A Review on Different Techniques of Privacy Preserving Mining with Attacks

Anand Bhushan Pandey[1] , S R Yadav[2]

[1]M. Tech Scholar CSE, [2]Prof. and HOD CSE,

Department of computer science and Engineering, Millennium Institute of Technology, RGPV, Bhopal

Abstract - Privacy Preserving Data Mining (PPDM) extracts knowledge from large databases and also protects the sensitive information from being misused. By the development of the techniques that incorporate privacy issues, the data mining research will be enhanced. As the importance of business transaction data has increased manifolds and the data has become an essential part of any business. This paper focus on various approaches implement by the miners for preserving of information at individual level, class level, etc. A detail description with limitation of different techniques of privacy preserving is explained. This paper explains different evaluation parameters for the analysis of the preserved dataset.

Keywords: Privacy Preserving Mining, Association Rule Mining, Data Perturbation.

I. INTRODUCTION

Extracting information from large data bases is known as data mining. Data mining discovers new patterns from large data sets which are used in different fields like research, marketing analysis, medical diagnosis, atmosphere forecast etc. Privacy advocates criticize data mining because of a misunderstanding about what it actually is and about how it is normally done. This is also a cause of concerns that for a variety of intrusive or malicious purposes ones personal data may be used. Privacy preserving data mining is used to fulfil data mining objectives without any compromise with the privacy of the individuals and the privacy of the underlying original data values.

Association rule mining is a technique in data mining that identifies the regularities found in large volume of data [1, 2]. This technique could be compromised when allowing third party to identify and reveal hidden information that is private for an individual or organization. Privacy-preserving data mining using association rules is the area of data mining that protects sensitive information from unsanctioned disclosure.

As with the advancement of technology and worldwide connectivity through internet the privacy of dataset stored at different stations, whether they are stored in a centralized server for ease of access, has become important. The privacy of individual data or the dataset as whole that might be used for data mining has become so important and hence increasing the need for extensive research towards their privacy that could be done in different ways.

A company that is in lack of expertise and computational resources can take third party service provider’s help by outsourcing its mining needs. The association rules and also the items of the outsourced database are private properties of the company (data owner).The data owner modifies its data to protect corporate privacy of business transaction database and ships it to the server. Normally the dataset is in table format. Adversaries can deduce any relations or any sensitive information from that data by applying linking attacks on quasi identifiers and sensitive attributes.

Protecting sensitive information in the context of our research has two important goals: knowledge protection and privacy preservation. Knowledge protection belongs to privacy preserving association rule mining and privacy preservation refers to privacy-preserving clustering. Knowledge protection and privacy preservation have a common characteristic. For instance, in knowledge protection, an organization is the owner of the data so it must protect the sensitive knowledge discovered from such data, while in privacy preservation individuals are the owner of their personal information.

II. RELATED WORK

D. Pedreschi, S. Ruggieri in [10] works on secure mining of association rules over horizontally partitioned data. The method uses cryptographic techniques to minimize the information shared, and adding less overhead to the mining task. Privacy concerns may prevent the parties from directly sharing the data, and some types of information about the data that allows parties to choose their desired level of security needed, allowing efficient solutions that maintain the desired security.

Tzung Pei et al in [4] presented Evolutionary privacy preserving in data mining. Collection of data, dissemination and mining from large datasets are responsible for the threats to the privacy of the data. Some sensitive or private information about the individuals and businesses or organizations had to be masked before it is disclosed to users of data mining. An evolutionary privacy preserving data mining technique was proposed which helps in finding out the transactions that were to be hidden from a database. Based on the reference and sensitivity of the individuals data in the database different weights were assigned to the attributes of the individuals. To reduce the cost of rescanning whole database and to accelerate the evaluation process of chromosomes, the concept of pre large item sets was proposed. The proposed approach [4] was used to make a good tradeoff between privacy preserving and running time of the data mining algorithms.

Verykios in [3] presents a survey of association rule mining techniques for market basket analysis, and also mentions the strengths of each association rule mining technique. As well as challenging issues need to be addressed by an association rule mining technique. The results of this evaluation will help decision maker for making important decisions for association analysis.

Y-H Wu et al. [11] proposed method to reduce the side effects in sanitized database, which are produced by other approaches. They present an approach that modifies some transactions in the transaction database to reduce the supports and confidences of sensitive rules and creating no side effects.

Hajian, S., Domingo-Ferrer, J in [2] survey on classification of privacy preserving techniques is presented and major algorithms in each class. The merits and demerits of different techniques were pointed out. The algorithms for hiding sensitive association rules like privacy preserving rule mining using genetic algorithm.

Chung-Min Chen, [8] present dithered B-tree and a B-tree index structure which can be use for realizing efficient system implementations in the field of secure and private database outsourcing. The dithered tree insert algorithm [8] can be further optimized to incur only one traversal from the root to the leaf, instead of two. The index structure from learning whether or not the search term (i.e., key) is present in the database and check the data for secure and private database outsourcing.

. C.Clifton in [9] has works on data perturbation method that adds 'noise' to databases for making individuals records confidential. This technique allows users to ascertain key summary information about the data while preventing a security breach. Four biases are proposed that assess the effectiveness of such a method. These biases deal with simple aggregate methods (averages, etc.) that are in the database. A fifth bias is also proposed which may be added through perturbation techniques (Data mining Bias), and empirically test for its existence. In e-commerce fields, organizations apply data mining approaches to databases to get additional knowledge about their customers.

Pedreschi in [1] proposed a privacy Preserving mining of frequent patterns on encrypted outsourced Transaction Database (TDB). They proposed an encryption scheme and adding fake transaction in the original dataset. Their method proposed a strategy for incremental appends and dropping of old transaction batches and decrypt dataset. They also analyze the crack probability for transactions and patterns. The Encryption/Decryption (E/D) module encrypts the TDB once which is sent to the server. Mining is conducted repeatedly at the server side and decrypted every time by the E/D [1] module. Thus, we need to compare the decryption time with the time of directly executing a priori over the original database.

III. TECHNIQUES OF PRIVACY PRESERVING

Additionally, 94% of the respondents consider acquisition of their personal information by a business they do, Protection Methods Privacy can be protected through different methods such as Data Modification and Secure Multi-party Computation. The classification of privacy preserving techniques depends on the protection methods they use.

A. Data Modification technique:

Data Modification techniques transform a data set before the release of it to the users [1, 2]. The modification is done in such a way that the privacy is preserved, whereas the data quality remains high enough to serve the purpose of the release. A data modification technique could be developed that protects the privacy of individuals and sensitive underlying patterns. These techniques include noise addition, data swapping, aggregation, and suppression etc.

B. Addition in Statistical Database:

The techniques of noise addition were used for statistical databases that were expected to maintain data quality in parallel to the privacy of individuals [3]. Later on noise addition techniques were also found useful in privacy preserving data mining. The incorrectness in the statistic of a perturbed data set with respect to the statistic of the unperturbed data set is termed as bias.

C. Attribute Value Swapping:

Data swapping techniques makes values modification in the context of secure statistical databases [7]. The main appeal of the method was it keeps all original values in the data set, while at the same time makes the record re-identification very difficult. The method actually replaces the original data set by another one, where some original values belonging to a sensitive attribute are exchanged between them. This swapping can be done in a way so that the t-order statistics of the original data set are preserved.

A t-order statistic is a statistic that can be generated from exactly t attributes. A new concept called \approximate data swap" was introduced for practical data swapping. It computes the t-order frequency table from the original data set, and finds a new data set with approximately the same t-order frequency.

D. Attribute Value Suppression:

In suppression technique sensitive data values are deleted or suppressed prior to the release of a micro data. Suppression is used to protect an individual privacy from intruders' attempts to accurately predict a suppressed value [10]. An intruder can take various approaches to predict a sensitive value. For example, a classifier, built on a released data set, can be used in an attempt to predict a suppressed attribute value. Therefore, sufficient number of attribute values should be suppressed in order to protect privacy. However, suppression of attribute values results in information loss. 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 over noise addition in order to reduce the chance of having misleading patterns in the perturbed data set. Suppression has also been used for association and classification rule confusion.

E. Distributed Privacy Preserving:

The key goal in most distributed methods for privacy - preserving data mining is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants [16]. Thus, the participants may want to collaborate to get aggregate results, but may not trust each other in terms of the distribution of their own data sets. The data sets may be partitioned either horizontally or vertically.

F. Horizontally Partitioned:

In horizontally partitioned data sets, a different set of records with the same set of attributes which are used for mining purposes [16]. A horizontally partitioned case is explained, in which privacy preserving classification is done in a fully distributed setting, where each individual have private access merely to their own records . A number of other data mining applications are generalized to the horizontally partitioned data sets. Many applications of data mining can be perform i.e. clustering, filtering and association rule mining.

G. Vertically Partitioned:

The vertically partitioned [16] have a number of primitive operations like computing the scalar product or the secure set size intersection may be used to compute the results of data mining algorithms. Vertically partitioned data performs linear regressions without sharing their data values. The approach of vertically partitioned can be extended to a variety of data mining applications i.e. k means clustering, decision trees, SVM Classification and Naïve Bayes Classifier.

IV. ATTACKS ON PERTURBED DATA

K-Anonymity:

A data set T is said to satisfy K-anonymity if after dividing it into a partition, each group Gi ($1 \leq i \leq p$) in the partition contains at least K records, and T is either generalized or anatomized [13].

a) Homogeneity Attack: A and B are neighbours. One day B falls ill and is rushed to a hospital. A decides to find out what ailment B has. A finds the 4- unacknowledged table of current inpatient records, thus she knows that one of the records in this table is having B’s data [14]. As A is B’s neighbour, she knows that B is a 31-year-old American male who resides in the postal division 13053. Thus, A knows that B’s record number one among 9, 10, 11, or 12. Presently, all of the patients have the same medical condition (disease), thus A concludes that B has cancer.

b) Background Knowledge Attack: A has a pen pal U who is also in the same hospital as B, and has some patient records. A knows that U is a 21 year old Japanese female who resides in postal district 13068. Thus A knows that U’s information is present in record number 1, 2, 3, or 4. With no additional information, A is not sure if U contracted an infection or has heart disease [15]. But, it is also well-known that the Japanese have too low chances of heart disease. So A concludes that U has a viral infection.

L-diversity:

For a single sensitive attribute an equivalence class is having l-diversity if there are at least l- “well-represented” values for the sensitive attribute. A table is having l-diversity if every equivalence class of the table has l-diversity [12, 14].

a) Skewness Attack: At the point when the overall dispersion is skewed, satisfying l-diversity does not counteract characteristic disclosure.

b) Similarity Attack: At the point when the touchy attributes values in an equivalence class are distinct however semantically similar, an adversary can learn essential information.

V. CONCLUSION

Preserving privacy in data mining activities is a very important issue in many applications. Randomization-based techniques are likely to play an important role in this domain. Paper detailed various methods like perturbing, swapping, etc. for privacy preserving, where each has its own importance. Researcher's works find knowledge in dataset by Aprior and other mining algorithm then apply preserving technique on them. Hiding information at different level is also term as multi-level privacy which provides only numeric data hiding.

REFERENCES

[1] Pedreschi, D., Ruggieri, S. & Turini, F. "Discrimination-aware data mining." Proc. of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008), pp. 560-568. ACM, 2008.

[2] Hajian, S., Domingo-Ferrer, J. & Martinez-Ballesté, "Discrimination prevention in data mining for intrusion and crime detection." Proc. of the IEEE Symposium on Computational Intelligence in Cyber Security (CICS 2011), pp. 47-54. IEEE, 2011.

[3] Verykios, V. & Gkoulalas-Divanis, "A survey of association rule hiding methods for privacy." In C. C. Aggarwal and P. S. Yu (Eds.), Privacy- Preserving Data Mining: Models and Algorithms. Springer, 2008.

[4] Meij, J. "Dealing with the data flood mining data, text and multimedia" The Hague: STT Netherlands Study Centre for Technology Trends, 2002.

[5] Calders, T., & Verwer, S. " Three naive Bayes approaches for discrimination-free classification." Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[6] Sara Hajian and Josep Domingo-Ferrer "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013

[7] Pedreschi, D., Ruggieri, S. & Turini, F. " Measuring discrimination in socially-sensitive decision records." Proc. of the 9th SIAM Data Mining Conference (SDM 2009), pp. 581-592. SIAM, 2009

[8] Hajian, S. & Domingo-Ferrer, J. "A methodology for direct and indirect discrimination prevention in data mining." Manuscript, 2012.

[9] C. Clifton. "Privacy preserving data mining: How do we mine data when we aren't allowed to see it?" In Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2003), Tutorial, Washington, DC (USA), 2003.

[10] D. Pedreschi, S. Ruggieri and F. Turini, "Discrimination-aware Data Mining," Proc. 14th Conf. KDD 2008, pp. 560-568. ACM, 2008.

[11] D. Pedreschi, S. Ruggieri and F. Turini, "Measuring discrimination in socially-sensitive decision records," SDM 2009, pp. 581-592. SIAM, 2009.

[12] Jian-min, Han, Cen Ting-ting, and Yu Hui-qun. "An improved V-MDAV algorithm for l-diversity." Information Processing (ISIP), 2008 International Symposiums on. IEEE, 2008.

[13] Snehal M. Nargundi, Rashmi Phalnikar, "k-Anonymization using Multidimensional Suppression for Data De-identification" International Journal of Computer Applications (0975 – 8887) Volume 60– No.11, December 2012.

[14] Hamza, Nermin, and Hesham A. Hefny. "Attacks on anonymization-based privacy-preserving: a survey for data mining and data publishing." Journal of Information Security 4: 101, 2013.

[15] LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan. "Incognito: Efficient full-domain kanonymity." Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005.

[16] Wenfei Fan, Jianzhong Li, Nan Tang, And Wenyuan Y. "Incremental Detection Of Inconsistencies In Distributed Data". Ieee Transactions On Knowledge And Data Engineering, Vol. 26, No. 6, June 2014.