A Technological Survey on Privacy Preservation Data Mining Techniques

Ms. Suman Madan, Astt. Prof(IT), JIMS, Sec-5, Rohini, INDIA, madan.suman@gmail.com

Ms. Lakshita Aggarwal, Student Scholar, JIMS, Rohini, INDIA, lakshitaaggarwal31@gmail.com

Abstract: In recent few years, lots of data is being created and collected from so many devices like sensing mobile devices, aerial devices, radio frequency readers etc. So, most of the enterprises are actively collecting data and storing data in large databases for that we need some tools to process those raw data into information. The data is analyzed through various data mining techniques which help us to analyze and process the data into some useful information and patterns from large amount of data. The data might contain private information about people or business or some personal information. Privacy is an important issue when one wants to make use of data that might involves individual's sensitive information. Nowadays privacy has been important aspect for all. Privacy is important area of concern even in healthcare, financials, commercial, academic, government organizations etc. A large amount of data is being collected and processed using data mining techniques so that it prevents individual's privacy violation. Privacy-preserving data publishing (PPDP) provides methods and tools for publishing useful information while preserving data privacy. In this paper, a brief yet systematic review of several anonymization techniques and algorithms have been designed for preserving the personal information of person's identity. This paper focus on effective method that can be used for providing better data utility and also provides better methods for protecting the privacy of the data.

Keywords—Anonymization, Encryption, AnonymityApproches, Distributed database, k-anonymity, l-diversity, t-closeness.

I. INTRODUCTION

Managing privacy of data is becoming an increasingly difficult challenge nowadays. Preserving privacy [1],[3] is an important concern of every organization, individual or an enterprise. The main objective of privacy[17] preserving algorithms is to mine the appropriate information from the raw data and protecting it at the same time. There are many technology that converts clear text into a non-human readable data for privacy preserving data publishing .This helps in protecting data from getting pirated. Many secure algorithms[20] have been proposed so far for protecting the data.

There is no specific algorithm or technique that performs better on possibly all above stated criterion. It is reasonably true that one algorithm might be better in one criteria but doesn't fulfills the other necessities. Privacy concerns avoid building of centralized warehouse i.e. data doesn't remains at one location it remains scattered in different locations at several places and no one is allowed to transfer data at one's other place. Suppose, some hospitals wants to get some aggregations about a specific diagnosis of their patient's records while each hospital is not allowed to disclose individual private data. Then for reaching on a conclusion all hospitals would run a joint and secure protocol on their distributed database to reach to the desired information. Distributed database [4],[5] provides more security because data is not at a single point of location. Distributed database is divided in two parts:-

- Horizontal Partitioning – It divides databases into a number of horizontal patterns. The records are placed at different places of the same entities.
- Vertical Partitioning – It divides databases into a number of vertical patterns. All the values of different attributes resides in different places.

II. PRIVACY PRESERVATION DATA MINING

In privacy preserving data mining (PPDM), the data is collected by various organizations and stored in various databases. PPDM framework[8] has 3 levels:

1. The raw data is collected from single or multiple databases. Privacy concerns or rules are applied on it for analytical purposes. Number of transactions takes place for the transformation of raw data and then it is further stored in warehouses.
2. Then, the data is taken from warehouses to process the data number of processes are applied on this stage such as blocking, suppression, perturbation, modification, generalization , sampling etc. Data mining algorithms are modified in such a manner that protecting data privacy is not sacrificed in any manner.
3. At the last level after applying various algorithms raw data is converted into useful information.

 DOI : 10.18231/2454-9150.2018.0992

A. PPDM Techniques

PPDM techniques [20] tends to transform the original data so that privacy techniques are applied which doesn't defies any constraints. Different PPDM techniques classified are:-

1. *Data or rule hiding:* Data hiding means protecting sensitive data values. This group tells that the grouped data should be hidden. And rule hiding states that protecting the confidentiality in data. Ex – names, social security numbers etc.
2. *Data Distribution:* It refers to the distribution of data in vertical and horizontal partitioned data sets. Horizontal partitioned data sets states different sets of records exists in different places while in vertical data sets all the values for different attributes resides in different places.
3. *Data modification :* Data modification refers to changing the unique values of a database that is shown in public this guarantees high privacy protection. Different techniques are used such as:-
   - Perturbation – In this we replace an old attribute with the new attributes such as changing 1 with 0 or 0 with 1 or some other kind of replacements.
   - Blocking – In this we replace an old attribute value with the "?".
   - Swapping – It refers to show just few samples of data to the population.
   - Sampling – It refers to show data only to a sample population.
   - Encryption – Cryptography techniques are used for the encryption.
4. *Data mining algorithms:* It is an algorithm for which data preserving technique is designed such as classification, association rule and clustering algorithms.
5. *Privacy Preserving Techniques:* Privacy preserving techniques [8] includes protecting the data from various thefts. Data anonymization is the process of removing personally identifiable information from data. The complete privacy publishing process is shown in Fig 1. It is done in order to release information in such a way that the privacy of individuals is maintained.It is a technique that is used to protect private information in your data while preserving, to varying degrees, the utility of that data; however, as we'll see, this tool is only best put to use in combination with others, and not as a standalone strategy to protect your data. Here, collector collects data from the original database, some anonymization techniques have been performed to prevent the loss of someone's personal information then the data is published by the recipient which prevents the data from being exposed to all. So, it preserves the data loss. For example, census data might be released for the purposes of research and public disclosure with all names, postal codes and other identifiable data removed. The following are common types of data anonymization :-
   - Removal- It is a technique in which we completely remove fields that could be used in any way to identify a person. It is considered a strong form of data anonymization.
   - Redaction - It includes removal and other techniques for anonymization such as blacking data out on paper with a marker and making a photocopy of the result. This means that original data is blackened with marker so as to avoid loss of personal information.
   - Encryption- Encryption is strong and difficult to reverse. It present a generation of strong decryption key. Data anonymization isn't intended to be reversible so the management of decryption keys is also a concern. Ideally, a strong and fully random key would be generated and then immediately erased from memory when encryption completes. So, that no loss to data can however occur.
   - Data Masking- Data masking [9],[10] is a potentially weak form of data anonymization that may include data scrambling and character replacement. The advantage of data masking is that it maintains the structure of data such that numbers remain numbers and dates remain dates. This allows anonymized data to be used for system testing without triggering application errors.

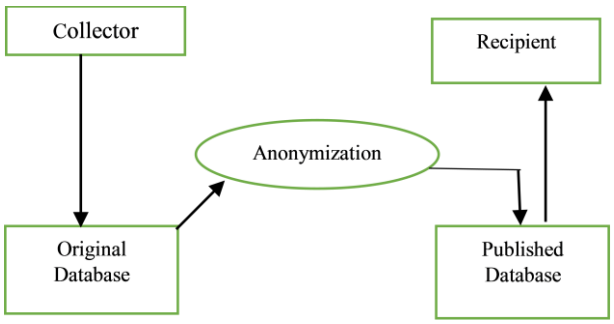

Figure 1: Privacy Publishing Process

B. Algorithms

There exists various privacy preserving data publishing (PPDP) algorithms which are used to de-identify the data [7],[9],[17].

1. *K-anonymity:* k-anonymity is one of the technique for privacy preserving data publishing .It has a database consisting of rows and columns. We have the data which shows the personal details of the people. We have data in two different tables, namely Table I and Table II, wherein we have general data collected by collector and voter data respectively.
   Thus, k-anonymity says:
   Pt= Private Table
   Rt=Released Table
   Qi=Quasi Identifier
   A1, A2 ….An= Attributes
   Let Rt (A1, A2 ….An) be the attributes of table Qi (A1, A2 ….An) be a quasi-identifier associated with Rt. Then

 DOI : 10.18231/2454-9150.2018.0992

each sequence of values in Rt[Ax] appears with at least K occurrences in Rt. So, we conclude after reading the data from these tables that the two different tables can be correlated to draw some conclusions which shows" ALICE is a teacher", as shown in Table III. Therefore, the sensitive data should be securely displayed so that it must be saved from inference attacks.

Table I: General Data Collected

Dob	Sex	Zip code	Profession
1/21/76	M	53715	Teacher
4/13/86	F	53703	Doctor
2/28/76	M	53704	Businessman
1/21/76	M	53715	Engineer
4/13/87	F	53706	Bank manager
2/05/89	F	53708	Clerk

Table II: Voter Data Collected

Name	Dob	Sex	Zipcode
Alice	1/21/76	M	53715
Cathy	4/13/86	F	53703
Bob	2/28/76	M	53704
Andy	1/21/76	M	53715
Dan	4/13/87	F	53706
Ellen	2/05/89	F	53708

Table III: Co-related Rows

Extracted data of table I

Dob	Sex	Zip Code	Profession
1/21/76	M	53715	Teacher

Extracted data of table II

Name	Dob	Sex	Zipcode
Alice	1/21/76	M	53715

The most common technique is Generalization [9], wherein we would partition the data into disjoint groups of transactions such that each group contains the sufficient records with at least one distinct sensitive information so that the privacy of information is sustained. To perform data analysis or data mining tasks on the generalized table, the data analyst has to make the uniform distribution assumption that every value in a generalized interval/set is equally possible, as no other distribution assumption can be justified. This significantly reduces the data utility of the generalized data.

2. *L-diversity :* l-diversity[11],[15] has a group of 'k' different records that all share a particular quasi identifier which are pieces of information that are not themselves unique identifiers but are correlated with an entity that can be combined with other identifiers to create a unique identifier. The notion of l-diversity has been proposed to address this; l-diversity requires that each equivalence class has at least l well-represented values for each sensitive attribute. This technique is efficient as attacker cannot identify the individual records in a database. If we see the above two tables so it shows that the tables are combined to have the record of an individual person completely as shown in Fig: So, we conclude that two tables when combined produced a record that ALICE is a teacher whose DOB is 1/21/76 ,sex is F and zipcode is 53715.
3. *T-closeness:* *t*-closeness [13],[14] is a further refinement group based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation. This reduction is a tradeoff that results in some loss of effectiveness of data management or mining algorithms in order to gain some privacy. The *t*-closeness model extends the *l*-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute. To summarize, an equivalence class is said to have *t*-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold *t*. A table is said to have *t*-closeness if all equivalence classes have *t*-closeness. It is one of the good technique for privacy of data.

The Table IV shows comparison between various techniques.

C. *Privacy preserving data mining approaches*

On the basis of database selection, Privacy preserving data mining techniques [16],[17] are broadly categorized in three ways:-

1. *Heuristic approach*: Heuristic method is used for centralized database. In this approach we view two varieties of data: Raw knowledge and aggregated information. Over both forms of knowledge Classification, Association rule mining, Clustering methods are applied, after that hiding procedures are used to preserve it from incorrect utilization.
2. *Reconstruction approach*: Reconstruction approach is also used for centralized database. Raw data approach is used in it. The data mining methods are applied over the raw data and when the outcome comes, the statistical distributed based method is used over them.
3. *Cryptography approach*: Cryptography approach basically works on distributed database i.e. data is stored in different places. The data which is being stored, may be raw data or aggregated data or both. On applying data mining methods on both the type of data some results are obtained and further encryption technique will be used.

On the basis of method for preserving privacy, the PPDM techniques are further categorized as follows:

1. *Anonymization based approach*: The aim of anonymization procedure is to deal with sensitive or private information about an individual. It is a strategy to retain the data in the original form and hide data with the help of several other approaches. The k-anonymity

 DOI : 10.18231/2454-9150.2018.0992

method says that data should be distinguished in the k records. This can be done using Generalization and Suppression techniques. Due to the some limitation of the k-anonymity method further other methods such as L-diversity, T-closeness methods are derived.

2. *Randomization response approach*: The randomized response approach is to mask the original information by adding some random data or noise in it, so one is not able to distinguish between the original data and the noise. The added random data or noise must be as big as possible so that someone cannot recover the data especially by the un-trusted one. This statistical approach was first proposed by Warner. The randomized response process is done in two phases:-
   - In the primary phase, the original information is randomized and it is transferred to the receiver side.
   - In the secondary phase, the receiver reconstructs the original data from randomized data by distribution reconstruction algorithm.
3. *Perturbation approach:* The perturbation approach modifies the original values with other values, in a manner that the data computed from this perturbed approach data does not distinguish from the other perturbed computed data and the original data. The perturbation approach is of two types:-
   - Additive perturbation: In additive type, random noise is added to the original data.
   - Multiplicative perturbation: In multiplicative type, random rotation method is used to perturb data.
4. *Condensation approach*: Condensation method constructs restricted clusters in dataset after which it generates pseudo information. Pseudo information is then further analysed to produce the information.
5. *Cryptography approach:* Cryptographic procedures are meant for situations where the multiple parties collaborate to compute outcome. It provides two motives:-
   - It offers well-defined model.
   - It provides large set of cryptographic algorithms.

The information could also be distributed among special collaborators in vertical or horizontal dataset.

Table I: Comparison between Privacy Techniques

Privacy measures	Definitions	Advantages	Limitations	Computational complexity
K – anonymity	Framework for constructing & evaluating algorithms for the events or entities that needs to be protected.	• Easy to implement. • Chances of re-identification is less when value of k is high.	• Long processing time. • Homogeneity attack. • No background knowledge.	O(K log k)
L – diversity	Equivalence class is said to have L- diversity if there are well represented sensitive values for the attributes. Every equivalence class must have L-diverse values.	• Reduce the data set like a summary format. • Sensitive attributes would have at most same frequency.	• Difficult • Unnecessary to achieve. • Prone to similarity attack and doesn't prevent attribute disclosure.	O((n^2)/k)
T – closeness	Distance between distribution of a sensitive attribute in class and distribution of attribute in a whole table is no more than a threshold t. A table is said in t-closeness if all equivalence classes have t-closeness**.**	• Prevents data from various malicious and skew-ness attacks.	•Complex computational procedure. •Utility not effective when value of t is so small. •Distribution of sensitive attribute in equivalence class is close to distribution of sensitive attribute in overall table.	2^O(n)O(m)

D. Evaluation Criteria for privacy protection algorithms

Evaluating is to select the appropriate evaluation criteria[18,19,20] for data mining algorithms in terms of performance it provides users with a set of metrics to enable them to choose the best appropriate algorithms for preserving their sensitive data. We distinguish on the basis of algorithm performance, data utility, privacy protection degree and the difficulty of data mining.

1. *Algorithm Performance-* algorithm with $O(n2)$ complexity polynomial time is more efficient than those of $O(en)$ index complexity. An alternative approach is to evaluate the time requirements in terms of the average number of operations of specific sensitive information. These values are considered in order to perform a fast comparison among different algorithms.
2. *Data Utility* - To hide sensitive information, false information should be inserted in the database, or block data values. Although sample Techniques do not modify the information stored in the database. More changes to the database, less data utility of the database.[15],[16]

 DOI : 10.18231/2454-9150.2018.0992

3. *Degree of Privacy Protection* - Privacy protection policy is to protect the information by a certain threshold, but hidden information can also be derived out by some uncertainty. The uncertainty reconstructed by hidden information can evaluate algorithm.
4. *Difficulty of Data Mining* – We need to measure difficulty of data mining algorithms which is different with purification method, and this called parameter horizontal difficulty. We may need to develop a formal framework that upon testing an algorithm whole transitive data outputs are obtained.

III. CONCLUSION

In this paper,we have presented the major functions of how the privacy preserving data mining helps in developing methods to provide privacy to the sensitive information so that they can't be revealed to unauthorized people. In this, we made a try to check a good quantity of current PPDM methods and we conclude stating that there does not exists a single privacy preserving knowledge mining algorithm that can performs all different algorithms on all viable criteria like efficiency, utility, cost, complexity, tolerance in opposition to data mining methods and so on. Various algorithms may perform better than a further on one exact criteria.

IV. REFERENCES

[1] Hua, J., Tang, A., Pan, Q., Choo, K.K.R., Ding, H. and Ren, Y., "Practical--Anonymization for Collaborative Data Publishing without Trusted Third Party," Security and Communication Networks, 2017.

[2] Jyothi, M. and Rao, M.C.S., "Preserving the Privacy of Sensitive Data using Data Anonymization," International Journal of Applied Engineering Research, vol. 12, no. 8, pp.1639-1663, 2017.

[3] Ilavarasi, A.K. and Sathiyabhama, B., "An evolutionary feature set decomposition based anonymization for classification workloads: Privacy Preserving Data Mining," Cluster Computing, vol. 20, no. 4, pp.3515-3525, 2017.

[4] J. Zhang *et al.*, "On Efficient and Robust Anonymization for Privacy Protection on Massive Streaming Categorical Information," in IEEE Transactions on Dependable and Secure Computing, vol. 14, no. 5, pp. 507-520, Sept.-Oct. 1 2017.

[5] Wu, X., Zhu, X., Wu, G. Q., & Ding, W., "Data mining with big data," IEEE Transactions on knowledge and data engineering, vol. 26, no. 1, pp. 97-107, 2014.

[6] Jun Yang, Zheli Liu, Chunfu Jia, Kai Lin, Zijing Cheng, "New Data Publishing Framework in the Big Data Environments," In Proceedings of IEEE International Conference on P2P, Parallel, Grid, Cloud, and Internet Computing, pp. 363-366, 2014.

[7] Huang Xuezhen, Liu Jiqiang, Han Zhen, Yang Jun, "A New Anonymity Model for Privacy-Preserving Data Publishing," Communications System Design, pp. 47-59, 2014.

[8] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, and Yong Ren, "Information Security in Big Data: Privacy and Data Mining," IEEE Access, vol. 2, pp. 1149-1176, 2014.

[9] Yang Xu, Tinghuai Ma, Meili Tang and Wei Tian, "A Survey of Privacy Preserving Data Publishing using Generalization and Suppression," An International Journal of Applied Mathematics & Information Sciences, vol. 8, no. 3, pp. 1103-1116, 2014.

[10] B. Fung, K. Wang, R. Chen, P. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Computing Surveys, vol. 42, pp. 1-53, 2010

[11] Peng Liu , Xianxian Li, "An Improved Privacy Preserving Algorithm for Publishing Social Network Data," In Proceedings of International Conference on High-Performance Computing and Communications & Embedded and Ubiquitous Computing, IEEE Computer Society, pp. 888-895, 2013.

[12] Jun Tang, Yong Cui and Qi Li, KuiRen, Jiangchuan Liu, RajkumarBuyya, "Ensuring Security and Privacy Preservation for Cloud Data Services," ACM Computing Surveys, vol. 49, no. 1, 2016.

[13] Zakerzadeh, H., Aggarwal, C.C., and Barker, K., "Privacy-preserving big data publishing," In Proceedings of the 27th International Conference on Scientific and Statistical Database Management, pp. 26, ACM. 2015.

[14] Konan Martin, Wenyong Wang, Brighter Agyemang, "Efran: "Efficient Scalar Homomorphic Scheme on MapReduce for Data Privacy Preserving," In Proceedings of IEEE International Conference on Cyber Security and Cloud Computing, pp. 66-74, 2016.

[15] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.

[16] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," In Proc. of 32nd Int"l Conf. Very Large Data Bases (VLDB "06), pp. 139-150, 2006.

[17] Puneet Goswami and Suman Madan, "Privacy preserving data Publishing and data anonymization techniques: A Review" In Proceedings of IEEE International Conference on Computing, communication and Automation, pp. 139 – 142, 2017.

[18] Seyedali Mirjalili, "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems," Neural Computing and Applications, Volume 27, Issue 4, pp. 1053–1073, May 2016.

[19] M. Juneja and S. K. Nagar, "Particle swarm optimization algorithm and its parameters: A review," 2016 International Conference on Control, Computing, Communication and Materials (ICCCCM), Allahbad, 2016, pp. 1-5.

[20] Suman Madan and Puneet Goswami, " Hybrid Privacy Preservation Model for Big Data Publishing on Cloud" in proceedings of International Conference on Internet of Things and Challenges(ICITC)2017, pp.113-119.

 DOI : 10.18231/2454-9150.2018.0992