Various Aspects of Privacy Preserving Data Mining: A Comparative Study

Jayram Dwivedi
Department of IT, SIRT Bhopal, India, jayipst@yahoo.com

Abstract—Privacy-preserving data-mining works with concealing a person's privacy without losing the capacity of using the data. This approach has now become one of the very significant domain of research, though it is in its inception stage. Various approaches like secured sum algorithm, randomization, and the k-anonymity approach have been proposed to carry out the privacy-preserving data-mining. Many algorithms for data-mining, consisting of the privacy-preserving approach, have been designed which enables one to retrieve the accurate information from the huge volume of the data. It protects the private information or data from leakage or interference. In this paper we provide a review of state-of-the-art approach for the privacy and to analyze the corresponding methods for the privacy-preserving data-mining.

Keywords: Many algorithms for data-mining.

1. INTRODUCTION

Data mining involves a combination of database technology, machine learning, information retrieval, knowledge-based system, high performance computing and data visualization. Data mining applications are used in various fields including industry due to the wide availability of data and imminent need of converting data into useful information and knowledge. The information and knowledge gained could be used for various applications ranging from market analysis, fraud detection and customer retention.

Due to the widespread nature of useful data, data mining has been viewed as a threat to privacy to confidential data that are maintained by the industry. Data mining incorporates privacy as a functional component for the gained information and knowledge. Information preservation of every individual is must for data owners to ensure privacy of the organization and its employees. 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 an important field of research

Data Mining [1] refers to extracting or "mining" knowledge from large amounts of data. Data mining refers to a technique of collecting essential knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. By performing data mining, interesting knowledge, regularities, or high-level information can be extracted from database and viewed or browsed from different angles. The discovered knowledge can be applied to decision making, process control, information management and query processing. Data mining is considered one of the most important frontiers in database systems and one of the most promising interdisciplinary developments in the information industry. Data mining, with its promise to efficiently discover valuable, non-obvious information from large databases, is particularly vulnerable to misuse. So, there might be a conflict between data mining and privacy.

According to the definition, Privacy is taken as a security condition which is being restricted from other interferences [2]. On relating privacy with data mining, privacy implies to keep information about individual from being available to others [3]. Privacy is a matter of concern because it may have adverse affect on someone's life. Privacy is not violated till one feels his personal information is being used negatively. Once personal information is revealed, one cannot prevent it from being misused. Let us take an example, date of birth, mother's maiden name, or sex etc. may not become a threat for an individual, but if one more attribute like the unique identification number or voter ID are also known then it may cause a serious effect like identity theft.

Privacy preserving [4] is being emerged as an essential concern behind the success of data mining. Privacy preserving data mining (PPDM) is taken for securing the privacy of one's data or crucial information not sacrificing the use of that data. People are getting more familiar with privacy intrusions of their personal data and are unwilling to share their important information. All these features make Privacy preserving data mining a crucial aspect. Regarding the constraints of privacy, various methods are being proposed. But this is still the beginning of the research.

2. PRIVACY PRESERVING DATA MINING TECHNIQUES

In this section, we focus on the different PPDM techniques which are developed like data perturbation, blocking based technique, cryptographic techniques etc.

A. Data Perturbation

Data Perturbation [5][6] is a technique for modifying data using random process. This technique apparently distorts sensitive data values by changing them by adding, subtracting or any other mathematical formula. This technique can handle different data types: character type, Boolean type, classification type and integer. In discrete data [5], it is required to preprocess the original data set. The pre processing of data is classified into attribute coding and obtaining sets coded data set. The method of average region to disperse the continuous data is used here. Discrete formula given by Sativa Lohiya and Lata Ragha [7] is: A (max) - A (min)/n =length. A is continuous attribute, n is number of discrete, and length is the length of the discrete interval. The technique does not reconstruct the original data values, it only reconstructs the distribution.

B. Blocking based technique

In blocking based technique [7][8], authors state that there is a sensitive classification rule which is used for hiding sensitive

Published By:
International Journal of Engineering Research in Current Trends

data from others.

In this technique, there are two steps which are used for preserving privacy. First is to identify transactions of sensitive rule and second is to replace the known values to the unknown values. In this technique, there is scanning of original database and identifying the transactions supporting sensitive rule. And then for each transaction, algorithm replaces the sensitive data with unknown values. This technique is applicable to those applications in which one can save unknown values for some attributes. Authors in [7] want to hide the actual values, they replace '1' by '0' or '0' by '1' or with any unknown values in a specific transaction. The replacement of these values does not depend on any specific rule. The main aim of this technique is to preserve the sensitive data from unauthorized access.

C. Cryptographic Technique

Cryptography is a technique through which sensitive data can be encrypted. It is a good technique to preserve the data. In[9] , authors introduced cryptographic technique which is very popular because it provides security and safety of sensitive attributes . There are different algorithms of cryptography available .But this method has many disadvantages. It fails to protect the output of computation. It prevents privacy leakage of computation. This algorithm does not give fruitful results when it talks about more parties. It is very difficult to apply this algorithm for huge databases. Final data mining result may break the privacy of individual's record.

D. Condensation Approach

Another approach used is Condensation approach .It was introduced by Charu C. Aggarwal and Philip [10] which builds constrained clusters in the data set and after that produces pseudo-data. The basic concept of the method is to contract or condense the data into multiple groups of predefined size. For each group, certain statistics are maintained. This approach is used in dynamic data update such as stream problems. Each group has a size of at least 'k', which is referred to as the level of that privacy-preserving approach. The higher the level, the high is the amount of privacy. They use the statistics from each group in order to generate the corresponding pseudo-data. This is a simple privacy preservation approach but it is not efficient because it leads to loss of the information.

E. Hybrid technique

Privacy preservation is a very huge field. Many algorithms have been proposed in order to secure the data. Hybrid technique is a new technique through which one can combine two or more techniques to preserve the data. Sativa Lohiya and Lata Ragha [7] proposed a hybrid technique in which they used randomization and generalization. In this approach first they randomize the data and then generalized the modified or randomized data. This technique protects personal data in an efficient manner; sometimes its recover the actual data and offer data in absence of loss of information. Many other techniques can also be combined to make a hybrid technique such as Data perturbation, Blocking based method, Cryptographic technique, Condensation approach etc.

3. APPLICATION

3.1 Medical Database: Scrub System

As per the author perceptions [11], clinical information would be in the form of text, which contains information of patients like his family members, address, blood group and phone number. Traditional techniques have been used only for global search and replace procedure inorder to maintain privacy [11].

3.2 Bioterrorism Application

It is essential to analyze the medical data for privacy preservation. For example, Biological agents are widely found in the natural environment such as anthrax. It is important to find the anthrax attack from the normal attack [12].It is necessary to track incidences of the common diseases. The corresponding data would be reported to the public health agencies. The respiratory diseases were not reportable-diseases. This provides a solution for more identifiable information in accordance with public health law [12].

Following are the general examples

1. First advantage of proposed algorithm is that support for the sensitive item is unchanged. Instead, only the position of the sensitive item set is changed.
2. The second advantage is the proposed approach uses a different approach for modifying the database transactions so that the confidence of the sensitive rules can be reduced but without changing the support of the sensitive item.
3. Simple and efficient technique for building data mining models from perturbed data.
4. As the distribution of the added noise is known, the data miner could rebuild the original distribution usingvarious statistical methods and mine the rebuilt data.
5. It provides a very high security of database as well as it keeps the utility and assurance of mined rules at highest level.
6. PPDM is very advantageous in development of various data mining techniques.
7. It allows sharing of large amount of privacy sensitive data for analysis purposes.
8. It has a ability to track and collect large amounts of data with the use of current hardware technology.

4. LITERATURE REVIEW

There is some good literature on secrecy in statistical databases. An excellent Survey of work prior to the late 1980's was done by Adam and Wortmann. With the help of taxonomy, this work goes down in category of output perturbation. The work [13] is responsible to finish the opportunities for privacy. This is done naturally due to the reason that it exploits databases and it's original number of crises.

Fanconi and Merol did a survey recently, with concentration on aggregated data released through web access [14]. Gehrke, Evfimievski and Srikant in [15], give a better discussion of work in randomization of data, in Which data contributors (e.g., respondents to a survey) separately add noise with their personal responses. A special issue (Vol.14, No. 4, 1998) of the Journal of Official Statistics is dedicated to disclosure control in statistical data. A discussion of some of the trends in the statistical research, accessible to the non-statistician, can be found.

5. CONCLUSION

Data Mining is very useful technique provided one can preserve the privacy of the client from any other party. While performing any privacy preservation technique, we need to understand various techniques for this. In this paper we also talked about the various application of the privacy preservation techniques along with literature review. Hence, major demands of any privacy-preserving data mining algorithm are privacy preservation of data and Data utility.

REFERENCES

[1] J. Han and M. Kamber , "Data Mining: Concepts and Techniques", 2nd ed.,The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor 2006.
[2] The free dictionary.Homepage on Privacy [Online]. Available: http://www.thefreedictionary.com/privacy.
[3] M. Prakash, G. Singaravel, "A New Model for Privacy Preserving Sensitive Data Mining", in proceedings of ICCCNT Coimbatore, India, IEEE 2012.
[4] M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.
[5] J. Liu, J. Luo and J. Z. Huang, "Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity requirements", in proceedings of 11th IEEE International Conference on Data Mining Workshops, IEEE 2011.
[6] H. Kargupta and S. Datta, Q. Wang and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", in proceedings of the Third IEEE International Conference on Data Mining, IEEE 2003.
[7] S. Lohiya and L. Ragha, "Privacy Preserving in Data Mining Using Hybrid Approach", in proceedings of 2012 Fourth International Conference on Computational Intelligence andCommunication Networks, IEEE 2012.
[8] A. Parmar, U. P. Rao, D. R. Patel, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database" , in proceedings of International Symposium on Computer Science and Society, IEEE 2011.
[9] Y. Lindell, B.Pinkas, "Privacy preserving data mining", in proceedings of Journal of Cryptology, 5(3), 2000.
[10] C. Aggarwal , P.S. Yu, "A condensation approach to privacy preserving data mining", in proceedings of International Conference on Extending Database Technology (EDBT), pp.183–199, 2004. 746
[11] Sweeney L, (1996), "Replacing Personally Identifiable Information in Medical Records, the Scrub System". Journal of the American Medical Informatics Association.
[12] CharuC.Aggarwal, "A General survey of privacy preserving Data Mining Models and Algorithms", IBM,T. J. Watson Research Centre
[13] A. V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke.Privacy preserving mining of association rules. In Proceedings of the Eighth ACM SIGKDD International Conferenceon Knowledge Discovery and Data Mining, pages 217–228, 2002
[14]. L. Franconi and G. Merola, Implementing Statistical Disclosure Control for Aggregated Data Released Via Remote Access, Working Paper No. 30, United Nations Statistical Commission and European Commission, joint ECE/EUROSTAT work session on statistical data confidentiality, April, 2003, available at http://www.unece.org/stats/documents/2003/04/confidentiality/wp.30.e.pdf
[15]. A. V. Evfimievski, J. Gehrke and R. Srikant, Limiting privacy breaches in privacy preserving data mining, Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 211-222, 2003.