Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
SURVEY REPORT
Available Online at www.ijarcs.info

A Survey on Privacy Preserving Data mining

Nivedita Bairagi
Depatment of CSE & IT
MITS, gwalior, MP, India

Punit k. Johari
Depatment of CSE & IT
MITS, gwalior, MP, India

Abstract:Data mining objectives to seek out valuable patterns from big quantity of data. These patterns signify information and are conveyed in clusters or organization ideas. 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.

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

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 isoftentimes 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 concern over enormouscollection 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 microdata (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.

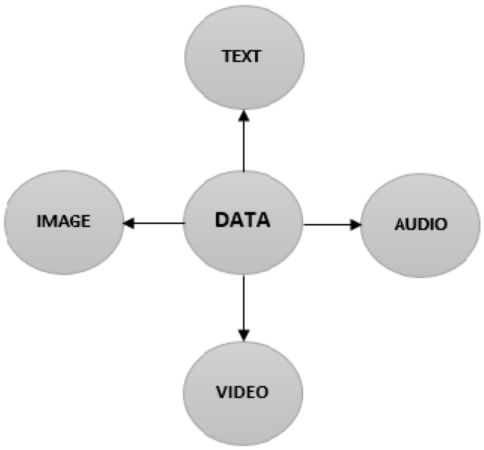

Fig. 1. Varieties of data

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.

PRIVACY P RESERVING DATA M INING FRAMEWORK

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.

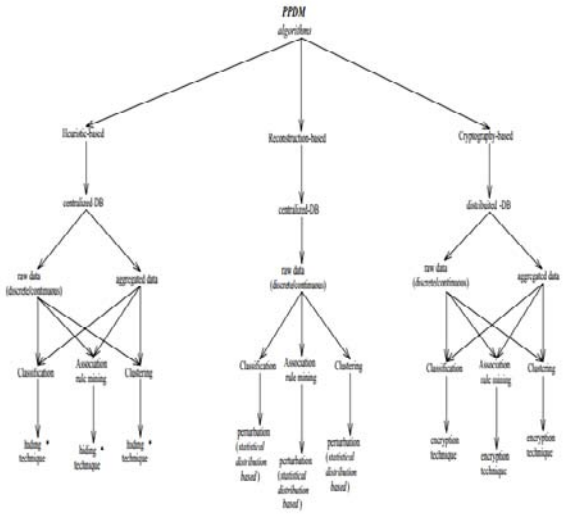

Fig.2. PPDM Framework [12]

PRIVACY PRESRVING DATA MINING 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–Cryptographyapproach 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 ba sed a pproach: 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.

Table 1. Original data

Age	Weight	Name
35	50	Ramesh
60	55	Shweta
65	50	Sham

Table 2. K-Anonymous data

Age	Weight	Name
[35,45]	[50,65]	Ramesh
[35,45]	[50,65]	Shweta
[55,65]	[50,65]	Sham

Here, Table 1 shows the original data which will be used for anonymization approach. Table 2 shows the k-anonymous data.

Randomization r esponse 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 fig 3.

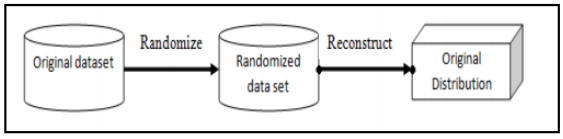

Fig.3 Randomized response approach [13]

Perturbation a pproach: 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 pe rturbation: 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 a pproach: 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.

Cryptographyapproach: 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.

LITERATURE REVIEW

Data anonymization is a promising process within the discipline of privacy preserving data mining used to protect the information in opposition to identity disclosure. Information loss and long-established attacks possible on the anonymized information are critical challenges of anonymization. Not too long ago, knowledge anonymization utilizing information mining strategies has showed gigantic improvement in information utility. Nonetheless the prevailing approaches lack in robust handling of attacks. As a result J. Jesu Vedha Nayahi et al. Proposed an anonymization algorithm established on clustering and resilient to similarity attack and probabilistic inference attack is proposed [16].

R. Rajeswari et al. Proposes a privacy persevered access control mechanism for data streams. For the privacy security mechanism it makes use of the combination of both the K-anonymity procedure and fragmentation system. The k-anonymity procedure makes use of the suppression and generalization. It prevents the privacy revelation of the sensitive information. The privacy defense mechanism avoids the identity and attributes disclosure. The privateness is executed by means of the high accuracy and consistency of the person expertise, i.e., the precision of the personal data. [17].

For addressing the drawback of identical privateness safety for all relocating objects in trajectory knowledge, Elahe Ghasemi Komishani et al. proposed PPTD, a novel process for keeping privateness in trajectory data publishing established on the concept of personalized privacy. They targets to strike a stability between the conflicting objectives of information utility and knowledge privacy in line with the privateness standards of relocating objects. They combines sensitive attribute generalization and trajectory nearby suppression to achieve a tailored personalized privacy model for trajectory data publishing. They performed experiments on two artificial trajectory datasets and concluded that PPTD is powerful for maintaining personalized privateness in trajectory information publishing [18].

The usual data publishing ways will do away with the sensitive attributes and generate the considerable records to attain the goal of privacy safety. . In the big data environment, the requirement of using information (e.g., data mining) come to be more and more quite a lot of, which is beyond the scope of the normal procedure. Tong Li et al. Presents a cryptographic data publishing system that preserves the information integrity (i.e., the long-established knowledge structure is preserved) and achieves anonymity without deletion of any attribute or utilization of redundancy. The safety analysis suggests that their process is secure underneath proposed security model [19].

Surbhi Sharma et al. Show how the exclusive departments of same group combine their data without harming the privateness of the client for making robust selections in efficient and correct manner. For that reason the approaches vertically information combination, cryptography and decision mining is established. To mine the choices from the information a C4.5 resolution tree is used. The implementation of the proposed privateness preserving data mining and decision making method is carried out using JAVA technology. Additionally the efficiency of the method is computed in phrases of accuracy, error rate, memory consumption and time consumption. In the end to justify the effects of the proposed data mining system the normal J4.5 tree utilizing WEKA instrument is used with same knowledge for comparative performance learn. The experimental results show the mighty performance and protection within the given privacy preserving procedure [20].

Table 3. A dvantages an d D isadvantages of P PDM Technique

Approaches	Advantages	Disadvantages
Anonymization based approach	This approach protects individual's identity while releasing sensitive information.	Linking attack. Information loss.
Randomization response approach	Simple technique. More efficient.	Not appropriate for several attribute database. More data loss.
Perturbation approach	Simple technique. Independent treatment of distinct attributes.	Distortion is the merely way to reconstruct the original value. Ambiguity in degree of equivalence of different records.
Condensation approach	Suitable for Pseudo-data. Better approach than modification in original data.	Pseudo-data have same format as the original data.
Cryptography	Well suitable	Scaling is difficult

approach	approach. Provide vast toolset for protecting sensitive information.	when more parties are involved.

Table 4. Evaluation dimensions of PPDM Techniques

Efficiency	Capability to execute all to be had assets with good performance.
Scalability	Evaluates the efficiency for rising data size.
Data quality	Comprises accuracy, completeness and consistency.
Hiding failure	Obtain zero hiding failure of sensitive information.
Privacy level	Define degree of uncertainty.

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 atry 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 algorithm may perform better than a further on one exact criterion.

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).

[20] 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.