A Survey on Privacy Preserving Data Mining Techniques

Prof. Hetal M Shah
Assistant Professor
Department of Computer Engineering
Vadodara Institute of Engineering, Kotambi, India

Abstract— Data mining is a process of extracting useful knowledge and an important data from large data sets. The typical process of data collection and data dissemination result in a possible risk of privacy threats and attacks. Some private information about individuals, businesses and organizations has to be suppressed before it is shared or published. In recent years, privacy preserving data mining (PPDM) has been studied extensively, because of the wide proliferation of sensitive information on the internet. Privacy preserving data mining has become increasingly popular because it allows sharing of privacy sensitive data for analysis purposes. It is essential to maintain a ratio between privacy protection and knowledge discovery. The primary goal of this survey paper is to understand the existing privacy preserving data mining techniques and to achieve efficiency.

Key words: Data Mining, Privacy Preserving, Randomization, Suppression, K-Annoymity

I. INTRODUCTION

In recent years, data mining has been viewed as a threat to privacy because of the widespread proliferation of electronic data maintained by corporations. Data is passed through many phases during the life cycle of data management. There should be privacy is very necessary in each stage of life cycle. Data contains lots of sensitive information which are very necessary for users so there is privacy required. This has lead to increased concerns about the privacy of the underlying data.

In recent years privacy, security and data integrity are considered as challenging problem in data mining. Data mining is extensively used for knowledge discovery from large datasets. There are numbers of techniques and algorithms are available for this purpose. Privacy preserving is very necessary in secure multi party computation. Despite its benefit in a wide range of applications, data mining techniques also have raised a number of ethical issues. Some such issues include those of privacy, data security, and many others. Data mining incorporate privacy as a functional component for gain information and knowledge. Preservations of individuals information is an essential for the data owners to ensure his privacy. Privacy plays an important role in data publishing. Data mining process allows a company to use large amount of data to develop correlations and relationships among the data to improve the business efficiency. Therefore privacy preserving data mining has become important field of research. In recent years, a number of techniques have been proposed for modifying or transforming the data in such a way so as to preserve privacy. Preservations of individuals information is an essential for the data owners to ensure his privacy. Privacy plays an important role in data publishing. Data mining process allows a company to use large amount of data to develop correlations and relationships among the data to improve the business efficiency. Therefore privacy preserving data mining has become important field of research. The original data is modified by the sanitization process to conceal sensitive knowledge before release so the problem can be addressed. Privacy preservation of sensitive knowledge is addressed by several researchers in the form of association rules by suppressing the frequent item sets. As the data mining deals with generation of association rules, the change in support and confidence of the association rule for hiding sensitive rules is done. A new concept named „not altering the support" is proposed to hide an association rule. Confidentiality issues in data mining. A key problem that arises in any en masse collection of data is that of confidentiality. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. The irony is that data mining results rarely violate privacy. The objective of data mining is to generalize across populations, rather than reveal information about individuals.

II. PRIVACY PRESERVING APPROACHES

Privacy is an important concern while disclosing various categories of electronic data including business data and medical data for data mining. Especially for doing medical data mining the original data should be available for making accurate predictions otherwise lead to impractical solutions. Any kind of disclosure related to the person-specific information leads to many problems including ethical issues. Therefore extra care should be taken to protect privacy of individuals before publishing such data[3]. The privacy can be interpreted as preventing unwanted disclosure of information while performing data mining on aggregate results. Thus, privacy can be addressed at various levels in the process of data mining. For entire database security both privacy and security are required.

III. RELATED WORK

A. Randomization

Randomization technique is an inexpensive and efficient approach for privacy preserving data mining (PPDM). In order to assure the performance [12] of data mining and to preserve individual privacy, this randomization schemes need to be implemented. The randomization approach protects the customers' data by letting them arbitrarily alter their records before sharing them, taking away some true information and introducing some noise. It can deal with character type, Boolean type, classification type and number types of discrete data, and to facilitate conversion of data sets, it is necessary to preprocess the original data set. The data preprocessing is divided into discrete data, attribute coding, data sets coded data set. Some methods in randomization are numerical randomization and item set

randomization Noise can be introduced either by adding or multiplying random values to numerical records (Agrawal & Srikant, 2000) or by deleting real items and adding "fake" values to the set of attributes. The randomized data after balanced would transmit to the recipient. The recipient would receive the data using distribution reconstruction algorithm.

B. Suppression

Suppression have not involved in releasing the actual value of the data. Suppression would replace the value of specific attribute description, typically the Quasi Identifiers as the attributes, with less specific description. Like generalization, Suppression would hide some details of Quasi Identifiers A specific value could be replaced by a generic value at the time of suppression. Suppression indicates that replaced value was not disclosed.

C. Anonymization

To protect individuals' identity when releasing sensitive information, data holders often encrypt or remove explicit identifiers, such as names and unique security numbers. However, unencrypted data provides no guarantee for anonymity. In order to preserve privacy, k-anonymity model has been proposed by Sweeney [8] which achieves k-anonymity using generalization and suppression [6], In K-anonymity, it is difficult for an imposter to determine the identity of the individuals in collection of data set containing personal information. Each release of data contains every combination of values of quasi-identifiers and that is indistinctly matched to at least k-1 respondents [16]. Generalization involves replacing a value with a less specific (generalized) but semantically reliable value. For example, the age of the person could be generalized to a range such as youth, middle age and adult without specifying appropriately, so as to reduce the risk of identification. [6] Suppression involves reduce the exactness of applications and it does not liberate any information .By using this method it reduces the risk of detecting exact information. When releasing micro data for research purposes, one needs to limit disclosure risks to an acceptable level while maximizing data utility. To limit disclosure risk, Sweeney introduced the *k*-anonymity privacy requirement, which requires each record in an anonymize table to be indistinguishable with at least *k*-1 other records within the dataset, with respect to a set of quasi-identifier attributes. To achieve the *k*-anonymity requirement, they used both generalization and suppression for data anonymization.

D. Blocking Based Method

Blocking technique applies to applications where we can store unknown values for some attributes, when actual values are not available or confidential [1] .This method replaces the 1's or 0's by unknowns ("?") in selected transactions. So, that rule will not be generated from the dataset. The goal of the algorithm presented here is to obscure a given set of sensitive rule by replacing known values with unknown values. For each sensitive rule, it scans the original database and find outs the transactions supporting sensitive rules.

E. Cryptographic Technique

Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy - preserving data mining algorithms. Recent work has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. This approach is especially difficult to scale when more than a few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records.

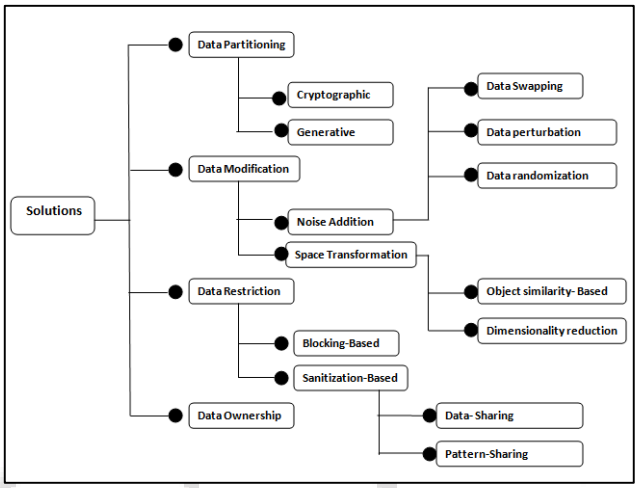

Fig. 1: Taxonomy of PPDM Techniques [4]

IV. Conclusion

In this paper we carried out some well-known algorithms concerned with privacy preserving data mining and analyses the major algorithms available for each method and points out the existing drawback. While all the purposed methods are only approximate to our goal of privacy preservation.

References

[1] Natarajan, R., et al. "A survey on Privacy Preserving Data Mining. "International Journal of Advanced Research in Computer and Communication Engineering 1.1 (2012).

[2] Lohiya,Savita, and Lata Ragha. "Privacy Preserving in Data Mining Using Hybrid Approach." Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on. IEEE, 2012.

[3] Kalita, M., D. K. Bhattacharyya, and M. Dutta. "Privacy Preserving Clustering-A Hybrid Approach." Advanced Computing and Communications, 2008. ADCOM 2008. 16th International Conference on. IEEE, 2008.

[4] Chidambaram, S., and K. G. Srinivasagan. "A combined random noise perturbation approach for multi level privacy preservation in data mining."Recent Trends in Information Technology (ICRTIT), 2014 International Conference on. IEEE, 2014.

[5] R.Hemalatha, M.Elamparithi,"Privacy Preserving Data Mining Using Sanitizing Algorithm", (IJCSIT) International Journal of Computer Science and Information Technologies, 2015.

[6] Pingshui WANG," Survey on Privacy Preserving Data ining", International Journal of Digital Content Technology and its Applications, Volume 4, Number 9, December 2010.
[7] Abeysekara, Ruvan Kumara, and Weishi Zhang. "Hybrid framework for privacy preserving data sharing." Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on. IEEE, 2013.
[8] L. Sweeney, "K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems", 10 (5), 2002.
[9] Yogendra Kumar Jain, Vinod Kumar Yadav& Geetika S. Panday, An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining, International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 7 July 2011.
[10] Alexandre Evfimievski, "Randomization in Privacy Preserving Data Mining", SIGKDD Explorations. Volume -4, Issue - 2, 2002.
[11] Jaideep Vaidya & Chris Clifton, "Privacy-Preserving Data Mining: Why, How, and When", the IEEE computer society, 2004.
[12] Yu Zhu& Lei Liu, "Optimal Randomization for Privacy Preserving Data Mining", ACM, August 2004.
[13] Aris Gkoulalas-Divanis, & Grigorios Loukides, "Revisiting Sequential Pattern Hiding to Enhance Utility", ACM, August 2011.
[14] Amruta Mhatre, Durga Toshniwal, "Hiding Co-occurring Sensitive Patterns in Progressive Databases", ACM, March 22, 2010.
[15] Shikha Sharma & Pooja Jain, "A Novel Data Mining Approach for Information Hiding", International Journal of Computers and Distributed Systems, Vol. No.1, Issue 3, October 2012.
[16] L.Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5.

