IJICTR
International Journal of Information & Communication Technology Research

Volume 12- Number 3 – 2020 (27 -37)

Classification and Evaluation of Privacy Preserving Data Mining Methods

Negar Nasiri
Department of Computer Engineering
Faculty of Engineering
Al-Zahra University
Tehran Iran
negarnasiri97@yahoo.com

MohammadReza Keyvanpour *
Department of Computer Engineering
Faculty of Engineering
Al-Zahra University
Tehran Iran
keyvanpour@alzahra.ac.ir

Received: 10 March 2020 - Accepted: 28 May 2020

Abstract—**In the last decades a huge number of information is produced per hour. This collected data can be used in some different fields such as business, healthcare, cybersecurity, after some process etc. in step two, the important process is that when this data is gathered, extraction of useful knowledge should be done from raw information. But the challenge that we face within this process, is the sensitivity of this information, which has made owners reluctant to share their sensitive information. This has led the study of the privacy of data in data mining to be a hot topic today. In this paper, an attempt is made to provide a framework for qualitative analysis of methods. This qualitative framework consists of three main sections: a comprehensive classification of proposed methods, proposed evaluation criteria, and their qualitative evaluation. In this case, we have a most important purpose of presenting this framework:1) systematic introduction of the most important methods of privacy-preserving in data mining 2) creating a suitable platform for qualitative comparison of these methods 3) providing the possibility of selecting methods appropriate to the needs of application areas 4) systematic introduction of points Weakness of existing methods as a prerequisite for improving methods of PPDM.**

Keywords—Information, Privacy, Data Mining, Privacy preserving Data Mining, PPDM.

I. INTRODUCTION

Data collecting and analysis, the number of which is increasing very fast every moment in these days, has become one of the most important parts of many jobs since its owner found that data analysis has a positive impact on the growth of their activities. Analysis of this data has shown that it can be useful for thousands of services such as healthcare, banking, cybersecurity, commerce, transportation, and many more [1]. Analysis of such information makes it possible to increase their productivity in the mentioned areas by using this information. However, this storage and use of this data have raised serious concerns about data privacy. This concern is due to the sensitivity of this information which is important to the user and relates to the privacy of individuals. The definitions of data mining are different, but all of these definitions refer to a common concept data mining is a process of discovering or extracting interesting patterns, associations, changes, anomalies and significant structures from large amounts of data which is stored in multiple data sources such as file systems, databases, data warehouses, or other information repositories [2].

However, data is often collected from several different sites [4]. There is a lot of concern today about the privacy of sensitive data, which limits access to some data, especially in distributed data. Methods that allow us to extract the knowledge from data while maintaining privacy are known as privacy techniques in data mining. In previous studies, none of the categories have been done completely and they didn't overview to

* Corresponding Author

all categories that exist in privacy-preserving data mining, we have tried to provide the most comprehensive classification because in most of the available articles only the steps of privacy-preserving data mining are divided and the following methods are not mentioned so we try to define a classification that to consider all groups and also to be examined from a new perspective. We tried to have a brief explanation of all the methods and sub-methods and also to provide a category in this regard.

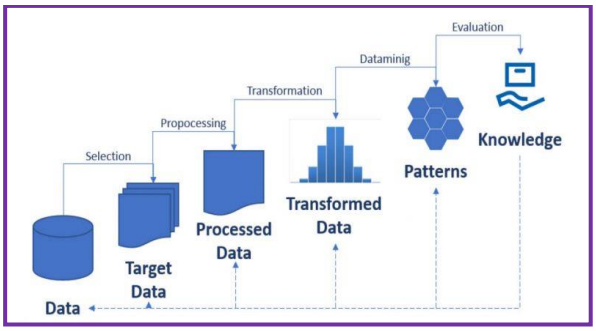

Figure. 1. Data mining process [3].

The rest of the paper is organized as follows. In section 2 the background of the researches will be described and in the third part classification of different techniques will be reviewed. The fourth part would consist of evaluation criteria and the last part would be the summary of the article. And at the end, you will see a table of the latest methods, which is the result of reading recent articles.

II. RELATED WORKS

In [9] the classification of privacy techniques has been done. In this article, privacy in the data mining cycle is examined. First Privacy at the time of data collection, second at the time of data dissemination third at the time of Data distribution, and fourth at the time that data exits from the data mining process using techniques that extract data privacy without compromising data privacy. Jayram Dwivedi has divided privacy techniques into five main sub-categories, Data Perturbation, Blocking based technique, Cryptographic Technique, Condensation Approach [10]. Hyma, Varma, Gupta, and Salini [11] proposed a technique in which classification is done with the help of Support Vector Machine while preserving privacy. Privacy is preserved by distorting the data heterogeneously according to the requirement of data. This technique maintains the privacy as well as the usefulness of the data. In [23] [24] Shweta Tanja, Shashank Khanna, Sugandha Tilwalia, Ankita, have reached results Cryptography and Random Data Perturbation methods perform better than the other existing methods results by Cryptography, Anonymization, Perturbation and A tabular comparison of work done by a different approach. Jun Liu, Yuan Tian, Yu Zhou, Yang Xiao, Nirwan Ansari in [26] used secure multi techniques party computation technique and they could find important. They have found when they use MPC and SPDZ protocol. that the performance of their implementation could be improved by utilizing graphic processing unit (GPU) acceleration. M. Antony Sheela and K. Vijayalakshmi [25] in 2018 concluded that using the Partition Based Perturbation technique wasn't that perfect, so choose differential privacy model. They Presented privacy classification methods are based on data distribution, data disruption, data mining algorithms, data or hidden rules, and privacy protection. In [33], different privacy preservation distributed data mining techniques commonly known as cryptographic approaches like Secure Multiparty Computation Homomorphic Encryption and Secret Sharing methods were discussed and methods like homomorphic encryption and secret sharing are implemented on medical and business data. In this paper experimental result shows that the secret sharing method performs better than homomorphic encryption.

III. CLASSIFICATION OF PRIVACY PRESERVING TECHNIQUES

In many studies, privacy methods have been considered from different perspectives. Privacy methods (in data mining) from a data mining perspective have different aspects that need to be considered in different situations. Data mining methods are important to protect privacy from three different perspectives [8]. We considered that there are three basic views on privacy related to the part of the Data Mining process where exactly we preserve privacy [7].1. Data Viewer or Responsible 2. Owner's Perspective 3. User Viewer.

On the other side Data mining methods along with privacy can be categorized from different perspectives:

- *Data mining algorithms***:** in point of view, extracting association rules, classifying, or clustering.
- *Data Distribution:* Data mining methods can be divided into two centralized or distributed categories based on data distribution (figure 2).
- *Privacy Protection Approach:* There are two general approaches to data privacy protection in data mining methods. Disruption-based or encryption-based methods and cryptography-based methods.

From the perspective of data mining, privacy is categorized into four perspectives. This classification is as follows:

1. Privacy at the time of data collection before data mining.
2. Privacy at the time of data publishing
3. After the completion of the data mining algorithms process.

In many cases, output must be limited to prevent the release of sensitive information since it can contain useful information

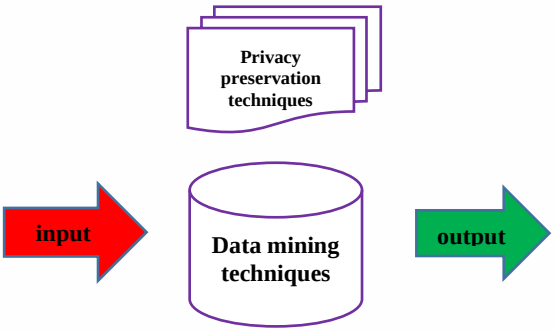

Figure. 2. PPDM Architecture [49].

With sensitive minerals. But all of these methods are appropriate when the data is aggregated in a centralized database, so an important issue is called 4.data privacy when the database is distributed in a way that requires privacy methods. Privately, we have this data in distributed mode. In this article, after reviewing articles and recent studies, we have tried to present a new classification (Figure 4) in the field of data mining privacy and have an overview of recent approaches, advantages, disadvantages, and evaluation criteria. The methods are then categorized as mentioned in [5], [6] [22].

A. Privacy while collecting data

To ensure privacy at data collection time, the sensory device transforms the raw data by randomizing the captured values, before sending them to the collector.

These techniques mainly use data alteration or disruption of the original data to prevent the disclosure of any sensitive information in that data. These are sometimes called ambiguity methods or concealment techniques. These methods are often related to the data perspective or respondent and mostly include data correction methods [7]. Methods based on this principle such as methods based on confusion, k-anonymity, data swapping, blocking, and sampling for Data modification are used. The simplest randomization approach may be formally described as follows. Let A be the original data distribution, B, a publicly known noise distribution independent of A, and C the result of the randomization of A with B. That is:

$$C = A + B \quad (1)$$

The collector estimates the distribution C from the received samples c_1, c_2,…,c_n, with n the number of samples. Then, with the noise distribution B (B has to be provided with the data), A may be reconstructed using:

$$A = C - B \quad (2)$$

Equation 1 corresponds to the randomization process at data collection, while equation 2 corresponds to the reconstruction of the original distribution by the collector entity. However, note that the reconstruction of A using equation 2 depends on the estimation of the distribution C. If B has a large variance and the number of samples (n) of C is small, then C (and consequently A) cannot be estimated precisely [10]. A better reconstruction approach using the Bayes formula may be implemented [9]. Additive noise is one of the ways in randomization method that can be used at collection time. The other techniques, that we can use in this way is multiplicative noise to randomize the data also exist [9].

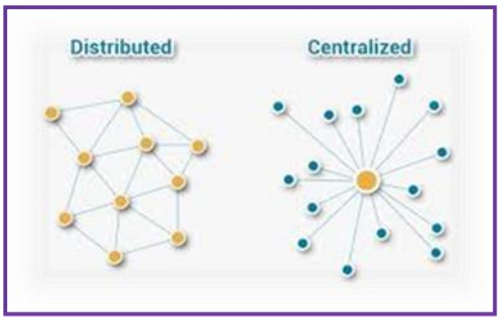

Figure. 3. Data distribution

On the other hand, Data modification can be applied at other phases than at data collection, and other methods besides additive and multiplicative noise do exist. Randomization is a subset of the perturbation operations as you can see in our classification.

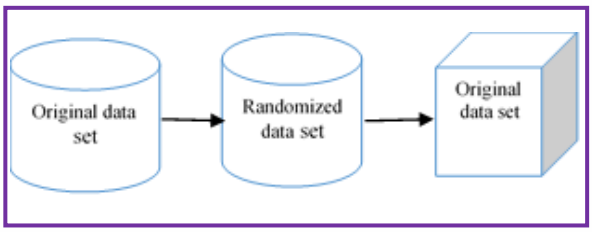

Figure. 4. Model Randomization Method [25].

B. Privacy while Data Publishing

Entities may wish to release data collections either publicly or to third parties for data analysis without disclosing the ownership of the sensitive data. In these circumstances, privacy may be established by anonymizing records before publication. PPDM in data dissemination is also known as privacy information dissemination (PPDP). Some of these methods are described below:

- *Perturbation method:* This method was proposed in 2000 by Agrawal R et al. These techniques are used to change values throughout the data set, which is an example of a data ambiguity technique [7]. These changes are created by adding noise to the data.

a) Random noise was selected independently by a known distribution such as Gaussian distribution [14]. Data confusion is a very simple and effective way to protect sensitive electronic information from unauthorized users or hackers [15] 1Probability Distribution Method.2 Value Distortion Approach. One of the disadvantages of perturbation-based PPDM is that each dimension is reconstructed independently. Therefore, the loss of implicit data in multidimensional records as each distribution-based data mining algorithm deals with different features independently [15].

b) *Swapping-Based Approach*: In data Swapping techniques, the values in different records are changed to maintain privacy in data mining [17]. One of the advantages of this technique is that the lower part of the data is completely preserved and not worried at all. Therefore, certain types of general calculations can be performed without violating the privacy of the data [18].

c) *Masking-Based Approach*: In this method, attributes which are sensitive, are substituted with different symbols such as "*" and privacy is maintained [16].

- *Anonymization method:* aims at making the individual record indistinguishable from group records by using techniques of generalization and suppression. Its representative approach is k-anonymity. The motivating factor behind the k-anonymity approach is that many attributes in the data can often be considered quasi-identifiers which can be used in conjunction with public records to uniquely identify the records. Many methods have been proposed, e.g, k-anonymity, p-sensitive k, (a, k)-anonymity, l-diversity, t-closeness, M-invariance, etc. [30].

1. *K _anonymity approach:* One of the most popular privacy models is the K_anonymity model that provides by Samarati and Sweeny [9]. To be able to use the k-anonymity method before the data mining process, we should follow the algorithms used in this method, which usually use methods such as repression. In other words, A particular data release possesses *k*-anonymity if an individual's record can't be differentiated from k-1 other records at the minimum [16]. In the k-anonymity, the value of k may be used as a measure of privacy: the biggest value of k could be a way that makes it harder to de-anonymize records. In this theory, in an equivalence class, the probability of de-anonymizing a record is 1/k. However, increasing k will also decrease the efficiency of the data science higher generalization will have to occur. Some of the merits of the k-anonymity model are the simplicity of definition and the number of available algorithms. Nevertheless, this privacy model has two important issues. At first, each record represents a unique individual, or in other words, that each represented individual has one, and only one record. If this is not the case, an equivalence class with k records does not necessarily link to k different individuals. The second problem relates to the fact that sensitive attributes are not taken into consideration when forming the k-anonymized dataset. This may lead to equivalent classes where the values of some sensitive attributes are equal for all the k records and consequently, disclosure of private information of any individual belonging to such groups. Another consequence of not taking into account sensitive attributes when forming the classes is the possibility of de-anonymizing an entry (or at least narrow down the possibilities) by associating QIDs with some background knowledge over a sensitive attribute [9].

a) *Generalization*: In this approach that is under anonymization, every attribute must be arranged to more than one common attribute. The main step in this approach refers to changing a respective value/attribute with a more common term. when generalization is finished, then anybody can use the original database value of quasi identifier must be specialized to sample quantity and this refers to the key of full domain generalization. In case of a parent node gets generalized, anything held should be generalized to a parent node.

b) *Suppression*: With appreciation to suppression under anonymity; the dataset can be comprised into two instances namely, suppressed attributes Non suppressed attributes If the tuple is anonymous (k), then each tuple is a member of T. In the dataset, the respective value is changed via *. Suppression is used to reduce the size of the dataset [34].

c) *L-Diversity Approach:* This approach has been proposed to prevent homogenous attacks from the K_anonymity technique, which not only emphasizes saving k values but also considers saving a variety of sensitive characteristics of each group. In this

technique, each anonymous group must consider the minimum / best value for each sensitive attribute [27]. However, this technique has some shortcomings too: e.g, it might be unnecessary and difficult to achieve that. On the other hand, this technique is insufficient to prevent attribute disclosure, Such as Similarity Attack. If the sensitive attribute values in an anonymized group are distinct but semantically the same, the adversary can learn important information [27].

d) *T-Closeness Approach:* The idea of this method is about the sharing of sensitive records in each team is no longer too away from the distribution in the full population. The "t" says that the distributions be no extra than a distance t apart. If the sensitive record in a group does not stand out, this thwarts the homogeneity attack and the historical past expertise assault. The dataset tuple is said to have t closeness if all same instructions have similar conditions. A piece of work responds to moving a piece of the earth via a piece of base distance.

$$WORK(p, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij} \quad (1)$$

Therefore, t-closeness protects towards attribute disclosure, however, it no longer identification disclosure too [35].

e) The $\epsilon -$differential: is one of the models from anonymized categorize that a single record does not considerably affect the outcome of the analysis over the dataset. From this perspective, individual privacy will not be affected by participating in the data collection because it will not make much difference in the outcome [9]. The ϵ-differential privacy model can be formalized as follows. Let T(.) be a randomized function, and A_1 and A_2 two databases differing at most on one record, so:

$$ln\left(\frac{P_r[K(A_1)\epsilon S]}{P_r[K(A_2)\epsilon S]}\right) \leq \epsilon \forall \subseteq Range\ (K) \quad (2)$$

- *Sampling-Based Approach:* This means that larger parts of the database are hidden and only the rest of it is displayed for data mining purposes [8]. This method results in the loss of a large part of the data [18]. If you want to know more, we suggest reading those references [37] and [38] for detailed descriptions of some of the referred group anonymization privacy models.

- ❖ k-anonymity variants: k m-anonymization [39], (α,k)-anonymity [40], p-sensitive k-anonymity [41], (k, e)-anonymity [42], MultiR (MultiRelational) k-anonymity [43] and (X, Y)-anonymity [44];
- ❖ l-diversity variants: (τ ,l)-diversity [45] and (c, l)- diversity [45];
- ❖ t-closeness variants: closeness [46];
- ❖ $\epsilon -$differential privacy variants: differential identifiability [47] and membership privacy [48].2017

C. *Privacy after data mining process (data mining output)*

These techniques are usually related to after the data mining process and the time of data exit. However, this method is not very common and is often used to determine which sensitive information can be extracted from the data mining and what information must be removed before the data mining process can begin [9].

- *Association Rule Hiding:* The rule of associations is a privacy technique that aims to identify all insensitive rules, while no sensitive rules have been discovered. [20] [21] These algorithms often encrypt important business information and work with Hidden algorithms to prevent sensitive rules from being revealed. The algorithms for concealing association rules can be divided into three different categories called disclosure approaches, border-based approaches, and precision approaches [19].
- *Downgrading Classier Effectiveness:* To maintain privacy in classification programs, techniques are used to reduce classification accuracy. Because some rule-based classifiers use the mining methods of the law as subroutines, the methods of concealing the law of communication are also used to reduce the effectiveness of the class [9].
- *Query Auditing and Inference Control:* Sometimes people may have access to the original data set, allowing exclusively statistical queries to the data. Specifically, users can only search for data collected from the data set, not individual or group records. However, some queries (or sequences of queries) can still display private information. There are two main approaches to addressing these inquiries: controlling the inference of the inquiry, in which either the original data or the output of the inquiry is disrupted. And handling inquiries, where one or more inquiries are ignored from existing sequences [9].

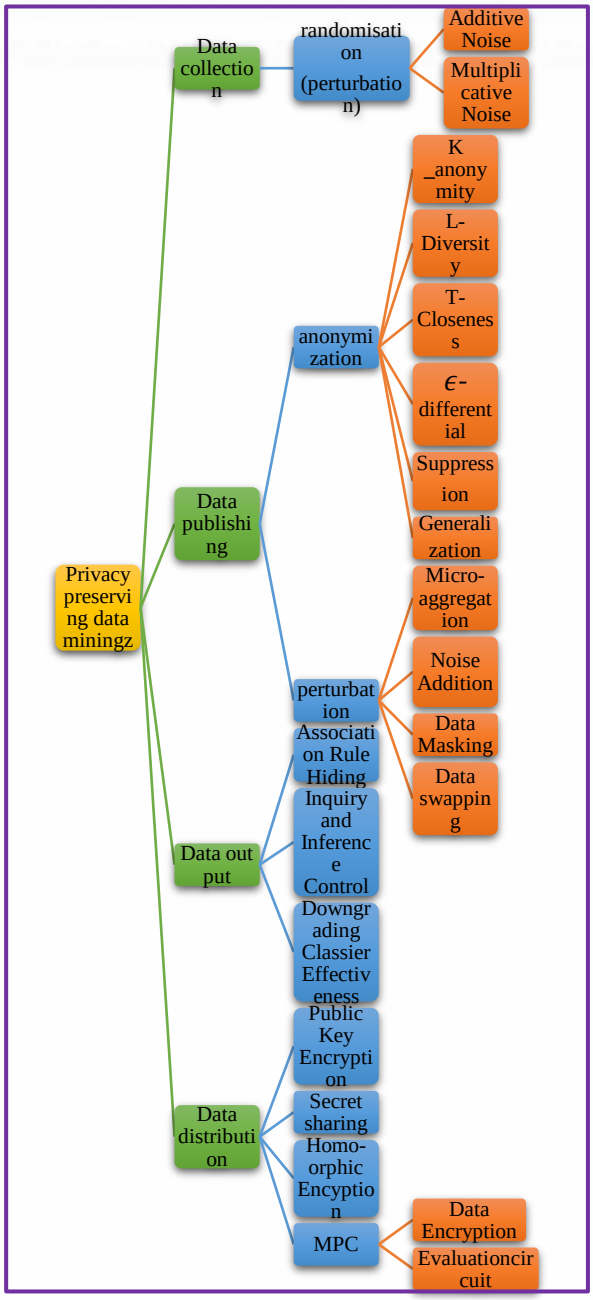

Figure. 5. Classification Privacy Preserving Data Mining.

D. *Privacy while distributing data*

The techniques used during this data mining process do not change the data, but try to manipulate the data mining process to avoid disclosing the sensitive knowledge that follows the process itself [7]. It is clear that the use of these techniques is appropriate when more than one party is involved in the data mining process, for instance, when we need to perform some distributive calculations during the data mining process [8].

- *secure multi-party computing _based (SMC):* It is a technique that gives various protocols through which various collaborative untrusted parties collectively calculate the function using their inputs while keeping the individual inputs (sensitive data) private. The calculation is such that the final output is correct and consistent [13] [5]. Secure Multi-Party Computation (MPC) is regarded as one of the superior and advanced cryptography-based security techniques wherein it is employed for privacy preservation in distributed data mining. It consists of an evaluation of statistical function along with multiple parties. None of these parties are allowed to access any other information except the required one [34].

Algorithm 1: Secure Multiparty Computation [34]
i. Setting 1: Two or more parties P_i (i=1,.... ,n) with private inputs X_i 2: Join and Compute f (x_1, x_2 , , x_n) = (y_1,y_2,.....y_n) on X_i Step 3: Each party P_i should obtain Y_i ii. Security Model 3: Preserve Security models 4: Modelled by an external adversary *A*

a) *the circuit evaluation method* is secure, but it poses significant computational problems since the computational complexity of this method depends on the input size, and then it is expensive since they require complicated encryptions for each bit. The computational cost of the approach for data mining tasks is very high, so that precludes using this method. Then some PPDDM methods use the idea only as sub-protocols to compute certain simple functions [27].

b) *Data Encryption* :Another way for privacy preserving in SMC is processing encrypted data and using homomorphic and commutative properties of encryption systems. In particular, based on homomorphic encryption, solutions for scalar product computation and based on commutative encryption, solutions for secure sum computation and secure size of set intersection are offered [27].

- *Homo-morphic Encryption approach:* Using this, calculations are performed on encrypted data, and results are obtained which are also encrypted. These acquire results, after decryption, are equal to the output of computations when performed on the plain text [16]. Homomorphic encryption systems are specific types of public-key encryption systems. As an example, in public-key encryption, Paillier that is additively homomorphic, the equation 1 holds [27].

Definition: Let M message, a security parameter, a homomorphic encryption M is a quadruple for (K, E, D, Eval). Where, K-key, E-encrypted, DDecrypted, Eval-Evaluation. It is denoted by M = {K, E, D, Eval}. Let say an encryption scheme is homomorphic concerning a function A on M given by. Let p and q are prime numbers and sets n=pq where, Carmichael function *X* given by Z(n)=lcm (p- 1,q-1), where, function L(u)=(u-1)/n. For plaintext x and cyphertext y,

$\forall\ m_1, m_2, r_1, r_2 \in Z_\mu$: $D_{sk}\ (E_{pk}\ (m_1,r_1)\ E_{pk}\ (m_1, r_2)\ \text{mod}\ \mu^2) = m_{1+} m_2\ \text{mode}\ \mu^2$ (3)

- *Public Key Encryption:* In this cryptography, two keys are used: the public key and the private key. The public key is used for encryption and a private key is used for decryption. There is no need of sharing the keys which also helps to increase the security and privacy of the system [16].

Algorithm 2: Secret Sharing [34]

i. Generation of Shares
1: (n) number of participants, (t) threshold, (s) secret value
2: Construct f(x) random polynomial with (k-1) random coefficients
3: Pick random n points for generating n shares
4: Distribute shares among the participants
ii. Reconstruction of Secret
5: Collect shares
6: Reconstruct using Legrange's Basis polynomial f(x)
7: Calculate f (0).

- *Secret sharing -based approach:* refers to the methods of distributing secrets among participants, each of which is assigned a long-term share. None of the participants can figure out the secret on their own, and the secret is only reconstructed when the secret sharing is combined [21] [5]. The scheme secret sharing A (t, n) is a set of two functions of S and R. The function S is a sharing function that takes a secret as input and produces n secret shares in the form of $S(s)= (S_1 , ..,S_n)$The two functions are selected in the manner that for any collection $I \subseteq \{1, ..., n\}$ of t indices, would hold the relation $R (I,S_{I.....}S_I) = S$. In addition, it is necessary that recovering s from a set of t-1 secret shares would be impossible [9].

Table 1. Classifies privacy-preserving data mining based on the data mining cycle. The advantages and disadvantages of each method are analyzed.

TABLE I. ADVANTAGES AND LIMITATIONS OF PPDM TECHNIQUES [12]

Technique	Advantages	Limitations
Anonymization technique of PPDM	Data owner's sensitive or private data are to be secreted.	More information loss, Linking attack
Perturbation technique of PPDM	Preserves various attributes independently.	Information loss and cannot regenerate original data values.
Randomized Response technique of PPDM	It provides good efficiency. Simple and useful for keeping the individual information secretly.	Loss in individual's information. Not much good for database containing several attributes.
Cryptography technique of PPDM	Data transformation is accurate and protected. Provides better privacy and data utility.	It is particularly hard to scale if multiple parties are involved

IV. EVALUATION OF PRIVACY PRESERVING TECHNIQUES

Since privacy has no single standard definition. It is important to note that this is not a quantitative evaluation based on scientific experiments, but a qualitative assessment based on a detailed study e.g[32][9]. Unfortunately, no single metric is enough, since multiple parameters may be evaluated [32]so we categorized privacy level metric in Table 2. The existing metrics may be classified into three main categories, differing on what aspect of the PPDM is being measured: **privacy level** metrics measure how secure is the data from a disclosure point of view, **data quality** metrics quantify the loss of information or utility, and **complexity** metrics, which measure efficiency and scalability of the different techniques [9].

- *Privacy level:* The level of privacy metrics gives a sense of how secure is the data from possible privacy breaches. Recall from the aforementioned discussion that privacy level metrics can be categorized into data privacy metrics and result privacy metrics.
  - One of the first metrics to evaluate data privacy is the confidence level. we use This metric in additive noise-based randomization techniques and calculate how well the original values may be estimated from the randomized data.
  - one of the important results for privacy metric is the hidden failure (HF). this metric when use measures the balance between privacy and knowledge discovery. The hidden failure may be

defined as the ratio between the sensitive patterns that were hidden with the privacy-preserving method, and the sensitive patterns found in the original data. This metric can measure whit this formula:

$$\mathrm{HF} = \frac{\#R_P\ (A')}{\#R_p\ (A)} \qquad (4)$$

HF is the hidden failure, A' and A are the sanitized dataset and the original dataset, and #Rp (0) is the number of sensitive patterns. If HF= 0, all sensitive patterns are successfully hidden, however, more non-sensitive information may be lost in the way. This metric can have used in any pattern recognition data mining technique (*e.g.*, classifier or an association rule algorithm). Note that this metric does not measure the amount of information lost [9].

- when we are faced with the issue of not taking into account the distribution of the original data, the average conditional entropy metric is proposed based on the concept of information entropy. [50].

$$h(x|z) = -\int_{\Omega_{x,z}} fx,z(x,z)\log_2 f(x|z) = z(x)dxdz \qquad (5)$$

where fx (0) and fz (0) are the density functions of X and Z, respectively.

- *Data Quality:* Privacy-preserving techniques often cause decreases the quality of the data. Data quality metrics try to quantify this loss of utility. the calculation is made by comparing the results of a function over the original data, and over the privacy preserved transformed data.

  - The MD metric is a simple counter that increments every time a value is generalized to the parent value. The higher the MD value, the more generalized is the data, and consequently, more information was lost [9].
  - The LM (**Loss Metric**) and ILoss **(Information Loss)** metrics measure the average information loss overall records, by taking into account the total number of original leaf nodes in the taxonomy tree. The ILoss be different from the LM metric by applying dissimilar weights to dissimilar attributes, for the average. The weight may be used to differentiate higher discriminating generalizations [51].
  - For the equality class algorithms, the Discernibility Metric (DM) [120] was described. This metric calculates how many manuscripts are equal to a given record, due to the generalizations. The greater the value, the more information is lost. As an illustration, in the k-anonymity, at least k − 1 other records are identical to any given record, thus the discernibility value would be at least k − 1 for any record. Growing k, will rise generalization and suppression, and consequently the discernibility value. For this purpose, this metric is considered to be the opposite concept of k-anonymity [9].
  - Two metrics to measure data quality loss from the results of pattern recognition algorithms are the Misses Cost (MC) and the Artifactual Patterns (AP). The MC measures the number of patterns that were incorrectly hidden. Those are non-sensitive patterns that were lost in the process of privacy preservation. This metric is defined as follows. A is the original database and A' the sanitized database. The misses cost [32].

$$MC = \frac{\#\sim R_p(A) - \#\sim R_p(A')}{\#\sim R_P(A)} \qquad (6)$$

In the best situation, an *MC* = 0% is desired, which means that all non-sensitive patterns are present in the transformed database.

  - The AP metric measures artifact patterns, i.e., the number of patterns that did not exist in A, but were created in the process that led to A'. The following equation defines the AP metric.

$$AP = \frac{|P'| - |P \cap P'|}{P'} \qquad (7)$$

In the best-case situation, *AP* should be 0, indicating that no artificial pattern was introduced in the sanitization process.

  - For clustering techniques, the Misclassification Error (ME) metric introduced in [53] estimates the

percentage of data circumstances that ‘‘are not well classified in the distorted database’’. That is the number of points that were not categorized in the same cluster with the main data and with the sanitized data. The misclassification is defined by the following equation [9]:

- $M_E = \frac{1}{N} \times \sum_{i=1}^{k} (|\text{Cluster}_i(D_1)| - |\text{Cluster}_i(D_0)|)$ (8)

- *Complexity:* The complexity of PPDM techniques mostly concerns the efficiency and the scalability of the completed algorithm [52]. These metrics are well-known to all algorithms [9].

On the other hand, any of the methods available in one of the fields related to qualitative evaluation criteria may work high, low or average. We reached Table 3 in the surveys [28]. This table is based on some studies and only slightly evaluates the evaluation criteria described above for each of the techniques and shows the criteria by which each of these methods has a weakness and better More work is needed on them.

TABLE II. EVAUATION OF PRIVACY PRESERVING DATA MINING TECHNIQUES [28]

Criteria **PPDM Technique**	**Cryptography**	**Perturbation**	**Anonymization**
Computational Cost	*High*	*Low*	*Low*
Privacy Preservation	*High*	*High*	*Average*
Accuracy of mining	*High*	*High* *Average* *Low*	*Average*
Scalability	*Low*	*High*	*Average*

TABLE III. PRIVACY LEVEL METRIC [32]

Data Metrics	**Results Metrics**
- Confidence Level - Average Conditional Entropy - Variance - Privacy Model Specific (K, L, T)	- Hidden Failure

TABLE IV. DATA QUALITY METRIC [32]

Data Metrics	**Results Metrics**
- Minimal Distortion - Loss Metric - Information Loss - Discernibility Metric	Misses Cost - Artifactual Patterns - Misclassification Error

V. CONCLUSION AND DISCUSSION

The main purpose of privacy protection in data mining processes is to develop algorithms that can hide or provide privacy to some sensitive information to prevent unauthorized access by profiteers. However, privacy and accuracy in data mining conflict. In this regard, we have tried to review the number of techniques available in privacy in data mining and examine some of their advantages and disadvantages.

In this article, we provide a brief but useful overview of existing privacy techniques, namely perturbation, anonymity, and cryptography, and analyze their competencies and differences in different scenarios. In most recent articles, only the classification of techniques into four categories before the data mining process during the data mining process after the data mining process and data distribution is mentioned, and in none of the methods is it fully included in these classifications. After reviewing and studying recent articles, we were able to provide a community classification of these methods. We have also summarized a set of advantages and disadvantages of these techniques in Table 1it is suggested that more attention be paid to evaluation metrics in future work and we can improve classification and work especially on every category and explain more in detail.

VI. REFERENCES

[1] I. NATGUNANATHAN, Y. XIANG, G. HUA and S. GUO, "Protection of Big Data Privacy," IEEE, vol. 4, p. 14, 2016.

[2] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," springer 2020

[3] S.D.Gheware, A. Kejkar and S. Tondare, "Data Mining: Task, Tools, Techniques and," International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 10, 2014.

[4] P. A. Head and S. R. Sowdamibai, "A Fast and Efficient Privacy Preserving Data Mining Over Vertically Partitioned Data.," International Journal of Computer Applications, 2013.

[5] P. Taghavi, "Mining sensitive data with privacy protection based on perturbation method," ‹‹M.Sc.›› Thesis, 2014.

[6] J. Domingo-Ferrer, A. Solanas and F. Sebé, "An Anonymity Model Achievable Via Microaggregation," Secure Data Management.Springer, vol. 5159, 2008.

[7] O. V´yborn´y, "Time, Data Mining and Security," Ph.D. Thesis Proposal, 2006.

[8] M. keyvanpour, F. Hasanzadeh and M. Moradi, Advanced topics in data mining, Tehan: kian, 2019.

[9] R. MENDES and . J. P. VILELA, "Privacy-Preserving Data Mining: Methods,Metrics, and Applications," IEEE, vol. 5, p. 21, 2017.

[10] 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.

[11] 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., pp. pp. 459-468, 2019.

[12] G. J. Taric and E. Poovammal, "A Survey on Privacy Preserving Data Mining Techniques," Indian Journal of Science and Technology, vol. 10, 2017.

[13] C. Zhao, S. Zhao, M. Zhao, . Z. Chen, H. Li and . Y.-a. Tan, "Secure Multi-Party Computation: Theory, Practice and," , Information Sciences, 2018.

[14] S. K. BHANDARE, "Data Distortion Based Privacy Preserving Method for Data Mining System," IJETTCS, vol. 2, no. 3, 2013.

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

[16] P. K. Kaur and K. Singh Attwal, "Privacy Preserving Data Mining: Approaches, Applications And Research Directions," International Journal of Advanced Science and Technology, vol. 28, p. pp. 718 – 729, 2019.

[17] D. Laskar and G. Lachit, "A Review on "Privacy Preservation Data Mining (PPDM),

[18] " International Journal of Computer Applications Technology and Research, vol. 3, no. 7, 2014.

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

[20] D. Thakur and P. H. Gupta, "An Exemplary Study of Privacy Preserving Association Rule Mining Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 11, 2013.

[21] D. A. A. Hassan and H. Qassim Jaleel, "A Survey on Privacy Preserving Data Mining PPDM concepts and methods.," vol. 5, no. 9, 2018.

[22] Z. Xu and X. Yi, "Classification of Privacy-preserving Distributed Data Mining Protocols," International Conference on IEEE, 2011.

[23] J. Domingo-Ferrer, "A Three-Dimensional Conceptual Framework for Database Privacy," Springer, vol. 4721, pp. 193-202, 2007.

[24] Y. Zhang and S. Zhong, "A privacy-preserving algorithm for distributed training of neural network ensembles," Neural Comput&Applic, 2013.

[25] C. Mathew, "A Survey on Privacy Preserving Data Mining Techniques," IJERT, vol. 7, no. 5, 2019.

[26] M. A. Sheela and . K. Vijayalakshmi, "Partition Based Perturbation for Privacy Preserving Distributed Data Mining," CYBERNETICS AND INFORMATION TECHNOLOGIES, vol. 17, 2017.

[27] J. Liu, Y. Tian, Y. Zhou, Y. Xiao and N. Ansari, "Privacy preserving distributed data mining based on secure multi-party computation," Elsevier , 2020 .

[28] 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.

[29] A. Senosi and G. Sibiya, "Classification and Evaluation of Privacy Preserving Data Mining: A Review," IEEE Africon, 2017.

[30] E. BERTINO, I. N. FOVINO and L. PARASILITI PROVENZA, "A FRAMEWORK FOR EVALUATING PRIVACY PRESERVING DATA MINING ALGORITHMS," Springer, 2005.

[31] P. Wang, T. Chen and Z. Wang, "Research on Privacy Preserving Data Mining," JIHPP, vol. 1, pp. pp.61-68, 2019.

[32] M. B. Malik, M. A. Ghazi and R. Ali , "Privacy Preserving Data Mining Techniques: Current Scenario and Future," Third International Conference on Computer and Communication Technology, pp. 26-32, 2012.

[33] E.-J. Hong, D.-w. Hong and C. Ho Seo, "Privacy Preserving Data Mining Methods and Metrics Analysis," *Journal of Digital Convergence,* vol. 16, 2018.

[34] V. K. Marimuth and C. Lakshmi, "Performance analysis of privacy preserving distributed data mining based on cryptographic techniques," *7th International Conference on Electrical Energy Systems (ICEES 2021),* pp. 635-640, 2021.

[35] S.Shimona and D. Mahalingam, "Survey on Privacy Preservation Technique," *Proceedings of the Fifth International Conference on Inventive Computation Technologies (ICICT-2020),* pp. 64-68, 2020.

[36] . Aggarwal, Charu C. "On k-anonymity and the curse of dimensionality." In VLDB, vol. 5, pp. 901-909. 2005Aggarwal, Charu C. "On k-anonymity and the curse of dimensionality." In VLDB, vol. 5, pp. 901-909. 2005

[37] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, ''Privacy-preserving data publishing: A survey of recent developments,'' ACM Comput. Surveys., vol. 42. no. 4, pp. 14:1–14:53, 2010

[38] ao, M. Du, J. Le, and Y. Luo, ''A survey on privacy preserving approaches in data publishing,'' in Proc. IEEE 1st Int. Workshop Database Technol. Appl., Apr. 2009, pp. 128–131.

[39] M. Terrovitis, N. Mamoulis, and P. Kalnis, ''Privacy-preserving anonymization of set-valued data,'' Proc. VLDB Endowment, vol. 1, no. 1, pp. 115–125, 2008

[40] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, ''(α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing,'' in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2006, pp. 754–759

[41] T. M. Truta and B. Vinay, ''Privacy protection: p-sensitive k-anonymity property,'' in Proc. IEEE 22nd Int. Conf. Data Eng. Workshops, Apr. 2006, p. 94.

[42] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, ''Aggregate query answering on anonymized tables,'' in Proc. IEEE 23rd Int. Conf. Data Eng. (ICDE), Apr. 2007, pp. 116–125.

[43] M. E. Nergiz, C. Clifton, and A. E. Nergiz, ''Multirelational k-anonymity,'' in Proc. IEEE 23rd Int. Conf. Data Eng. (ICDE), Apr. 2007, pp. 1417–1421.

[44] K. Wang and B. Fung, ''Anonymizing sequential releases,'' in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2006, pp. 414–423

[45] H. Tian and W. Zhang, ''Extending `-diversity to generalize sensitive data,'' Data Knowl. Eng., vol. 70, no. 1, pp. 101–126, 2011

[46] N. Li, T. Li, and S. Venkatasubramanian, ''Closeness: A new privacy measure for data publishing,'' IEEE Trans. Knowl. Data Eng., vol. 22, no. 7, pp. 943–956, Jul. 2010.

[47] J. Lee and C. Clifton, ''Differential identifiability,'' in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 1041–1049.

[48] N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang, ''Membership privacy: A unifying framework for privacy definitions,'' in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2013, pp. 889–900

[49] R. Ratra and P. Gulia, "Privacy Preserving Data Mining: Techniques and Algorithms," *International Journal of Engineering Trends and Technology*, vol. 68, no. 11, 2020.

[50] D. Agrawal and C. C. Aggarwal, ''On the design and quantification of privacy preserving data mining algorithms,'' in Proc. 20th ACM SIGMODSIGACT-SIGART Symp. Principles Database Syst., 2001, pp. 247–255.

[51] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, ''`-diversity: Privacy beyond k-anonymity,'' ACM Trans. Knowl. Discovery Data, vol. 1, no. 1, p. 3, 2007.

[52] E. Bertino, D. Lin, and W. Jiang, ‘‘A survey of quantification of privacy preserving data mining algorithms,’’ in Privacy-Preserving Data Mining. New York, NY, USA: Springer, 2008, pp. 183–205.

[53] S. R. M. Oliveira and O. R. Za, ‘‘Privacy preserving clustering by data transformation,’’ J. Inf. Data Manage., vol. 1, no. 1, pp. 37–52, Feb. 2010.

Negar Nasiri received the B.Sc. degree in Information Technology Engineering from Payam Noor University in 2018, and Now studying M.Sc. degree in Software Engineering at Alzahra University, Tehran, Iran. her current research interests include Cryptography, Privacy Preserving Data mining, Data mining and Machine Learning.

Mohammad Reza Keyvanpour received the B.Sc. degree in software engineering from the Iran University of Science & Technology, in 1997, and the M.Sc. and Ph.D. degrees in Software Engineering from Tarbiat Modares University, Tehran, Iran, in 2000 and 2007, respectively. Currently, he is an associate professor with Alzahra University, Tehran, Iran. His current research interests include E-health and Data Mining.