Review on Novel Protocol for Secure Mining in horizontally Distributed Database

Nidhi Kumari
Computer Science and Engineering
G. H. Raisoni Academy of Engineering and Technology
Nagpur, India
singhnidhik@gmail.com

Sonali Bodkhe
Computer Science and Engineering
G. H. Raisoni Academy of Engineering and Technology
Nagpur, India
sonali.mahure@gmail.com

***Abstract*—**With the existence of large amount of data to be stored in databases and different repositories. It is very essential to develop a controlling and effective means for study and elucidation of such data for mining the motivating and valuable knowledge that may help in decision making. Data mining is a kind of procedure which excerpts the useful information from the large repositories. Another name of data mining is also known Knowledge Discovery Database. The techniques of secure data mining are introduced with the aim of taking out the related knowledge from the large amount of data while guarding the reasonable information at the same time. This paper reviews the information and also reviews the various secure data mining (SDM) techniques like data modification and secure multiparty computation based on the various characteristics simultaneously. We have analyzed the comparative study of all the existing techniques.

Keywords- *Secure Data Mining, Mining Techniques, Privacy Preserving, Secure Multiparty Computation (SMC) and Data Modification*

*****

I. INTRODUCTION

In today's scenario we have E-Governance, E-Commerce and environments where personal data is dispersed online. Such environments require more focus to be given on privacy of data. All the information in the distributed system contains most important property as security. The useful information found in mining can be susceptible or it can be exploited by anyone. Here we consider a situation in which two or more parties own their secure databases wants to run a their data mining algorithm on the union of their database without highlighting any confidential information. For example, consider medical institutes located at different places that wish to conduct a join medical research while securing the important or private records of their various patients. It is required to protect the sensitive information in this situation, but it is also required to enable its use for prospect research work. All parties are realizing that combining their data gives common profit but none of them is keen to reveal its database to other parties. For this reason various secure mining techniques are applied with data mining algorithm to protect the mining of sensitive information during the knowledge discovery.

Main objective of secure data mining is how to protect the private information or sensitive knowledge from leaking in the knowledge discovery process, meanwhile obtain the precise results of data mining.

The secure data mining is divided into two levels [1]:

- First level of SDM is focus on securing the sensitive data such as address, income, disease and other important information.
- Second level of SDM is focus on protecting the sensitive information which is showed by data discovery.

Various secure mining techniques are using some or other form of transformation in order to achieve privacy. Secure mining is mainly focused on data reconstruction, data encryption and data distortion technology. The SDM techniques implementation has become the demand now a days. The main aim of this paper is to present the review on Secure mining techniques which is very useful while mining process over large repositories with reasonable effectiveness and maintain security.

II. SECURITY ISSUES RELATED TO DATA MINING

Data mining is a extensively recognized technique for enormous range of organizations. Now a days Knowledge discovery is included in day to day operating activities of every organization. While from gathering of data to discovery of knowledge i.e. data mining we get the final outcome as data. These discovered data may contain all useful private information of individual one. So this private information may expose to various different entities including miners, users and data collector. Exposure of such information results in breaching of the individual's privacy. We take a simple example, revealing a patient's health related information can affect his social and personal life. Secure information of one can also be disclosed by linking various databases belong to large data warehouse [2] and accessing web data [3].

An unauthorized data miner can learn sensitive attributes or data values such as disease type or id of a certain individual through re-verification of record from an exposed data set. The combination of other data values also provides help to the malicious miner to recognize the sensitive data values. It is not guaranteed to provide the secrecy of private information if we remove other attributes. Satisfactory additional knowledge is also helpful for hikers to identify sensitive data values.

A. Public Awareness

Secondary use of data becomes very common these days. It means the use of data other than the purposes for which it was collected in the initial phase. The various misuse of personal information of public is increasing now days in rapid manner. The possibility of sensitive data is not limited to financial or medical records it may be phone calls made by an individual, buying patterns an individual has gone through and many more. No individual will agree selling his personal data to any other unwanted party without prior permission from him. Cautious approach of some of the individuals in sharing their information that incurs an additional discomfort in collecting

the true information. Awareness of public is of high importance if private information is shared between multiple entities. Awareness about security and lack of public trust in organization may lead to an additional complexity in the process of data gathering. Strong concern of public may force law forcing agencies and government to introduce new privacy protecting policy. For example according to US Executive Order federal employees being prepossess, on the basis of protected genetic information [4].

B. Secure Data Mining

Suitably the tremendous benefit of knowledge discovery and high concern of public regarding individual data security, implementation of Secure data mining has become demand of today's environment. Such technique provides individual security along while simultaneously allowing extraction of useful information from data.

There are variety of methods which can be implemented to enable secure data mining. Majority of the techniques use some or other form of transformation or modification. These techniques modify the collected data sets before it is being released in an attempt to protect individual records from being re-identified [5]. A malicious data miner or intruder even with added knowledge cannot be sure about the correctness of a re-identification, even in case the data set has been altered. Apart from the context of data mining it is important to maintain patterns in data set.

High data quality with privacy/security is the major requirement of good secure mining techniques.

III. CLASSIFICATION SCHEME AND EVALUATION CRITERIA FOR SECURE DATA MINING TECHNIQUE

There are various techniques for secure data mining. Each one of the techniques found to be appropriate for specific type of scenarios and achieving variety of objectives. Here we are presenting a differentiating schemes and evaluation criterion for those techniques. However these schemes and criteria are fabricated on the scheme and criteria proposed in [1].

A. Classification Scheme

Secure Data Mining techniques can be classified based on

- Data Mining Scenario
- Data Mining Tasks
- Data Distribution
- Data Types
- Privacy Definition
- Privacy Method

We describe these classifications characteristics as follows:

1) Data Mining Scenario: Mostly two data mining scenarios are found presently. The first one involves organizations to release their available data sets for mining purpose and permit the unhindered access to it. Data modification is the methodology used to attain the security in such scenarios. The second type of scenario does not require organizations to release their data sets but allows the data mining tasks to be accomplished. Encryption techniques are employed to for privacy preserving in such scenarios.

2) Data Mining Task: The patterns available in Data sets are extracted out using different types of data mining tasks like association rule mining, classification, clustering and outlier analysis, evolution analysis [6]. Maintaining data quality is the basic achievement of all secure mining techniques in order to support every possible data mining tasks and statistical analysis. However it is found that it maintains data quality parameter to support only a certain group of data mining tasks. This forms the basis on what the secure mining techniques are categorized.

3) Data Distribution: Data sets which are given as an input to data mining process are either centralized or distributed. It is independent of the physical location i.e. the sectors where data is stored but depends on the ownership/ availability of the data. The centralized data sets are managed by a individual organization. Its availability is either at computational site or it can be sent to the desired site. In contrast, the distributed data sets are shared among two or more organizations which may or may not trust on secure data of each other but they generally pay significant amount of attention on performing data mining of joint data. The data sets used can be of heterogeneous type meaning they are vertically partitioned where each organization owns the same set of attributes but diverse subset of attributes. Instead they can be of homogeneous nature which means they are horizontally partitioned where each organization owns the same set of attributes but diverse subset of records.

4) Data Types: Fundamentally there are two attributes of data set: Numerical and Categorical. There is special case of Categorical type that takes only two possible values 0 and 1. Such a type is called Boolean. Natural Ordering is the factor which is deficit in Categorical type of values. This constitutes the elementary difference between the categorical and numerical type of values which forces the privacy preservation technique to take dissimilar kinds approaches for them.

5) Privacy Definition: The definition of privacy varies depending on the context in which it is being signified. In some environments, individual's data values are private, whereas in others a certain association or classification rules are private. Depending upon the privacy definition we work on different Secure mining techniques.

A. Evaluation Criteria

It is essential to formulate the evaluation criteria and related standards. Some of the evaluation criteria are as follows:

1) Versatility: It denotes the capability of the techniques to provide for various data mining task, privacy requirements and types of data set. The technique is more useful if it is more private i.e. data and attribute.

2) Disclosure Risks: It cites to the probabilities of delicate information being derived by a malicious data miner. It is inversely proportional to the level of security offered by the technique under consideration. Development of the disclosure risks is tough task, since it depends on various factors like supplementary knowledge of an intruder and nature of the techniques. Secure mining primarily aims to minimize the disclosure risk and hence the risk evaluation is essential.

3) Information Loss: Information loss is generally proportional to the amount of noise or disturbance sustained and level of security employed in the technique. However it is inversely proportional to the quality of data incurred by the technique. The principal requirement of the secure mining

technique is maintenance of a high data quality in released data sets. High level of security is of no use if data quality is not maintained in desired manner.

4) Cost: Cost as a whole sums up to both the computation cost and the communication cost between the collaborating parties [1]. Computational cost includes both preprocessing cost (e.g., initial perturbation of values) and running cost (e.g., processing overhead). If data set is of distributed type then the communication cost becomes significant concern. As the cost of the technique is on the higher side, the efficiency of technique is found to be on the lower side.

IV. TECHNIQUES OF SECURE MINING

Now we see some techniques of secure data mining:

A. Data Modification

For centralized databases, the existing secure mining method can be categorized in three main groups based on the approaches they take, such as data modification, query restriction, and output perturbation [7]. Data modification is a straightforward technique to implement from all these techniques. In this technique before the release of a dataset for various data mining tasks and analysis, it modifies the data set for security of individual privacy by keeping high quality of released data. After this modification we can use any off the shelf software (i.e. See5) to analyze or manage the data. It is not with the case of output perturbation and query restriction. This technique is made simple, attractive and widely used in the context of arithmetical database in data mining. Various number of ways of doing this data modification such as swapping, aggregation, noise addition, suppression. Following are some basic idea of these techniques.

1) Data Swapping: Dalenius and Reiss in 1982 were introduced the data swapping technique, in the context of secure statistical databases for categorical values modification [8]. The main idea of this method was it makes the record re-identification very complex and keeps all original value in the data set. It replace the original data set by another different data where some original values belonging to a sensitive attributes are exchanged between them. Existing data swapping technique introduction can be found in [9], [10].

Inspired by existing techniques a new data swapping techniques is introduced for secure data mining. This technique emphasizes on the pattern preservation instead of finding out unbiased statistical parameters. It performs the task of preserving the most classification rule even if they are acquired by different classification algorithm.

2) Aggregation: It is also known as generalization or global recording and is used for securing an individual privacy in a related data set before its releasing by perturbing the original dataset. Aggregation change k number of records of a data by representative records. An attributes value in such a representative record is derived by taking the average of all values, for the attributes, belonging to the records that are replaced. Other method of generalization or aggregation is attribute values transformation. For example- a correct birth date can be changed by the year of birth. Generalization makes an value of attribute less informatics. For example- if accurate birth date is changed by the century of birth then the released data can became useless to the miners [11].

3) Noise Addition in Data Mining: Noise addition phenomenon in data mining refers to addition of a noise (random number) to numerical attributes. Such a random number is usually drawn from a normal distribution with zero mean or standard deviation. Noise addition is performed in a controlled way in order to maintain variance, co-variance and means of the attributes of a data set. However due to absence of natural ordering in categorical values, adding of noise in categorical attributes is not as simple as like addition of noise in numerical attributes. Various techniques are proposed for addition of noise in data mining. Evfimievski et al. suggested a novel noise addition technique for privacy preserving association rule mining in 2002 [12]. Agarwal and Srikant advised a noise addition technique in 2000 which is based on addition of random noise to attribute values in such a way that the distributions of data values that belongs to original and disordered or perturbed kind of data set were very challenging [5]. Du and Zhan proposed a decision tree building algorithm which is used to unhinge multiple attributes [13]. In 2004 Zhu and lieu [14] proposed a generalized framework for randomization using a well studied statistical model called mixture model. As per this scheme, the data are generated from a distribution that relies on certain factors including original data as well. Their randomization framework confirms secure density estimation.

B. Secure Multiparty Computation (SMC)

A secure Multi-party Computation (SMC) technique encodes the data sets, while as the same time allowing data mining operations to be carried out in smooth fashion. These techniques are not speculated to disclose any new type of information other than the final outcome of the computation to a participating party. Such techniques are naturally based on available cryptographic protocols and are implemented on distributed type of data sets. Parties taking part in a distributed data mining, encrypts their data and sends it to others parties. These encrypted data are used as an input in the process of computing the aggregate data which belongs to the joint data set and which is in turn used for data mining purpose. Secure Multipart Computation was originally proposed by Yao in 1982 [15]. Basically, SMC is considered to disclose to a party only the result of the computation and the data owned by the party. There are several SMC algorithms formulated. Most of the algorithms deploy some primitive computations such as secure sum, secure set union, secure size of set intersection and secure scalar product.

V. COMPARATIVE STUDY

In this section we present the comparative study of all the secure mining techniques based on the evaluation criterion which we have already discussed above. The comparative study depicted in following table will give us a clear idea about which technique is best suitable for which scenario. We present the comparative study in the tabular form which is shown in Table 1:

Basically Secure Multiparty Computation techniques tend to suffer a significantly higher running cost, but they are able to provide much higher level of security. It doesn't reveal anything other than the final results such as the classification rules, cluster and association rules. Hence, such techniques are

suitable in a particular scenario where multiple parties agree to cooperate for only the final result extraction from their combined data set. However, in a situation where a data set is required to be released to simplify the process for various researches and extract general knowledge, Data modification is the evident and mostly preferred choice. Data Modification usually experiences less computational costs and less amount of information loss as well.

Name	Private: Data/Rules	Dataset: Central/Distribute (Vertical/Horizontal)	Attributes: Categorical/ Numerical/ Boolean	Data Mining Task	Disclosure Risk	Information Loss	Cost
Outlier Detection	Data(Both)	Distributed	Both	Outliers	Very Low	None	High
Association Rules	Data	Distributed	Boolean	Association Rules	Very Low	None	Moderate
Randomized Noise	Data	Both	Numerical	Association	Very Low	Moderate	Low
Secure Multiparty Computation	Data	Diatributed	Numerical	Association Rule	Very Low	Moderate	Moderate

Table 1: Secure Mining Techniques- Comparative Study

VI. CONCLUSION

In this paper we made an attempt to present a detailed study of security threats which are incurred during mining of association rule in horizontally distributed databases. And briefly review security including cryptographic technique to minimize security. We also sought to present the comparative study of privacy preserving techniques which can prove to be obliging means to select the best techniques depending on the scenarios in which it is to be employed. Secure mining of data is viewed as an important need for all the organizations (or parties) since they have their own data which is required to be managed in an efficient manner. Secure mining also finds its usefulness in order to refine and develop own techniques or protocols for achieving a high level of efficiency and security.

In this paper we present the detailed theoretical study in the context of secure data mining and concisely reviewed the techniques of Data Modification and Secure Multiparty Computation. Secure data mining is becoming a essential need for all the organization, since the organizations are very typical in taking care of their data due to which they can formulate the strategies that are useful in business expansion and enhance the revenues. All the methods discussed here are only approximate to our aim of privacy preservation, it will be needed to further refine these approaches or design some efficient techniques. For considering these

- Efforts are required to be made to device more efficient methodologies which will make an attempt to achieve balance between computation and communication cost.
- A tradeoff between Accuracy and privacy is required to be achieved as both of the constraints are closely related to each other and attempts to improve one affects the other significantly.

REFERENCES

[1] V.S. Verykios, E. Bertino, I. N. Fovino, L. P. Provonza, Y.Saygin and Y. Theodoridis. State-of-the-art in privacy preserving data mining. SIGMOD Record, 33 (1): 50-57, 2004.

[2] S. E. Fienberg. Privacy and confidentiality in an e-commerce world: Data mining, data warehousing, matching and disclosure limitation. Statistical Science, 21:143-154, 2006.

[3] B. M. Thuraisingham. Data mining, national security, privacy, and civil liberties. SIGKDD Explorations, 4 (2): 1-5, 2002.

[4] US Department of Labor. Executive order 13145. Available from http://www.dol.gov/oasam/regs/statutes/eo13145.htm, Feb 8, 2000.

[5] R. Agarwal and R. Srikant. Privacy –preserving data mining. In Proc. Of the ACM SIGMOD Conference of Management of Data, pages 439-450. ACM Press, May 2000.

[6] J. Han and M. Kamber, Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, San Diego, CA92101-4495, USA, 2001.

[7] N. Adam and J. C. Wortmann. Security control methods for statistical databases: A comparative study. ACM Computing Surveys, 21 (4): 515-556, 1999.

[8] T. Dalenius and S. P. Reiss. Data Swapping: A technique for disclosure control. Journal of Statistical Planning and Inference, 6(1):73-85, 1982.

[9] S. E. Fienberg and J. McIntyre. Data swapping: Variations on a theme by Dalenius and Reiss. Journal of Official Statistics, 21:309-323, 2005.

[10] K. Murlidhar and R. Sarathy. Data Shuffling – a new masking approach for numerical data. Management Science, Forthcoming, 2006.

[11] V. S. Iyenger. Transforming data to satisfy privacy constraints. In ***Proc. Of SIGKDD'02***, Edmonton, Alberta, Canada, 2002.

[12] S. Rizvi and J.R Hartisa. Maintaining data privacy in association rule mining. In Proc. of the 28th VLDB Conference, pages 682-693, Hong-Kong, China, 2002.

[13] W. Du and Z. Zhan. Using randomized response techniques for privacy preserving data mining. In Proc. of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 515-510, Washington DC, USA, August 2003.

[14] Y. Zhu and L. Liu. Optimal randomization for privacy preserving data mining. In Proc. of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 761-766, Seattle, Washington, USA, August 2004.

[15] A. C Yao. Protocols for secure computations. In Proc. of the 23rd Annual IEEE Symposium on Foundation of Computer Science, 1982.