JCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

# Review paper on Privacy Preserving Data Analysis

**Yuvraj Singh[1*], Pankaj Pratap Singh[2], Anirudh Tripathi[3], Amit kishor[4]**

[1,2,3,4]CSE Department, Swami Vivekanad Subharti University, Meerut, India

*Corresponding Author: yuvrajsinghsolank03@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i6.11351138 | Available online at: www.ijcseonline.org

Accepted: 21/Jun/2019, Published: 30/Jun/2019

***Abstract***- Privacy-Preserving Data Mining (PPDM), as an important branch of data mining and an interesting topic in privacy preservation, has gained special attention in recent years. In addition to extracting useful information and revealing patterns from large amounts of data, PPDM also protects private and sensitive data from disclosure without the permission of data owners or providers. In recent years, privacy preserving data mining has been studied extensively, because of the wide proliferation of sensitive information on the internet. The major area of concern is that non-sensitive data even may deliver sensitive information, including personal information, facts or patterns. K-anonymity is a property that models the protection of released data against possible re-identification of the respondents to which the data refers. Anonymization approach makes the data owners anonymous but vulnerable to attacks like linking attacks. The paper presents various techniques which are used to perform PPDM technique and also tabulates their advantages and disadvantages.

***Keywords*** - Anonymization, Privacy Preserving Data Mining, k-anonymity, Randomization

## I. INTRODUCTION

The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. A number of techniques such as randomization and k-anonymity have been suggested in recent years in order to perform privacy-preserving data mining[1,3]. Furthermore, the problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community. In some cases, the different communities have explored parallel lines of work which are quite similar.

K-anonymity is one of the well-known anonymizing approaches proposed by Samarati and Sweeney [16]. We can say that if a data complies with k-anonymity protection if each individual's record stored in the released data set cannot be distinguished from at least k-1 records whose data also appear in the data set. For the checking the kanonymity requirements use the generalization and suppression methods are used for different datasets [1]. Now we conclude some basic thing about that how to we preserve the privacy in the k-anonymity model for that we using the PPDM techniques. There is many privacy preserving techniques available in the data mining. In all that technique k-anonymity is the one of PPDM technique What is the PPDM the basic thing is the extend traditional data mining technique to work with the modified data for hiding the sensitive attributes there are two approach is available one is the SMC and second is the Anonymization. In the secure multiparty computation it use the cryptographic approach for privacy preservation goal to create methods for parties to jointly compute a function over there inputs while keeping those input privates. In the anonymization some data is replace with the some modified related attribute from the overall data base.

The usage of randomization for preserving privacy has been studied extensively in the framework of statistical databases [9, 10,12]. In that case, the server has a complete and precise database with the information from its clients, and it has to make a version of this database public, for others to work with. One important example is census data: the government of a country collects private information about its inhabitants, and then has to turn this data into a tool for research and economic planning. However, it is assumed that private records of any given person should not be released nor be recoverable from what is released. In particular, a company should not be able to match up records in the publicly released database with the corresponding records in the company's own database of its customers. In the case of statistical databases, however, the database is randomized when it is already fully known. This is different from our problem, where the randomization procedure is run on the client's side, and must be decided upon before the data is collected. A randomization for a statistical database is usually chosen so that it preserves certain aggregate characteristics (averages and covariance matrices for numerical data, or marginal totals in contingency tables for categorical data), or changes them in a predetermined way

[12; 15]. Besides randomization, other privacy preserving transformations are used such as sampling and swapping values among records [15].

## II. 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].

## III. RELATED WORK

Many Privacy preserving techniques were developed, but most of them are based on anonymization of data. The list of privacy preservation techniques is given below.

- K anonymity
- L diversity
- T closeness
- Randomization
- Data distribution

**Anonymization**

Anonymization is the process of modifying data before it is given for data analytics [11], so that de identification is not possible and will lead to K indistinguishable records if an attempt is made to de identify by mapping the anonymized data with external data sources. K anonymity is prone to two attacks namely homogeneity attack and back ground knowledge attack. Some of the algorithms applied include, Incognito [12], Mondrian [13] to ensure Anonymization. K anonymity is applied on the patient data shown in Table 1. The table shows data before anonymization.

**Table 1 Patient data, before anonymization**

| Sno | Zip | Age | Disease |
|---|---|---|---|
| 1 | 57677 | 29 | Cardiac problem |
| 2 | 57602 | 22 | Cardiac problem |
| 3 | 57678 | 27 | Cardiac problem |
| 4 | 57905 | 43 | Skin allergy |
| 5 | 57909 | 52 | Cardiac problem |
| 6 | 57906 | 47 | Cancer |
| 7 | 57605 | 30 | Cardiac problem |
| 8 | 57673 | 36 | Cancer |
| 9 | 57607 | 32 | Cancer |

K anonymity algorithm is applied with k value as 3 to ensure 3 indistinguishable records when an attempt is made to identify a particular person's data. K anonymity is applied on the two attributes viz. Zip and age shown in Table 1. The result of applying anonymization on Zip and age attributes is shown in Table 2.

**Randomization technique**

Randomization is the process of adding noise to the data which is generally done by probability distribution [21]. Randomization is applied in surveys, sentiment analysis etc. Randomization does not need knowledge of other records in the data. It can be applied during data collection and pre processing time. There is no anonymization overhead in randomization. However, applying randomization on large datasets is not possible because of time complexity and data utility which has been proved in our experiment described below. We have loaded 10k records from an employee database into Hadoop Distributed File System and processed them by executing a Map Reduce Job. We have experimented to classify the employees based on their salary and age groups. In order apply randomization we added noise in the form of 5k records which are randomly added to make a database of 15k records and following observations were made after running Map Reduce job.

- More number of Mappers and Reducers were used as data volume increased.
- Results before and after randomization were significantly different.
- Some of the records which are outliers remain unaffected with randomization and are vulnerable to adversary attack.
- Privacy preservation at the cost of data utility is not appreciated and hence randomization may not be suitable for privacy preservation especially attribute disclosure.

## IV. RESULTS AND DISCUSSION

As part of systematic literature review, it has been observed that all existing mechanisms of privacy preservation are with respect to structured data. More than 80% of data being generated today is unstructured. As such, there is a need to address following challenges.

**Table 2 After applying anonymization on Zip and age**

| Sno | Zip | Age | Disease |
|---|---|---|---|
| 1 | 576** | 2* | Cardiac problem |
| 2 | 576** | 2* | Cardiac problem |
| 3 | 576** | 2* | Cardiac problem |
| 4 | 5790* | > 40 | Skin allergy |
| 5 | 5790* | > 40 | Cardiac problem |
| 6 | 5790* | > 40 | Cancer |
| 7 | 576** | 3* | Cardiac problem |
| 8 | 576** | 3* | Cancer |
| 9 | 576** | 3* | Cancer |

i. Develop concrete solution to protect privacy in both structured and unstructured data.

ii. Scalable and robust techniques to be developed to handle large scale heterogeneous data sets.

iii. Data should be allowed to stay in its native form without need for transformation and data analytics can be carried out while ensuring privacy preservation.

iv. New techniques apart from Anonymization must be developed to ensure protection against key privacy threats which include identity disclosure, discrimination, surveillance etc.

v. Maximizing data utility while ensuring data privacy.key which is used for encryption is to used for decryption. Thus as long as the symmetric key is kept confidential by both the communicating parties can be sure that they are in communication with the authenticated party [8].

## V. CONCLUSION

No concrete solution for unstructured data has been developed yet. Conventional data mining algorithms can be applied for classification and clustering problems but cannot be used in privacy preservation especially when dealing with person specific information. Machine learning and soft computing techniques can be used to develop new and more appropriate solution to privacy problems which include identity disclosure that can lead to personal embarrassment and abuse. There is a strong need for law enforcement by governments of all countries to ensure individual privacy. One of the serious privacy threats is smart phone. Lot of personal information in the form of contacts, messages, chats and files are being accessed by many apps running in our smart phone without our knowledge. Most of the time people do not even read the privacy statement before installing any app. Hence there is a strong need to educate people on the various vulnerabilities which can contribute to leakage of private information.

## REFERENCES

[1] Agrawal R., Srikant R. Privacy-Preserving Data Mining. ACM SIGMOD Conference, 2000.

[2] K David Raju, L Vijay Kumar, K Anthony Rahul Showry, B LhoitKrishn ,(2018) ." techniques of providing data integrity in cloud computing".

[3] Aggarwal C. C., Yu P. S. On Variable Constraints in Privacy Preserving Data Mining. ACM SIAM Data Mining Conference, 2005.

[4] Alexandre Evfimievski, Tyrone Grandison, “Privacy Preserving Data Mining”.

[5] Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke. “Privacy preserving mining of association rules”. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-25, ACM Press, Edmonton, AB., Canada, pp. 1-12,2002.

[6] 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.

[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] G. T. Duncan and S. Mukherjee. Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise. Journal of the American Statistical Association, 95(451):720–729, 2000.

[10] T. Evans, L. Zayatz, and J. Slanta. Using noise for disclosure limitation of establishment tabular data. Journal of Official Statistics, 14(4):537–551, 1998.

[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] S. E. Fienberg, U. E. Makov, and R. J. Steele. Disclosure limitation using perturbation and related methods for categorical data. Journal of Official Statistics, 14(4):485–502, 1998.

[13] Mahima Joshi, Yudhveer Singh Moudgil,"secure Cloud Storage."

[14] Yogendra Kumar Jain, Vinod Kumar Yadav& Geetika S. Panday, An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining, International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 7 July 2011.

[15] J. J. Kim and W. E. Winkler. Masking microdata files, 1995.

[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).