ISSN:0975-9646

A Survey on Privacy Preservation Used in Data Mining Techniques

Haina Mahilary

Department of Information Technology
Gauhati University Institute of Science and Technology, Assam, India

Abstract-- Increasing **network complexity, affording greater access, sharing information and a growing emphasis on the Internet have made information security and privacy a major concern for individuals and organizations. Data mining is a well-known technology for automatically and intelligently extracting knowledge from large amount of data. Such a process, however, can also disclosure sensitive information about individuals compromising the individual's right to privacy. Moreover, data mining techniques can reveal critical information about business transactions, compromising the free competition in a business setting, shopping habits, criminal records, medical history, credit records etc. Privacy preserving data mining (PPDM) is a new era of research in data mining for providing privacy of sensitive knowledge of information extracted from data mining system. Its ultimate goal is to develop efficient algorithms that allow one to extract relevant knowledge from large amount of data, while prevent sensitive information from disclosure or inference.**

***Keywords*— Data Mining, PPDM, Security.**

I. INTRODUCTION

Data mining is a non-trivial extraction of implicit, previously unknown and potentially useful information from the large amount of data. Data mining uses variety of techniques to identify important information or decision-making knowledge in the database and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting and estimation [6]. The data is often voluminous, it has low value and no direct use can be made of it. It is the hidden information in the data that is useful. As data mining is a well-known technology for automatically and intelligently extracting knowledge from large amount of data, however, it can also disclose sensitive information about individuals compromising the individual's right to privacy. Moreover, data mining techniques can reveal critical information about business transactions, compromising the free competition in a business setting. Therefore, privacy preserving data mining has becoming an increasingly important field in data mining. Privacy preserving data mining is a novel research direction in data mining. In recent years, with the rapid development in Internet, data storage and data processing technologies, privacy preserving data mining has been drawn increasing attention. In order to make a publicly available system secure, we must ensure not only the private sensitive data that have been trimmed out, but also to make sure that certain inference channels have been blocked as well. The main goal of privacy preserving data mining is to develop data mining methods that allow one to extract relevant knowledge from large amount of data, while prevent sensitive information from disclosure or inference. A number of effective methods for privacy preserving data mining have been proposed now. Most methods use some form of transformation on the original data in order to perform the privacy preservation. The transformed dataset is made available for mining and must meet privacy requirements without losing the benefit of mining. But most of these methods might result in information loss and side-effects in some extent, such as data utility-reduced, data mining efficiency-downgraded, etc. Privacy Preserving Data Mining can be found more in [8], [9], [10], [11].

II. REVIEW OF LITERATURE

Since privacy preserving data mining has become an important research area due to the distributed nature of data storage, people had started working on it since few years back. To preserve the privacy of data there are basically two approaches. The first one aims to preserve customer privacy by perturbing the data values whereas the second one is based on cryptographic tools to build various models for data mining process. Some of the recent researches done on privacy preserving data mining in [1] are like, hiding association rule by using confidence and support where the researchers suggested some rules for hiding sensitivity by changing the support and the confidence of the association rule or frequent item set as data mining mainly deals with generation of association rules, privacy preserving clustering by data transformation where preserving the privacy of individuals when data are shared for clustering was a complex problem and the challenge was how to protect the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis and this method was designed to specify privacy-preserving clustering, in context where data owners must meet privacy requirements as well as guarantee valid clustering results and perturbation based privacy preserving data mining for real world data where the perturbation method in [2] has been extensively studied for privacy preserving data mining. In this method, random noise from a known distribution is added to the privacy sensitive data before the data is sent to the miner for data mining. Consequently, the data miner rebuilds an approximation to the original data distribution from the perturbed data and uses the reconstructed distribution for data mining purposes. There are many criteria or techniques where privacy preserving data mining has been used and according to this the privacy preserving data mining has been classified.

According to [3] work done in PPDM (privacy preserving data mining) can be classified based on the following criteria.

A. Data Distribution

Based on the distribution of data PPDM algorithms can be divided into two major categories i.e., centralized and distributed. In a centralized database environment, data are all stored in a single database; while, in a distributed database environment, data are stored in different databases. Distributed data scenarios can be further classified into horizontal and vertical data distributions. Horizontal distributions refer to the cases where different records of the same data attributes are resided in different places. While in a vertical data distribution, different attributes of the same record of data are resided in different places. Earlier research has been predominately focused on dealing with privacy preservation in a centralized database. The difficulties of applying PPDM algorithms to a distributed database can be attributed to: first, the data owners have privacy concerns so they may not willing to release their own data for others; second, even if they are willing to share data, the communication cost between the sites is too expensive.

B. Purpose of Hiding

The PPDM algorithms can be further classified into two types according to the purposes of hiding i.e., data hiding and rule hiding. Data hiding refers to the cases where the sensitive data from original database like identity, name, and address that can be linked, directly or indirectly, to an individual person are hided. In contrast, in rule hiding, the sensitive knowledge (rule) derived from original database after applying data mining algorithms is removed. Majority of the PPDM algorithms used data hiding techniques. Most PPDM algorithms hide sensitive patterns by modifying data.

C. Data Mining Task/Algorithm

Currently, the PPDM algorithms are mainly used on the tasks of classification, association rule and clustering. Association rule mining is one of the most important and well researched techniques of data mining, and it was first introduced in the year of 1993. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Clustering Analysis concerns the problem of decomposing or partitioning a data set into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups. In clustering, objects are grouped together based on their similarities. Similarities between objects are defined by similarity functions, usually similarities are quantitatively specified as distance or other measures by corresponding domain experts.

D. Privacy Preservation techniques

The techniques like – sanitation, blocking, distort, and generalization -- have been used to hide data items for a centralized data distribution. The idea behind data sanitation is to remove or modify items in a database to reduce the support of some frequently used item sets such that sensitive patterns cannot be mined. The blocking approach replaces certain attributes of the data with a question mark. In this regard, the minimum support and confidence level will be altered into a minimum interval. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges, the confidentiality of data is expected to be protected. Also known as data perturbation or data randomization, data distort protects privacy for individual data records through modification of its original data, in which the original distribution of the data is reconstructed from the randomized data. These techniques aim to design distortion methods after which the true value of any individual record is difficult to ascertain, but "global" properties of the data remain largely unchanged. Generalization transforms and replaces each record value with a corresponding generalized value.

The privacy preservation technique used in a distributed database is mainly based on cryptography techniques. SMC (Secure Multi-Party Computation) algorithms deal with computing any function on any input, in a distributed network where each participant holds one of the inputs, while ensuring that no more information is revealed to a participant in the computation than can be inferred from that participant's input and output. Data distort is the most popular method used in hiding data followed by data sanitation and generalization If one wants to obtain data mining results from different data sources, then the only method can be used is a cryptography technique. Since the parties who use SMC operators cannot reveal anything from others except final results, it can have benefits of both accuracy of data mining results and the privacy of the database.

The main PPDM (privacy preserving data mining) techniques are as follows:

1. Randomization Method:

The randomization technique [2] uses data distortion methods in order to create private representations of the records. In this noise is added to the data in order to mask the attribute values of records. The noise added is sufficiently large so that the individual values of the records can no longer be recovered. In most cases, the individual records cannot be recovered, but only aggregate distributions can be recovered. These aggregate distributions can be used for data mining purposes. Data mining techniques can be developed in order to work with these aggregate distributions. Two kinds of perturbation are possible with the randomization method: Additive Perturbation and Multiplicative Perturbation. In Additive Perturbation, randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records. Data mining and management algorithms re designed to work with these data distributions. In Multiplicative Perturbation, the random projection or random rotation techniques are used in order to perturb the

records. Randomization method can be found more in [4], [7].

2. Anonymization Method:

Anonymization method aims at making the individual record be indistinguishable among a group records by using techniques of generalization, suppression and random perturbation. Perhaps the most natural way of anonymizing numerical data is to perturb it. Rather than reporting a value x for an attribute, we report the value x' = x+r, where r is a random value drawn from an appropriate (usually bias-free) distribution. One must be careful with this approach however; if the value r is chosen independently each time x is queried, then simple averaging will eliminate its effect. Since introducing bias would affect any statistical analysis one might wish to perform on the data, a preferred method is to fix the perturbations in advance. If the attribute x has a domain other than R, then perturbation is more complex. If the data is categorical however, other methods, such as deleting items and inserting other, randomly chosen items, must be employed. There are two types of perturbation i.e., input perturbation and output perturbation. Input perturbation is the process of perturbing the source data itself, and returning correct answers to queries on this perturbed data. Output perturbation on the other hand perturbs the answers sent to a query, rather than modifying the input itself. The other method for anonymizing data is generalization, which is often used in conjunction with suppression. Suppose the data domain possesses a natural hierarchical structure. For example, ZIP codes can be thought of as the leaves of a hierarchy, where 8411* is the parent of 84117, and 84* is an ancestor of 8411*, and so on. In the presence of such a hierarchy, attributes can be generalized by replacing their values with that of their (common) parent. ZIP codes of the form 84117, 84118, 84120 might all be replaced by the generic ZIP 841*. The degree of perturbation can then be measured in terms of the height of the resulting generalization above the leaf values. Data suppression, very simply, is the omission of data. For example, a set of database tuples might all have ZIP code fields of the form 84117 or 84118, with the exception of a few tuples that have a ZIP code field value of 90210. In this case, the outlier tuples can be suppressed in order to construct valid and compact generalization. Another way of performing data suppression is to replace a field with a generic identifier for that field. In the above example, the ZIP code field value of 90210 might be replaced by a null value $\perp_{ZIP}$. Anonymization method can be found more in [4]

3. Encryption Method:

The main goal in most encryption or 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. 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. In horizontally partitioned data sets, the individual records are spread out across multiple entities, each of which has the same set of attributes. Here, different sites contain different sets of records with the same (or highly overlapping) set of attributes which are used for mining purposes. In vertical partitioning, the individual entities may have different attributes (or views) of the same set of records. 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. The approach of vertically partitioned mining has been extended to a variety of data mining applications such as decision trees, SVM Classification, Naive Bayes Classifier, and k-means clustering. Vertical partitioning method can be found more in [4], [5]. The problem of distributed privacy-preserving data mining overlaps closely with a field in cryptography for determining secure multi-party computations.

III. CONCLUSIONS

Different types of techniques and approaches to preserve privacy in data mining techniques like association rule mining has been discussed here. Classification of privacy preserving data mining based on different criteria like data distribution, purpose of hiding, data mining task/algorithm, and privacy preservation technique has also been discussed here. Privacy preservation technique or method like randomization, anonymization and encryption are used to preserve privacy of the data. As now a day's data are distributed among several networks so there is a need to preserve privacy of the data so that one party cannot learn anything that is sensitive from other parties. Due to its distributed nature of the data, data must be encrypted to ensure security so encryption technique has been mostly used now a day's.

REFERENCES

[1] Privacy Preserving Data Mining. http://www.ijarcce.com/upload/2013/april/10-sneha dhawale- privacy preserving data.pdf

[2] Li Liu, Murat Kantarcioglu, and Bhavani Thuraisingham, "The applicability of the perturbation based privacy preserving data mining for real-world data," Journal of Data & Knowledge Engineering, vol. 65, pp. 5–21, 2008.

[3] K. Srinivasa Rao & B. Srinivasa Rao "An Insight in to Privacy Preserving Data Mining Methods." The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 3, July-August 2013

[4] Charu C Aggarwal and Philip S.Yu, "Privacy-Preserving Data Mining: Models and Algorithm." IBM T.J. Watson Research Center, USA and University of Illinois at Chicago, USA, IL 60607-7053 .

[5] Cynthia Dwork and Kobbi Nissim, "Privacy Preserving Data Mining on Vertically Partitioned Databases." Microsoft Research, SVC, 1065 La Avenida, Mountain View CA 94043.

[6] Arun K Pujari, "Data Mining Techniques", Third Edition, pp. 98-120, April 2013.

[7] Alexandre Evfimievski, "Randomization in Privacy Preserving Data Mining." Cornell University Ithaca, NY 14853, USA

[8] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-Preserving Data Mining." IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120.

[9] Yehuda Lindell, Department of Computer Science, Weizmann Institute of Science, Rehovot, ISRAEL, and Benny Pinkas, STAR Lab, Intertrust Technologies 4750 Patrick Henry Drive, Santa Clara CA 95054, "Privacy Preserving Data Mining."

[10] Haitao Liu and Jing Ge, "Survey on Privacy Preserving Data Mining." Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana IL, 61801, USA.

[11] Alaa H Al-Hamami, and Suhad Abu Shehab, "An Approach for preserving Privacy and Knowledge in Data Mining Applications." College of Computer Sciences and Informatics, Amman Arab University, Amman-Jordan.