International Journal of Information Technology and Management
Vol. XI, Issue No. XVII, November-2016, ISSN 2249-4510

Case Study on Data Mining with Privacy Preservation

Mula Malyadri[1]* Dr. Arvind Kumar Sharma[2]

[1]Research Scholar

[2]Associate Professor

Abstract – The primary issue examined in this research is that privacy-preserving data mining (PPDM) research has produced theoretical solutions and many peer-reviewed articles claiming to solve the problem. In order to gain any real benefit from the theoretical solutions, practitioners must attempt to convert that theory into practical software- and hardware- based solutions. This article begins with a review of data mining, privacy, and privacy-preserving data mining. It then reviews and analyzes the barriers that prevent widespread adoption of privacy-preserving data mining solutions. The article concludes by presenting recommendations and ideas for future work. Our proposal has two main advantages. Firstly, as also suggested by our experimental results the perturbed data set maintains the same or very similar patterns as the original data set, as well as the correlations among attributes. While there are some noise addition techniques that maintain the statistical parameters of the data set, to the best of our knowledge this is the first comprehensive technique that preserves the patterns and thus removes the so called Data Mining Bias from the perturbed data set. Secondly, re-identification of the original records directly depends on the amount of noise added, and in general can be made arbitrarily hard, while still preserving the original patterns in the data set. The only exception to this is the case when an intruder knows enough about the record to learn the confidential class value by applying the classifier. However, this is always possible, even when the original record has not been used in the training data set. In other words, providing that enough noise is added, our technique makes the records from the training set as safe as any other previously unseen records of the same kind.

Keywords: Data Mining, Privacy Preservation, Solutions, Problem, Software, Hardware, Techniques, etc.

- - - - - - - - - - - - - - X - - - - - - - - - - - - - -

INTRODUCTION

The word privacy about information sharing and analysis is often indefinite and may be misleading. Traditional definitions for privacy can therefore describe in two definitions. Discussions about the concept of information privacy started in the 1960s when a number of researchers recognized the dangers of privacy violations by large collections of personal information in computer systems. Over the years a number of definitions of information privacy have emerged (Guang Li 2011).

Among those definitions, one of the definitions for information privacy is, it is the potential of an individual to manage and control the information of an individual which is published. Another widely used definition is, privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others. Sometimes privacy is confused with confidentiality and at other times with security. Privacy does involve confidentiality and security.

Basic Principles to Protect Information Privacy: As noted earlier, during the 1970s and 1980s many countries and organizations developed similar basic information privacy principles which were then enshrined in legislation by many nations. These principles are interrelated and partly overlapping and should therefore be treated together (Guang Li 2011). The OECD principles are:

1. Data Collection: One way of ensuring privacy is to limit the data while collecting from the users regarding their personal data, user can provide the data according to his/her requirements by not submitting the complete sensitive data

2. Data Quality: According to the principle 1, user can limit the personal data but it should be relevant so that, it can be effectively useful for which they are to be used and, the

data should be accurate, complete and kept up-to-date for getting mining results accurately.

3. Accountability: This is important principle and a data controller should be responsible for complying with measures and the result leads an effect to the principles stated above.

4. Purpose Specification: During data collection process, some sort of data will be collected based on individual's interest, but it should clearly represent the purpose of gathering that data.

5. Openness: In Openness we have a general policy and that is about developments, practices and policies with respect to personal information Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

6. Copyrighted Information: The data owned by a person should not be exposed, modified, or made available for any benefit or personal use as per the Principle. The only exception of using the information is when it is authorized for use by the law.

7. Security safeguards: Individual Personal data should be protected by security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data.

These privacy protection principles were developed for online transaction processing (OLTP) systems before technologies like data mining became available. In OLTP systems, the purpose of the system is quite clearly defined since the system is used for a particular operational purpose of an enterprise (e.g. student enrolment). Given a clear purpose of the system, it is then possible to adhere to the above principles.

REVIEW OF LITERATURE:

The overall generalizing is brought in to explain the fact that the theory of privacy of a person is not secured or unavailable online as in (B. Malin 2005), and it can be generalized into two main scenarios to understand how this theory works.

The first is the case of a medical database. Here, there is a need to collect and provide information on diseases all the while protecting the patient's identity from data aggregators or analyzers. Another scenario to consider is the classic "Market Basket" database. Here, transactions related to different client purchases are stored in a way that the client's purchasing can be identified. Using this data set, it is possible to extract some information about the client's behavior in the world of online shopping. This information is extracted by understanding forms of association rules, such as, "If a client buys product X, he/she will also purchase product Z with y% probability."

Depending on different factors that show how individual components in the matrix of online shopping correlate for various outcomes; it can be concluded on what trajectory a certain behavior can follow over a period of time. The first case is an example of wherein the individual's privacy has to be ensured by protecting it from unauthorized disclosure of sensitive information, which could go out in the form of specific data items that relate to specific individuals. Instead, the second scenario puts an emphasis on two primary aspects:

1. How the raw data contained in a database must be protected.

2. In certain scenarios, the high-level information that can be further derived from the available non-sensible raw data also needs to be protected so that it is not used out of context or for other purposes.

Given that such varied scenarios exist, the definition of privacy can be kept as a generalized definition (C. Cliftion 2004). Further, depending on the various considerations that have looked at, main goals of a PPDM algorithm can be defined and should be looking to enforce:

1. It should not compromise the use of non-sensitive data or access to the data.

2. A Privacy preserving data mining algorithm should be effective so that, identification of sensible information can be prevented.

3. When any protocol or approach is designed, that approach should avoid exponential memory requirements and computational complexities.

4. The algorithm should be flexible means; it should also perform better when different data mining tasks are applied not for single task.

The following are the various dimensions that can be adapted or to be followed for reaching privacy preserving data mining.

1. Data distribution

2. Data modification

3. Data mining algorithm

4. Data or rule hiding

Privacy preservation First dimension refers to data distribution; how data is located, one way is placing of data in one place called as centralized data based and second way is distributed database.

In distributed database data base can be distributed vertically or horizontally over the systems. Second dimension refers to modifying the original data to other form, so that we can prevent de-identification of sensitive data, here actual data will be modified by noising or multiplying the noise to some extent, there are several methods are there for data modification like randomization, swapping, sampling, anonymity, blocking etc.

Third dimension is Data mining algorithm, when mining is performed on data we could be able to preserve privacy of individuals. Fourth dimension refers to Hiding, some part of data or result of data mining will be taken and keep them in hidden state. Fifth dimension is the most important issue i.e., providing privacy during data mining This Thesis proposes and provides Privacy preservation during data mining based on vector quantization with codebook generation algorithms.

Classification of Data Sets: Data set is a collection of related data; it can have one or more records composing of set of attributes and its description. Data sets can be classified into two types.

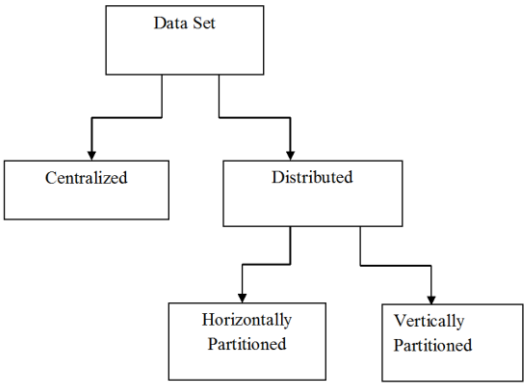

Figure 1: Types of Data Sets

Distributed Database: This is a database, which is managed under a central database management system (DBMS) in which every storage device is not automatically connected to a common CPU. The database may be stored across many computers that may or may not be located in the same physical area. The collected data—for example, in a database can also be distributed across many physical locations. The distributed database can be made to reside on different network servers, be they the ones on the Internet, or on corporate extranets or intranets, or on any other company network. By replicating and distributing the database, the performance at the end-user worksite improves.

Centralized Databases: In centralized database, data will be located and maintained at single place where as in distributed database; data may be distributed vertically or horizontally to various sources. When the database is centralized, all the data is stored in one place. This type of database is completely different from the distributed database. One of the issues the centralized database faces is that as the entire data resides at one central location, there can be problems with bottle-necks occurring at key points where the data is released or assimilated. As a result, when looking for the availability of data, the efficiency with which it is retrieved is not as strong as in the distributed database system. This thesis proposes a PPDM technique for centralized dataset.

Taxonomy of PPDM: There are many techniques proposed for maintaining privacy preserving data mining. “Privacy preserving data mining” term was first introduced by Agrawal & Srikanth, 2000, in a study they wrote on randomization techniques for centralized databases. (W. Clayton 2003) wrote on cryptographic protocols with regard to distributed databases, in which the data set is partitioned horizontally between two parties.

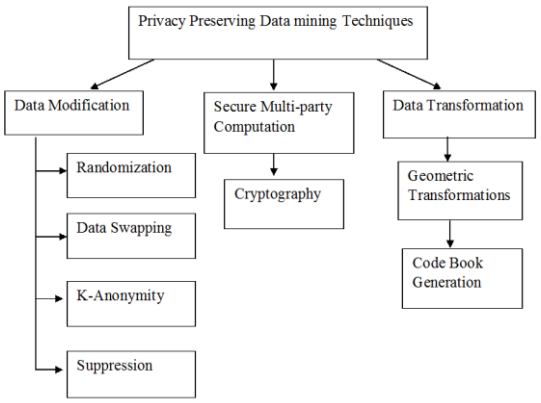

Figure 2: Taxonomy of Privacy Preserving Data Mining

(S. R. M. Oliveria 2005) brought out a protocol for secure association rule mining, k-means clustering. There were other researchers who worked on various aspects of privacy preserving data mining, such as (P. Samarati 2001) who worked on secure protocols for data that is vertically partitioned, which

was developed for mining association rules, k-means clusters, decision trees and so on. Other areas that are related to the influence and development of PPDM are cryptography, database management systems, steganography, e-commerce, secure multi-party computation, biological histories, and intrusion detection. Given that data mining can show quantitative aspects of a situation and probable ways in which the situation will develop, it has become an invaluable tool to make predictions and assess how the scenario may develop. Therefore, as an analytical tool, privacy preserving data mining's reach stretches to every domain where the components can be quantified and connections between the components noted; this is especially useful in dealing with large numbers and/or where there exist discrete behavioral components. Some domains where this form of data mining is indispensable are education systems, businesses, online world, media, the medical field and political groups.

Personalized Privacy-Preservation: Not everyone is concerned about privacy in the same way. For instance, an organization or institution considers the privacy of their employees in a different outlook. As a result, this means we would be required to treat the records for a given data set in different ways for the purposes of anonymization. In technical terms, it implies that the value of k is not fixed in anonymization. But the value can change according to the given record. In order to carry out privacy-preserving data mining when there are variable limitations on the security of the data records, a condensation- based method is proposed in (S. R. M. Oliveria 2005). When this technique is used, groups of non-homogeneous size from the data can be generated, in a way that it is guaranteed that each record lies in a group whose size is at least equal to its anonymity level. So, pseudo-data, which is obtained from each group is used to create a synthetic data store with the same aggregate distribution function as the original data store. One more paradigm of personalized anonymity is discussed in (K. C. Laudon 1996) in which an individual can identify the level of privacy for his or her "sensitive values. This approach has the advantage that it allows for direct protection of the sensitive values of persons than a vanilla k-anonymity method which may be susceptible to various kinds of attacks.

Utility-Based Privacy Preservation: It is observed that there is information loss in Privacy-preservation process for data-mining purposes. This loss of information can called as loss of "utility" for data-mining purposes. The work presented in (W. Clayton 2003) suggests that a lot of attributes may need to be suppressed to preserve anonymity; it is very important aspect to perform this carefully to preserve utility. There are several anonymization methods that use cost indicators for measuring the information loss during the anonymization process. The issue of utility-based privacy controlling data mining was reviewed in (G. Phanindra Babu 1994). The overall plan in (G. Phanindra Babu 1994) is to identify and reduce the impact of structuring by individually publishing tables which contain the utility based attributes, but it creates issues the basic need of preserving privacy. The details performed on the real tables and the marginal tables need not be same. It is now identified that this approach can improve managing and preserving the data set without negotiating on the privacy. Anonymization will differ based on the workload, for example, a workload in which some records are used more often than others would be different from one that is based on the full data set. In (B. Chandra 2009), the study proposes an effective and efficient algorithm to achieve workload-aware anonymization.

CONCLUSION:

The main focus of privacy preserving was to enhance traditional data mining techniques for masking sensitive information through data modification. The major issues were how to modify the data and how to recover the data mining result from the altered data. The reports were often tightly coupled with the data mining algorithms under consideration. Privacy preserving data publishing focuses on techniques for publishing data, not techniques for data mining. Review on Data mining privacy preserving in social network. Main objective of this review on privacy preservative technique is to protect different users and their identities in the social network along with obtaining originality. To achieve this goal, there is a need to develop perfect privacy models to specify the expected loss of privacy under different attacks, and deployed anonymization techniques to the data. So, the various techniques are surveyed.

REFERENCES:

B. Chandra, (2009). "Hybrid Clustering Algorithm", Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics.

B. Malin, (2005). "Protecting DNA Sequence Anonymity with Generalization Lattices", Methods of Information in Medicine, Vol.44, No.5, pp. 687-692.

C. Cliftion, A. Doan, A. Elmagarmid, M. Kantarcioglu, G. Schadow and D. Suciu, J. Vaidya, (2004). "Privacy- Preserving Data Integration and Sharing", DMKD'04, June 13.

G. Phanindra Babu and M. Narasimha Murthy, (1994). "Clustering with Evolution Strategies, Pattern Recognition, Vol. 27, No. 2, pp. 321 - 329.

Guang Li and Yadong Wang, (2011). "A Privacy-Preserving Data Mining Method Based on Singular Value Decomposition and Independent Component Analysis", Data Science Journal, Volume 9.

K. C. Laudon, (1996). "Markets and Privacy", Communication of the ACM.

P. Samarati, (2001). "Protecting Respondents' Identities in Micro data Release", IEEE Transactions on Knowledge and Data Engineering, Vol.13, No.6, pp. 1010-1027.

Rakesh Agrawal and RamaKrishnan Srikanth, (2000). "Privacy-Preserving Data Mining", In Proc. of ACM SIGMOD.

S. R. M. Oliveria, (2005). Data Transformation for Privacy-Preserving Data Mining, Ph. D. thesis, University of Alberta.

W. Clayton, (2003). "Ethical, Legal and Implications of Genomic Medicine", New England Journal of Medicine, Vol.349, No.6, pp. 562-569.

Corresponding Author

Mula Malyadri*

Research Scholar

E-Mail – malyadri.mtech@gmail.com