A Survey of Privacy Preserving Data Mining Approaches for Cyber Security

Qazaleh Sadat Mirhashemi[1], Mohammad Reza Keyvanpour[2]

[1]*Data Mining Laboratory, Department of Computer Engineering, Faculty of Engineering, Alzahra University, Tehran, Iran*
qzmirhashemi@gmail.com
[2]*Department of Computer Engineering, Faculty of Engineering, Alzahra University, Tehran, Iran*
keyvanpour@alzahra.ac.ir

Abstract

Recently, the amount of information is growing exponentially, so security and privacy protection have been a public concern for quite a long time. This data can be utilized in several fields, such as business, health care, and cybersecurity. Cyberspace is a virtual computer environment to ease online communication. Data mining applications can detect future cyber-attacks through analysis. Data mining techniques bring a number of privacy risks while also allowing users to access information that was previously hidden. However, there are various techniques and algorithms for data mining that preserve cyberspace and privacy for publishing data in data mining. These algorithms consist of perturbation and anonymization. In this paper, a framework has been developed for analyzing qualitative methods as a platform for data classification and evaluation based on the latest perspectives. Our aim is to present a systematic review of data dissemination methods to prevent cyber-attacks and privacy preserving data mining (PPDM) and provide a platform for qualitative comparison within this framework. Additionally, exposing existing method weaknesses is important for improving PPDM approaches and determining the appropriate methods according to the requirements of the fields to be studied.

Keywords: *Cyber Security, Privacy Preserving Data Mining, Data Publishing, Perturbation, Anonymization.*

1 Introduction

Combining and analyzing sensitive data from multiple sources offers considerable potential for knowledge discovery. However, there are a number of issues that pose problems for such analyses, including technical barriers, privacy restrictions, security concerns, and trust issues [1]. Cyber security is concerned with protecting computer and network

systems from corruption due to malicious software, including Trojan horses and viruses. Data mining has also proven a useful tool in cyber security solutions for discovering vulnerabilities and gathering indicators for baseline, as shown in fig.1 . Data mining is the process of identifying patterns in large datasets [2]. Data mining techniques are heavily used in scientific research as well as in business, mostly to gather statistics and valuable information to enhance customer relations and marketing strategies. Privacy preserving distributed data mining techniques (PPDDM) aim to overcome these challenges by extracting knowledge from partitioned data while minimizing the release of sensitive information. Numerous methods have been proposed in the field of privacy data mining [3]. These methods may lead to information loss or side effects, such as reducing the most recent classifications of perturbation and data usage. This article provides an overview of anonymization and perturbation techniques. Anonymization techniques prevent identifying the characteristics and identity of critical data to ensure privacy, while data perturbation techniques modify a piece of data or the entire dataset while maintaining the meaningful properties for creating data mining models [4]. This technique is chosen by some data owners since they do not want to expose their privacy. In the perturbation approach, there are two types of techniques: value-based perturbation and Multi-Dimensional Perturbation. It is a technique that maintains data privacy during data integration or before sending data to a data-mining program [5]. In this article, publishing techniques for privacy preserving data mining and a comprehensive overview of all the methods used so far are presented. We have attempted to provide the most comprehensive classification of the categories because in most articles available, only the steps of privacy preserving data mining are divided, whereas the following methods are not discussed. The aim of this article was to provide a brief description of each method, as well as a category for each method. This essay is organized as follows: After the introduction, a description of the research is provided in part two, followed by a classification of the various techniques in the next parts. The next part consists of evaluation criteria. Finally, at the end of the article, a list of the newest methods is provided based on recent research projects.

2 Related Works

Confidentiality is the main problem that arises in a large set of data. In this case, PPDM protects the privacy of data mining and its purpose is to achieve reliable data mining results without disclosing sensitive information. In [6], privacy techniques are classified into two categories, the Anonymous and the perturbation approach. After analyzing each approach, their significant features were identified, but only a small number were examined in the classification of methods. This paper provides an overview of existing privacy techniques, such as perturbation, anonymization and analyzes their strengths and weaknesses in various contexts [7],[8]. Privacy preservation is divided into the following types: Privacy preservation data mining (PPDM), Privacy preservation data publishing (PPDP), Privacy-preserving distributed data mining (PPDDM), and privacy

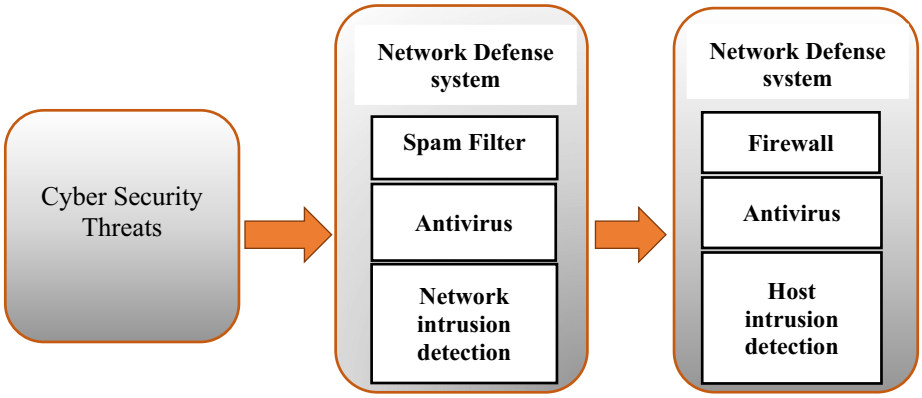

Figure 1: Conventional cyber security system [2]

preserving social network data publication (PPSNDP) [9]. Anushree Raj and Rio G.L. D'Souza have also said that two types of privacy attacks, called record linkage and attribute linkage, are prevalent. Several techniques are presented to preserve privacy [10]. According to Alpa Shah, the essence of PPDM lies in anonymization, perturbation, cryptography, fuzzy logic, and neural networks. Perturbation-based approaches have been discussed in [11]. It is also said that the data perturbation Approach is divided into two groups: the approach to probability distribution and the approach to value distortion. This paper aims to provide a complete and comprehensive classification of perturbation-based and anonymization-based algorithms according to the latest updates on data mining privacy. It also seeks to come up with an acceptable basis for more accurate classification and evaluation of data mining privacy techniques.

3 Algorithms and Techniques

To carry out data mining based on the desired results, complexity, and data properties, various used algorithms and techniques, including Association Regulation Learning, Classification, clustering, regression and outlier analysis [12] are explained in fig. 2.

3.1 Association Regulation Learning

This is also known as dependency modeling or market basket analysis. It is used to discover relationship rules and correlations between variables. The purpose of association rule mining algorithms is to identify relevant relationships between variables in a dataset [13].

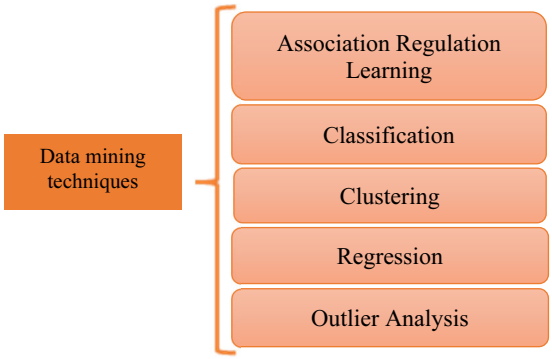

Figure 2: Various Data Mining Techniques

3.2 Classification

A classification algorithm is a powerful tool for analyzing and categorizing large amounts of data. It also enters data into a group belonging to a public class. It also enters data into a group belonging to a public class. This is also known as supervised classification [14].

3.3 Clustering

A clustering algorithm, known as Unsupervised classification, finds and creates groups of data elements that are similar [15]. Since each cluster can be considered as a class without a label, clustering is also known as automatic classification or classification that learns from observations rather than a training set [16].

3.4 Regression

A regression algorithm looks for a function that models the data with the fewest possible errors.Although algorithms have been created for a variety of tasks (such as clustering, association-rule mining, and classification), only two parties remain in the case of regression. The most interesting point about this algorithm is that it will keep the answer variable private [17].

4 PPDM Techniques for cyber security

Despite that, information discovered by data mining can be very valuable to many applications, people have shown increasing concern about the other side of the coin, namely the privacy threats posed by data mining [18].New strategies are discovered in PPDM to provide privacy for data mining knowledge. Furthermore, the process of knowledge discovery should not be impeded due to privacy. PPDM's major purpose

is to create algorithms that modify original data in several ways so that personal data remains private even after mining [19].

4.1 PPDM Techniques

There are several methods and a number of proposed conservation techniques to preserve privacy. Most techniques use some form of transformation in the main dataset to preserve privacy [20]. In this paper, the current techniques are divided into two general categories: perturbation and anonymization (fig.3).

Perturbation is a procedure for maintaining information confidentiality. This technique modifies the value of records without altering the significance of the input data Research shows that rotation disorders, projection disorders, and geometric data disorders are the three types of approaches to data disruptions [21]. The summary of these three types of data Perturbation is presented in Table 1.Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden.

4.2 Cyber security

Cyber security is the set of technologies and processes designed to protect computers, networks, programs, and data from attack, unauthorized access, change, or destruction. Cyber security systems are composed of network security systems and computer (host) security systems [22]. Data mining has also proven a useful tool in cyber security solutions for discovering vulnerabilities and gathering indicators for baseline. In this paper, we will focus on privacy preserving Data mining approaches for cyber security.

4.2.1 Value-based Perturbation

This will be explained in detail to preserve statistical characteristics and column distributions. Value-based Perturbation is caused by:

Table 1: Summary of different Data Perturbation Types [23]

Random Rotation perturbation	***Geometric Perturbation***	***Random Projection perturbation***
$R*X = Y$ For all three formulas, X is the original dataset. For all three formulas, Y is the perturbed dataset. The random rotation matrix is denoted by the letter R.	$RX + T + D = Y$ The secret rotation matrix is denoted by the letter R. (preserves Euclidean distances) The secret random translation matrix is denoted by the letter T. The secret random noise matrix is denoted by the letter D.	$A*X = Y$ The random projection matrix is denoted by the letter A.

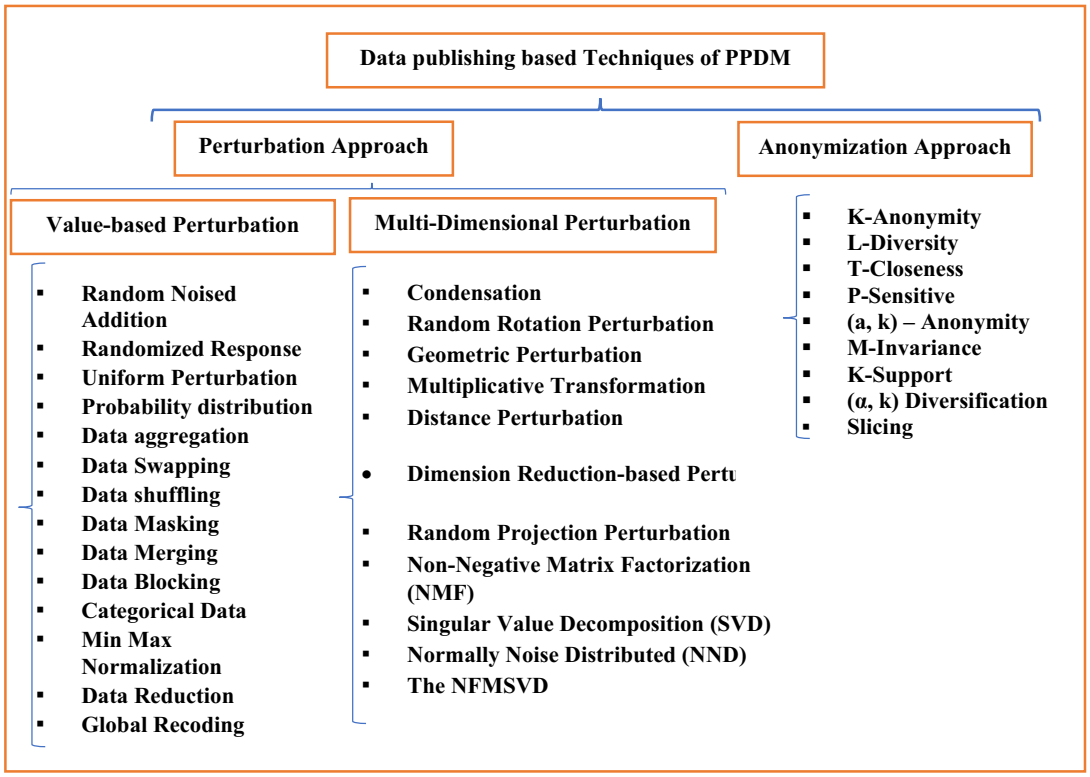

Figure 3: Privacy preserving data mining approaches for cyber-security

Random Noise Addition. This method is based on the fact that data owners may not want to preserve all values in a record equally [8]. We consider the original values $(x_1, x_2, \ldots, x_n)$ in a column to be drawn arbitrarily from a random variable x with a distribution. The randomization process alters the original data by adding random noises R to the original data values ($Y = X + R$ in column Y). It then makes the resultant record $(x_1 + r_1, x_2 + r_2, \ldots, x_n + r_n)$ public [24].

Randomized Response. As a technique developed in the statistical community, the RR scheme is designed to collect sensitive information from individuals in a way that interviewers and the data processors are unaware of the two alternative questions answered by the respondents [25].

Uniform Perturbation. To ensure that individual values are hidden during data collection, data providers can alter the value of each data item or attribute separately before sending it to collectors in two ways. 1.Addition fixed data perturbation or substituting an attribute value with a new one, 2. Generalizing the data values or aggregating based on the related domain hierarchy [26].

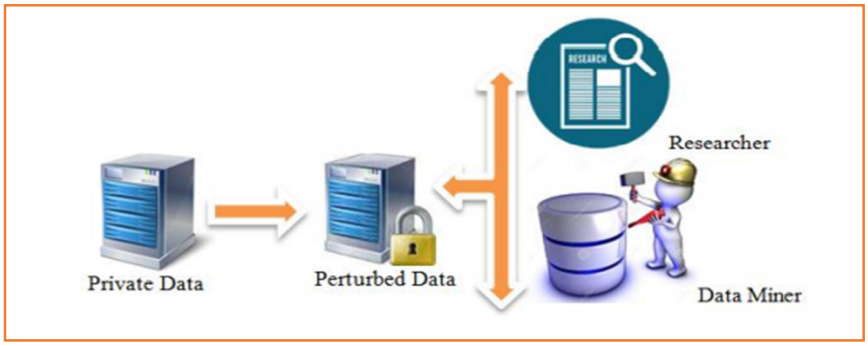

Figure 4: Data Micro Aggregation Perturbation Model [30]

Probability Distribution. This method endeavors to preserve data privacy for individuals by reconstructing the distributions. The point of this approach is that the owner of the data set publishes the resulting tuples by x_{i+r} instead of $x_i(x_1, x_2, \ldots, x_n)$. x_{i+r} is the original data value of a column (one-dimensional distribution) and is drawn from a random variable x and a random value of a certain distribution r [27].

Data Aggregation. This PPDM technique exposes aggregated data and allows the evaluation of particular aggregate query functions in the procedure of concealing an individual record [28].

Data Swapping. The values in distinct records are modified in data swapping strategies to ensure privacy in data mining. The lower half of the data is kept intact and is not impacted in any way, which confirms this method's benefit. As a result, certain general calculations can be performed without jeopardizing data privacy [29].

Data Shuffling. In addition to preserving summary statistics, data shuffling minimizes the risk of exposing confidential variables (X), a risk beyond that already present in the original data [31].

Data Masking. In this method, original sensitive attributes are restored with symbols like 'or' and '+'. It is very similar to the data blocking method but in this method, symbols are used for masking attributes [28].

Data Merging. The common approach to achieving data merging is sharing aggregated data rather than the subject's personal data [32].

Data Blocking. Methodology for falsifying association rules can be complicated by introducing unknown values into data. When uncertain values are entered, the Values of support and trust will fall within a certain range rather than a fixed number. It

means that important association rules are obscured by the final results set [33]. When data is given for mining, a blocking-based strategy suggests hiding certain sensitive information.

4.2.2 Multi-Dimensional Perturbation

The purpose is to hold Multi-Dimensional information. Multi-Dimensional Perturbation is caused by: **Data Mining Task-based Perturbation**. Data mining techniques modify the original data in a way in which the preserved properties can be utilized for tasks specific to data mining or even a particular model [34].

Condensation. This is one of the PPDM techniques that uses a methodology that condenses the data into various classes of the same size. The higher the indistinguishability level, the higher the amount of data privacy. Every group has at least a k value which alludes to the level of indistinguishability [34].

Multiplicative Transformation. Before publishing, the data owner perturbs the data using the multiplicative data perturbation method. The multiplicative data perturbation is a combination of random rotation and fuzzy logic. The original data is given as input to multiplicative data perturbation and obtains perturbed data as output [35].

Distance Perturbation. It also provides a high level of guarantee on data utility, particularly in terms of classification and clustering. As long as distance or inner product are preserved in data mining models, perturbed data will have the same accuracy as original data if such information is preserved [34].

4.2.3 Dimension Reduction-based

Data dimension-reduction-based strategy is an effective way to reduce data size in such a manner that it is compact and provides lower input data dimensions while keeping its geometric structure [36].

Dimension Reduction-based is caused by: **Non-negative Matrix Factorization (NMF)**. The NMF [37] method uses nonnegative constraints to obtain a matrices-based data representation. The NNF matrix (Equation (1)) is as follows:

$$A_{n\times m} = W_{m\times k} \times H_{k\times n} \tag{1}$$

In which W and H are nonnegative matrices of dimensions $m \times k$ and $k \times n$, respectively.

Singular Value Decomposition (SVD). A noted method of dimension reduction in data mining is the SVD [38]. M is the original matrix, while n represents records and m represents attributes. Equation (2) is as follows:

$$M = u\Sigma vT \tag{2}$$

U is an $m \times n$ orthogonal matrix, Σ is an $m \times n$ diagonal matrix whose diagonal elements are positive, and VT represents an $m \times n$ normal orthogonal matrix [39].

Normally Distributed Noise (NDN). As a result of this method, M is added to a noise matrix (Tu) with the same size and normal distribution. The matrix dimensions of Tu and M are the same. Tu's elements are arbitrary values with a standard deviation and a mean parameter. The distributed matrix Equation (3) is as follows [39]:

$$\bar{M} = M + Tu \tag{3}$$

NMFSVD. This method sequentially decomposes the initial matrix using NMF and SVD. The algorithm is too time-consuming, and the change of the data space distance results in undesired privacy protection [40].

4.2.4 Anonymization Approach

This method creates a system in which individual records cannot be distinguished from groups of records by data generalizing and suppressing [40].

The Anonymization Approach is caused by:

K-Anonymity . The k-anonymity approach is an extensively applied and recognized privacy technique [41]. The K value is used as a measure of privacy. The lower the K value, the lower the probability of anonymizing.

L-Diversity. This method saves k values in addition to a variety of sensitive characteristics about each group to prevent homogenous attacks [42].

T-Closeness. As a result of the disadvantages of the L technique, the T-Closeness Approach was developed. Using the T- Closeness technique, the space between a sentient property's distribution in an unknown group and its distribution in the entire table should not exceed the threshold t [42].

P-Sensitive. As an extension of k-anonymity, the p-sensitive model addresses several shortcomings of this model. It considers several sensitive attributes that must not be disclosed. Although initially designed to protect against homogeneity attacks, it also performs well against different types of background attacks [43].

M-Invariance . M-Invariance is a fundamentally privacy-preserving concept in microdata republication. Unfortunately, the existing generalization-based m-Invariances require changing microdata for big data releases. This leads to problems with data utility loss and poor querying performance [43].

5 Evaluation of Privacy Preserving Data Mining Techniques

Privacy measurement is difficult due to the lack of a single and universal definition. However, some metrics have been proposed in the context of PPDMS. Unfortunately, there is no such thing as a single metric because some parameters can be evaluated. The existing metrics can be categorized into three groups that are different in PPDM aspects that are being assessed: 1. Data quality metrics that calculate the loss of data 2. A complexity metric that measures a technique's efficiency and scalability 3. Privacy level metrics that measure how safe data is from the standpoint of disclosure. Metric results make similar evaluations, but the assessment is carried out through data mining outcomes developed with changed data. The following subpar presents a survey of PPDM metrics regarding the privacy level, data quality and complexity [43].

5.1 Privacy Level

The goal of PPDM is to maintain a certain level of privacy while maximizing the data's usefulness. According to data privacy metrics, the original sensitive data can be deduced from the altered information that results from using a privacy preserving approach [43]. The average conditional entropy measure is presented based on the information entropy idea to address the issue of not including the original data distribution [44]. The conditional differential entropy of X, $h(X|Z)$, is obtained from equation 4. in which $f_X()$ and $f_Z()$ are the X and Z density functions, respectively.

$$h(X|Z) = -\int_{\Omega_{x,z}(x,z)} f_{x,z}(x,z)\log_2 f_{x|z}(x)dxdz \tag{4}$$

A key privacy statistic is the hidden failure (HF), which is used to assess the balance between privacy knowledge discovery and security. The ratio of the hidden patterns compared to the original information hidden as a privacy- preserving method is known as the hidden failure [44] and is derived from Equation 5.

$$HF = \frac{\#RP(D')}{\#RP(D)} \tag{5}$$

In this equation, HF stands for hidden failure, D and D stand for sanitized and original data sets, respectively. $\#RP()$ stands for the sensitive Patterns. All sensitive patterns will be properly hidden if $HF = 0$, but no sensitive information will be lost in the process [45].

5.2 Data Quality

Data quality is frequently harmed by privacy-preserving measures. Data quality measurements (also known as Metrics of functionality loss) try to determine the extent of

Table 2: Advantages and Limitations of PPDM Techniques

Technique	*Advantages*	*Limitations*
Techniques according to perturbation	-Scalable -Efficient -In this technique different attributes are preserved independently	-Original data values cannot be regenerated. -Loss of information
Techniques according to anonymization	-Hide records with -Identity or sensitive data about record owners are to be hidden	-Linking attack. -Heavy loss of information

the utility loss. In most cases, the measurements are conducted by analogizing outputs of a function to the original data and the privacy maintained altered information [45].

A metric is defined for determining the accuracy of any reconstruction algorithm (such as randomization) [44]. The authors calculate the amount of data loss by comparing the reconstructed and original distributions using Equation 6, in which $f_x(x)$ is the original density function, and $\hat{f}$ is the reconstructed density function.

$$I(f_x(x), \hat{f}_x(x)) = \frac{1}{2} E[\int_{\Omega_x} |f_x(x) - \hat{f}_x(x)| dx] \tag{6}$$

The MC is a metric that counts how often patterns were hidden when they shouldn't have. As expected, during privacy protection process, no sensitive Patterns were re-lost [45]. This metric is obtained from Equation 7, in which $Rp(X)$ stands for the number of non-restrictive patterns found in database X.

$$MC = \frac{\# \sim Rp(D) - \# \sim Rp(D')}{\# \sim Rp(D)} \tag{7}$$

5.3 Complexity

The efficiency and the implemented algorithm's scalability are the most important aspects of the PPDM approach complexity. Metrics can be utilized for resource consumption, such as time and space to quantify efficiency. All algorithms use these measures. A brief overview is provided here [45].

6 Discussion and Conclusion

Privacy models that cleanse data are used to achieve data publication privacy. However, attackers may attempt to anonymize or infer sensitive information due to access to other publicly accessible sources Modeling background information on adversaries presents a number of challenges as the amount of published data keeps increasing in quantity and complexity. These challenges include determining what data can be used to de-anonymize and the number of public data sources that can be linked together. This

Table 3: Techniques for Data Mining that Preserve Privacy Analysis and Comparison Methods

Methods	***Scenario***	***Criteria***			
		Lots of computation	***Privacy preservation***	***Accuracy of mining***	***Scalability***
Anonymization	Central Commodity	Low	Average	Average	Average
Perturbation	Central Commodity and Distributed	Low	High	High Low Average	High

Table 4: Comparison of Perturbation-Based Privacy Techniques

Methods	***Processing***	***Advantages***
Uniform Noise Distorted	A uniformly determined perturbation matrix is added to the initial matrix	It is easy to use and has a high addition noise efficiency
Normally Noise Distorted	Add randomly selected noise values to the initial matrix	Add noise based on attribute value, randomness is strong
Singular Value Decomposition	In the original matrix of higher latitude, three matrix multiplications are required	Clustering is based on spatial distance and similarity between data
Non-negative Matrix Factorization	Dividing the initial matrix into two matrices and multiplying them	Optimization problems have a much lower computational overhead than SVD
Min Max Normalization	Normalize attribute values at uniform intervals	The prediction accuracy rate of normalization, standardization, and data mining
The NMFSVD	In this method, the initial matrix is successively decomposed by NMF and SVD	It is difficult to reconstruct the initial matrix, and it also has a high mining accuracy

necessitates the creation of more sophisticated and accurate adversarial background knowledge models, which can stimulate research on privacy safeguards that are effective against them.

In privacy preserving data mining, the key objective is to come up with a new algorithm that will hide or protect sensitive data from unauthorized parties. This paper presents a framework based on data publishing for categorizing and evaluating Privacy Preserving Data Mining methods. To begin with, these techniques were divided into two classes of anonymization and perturbation, and their key characteristics were examined. The perturbation process has an impressive computation cost efficiency, but it is difficult to achieve a balance between privacy and accuracy in data mining results. The purpose of this study is to review a wide range of privacy preserving data mining methods and their existing approaches. In this paper, the current state of privacy preserving techniques and cyber space in data mining discussed. We hope that the review presented in this paper can offer researchers different insights into the issue of privacy-preserving data mining, and promote the exploration of new solutions to the security of sensitive.

References

[1] Chang Sun, Lianne Ippel, Andre Dekker , Michel Dumontier and Johan van Soest, "A systematic review on privacy-preserving distributed data mining," Data Science , pp.121–150, 2021.

[2] Prof. K. P. Barabde and Prof. V. Y. Gaud, "A SURVEY OF DATA MINING TECHNIQUES FOR CYBER SECURITY," vol. 6, 2019.

[3] Supritha S, Sushmitha S, and Basavaraju S, "Privacy-Preserving Data Mining: Methods, Metrics and Applications," vol. 2, 2018.

[4] R. Raj and V. Kulkarni, "A Study on Privacy Preserving Data Mining: Techniques, Challenges and Future Prospects," IJIRCCE, vol. 3, no. 11, 2015.

[5] Hina Vaghashia and Amit Ganatra, "A Survey: Privacy Preservation Techniques in Data Mining," International Journal of Computer Applications, vol. 119 , no. 4, 2015.

[6] M. Keyvanpour and S. Seifi Moradi, "Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework," International Journal on Computer Science and Engineering (IJCSE), 2011.

[7] Negar Nasiri and MohammadReza Keyvanpour, "Classification and Evaluation of Privacy Preserving Data Mining Methods," IKT, 2020.

[8] Negar Nasiri and MohammadReza Keyvanpour, "Classification and Evaluation of Privacy Preserving Data Mining Methods," IJCTI, vol. 12, no. 3, 2020.

[9] Anushree Raj and Rio G.L. D'Souza, "Survey on Anonymization of Privacy Preserving Data Publishing," Computer Science and Engineering Department, vol. 3, 2018.

[10] Sri Satya Sai, Bhopal Indore Roa and Madhya Pradesh,"Efficient Model for Privacy Preserving Classification Of Data Streams," vol. 12, no. 2, 2021.

[11] Sangavi N, Jeevitha R, Kathirvel P and Dr. Premalatha K, "RANDOM DATA PERTURBATION TECHNIQUES IN PRIVACY PRESERVING DATA MINING," (IRJET), vol.7, 2020.

[12] Balkis Abidi, Sadok Ben Yahia and Charith Perera, "Hybrid microaggregation for privacy preserving data mining," Journal of Ambient Intelligence and Humanized Computing, 2019.

[13] G Ravi Kumar , Dr. Harsh Pratap Singh and Dr. N.Rajasekhar, "Security and Privacy Protection in Datamining," vol. 12 , no. 2, 2021.

[14] Supritha S, Sushmitha S, Dina Asghar and Kamal Mohan, "Privacy-Preserving Data Mining: Methods, Metrics and Applications," An International Journal, vol. 2, 2018.

[15] Parul Agarwal, M. Afshar Alam and Ranjit Biswas,"Analysing the agglomerative hierarchical clustering algorithm for categorical attributes," International Journal of Innovation, Management and Technology, vol. 2, no. 2, 2010.

[16] RICARDO MENDES and JOAO P. VILELA, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications, vol. 5, 2017.

[17] J. Dwivedi, "Various Aspects of Privacy Preserving Data Mining: A Comparative Study," International Journal of Engineering Research in Current Trends, vol. 1, no. 1, 2019.

[18] LEI XU, CHUNXIAO JIANG, JIAN WANG, JIAN YUAN, and YONG REN, "Information Security in Big Data: Privacy and Data Mining," vol. 2, 2014.

[19] Alpa Shah and Ravi Gulati, "Privacy Preserving Data Mining: Techniques,Classification and Implications-A Survey," vol. 137, no. 12, 2016.

[20] Mrs. Suchitra Shelke and Prof. Babita Bhagat, "Techniques for Privacy Preservation in Data Mining," International Journal of Engineering Research & Technology (IJERT), vol. 4, 2015.

[21] Benjamin Denhama, Russel Pears and M. Asif Naeema, "Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining," vol. 152, 2020.

[22] Anna L. Buczak and Erhan Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," vol. 18, no. 2, 2016.

[23] G. Srinivas Reddy, "DATA PROCESSING THROUGH AN ADDITIVE ROTATIONAL PERTURBATION TECHNIQUE IN A SECURED ENVIRONMENT OF PPRIVACY," vol. 9, no. 2, 2021.

[24] M.Mohanrao and Dr.S.Karthik, "Perturbation Based Privacy Preserving Data Mining," SSRG International Journal of Computer Science and Engineering (ICRTESTM), 2017.

[25] Musavir Hassan, Muheet Ahmed and Majid Zaman, "An Ensemble Random Forest Algorithm for Privacy Preserving Distributed Medical Data Mining," International Journal of E-Health and Medical Communications, vol. 12 , 2021.

[26] MOHAMMED BINJUBEIR, ABDULGHANI ALI AHMED, MOHD ARFIAN BIN ISMAIL, ALI SAFAA SADIQ AND MUHAMMAD KHURRAM KHAN, "Comprehensive Survey on Big Data Privacy Protection," vol. 8, 2020.

[27] Tanzeela Javid, Manoj Kumar Gupta and Abhishek Gupta, "A hybrid-security model for privacy-enhanced distributed data mining," vol. 34, pp. 3602-3614, 2022.

[28] Ajmeera Kiran and Dr. D. Vasumathi, "Data Mining: Random Swapping based Data Perturbation Technique for Privacy Preservingin," International Journal of Recent Technology and Engineering (IJRTE) , vol. 8, 2019.

[29] D. Laskar and G. Lachit, "A Review on Privacy Preservation Data Mining (PPDM)," International Journal of Computer Applications Technology and Research, vol. 3, no. 7, 2014.

[30] V. Jane Varamani Sulekha and Dr. G. Arumugam, "PMA for Privacy Preservation in Data Mining," (IJERT), vol. 6, 2017.

[31] Han Lia, Krishnamurty Muralidhar and Rathindra Sarathy, "The Effectiveness of Data Shuffling for Privacy-Preserving Data Mining Applications," vol. 8, 2014.

[32] Athos Antoniades, ohn Keane, Aristos Aristodimou, Christa Philipou, Andreas Constantinou, Christos Georgousopoulos, Federica Tozzi , Kyriacos Kyriacou, Andreas Hadjisavvas, Maria Loizidou, Christiana Demetriou and Constantinos Pattichis, "The effects of applying cell-suppression and perturbation to aggregated genetic data," (BIBE), 2012.

[33] Lokesh Patel and Prof. Ravindra Gupta, "A Survey of Perturbation Technique For Privacy-Preserving of Data," vol. 3, 2013.

[34] Desmond Ko Khang Siang, Siti Hajar Othman and Raja Zahilah Raja Mohd Radzi, "Comparative Study on Perturbation Techniques in Privacy Preserving Data Mining," vol. 8, 2018.

[35] Thanveer Jahan , G. Narasimha and V. Guru Rao, "A Multiplicative Data Perturbation Method to Prevent Attacks in Privacy Preserving Data Mining," no. 1 , 2016.

[36] J. Hyma, P. S. Varma, S. N. K. Gupta and R. Salini, "Heterogeneous Data Distortion for Privacy-Preserving SVM Classification," In Smart Intelligent Computing and Applications. Springer, Singapore, 2019.

[37] Tao Li, Yongzhen Ren, Yongjun Ren, Lina Wang, Lingyun Wang and Lei Wang, "NMF-Based Privacy-Preserving Collaborative Filtering on Cloud Computing," 2019.

[38] Afsana Afrin, Mahit Kumar Paul and A. H. M. Sarowar Sattar, "Privacy Preserving Data Mining Using Non-Negative Matrix Factorization and Singular Value Decomposition," EICT, 2019.

[39] Jinzhao Shan, Ying Lin and Xiaoke Zhu, "A New Range Noise Perturbation Method based on Privacy Preserving Data Mining," IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), 2020.

[40] Guangfu Chen, Chen Xu, Jingyi Wang, Jianwen Feng and Jiqiang Feng , "Nonnegative matrix factorization for link prediction in directed complex networks using PageRank and asymmetric link clustering information," vol. 148, 2020.

[41] Ruoxuan Wei, Hui Tian and Hong Shen, "Improving k-anonymity based privacy preservation for collaborative filtering," vol.67, pp. 509-519, 2018.

[42] Rupali Gangarde, Amit Sharma, Ambika Pawar, Rahul Joshi and Sudhanshu Gonge, "Privacy Preservation in Online Social Networks Using Multiple-Graph-Properties-Based Clustering to Ensure k-Anonymity, l-Diversity, and t-Closeness," vol. 10, 2021.

[43] R.KAYALVIZHI and Dr. K. RAMESHKUMAR, "PRESERVATION TECHNIQUES IN DATA MINING," vol. 8, 2019.

[44] LILI ZHANG, WENJIE WANG and YUQING ZHANG, "Privacy Preserving Association Rule Mining: Taxonomy, Techniques, and Metrics," vol. 7, 2019.

[45] Gayathiri. P and Dr. B Poorna, "Association Rule Hiding for Privacy Preserving Data Mining : A Survey on Algorithmic Classifications," vol. 12, no. 23, 2017.