A Review on Privacy Preserving Data mining

[1]Shabija M B, [2] Lekshmy P L
[1]M.Tech Student, [2]Assistant Professor
[1]Departmentof Computer Science and Engineering,
[1]LBS Institute of Technology for Women, Thiruvananthapuram,India

***Abstract*—Data mining technology has emerged as an identifying patterns and trends from large quantities of data. Data mining is being used in wide areas such as banking, medicine, scientific research and among government agencies. With increasing usage of data mining in the public and private sectors, privacy assumes paramount importance. In recent years, with the explosive development in Internet, data storage and data processing technologies, privacy preservation has been one of the greater concerns in data mining. A number of methods and techniques have been developed for privacy preserving data mining. In this survey describes various techniques used in privacy preserving data mining.**

***Index Terms*— Privacy Preserving data mining (PPDM), Secure Multi-party Computation, Perturbation, Classification Rule, Association Rule.**

I. INTRODUCTION

Data mining is concerned with extraction of non-trivial, novel and potentially useful knowledge from large databases. Continuous data mining techniques have been applied to a wide range of areas such as customer relationship management, web mining, banking and medicine. The extracted information could be in the form of patterns, clusters and classification models. Many security and counter-terrorism related decision support application need data mining techniques for identifying emerging behavior, link analysis, building predictive models, and extracting social networks. The power of data mining tools to extract hidden information from large collection of data lead to increased data collection efforts by companies and government agencies. Naturally this raised privacy concerned about collected data. In response to that, data mining researchers started to address privacy concerns by developing special data mining techniques under the framework of privacy preserving data mining. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. The main consideration in privacy pre-serving data mining is two folds. First, sensitive raw data should be modified or trimmed out from the original database, in order for recipient of the data not to be able to compromise privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded. In a nutshell, the privacy preserving mining methods modify original data in some way, so that the privacy of the user data is preserved and at the same time the mining models can be reconstructed from the modified data with reasonably accuracy.

II. PRIVACY PRESERVING DATA MINING FRAMEWORK

In the paper [1], describes the framework of the privacy preserving data mining. Data from different data sources or operational systems are collected and are preprocessed using ETL tools. T his transformed and clean data from level 1 is stored in the data warehouse. Data in the data warehouse is used for mining.

At level 2, data from data warehouse is subjected to various processes that make data sanitized so that it can be revealed even to untrustworthy data miners. The processes applied this stage are blocking, suppression, perturbation, modification, generalization, sampling etc. Then the data mining algorithms are applied to processed data for knowledge/information discovery. Even the data mining algorithms are modified for the purpose of protecting privacy without sacrificing the goals of data mining.

At level 3, the information/knowledge releaved by data mining algorithms is checked for its sensitiveness towards disclosure risks. The embedding of privacy concerns at three levels, but any combination of these may be used.

III. LITERATURE SURVEY

In [2] Xinjun Qi and Mingkui Zong describe privacy preserving data mining classified based on the following dimensions:

1. Data Distribution
2. Data Distortion
3. Data mining algorithms
4. Data or rule hiding
5. Privacy preservation

The first dimension refers to the distribution of data.Some of the approaches have been developed for centralized data,while others refer to distributed data scenario.Distributed data scenarios can be classified as horizontal data distribution and vertical data distribution.Horizontal distribution refers to these cases where different database records reside in different places,while vertical data distribution,refers to the cases where all the values for different attributes reside in different places.

The second dimension refers to the data distortion scheme. In general, data distortion is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection.It is important that a data modification technique should be in concert with the privacy policy adopted by an organization.Methods of modification include:

Perturbation:

This is accomplished by the alteration of an attribute value by a new value (i.e., changing 1 value to a 0 value, or adding noise.

Blocking:

In blocking, the entry is not modified, but is left in complete. Thus, unknown entry values are used to prevent discovery.

Aggregation:

Known as merging, this is the combination of several values into a coarser category.

Swapping:

This refers to interchanging values of individual records.

Sampling:

This refers to releasing data for only a sample of a population.

The third dimension refers to the data mining algorithm,for which the data modification is taking place.The most important ideas have been developed for classification data mining algorithms like decision tree inducers,association rule mining algorithms,clustering algorithms ,rough sets and bayesian networks.

The fourth dimension refers to the hide original data or rules of original data. Due to rules hidden of original data is very complex, some person proposed heuristic method to solve issue.

The last dimension refers to privacy preservation; in order to protect privacy there need to modify data carefully for achieving a high data utility. Do this for some reasons as:

(1) Modify data based on adaptive heuristics method, and only modify selected values of, but not all values, which make information loss of data is minimum.
(2) Encryption technologies, such as secure multiparty computation. If each site know only their input and input but nothing about others, the calculations are safe.
(3) Data reconstruction method can reconstruct original data distribution from random data.

Based on these dimensions, different PPDM techniques may be classified into following categories:

[1] Anonymization based PPDM
[2] Perturbation based PPDM
[3] Randomized Response Based PPDM
[4] Condensation approach based PPDM
[5] Cryptography based PPDM

The main contribution of paper [3] is Anonymization based PPDM.The basic form of the data in a table consists of four types of attributes:

(1) Explicit identifiers is a set of attributes containing in-formation that identifies a record owner explicitly such as name, SS number etc.
(2) Quasi identifier is a set of attributes that could potentially identify a record owner when combined with publicly available data.
(3) Sensitive Attributes is a set of attributes that contains sensitive person specific information such as disease, salary etc.
(4) Non-sensitive Attributes is a set of attributes that creates no problem if revealed even to untrustworthy parties.

Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden. It even assumes that sensitive data should be retained for analysis. It’s obvious that explicit identifiers should be removed but still there is a danger privacy intrusion when quasi identifiers are linked to publicly available data Such attack are called linking attack .For example attribute such as DOB, Sex, Race, and Zip are available in public records such as voter list. Such records are available in medical records also, when linked, can be used to infer the identity of the corresponding individual with high probability. Sensitive data in medical record is disease or even medication prescribed. The quasi-identifiers like DOB, Sex, Race, Zip etc. are available in medical records and also in voter list that is publicly available. The explicit identifiers like Name, SS number etc. have been removed from the medical records. Although the Anonymization method ensures that the tranformed data is true but suffers heavy information loss.Moreover it is not immune to homogeneity attack and background knowledge attack. Limitations of k-anonymity model stem from the two conventions.First, it may be very tough for the owner of a database to decide which of the attributes are available or which are not available in external tables. The second limitation is that the k-anonymity model adopts a certain method of attack, while in real situations; there is no reason why the attacker should not try other methods. However,as a research direction, k-anonymity in combination with other privacy preserving methods needs to be investigated for detecting and even blocking k-anonymity violations.

Anand Sharma et al[4] describes cryptographic based techniques in PPDM.A number of cryptography-based approaches have been developed in the context of privacy preserving data mining algorithms, to solve problems of the following nature.Two or more parties want to conduct a computation based on their private inputs, but neither party is willing to disclose its own output to anybody else. The issue here is how to conduct such a computation while preserving the privacy of the inputs. This problem is referred to as the Secure Multiparty Computation (SMC) problem. In particular, an SMS problem deals with computing a probabilistic function on any input, in a distributed network where each participant holds one of the inputs, ensuring independence of the inputs, correctness of the computation, and that no more information is revealed to a participant in the computation than that is participants input and output.

IV. CONCLUSION

The main objective of privacy preserving data mining is developing algorithm to hide or provide privacy to certain sensitive information so that they cannot be disclosed to unauthorized parties or intruder. Although a Privacy and accuracy in case of data mining is a pair of ambiguity. Succeeding one can lead to adverse effect on other. Finally, conclude there does not exists a single privacy preserving data mining algorithm that outperforms all other algorithms on all possible criteria like performance, utility, cost, complexity, tolerance against data mining algorithms etc. Different algorithm may perform better than another on one particular criterion.

IV. ACKNOWLEDGMENT

We are greatly indebted to our principal Dr. K. C. RAVEENDRANATHAN, Dr. SHREELEKSHMI R., Professor, Head of the Department of Computer Science and Engineering ,Mr. MANOJ KUMAR G., Associate Professor, Department of Computer Science and Engineering, LBS Institute of Technology for Women who has been instrumental in keeping my confidence level high and for being supportive in the successful completion of this paper. We would also extend our gratefulness to all the staff members in the Department; also thank all my friends and well-wishers who greatly helped me in my endeavor. Above all, we thank the Almighty God for the support, guidance and blessings bestowed on us, which made it a success.

REFERENCES

[1] Shweta Taneja et al , A Review On Privacy Preserving Data Mining: Techniques and Research Challenges, , International Journal of Computer Science and Informatiom Technology, Vol 5(2),2014,2310-2315

[2] Xinjun Qi and Mingkui Zong, An Overview of privacy Preserving Data Mining, International Conference on Environmental Science and Engineering, 2011

[3] A Hussien et al, Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing, , Journal of Information Security, , 2013

[4] Anand Sharma and Vibha Ojha, Implementation of cryptography for privacy preserving data mining , International Journal of Database Management Systems, vol 2,No:3,Aug 2013.

[5] MohammadReza Keyvanpour et al, Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modifi-cationbased Framework,International Journal on Computer Science and Engineering , vol 3,No:2,Feb 2011.