393
C2	28, 5236	105
C3	9, 167	3445
C4	28, 21474	9
C5	15, 1002	1048

6.1.2. Multiparty Clustering

Table 2: Multiparty clustering detailed results

Sites	Initial Centroids
Site1	11,668
Site2	11,689
Site3	9,611
Site4	13,602
Site5	11,764

6.2. Merging multiparty results

32 iterations
Elapsed Time - 30.671875

6.3. Brief comparison results

Table 3: Time elapsed comparison results

Entities	Centralized	Distributed
6	1.28	1.28
10	3.55	3.16
500	10.54	9.53
1000	22.13	21.75
5000	34.47	30.67

Table 4: Iterations comparison results

Entities	Centralized	Distributed
6	3	2
10	6	5
500	10	11
1000	20	20
5000	32	34

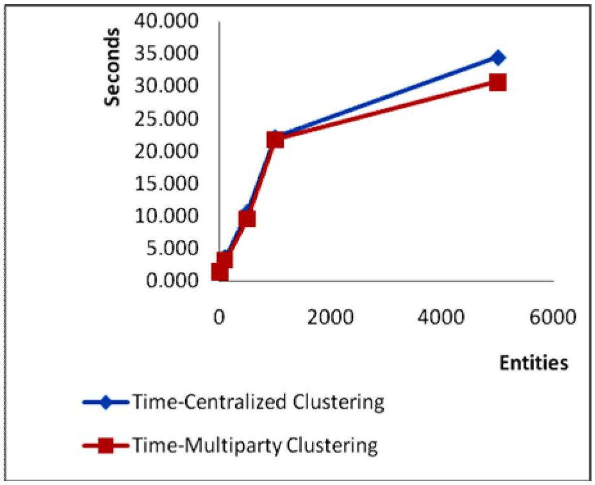


Fig. 7: Graphical Results: Entities vs. Time

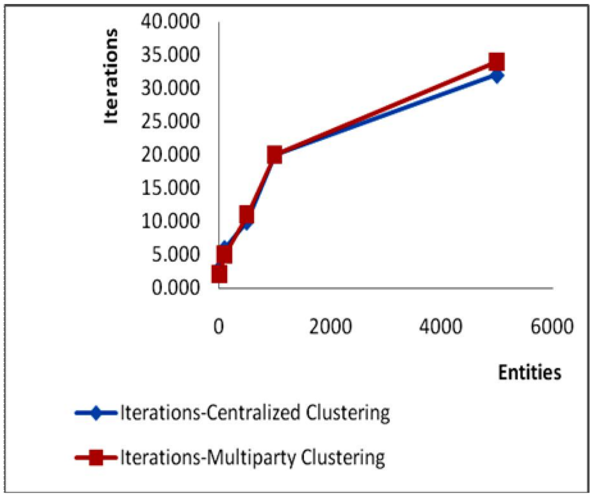


Fig. 8: Graphical Results: Entities vs. Iterations

Experiments show that this method can greatly improve the privacy quality without sacrificing accuracy.

7. Conclusion

It appears that complete privacy is impossible to maintain while allowing useful data-mining. To allow complete privacy makes data-mining results completely unreliable, whilst enabling accurate data-mining results in an unacceptable threat to the privacy of individuals. Data-mining is nevertheless a useful and vitally important pursuit, and thus techniques which maximize accuracy of results, while minimizing the threats to privacy, will become increasingly important.

To best knowledge this is the first effort toward a building block solution for the problem of privacy preserving data clustering. The performance evaluation experiments demonstrated that the methods are effective and provide practically acceptable values for balancing privacy and accuracy. The transformed database is available for secondary use and must hold the following restrictions: (a) the distorted database must preserve the main features of the clusters mined from the original database; (b) an appropriate balance between clustering accuracy and privacy must be guaranteed. The results of the investigation clearly indicate that the methods achieved reasonable results and are promising.

This work can be extended in two directions: (a) combining cryptography and perturbation to increase both accuracy and privacy; (b) designing new methods for privacy preserving clustering when considering the analysis of confidential categorical attributes, which requires further exploration.

Authors:

Kirubhakar.G currently working as Assistant Professor in Surya Engineering College, Erode, is a research scholar under Anna University of Technology, Coimbatore. He has received his Bachelor's degree in 2002 from Amrita Institute of Technology, Coimbatore and Master's degree in computer science from Kongu Engineering College, Perundurai during 2007. He secured **16th rank** in Anna University, Chennai during his Post graduate studies. His research interest includes data mining, web technology and security techniques.

Dr. C. Venkatesh, Dean, Faculty of Engineering, Erode Builder Educational Trust's group of Institutions, Kangayam, graduated in ECE from Kongu Engineering College in the year 1988, obtained his Master's degree in Applied Electronics from Coimbatore Institute of Technology, Coimbatore in the year 1990. He was awarded Doctorate in ECE from

Jawaharlal Nehru Technological University, Hyderabad in 2007. He has a credit of two decade of experience which includes around 3 years in industry. He has 16 years of teaching experience during tenure he was awarded **Best Teacher Award** twice. He was the founder Principal of Surya Engineering College, Erode. He is guiding 10 Ph.D., research scholars. He is a Member of IEEE, CSI, ISTE and Fellow IETE. He has Published 13 papers in International and National Level Journals and 50 Papers in International and National Level conferences. His area of interest includes Soft Computing, Sensor Networks and communication.

References

[1] P. Samarati, 2001. Protecting respondent’s privacy in micro data release. IEEE Transaction on Knowledge and Data Engineering. DOI: 10.1109/69.971193. pp: 1010 – 1027.

[2] D.Aruna Kumari, Dr.K.Rajasekhar Rao, M.Suman, May 2011. Distributed data mining: A new Approach using Steganography techniques. CIT International journal of Research and educations. ISSN: 2230-9144.

[3] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, 2004. State of the art in privacy preserving data mining. Proceedings of ACM SIGMOD. DOI: 10.1145/974121.974131. pp: 50 – 57.

[4] Ackerman, M. S., Cranor, L. F., and Reagle, J, 1999. Privacy in ecommerce: examining user scenarios and privacy preferences. Proceedings of Electronic Commerce. DOI: 10.1145/336992.336995. pp: 1-8.

[5] W. Du, Y. Han, and S. Chen, 2004. Privacy preserving multivariate statistical analysis: Linear regression and classification. Proceedings of the Fourth SIAM International Conference on Data Mining. DOI: 10.1.1.38.595. pp: 222-233.

[6] Agrawal, R. and Srikant, R, 2000. Privacy-preserving data mining. Proceedings of SIGMOD00. DOI: 10.1145/342009.335438. pp: 439-450.

[7] J. Vaidya and C. Clifton, 2002. Privacy preserving association rule mining in vertically partitioned data. Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, CA. DOI: 10.1145/775047.775142. pp: 639-644.

[8] S. Laur, H. Lipmaa, and T. Mielikainen, 2006. Cryptographically private support vector machines. Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. DOI: 10.1145/1150402.1150477. pp: 618-624.

[9] Kantardzic M., Djulbegovic B., Hamdan H., September 2002. A Data Mining Approach to Improving Polycythemia Vera Diagnostics. Computers and Industrial Engineering, Volume 43, Issue 4. pp. 765-777

[10] D.Aruna Kumari, Dr.K.Rajasekhar rao, M.Suman, May 2011. Privacy preserving clustering in data mining using vector quantization. Research journal of Computer science and engineering (RJCSE). ISSN : 2230-8563

10/22/2012