Indian Journal of Science and Technology, *Vol 10(5), DOI: 10.17485/ijst/2017/v10i5/111138, February 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Survey on Privacy Preserving Data Mining Techniques

G. Jelin Taric* and E. Poovammal

Computer Science and Engineering, SRM University, Chennai - 603203, Tamil Nadu, India; jelintaric@gmail.com, poovammals@gmail.com

Abstract

Objectives: Recently Privacy preserving data mining (PPDM) is known to be the most important aspect among researchers. As Privacy preserving data mining permits, sharing and exchanging of privacy susceptible data for analysis, it has grown more and more popular. Since one of the important aspects of data mining is safeguarding privacy, this paper aims to analyze different technique adopted for preserving privacy while maintaining the real characteristic of data under consideration. **Methods/Statistical Analysis:** In this paper, the authors evaluate the usefulness of PPDM techniques based on its performance, data usage, and uncertainty level and so on. The findings of authors and limitations in each technique are consolidated. **Findings:** Each technique has its unique way of usefulness apart from its limitations. Anonymization approach makes the data owners anonymous but vulnerable to attacks like linking attacks. Perturbation approach protects each and every attribute independently but unable to regenerate the original values from the perturbed data. Randomization technique provides good security for individual's private data but the utility of the data. The degradation of the utility of the data is due to the noise added. The cryptographic technique provides good security for the data while providing high utility. But it falls short in efficiency when compared with other methods. Anyhow, there is no single privacy protecting algorithm capable of outperforming every other algorithm in all possible yardsticks. On the contrary, one algorithm may do well when compared to another, on a particular criterion. **Novelty/Improvement:** The paper presents various techniques which are used to perform PPDM technique and also tabulates their advantages and disadvantages.

Keyword: Anonymization, Cryptography, Perturbation, Privacy Preserving Data Mining, Randomization

1. Introduction

Data Mining[1] indicates mining or deriving wisdom from voluminous data. Data Mining has been defined as the procedure of finding intriguing knowledge from huge volume of data that has been saved in databases, or any data archives. Through data mining, it becomes possible to extract consistencies, interesting facts, or advanced information out of the database checked or searched from various angles. Such extracted knowledge can then be used for query processing, decision making, data management, and process control. Considered as the most crucial benchmarks in database systems, Data mining certainly is one among the best reliable interdisciplinary advancements in the industry of Information. Data mining investigation involves extracting possibly fruitful information from voluminous selection of data that includes a wide range of application domains like client relationship management and market basket research. The information mined from large database may be clusters, rules, patterns, or even classification models. Throughout the entire data mining operation (starting with collection of data through finding knowledge) the data used normally consist of sensitive personal information about many individuals which they does not wants to disclose to any parties such as the owner of the dataset, collectors, users, and miners. There are lot of opportunities to mishandle the delicate information if the sensitive data about individuals are leaked[2]. Availability of voluminous data assures the possibility of learning a great deal of information regarding individuals out of the public data. And this fact naturally leads to more responsibility

**Author for correspondence*

with regard to privacy of person's individual data[3]. Entire information regarding any person often contains certain private information. Careless distribution of such data means instant violation of the privacy with regard to the individual. Privacy has been defined as the condition or quality of being blocked or secluded from others view or presence[4]. When data mining gets related with privacy, privacy suggests keeping an individual's information from freely being obtainable by other people[5]. As long as one does not feel his or her personal information has been negatively used, privacy is not considered to be violated. When once personal sensitive information has been revealed, one is not in a position to prevent misuse of the same. Privacy preservation has been treated very crucial for evading information spillage for the most effective usage of voluminous data. It involves storing data in electronic format with no disturbance to the individual. Hence, privacy has to be preserved before, during and after the data mining procedure. Privacy Preserving is found to have emerged as one major concern with regard to the data mining process success[6]. PPDM is used to protect the privacy of an individual's personal data or classified knowledge without having to sacrifice complete usage of the required data. Privacy intrusions on the part of personal data are common, people have become aware of this, and they are naturally very hesitant of sharing their classified information. Importance of the issues in PPDM have been realized more pronouncedly during the recent period. This is due to the fact that the ability to save users' personal data has increased and data mining algorithm leveraging these information have become increasingly sophisticated. It is not possible to apply privacy constraints in a single step. One must remember that PPDM technique has to be applied throughout the data mining practice beginning with data collection through information/knowledge generation. The goal of PPDM consists of constructing procedures to transform the raw data in such a way as to maintain confidentiality of private knowledge and private data even subsequent to data mining process[7].

2. Techniques

Developing data mining practices that do not increase risk of misuse[8] of data of an individual and not degrading the use of information is the primary goal of PPDM. A majority of these techniques modify the raw data in some form for attaining privacy preservation. Then, the modified data which is ready to mine should comply with privacy needs and at the same time not lose the[8] benefit of the process of mining. The particular information regarding a person get stored in the form of table (that is, a relation) of rows (or records) and columns (or attributes), as stated by[9]. Many privacy preserving techniques with regard to data publishing like cell suppression, randomization, data swapping, sampling, and perturbation are constructed for publication of micro data[10]. Privacy preservation has evolved through different stages of development. As the existing technique involves a level of complexity, the intention of privacy preservation is treated as a new research area. Personal identifications are generally removed prior to the publication of information in order to mine data. Preservation of privacy which has been considered as a crucial matter is attained by using various techniques. Privacy preserving data mining techniques can be classified under three major categories such as Perturbation, Anonymization and Cryptography is shown in Figure 1.

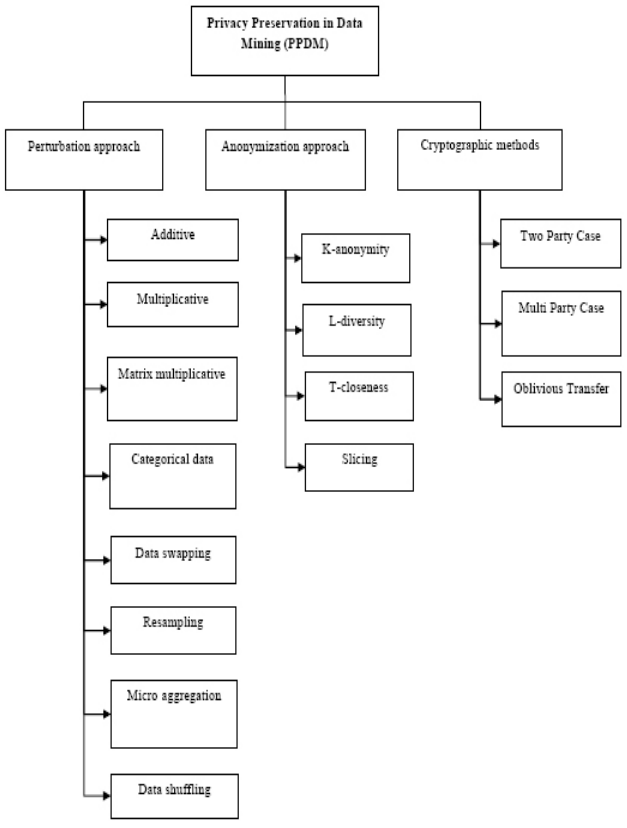

Figure 1. Classification of Privacy Preserving Data Mining Techniques.

2.1 Perturbation

Data Perturbation[11,12] is a method used to modify data with the use of random process. Apparently, this method disfigures delicate data values through altering them by subtracting, adding, or by some other mathematical procedure. This method may be able to cope with various data types: Boolean type, character type, integer, and classification type. It is essential to preprocess the raw data set before entering to perturbation method. Perturbation of data is known by other names such as data noise and data distortion. Securing sensitive data is vital and critical and the data perturbation process performs crucial role in the preservation of delicate data. Distortion can be applied using various techniques like data rearrangement matrix, adding noise, by adding unfamiliar values, and so on.

2.2 Anonymization

Information is frequently published through the removal of vital identity indicators like social security number and name from individual records. Even though, the combination of different attributes from different datasets (quasi-identifiers) may be used for identifying individual records accurately. For instance, certain attributes like birth, race, zip, and sex appear in voter list. If such indicators appear in sensitive database such as medical data, quasi identifiers are employed for gathering identification of the concerned person by linking the two datasets together. For preserving privacy, k-anonymity model that used suppression and generalization was proposed by[10] says, that any individual becomes indistinguishable if there is a minimum of k-1 other individuals having same details with respect to the quasi-identifiers. The process of Generalization requires recoding a certain entry using a less distinct but meaningfully consistent entry. For instance, for reducing the risk of identification, birth date can get generalized to certain range like year of birth. Suppression implies totally not revealing a value. It is obvious that while such strategies reduce identification risk in the public records usage, they reduce the precision of operations on the modified data. Including this two more attacks called homogeneity attack and background knowledge attack[13] is also possible in this method.

2.2.1 L-diversity

The two major attacks called as Homogeneity attack and background knowledge attack leads to the creation of a new technique called l-diversity which is an advancement of k-anonymity model where it protects privacy even though the data owner is not aware of any information that the intruder holds[13]. This method is derived from k-anonymity model where k records in the dataset will match with k-1 other data in the records with reduction of the scale or level of detail from the dataset to form an l-diverse dataset.

2.2.2 T-closeness

To achieve l-diversity, every set of records in the dataset which approve k-anonymity needs to have l well represented values for each sensitive attributes. In addition to this the previous technique cannot safeguard the dataset from disclosure of the attributes. For these cases t-closeness method was discovered and it overcomes the problem of k-anonymity and l-diversity.

The t-closeness concept which we have introduced in this section is defined by[14]. Every dataset assumes to satisfy t-closeness if every equivalence classes has t-closeness. The model of t-closeness is an enhancement from the l-diversity model. An important feature of the l-diversity model is that it takes all given value attribute in a same way irrespective of its dispensation in the data.

2.3 Randomization

Randomization is considered as one of the frequently used approaches in PPDM research. This method involves adding noise on to the authentic data for creating values of each record. Perturbation mixed with authentic data are sufficiently huge for maintaining the privacy and hence one cannot recover the actual data. Randomized Response scheme and random-noise-based perturbation help Randomization techniques to achieve both knowledge discovery and privacy preservation. Although involving huge loss of information, this technique is comparatively a better and efficient process. Randomization proves to have the ability of preserving some semantics and anonymizing the entire dataset. Among the currently used privacy preserving data mining methods. Randomization is treated as the crucial method. Harmony between utility and privacy[15] as well as knowledge discovery are provided by this. After being balanced, the randomized data gets transmitted to the concerned recipient. Using distribution reconstruction algorithm, the recipient will receive the data. This method offers effective and simple way for ensuring the person's privacy and also preserve is used of data to some extent.

2.4 Cryptography

Cryptography is one method used to preserve sensitive data. Cryptographic technique is very much favored as it offers safety and security of sensitive attributes, and was suggested by authors in[16]. The privacy of a person's record may be broken by final data mining. Consider for example a situation in which several medical institutions look for conducting a joint study on the datasets for certain mutual benefits, while not disclosing unwanted information. It is possible that sometimes when a data mining algorithm is passed to a dataset formed by combining two data sets there is some possibility that the results may disclose private information about the individual. But, this kind of leakage is inevitable.

2.4.1 The Two Party Case

A protocol called constant-round was proposed by[17] for calculating any probabilistic polynomial time function (the opponent is malicious or partly honest). Consider two parties with inputs a and b as an example. These two parties are very much interested in performing a functionality jointly to their inputs for some mutual benefits. Let the functionality be g(x,y)= (g1(x,y),g2(x,y)). Finally g1(x,y) is given to the first party and g2(x,y) is given to the second party. The security here is the only the output is shared with the parties. Other than the output, parties can't learn anything about the protocol.

2.4.2 The Multiparty Case

The protocols of multiparty case allows the participants to calculate their inputs with a combined method as well as not leaking related data regarding the inputs. Which means the parties can able to evaluate the function by protecting the privacy of the input as in the previous model. This was achieved and demonstrated by many researchers in[18-20], for various scenarios. The protocols of multi-party case needs every pair of parties to interchange messages so that each gate of the circuit can compute functions effectively. But this is impossible in some situations like web applications. Because the application running between the server and the client does not support effective communication between every pair of parties. Another overhead in this scenario is communication and computation are linear to each other with respect to the size of the circuit.

2.4.3 Oblivious Transfer

This protocol is considered to be an important building block for protected computation. The idea of 1-out-2 oblivious protocol was proposed by[21]. In the protocol of oblivious transfer two parties are required, a sender and a recipient. Sender's input is a pair (X0,X1) and receiver's input is Q € {0,1}. After the protocol ends sender cannot learn anything and receiver can only learn X_Q. Although the accuracy and security of altered data is ensured in Cryptographic methods, when several participants are involved, this approach falls short on delivering. Furthermore, confidentiality of individual records may be breached by the data mining results. Although there are a lot of solutions while using semi-honest models, very low number of studies has been conducted when it comes to malicious models.

2.5 Generalization

Generalization is found to be one among the traditional anonymization methods. Generalization can make a more person specific dataset to a less person specific dataset. It has been a vastly used method that takes the place of QI values using less-particular but semantically constant value. Generalization causes huge information loss because of high amplitude of the QI. For avoiding information loss, it is necessary to keep records in conformity class close to one another. Another deficiency is that generalization renders the data fruitless. Efficient study of attribute interrelationship may also get lost because of discrete generalization of every attribute. The advantages and limitations of all PPDM techniques are tabulated in Table 1.

Table 1. Advantages and Limitations of PPDM Techniques

Technique	Advantages	Limitations
Anonymization technique of PPDM	Data owner's sensitive or private data are to be secreted.	More information loss, Linking attack
Perturbation technique of PPDM	Preserves various attributes independently.	Information loss and Cannot regenerate original data values.
Randomized Response technique of PPDM	It provides good efficiency. Simple and useful for keeping the individual information secretly.	Loss in individual's information. Not much good for database containing several attributes.
Cryptography technique of PPDM	Data transformation is accurate and protected. Provides better privacy and data utility.	It is particularly hard to scale if multiple parties are involved.

3. Conclusion

The primary objective of PPDM is promoting algorithm to conceal sensitive data or offer privacy. These sensitive data do not get revealed to unapproved parties or invader. In data mining there exists a trade of between utility and privacy of data. When we accomplishes one it inevitably leads to the detrimental impact on the other. Many PPDM techniques in existence are reviewed in the paper. Ultimately, it is concluded with the fact that there is no single PPDM technique in existence that outshines every other techniques with relation to each possible criteria such as use of data, performance, difficulty, compatibility with procedures for data mining, and so on. A particular algorithm may function better when compared to another, on a specific criterion. Various algorithms may be found to function better than one another on given criterion. Researchers are doing extensive research in ensuring that the sensitive data of a person is not revealed as well as not compromising the utility of data so that the data can be useful for many purposes.

4. References

1. Ghalehsefidi, Narges J, Mohammad ND. A Hybrid Algorithm based on Heuristic Method to Preserve Privacy in Association Rule Mining. Indian Journal of Science and Technology. 2016 Jul; 9(27):1–10. https://doi.org/10.17485/ijst/2016/v9i27/97476
2. Benjamin CMF, Ke W, Rui C, Philip SY. Privacy-Preserving Data Publishing: A Survey of Recent Developments. ACM Computing Surveys. 2010 Jun; 42(4).
3. Charu CA, Philip SY. A General Survey of Privacy-Preserving Data Mining Models and Algorithms, Springer US. 2008; 11–52.
4. Privacy. Available from: https://en.wikipedia.org/wiki/Privacy. Date Accessed: 29/09/2016.
5. A New Model for Privacy Preserving Sensitive Data Mining. Available from: http://ieeexplore.ieee.org/document/6396017/. Date Accessed: 26/07/2012.
6. Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects. Available from: http://ieeexplore.ieee.org/document/6394662/. Date Accessed: 23/11/2012.
7. Aruna Kumari D, Rajasekhara Rao K, Suman M. Privacy Preserving Data Mining. Springer International Publishing. 2014; 517–24.
8. Wang PS. Survey on Privacy Preserving Data Mining. International Journal of Digital Content Technology and its Applications. 2010; 4(9):1–7. https://doi.org/10.4156/jdcta.vol4.issue9.1
9. Latanya S. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge based Systems. 2002 Oct; 10(5):557–70. https://doi.org/10.1142/S0218488502001648
10. Yan Z, Ming D, Jiajin L, Yongcheng L. A Survey on Privacy Preserving Approaches in Data Publishing. IEEE Computer Society. 2009; 128–31.
11. Jinfei L, Jun L, Joshua ZH. Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity requirements. Proceedings of 11th IEEE International Conference on Data Mining Workshops, China, IEEE. 2011. p. 666–73.
12. Kargupta H, Datta S, Wang Q, Krishnamoorthy S. On the Privacy Preserving Properties of Random Data Perturbation Techniques. Proceedings of the Third IEEE International Conference on Data Mining USA. 2003. p. 99. https://doi.org/10.1109/ICDM.2003.1250908
13. Ashwin M, Johannes G, Daniel K, Muthuramakrishnan V. ℓ-Diversity: Privacy beyond k-Anonymity. ACM Transactions on Knowledge Discovery from Data. 2007 Mar; 1(1).
14. Ninghui L, Tiancheng L, Suresh V. t-Closeness: Privacy beyond K-Anonymity and l-Diversity. Proceedings of the IEEE 23rd International Conference on Data Engineering, Istanbul. 2007 Apr. p. 106–15.
15. Clifton C, Murat K, Jaideep V, Xiadong L, Michale YZ. Tools for privacy-preserving distributed data mining. ACM SIGKDD Explorations. 2002 Dec; 4(2):28–34. https://doi.org/10.1145/772862.772867
16. Data Perturbation and Features Selection in Preserving Privacy. Available from: http://ieeexplore.ieee.org/document/6335531/. Date Accessed: 20/09/2012.
17. Andrew CCY. How to generate and exchange secrets. Proceedings of the 27th Annual Symposium on Foundations of Computer Science (FOCS). 1987. p. 218–29. PMid:3572436 PMCid:PMC1031495
18. Goldreich O, Micali S, Wigderson A. How to Play any Mental Game - A Completeness Theorem for Protocols with Honest Majority. Proceedings of the 19th Annual Symposium on the Theory of Computing, ACM, USA. 1987; 218–29.
19. Michale BO, Shafi G Wigderson A. Completeness theorems for non-cryptographic fault tolerant distributed computation, Proceedings of the 20th Annual Symposium on the Theory of Computing (STOC), ACM, Israel. 1988; 1–10.
20. Bhanumathi S, Sakthivel P. Preservation of Private Information using Secure Multi-Party Computation. Indian Journal of Science and Technology. 2016 Apr; 9(14):1–6. https://doi.org/10.17485/ijst/2016/v9i14/74588
21. Shimon E, Oded G, Abraham L. A Randomized Protocol for Signing Contracts. Communications of the ACM. 1985 Jun; 28(6):637–47. https://doi.org/10.1145/3812.3818