International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 09 | Sep -2017 www.irjet.net p-ISSN: 2395-0072

A Survey on Features and Techniques Description for Privacy of Sensitive Information

Archana Singh[1], Varun Singh[2]

[1]*M.Tech. Scholar, Computer Science Dept. RIT College Rewa, M.P.*
[2]*Asst. Prof. Computer Science Dept. RIT College Rewa, M.P.*

Abstract - User privacy protection is of the utmost importance. Privacy has been acknowledged as a human right that is beneficial not only to each individual but also to society as a whole. So Privacy protection has become a necessary requirement in many data mining applications due to emerging privacy legislation and regulations. A detailed survey of various researcher work is summarized in this paper. Different techniques of privacy preserving mining are explained with their restriction on various systems of privacy preserving. This paper clarify distinctive threats of data mining for the examination of the preserved dataset.

Key Words: *Privacy Preserving Mining, Association Rule Mining, Data Perturbation, Threats*

1.INTRODUCTION

Every organization gather facts about their clients or users for exploration or any other intent. Information being collected may be audio, videos, images and text etc fig. 1. Since privacy concerns related to a possible misuse of knowledge discovered by means of data mining techniques have been raised [3], many attempts have been made to provide privacy preserving techniques for data mining [12, 7, 98]. Thus, a new (sub)domain of data mining, privacy preserving data mining, emerged in the last decade. In order to provide sufficient privacy protection in data mining, several methods for incorporating privacy have been developed. Privacy itself is not an easy term to define and can be preserved on different levels in different scenarios [8, 1]. In spite of enormous diversity in privacy aspects of data mining, three main approaches can be distinguished: heuristic-based, reconstruction-based and cryptography-based [11].

In the first approach, the heuristic algorithms are used to hide knowledge an organization does not want to reveal, for instance, individual values in data are changed according to a heuristic algorithm to hide sensitive knowledge such as important rules in the case of association rules mining.

The reconstruction-based approach is used to incorporate privacy on an individual level by changing original individual values (for instance, users' answers) in a random way by means of a randomization-based method and revealing only modified values.

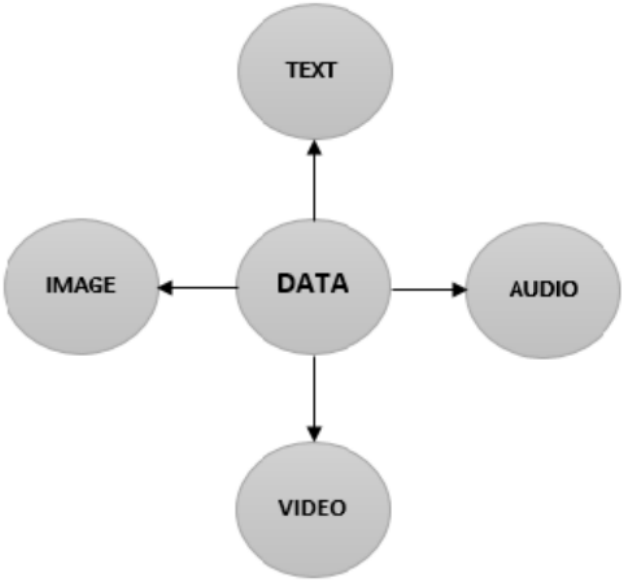

Fig. 1. Types of data in the datasets.

The distorted data as well as parameters of a randomization-based method used to distort them can be published or passed to a third party. Knowing distorted individual values and parameters of a randomization-based method, one is able to perform data mining tasks. To this end, first original distributions of values of attributes are reconstructed (estimated) based on the distorted values and the parameters of the distortion method, and a data mining model is built based on the distorted data. The creation of a model is carried out without the need to access original individual data. The third approach, which is based on cryptography, uses secure multiparty computations (SMC) to carry out data mining tasks based on distributed data, that is, data possessed by different organizations that do not want to disclose their private input. Furthermore, encryption techniques which enable one to perform computations over encrypted data without being able to decrypt can be used in privacy preserving. The heuristic approach is designed for centralized data. The cryptography-based approach is used for the distributed data, while the reconstruction-based approach can be applied to both distributed and centralized data.

2 Privacy Preserving Technique

Data Swapping

In this techniques is data maintains as a order basically data e evolve as a textual form, text data perturbation as a textual data form .textual data means addition new values and may not possible in all cases of textual datasets. so swapping technology is better option for the same In which most frequent values are observed and replace with the least or lesser frequent values so that original values or decision cannot be taken from the perturbed copy of the dataset.

In some case if the replacement of the single item is done for the most frequent item then detection of that hide technique can be easily breakable. So it is necessary to choose the item from a set randomly for replacing the frequent one.

Suppression

In some data set have some information ,that information is directly identify by the individuals person or individual class then those has to remove from the data set. So columns or items are delete from the original data set ,the is such types of sensitive data set, Suppression is used for protecting for information ,As Example: We have data set contain a driving license number, the only one person can detectable and we cannot add or delete in driving license. as format of that driving license number is define. So such data is removed from the original dataset.

Noise Addition

In this approach data set change as a change in a numeric value where amount is change is a sequence of random value, that value reflected as a original values but not in original data set order. In [5] noise is generate by a Gaussian function that create number as a sequence form then add there sequence in the original value. so a kind of variation is develop over here for the privacy of the original one. While data can add a single value but it can be detect easily or observed also if intruder will present in data set.

There are different numeric category involving as : involving percentiles, sums, conditional means etc. Some noise addition techniques, Random Perturbation Technique, Probabilistic Perturbation Technique , etc.

Data Perturbation

In data Perturbation on data set is transformed in to perturbation and selecting random position data then add, subtraction the value into the original in order produce new value that is differ from the previous data. One is important information is here whatever you want add or subtraction delete from that value should not cross the limits of the original lets understand an age value is perturbed by adding or subtracting from original data but the resultant value or the perturbed value should not be less then 0 or greater then a normal life of 120. In order to perform perturbation some kinds of random value that by original value change randomly. There are generate two approaches.

First is probability distribution approach and Second is Value distortion approach

- probability distribution approach **:-**The approach of probability distribution, In this approach data replace with another sample from the same (estimated) distribution or by the distribution itself.
- Value distortion approach:- The approach of Value distribution, perturbed the value of data and elements or directly by adding or multiplicative some noise before releasing of the data.

3 Related Work

This paper addresses [10] secure mining of association rules over horizontally partitioned data. The methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task. Privacy concerns may prevent the parties from directly sharing the data, and some types of information about the data. That allow parties to choose their desired level of security are needed, allowing efficient solutions that maintain the desired security.

Tzung Pei et al presented Evolutionary privacy preserving in data mining [4]. Collection of data, dissemination and mining from large datasets introduced threats to the privacy of the data. Some sensitive or private information about the individuals and businesses or organizations had to be masked before it is disclosed to users of data mining. An evolutionary privacy preserving data mining method was proposed to find about what transactions were to be hidden from a database. Based on the reference and sensitivity of the individuals data in the database different weights were assigned to the attributes of the individuals. The concept of pre large item sets was used to minimize the cost of rescanning the entire database and speed up the evaluation process of chromosomes. The proposed approach [4] was used to make a good tradeoff between privacy preserving and running time of the data mining algorithms.

This authors [3] presents a survey of different association rule mining techniques for market basket analysis, highlighting strengths of different association rule mining techniques. As well as challenging issues need to be addressed by an association rule mining technique. The results of this evaluation will help decision maker for making important decisions for association analysis.

Y-H Wu et al. [11] proposed technique to decrease the reactions in sterilized database, which are delivered by

different methodologies. They exhibit a novel approach that deliberately alters a couple of exchanges in the exchange database to diminish the backings or confidences of touchy guidelines without creating the reactions.

In [12] In this paper, a novel ef_cient anonymization system called PTA is proposed to not only anonymize transactional data with a small information loss but also to reduce the computational complexity of the anonymization process. The PTA system consists of three modules, which are the Pre-processing module, the TSP module, and the Anonymity model, to anonymize transactional data and guarantees that at least *k*-anonymity is achieved: a pre-processing module, a traveling salesman problem module, and an anonymization module.

A characterization of security protecting strategies is displayed and significant calculations in each class is studied. The benefits and bad marks of various strategies were brought up [2]. The calculations for concealing touchy affiliation rules like protection preserving guideline mining utilizing hereditary calculation.

Chung-Min Chen, [8] introduce dithered B-tree, a B-tree file structure that can fill in as a building obstruct for acknowledging productive framework usage in the zone of secure and private database outsourcing. The dithered tree embed calculation [8] can be additionally upgraded to bring about just a single traversal from the root to the leaf, rather than two. The file structure from learning regardless of whether the inquiry term (i.e., key) is available in the database and check the information for secure and private database outsourcing.

In Privacy Preserving Data Mining, information irritation is an information security strategy that includes "clamor" to databases to permit singular record secrecy. This method [9] enables clients to determine key rundown data about the information while keeping a security rupture. Four predisposition sorts have been proposed which evaluate the adequacy of such a system. Be that as it may, these predispositions manage basic total ideas (midpoints, and so forth.) found in the database. The creator propose a fifth kind of inclination that might be included by irritation procedures (Data mining Bias), and observationally test for its reality. In internet business applications, associations are occupied with applying information mining ways to deal with databases to find extra learning about clients.

The creator idea in this paper is Privacy Preserving mining of incessant examples on scrambled outsourced Transaction Database (TDB) [1]. They proposed an encryption plot and including fake exchange in the first dataset. Their technique proposed a system for incremental affixes and dropping of old exchange clusters and decode dataset. They additionally investigate the break likelihood for exchanges and examples. The Encryption/Decryption (E/D) module encodes the TDB once which is sent to the server. Mining is directed over and again at the server side and decoded each time by the E/D [1] module. Accordingly, we have to contrast the unscrambling time and the season of straightforwardly executing from the earlier finished the first database.

4. Privacy Threats and Framework

The main goal of privacy threat is to disclose the identity and personal information, which is sensitive for the respective one. There are some kind of privacy threats which may disclose ones sensitive information:

• Identity disclosure [8]: In identity disclosure threat, intruder can get the individual identity from published data. This threat is affined to direct identifier attribute.

• Attribute disclosure [9]: In attribute disclosure threat, intruder can reveal individual's sensitive information. This threat is affined to sensitive attribute.

• Membership disclosure [10]: Any information concerning individual is disclosed from data set, known as membership disclosure. This may happen when data is not protected from identity disclosure.

Plenty of privacy preserving techniques are existingto solve the secrecy breaching problems. The general outline for these techniques can be classified in five phases in which data is goes through [11]: • Distribution: The distribution of data can be either centralized or distributed. In centralized distribution, all the data kept in repository on central server, whereas all data are stored on different databases.

• Modification: This describes how data is modified for concealing the original data. To fulfill this requirement, various ways of modification applied on data like perturbation, aggregation, swapping, sampling, suppression, noise addition.

• Data Mining Algorithm: The data mining approaches comprises the ways of generating decision making results from the data. This phase\stage deals with various algorithms like decision tree, clustering, rough sets, association rule, regression, classification.

• Data hiding: The data hiding entails raw knowledge or aggregate data which desires to be hidden.

• Privacy Preservation Technique: The privacy preservation approach includes different approaches to attain privacy, which are, generalization, data distortion, data sanitation, blocking, cryptographic and anonymization.

5. CONCLUSIONS

This paper addresses the design issues for extracting knowledge from large amounts of data without violating the privacy of data owners. So for privacy preserving researcher first introduce an integrated baseline architecture, design principle, and implementation techniques for privacy-preserving data mining systems. Here detailed discussion of different techniques and combination of those are done. In few works both numeric and text information was protected, so the time and space required for those calculation is similarly high. So a proper method need to develop for anomaly detection and there thoroughly investigation issues related to the design of privacy-preserving data mining techniques.

REFERENCES

[1]. Pedreschi, D., Ruggieri, S. & Turini, F. (2008). Discrimination-aware data mining. Proc. of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008), pp. 560-568. ACM.

[2]. Hajian, S., Domingo-Ferrer, J. & Martinez-Ballesté, A. (2011a). Discrimination prevention in data mining for intrusion and crime detection. Proc. of the IEEE Symposium on Computational Intelligence in Cyber Security (CICS 2011), pp. 47-54. IEEE.

[3]. Verykios, V. & Gkoulalas-Divanis, A. (2008). A survey of association rule hiding methods for privacy. In C. C. Aggarwal and P. S. Yu (Eds.), Privacy- Preserving Data Mining: Models and Algorithms. Springer.

[4]. Meij, J. (2002) Dealing with the data flood; mining data, text and multimedia, The Hague: STT Netherlands Study Centre for Technology Trends.

[5]. Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 21(2):277-292.

[6]. Sara Hajian and Josep Domingo-Ferrer "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013

[7]. Pedreschi, D., Ruggieri, S. & Turini, F. (2009a). Measuring discrimination in socially-sensitive decision records. Proc. of the 9th SIAM Data Mining Conference (SDM 2009), pp. 581-592. SIAM

[8]. Hajian, S. & Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. Manuscript.

[9]. C. Clifton. Privacy preserving data mining: How do we mine data when we aren't allowed to see it? In Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2003), Tutorial, Washington, DC (USA), 2003.

[10]. D. Pedreschi, S. Ruggieri and F. Turini, "Discrimination-aware Data Mining," Proc. 14th Conf. KDD 2008, pp. 560-568. ACM, 2008.

[11]. D. Pedreschi, S. Ruggieri and F. Turini, "Measuring discrimination in socially-sensitive decision records," SDM 2009, pp. 581-592. SIAM, 2009.

[12]. Jerry Chun-Wei Lin, (Member, Ieee)1, Qiankun Liu1, Philippe Fournier-Viger2, And Tzung-Pei Hong. "Pta: An Efficient System For Transaction Database Anonymization". August 25, 2016, Date Of Current Version October 31, 2016. Digital Object Identifier 10.1109/Access.2016.2596542.