ISSN 2047-3338

A Review Study on the Privacy Preserving Data Mining Techniques and Approaches

Manish Sharma[1], Atul Chaudhary[2], Manish Mathuria[3] and Shalini Chaudhary[4]
[1,3,4]Department of CSE, Govt. Engineering College, Ajmer
[2]Department of I.T, Govt. Engineering College, Ajmer
[1]manu13raj@gmail.com, [2]atul.chaudhary83@gmail.com, [3]manish.4589@gmail.com, [4]cshalini14@yahoo.in

***Abstract*– With the extensive amount of data stored in databases and other repositories it is very important to develop a powerful and effective mean for analysis and interpretation of such data for extracting the interesting and useful knowledge that could help in decision making. Data mining is such a technique which extracts the useful information from the large repositories. Knowledge discovery in database (KDD) is another name of data mining. Privacy preserving data mining techniques are introduced with the aim of extract the relevant knowledge from the large amount of data while protecting the sensible information at the same time. In this paper we review on the various privacy preserving data mining techniques like data modification and secure multiparty computation based on the different aspects. We also analyze the comparative study of all techniques followed by the future research work.**

***Index Terms*– Privacy and Security, Data Mining, Privacy Preserving, Secure Multiparty Computation (SMC) and Data Modification**

I. INTRODUCTION

IN today's scenario we have E-Commerce, E- Governance and personal data is distributed online, privacy of data is become the most important issue. All the distributed information system contains most important property as privacy. The information found in mining can be sensitive or it can be misuse by anyone. We consider a scenario in which two or more parties own their confidential databases wishes to run a data mining algorithm on the union of their database without revealing any private information. For example, consider separate medical institutes that wish to conduct a join medical research while preserving the privacy records of their patients. In this condition it is required to protect the sensitive information, but it is also required to enable its use for future research work. Involving parties are realizing that combining their data gives mutual benefit but none of them is willing to reveal its database to other parties. For this reason various privacy preserving techniques are applied with data mining algorithm in order to protect the extraction of sensitive information during the knowledge finding.

Main research objective of privacy preserving data mining (PPDM) is how to protect the sensitive information or private knowledge from leaking in the mining process, meanwhile obtain the accurate results of data mining.

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

- First level of PPDM is focus on protecting the sensitive data such as id, name, address and other sensitive information.
- Second level of PPDM is focus on protecting the sensitive knowledge which is showed by data mining.

Many privacy preserving techniques are using some form of transformation to achieve privacy. Privacy preserving is mainly focused on data distortion, data reconstruction and data encryption technology. The implementation of PPDM techniques has become the demand of the moment. The goal of this paper is to present the review on privacy preserving techniques which is very helpful while mining process over large data sets with reasonable efficiency and preserve security.

II. PRIVACY ISSUES RELATED TO DATA MINING

As we know that data mining is a widely accepted technique for vast range of organizations. Data mining is included in day to day operational activities of every organization. During the whole process of data mining (from collection of data to discovery of knowledge) we get the data. These data may contain sensitive information of individual one. This information may expose to several other entities including data collector, users and miners. Disclosure of such information is results breaking the individual privacy. We take a simple ex: exposed credit card details of a user will affect its social and economic life. Private information can also be disclosed by linking multiple databases belong to giant data warehouse [2] and accessing web data [3].

A malicious data miner can learn sensitive data values or attributes such as income ($8500) or disease type (HIV Positive) of a certain individual through re-identification of record from an exposed data set. The combination of other attributes also provides help to the intruders to identify the

sensitive data values. If we remove other attributes then it's not guaranteed to provide the confidentiality of private information. Sufficient supplementary knowledge is also helpful for intruders to identify sensitive data values.

A. Public Awareness

Now days secondary use of data become very common. Secondary use of data means data is used for some other purpose not for which data is collected initially. The potential misuse of personal information of public is increasing rapidly. The scope of sensitive data is not limited to medical or financial records it may be phone calls made by an individual, buying patterns and many more. No one wants that his/her personal data is sold to any other party without their prior consent. Some individuals become hesitant to share their information which results additional difficulty to obtaining correct information. Public awareness is so much important if private information is shared between different entities. Public awareness about privacy and lack of public trust in organization may introduce additional complexity to data collection. Strong public concern may force government and law forcing agencies to introduce new privacy protecting regulations. For example federal employees being prepossess, on the basis of protected genetic information, according to US Executive Order (2000) [4].

B. Privacy Preserving Data Mining

Due to the tremendous benefit of data mining and high public concern regarding individual data privacy, implementation of privacy preserving data mining has become demand of today's environment. This technique provides individual privacy while at the same time allowing extraction of useful knowledge from data.

There are several methods which can be used to enable privacy preserving data mining. Most of the techniques use some form of transformation or modification. These techniques modified the collected data set before its release in an attempt to protect individual records from being re-identified [5]. A malicious data miner or intruder even with additional knowledge cannot be certain about the correctness of a re-identification, even when the data set has been modified. 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 privacy preserving techniques.

III. CLASSIFICATION SCHEME AND EVALUATION CRITERIA FOR PRIVACY PRESERVING DATA MINING TECHNIQUES

There are various techniques for privacy preserving data mining. Each of the technique is suitable for particular type of scenario and objectives. We are here presented a classification scheme and evaluation criteria for those techniques. However these schemes and criteria build on the classification scheme and evaluation criteria proposed in [1].

A. Classification Scheme

Privacy preserving techniques can be classified based on following characteristics:

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

We describe these classifications characteristics as follows:

1) Data Mining Scenario: There are basically two major data mining scenario present. In the first one organization release their data sets for data mining and allowing unrestricted access to it. Data modification is used to achieve the privacy in this scenario. In the second one organization do not release their data sets but still allow data mining tasks. Cryptographic techniques are basically used for privacy preserving.

2) Data Mining Task: Data set contains various patterns. These patterns are taken out through different types of data mining tasks like classification, association rule mining, outlier analysis, clustering and evolution analysis [6]. Basically, all privacy preserving techniques should maintain data quality to support all possible data mining tasks and statistical analysis but it usually maintain data quality to support only a group of data mining tasks. Basis on that task we categorize the privacy preserving techniques.

3) Data Distribution: Data sets used for data mining can be either distributed or centralized. It is not depending on the physical location where data is stored but to the availability/ownership of data. The centralized data set is owned by a single party. It is either available at computational site or it can be sent to the site. However, distributed data set is shared between two or more parties which do not necessarily trust each other private data but interested to perform data mining on joint data. The data set can be heterogeneous means vertically partitioned where each party owns the same set of attributes but different subset of attributes. Alternatively it can be homogeneous means horizontally partitioned where each party owns the same set of attributes but different subset of records. In Fig. 1 we shows the classification based on distribution.

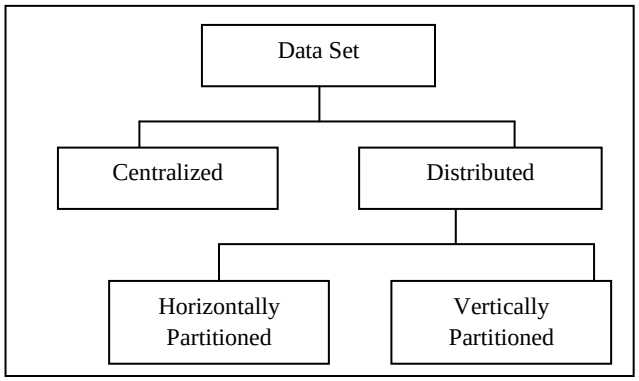

Fig. 1: Classification of Different Dataset Based on Distribution

1) Data Types: There are basically two attributes in data set: Numerical and Categorical. Boolean data are the special case of categorical data which takes two possible values 0 and 1. Categorical values lack natural ordering in them. This is the basic difference between categorical and numerical values and its force the privacy preservation technique to take different approaches for them.

2) Privacy Definition: The definitions of privacy are different in different context. In some scenario individuals data values are private, whereas in other scenario certain association or classification rules are private.. Depend on the privacy definition we work on different privacy preserving techniques.

3) Protection Methods: Privacy in data mining is protected through different methods such as data modification and secure multiparty computation (SMC). On the basis of protection method we can also categorize the privacy preserving techniques. The classification is shown in Fig. 2.

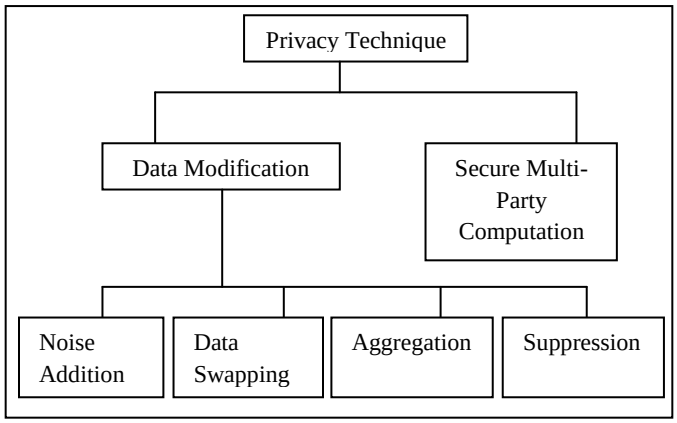

Fig. 2: A Classification of Privacy Preserving Techniques

We see the detailed study of all the techniques in the later section.

Now we discuss the evaluation criteria as follows:

B. Evaluation Criteria

It is important to determine the evaluation criteria and related benchmarks. Some evaluation criteria are:

1) Versatility: It refers to the ability of the techniques to cater for various data mining task, privacy requirements and types of data set. The technique is more useful if it is more versatile. Versatility includes the following:

- Private: Data vs. Patterns
- Data Sets: Distributed or Centralized (Vertical or horizontal)
- Attributes: Numerical or Categorical
- Data Mining Tasks

2) Disclosure Risks: It refers to the chances of sensitive information being inferred by a malicious data miner. It's inversely proportional to the level of security which is offered by the technique. Development of the disclosure risks is difficult task, since it depends on many factors like supplementary knowledge of an intruder and nature of the techniques. Primary objective of privacy preservation technique is minimizing the disclosure risk, so risk evaluation is essential.

3) Information Loss: Information loss is usually proportional to the amount of noise added and level of security. It is inversely proportional to data quality. The primary requirement of privacy preserving technique is to maintain high data quality in released data sets. If data quality is not maintained then high security will be useless.

4) Cost: Cost refers to both computation cost and communication cost between the collaborating parties [1]. Computational costs contain both preprocessing cost (e.g., initial perturbation of values) and running cost (e.g., processing overhead). If data set is distributed then communication cost becomes important issue. The higher the cost, the lower the efficiency of the technique.

IV. TECHNIQUES OF PRIVACY PRESERVING

Now we see some techniques of privacy preserving data mining:

A. Data Modification

Existing privacy preserving techniques method for centralized databases can be categorized in three main groups based on the approaches they take, such as query restriction, output perturbation and data modification [7]. From all these techniques data modification is a straightforward technique to implement. In data modification before the release of a dataset for various data mining tasks and analysis, it modifies the data set for protection of individual privacy while the quality of released data remains high. After this modification of the data we can use any off the shelf software such as See5 to manage or analyze the data. It's not with the case of query restriction and output perturbation. The simplicity of this technique made it attractive and widely used in the context of statistical database in data mining. There are number of ways of doing data modification such as suppression, swapping, aggregation and noise addition. The basic idea of these techniques is given below.

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

Inspired by existing techniques a new data swapping techniques is introduced for privacy preserving data mining. This technique focus on the pattern preservation instead of obtaining unbiased statistical parameters. Basically it preserves the most classification rule even if they are obtained by different classification algorithm.

2) Aggregation: Aggregation is also known as generalization or global recording. It is used for protecting an

individual privacy in a released data set by perturbing the original data set before its releasing. Aggregation change k no. of records of a data by representative records. The value of an attribute in such a representative record is generally derived by taking the average of all values, for the attributes, belonging to the records that are replaced. Another method of aggregation or generalization is transformation of attribute values. For ex- an exact birth date can be changed by the year of birth. Such a generalization makes an attribute value less informatics. For ex- if exact birth date is changed by the century of birth then the released data can became useless to data miners [11].

3) Suppression: In this technique sensitive data value are deleted or suppressed prior to the release of a micro data. Suppression is used to protect an individual privacy from intruders attempt to accurately predict a suppressed value. To predict a sensitive value an intruder can use various approaches. An important issue in suppression is to minimize the information loss by minimizing the number of values suppressed. For some applications, such as medical, suppression is preferred mainly over noise addition in order to reduce the chance of having misleading pattern in perturbed data set. Suppression technique is also been used for association and classification rule confusion [12], [13].

4) Noise Addition in Data Mining: Noise addition is usually adds a random number (noise) to numerical attributes. This random number is generally drawn from a normal distribution with zero mean or standard deviation. Noise is added in a controlled way so as to maintain variance, co-variance and means of the attributes of a data set. Due to the absence of natural ordering in categorical values, addition of noise in categorical attributes is not straightforward as like addition of noise in numerical attributes. Many techniques proposed for noise addition in data mining. Evfimievski et al. proposed a novel noise addition technique for privacy preserving association rule mining in 2002 [14]. Agarwal and Srikant proposed 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 belonging to original and perturbed data set were very difficult [5]. Du and Zhan presented a decision tree building algorithm which is used to perturb multiple attributes [15]. In 2004 Zhu and lieu [16] proposed a general framework for randomization using a well studied statistical model called mixture model. According to this scheme data are generated from a distribution that depends on some factors including original data as well. Their randomization framework supports privacy preserving density estimation.

C. Secure Multiparty Computation (SMC)

A Secure Multi-party Computation (SMC) technique encrypts the data sets, while still allowing data mining operations. SMC techniques are not supposed to disclose any new information other than the final result of the computation to a participating party. These techniques are typically based on cryptographic protocols and are applied to distributed data sets. Parties involved in a distributed data mining encrypt their data and send to others parties. These encrypted data are used to compute the aggregate data, belonging to the joint data set, which is used for data mining purpose. Secure Multipart Computation was originally introduced by Yao in 1982 [17]. Basically, SMC is supposed to reveal to a party just the result of the computation and the data owned by the party. There are various SMC algorithms developed. Most of the algorithms make use of some primitive computations such as secure sum, secure set union, secure size of set intersection and secure scalar product.

V. COMPARATIVE STUDY

Now we present the comparative study of all the privacy preservation techniques on the basis of evaluation criteria which we already discussed above. This comparative study gives us the clear idea about which technique is better in which scenario. We present the comparative study in the tabular form which is shown in Table 1:

TABLE 1: PRIVACY PRESERVING TECHNIQUES – COMPARATIVE STUDY

Name	**Versatility**				**Disclosure Risk**	**Information Loss**	**Cost**
	Private: Data/Rules	**Dataset: Central/Distributed (Vertical/Horizontal)**	**Attributes: Categorical/ Numerical/ Boolean**	**Data Mining Task**			
Outlier Detection	Data (Both)	Distributed	Both	Outliers	Very Low	None	High
Association Rules	Data	Distributed (Vertical)	Boolean	Association Rules	Very Low	None	Moderate
Randomized Noise	Data	Both	Numerical	Classification	Low	Moderate	Low
Correlated Noise	Data	Centralized	Numerical	Clustering, Classification	Moderate	Low	Low
Decision Tree Noise	Data	Centralized	Both	Classification	Moderate	Low	Low
Randomized Response	Data	Both	Numerical	Density Estimation	Low	Moderate	Low
Secure Multiparty Computation (SMC)	Data	Distributed (Vertical)	Numerical	Association Rules	Very Low	None	Moderate

Generally Secure Multiparty Computation techniques tend to incur a significantly higher running cost, but also to provide much higher level of security. It does not disclose anything other than the final results such as the classification rules, cluster and association rules. Therefore, it is suitable in a particular scenario where multiple parties agree to cooperate for just the final result extraction from their combined data set. However, in a scenario where a data set is supposed to be released to facilitate to various research and extract general knowledge, Data modification is the obvious choice. Data Modification usually incurs less computational cost and less information loss as well.

VI. CONCLUSION

In this paper we present the detailed study about the privacy preserving data mining and briefly review the techniques Data Modification and Secure Multiparty Computation. We tried to present the comparative study of privacy preserving techniques which is helpful to understand that which technique is better in which scenario or environment. Privacy preserving data mining is an important need for all the organization because every organization has their own personal data and they care about their data. All the methods discussed here are only approximate to our goal of privacy preservation now we need to further refine those approaches or develop some efficient techniques. For considering these issues, following problem should be widely studied:

- As we know data mining is a very important issue in distributed privacy preserving data mining. We should try to develop more efficient algorithms and achieve a balance between computation, communication and disclosure cost.
- Accuracy and privacy are closely related to each other. If we tried to improving one then other effected accordingly. So how to achieving tradeoff between both is an important research.

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, CA 92101-4495, 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] Y. Saygin, V. S. Verykios and A. K. Elmagarmid. Privacy preserving association rule mining. In *RIDE*, pages 151-158, 2002.

[14] A. V. Evfimievski, R. Srikant, R. Agarwal and J. Gehrke. Privacy preserving mining of association rules. In *Proc. Of the Eighth ACM SIGKDD International Conference on Knowledge and Data Mining*, pages 217-228, 2002.

[15] 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.

[16] 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.

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

[18] M. Kantarcioglu and C. Clifton. Privacy preserving distributed mining of association rules on horizontally partitioned data. *IEEE Trans. Knowl. Data Eng*. , 16 (9):1026: 1036, 2004.

[19] M. Kantardzic. *Data Mining Concepts, Models, Methods and Algorithms*. IEEE Press, NJ, USA, 2003.

[20] H. Kargupta, S. Datta, Q. Wang and K. Sivakumar. Random data perturbation techniques and privacy preserving data mining. *Knowledge and Information Systems*, 7:387-414, 2005.