A review on security in Data Mining

Amarpreet Singh[#1], Vinay Bhardwaj[*2]

[#]*Computer Science Department, Sri Guru Granth Sahib World University*
Fatehgarh Sahib, Punjab, India

***Abstract*—Privacy is one of the most important properties of an information system must satisfy, in which systems the need to share information among different, not trusted entities, the protection of sensible information has a relevant role. Thus privacy is becoming an increasingly important issue in many data mining applications. For that privacy secure distributed computation, which was done as part of a larger body of research in the theory of cryptography, has achieved remarkable results. These results were shown using generic constructions that can be applied to any function that has an efficient representation as a circuit. A relatively new trend shows that classical access control techniques are not sufficient to guarantee privacy when data mining techniques are used in a malicious way. Privacy preserving data mining algorithms have been recently introduced with the aim of preventing the discovery of sensible information.**

***Keywords*—** data mining

I. Introduction

Privacy preserving data mining is an important property that any mining system must satisfy. So far, if we assumed that the information in each database found in mining can be freely shared. Consider a scenario in which two or more parties owning confidential databases wish to run a data mining algorithm on the union of their databases without revealing any unnecessary information. For example, consider separate medical institutions that wish to conduct a joint research while preserving the privacy of their patients. In this scenario it is required to protect privileged information, but it is also required to enable its use for research or for other purposes. In particular, although the parties realize that combining their data has some mutual benefit, none of them is willing to reveal its database to any other party. The common definition of privacy in the cryptographic community limits the information that is leaked by the distributed computation to be the information that can be learned from the designated output of the computation. Although there are several variants of the definition of privacy, for the purpose of this discussion we use the definition that compares the result of the actual computation to that of an "ideal" computation: Consider first a party that is involved in the actual computation of a function (e.g. a data mining algorithm). Consider also an "ideal scenario", where in addition to the original parties there is also a "trusted party" who does not deviate from the behavior that we prescribe for him, and does not attempt to cheat. In the ideal scenario all parties send their inputs to the trusted party, who then computes the function and sends the appropriate results to the other parties. Loosely speaking, a protocol is secure if anything that an adversary can learn in the actual world it can also learn in the ideal world, namely from its own input and from the output it receives from the trusted party. In essence, this means that the protocol that is run in order to compute the function does not leak any "unnecessary" information.

II. CLASSIFICATION OF PRIVACY DATA MINING

Data Hiding	Data Perturbation	Value Distortion	Additive Perturbation
			Multiplicative Perturbation
			Data Microaggregation
			Data Anonymization
			Data Swapping
			Other Randomization Techniques
		Probability Distribution	Sampling Method
			Analytical Method
	Secure Multi-Party Computation (SMC) / Cryptographic Protocols		
	Distributed Data Mining (DDM)		
Rule Hiding	Association Rule Hiding		Data Perturbation
			Data Blocking
	Classification Rule Hiding		Parsimonious Downgrading

Table 1. A brief overview of privacy preserving data mining techniques

2.1 Privacy Preservation Data Mining

Privacy has been gaining more attention to handle the terrorism, the government needed to examine, using data mining technology, more information about individuals to detect unusual disease outbreaks, financial fraudulent behaviors, network intrusions, etc. While all of these applications of data mining can benefit our society, there is also a negative side to this technology because it could

Manuscript received June 5, 2015
Manuscript revised June 20, 2015

be a threat to the individuals' privacy. Overcome the "limitations" of data mining techniques including areas like data security and privacy preserving data mining, which are actually active and growing research areas

2.2 Data Distribution

The PPDM algorithm can be divided into two major categories, Centralized and distributed data. In a centralized database, data are stored in a single database, in distributed data can further classified into horizontal a vertical data distributions. In Horizontal data distribution from different records of the same data attributes are resided in different places. In vertical data distributor different attributes of the same record of data are resided in different places most research occurred on a centralized data base. Applying PPDM algorithm to a distributed database privacy concerns, communication cost is too expensive

2.3 Purpose of PPDM

The PPDM algorithms main purpose is hiding is data hiding and rule hiding. In Data Hiding the sensitive data from original database like indentify name and address are linked, directly or indirectly to an individual person are hided. In rule hiding the sensitive data (rule) from original database after applying data mining algorithm is removed. Most of the PPDM algorithms hide sensitive patterns by modifying data hiding.

2.4 PPDM Algorithm

The PPDM algorithm are specifically on the tasks of classification, association rule and clustering classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of the model is used for prediction the class of objects whose label is unknown clustering analysis concerns the problem of separating a data set in one group which are similar top each other and are different as possible in other group.

2.5 PPDM Techniques

PPDM techniques used by four categories Sanitation, it can remove or modify items for a database to reduce the support of some frequently used items sets that sensitive patterns are not to be mined. Blocking it can replace certain attributes the data with a question mark. According to this the minimum support and confidence level will be altered into a minimum interval. In distort, the support and the confidence of a sensitive rule lie below the middle the two and the confidentiality of data is expected to be protected and also known as data perturb action or data randomization, where individual data records are modified from original data, and reconstructed from randomized data. This technique aims to design for distortion methods after which the true value of any individual record is difficult to ascertain, but unchanged for danger data. In generalization transforms and replaces each record value with a correspondery generalized value.

III. Related Study

Abusukhon et al.[1] describes security is an important issue when secure information is sent over a network .Data encryption is done and method used is text to image encryption .To investigate dividing the text into blocks and then transfer each block into an image and create an individual key for each block.

Chauhan et al.[3] provides an overview of the different techniques that are used for privacy preserving in data mining. There are many data mining applications which deal with privacy sensitive data. Data mining in such privacy sensitive domains is facing growing concerns. Therefore it is needed to develop data mining techniques that are sensitive to the privacy issue and also address the issue of the privacy preserving data mining. This paper provides an overview of the RSA encryption for the privacy preserving data mining. The aim of the RSA encryption is to encrypt the data so that the customer may not lose his\her personal or valuable data. It also provides an overview of the different techniques and how they are related to each other.

Ge et al.[5] describes data base system have two goals to achieve one is security of encryption and second fast performance of queries .The order preserving techniques are well suited for database but they support a simple and efficient way to build indices ,FCE is used to provide security to database .

Gupta et al.[6] explains data mining is a technique to dig the data from the large databases for analysis and executive decision making. Security aspect is one of the measure requirements for data mining applications. This paper presents security requirement measures for the data mining and summarized the requirements of security for data mining in tabular format. The summarization is performed by the requirements with different aspects of security measure of data mining. The performances and outcomes are determined by the given factors under the summarization criteria. Effects are also given under the tabular form for the requirements of different parameters of security aspects.

Kumar et al.[3] describes multimedia security is an important field of research in the area of information sharing. In this paper, Fast Encryption Algorithm (FEAL), an encryption/decryption strategy for gray scale images is proposed. The FEAL is a block cipher, also called as Japanese Encryption algorithm. FEAL works almost

similar to Data Encryption Standard (DES) algorithm, but it is faster than DES. To encrypt the images, the input image is split into 16x16 blocks of information. Encryption/Decryption is carried out using 12 keys, each of length 16-bits.

Li et al.[7] explains one of the main efficient drawbacks of IBE is the overhead computation at private key generator during user revocation. It aims at tacking the critical issue of identity revocation; introduce outsourcing computation into IBE for the first time and proposed a revocable IBE scheme in the server aided setting.

Nguyen et al.[9] describes network attacks have increased in number and severity over the past few years, intrusion detection system (IDS) is increasingly becoming a critical component to secure the network. Due to large volumes of security audit data as well as complex and dynamic properties of intrusion behaviors, optimizing performance of IDS becomes an important open problem that is receiving more and more attention from the research community. The uncertainty to explore if certain algorithms perform better for certain attack classes constitutes the motivation for the reported herein. This paper evaluates performance of a comprehensive set of classifier algorithms using KDD99 dataset. Based on evaluation results, best algorithms for each attack category is chosen and two classifier algorithm selection models are proposed.

Thuraisingham et al.[10] discuss the issue of privacy preserving data mining and present the technique that provide the privacy on data mining application and provided an overview of the different techniques and how they relate to one another. It used the asymmetric encryption to provide the privacy and used the RSA encryption to encrypt the data and also presented a client server architecture that connects to the multiple clients. Server need to receive data from client .Connect server to a data base and enter the data received from the client into database with client id. The proposed protocol is to encrypt the data so use of the encryption technique to encrypt the data, also used homomorphic encryption to secure the information. Without security our data may stand compromised.

Wang et al.[11] explains terrorist attacks have awakened the awareness of nations on transportation security and safety .The results support conceptual framework that there are two sources of influences on the allocation of transportation security grants environmental changes and project characteristics .

IV. Method of randomization

The randomization method provides an effective yet simple way of preventing the user from learning sensitive data, which can be easily implemented at data collection phase for privacy preserving data mining, because the noise added to a given record is independent of the behavior of other data records. When the randomization method is carried out, the data collection process consists of two steps [3]. The first step is for the data providers to randomize their data and transmit the randomized data to the data receiver. In the second step, the data receiver estimates the original distribution of the data by employing a distribution reconstruction algorithm. The model of randomization is shown in Figure 1.

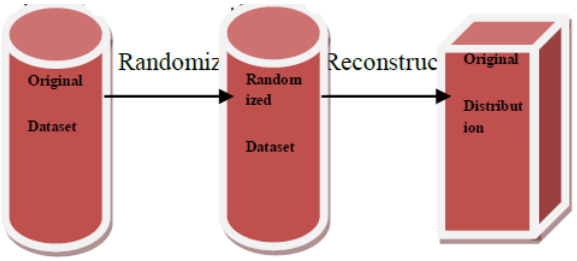

Fig 1. The Model of Randomization.

Representative randomization methods include random-noise-based perturbation and Randomized Response scheme. Agrawal and Srikant proposed a scheme for privacy preserving data mining using random perturbation and discussed how the reconstructed distributions may be used for data mining [4]. In their randomization scheme, a random number is added to the value of a sensitive attribute. For example, if $i\ x$ is the value of a sensitive attribute, $i\ i\ x \square\ r$, rather than $i\ x$, will appear in the database, where $i\ r$ is a random noise drawn from some distribution. It is shown that given the distribution of random noises, reconstructing the distribution of the original data is possible. Subsequently, Evmievski et al. proposed an approach to conduct privacy preserving association rule mining [5]. Kargupta et al. [6] proposed a random matrix-based spectral filtering technique to recover the original data from the perturbed data. Huang et al. further proposed two other data reconstruction methods: PCA-DR and MLE-DR in [7]. In addition, several distribution reconstruction algorithms have been proposed in correspondence to different randomization operators [8-10]. The basic idea of most algorithms is to use Bayesian analysis to estimate the original data distribution based on the randomization operator and the randomized data. For example, the expectation maximization (EM) algorithm [8] generates areconstructed distribution that converges to the maximum likelihood estimate of the original distribution. The Randomized Response (RR) was firstly proposed by Warner [11]. The RR scheme is a technique originally developed in the statistics community to collect sensitive information from individuals in such a way that survey interviewers and those who process the data do not know which of two alternative questions the respondent has

answered. In data mining community, Rizvi and Haritsa presented a MASK scheme to mine association rules with secrecy constraints [12]. Du and Zhan proposed an approach to conduct privacy preserving decision tree building [13]. Guo et al. addressed the issue of providing accuracy in terms of various reconstructed measures in privacy preserving market basket data analysis [14]. The randomization method is a simple technique which can be easily implemented at data collection time. It has been shown to be a useful technique for hiding individual data in privacy preserving data mining. The randomization method is more efficient. However, it results in high information loss.

V. Method of Cryptography

The growth of Internet has triggered tremendous opportunities for distributed data mining, where people jointly conducting mining tasks based on the private inputs they supplies. These mining tasks could occur between mutual un-trusted parties, or even between competitors, therefore, protecting privacy becomes a primary concern in distributed data mining setting. Distributed privacy preserving data mining algorithms require collaboration between parties to compute the results or share no-sensitive mining results, while provably leading to the disclosure of any sensitive information. In general, distributed data mining involves two forms: horizontally partitioned data and vertically partitioned data. Horizontally partitioned data means that each site has complete information on a distinct set of entities, and an integrated dataset consists of the union of these datasets. In contrast, vertically partitioned data has different types of information at each site; each has partial information on the same set of entities. Most privacy preserving distributed data mining algorithms are developed to reveal nothing other than the final result. Kantarcioglu and Clifton studied the privacy-preserving association rule mining problem over horizontally partitioned data. Their methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task. Lindell et al. researched how to privately generate ID3 decision trees on horizontally partitioned data. The problemof privately mining association rules on vertically partitioned data was addressed. Vaidya and Clifton first studied how secure association rule mining can be done for vertically partitioned data by extending the Apriori algorithm. Du and Zhan developed a solution for constructing ID3 on vertically partitioned data between two parties. Vaidya and Clifton developed a Naive Bayes classifier for privacy preservation on vertically partitioned data and proposed the first method for clustering over vertically partitioned data. All these methods are almost based on the special encryption protocol known as Secure Multiparty Computation (SMC) technology. SMC originated with Yao's Millionaires' problem. The basic problem is that two millionaires would like to know who is richer, with neither revealing their net worth. Abstractly, the problem is to simply compare two numbers, each held by one party, without either party revealing its number to the other. The SMC literature defines two basic adversarial models:

Semi-Honest Model

Semi-honest adversaries follow the protocol faithfully, but can try to infer the secret information of the other parties from the data they see during the execution of the protocol.

Malicious Model

Malicious adversaries may do anything to infer secret information. They can abort the protocol at any time, send spurious messages, spoof messages, collude with other (malicious) parties, etc.SMC technology used in distributed privacy preserving data mining areas mainly consists of a set of secure sub-protocols, such as, secure sum, secure comparison, dot product protocol, secure intersection, secure set union and so on. In the following, we will briefly describe the basic idea of two kinds of secure sub-protocols used in horizontally partitioned and vertically partitioned setting.

Secure Sum

Secure Sum can securely calculate the sum of values from different sites. Assume that each site i has some value $i\ v$ and all sites want to securely compute $S\ v\ v, , vn\ 1\ 2$, where $i\ v$ is known to be in the range [0..m]. For example, in horizontally partitioned association rule mining setting, we can securely calculate the global support count of an itemset by the secure sum sub-protocol.

Dot Product Protocol

To present, many secure dot product protocols have been proposed. The problem can be defined as follows: Alice has a n-dimensional vector $(, , ,)\ 1\ 2\ n\ X\ x\ x\ x$, while Bob has a n-dimensional vector $(, , ,)\ 1\ 2\ n\ Y\ y\ y\ y$. At the end of the protocol, Alice should get $a\ b\ r\ X\ Y\ r$ where $b\ r$ is a random number chosen from uniform distribution that is known only to Bob, and $n\ X\ Y\ x\ y\ x\ y, ,$ $x\ y\ 1\ 1\ 2\ 2$. For example, using the dot product protocol we can securely calculate the global support count of an itemset whose items are located at different sites in vertically partitioned setting. The encryption method can ensure that the transformed data is exact and secure, but it is much low efficient. Moreover, most existing work on very efficient privacy preserving data mining only provides the protocols against semi-honest adversaries. An important area for future research is to develop efficient mining protocols that remain secure and private even if some of the parties involved behave maliciously.

VI. Conclusion

This paper carries out various approaches for privacy preservation data mining and analysis techniques and method what are existing. All the proposed methods are just approximate to achieve the goal of privacy upon some extend. Firstly, new algorithm with better approximation ratio and/or
time complexity in this framework needs to be under development , still introduce considerable information loss with high-dimensional metric space involved. Drawback can be try to solve in next research paper .

VII. References

[1] Privacy, Security, and Data Mining, pp.1-8, 2002. Han Jiawei, M. Kamber, and Data Mining: Concepts and Techniques, Beijing: China Machine Press, pp.1-40, 2006.
[2] V.S.Verykios, E.Bertino, I.N.Fovino, L.P.Provenza, Y.Saygin, Y.Theodoridis, "State-of-the-art in Privacy Preserving Data Mining", New York, ACM SIGMOD Record, vol.33, no.2,Pp.50-57, 2004.
[3] N. Zhang, "Privacy-Preserving Data Mining", Texas A&M University, pp.19-25, 2006.
[4] R. Agrawal, R. Srikant, "Privacy-Preserving Data Mining", ACM SIGMOD Record, New York, vol.29, no.2, pp.439-450,2000.
[5] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, "Privacy Preserving Mining of Association
[6] Rules", Information System, vol.29, no.4, pp.343-364, 2004.
[7] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", In Proceedings of the 3rd International Conference on Data Mining, pp.99-106, 2003.
[8] Z. Huang, W. Du, B. Chen, "Deriving Private Information from Randomized Data", In Proceedings of the ACM SIGMOD Conference on Management of Data, Baltimore, Maryland,USA, pp.37-48, 2005.
[9] D. Agrawal, C.C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms", In Proceedings of the 20th ACM SIGMOD-SIGACTSIGART Symposium on Principles of Database Systems, pp.247-255, 2001.
[10] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrk, "Privacy Preserving Mining of Association Rules", In Proceedings the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, pp.217-228, 2002.
[11] S. Rizvi, J. Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th International Conference on Very Large Data Bases, pp.682-693, 2002.
[12] S. L. Warner, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", J. Am. Stat. Assoc., vol.60, no.309, pp.63-69, 1965.
[13] S.J. Rizvi, J.R. Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th VLDB conference, pp.1-12, 2002.
[14] W. Du, Z. Zhan, "Using Randomized Response Techniques for Privacy Preserving Data Mining", In Proceedings 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.505-510, 2003.
[15] Guo, S. Guo, X. Wu, "Privacy Preserving Market Basket Data Analysis", In Proceedings the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases,Pp.103-114, 2007.
[16] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5, pp.557-570, 2002.
[17] R. Bayardo, R. Agrawal, "Data Privacy Through Optimal k-Anonymization", In Proceedings the 21st International Conference on Data Engineering, pp.217-228, 2005.
[18] K. Lefevre, J. Dewittd, R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity", In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Pp.49-60, 2005.
[19] B. Fung, K. Wang, P. Yu, "Top-down Specialization for Information and Privacy Preservation", In Proceedings of the 21st IEEE International Conference on Data Engineering, pp.205-216, 2005.
[20] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5, pp.571-588, 2002.