A General Survey of Privacy-Preserving Data Mining Techniques

Smita Rathod[1] ,S.S.Hatkar[2]

[1]CSE, SGGSIE&T, Nanded
[2]CSE, SGGSIE&T, Nanded

---***---

Abstract - Privacy in Data Mining has become more prominent and popular due to the fact that it maintains the privacy of sensitive data for analysis purposes. The typical data collection and distribution process results in a potential risk of threats and attacks on privacy. Any private information about individuals, corporations and organizations must be suppressed before it is shared or re-identified by linking public data like voter's data. The paper giving review of k-anonymization and different privacy preserving data mining techniques

Key Words: Privacy-Preserving, K-anonymity, Re-identification

1.INTRODUCTION

Data mining requires processing vast quantities of knowledge. This is to minimize randomness and discover the pattern that is hidden. In the information industry and in culture, it has gained a great deal of interest. Preserving the integrity of data so that confidential information is not exposed to third parties is the main challenge when the data goes through any of the process. So, in data mining, privacy plays an important role.

In order to identify previously unknown, legitimate patterns and associations in large data sets, data mining requires the use of refined data analysis techniques.. The main objective of the PPDM is to reduce the risk of misuse of sensitive data and information which helps to produce the same result generated in the absence of privacy techniques.

1.1 **Privacy Preserving Data Mining** :

Privacy has become key for knowledge-based applications. This data mining based on privacy is critical for data types such as data from health reports, data from medical diagnosis and data from the funding system. Five dimensions need to be considered and listed below to achieve optimized results while protecting the privacy of data subjects effectively:

1. The distribution of fundamental data
2. Modifying the basic data
3. Method of mining used
4. If the basic data or rules need to be concealed and
5. Additional approaches used to preserve privacy.

1.2 **Data Partition Model:**

In this scenario data sets can be distributed on various sites.There are two types of data distribution, such as homogeneous distribution (horizontal partitioning) and vertical partitioning (heterogeneous distribution).In Horizontal Partitioning, the different sites or places which have different data sets of records containing with the same attributes.The different sites or locations can have different attributes of the same record data sets in vertical partitioning.

2.Techniques:

There are many techniques can be used to protect data from unauthorized user such as anonymization, perturbation, randomization, condensation and cryptography based approach which are explained in below.

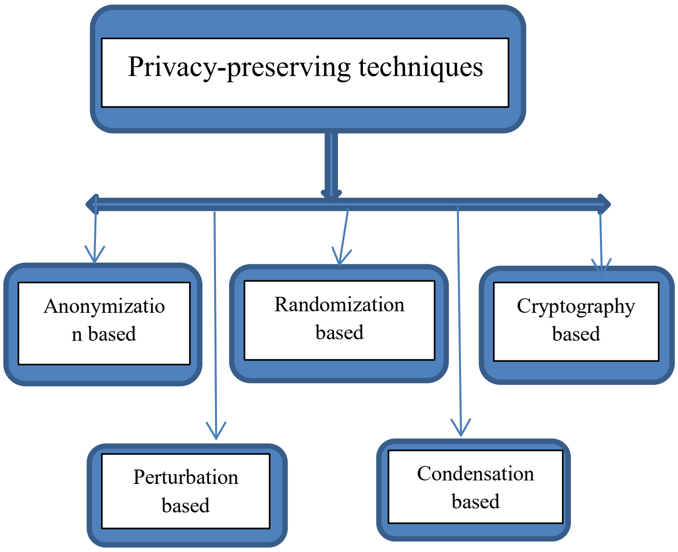

Fig. Classification of privacy-preserving techniques

A. **Anonymization based PPDM**

The primary objective of the Anonymization method is to protect individual records separated by generalization, suppression, and suppression within a community of records.Manipulating the content such that the records connect to k-anonymity is one way to anonymize the data collection.

There are two approaches used in this method to achieve k-anonymity of a data set are generalization and suppression.The basic form of anonymization consists of following types of attributes:

i. Explicit Identifiers are a list of data-containing attributes that define a record owner uniquely, such as name, social security, etc.
ii. Quasi Identifiers is a set of features that, when combined with publicly available data, could potentially identify the owner of a record.
iii. Sensitive characteristics are a series of characteristics containing detailed information from sensitive individuals such as no ID card, wages, etc.

The original data set or records are manipulated by k-anonymous in the Anonymization process with respect to quasi-identifier attributes.

The following table shows the original data set and k-anonymous data

Emp-ID	Name	Age	Salary
201	ABC	30	20000
202	XYZ	45	30000
203	PQR	27	50000

Table.1(a) Original Data

Emp-ID	Name	Age	Salary
2**	ABC	3*	20000
2**	XYZ	4*	30000
2**	PQR	2*	50000

Table.1(b) K-anonymous Data

B. **Perturbation Based PPDM**

It is a convenient and reliable way to secure confidential electronic data from unauthorized users or hackers.Data Perturbation allows individual values to be randomly disrupted to protect privacy before data is released. In perturbation, the original values are changed by such synthetic data values so that the statistical information computed from the disrupted data does not differ to a greater extent from the statistical information computed from the original data.. Additive noise or data swapping or synthetic data generation may be used for perturbation.

C. **Randomized Response Based PPDM**

Randomization response is the research approach used in survey interviews enables sensitive problems such as illegal activity to be discussed while retaining confidentiality.This randomization technique is very simple and does not require any prior knowledge of data distribution. Therefore, randomizing data is carried out during data collection process.This approach is useful for the classification of the decision tree, which is based on aggregate values.

In PPDM algorithms, several randomization techniques have been established, including

i. Adding numbers randomly
ii. By generating random vectors and
iii. Random sequence permutation.

D. **Condensation approach based PPDM**

This approach works best with pseudo-data than with original data changes, which helps protect privacy rather than original data strategies.

E. **Cryptography Based PPDM**

It is a technique of hiding and transmitting information in a way that can be read and accessed only by those who have access.It is used for contact between two parties, along with the participation of third parties, in a safe manner.For two reasons, cryptographic techniques find their usefulness in such scenarios: First, it offers a well-defined privacy model that includes methods to prove and measure it. Second, in this area, a wide set of cryptographic algorithms and constructs are available to implement privacy-protecting data mining algorithms.

The data can be distributed vertically or horizontally among various collaborators.Both of these techniques are based on a basic encryption protocol known as the Secure Multiparty Computation (SMC) technology. Moreover, data mining results will violate the privacy of individual records.

ADVANTAGES OF PPDM

- In developing different data mining strategies, PPDM is very advantageous.
- It allows large quantities of privacy-sensitive data to be shared for analysis purposes.
- It can track and collect huge amount of data using current hardware technologies.

DISADVANTAGES OF PPDM

- One of the major challenges of preserving the privacy of data mining is the abundant availability of personal data.
- In order to facilitate proper data processing, many systems exist, but a lot of work remains, and certain challenges must be addressed in order to be deployed.

2. Survey Table

The table below, will give overall information about the available methods about the privacy-preserving data mining techniques. It will help the new author for selecting algorithms, give knowledge about present methods and new challenges in this domain. It will knowledge about the present work and the work done by other authors.

References	Approach	Technique used	Publication	Result & Accuracy

		For PPDM	Year	
Aristides Gionis and Tamir Tassa [1]	NP-hard and Polynomial approximation for optimal solution	Condense the data into several predefined size groups.	2009	Two information-theoretic measures for capturing the amount of information lost during the anonymization process is obtained.
N. Kumbhar and R. Kharat N [2]	Association Rule By Horizontal and Vertical Distribution	Various methods in the area of the law of association have been studied.	2012	The output of all models is evaluated in terms of privacy, protection and communication.
George Mathew, ZoranObradovic [3]	Distributed Privacy Preserving	Provides an algorithm to construct a better model for decision making	2012	The overall accuracy of a classification model is improved.
G. Mathew, Z. Obradovic [4]	Technical & methodological approach and give judgmental knowledge	Decision Tree	2011	A graph-based framework To protect patient's sensitive data.
Shweta Taneja, Shashank Khanna, [5]	A tabular comparison of different methods	Description of PPDM Challenges and methods	2014	The methods of Cryptography and Random Data Perturbation Perform better than the other approaches.
M. Antony Sheela, K.Vijayalakshmi N. [6]	Applied methods on the vertically partitioned data.	Partition Based Perturbation	2017	Individual data is changes when threshold value is reaches
Savita Lohiya and LataRagha [7]	Hybrid Approach	combination of K Anonymity and Randomization used	2012	It has more precision and it is possible to regain original data.
JalpeshVasa, PanthiniModi [8]	Anonymization based techniques used to Protect privacy by reducing the granularity.	t- closeness	2018	Anonymization is used to protect privacy by reducing granularity are used.
Abhijit Patankar [9]	Protect privacy with less information loss	K-nearest neighbour	2019	It proves that k-anonymity is better to protect from attacks.

Survey Table

3. CONCLUSIONS

This paper, discussed about various approaches and techniques which are used for protection of confidential information. Due to huge amount of data collection of information, it is important to maintain the privacy of sensitive information. Each approach has its own benefits and drawbacks. We are collecting is some methods and work done by other user for ease of use to new users.

REFERENCES

1. Alpa Shah and Ravi Gulati, "Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey",

International Journal of Computer Application, Vol. 137 – No 12, March 2016, 40-46.
2. J. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data”, in The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, CA, July 2002, IEEE 2002.
3 Charu C. Aggarwal, “A General survey of Privacy-Preserving Data Mining Models And Algorithms”, IEEE, pp 11-52,2008.”
4 Gionis A, Tassa T., “k-Anonymization with Minimal Loss of Information”, Knowledge and Data Engineering, IEEE Transactions, pp. 206-219, 2009.
5. 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.
6. Accuracy-Constrained Privacy- Preserving Access Control Mechanism for Relational Data Zahid Pervaiz, Walid G. Aref, Senior Member, IEEE, Arif Ghafoor, Fellow, IEEE, and Nagabhushana Prabhu, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.26, NO. 4, APRIL 2014
7. Samarati, Latanya Sweeney, “Protecting privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression”, IEEE Transactions on Knowledge and Data Engineering, 2001
8. Datta, A., Joye, M., 2016. Cryptanalysis of a privacy-preserving aggregation protocol. IEEE Transactions on Dependable and Secure Computing 82, 23– 30.
9. K. Saranya, K. Premalatha, S. Rajasekar, A survey on privacy preserving data mining, in International Conference on Electronics & Communication System (IEEE, 2015)
10. Amit Datta and Marc Joye “Cryptanalysis of a Privacy-Preserving Aggregation Protocol”IEEE Transactions on Dependable and Secure Computing 14(6):693–694, 2017.
11. Manish Shannal Atul Chaudhar/ Manish Mathuria3 Shalini Chaudhar/ Santosh Kumar5 “An Efficient Approach for Privacy Preserving in Data Mining” 978-1-4799-3140-8/14/$31.00 ©2014 IEEE 2010
12. Sweeney L, "Achieving k-Anonymity privacy protection uses generalization and suppression" International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 571-588, 2002.
13.Shilpa Rathod , “ Survey On Privacy preserving Data Mining Techniques”,International Journal of Engineering Research and Technology, Vol.9,2278-0181,2020.

BIOGRAPHIES

Smita Rathod is currently pursuing master’s degree program in Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded (MH),India.

Prof. S.S.Hatkar is Associate Professor in Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded (MH),India.