ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

A Literature analysis on Privacy Preserving Data Mining

Tamanna Kachwala, Dr. L. K. Sharma

Assistant Professor, Dept. of BCA, DNICA, S.P. University, Anand, India.

Scientist 'B', National Institute of Occupational Health, Ahemdabad, India.

ABSTRACT: Privacy Preserving Data Mining (PPDM) is a research area concerned with the privacy driven from personally identifiable information when considered for data mining. Therefore, PPDM has become an increasingly important field of research. PPDM is a novel research direction in data mining. A number of methods and techniques have been developed for privacy preserving data mining. This paper provides a complete review on PPDM and different techniques such as data partition, data modification, data restriction technique which could be used to prevent the data access from unauthorized users. Privacy preserving data mining has become increasingly popular because it allows sharing of privacy Sensitive data for analysis purposes. Several data mining algorithms, incorporating privacy preserving mechanisms, have been developed that allow one to extract relevant knowledge from large amount of data, while hide sensitive data or information from disclosure or inference. We provide a review of the state-of-the-art methods for privacy and analyze the representative technique for privacy preserving data mining.

KEYWORDS: *Data* Mining, Privacy Preserving, Knowledge, Protection, data modification, data restriction

I. INTRODUCTION

Data mining aims to take out useful information from multiple sources, whereas privacy preservation in data mining aims to preserve these data against disclosure or loss. Privacy preserving data mining (PPDM) [1,2] is a new research direction in data mining and statistical databases [3], where data mining algorithms are analyzed for the side effects they acquire in data privacy. The main consideration of the privacy preserving data mining is twofold. First, sensitive raw data like identifiers, name, addresses and the like should be modified out from the original database, in order for the recipient of the data not to be able to cooperation another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well cooperation data privacy. The main purpose of privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and the private knowledge stay private even after the mining process. In this paper, we provide a classification and description of the various techniques and methodologies. Agarwal and Srikant [3] and Lindell and Pinkas [4] introduced the first Privacy -preserving data mining algorithms which allow parties to collaborate in the extraction of knowledge, without any party having to reveal individual items or data. The goal of this paper is to give a review of the different dimension and classification of privacy preservation techniques used in privacy preserving data mining. Also aim is to give different data mining algorithms used in PPDM and related research in this field.

II. PRIVACY PRESERVING TECHNIQUES

The main objective of privacy preserving data mining is to develop data mining methods without increasing the risk of mishandling [5] of the data used to generate those methods. Most of the techniques use some form of alteration on the original data in order to attain the privacy preservation. The altered dataset is obtainable for mining and must meet privacy requirements without losing the [5] benefit of mining.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

A.Randomization

Randomization technique is an inexpensive and efficient approach for privacy preserving data mining (PPDM). In order to assure the performance [7] 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. 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.

B. 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 [5] which achieves k-anonymity using generalization and suppression [5], In K-anonymity, it is difficult for an imposter to decide 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. 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. [5] 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.

C. Secure multi-party computation

A substitute approach based on the multiparty computation is that every part of private data is validly known to one or more parties. Revealing private data to parties such as by whom the data is owned or the individual to whom the data refers to is not a condition of violating privacy. The problem arises when the private information is revealed to some other third parties. To deal with this problem, we use a specialized form of privacy preserving distributed data mining. Parties that each knows some of the private data contribute in a protocol that generates the data mining results, [8] that guarantees no data items is revealed to other parties. Thus the process of data mining doesn't cause, or even increase the chance for breach of privacy.

D. Sequential pattern hiding

Sequential pattern hiding method is necessary to conceal sensitive patterns that can otherwise be extracted from published data, without critically affecting the data and the non sensitive interesting patterns. [9]Sequential pattern hiding is a challenging problem, because sequences have more composite semantics than item sets, and calls for efficient solutions that offer high utility.

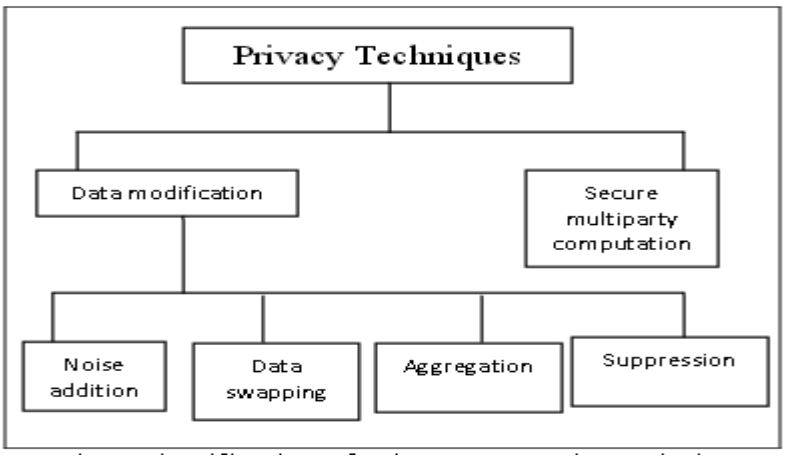

Fig.1 Classification of Privacy preserving techniques

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

III.PRIVACY PRESERVING DATA MINING ALGORITHM

The followings are some of the data mining algorithms that have been used for privacy preservation.

1.Association Rule Mining

The association rule mining problem can be formally stated as follows [10]: Let I = {i1, i2, ---, im} be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that T $\subseteq$I. Associated with each transaction is a unique identifier, called its TID. We say that a transaction T contains X, a set of some items in I, if X $\subseteq$T. An association rule is an implication of the form, X $\Rightarrow$Y, where X $\subset$I, Y$\subset$I, and X $\cap$ Y = φ. The rule X $\Rightarrow$Y holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y . The rule X $\Rightarrow$Y has support s in the transaction set D if s% of transactions in D contains X U Y. To find out if a particular item set is frequent, it counts the number of records where the values for all the attributes in the item set are 1.

2. Clustering

Clustering [11] is a data mining method that has not taken its real part in the works already quoted although, the most important algorithm of this method was very studied in the context of privacy preserving, which is k -means algorithm [12]. Surveying privacy preserving k -means clustering approaches apart from other privacy preserving data mining ones is important due to the use of this algorithm in important other areas, like image and signal processing where the problem of security is strongly posed [13]. Most of works in privacy preserving clustering are developed on the k-means algorithm by applying the model of secure multi-party computation on different distributions (vertically, horizontally and arbitrary partitioned data). Among the formulations of Partition clustering based on the minimization of an objective function, k-means algorithm is the most widely used and studied. Given a dataset D of n entities (objects, data points, items,…) in real p-dimension space R^p and an integer k. The k -means clustering algorithm partitions the dataset D of entities into k-disjoint subsets, called clusters. Each cluster is represented by its center which is the central id of all entities in that subset. The need to preserve privacy in k-means algorithm occurs when it is applied on distributed data over several sites, so called “parties” and that it wishes to do clustering on the union of their datasets. The aim is to prevent a party to see or deduce the data of another party during the execution of the algorithm. This is achieved by using secure multi-party computation that provides a formal model to preserve privacy of data.

3. Classification Data Mining

Classification is one of the most common applications found in the real world. The goal of classification is to build a model which can predict the value of one variable, based on the values of the other variables. For example, based on financial, criminal and travel data, one may want to classify passengers as security risks. In the financial sector, categorizing the credit risk of customers, as well as detecting fraudulent transactions is classification problems. Decision tree classification is one of the best known solution approaches. The decision tree in ID3 [14] is built top-down in a recursive fashion. In the first iteration it finds the attribute which best classifies the data considering the target class attribute. Once the attribute is identified in the given set of attributes algorithm creates a branch for each value. This process is continued until all the attributes are considered. In order to calculate which attribute is the best to classify the data set information gain is used. Information gain is defined as the expected reduction in entropy. Another most actively developed methodology in data mining is the Support Vector Machine (SVM) classification [15]. SVM has proven to be effective in many real-world applications [16]. Like other classifiers, the accuracy of an SVM classifier crucially depends on having access to the correct set of data. Data collected from different sites is useful in most cases, since it provides a better estimation of the population than the data collected at a single site.

4. Bayesian Data Mining

Bayesian networks are a powerful data mining tool. A Bayesian network consists of two parts: the network structure and the network parameters. Bayesian networks can be used for many tasks, such as hypothesis testing and automated

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

scientific discovery. A Bayesian network (BN) is a graphical model that encodes probabilistic relationships among variables of interest [17]. Formally, a Bayesian network for a set V of m variables is a pair (Bs, Bp). The network structure Bs = (V, E) is a directed acyclic graph whose nodes are the set of variables. The parameters Bp describe local probability distributions associated with each variable. The graph Bs represents conditional independence assertions about variables in V: An edge between two nodes denotes direct probabilistic relationships between the corresponding variables. Together, Bs and Bp define the joint probability distribution for V.

IV.RECENT TECHNOLOGY

GENETIC ALGORITHM

In Genetic Algorithms, a population consists of a group of individuals called chromosomes that represent a complete solution to a distinct dilemma. Every chromosome represents a sequence of 0s or 1s. The first set of the population is set of individuals that are randomly generated. There are two methods to generate new population: steady state Genetic Algorithm and generational Genetic Algorithm. The steady-state Genetic Algorithm replaces one or two members of the population; whereas the generational Genetic Algorithm replaces all of them at each generation of evolution. In this work a generational Genetic Algorithm is adopted as population replacement method. In this method tried to keep a certain number of the best individuals from each generation and copies them to the new generation.

Advantages

- It provides a very high security of database as well as it keeps the utility and declaration of mined rules at highest level. Here a new multi-objective method for hiding sensitive association rules based on the concept of genetic algorithms is used.

Disadvantages

- One big issue is the risk of information leakage and its confidence.
- It emphasizes on alteration of original data in such a way that it would be impractical for the opponent to mine the sensitive rules from the modified data set.

V.SUGGESTIONS FOR FUTURE WORK

There are many future research directions for privacy preserving data mining. First, present studies tend to use different terminology to describe similar or related practice. For instance, people used data modification, data perturbation, data sanitation, data hiding, and preprocessing as possible methods for preserving privacy; however, all are in fact related to the use of some types of technique to modify original data so that private data and knowledge remain private even after the mining process. Lacking a common language for discussions will cause misunderstanding and slow down the research breakthrough. Therefore, there is an emerging need of standardizing the terminology and PPDM practice. Second, most prior PPDM algorithms were developed for use with data stored in a centralized database. However, in today's global digital environment, data is often stored in different sites. With recent advances in information and communication technologies, the distributed PPDM methodology may have a wider application, especially in medical, health care, banking, military and supply chain scenarios. Third, data hiding techniques have been the dominated methods for protecting privacy of individual information. However, those algorithms do not pay full attention to data mining results, which may lead to sensitive rules leakages. While some algorithms are designed for preserving the rule such as with sensitive information, it may degrade the accuracy of other non-sensitive rules. Thus, further investigation, focusing on combining data and rule hiding, may be beneficial, specifically, when taking into account the interactive impact of sensitive and non-sensitive rules. Fourth, although many machine learning methods have been used for classification, clustering, and other data mining tasks (e.g., diagnose, prediction, optimization), currently only the association rules method has been predominately used for classification. It would be interesting to see how to extend the current technique and practice into other problem domains or data mining tasks. Furthermore, it is important to find the privacy preserving technique that is independent of data mining task so that after applying privacy preserving technique a database can be released without being constrained to the original task. Finally, identifying suitable evaluation criteria and developing benchmarks for algorithm selection are two important aspects in PPDM research. A

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

framework for evaluating selected association rule hiding algorithms has been proposed by Bertino [18]. Future research can consider testing the proposed evaluation framework for other privacy preservation algorithms, such as data distortion or ccryptography methods.

VI.CONCLUSIONS

PPDM has recently emerged as a new field of study. As a new comer, PPDM may offer a wide application prospect but at the same time it also brings us many issues / problems to be answered. In this study, we conduct a comprehensive survey on 29 prior studies to find out the current status of PPDM development. We propose a generic PPDM framework and a simplified taxonomy to help understand the problem and explore possible research issues. We also examine the strengths and weaknesses of different privacy preserving techniques and summarize general principles from early research to guide the selection of PPDM algorithms. As part of future work, we plan to apply the proposed evaluation framework to formally test a complete spectrum of PPDM algorithms.

REFERENCES

[1] Chris Clifton and Donald Marks, "Security and privacy implications of data mining", In Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (1996), 15–19.
[2] Daniel E. O'Leary, "Knowledge Discovery as a Threat to Database Security " , In Proceedings of the 1st International Conference on Knowledge Discovery and Databases (1991), 107–516.
[3] R. Agrawal and R. Srikant, "Privacy-preserving data mining", In ACM SIGMOD, pages 439–450, May 2000 Y. Lindell and B. Pinkas,"Privacy preserving data mining", J. Cryptology, 15(3):177–206, 2002.
[4] Y. Lindell and B. Pinkas,"Privacy preserving data mining " , J. Cryptology, 15(3):177–206, 2002.
[5] Pingshui WANG," Survey on Privacy Preserving Data Mining", International Journal of Digital Content Technology and its Applications, Volume 4, Number 9, December 2010.
[6] L. Sweeney, " K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems", 10(5), 2002.
[7] Yu Zhu& Lei Liu, "Optimal Randomization for Privacy Preserving Data Mining", ACM, August 2004.
[8] Jaideep Vaidya & Chris Clifton, " Privacy-Preserving Data Mining:Why, How, and When",the IEEE computer society, 2004.
[9] Aris Gkoulalas-Divanis, & Grigorios Loukides, " Revisiting Sequential Pattern Hiding to Enhance Utility", ACM, August 2011.
[10] R. Agrawal, T. Imielinski, and A. N. Swami. "Mining association rules between sets of items in large database"s. In P. Buneman and S. Jajodia editors, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pages 207{216,Washington, D.C., May26-28, 1993.
[11] Jain A., Murty M., and Flynn P." Data Clustering: A Review", ACM Computing Surveys , vol. 31, no. 3, pp. 264 -323, 1999.
[12] MacQueen J., "Some Methods for Classification and Analysis of Multivariate Observations," in Proceedings of 5th Berkley Symposium Math. Statistics and Probability, California, USA, pp. 281 - 296, 1967.
[13] Erkin Z., Piva A., Katzenbeisser S., Lagendijk R., Shokrollahi J., Neven G., and Barni M., "Protection and Retrieval of Encrypted Multimedia Content: When Cryptography meets Signal Processing," EURASIP Journal of Information Security, vol. 7, no. 17, pp. 1 - 20, 2007.
[14] Lindell Y. , Pinkas B., "Privacy Preserving Data mining", In ternational Journal of Cryptology, Citesheer, 2000.
[15] Yu H., Vaidya J., Jiang X.: "Privacy - Preserving SVM Classification on Vertically Partitioned Data", PAKDD Conference, 2006.
[16] V. N. Vapnik, " Statistical Learning Theory" , John Wiley and Sons, 1998.
[17] G. Cooper and E. Herskovits, " A Bayesian Method for the Induction of Probabilistic Networks from Data," Machine Learning, vol. 9, no. 4, pp. 309-347, 1992.
[18] E. Bertino, I. Fovino & L. Provenza (2005), "A Framework for Evaluating Privacy Preserving Data Mining Algorithms", Data Mining and Knowledge Discovery, Vol. 11,No. 2, Pp. 121–154.