A CONCEPTUAL OVERVIEW OF DATA MINING AND PRIVACY PRESERVING DATA MINING

[1]DEVENDRASINH I. VASHI

[1]Assistant Professor, Department of Computer Science and Engineering, Institute of Technology, Nirma University, Gujarat, India

devendra.vashi@gmail.com

***ABSTRACT* :**. *Data mining allow fetching crucial info from the data base. Some data are very sensitive and they should not be disclose to third party. Privacy Preserving Data Mining is the technique which is useful to safe these sensitive data and hiding sensitive information before publishing the database. Privacy Preserving Data Mining is also useful to avoid or securing the sensitive and private data set. privacy preserving data mining allows publishing a micro data without disclosing private data to a third party.*

Key Words: Data Mining, Privacy Preserving Data Mining

1. INTRODUCTION TO DATA MINING

Data mining is the process that uses a variety of data analysis tools to discover patterns and relationships, which are hidden among the vast amount of data. From these patterns and relationships, the company will be able to make valid predictions in the future trends. Data mining tools can answer business questions that traditionally were too time consuming to resolve. Data mining find the hidden patterns and predictive information that experts may miss because those information might lie outside their expectations. Data mining is also known as knowledge data discovery (KDD). Those seeking to make a distinction between the terms data mining and KDD generally use KDD to refer to the process of discovering useful knowledge from data, while using data mining to refer to the application of algorithms for extracting patterns from data.

Here are some examples for data mining:

i. Classification: Following are the examples of cases where the data analysis task is Classification :

- A bank loan officer wants to analyse the data in order to know which customer (loan applicant) are risky or which are safe.
- A marketing manager at a company needs to analyze to guess a customer with a given profile will buy a new computer.

In both of the above examples a model or classifier is constructed to predict categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.

ii. Association: Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to process of uncovering the relationship among data and determining association rules.

For example A retailer generates association rule that show that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

iii. Sequence: It is essentially a time-ordered association, although the associated events may be spread for apart in time. For example, most people will buy insurance after the marriage.

iv. Clustering: Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters. It is like classification except that the categories are not normally known beforehand. It shows the collection of shopping baskets and discovers clusters corresponding to health food buyers, convenience food buyers, luxury food buyers and so on.

1.1 Data mining process

1. Develop an understanding of the application, relevant prior knowledge, and the end user's goals.
2. Create a target data set to be used for discovery.
3. Clean and preprocess data (including handling missing data fields, noise in the data, accounting for time series, and known changes).
4. Reduce the number of variables and find invariant representations of data if possible.
5. Choose the data mining task (classification, regression, clustering, etc.).
6. Choose the data mining algorithm.

7. Search for patterns of interest (this is the actual data mining).
8. Interpret the pattern mined. If necessary, iterate through any of steps 1 through 7.
9. Consolidate knowledge discovered and prepare a report.

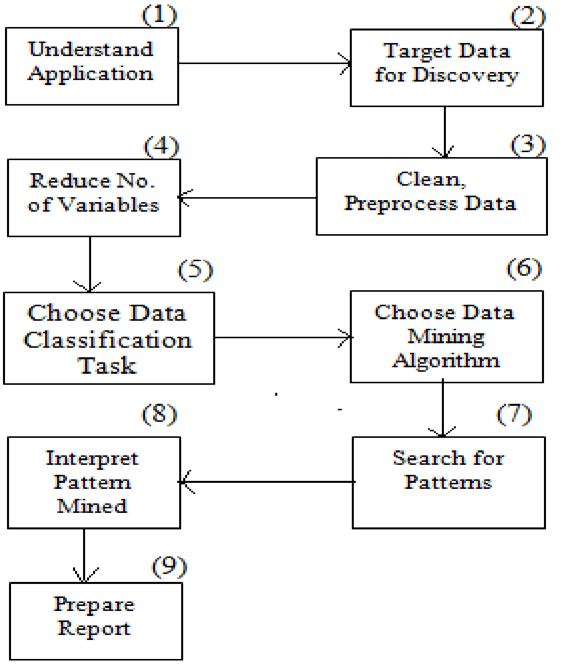

Figure 1: Steps in the Data Mining Process

1.2 Data Mining Issues:

Data mining is not that easy. The algorithm used are very complex. The data is not available at one place it needs to be integrated form the various heterogeneous data sources. These factors also creates some issues. Here in this tutorial we will discuss the major issues regarding:

i. Mining Methodology and User Interaction:

- Mining different kinds of knowledge in databases.
- Data mining query languages and ad hoc data mining.
- Handling noisy or incomplete data.
- Pattern evaluation.

ii. Performance Issues:

- Efficiency and scalability of data mining algorithms.
- Parallel, distributed, and incremental mining algorithms.

iii. Diverse Data Types Issues:

- Handling of relational and complex types of data.
- Mining information from heterogeneous databases and global information systems.

1.3 Privacy Concerns

One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences.

What if every telephone call you make, every credit card purchase you make, every flight you take, every visit to the doctor you make, every warranty card you send in, every employment application you fill out, every school record you have, your credit record, every web page you visit ... was all collected together? A lot would be known about you! This is an all-too-real possibility. Much of this kind of information is already stored in a database. Remember that phone interview you gave to a marketing company last week? Your replies went into a database. Remember that loan application you filled out? In a database. Too much information about too many people for anybody to make sense of? Not with data mining tools running on massively parallel processing computers! Would you feel comfortable about someone (or lots of someone's) having access to all this data about you? And remember, all this data does not have to reside in one physical location; as the net grows; information of this type becomes more available to more people.

2. PRIVACY PRESERVING DATA MINING

Data mining technology has emerged as a means for identifying patterns and trends from such large quantities of data. For instance, shopping centers conclude that male customers who buy diaper usually shop beers by analyzing consuming lists. This forms the relation between diaper and beer through rearranging these goods. In addition, there is also the relation between milk and bread. This improvement of goods arrangement after analysis not only makes customers convenient but increases expenditure. Credit card centers of banks find out behavior features and consuming models of high-quality clients from lots of trade data so as to seek potential clients, stimulate clients' consumption and create more opportunities of overlapping sell.

However, data mining also brings some problems. For example, credit card centers may intentionally or unconsciously make sensitive information of clients leak while mining relating information of clients. With the Internet popularity, because more and more information can be obtained in electronic form, that people have their own privacy confidential is becoming increasingly urgent. According to statistics, even if privacy protection measures, about one-fifth of Internet users don't like to provide their own information to the Web site and more than the half investigators only in good privacy-preserving measures are willing to provide their own information to the Web site. Among the potential consumers shopping in internet browser, there are almost half who gave up the hope for internet shopping because of worrying about no protection of their privacy. Therefore, how to ensure personal privacy in data mining has become a need to be addressed. It requires significant research on how to extract valuable knowledge in data and at the same time, prevent private or sensitive information in data mining process from leaking. Thus techniques of data mining without leaking the private information are needed. Research on privacy preserving data mining is developed for this purpose. Correspondingly the privacy preserving data mining and knowledge discovery should be developed aimed at these problems.

2.1 Dimensions of Privacy for Data Mining

In general, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. We refer to the former as individual privacy preservation and the latter as collective privacy preservation, which is related to corporate privacy in (Clifton et al., 2002).

- Individual privacy preservation: The primary goal of data privacy is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual.
- Collective privacy preservation: Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. We refer to the protection of sensitive knowledge as collective privacy preservation. The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to protect sensitive knowledge that can provide competitive advantage in the business world.

2.2 Privacy preserving techniques Classification

Privacy preserving techniques can be classified based on following characteristics:

1 Data Mining Scenario
2 Data Mining Tasks
3 Data Distribution
4 Data Types
5 Privacy Definition
6 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. 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 sets 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 setoff attributes but different subset of records. In Figure. 2 we shows the classification based on distribution

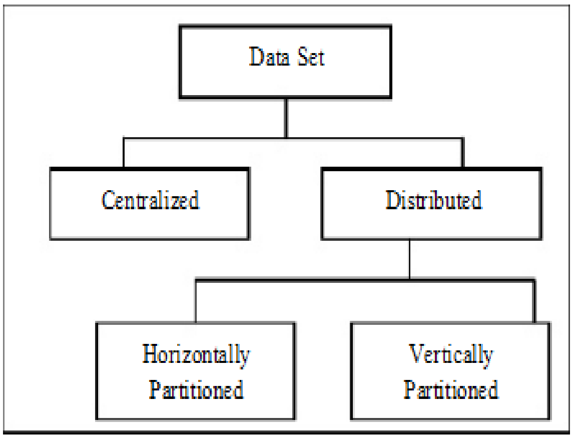

Figure 2 : Classification od different dataset

4) Data Types:

There are basically two attributes in dataset: Numerical and Categorical. Boolean data are the special case of categorical data which takes two possible values 0 and1. 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.

5) 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.

6) 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. 3

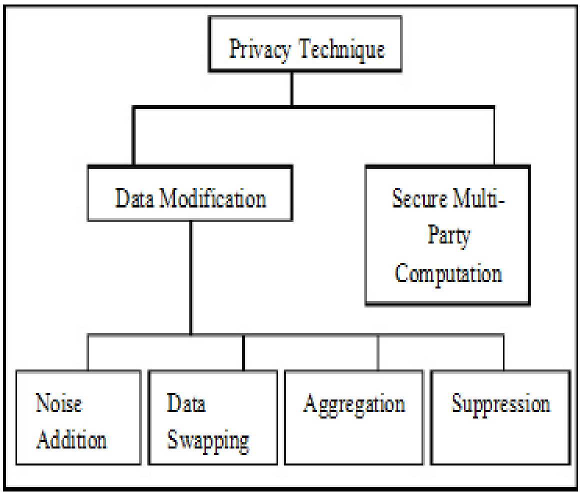

Figure 3 : A Classification of Privacy Preserving Techniques

.

3. SOME COMMONLY USED TOOLS AND ALGORITHMS FOR PRIVACY PRESERVING DATA MINING AND ITS ISSUES

3.1. Secure Multi Party Communication

Al most all PPDM techniques rely on secure multi party communication protocol. Secure multi party communication is defined as a computation protocol at the end of which no party involved knows anything else except its own inputs the results, i.e. the view of each party during the execution can be effectively simulated by the input and output of the party. In the late 1980s, work on secure multi party communication demonstrated that a wide class of functions can be computed securely under reasonable assumptions without involving a trusted third party. Secure multi party communication has generally concentrated on two models of security. The semi-honest model assumes that each party follows the rule of the protocol, but is free to later use what it sees during execution of the protocol. The malicious model assumes that parties can arbitrarily cheat and such cheating will not compromise either security or the results, i.e. the results from the malicious party will be correct or the malicious party will be detected. Most of the PPDM techniques assume an intermediate model, - preserving privacy with non-colluding parties. A malicious party may corrupt the results, but will not be able to learn the private data of other parties without colluding with another party. This is a reasonable assumption in most cases.

3.2 Secure Set Union

Secure set union methods are useful in data mining where each party needs to give rules, frequent item sets, etc without revealing the owner. This can be implemented efficiently using a commutative encryption technique. An encryption algorithm is commutative if given encryption keys, the final encryption of a data M by applying all the keys is the same for any permuted order of the keys. The main idea is that every site encrypts its set and adds it to a global set. Then every site encrypts the items it hasn't encrypted before. At the end of the iteration, the global set will contain items encrypted by every site. Since encryption technique chosen is commutative, the duplicates will encrypt to the same value and can be eliminated from the global set. Finally every site decrypts every item in the global set to get the final union of the individual sets. One addition is to permute the order of the items in the global set to prevent sites from tracking the source of an item. The only additional information each site learns in the case is the number of duplicates for each item, but they cannot find out what the item is.

3.3 Secure Size of Set Intersection

In this case, every party has their own set of items from a common domain. The problem is to securely compute the cardinality/size of the intersection of these sets. The solution to this is the same technique as the secure union using a commutative encryption algorithm. All k parties locally generate their public key-part for a commutative encryption scheme. The decryption key is never used in this protocol. Each party encrypts its items with its key and passes it along to the other parties. On receiving a set of encrypted items, a party encrypts each item and permutes the order before sending it to the next party. This is repeated until every item has been encrypted by every party. Since encryption is commutative, the resulting values from two different sets will be equal if and only if the original values were the same. At the end, we can count the number of values that are present in all of the encrypted item sets. This can be done by any party. None of the parties can find out which of the items are present in the intersection set because of the encryption.

3.4 Philosophical issues in privacy preserving data mining

1. the inherent technological conflict between the desire for privacy by World Wide Web users, and the need of Web content providers and advertisers to more fully collect and utilize data about users.
2. Imagine going to a shopping mall in which researchers follow you from store to store, taking notes on every product you examine or buy. Would you shop in such a place? Chances are, you already do. Welcome to the Internet.
3. from the perspective of the Web user who may be unaware of the degree to which identifying information can be inadvertently disclosed.
4. from the perspective of a Data Miner we consider the extent to which privacy enhancing technologies could substantially invalidate data mining results.

4. CONCLUSIONS

Due to different types of data in a variety of repositories it is very difficult to do data mining effectively and efficiently to achieve good mining results on all kinds of data and sources. this paper covers a brief explanation about the typical architecture of data mining and explained the steps of the data mining process. This paper also covers Privacy preserving techniques Classification. Publishing the data of individuals without revealing sensitive information to a third party is an important problem.

5. REFRENCES

1. M. Prakash, Dr. G. Singaravel, " A New Model for Privacy Preserving Sensitive Data Mining", ICCCNT -12, IEEE 2012
2. Hand, D., Mannila, H., & Smyth, P. (2001). "Principles of Data mining", The MIT Press: Cambridge, Massachusetts.
3. Sunita M. Mahajan and Alpa K. Reshamwala, "Data Mining Ethics in Privacy Preservation - A Survey".
4. R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," ACM SIGMOD Record, New York,1993,pp. 207-216.
5. R. Agrawal and R. Srikant, Mining sequential patterns, In Proc. Of the 11th Int'l Conference on Data Engineering, pp3-14, Taipei, Taiwan, March 1995.
6. Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data Murat Kantarcıoglu and Chris Clifton, Senior Member, IEEE
7. Assuring Privacy when Big Brother Murat Kantarcıoglu Chris Clifton
8. Privacy Preserving Association Rule Mining in Vertically Partitioned Data Jaideep Vaidya & Chris Clifton
9. Privacy Preserving Data Mining Yehuda Lindell & Benny Pinkasy
10. k-anonymity: Algorithm and Hardness, Gagan Aggarwal, Tomas Feder, Stanford University

11. Towards Standardization in Privacy Preserving Data Mining, Stanley R. M. Oliveira and Osmar R Zaiane, University of Alberta, Edmonton, Canada

12. Tools for Privacy Preserving Data Mining, Chris Clifton, Murat Kantarcioglu and Jaideep Vaidya, Purdue University.

13. Privacy Presercing Classification of Customer Data without Loss of Accuracy, Zhiqiang Yang, Sheng Zhong, Rebecca N. Wright.

14. Manish Sharma, Atul Chaudhary, Manish Mathuria and Shalini Chaudhary, "A Review Study on the Privacy Preserving Data Mining Techniques and Approaches", International Journal of Computer Science and Telecommunications

15. Alan J. Broder, "Data Mining, the Internet, and Privacy," International WEBKDD'99 Workshop San Diego, CA, USA, August 15, 1999.

16. http://www.tutorialspoint.com

17. Mary J. Cronin, "e-Privacy?," HOOVER DIGEST, No. 3, 2000. Adapted from the essay "Privacy and Electronic Commerce," by Mary J. Cronin, in the new Hoover Press book Public Policy and the Internet: Privacy, Taxes, and Contract, edited by Nicholas Imparato.