Analytical Study of Structure of Models and Techniques of Privacy Preserving Data Mining

Dr. Devendra Kumar
ABES Engg. College, Ghaziabad
devendra.arya@abes.ac.in

ABSTRACT: ***Privacy preserving becomes an important issue in the development progress of data mining techniques. Privacy preserving data mining has become increasingly popular because it allows sharing of privacy-sensitive data for analysis purposes. There has been an important research area that how to protect private information or sensitive knowledge from leaking in the mining process. The goal in investigating privacy preservation issues is to take a systemic view of architectural requirements and design principles. This paper describes the meaning of data mining, Privacy, Design architecture for privacy preserving, types of privacy preserving, privacy preserving data mining, data distortion, data encryption, and reconstruction techniques in detail.***

KEYWORDS: ***Data Dining, Privacy Preserving, Data Distortion, Data Encryption, Data Reconstruction.***

I. INTRODUCTION

With database technology and network technology development, people generated and collected data has increased dramatically. Data mining as a powerful data analysis tool, can find the potential models and rules in data, and is applied more and more in-depth in business decisions, scientific and medical research areas. At the same time, data mining is directly on the original data set which also produces inevitable leakage of privacy. The main research direction of privacy preserving data mining is how to protect private information or sensitive knowledge from leaking in the mining process, to obtain accurate results of data mining.

Privacy preserving data mining can be divided into two levels [17]. The first level of privacy preserving data mining is the protection of sensitive data, such as name, id number, address and other sensitive data. The second level is knowledge hiding in the database is the protection of sensitive knowledge that is shown by data mining. It is a problem to be solved is how to effectively hide sensitive rules of the data set, with minimal impact on non-sensitive rules and the usefulness of data sets. Privacy protection technology is used to hide sensitive data or sensitive knowledge which mainly focuses on data distortion, data encryption and data reconstruction technology to the study at present.

A. DATA MINING

The efficient database management systems have been important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to the recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transaction and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the “essence” of information stored, and the discovery of patterns in raw data.

In this information age, information leads to power and success, and thanks to sophisticated technologies such as computers, satellites etc. we have been collecting tremendous amounts of information. Initially, with the advent of computers and means of a mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored in disparate structures very rapidly became over shelling.

B. PRIVACY

According to Berry and Linoff (2000:468) [17] privacy is a complex issue but because of technology, it is increasingly becoming a social issue. The Cambridge Advance Learner’s Dictionary (2004) defines the word privacy as someone’s right to keep their personal matters and relationships secret. Today, every form of commerce leaves an electronic trail, and acts that were once considered private or at least quickly forgotten, are stored for future reference.

It is an important issue to consider both as individuals and in the work we do that may intrude on the privacy of others.

- Limits are already placed on privacy by the social contacts, and the issue is really how much information should be collected and who is in control of the information.
- Every person has a different perspective on privacy.
- Different levels of tolerance with regard to information about them being available to others. Technology plays a role in defining privacy, protecting it, and intruding on privacy.

C. PRIVACY PRESERVING DATA MINING

Privacy Preserving Data Mining (PPDM) is a novel research direction in data mining and statistical databases where data mining algorithms are analyzed for the side-effects they incur in data privacy. The main objective of PPDM is to develop algorithms for modifying the original data in some way so that the private data and private knowledge remain private even after the mining process, [10]. In essence this implies that the data to be mined would be stripped of all information that could be used to identify a specific individual, and that the same would be done with the resulting knowledge gained from the data mining effort.

PPDM is still in its infancy, and whether it will be able to address all the privacy concerns in data mining is debatable. Another question that one needs to ask is how valuable data mining results would be too marketable, if individual customers that the marketing efforts should be directed at, cannot be identified. In the meantime, organizations should consider the following in order to protect themselves from legal liabilities or bad press resulting from irresponsible data mining efforts:

- Provide customers with an opt-out option whereby they have the ability to exclude themselves from being used in data mining or from being the target of direct marketing.
- Ensure that you only buy data from reputable organizations, and that the necessary permission has been obtained for making use of that data.
- Inform your customers of potential use of their information for data mining purposes, and obtain their consent prior to releasing this information to other organizations.

II. DATA MINING MODELS

The goal of data mining is to extract knowledge from data. David Hand, Heikki Mannila, and Padhraic Smyth2 categorize data mining into five tasks:

- **Exploratory Data Analysis (EDA):** Typically interactive and visual, EDA techniques simply explore the data without any preconceived idea of what to look for.

- **Descriptive Modeling:** A descriptive model should completely describe the data (or the process generating it); examples include models for the data’s overall probability distribution (density estimation), partitions of the *d*imensional space into groups (cluster analysis and segmentation), and descriptions of the relationship between variables (dependency modeling).

- **Predictive Modeling: classification and regression:** The goal here is to build a model that can predict the value of a single variable based on the values of the other variables. In classification, the variable being predicted is categorical, whereas in regression, it’s quantitative.

- **Discovering Patterns and Rules:** Instead of building models, we can also look for patterns or rules. Association rules aim to find frequent associations among items or features, whereas outlier analysis or detection focuses on finding “outlying” records that differ significantly from the majority.

- **Retrieval of Content:** Given a pattern, we try to find similar patterns from the data set.

III. DESIGN ARCHITECTURE FOR PRIVACY PRESERVING

As Figure 1 shows, privacy-preserving data mining has multiple steps that translate into a three-tiered architecture: At the bottom tier are the data providers, the data owners, which are often physically distributed.

The data providers submit their private data to the data warehouse server. This server which constitutes the middle tier supports online analytical data processing to facilitate data mining by translating raw data from the data providers into aggregate data that the data mining servers can more quickly process.

The data warehouse server stores the data collected in disciplined physical structures such as a multidimensional data cube, and aggregates and recomputed the data in various forms, such as sum, average, max, and min.

At the top tier are the data mining servers which perform the actual data mining. In a privacy-preserving data mining system, these servers do not have free access to all data in the data warehouse.
In a hospital system, the accounting department can mine patients' financial data, for example, but cannot access patients' medical records. Developing and validating effective rules for the data mining servers' access to the data warehouse is an open research problem [4]

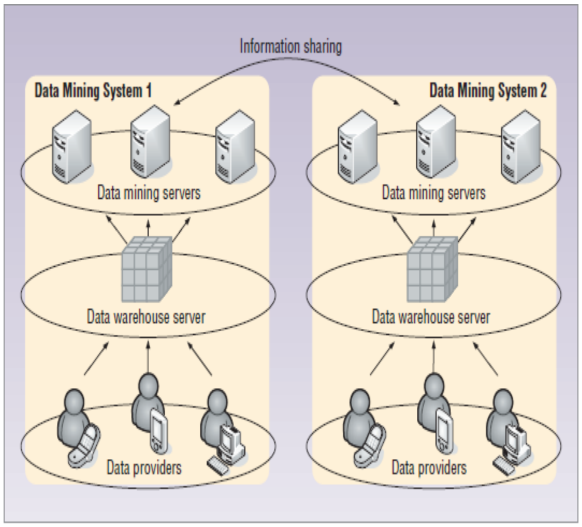

Figure1. Basic architecture for privacy-preserving data mining

The architecture typically has three tiers: data providers which are the data owners; the data warehouse server, which supports online analytical processing; and the data mining servers that perform data mining tasks and share information. The challenge is to control private information transmitted among entities without impeding data mining.

As Figure 1 shows, sharing occurs in the top tier, where each data mining server holds the data mining model of its own system. Thus, "sharing" means sharing local data mining models rather than raw data.

IV. TYPES OF PRIVACY PRESERVING

A. THE DISTORTION-BASED PRIVACY PRESERVING

Data distortion perturbs the original data to achieve privacy preserving. The perturbed data would meet the two conditions. First, an attacker cannot discover the real original data. In other words, the attacker cannot reconstruct the real original data from the issuance of the distortion data.

Second, the distorted data is still to maintain some properties of the original data, namely some of the information derived from the distorted data are equivalent to data obtained from the original information. So it ensures that some applications based on the distorted data are feasible. At present, the techniques of privacy preserving based on data distortion include randomization [9], data blocking [12] and so on. Data randomization is the technique that adds random noise to the original data, and then distributes the disturbed data.

Randomization techniques include two types. The first type of randomization is called random perturbation. The other type of randomization is called randomized response.

- **Random Distruption Based on Association Rule Hiding**

Random perturbation modifies sensitive data in a random process, thus achieving data privacy preserving. The basic idea of the data transformation method is: to find the sensitive transactions to support the sensitive rules in the original database, to delete them or add items, so support or the confidence of the sensible rule reduced to below the threshold specified in order to achieve the sensitive rule hidden.

The problem of mining association rules was introduced in [13]. Let I = {i1, i2, in} be a set of literals, called items. Let D be a set of transactions, which is the database that is going to be disclosed. Each transaction t ∈ D is an item set such that t ⊆ I, an association rule is an expression X ⇒ Y where X⊆ I, Y⊆ I, and X ∩ Y =Φ. The X and Y are called respectively the left hand side and right hand side of the rule. The confidence of the rule is calculated as |X ∪ Y |/|X|, where |X| is the number of transactions containing X and | X ∪ Y | is the number of transactions containing both X and Y. The support of the rule is calculated as | X ∪ Y |/|N|, where |N| is the number of transactions in D. The problem of mining association rules is to find all rules that are greater than the user-specified minimum support and minimum confidence.

- **Data blocking**

Data blocking is different from data transformation modifying the data and provide non-real data, it don't distribute any specific data to achieving data privacy preserving because some applications hope more to conduct studies based on real data. Blocking specific responses to the data table that the certain value of the data table is replaced with an uncertain symbol. For example, an uncertain symbol "?" with the exception of {0,1} is introduced to realize the Boolean association rule hiding. Because some values are instead of"?", then the count of some item sets is an uncertain value that is located between a minimum estimate and the maximum estimated value range. So achieving sensitive association rule hiding is to design an algorithm, in case of the data values as little as possible blocking, the Third International Symposium on Intelligent Information Technology and Security Informatics support and confidence of a sensitive association rule is controlled at below a predetermined threshold.

- **Random disruption based on classification rule**

Classification is a process, which can identify the model or function to describe and distinguish data classes or concepts, in order to be able to use predictive models to mark the unknown object classes. Classification goal is to construct a classification model to predict future data trends.
At present classification methods are mainly used in classification rules, decision trees, neural networks and so on [18].

Parsimonious downgrading is a combination of rules and decision tree classification of the privacy preserving method [15].

Parsimonious downgrading is such an algorithm that the information in the data to be downgraded will be removed as a formal description. The so-called downgrading is to make the sensitivity level or the privacy level reduced to level can be announced. In other words, downgrade is right to make public release of information privacy preserving handling process. The parsimonious downgrading uses θ instead of the data to be blocked, while the other blocking methods are often used "?" instead of the data to be blocked. θ values are between 0 and 1, which represents the blocked property to take the probability of a certain value.

- **Randomized Response**

The basic idea of randomized response is: the data owner distributes the original perturbed data so that an attacker cannot be higher than the predetermined threshold probability to obtain the original data whether to include some real or false information. Randomized response technique and random perturbation techniques difference is that sensitive data is through a kind of indirect way of answering specific questions provided to the outside world. The randomized response model has two types: related-question model and unrelated-question model. Related-question model designs the two opposing issues of sensitive data **[6].** For example:

- I contain sensitive values A;
- I don't contain sensitive value A.

Data owners according to own data select at random a question to be answered. But the questioner isn't aware of the specific problems to be answered by data owners. After large amounts of data owners answered questions, the proportion of respondents containing sensitive values and the proportion of respondents not containing sensitive values can be obtained by calculating. It is supposed that The probability of respondents randomly selecting the first question is θ, and then the probability of the second question to be answered is 1-θ. In various proportions stated as follows:

• P (A=yes), The proportion of the data owners containing sensitive values A;
• P (A=no), The proportion of the data owners not containing sensitive values A;

- P*(A=yes), The proportion of respondents answered yes;
- P*(A=no), The proportion of respondents answered no.

The following equation set:

- P* (A=yes)=P (A=yes)×θ+P (A=no)×(1-θ);
- P*(A=no)=P (A=no)×θ+P (A=yes)×(1-θ).

Based on the above two equations, and a combined estimate of all respondents drawn from p * (a = yes) and p * (a = no), p (a = yes), p (a = no) can be obtained.

Throughout this process, the inability to identify issues related to the respondents to answer and therefore cannot determine whether they contain sensitive data values.

A randomized response technique used to provide information with response model for processing categorical data.

B. THE ENCRYPTION - BASED PRIVACY PRESERVING

In a distributed environment the primary issue to achieve privacy preserving is the security of communications and encryption technology just to meet this demand. Therefore, privacy preserving based on data encryption technology commonly applies to distributed applications. Distributed applications store data using two models: vertically partitioned data model and horizontally partitioned data model. Vertically partitioned data refers to data by property located in different sites, all sites stored data does not overlap. Horizontally partitioned data refers to data distributed in each site according to records in this condition, the various sites without having to know the specific record information to other sites; we can calculate the overall association rules [5].

There are a lot of data mining algorithms on cryptography technology to solve real privacy issues, for example: secure multi-party computation, SMC. SMC is defined as in a distrust of the multi-user networks; each user can be coordinated through the network to complete the reliable computing tasks, while maintaining the security of their data [20] [14]. B. Pinkas put forward a theoretical study of cryptography used in data mining privacy preserving, and demonstrated different kinds of data mining problems can be transformed into SMC [20].

For the vertically partitioned data mining association rules, the difficulty is how to calculate the support of the item set, while using the secure scalar product or secure size of set intersection of the problem can be resolved [2]. Literature [2] described an algorithm applied to Expectation Maximization clustering without disclosure of information on each site because the algorithm used secure sum computing. In the multi-classification mining areas, Wenliang Du proposed security classification algorithm on Vertically partitioned data using the secure scalar product [3]; while Liddell made the use of encryption methods to establish the horizontally partitioned data decision tree and translate the search for the best classification property into secure multi-party computation [7].

C. THE RECONSTRUCTION - BASED PRIVACY PRESERVING

Data reconstruction includes numerical data reconstruction and binary data and classification data reconstruction.

- **Numerical data of the reconstruction techniques*:***

The original data are modified by discrimination methods and the value of the deformation method, then use reconstruction algorithm to construct the distribution of original data [8].

Literature [16] proposed an improved method based on Bayesian reconstruction with the Expectation Maximization algorithm. In the case of a large enough data set, the Expectation Maximization algorithm can get the original data distribution of the maximum likelihood estimate. In addition, numerical data reconstruction can achieve the original data protection, but sensitive knowledge is not protected. For the binary data and classification data reconstruction: literature [11] used a randomization technology to modify the binary data and classification data of the association rules as dimensions.

V. ASSESSMENT OF THE PRIVACY PRESERVING DATA MINING ALGORITHMS

So far, there is no privacy preserving data mining algorithm to effectively hide the various data sets. The current algorithms are mostly designed for specific data sets; therefore, there is no specific criterion to obtain an accurate assessment of the performance of each algorithm but generally speaking, privacy preserving algorithms can be evaluated and compared from the following areas [1]:

• **Efficiency of the algorithm:** The main is the algorithm running time to hide sensitive data or sensitive information. It is necessary to evaluate an important indicator of various algorithms.

• Availability of the algorithm: Availability of the algorithm refers to data sets processed by privacy preserving technology; it contains information that should be as far as possible meet the needs of data mining. If the global association rules derived from data sets processed by privacy preserving algorithms is wrong or does not reflect the true situation, so that the algorithms lost availability.

• **Level of privacy preserving:** Level of privacy preserving refers to the extent to which the success of sensitive information hidden and in the data set to be distributed for the use of varieties of data mining algorithms excavated the success rate of private information.

• Scalability of the algorithm: Scalability of the algorithm refers to the ability to handle massive data sets or the variation trend in processing efficiency when the amount of data increases. The efficiency of change is relatively slow for a good scalability of the algorithm when the amount of data increases.

VI. ALGORITHM

Input: Table T
Output: Table T* (Having all the required attributes of T and an individual tuple from T is not identifiable in T*)

- Select only the required attributes from Table T.
- Categorize the type of attributes (Identifier (Ai), Sensitive (As) or Quasi-identifier (Aq))
- Identifier attributes (e.g. Name) should not be disclosed so these attributes values, if present, with auto generator id numbers.
- Find the member of quasi identifier attributes whose data type is numeric (e. g. Age). By deciding the size of fuzzy set (k), min, max and midpoints of each fuzzy set ml to make, transform the actual value (x) using the functions fl (x), fi (x) and fk (x) into new value (xn) and replace the actual value with xn plus category number.
- For categorical sensitive attribute (e.g. Disease) transformation is performed using the mapping table prepared with domain knowledge considering privacy disclosure level set by the user.
- Till the end of the record, the transformation is done and Table T* is generated.

VII. CONCLUSION

Privacy preserving is applied widely in many fields and is the research subject of the emerging academic in recent years. This paper describes the Data Mining, privacy, privacy preserving data mining, distortion-based privacy preserving, the encryption - based privacy preserving and the reconstruction - based privacy preserving. At present a variety of privacy preserving data mining algorithms are still some shortcomings, and are targeted at specific applications and data sets, rather than to be extended to the general. The premise of ensuring the privacy of how to reduce the loss of accuracy, how to further improve the algorithm efficiency and privacy preserving generality in different types, and distribution characteristics of different data sets are the direction of the future worthy of further study. This paper reiterates several privacy preserving data mining technologies clearly and then proceeds to analyze the merits and shortcomings of these technologies.

VIII. REFERENCES

[1] Agrawal D, Aggarwal C C. On the design and quantification of privacy preserving

[2] Agrawal R, Srikant R. Privacy-preserving data mining [A]. Proceedings of the 2000 ACM SIGMOD international conference on Management of data [C]. Dallas, Texas, United States: ACM, 2000. 439-450

[3] Clifton C, Kantarcioglou M, Zhu Y M. Tools for privacy preserving distributed data mining [J]. SIGKDD Explorations 4, 2002

[4] Data mining algorithms [A]. Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of Study of Privacy Preserving Data Mining Haisheng Li East China Jiaotong University, Nachang330013, China

[5] Du Wenliang, Attalah M J. Secure multi problem computation problems and their applications: A review and open problems

[R]. CERIAS Tech Report 2001-51, Center for Education and Research in Information Assurance and Security and Department of Computer Sciences, Purdue University, West Lafayette, IN, 2001.

[6] DU Wen-liang, ZHAN Zhi-jun. Building decision tree classifier for private data [C] Proc of IEEE International Conference on Privacy, Security and Data Mining. Darlinghurst: Australian Computer Society, 2002: 1-8.

[7] Du Wenliang, Zhan Zhijun. Building decision tree classifier for private data [C]. In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining, 2002

[8] Kantarcioglu M, Clifton C· Privacy preserving Distributed Mining of Association Rules on Horizontally Partitioned Data In: ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 2002

[9] Kargupta H, Datta S, Wang Q, Sivakumar K. On the privacy preserving properties of random data perturbation techniques, Proceedings of the IEEE International Conference on Data Mining

[10] L. Wang, S. Jajodia, and D. Wijesekera, "Securing OLAP Data Cubes against Privacy Breaches," *Proc. 25th IEEE Symp. Security and Privacy*, IEEE Press, 2004, pp. 161-175.

[11] Lindell Y , Pinkas B. Privacy preserving data mining [C]. In Advances in Cryptology-CRYPTO 2000, 2000 : 36-54.

[12] Melbourne, Florida, 2003: 99-106 (ICDM).

[13] Moskowitz I S, Chang L W. A decision theoretical based system for information downgrading/Proceedings of the 5th Joint Conference on Information Sciences (JCIS). Atlantic City, NJ USA, 2000: 82-89

[14] Pinkas B. Cryptographic techniques for privacy-preserving data mining [J]. ACM SIGKDD Explorations Newsletter, 2002,4 (2): 12-19.

[15] Pinkas B. Cryptographic techniques for privacy-preserving datamining· SIGKDD Explorations, 2002, 4 (2)

[16] Rizvi S J, Haritsa J R. Maintaining data privacy in association rule mining [A]. In Proceedings of the 28th International Conference on Very Large Databases (VLD) [C]. Hong Kong,Chi-na: [s.n.], 2002. 682-693.

[17] V. Verykios, E. Bertino, I.G. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis, State-of-the-art in Privacy Preserving Data Mining? SIGMOD Record, Vol. 33, No. 1, 50-57, March 2004.

[18] VERYKIOS V S, ELMAGARMID A, BERTINO E, et al. Association rule hiding [J]. IEEE Trans on Knowledge and Data Engineering, 2004,16 (4): 434-447

[19] Verykios, VS; Bertino, E; Fovino, IN; Provenza, LP; Saygin, and Theodoridis, Y. 2004. State-of –the-art in Privacy Preserving Data Mining. SIGMOD Record. Volume 33, Issue 1:50-57.

[20] ZHOU Shui-Geng, LI Feng1, TAO Yu-Fei, XIAO Xiao-Kui. Privacy Preservation in Database Applications: A Surve. CHINESE JOURNAL OF COMPUTER, 2009,32 (5)