Reviewing Privacy Preserving Distributed Data Mining

Sajjad Baghernezhad, Parisa Arjmand
Islamic Azad University
Iran
Sbaghernezhad@yahoo.com, parisa.arj22@gmail.com

ABSTRACT: *Nowadays considering human involved in increasing data development some methods such as data mining to extract science are unavoidable. One f the discussions of data mining is inherent distribution of the data usually the bases creating or receiving such data belong to corporate or non-corporate persons and do not give their information freely to others. Yet there is no guarantee to enable someone to mine special data without entering in the owner's privacy. Sending data and then gathering them by each vertical or horizontal software depends on the type of their preserving type and also executed to improve data privacy. In this study it was attempted to compare comprehensively preserving data methods; also general methods such as random data, coding and strong and weak points of each one are examined.*

Keywords: Data Mining, Distributed Data Mining, Privacy Protection, Privacy Preserving

Received: 12 October 2014, Revised 14 November 2014, Accepted 19 November 2014

1. Introduction

Recently a new research field known as PPDDM appeared whose general goal is to solve following problem:
Some participants may wish to mine, execute and perform data jointly based on privacy data of which each part is preserved by one of them. The researches, personnel and developers have been interested such difficult adjustments in both data mining and security aspects. They have had many developments in solutions design and their durability to achieve the scenario. However, at present time the researchers encounter some challenges to create some standard to compose and assess different PPDDM protocols because they are confused about many developed techniques[6].

Data privacy preserving is so important that sometimes the authorities prevent to gather their data in order to mine data and analyze and discover the relations. Sometimes distributing the findings from data mining leads to serious commercial competition between the businessmen in a way that in some case it is a security subject to preserve and use such findings[7].

Deleting private data is not practical. It is not clear how we may identify private data and even if it was possible, the quality and utility of the data would most probably decrease. On the other hand, it is not enough to use only one secured source to save data though high security for a data source may protect it in some degree against abuse risks, but it should be noted that each attack against the source involves the data; data distribution mechanism may guarantee relatively all the data[8].

2. DDM (Distributed data mining)

DDM is to find semi-automatically hidden patterns in the data while the data or conclusion mechanisms are distributed between two or several sites decentralized data mean the charge to transfer all or some data to a central site is not ignorable[1&3]. Although DDM is a key solution for main problems encountered by data mining it creates some other challenges and problems, too so if such problems are solve effectively, data mining may be used more and we have new possibilities and potentials in a way that notwithstanding necessity of data mining it would have limited use [1&2].

3. DDM Architecture

Spatial data mining is used both locally and throughout the distributed sites. A sample of DDM architecture is shown in Figure 1. The first phase includes usually local database analysis in each distributed site; then the found information which is usually transferred to an integrated site where the local distribution models are integrated. The findings are transferred to the distributed database. In some methods instead of an integrated site local models are distributed to all other sites in a way that it is possible to compute in each site parallel to throughout model [1].

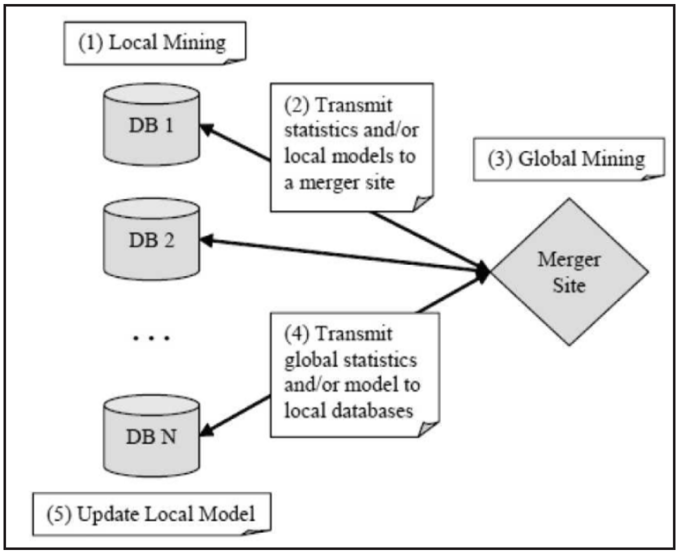
(1) Local Mining
DB 1
(2) Transmit statistics and/or local models to a merger site
(3) Global Mining
DB 2
Merger Site
. . .
(4) Transmit global statistics and/or model to local databases
DB N
(5) Update Local Model

Figure 1. A sample of DDM architecture

4. The Role of Intelligent Factors in Ddm

In DDM all methods are based on transferring some middle data by maximizing science discovery possibility and minimizing the possibility to guess raw information by others and the most important (And perhaps the only one) approach to DDM is to use the factors; they are independent software or hardware units which do some users’ duties with some autonomy.

A computer system (Software) factor which is autonomous and has some social, reaction and programming potentials for future.

5. Multi-factor Systems

The multi-factors systems are a subgroup developing from artificial intelligence to provide the principles to create complicated systems including several factors and mechanisms to harmonize the behaviors of the factors. In distributed artificial intelligence view the multi-factors system is consist of independent factors to solve problem and each factor has all the proposed specifications [4&5]. The multi-factors systems have following specifications:

- The science enough and necessary to solve a problem which is not in the factor.
- Controlling the distributed system (There is no general control system).
- The data are decentralized.
- The factors need mutual interaction to achieve their goals.

6. Benefiting Factor to Mine Data

Multi-factor systems may be taken into consideration as a type of open systems creating possibility to cooperation or competition under defined environmental rules to access common or special goal of each factor. The multi-factor systems Multi-factor systems have inherently adaptation to DDM environment so they are used vastly in the field[5].

7. DDM Systems Division

At present time most databases have been distributed in the networks. The DDM systems may be divided in three groups as follows:

7.1 DDM Systems Based on Parallel Data Mining Agents

In principle, they were designed to harmonize parallel data mining by multi-factor technology in order to increase work efficiency.

7.2 DDM Systems Based on Meta-Learning

In principle, they were designed to improve choice quality and the composition of data mining algorithms and select the appropriate data mining model by virtue of the data correlation from the sources belonging to the website.

7.3 DDM Systems Based on Grid

Nowadays the new DDM projects are to extract the data in a geographical distributed environment. They are based on Grid network standards and platform in order to hide the complexity of heterogonous data and lower details.

8. Examining the Methods Protecting Privacy in DDM

8.1 Perturbation in the Data

Perturbation of disturbance creation is one of the methods to protect data privacy. On the basis of this method the data of each record have not exactly their real amounts. So if a record is identified in some way, the data related to it are not the essential ones; then by some mechanisms it is possible to identify the essential distribution and used to mine data [16]. This method will have different implementation models.

8.2 Data Displacement

It is possible only by displacing the amounts of a quality among all records, but when most qualities or all of them are displaced in the records groupings creation is inefficient and by virtue of [11] efficient data mining execution is prevented.

8.3 Making Data Random

‘Making data random’ (Or ‘Data Manipulation’) is to use a function adding some amounts to the essential quality; such functions create some perturbation in the relation channel to change essential data in line with preserving personal privacy. Such functions are used to modify data by using probability distribution functions. If total records are X = X1, X2, XN, each X record creates a noise element with FY(y) distribution function as Y1-YN and we add to it to make it as X1+Y1,…,XN+YN.

Depending on the different functions creating random there are different mechanisms to reform essential data distribution and most of them use Byes Theorem such as EM.

EM method may group properly the data and also guess the essential data distribution[9]. This method is a favorite way to approximate imperfect data or when the data have changed to the same form and generally to any type of approximation. The algorithm is consist of two parts. Approximating the lost or hidden data and in second phase by virtue of the our suppositions we select the mode with the most similarity to reality. This step may be repeated and guaranteed that the similarity increases in each repetition.

The Advantages of making the data random have following specifications:

- In the method it is possible to create the random operation for each record whenever you like and contrary to other methods it is not necessary to collect the data before the work commencement.
- it is not necessary to collect the data in a separate server to make them random so there is no potential attack for the server.
- When the data are sent to the database only by the user and there is no other exchange and operation between the sender and receiver there is no reason to use complicated computations and interactive protocols of coding method. Then the simple

method to make the data random is the best mechanism for the unilateral relation cases[18].

• Other applicable cases of the method are in OLAP and also in grouping data or discrete data mining.

8.4 Multiple Perturbation

If the number of the qualities creating perturbation increases, it is more probable to have guarantee preserve privacy, but the efficiency of the method making random decreases highly[9]. So it seems necessary to analyze the data set into details by ray operation and then it is possible to merger them by the applied software: data mining. The method is applicable in data mining; above mechanisms are shown in Table 1.

Method title	**Use charge**	**Specifications**	**Applications**
Displacement	Low	Simplicity – No change in essential amounts	DDM – OLAP – Unilateral interaction – Limited sources to preserve privacy
Making random	Appropriate	Using during data collection simplicity – little computation – efficiency	
Multiple perturbation	More than making random mode (For ray operation & reform)	More privacy preservation, if the qualities are not too much- more overflow	

Table 1. A summary of perturbation mechanisms in data

9. K-anonymity

Each composition of amounts of each data version should be at least indistinguishably in accord with K amount namely at least each record is relation to (k-1) amount of data in other records[7&15]. In this method the data are converted by two generalization and suppression methods:

• Generalization: In this mode the data exposition granulation decreases; for instance, the exact date of birth decreases to the year of the birth.

• Suppression: All amounts of a quality are removed. Although this mode execute well all confidential duties, but it becomes less probable to have proper results from the method.

Figure 3 shows the sample of the amounts only on a quality before and after using the method.

Optimization response for this method is of NP-hard algorithms, but if heuristic methods are used, related complexity will decrease[15].

If there is ‘*K*’ record and all of them have an amount for a sensible quality, it is possible to find exactly the quality and also the science background may make the method vulnerable because total amounts may decrease and it may be abused

10. Code Technique

Generally this method are used in data mining duties (Extracting patterns, correlation rules, …) while there are special data; such duties are necessary when there is no confidence between the partners or when they compete with each other to be able to hide the data and sometimes local findings[17]. Most of distribution mechanisms use this method to extract their models because this method is a known model for security discussions and its improvement methods and assessment exist, too. On the other hand, they are known algorithms and tools to be used to mine data; of course, some weak points were seen in computations steps or their output[14].

11. Distribution or Secure Multiparty Computation

In many modes the database has been saved between two or more sites. Essentially in a concentrated data mining model it is supposed that we have all data necessary for data mining algorithm or at least, it is possible to send it to a central site to execute the data mining algorithm on the data set. Regardless data privacy protection problem the method is very insufficient because it is necessary to transfer a high amount of data namely total database to a central site[8].

Distributed privacy protection problem often uses coding protocols for secured computations of several bases. A protocol designs how data exchange without violating privacy field.

If the data are shared (Or in other words, distributed) between several bases, it is necessary to present some mechanisms to protect private data of each base in a way that each base is informed only of its data and final findings from data mining. In addition to the data private for each special base even sometimes this level of data for a base leads to reveal private data in another base which is considered as an insufficiency for secured computations between two bases. Each base is aware of its data and total result so it may be aware of such amount in another base, too. On the other hand, when the entered amounts are not too much this mechanism's efficiency decreases. All the interactions are done through sent and received messages so [16] it proposes to use the protocols which are able to codify and send each message by different keys; then even two equal messages seem differently. Like any another method at first of the way this method was presented between only two bases for secured computations in 1986 [19]. ID3 decision tree proposed to group the data between two bases became the base of many methods which has not the limits of some others [12&13].

Generally discrete data mining algorithms treat their executive process based on data distribution way because depending on their vertical or horizontal separation they need separated methods to extract rules.

11.1 Vertical Distribution

It is when actual data have different specifications in each database. Usually in such cases the data have been distributed between different bases and are not inherently distributed; that is why it is known as vertical distribution.

11.2 Horizontal Distribution

It is when the data have been distributed equally between different bases and the data of each base have all specifications; the data may be inherently distributed or due to security reasons the concentrated data have been distributed between different bases. This mode is known as horizontal distribution.

11.3 Authors and Affiliations

Having distributed data between several bases general grouping methods, mining correlation rules, codification, etc. were created, but unfortunately these general mechanisms have not enough efficiency specially when their entrance volume is high[18].

12. Conclusion

Privacy protection is necessary in many steps including data to disclosing the findings from data mining. Perturbation methods help the data to be changed and saved. This simple method may extract properly rules and patterns, but its security is not guaranteed completely specially when no main data are known to the attackers. Although anonymity method is effective to generalize and then change the main data, but it has some inefficiencies such as case applications and weak points against attacks. Due to its history and reliable methods in data mining the codification technique is used vastly, but complicated computations and expensive communications decrease its efficiency. Three efficiency parameters namely efficiency,

Method	Efficiency (5-9)	Security (5-10)	Authenticity (6-10)
Secured multi-detail computations	0 – 5	8 – 10	8.5 – 9.5
Randomly	6 - 9	7 - 5	7.5 – 9.5

Table 2. Examining Some Parameters Effective on Selecting the Methods to Protect Privacy[16]

Method	Use cost	Specifications	Weak points
Making Anonym	Finding complete solution of NP-hard group, but using heuristic methods decreases costs	Generalization & privacy protection- developing new methods to remove deficiencies- efficiency for text and verbal data	If there are too much vulnerable views, the data quality decreases
Secured multi-detail computations	Computations & high communications-overflow decrease, if composed with other methods	Security- independence of bases for local operations such as making random	Complicated computations = Lack of protection of middle findings in some cases

Table 3. Comparing two methods making anonym and Secured multi-detail computations

and security should be chosen in harmony with each other in order to select the method to protect the privacy; in Tables 2 and 3 you see a comparison between the methods[16]. Until there is no comprehensive standardization the most important subject is to select a method appropriate to real environment of the problem after taking into consideration presuppositions of each mechanism.

References

[1] Tsoumakas, G., Vlahavas, I. (2009). Distributed Data Mining, IGI Global, distributing in print or electronic forms without written permission of IGI Global, Section D.

[2] Liu, B., Cao, S., Jia, X., Hua, Z. (2010). Data Mining In Distributed Data Environment, IEEE, Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, p.421-426, 11-14 July.

[3] Ashrafi, Zaman, M., Taniar, D., Smith, K. (2002). A Data Mining Architecture for Distributed Environments, Springer-Verlag Berlin Heidelberg, p.27-38.

[4] Hoonpark, B., Kargupta, H. (2005). Distributed Data Mining: Algorithms, Systems and Applications, Department of Computer Science and Electrical Engineering University of Maryland Baltimore Country 100 Hilltop Circle Baltimore.

[5] Da Silva, C., Giannella1, C., Bhargava, R., Kargupta1, H., Klusch, M. (2006). Distributed Data Mining and Agents, Unpublished.

[6] Jiang, W.N., Yu, J. (2005). Distributed Data Mining on The Grid, Machine Learning and Cybernetics. Proceedings of 2005 International Conference on, p. 2014, 4, 18-21 August, 2005.

[7] Shen, Y., Shao, H., Li, Y. (2009). Research on the Personalized Privacy Preserving Distributed Data Mining, IEEE, Future Information Technology and Management Engineering, FITME '09. Second International Conference, p.436 –439, 13-14 Dec. 2009.

[8] Li, F., Ma, J., Li, J. (2008). An Adaptive Privacy Preserving Data Mining Model Under Distributed Environment, Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, p.60-68.

[9] Charu, A., Yu, C. (2008). Privacy-Preserving Data Mining Models and Algorithms ,1st edition, Springer, p.3-51.

[10] Borman, S. (2004). The Expectation Maximization Algorithm A short tutorial, Unpublished.

[11] Duan, Y., Canny, J., Zhan, J. (2010). P4P: Practical Large-Scale Privacy-Preserving Distributed Computation Robust Against Malicious Users, 9th USENIX Security Symposium Washington, D.C.

[12] Estivill-Castro, V., Brankovic, L. (1999). Data Swapping: BalancingPrivacy Against Precision in Mining for Logic Rules, *In*: Proc. 1st Conf. Data Warehousing, Knowledge Discovery (DaWaK-99), LNCS 1676, *Springer-Verlag*, p.389–398.

[13] Kantarcioglu, M., Clifton, C. (2004). Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data, *IEEE Transactions on Knowledge and Data Engineering*, 16 (9) 1026-1037.

[14] Nayak, G., Devi, S. (2011). A Survey on Privacy Preserving Data Mining: Approaches and Techniques , *International Journal of Engineering Science and Technology* (IJEST).

[15] Sweeney, L. (2001). Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, International Journal of Uncertainly, Fuzzinessand Knowledge-Based Systems, p.571-587.

[16] Vaidya, J., Clifton, C. (2004). Privacy-Preserving Data Mining: Why, How, and When, Published by The IEEE *Computer Society*, 1540-7993904/ IEEE, 2004.

[17] Wang, P., Survey on Privacy Preserving Data Mining, *International Journal of Digital Content Technology and Its Applications*.

[18] Wu, C. (2005). Privacy Preserving Data Mining with Unidirectional Interaction, *International Symposium on Circuits and Systems (ISCAS 2005)*, Kobe, Japan, 2005.

[19] Yao, A. (1986). How to Generate and Exchange Secrets, *In*: Proc. 27th IEEE Symp. Foundations of Computer Science, *IEEE CS Press*, p.162–167.