IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY
SURVEY PAPER ON PRIVACY PRESERVING DATA MINING TECHNIQUES

Shreyans Gupta[*1] & Hrushabh Bhadkamkar[*2]
[*1]Information Technology, Institute of Engineering & Technology (DAVV), Indore (M.P.), India
[*2]Computer Science & Engineering, Institute of Engineering & Technology (DAVV), Indore (M.P.), India

ABSTRACT
The main aim of Data mining techniques are to try to find out helpful patterns from the data that is big in quantity. These ideas or patterns are useful to find out some useful information. The abilities learned through fully knowledge mining approaches may contain confidential information about persons or trade. Upkeep of secrecy is a gigantic aspect of information mining also as a result be taught of attaining some information mining ambitions without dropping the secrecy of the individuals .The assessment of privacy preserving data mining (PPDM) algorithms must don't forget the penalties of those algorithms in mining the outcome along with retaining privacy. Inside the constraints of privateness, a couple of ways have been introduced however nonetheless this branch of exploration is in its early life .The success of privateness preserving data mining procedures is measured in phrases of its efficiency, data utility, degree of uncertainty or resistance to data mining procedures and so on. Nevertheless no privateness maintaining algorithm exists that outperforms all others on all feasible standards. Rather, an algorithm could participate in better than one other on one exact criterion. So, the aim of this paper is to show the current situation of privacy preserving knowledge mining framework and tactics

KEYWORDS: Privacy threats, anonymization, randomization response, Perturbation.

I.   INTRODUCTION
In era of digitization, data security and dispensing of data is difficult to achieve. The security of user's sentient information is a vital anxiety. The day to day use up of word privacy about information security, data dispensing and analysis is often times vogue and may be caused to stray. 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. The resulting data size can consist of terabytes of data. The concerns over enormous collection of data are certainly expansion to analytic tools applied to data. With the progress of data analysis and processing method, businesses, industries and governments are more and more publishing micro data (i.e., data that comprise aggregated know-how about members) for data mining functions, finding out sickness outbreaks or economic patterns. While the released datasets provide valuable understanding to researchers, and also they include sentient data about particular whose privacy is also at threat [1]. Privacy [2] refers to the extraction of sentientdata using data mining. The most usual privacy problems are usage of person's respective information, handling false information and controlling access to personal information. Privacy preserving data mining (PPDM) [3,4] is a innovative research path in data mining and statistical records [5], where data mining procedures are analyzed for the aspect-effects they incur in data privacy. Agarwal and Srikant [6] and Lindell and Pinkas [7] presented the primary privateness-keeping knowledge mining algorithms which enable parties to collaborate within the extraction of knowledge, without any party having to disclose individual data.

PRIVACY THREATS 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.

II.   PRIVACY PRESERVING FRAMEWORK
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.
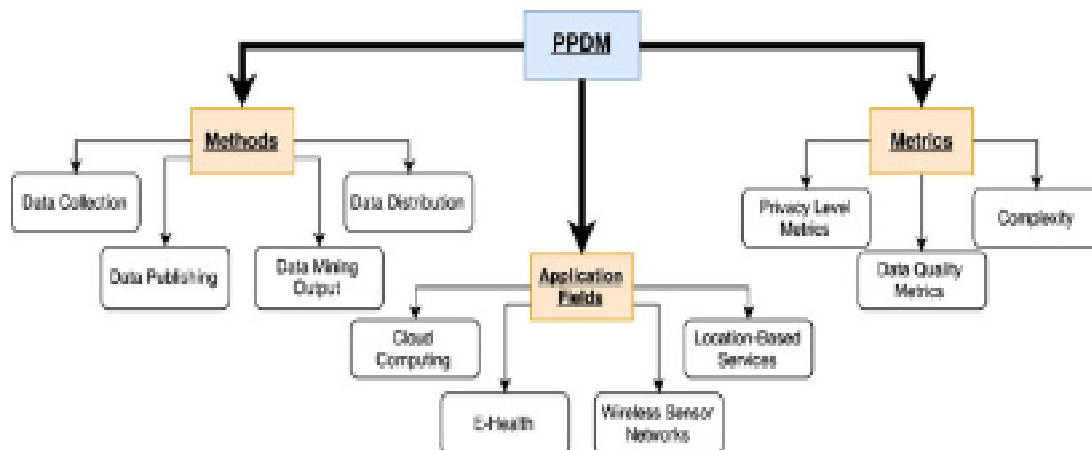
III.   PRIVACY PRESERVING TECHNIQUES
Privacy preserving data mining techniques can be broadly categorized as three ways [13]-

Heuristic approach – Heuristic method is just about used for centralized database, right here two varieties of data is viewed, which is, raw knowledge and aggregated information. Over each forms of knowledge Classification, Association rule mining, Clustering methods are applied, after that hiding procedures are used over the effect of them to preserve it from incorrect utilization.
Reconstruction approach – Reconstruction approach is also used for centralized database, but here, only one type of data is used, which is, raw data. The data mining methods are applied over the raw data. Whatever the outcome comes, the statistical distributed based method is used over them.

Cryptography approach–Cryptography approach is basically works on distributed database, which is the one, where 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 each type of data some results will come, on them encryption technique will be used. The PPDM techniques can be further categorized, which follows these approaches [14,15]. Those categories are –

Anonymization based approach: The aim of anonymization procedure is to conceal sensitive or private information about an individual. Anonymization is a strategy to retain the data in order that original information will be alternate into hid data with the help of several approaches. The k-anonymity method says that data should be undistinguishable within in the k records. This can be done using Generalization and Suppression techniques. Due to the some limitation of the k-anonymity method, L-diversity, T-closeness methods are derived.

Randomization response approach: The randomized response approach is a manner to mask the original information by adding some random data or noise in it, so One are not able to say that knowledge from a person contains genuine know-how or now not. The added random data or noise must be as big as possible hence the data about someone cannot be recovered by the un-trusted one. This is statistical approach first proposed by Warner. The randomized response process is done in two phases. In the primary phase, the original information is being randomized and transfer to the receiver side. In the secondary phase, the receiver reconstruct the original data from randomized data by distribution reconstruction algorithm. The approach is shown in figure given below:
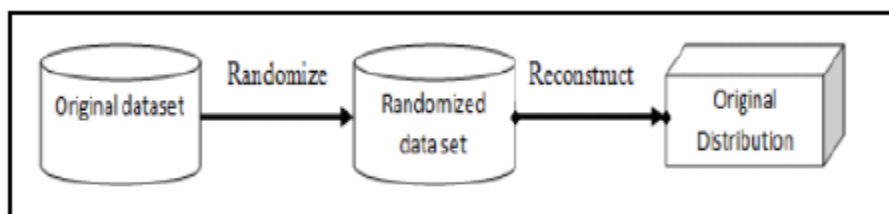
Perturbation approach: The perturbation approach modified the normal information values with synthetic information values, in order that the data computed from the perturbed data does now not distinguish from the know-how computed from original data. The perturbation approach are of two type.

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.

Condensation approach: Condensation method constructs restricted clusters in dataset after which generates pseudo. Knowledge from the information of these clusters. It is known as condensation because of the sooth that of its strategy of applying condensed facts of the clusters to generate pseudo data. It creates units of multiple size from the data, such that it is definite that each and every record lies in a suite whose size is at least alike to its anonymity level. Evolved, pseudo knowledge are generated from each and every set so that you can create a synthetic information set with the equal mixture distribution as the designated information. This approach can also be simply used for the classification hindrance.

Cryptography approach: Cryptographic procedures are ideally meant for such situations the place multiple parties collaborate to compute outcome or share non sensitive mining outcome and thereby averting disclosure of touchy knowledge. Cryptographic procedures to find its utility in such situations given that of two motives: First, it offers a well-defined model for privateness that includes methods for proving and quantifying it. Second, a large set of cryptographic algorithms and constructs to put in force privacy preserving data mining methods are to be had on this area. The information could also be distributed among special collaborators vertically or horizontally.

IV. CONCLUSION

The major function of privacy preserving data mining is developing methods to cover or provide privacy to specific sensitive information so that they can't be revealed to unauthorized one or intruder. Despite the fact that a privateness and accuracy in case of data mining is a pair of ambiguity. Succeeding possible result in opposed outcomes on other. On this, we made a try to check a good quantity of current PPDM methods. Sooner or later, we conclude there does now not exists a single privacy preserving knowledge mining algorithm that outperforms 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 criterion

V.   REFERENCES

[1] P. Samarati, "Protecting respondent's privacy in micro data release", in IEEE Transaction on Knowledge and Data Engineering, 2001, pp.1010-1027.

[2] M.B. Malik, M.A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.

[3] 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.

[4] 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.

[5] Nabil Adam and John C. Wortmann, Security Control Methods for Statistical Databases: A Comparison Study, ACM Computing Surveys 21 (1989), no. 4, 515–556.

[6] R. Agrawal and R. Srikant, "Privacy-preserving data mining", In ACM SIGMOD, pages 439–450, May 2000/

[7] Y. Lindell and B. Pinkas, "Privacy preserving data mining", J. Cryptology, 15(3):177–206, 2002.

[8] L. Sweeney, "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY," Int. J.Uncertain., vol. 10, no. 5, pp. 557- 570, 2002.

[9] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy beyond k-anonymity," Proc. -Int. Conf Data Eng., vol. 2006, p. 24, 2006.

[10] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," Proc. 2007 ACM SIGMOD Int. Conf. Manag. data, pp. 665- 676, 2007.

[11] M. Prakash and G. Singaravel," An approach for prevention of privacy breach and information leakage in sensitive data mining", Computers and Electrical Engineering 2015.

[12] Bertino E, Fovino I, Provenza L. A framework for evaluating privacy preserving data mining algorithms. J Data Min Knowl Discovery 2005;11(2):121–54.

[13] Hina Vaghashia and Amit Ganatra, "A Survey: Privacy Preservation Techniques in Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 119 – No.4, June 2015.

[14] S.V. Vassilios , B. Elisa, N.F. Igor, P.P. Loredana, S.Yucel and T. Yannis, 2004, "State of the Art in Privacy Preserving Data Mining" Published in SIGMOD Record, 33, 2004, pp: 50-57.

[15] Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", International Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011.

[16] J. Jesu Vedha Nayahi and V. Kavitha," Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop",Future Generation Computer Systems, 0167-739X/© 2016 Elsevier.

[17] R. Rajeswari and Mrs R. Kavitha ,"Privacy Preserving Mechanism for anonymizing data streams in data mining", International conference on current research in Engineering Science and Technology(ICCREST-2016).

[18] Elahe Ghasemi Komishani, Mahdi Abadi, and Fatemeh Deldar," Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression", Future Generation Computer Science(2016). [19] Tong Li, Zheli Liu, Zin Li, Chunfu Jia and Kuan-Ching Li, "A Cryptographic Data Publishing System",J. Computer System Science(2016).

[19] Surbhi Sharma and Deepak Shukla, "Efficient multi-party privacy preserving data mining for vertically partitioned data",Inventive Computation Technologies (ICICT), 10.1109/INVENTIVE.2016.7824852, © 2017 IEEE

CITE AN ARTICLE