INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A SURVEY ON PRIVACY PRESERVING MINING VARIOUS TECHNIQUES WITH ATTACKS

[1]Nidhi Jain, [2]Prof.Angad Singh

[1]Information Technology, NRI Institute of Science and Technology, Bhopal, India. Nidhijain173@yahoo.co.in
[2]Head of Department Information Technology, NRI Institute of Science and Technology, Bhopal, India.
Angada2007@gmail.com

Abstract: - Privacy Preserving Data Mining (PPDM) is used to extract relevant knowledge from large amount of data and at the same time protect the sensitive information from the data miners. The enhancement of data mining research will be the development of techniques that incorporate privacy concerns. 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: Association Rule Mining, Data Perturbation, Privacy Preserving Mining.

1. Introduction

DATA mining is to extract information from large databases. Data mining is the process of discovering new patterns from large data sets which gives advantages for research, marketing analysis, medical diagnosis, atmosphere forecast etc. Data mining is under attack from privacy advocates because of a misunderstanding about what it actually is and a valid concern about how it's generally done. This has caused concerns that personal data may be used for a variety of intrusive or malicious purposes. Privacy preserving data mining help to 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 rule refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or 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 (data owner) lacking in expertise or computational resources can outsource its mining needs to a third party service provider (through server). However, both the items and the association rules of the outsourced database are considered private property of the corporation (data owner). To protect corporate privacy of business transaction database, the data owner transforms its data and ships it to the server. Normally the dataset is in table format. Adversaries can use that data for deducing any relations or any sensitive data from it by applying linking attacks on quasi identifiers and sensitive attributes.

Protecting sensitive information in the context of our research encompasses two important goals: knowledge protection and privacy preservation. The former is related to privacy preserving association rule mining, while the latter refers to privacy-preserving clustering. An interesting aspect between knowledge protection and privacy preservation is that they 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.

2. Related Work

This paper addresses [10] secure mining of association rules over horizontally partitioned data. The methods incorporate cryptographic techniques to minimize the information shared, while adding little 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 are needed, allowing efficient solutions that maintain the desired security.

Tzung Pei et al presented Evolutionary privacy preserving in data mining [4]. Collection of data, dissemination and mining from large datasets introduced 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 method was proposed to find about what transactions were to be hidden from a database. Based on the reference and sensitivity of the individual’s data in the database different weights were assigned to the attributes of the individuals. The concept of pre large item sets was used to minimize the cost of rescanning the entire database and speed up the evaluation process of chromosomes. The proposed approach [4] was used to make a good tradeoff between privacy preserving and running time of the data mining algorithms.

This authors [3] presents a survey of different association rule mining techniques for market basket analysis, highlighting strengths of different association rule mining techniques. 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 a novel approach that strategically modifies a few transactions in the transaction database to decrease the supports or confidences of sensitive rules without producing the side effects.

A classification of privacy preserving techniques is presented and major algorithms in each class is surveyed. The merits and demerits of different techniques were pointed out [2]. The algorithms for hiding sensitive association rules like privacy preserving rule mining using genetic algorithm.

Chung-Min Chen, [8] present dithered B-tree, a B-tree index structure that can serve as a building block for realizing efficient system implementations in the area 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.

In Privacy Preserving Data Mining, data perturbation is a data security technique that adds 'noise' to databases to allow individual record confidentiality. This technique [9] allows users to ascertain key summary information about the data while preventing a security breach. Four bias types have been proposed which assess the effectiveness of such a technique. However, these biases deal with simple aggregate concepts (averages, etc.) found in the database. The author propose a fifth type of bias that may be added by perturbation techniques (Data mining

Bias), and empirically test for its existence. In e-commerce applications, organizations are interested in applying data mining approaches to databases to discover additional knowledge about customers.

The author concept in this paper is Privacy Preserving mining of frequent patterns on encrypted outsourced Transaction Database (TDB) [1]. They proposed a 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.

3. 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. Privacy preserving techniques can be classified based on the protection methods used by them.

3.1 Data Modification techniques

Data Modification techniques modify a data set before releasing it to the users [1, 2]. Data is modified in such a way that the privacy is preserved in the released data set, whereas the data quality remains high enough to serve the purpose of the release. A data modification technique could be developed to protect the privacy of individuals, sensitive underlying patterns, or both. This class of techniques include noise addition, data swapping, aggregation, and suppression.

3.2 Addition in Statistical Database

Noise addition techniques were originally used for statistical databases which were supposed 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.

3.3 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.

3.4 Attribute Value Suppression

In suppression technique sensitive data values are deleted or suppressed prior to the release of a microdata. 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

3.5 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 wish to collaborate in obtaining aggregate results, but may not fully trust each other in terms of the distribution of their own data sets. For this purpose, the data sets may either be horizontally partitioned or be vertically partitioned.

3.6 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 discussed, in which privacy preserving classification is performed in a fully distributed setting, where every individual have private access to only their own record. A host of other data mining applications have been generalized to the problem of horizontally partitioned data sets. Many applications of data mining can be perform i.e. clustering, filtering and association rule mining.

3.7 Vertical partitioned

The vertically partitioned [16] have many primitive operations such as computing the scalar product or the secure set size intersection can be useful in computing the results of data mining algorithms. Vertically partitioned data to perform 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.

4. Attacks On Perturbed Data

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

4.1 Homogeneity Attack

Alice and Bob are hostile neighbors. One day Bob falls ill and is taken by ambulance to the hospital. Having seen the ambulance, Alice sets out to find what ailment Bob is suffering from. Alice finds the 4- unacknowledged table of current inpatient records published by the hospital, thus she knows that one of the records in this table contains Bob's data [14]. Since Alice is Bob's neighbor, she knows that Bob is a 31-year-old American male who lives in the postal division 13053. In this way, Alice knows that Bob's record number one among 9, 10, 11, or 12. Presently, all of those patients have the same medical condition (disease), along these lines Alice concludes that Bob has cancer.

4.2 Background Knowledge Attack

Alice has a pen companion named Umeko who is admitted to the same hospital as Bob, and who has some patient records. Alice knows that Umeko is a 21 year old Japanese female who currently lives in postal district 13068. In light of this information, Alice learns that Umeko's information is present in record number 1, 2, 3, or 4. Without additional information, Alice is not certain whether Umeko contracted an infection or has heart disease [15]. On the other hand, it is also well-known that the Japanese have a too low incidence of heart disease. Along these lines Alice concludes with close certainty that Umeko has a viral infection.

L-diversity for a single sensitive attribute) an equivalence class is said to have l-diversity if there are at least l-"well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity [12, 14].

4.2.1 Skweness Attack

At the point when the overall dispersion is skewed, satisfying l-diversity does not counteract characteristic disclosure.

4.2.2 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.

5. Conclusions

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. (2008). Discrimination-aware data mining. Proc. of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008), pp. 560-568. ACM.

[2] Hajian, S., Domingo-Ferrer, J. & Martinez-Ballesté, A. (2011a). 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.

[3] Verykios, V. & Gkoulalas-Divanis, A. (2008). 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.

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

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

[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. (2009a). Measuring discrimination in socially-sensitive decision records. Proc. of the 9th SIAM Data Mining Conference (SDM 2009), pp. 581-592. SIAM

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

[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 1367.