DOI: http://dx.doi.org/10.26483/ijarcs.v13i5.6906

Volume 13, No. 5, September-October 2022

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

REVIEW PAPER

Available Online at www.ijarcs.info

A STUDY ON PRIVACY PRESERVING BIG DATA MINING: TECHNIQUES AND CHALLENGES

Anuradha
Assistant Professor of Computer Science
Kanya Mahavidyalaya Kharkhoda
Sonipat,India

Abstract: The basic goal of data mining algorithms is to extract previously undiscovered patterns from the data. When mining the data, sensitive and confidential information should be secured simultaneously to protect privacy. Due to the widespread use of information technology, enormous amounts of data are being produced at an exponential rate by several organisations, including hospitals, insurance providers, banks,e-commerce, and stock exchanges, making privacy a crucial concern in data mining. Anonymization, Perturbation, Generalization, and Cryptography are some of the privacy-preserving data mining techniques that have been proposed in the literature. In this study, we have reviewed all of these state of art techniques and presented a tabular comparison of work done by different authors as well as discussed the challenges of privacy preserving data mining.

Keywords: Big data, Data mining, PPDM, Anonymization, Cryptography, Perturbation.

1.INTRODUCTION

Big data is a collection of information that comes from numerous, heterogeneous, autonomous sources. Every second, more data is being collected from many websites. The data expands in size as a result. Social networking sites generate a lot of data. In 2010, **Apache Hadoop** defined big data as “datasets which could not be captured, managed, and processed by general computers within an acceptable scope.” On the basis of this definition, in May 2011, **McKinsey & Company**, a global consulting agency announced Big Data as the next frontier for innovation, competition, and productivity[1]. Big data sets that are commonly available greatly enhance our knowledge, productivity, and services in a variety of societal sectors. For example,big GPS data sets can be used to find the best route for going to a particular location and big medical data sets helps in finding the best treatment for a particular disease[2].Extraction of knowledge from the data is referred to as "data mining". Data analysis and application-oriented pattern extraction are the main goals of data mining. Data mining is frequently used interchangeably with the phrase "knowledge discovery from data (KDD)”. However, other individuals see data mining as just one phase in the process of knowledge discovery. The knowledge discovery process has following steps: 1. Data cleaning: is used to eliminate ambiguous and noisy data 2. Data integration: in this phase various sources of data is combined 3. Data selection: here information from the database that is pertinent to the analysis activity is retrieved 4. Data transformation: here data is transformed into form that is appropriate for mining 5. Data mining: extract useful information from the data 6. Patternevaluation: used to determine the highly interesting patterns of information using measurements of interestingness7. Knowledge presentation: present mined knowledge to user using visualization tools[3].

Data mining—searching for knowledge (interesting patterns) in data. [3]

People are very hesitant to share their sensitive information because they are fully aware of the privacy violations of their personal data. Due to this, data mining may produce unintended results.So privacy preserving data mining(PPDM) is introduced.Data mining that protects privacy with maintaining the usefulness of the data is known as privacy-preserving data mining (PPDM). Privacy-preserving data mining (PPDM) strategies are ways to extract knowledge from data with protecting individual privacy [4].Big data mining requires a robust mechanism for protecting privacy because the rate of data generation is increasing quickly and the data is unstructured, making it impossible to handle it using conventional systems. We describethe concepts of data mining and privacy preserving big data mining in introduction Section.Insection 2 we discuss various techniques for preserving privacy when mine large data sets.Following that, a tabular comparison of several PPDM methods given by various authors is displayed in section 3. In the section 4 we discussvarious challenges of PPDM. And finally in section 5 we conclude.

2.PRIVACY PRESERVING DATA MINING(PPDM) METHODS

Privacy is divided into four categories from the data mining perspective. These categories are as follows: 1. Before data mining ensure privacy at the time of data collection from various sources 2. Ensuring privacy of data before publishing data for analysis. 3. Ensuring privacy after data mining process. 4. Ensuring privacy of data when distributing data [4].Here we describe various developed techniques of privacy preserving of data, such as Data anonymization based, cryptographic based,and Data perturbation based.

2.1 Data anonymization based approach: Data is transformed using anonymization-based PPDM techniques to ensure privacy by changing more specific data with less specific data through generalisation, Character Replacement, Noise Addition, Shuffling, and Suppression.These anonymous data sets supplied as data mining input.Substituting a value with more general value is called generalization for example, change the attribute “date” (month/day/year) with attribute “year”. In Shuffling, the data is mixed or restructured at random, but the original attribute values are kept in the dataset. Shuffling is useful when we are analysing only one attribute we do not need to relate it with other attributes. Noise addition gently alternate the attributes makes them less accurate. Character replacement hide specific part of attributes values with predefined symbol.Suppression blocks an attribute that is not relevant for current analyse[5]. Figure1 shows anonymization using character replacement.

Before Anonymization

Roll number	Marks
10010111	65%
20010222	78%
30010333	56%

After anonymization using character replacement

Roll number	Marks
100*****	65%
200*****	78%
300*****	56%

Figure1: character replacement

Privacy preservation techniques based on anonymization are K anonymity, L diversity, T closeness

K anonymity: Latanya Sweeney and Pierangela Samarati [6]introduced the concept of k-anonymity by using generalization and Suppression.K anonymity links anonymized data to other data sets to address the possibility of re-identification.Using this method, K individual' records are grouped so that they all come under the same combinations. Group of these k undistinguishable records is known as equivalence class.In k anonymized dataset it is difficult to determine the identity of a record because there exitk-1 similar records. The value k in the k-anonymity is used as a privacy indicator; the greater the value of k, it is more difficult to de-anonymize records. The possibility of finding a particular record is 1/K or less likely. It is vulnerable to two types of threats: background knowledge attacks and homogeneity attacks [2]. Ajmeera kiran et al.[7]proposed aprivacy preserving Multidimensional Geometric data perturbation approach using data anonymization that alter sensitive values in a dataset using K-anonymization and work with both type of datanumerical as well as categorical. This proposed approach capable to achieve data accuracy and privacy while reducing data loss compared to existing z-score or Gaussian noise normalisation methods.Suman Madan et al.[8]proposed anadaptive Dragon-PSO algorithm based K-anonymization model for privacy preservation of sensitive data during data publishing in cloud environment.The suggested method creates a fitness function for the suggested adaptive Dragon-PSO algorithm to maximise utility and privacy.

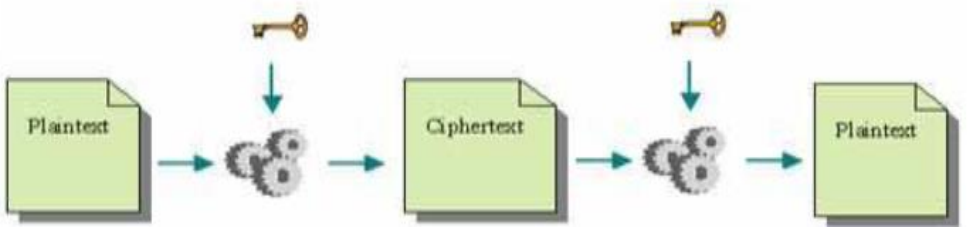
L diversity has been introduced by Machanavajjhalaet al. [9]to overcomes the problems of k anonymity. L diversity requires that every equivalence class hold at least l "well-representing" values for the sensitive identifier, which enhances the k anonymity. l -diverse equivalence class meets k-anonymity with k = l includes at least l records (because l distinct values are necessary).All currently existing equivalence classes must be l-diverse in order for a table to be l-diverse. This method aims to reduce the appearance of equivalence classes having minimal attribute variability. As a result, there is always some degree of doubt for an intrusion who seems to have access to records for a certain person. l-diversity is also vulnerable to skewness threats. It does not account for how the sensitive data are distributed, which can result in privacy violations when the distribution of the sensitive data is skewed. Brijesh B.Mehta et al.[10] proposed improved scalable l-diversity approachimplement using apache pig and python as an extension to improved scalable k-anonymization. This scalable approach preserves privacy of sensitivity information in big data mining. Odsuren Temuujin et al.[11] proposed L-Diversity Algorithm for Preserving Privacy ofdynamically modified datasets achieve anonymity using Anatomy instead of generalization or suppression. Proposed method use a new probabilistic data structure Cuckoo filter was specifically efficient, dramatically minimising operation

execution times while preserving privacy of dynamically changing datasets.

T closenessintroduced by Ninghui Li et at.[12] to address the shortcomings of the current k-anonymity and l-diversity methods. T closenessprotects privacyby producing the difference between the sensitive value distribution in each equivalence class and sensitive value distribution in main tableby less than threshold t.In other words, t-closeness principle states that distance between a sensitive attribute distribution in the original table and the same attribute distribution in any equivalence class is less than equal to t.All equivalence classes of a table must be *t*-closeness in order for table to be *t*-closeness [4]. Debaditya Roy et at. [13] proposed a scheme that utilized the multiple sensitive attributes of gave dataset to determine t. The proposed method only works with datasets that have multiple sensitive attributes; it is not applicable to datasets that only have one sensitive attribute.

2.2 Cryptography based approach: Cryptography is a method of encrypting sensitive data. Cryptography-based methods use cryptographic techniques to perform data mining. Public Key Encryption, Secret sharing, Homomorphic Encryption, Secure Multi-party computation **(**MPC), etc. are cryptography based techniques for preserving privacy in distributed data mining[14].

Public Key Encryptionuses two different keys (public and private) for encryption and decryption of data. Public key (known to everyone) is used for encryption of plaintext and private key (known to only receiver) is used for decryption of ciphertext. These two keys are related to each other mathematically. Rivest-Shamir-Adelman (RSA) Data Security algorithms are used commonly to implement public key cryptography [15].



[15]

Encryption with public key Decryption with private key

Secret sharing method distributes a secret to a group in such a way that no one holds sensual information about the secret, but if all individuals combine their 'share' secret is recreated. Secret sharing method is a set of two functions sharing function and relation function. Sharing function takes secret as an input and divides secret into n secret shares and relation function reproduce (recovery of secret from n-1 shares is impossible) secret from n shares of secret [14].

Homomorphic Encryption performsalgebraic operation on ciphertext in such a way that discovered result would be same to the result of operation performed with the plaintext from that ciphertext is generated [4]. Traditional encryption methods decrypt data into its original form before doing computation. As a result,privacy of sensitive information may be violated but homomorphic encryption allows to compute encrypted data and results are also provided in encrypted form to user. Partial homomorphic encryption, somewhat homomorphic encryption, and Fully homomorphic encryption are schemes of homomorphic encryption. Partially homomorphic systems were only able to implement certain algebraic operations over the ciphertext.Somewhat homomorphic systems can perform arbitrary operations but only a limited number of times. Fully homomorphic systems allow any arbitrary function to be applied to the ciphertext[16].

Secure Multi-party computation(MPC)method provides various protocols that allow multiple collaborative untrusted parties (Only required information access is allowed to these parties) to calculate the function collectively using their inputs at the same time maintaining individual sensitive data private[17]. MPC has been widely applied in recent years to preserve privacy in distributed data mining because it is more efficient and effective than other strategies. MPC support linear complexity for arithmetic, fixed-point, and floating-point operations but risk of privacy leakage occur when statistical calculations are performed on different types of data and study the relationship between variables using linear regressionto address these problems Jun Liu et at. [17] proposed solution basedon matrixcomputationwithone-hotencodingandLUdecomposition by utilizing MP-SPDZ secure multi-party computation protocol for implementation.

2.3Data perturbation based approach: Data perturbation replace original values with some artificial data values so that statisticallycomputed information from perturbed data is not significantly different from statistically computed information from original data. Perturbation of data is performed with the help of data swapping, additive noise, and data Masking [14].Data perturbation approach classified into probability distribution and value distortion [18]. In probability distribution approachdata is replaced with the

distribution itself or with other sample taken from same distribution, and in value distortion approach data elements are perturbed by randomization procedures (either multiplicative noise, additive noise, or some other). Some existing perturbation methods include Additive perturbation, multiplicative perturbation,geometric perturbation, Projection Perturbation, data condensation, and rotation perturbation[18][19][20].Additive perturbationconcentrates on changing a single sensitive attribute of statistical databases. In this perturbation noise (amount of noise is independent to sensitive attribute) is added to sensitive attributes to generate perturbed data.The main weakness of additive perturbation is that by utilizing Spectral properties of data sensitive data can be separated from noise.Multiplicative Perturbation multiply noise to sensitive attribute. Data condensation is a multi-dimensional perturbation approach perturbs multiple columns to generate perturbed dataset and also preserve covariance matrix to multiple columns. In data condensation existing data mining algorithms can be apply for mining the perturbed dataset without requiring modification because it preserves covariance matrix.Rotation perturbation is also a multidimensional perturbation method that rotates valueseither clockwise or anti-clockwise by multiplying original data with orthonormal rotation matrix. In projection perturbation a set of data points are projected from multidimensional space to other randomly selected space. Geometric Data Perturbation(GDP) method combines three perturbation techniques: rotation, translation, and distance. GDP includes sequence of random geometric transformations that use multiplicative perturbation, translation transformation, and noise addition based on distance perturbation. Alpa Shah et al. [19] presents comparative analysis of performance of Additive, Multiplicative and Geometric Data Perturbation using various statistical metrics like Mean, Mean Square Error, Standard Deviation, Root Mean Square, and Mean Absolute Error on two datasets Adult and Breast Cancer-w (both of them have large number of records) from UCI Repository.

Table 1

Technique	Advantages	Problems
Anonymization Based Approach	Sensitive information of user is secured	Suffer from heavy data loss and linking attack
Cryptography Based Approach	Transformed data are accurate and secure	Complex to scale when multiple parties are involved
Perturbation Based Approach	Preserves different attributes independently	Data loss, chancesof adversarial attack,and original data cannot be regenerated

2.4 Hybrid approach: Various techniques for preserving the privacy of sensitive information are developedamong themadvantages and problems of some are analysed in table1. There seems to be no single technique that is reliable in all domains so some hybrid techniques based on a combination of two or more than two techniques to ensure the privacy are developed.Akash Siddhpura et al. [21] proposed a hybrid approach using perturbation and cryptography for privacy preservation and also maintain data utility. In this approach first data is converted into their asci values then perturbation on data is applied after that cryptography using ECB algorithm is applied on perturbed data and for achieving original data back reverse process is performed. Many other techniques, such as data perturbation, blocking-based methods, cryptographic techniques, and data anonymization, can also be combined to create a hybrid technique.

3.Comparison between different techniques

Many different techniques for protecting sensitive information (private data) have been proposed in the field of Privacy Preserving Data Mining.Performances of different techniques are differing according to criteria. One technique can outperform over other technique or vice-versa depends on different criteria. In below table we shows a tabular comparison of the work done by various authors in a time scale(from past to present). We used parameters such as the PPDM technique, Year of publication, dataset used, and performance of technique and research gaps for comparison.

Table 2: Tabular comparison of different PPDM techniques

S.NO	Authors	Year of Publication	PPDM Techniques	Datasets	Approach and Performance	Research Gaps
1	Jaideep Vaidyaet al.[22]	2014	randomization and cryptographic	Mushroom, Image Segmentation, Nursery, and Car from UCI repository	The authors develop methods to safely construct random decision trees for both vertically and horizontally partitioned datasets.Proposed solution enhancesboth efficiency and security.	Future work will focus on developing general solutions for arbitrarily partitioned data.
2	Rupinder	2015	Transformation	Adult	The authors proposed a	Performance of

	Kaur et al.[23]			data(30,718 instances and 9 attributes)	method topreserve the privacy of sensitive Boolean attributesusing transformation techniques. Result shows that there is no information loss in transformed data set.	Proposed method needs to be tested against other data mining tasks and can be expanded to horizontally and vertically partitioned data.
3	A S M Touhidul Hasan et al.[24]	2016	Anonimization using slicing and swapping	Synthetic dataset (10,000 instances and 4 attributes) and adult dataset (45,222 instances and 8 attributes)	The suggested method uses value swappingmechanism to enhance slicing based anonymization to ensure that the released microdata table supports l-diverse slicing.Compared withexisting slicing methods, it hasbetter data utility and smaller relative query error.	In future research, we will try to make the swapping algorithm faster and reduce the identity disclosure risk from sliced table with maximizing data utility and privacy.
4	K. Abrar Ahmed et al.[25]	2017	fuzzy data modification(FDM) and Random Rotation Perturbation(RRP)	Adult dataset (10,000 instances, 14 attributes) from UCI repository	In proposed approach original data is anonymized using combination of FDM and RRP. FDM includes fuzzy K-member clustering together with membership function to be performed to distorted data and geometric structure on dataset is preserved using RRP. This approach has optimal information loss.	The goal of the next improvement is to find more methods to use in combination with FDM to distort textual and numerical data.
5	Akash Siddhpura et al.[21]	2018	Perturbation and Cryptography	Indian Liver Patient, Balance Scale,Ablon, andBank Marketing datasets from machine learning, UCI	Protect sensitive data as well as reduce data loss by apply first data perturbation on asci values of data then perform cryptography on perturbed data.	We can also protect sensitive information in audio and video data.
6	Guang Li et al.[26]	2018	Perturb data using hybrid (SVD+NMF) matrix decomposition	Two real data set WBC (Breast Cancer Wisconsin) and Iris from UCI repository	Proposed approach applies first Non-negative matrix factorization(NMF) for perturbing the data, then data are again perturbed using Singular value decomposition (SVD) for finding and deleting sensitive data from datasetto create a new dataset for data mining task.	Future research will focus on several privacy protection algorithms to find the best possible combination to maximize privacy protection.
7	Odsuren Temuujin et al.[11]	2019	L-diversity model and anatomy technique of anonymization	Real census data taken from IPUMS, USA	Preserve Privacy of dynamically modified datasets using Anatomy instead of generalisation and suppression. This method processed data more efficiently compare to other traditional methods.	In future research, we will concentrate on more strong data anonymization model, like t-closeness.
8	Ajmeera kiran et al.[27]	2019	perturbation based on random data Swapping	Adult Dataset(48,842 instances, 6	Proposed approach perturbs multiple column in one iteration. Thismethod offers	Future research willfocuses on more strong data

				attributes)	accuracy, reduced error rates with minimal information loss, and the ability to produce original data from perturbed data compared to existing methods.	transformationand normalization techniques as well as preserve privacy forthe data stream.
9	Devendrasinh Vashi et al.[28]	2020	Cryptography hybrid approach using (RSA+DES)	Medical dataset	Proposed approach horizontally partitioned database into two tables and apply different encryption method on different table. After thatmerge encrypted tables and perform data mining. Performance of the hybrid approach is best comparing to using the single encryption technique either RSA or DES.	For vertically partitioned database a new hybrid approach of cryptography can be developed.
10	Jun Liu et al. [17]	2020	Secure multi-party computation(MPC)	Collect vehicle driving logs and Cotton trading records for experiments and choose wine quality dataset for performance comparison with other	Authors proposed solution based on matrix computation with one-hot encoding and LU decomposition for supporting the tasks of statistical calculation of different types of data and relationship between variables studying using linear regression in MPC.	Implementation of proposed work could be improvedusing graphic processing unit acceleration and work will be extended to more machine learning task beyond simple statistical computation.
11	N. Kousika et al. [29]	2021	3D rotation data perturbation and singular value decomposition (SVD)	Adult dataset (48,842 instances and 14 attributes), heart disease (303 instances and 14 attributes) from UCI repository	Firstly, the original data is perturbed by SVD that take out sensitive attributes by using matrix decomposition and remove extraneous data than 3D RDP is repeatedly used to make sure all sensitive attributes are skewed to retain privacy along various axes. Here both data utility and privacy are balanced more effectively.	To further improve the classification accuracy, other perturbation approaches can be applied.
12	Thanveer Jahan et al.[30]	2021	Geometric Data Perturbationwith Euclidean distance preservation	Iris dataset (3 features and 50 instances), Water Treatment Plant dataset (8 features and 527 instances), Stone Flakes Dataset(8 features and 79 instances) from UCI	first perturb data using traditional geometric data perturbation and apply Euclidean distance preservation then sampling is performed on perturbed data using different interval basedon privacy levels. This method reduces the trade-offs that exist between accuracy,privacy levels, and information loss.	----------
13	Suman Madan et al.[8]	2021	K anonymization	Adult data set (48,842 instances, 14attributes)	proposed an adaptive Dragon-PSO algorithm based K-anonymization model for privacy preservation of sensitive data during data publishing in	Proposed method will be enhanced using multi-objective optimization method and will

					cloud environment. The suggested method creates a fitness function for the suggested adaptive Dragon-PSO algorithm to maximise utility and privacy.	be use larger dataset. Additionally, further improve to uncover the hidden patterns and correlations in the huge data.
14	Brijesh B.Mehta et al.[10]	2022	L diversity and k anonymity	Use synthetic big datasets generated bypoker dataset from UCI repository	Proposed improved scalable l-diversity approach implement using apache pig and python as an extension to improved scalable k-anonymization approach. Proposedmethod improve running time and lower information loss with same level of privacy compare to other approaches.	---------
15	Ajmeera kiran et al.[7]	2022	K anonymization	Adult data set(48,842 instances, 6 attributes)	This method alters sensitive values (multiple columns at a time) using k anonymization and capable to achieve data accuracy and privacy while reducing data loss compared to existing z-score or Gaussian noise normalisation methods.	This work can be expanded by incorporating additional transformation approaches, such as decimal scaling normalisation, to improve data accuracy while reducing data loss.

4. Challenges of PPDM

There areso many issues and obstacles ahead in terms of privacy preserving data mining. Depending on our current understanding, we describe the key ones here.

- Recently the amount of data generated from various heterogeneous sources is increasing day by day and these large amounts of real world data sets are modified dynamicallycreates challenges toexisting privacy preserving data mining techniques. Most of the traditional PPDM techniques preserve privacy of static data that remain unaffected after processing.Various techniques of preserving privacy of dynamically modified data sets are proposed among them some only work with incremental data updates.When large datasets are frequently modified, re-anonymization of entire datasets after modification is inefficient. So we still need a strong method for privacy preserving data mining of dynamically updated datasets [11].
- New technologies such as bioinformatics,wireless sensor networks, mash-up web applications,cloud computing, andlocation-based services enable users to gain access to previously unavailable knowledge and information. These technologies raise new privacy concerns, necessitating efficient privacy protection methods [8].Single machine does not have the capacity to store large amount of data so we use cloud for storage of large datasets. Privacy preservation of sensitive information in cloud will create new research challenges for us asour data is not under our control and we cannot believe on cloud services provider.
- In the field of PPDM, data privacy and data utility are a major trade-off. An increase in one of these leads to an observable decrease in the other. An improved privacy-utility trade-off in PPDM requires the development of stronger methods. Additionally, some practical-oriented solutions must be developed and implemented to speedily, effectively, and inexpensively maintaining privacy preservation of sensitive data in real-world problems[31].

5. Conclusion

Various organizations in the world regularly collect and store large amounts of data on individuals. These large amounts of datasets are used in various fields using data mining techniques. Performing data mining on these huge amount of data can lead to violation of privacy of sensitive and private information. As a result, Privacy Preserving Data Mining (PPDM) emerged as a crucial research area in the modern era.Various techniques of data mining have been developed for preserving privacy of sensitive information. An overview of privacy preserving data mining techniques based on anonymization, perturbation, cryptography, and hybrid along with a tabular comparison of work done by different authorsis presented here and the challenges in field

of privacy-preserving data mining is also discussed.The outcomes of this study show that privacy preservation of data when mine large datasets have potential challenges that encourage scholars to do additional research in the field of privacy preserving big data mining.Inthe end we concluded that no single technique is reliable across all domains. Depending on the type of data and the type of application, all methods work differently. Soin future we can do better using hybrid approach.

REFERENCES

[1] M. Chen, S. Mao, and Y. Liu, 'Big Data: A Survey', Mob. Netw. Appl., vol. 19, no. 2, pp. 171–209, Apr. 2014, doi: 10.1007/s11036-013-0489-0.

[2] S. Yu, 'Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data', IEEE Access, vol. 4, pp. 2751–2763, 2016, doi: 10.1109/ACCESS.2016.2577036.

[3] 'The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf'. Accessed: Jul. 06, 2022. [Online]. Available: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

[4] R. Mendes and J. P. Vilela, 'Privacy-Preserving Data Mining: Methods, Metrics, and Applications', IEEE Access, vol. 5, pp. 10562–10582, 2017, doi: 10.1109/ACCESS.2017.2706947.

[5] J. Marques and J. Bernardino, 'Analysis of Data Anonymization Techniques':, in Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Budapest, Hungary, 2020, pp. 235–241. doi: 10.5220/0010142302350241.

[6] P. Samarati and L. Sweeney, 'Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression', p. 19.

[7] A. Kiran and N. Shirisha, 'K-Anonymization approach for privacy preservation using data perturbation techniques in data mining', Mater. Today Proc., Jun. 2022, doi: 10.1016/j.matpr.2022.05.117.

[8] S. Madan and P. Goswami, 'Adaptive Privacy Preservation Approach for Big Data Publishing in Cloud using k-anonymization', Recent Adv. Comput. Sci. Commun. Former. Recent Pat. Comput. Sci., vol. 14, no. 8, pp. 2678–2688, Oct. 2021, doi: 10.2174/2666255813999200630114256.

[9] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, 'L -diversity: Privacy beyond k -anonymity', ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, p. 3, Mar. 2007, doi: 10.1145/1217299.1217302.

[10] B. B. Mehta and U. P. Rao, 'Improved l-diversity: Scalable anonymization approach for Privacy Preserving Big Data Publishing', J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 4, pp. 1423–1430, Apr. 2022, doi: 10.1016/j.jksuci.2019.08.006.

[11] O. Temuujin, J. Ahn, and D.-H. Im, 'Efficient L-Diversity Algorithm for Preserving Privacy of Dynamically Published Datasets', IEEE Access, vol. 7, pp. 122878–122888, 2019, doi: 10.1109/ACCESS.2019.2936301.

[12] N. Li, T. Li, and S. Venkatasubramanian, 't-Closeness: Privacy Beyond k-Anonymity and l-Diversity', in 2007 IEEE 23rd International Conference on Data Engineering, Apr. 2007, pp. 106–115. doi: 10.1109/ICDE.2007.367856.

[13] D. Roy and S. Jena, 'Determining t in t-closeness using Multiple Sensitive Attributes', Int. J. Comput. Appl., vol. 70, pp. 47–51, May 2013, doi: 10.5120/12179-8291.

[14] N. Nasiri and M. Keyvanpour, 'Classification and Evaluation of Privacy Preserving Data Mining Methods', in 2020 11th International Conference on Information and Knowledge Technology (IKT), Dec. 2020, pp. 17–22. doi: 10.1109/IKT51791.2020.9345620.

[15] D. Liestyowati, 'Public Key Cryptography', J. Phys. Conf. Ser., vol. 1477, no. 5, p. 052062, Mar. 2020, doi: 10.1088/1742-6596/1477/5/052062.

[16] K. Munjal and R. Bhatia, 'A systematic review of homomorphic encryption and its contributions in healthcare industry', Complex Intell. Syst., May 2022, doi: 10.1007/s40747-022-00756-z.

[17] J. Liu, Y. Tian, Y. Zhou, Y. Xiao, and N. Ansari, 'Privacy preserving distributed data mining based on secure multi-party computation', Comput. Commun., vol. 153, pp. 208–216, Mar. 2020, doi: 10.1016/j.comcom.2020.02.014.

[18] N. Patel and S. Patel, 'A Study on Data Perturbation Techniques in Privacy Preserving Data Mining', vol. 02, no. 09, p. 6.

[19] A. Shah and R. Gulati, 'Evaluating applicability of perturbation techniques for privacy preserving data mining by descriptive statistics', in 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Sep. 2016, pp. 607–613. doi: 10.1109/ICACCI.2016.7732113.

[20] K. Chen and L. Liu, 'Geometric data perturbation for privacy preserving outsourced data mining', Knowl. Inf. Syst., vol. 29, no. 3, pp. 657–695, Dec. 2011, doi: 10.1007/s10115-010-0362-4.

[21] A. Siddhpura and P. D. V. Vekariya, 'An approach of Privacy Preserving Data mining using Perturbation &

Cryptography Technique', Int. J. Future Revolut. Comput. Sci. Commun. Eng., vol. 4, no. 4, Art. no. 4, Apr. 2018.

[22] J. Vaidya, B. Shafiq, W. Fan, D. Mehmood, and D. Lorenzi, 'A Random Decision Tree Framework for Privacy-Preserving Data Mining', IEEE Trans. Dependable Secure Comput., vol. 11, no. 5, pp. 399–411, Sep. 2014, doi: 10.1109/TDSC.2013.43.

[23] R. Kaur and M. Bansal, 'Transformation approach for boolean attributes in privacy preserving data mining', in 2015 1st International Conference on Next Generation Computing Technologies (NGCT), Sep. 2015, pp. 644–648. doi: 10.1109/NGCT.2015.7375200.

[24] A. S. M. T. Hasan, Q. Jiang, J. Luo, C. Li, and L. Chen, 'An effective value swapping method for privacy preserving data publishing: An effective value swapping method for privacy preserving data publishing', Secur. Commun. Netw., vol. 9, Jul. 2016, doi: 10.1002/sec.1527.

[25] K. Abrar Ahmed, Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Chennai – 600017, Tamil Nadu, India, H. Abdul Rauf, and Sree Sastha Institute of Engineering and Technology, Chennai – 600113, Tamil Nadu, India, 'Privacy Preserving Data using Fuzzy Hybrid Data Transformation Technique', Indian J. Sci. Technol., vol. 10, no. 24, pp. 1–6, Jun. 2017, doi: 10.17485/ijst/2017/v10i24/114039.

[26] G. Li and R. Xue, 'A New Privacy-Preserving Data Mining Method Using Non-negative Matrix Factorization and Singular Value Decomposition', Wirel. Pers. Commun., vol. 102, no. 2, pp. 1799–1808, Sep. 2018, doi: 10.1007/s11277-017-5237-5.

[27] A. Kiran and D. D. Vasumathi, 'Data Mining: Random Swapping based Data Perturbation Technique for Privacy Preserving in Data Mining', DATA Min., vol. 8, no. 1, p. 15, 2019.

[28] D. Vashi, H. B. Bhadka, K. Patel, and S. Garg, 'An Efficient Hybrid Approach of Attribute Based Encryption For Privacy Preserving Through Horizontally Partitioned Data', Procedia Comput. Sci., vol. 167, pp. 2437–2444, Jan. 2020, doi: 10.1016/j.procs.2020.03.296.

[29] N. Kousika and K. Premalatha, 'An improved privacy-preserving data mining technique using singular value decomposition with three-dimensional rotation data perturbation', J. Supercomput., vol. 77, no. 9, pp. 10003–10011, Sep. 2021, doi: 10.1007/s11227-021-03643-5.

[30] T. Jahan, G. R. Reddy, K. Shekhar, and M. Swapna, 'Novel hybrid geometric data perturbation technique by means of sampling data intervals', Mater. Today Proc., Jul. 2021, doi: 10.1016/j.matpr.2021.06.420.

[31] S. A. Abdelhameed, S. M. Moussa, N. L. Badr, and M. Essam Khalifa, 'The Generic Framework of Privacy Preserving Data Mining Phases: Challenges & Future Directions', in 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), Dec. 2021, pp. 341–347. doi: 10.1109/ICICIS52592.2021.9694174.

