Classification of Privacy Preserving Data Mining Algorithms: A review

WORD COUNT 6164
TIME SUBMITTED 26-AUG-2020 11:09AM
PAPER ID 62521694

Classification of Privacy Preserving Data Mining Algorithms: A review

Dedi Gunawan*

Informatics Department
Universitas Muhammadiyah Surakarta
Jl. A. Yani Pabelan, Kartasura
Surakarta, Indonesia
**dedi.gunawan@ums.ac.id*

Abstract

Nowadays, data from various sources are gathered and stored in databases. The collection of the data does not give significant impact, unless database owner conducts certain data analysis such as using data mining techniques to the databases. Presently, the development of data mining techniques and algorithms provide significant benefits for information extraction process in terms of the quality, accuracy and precision results. Realizing the fact that performing data mining tasks using some available data mining algorithms may disclose sensitive information of data subject in the databases, an action to protect privacy should be taken into account by the data owner. Therefore, privacy preserving data mining (PPDM) is becoming an emerging field of study in data mining research group. The main purpose of PPDM is investigating the side effects of data mining methods that originate from the penetration into the privacy of individuals and organizations. In addition, it guarantees the data miners cannot reveal any personal sensitive information contained in a database, while at the same time data utility of a sanitized database does not significantly differ from that of the original one. In this paper we present a comprehensive review of the current PPDM techniques based on its taxonomy a long with the challenges and future development of PPDM techniques.

Keywords: Database, data mining, privacy preserving data mining, sensitive information.

I. INTRODUCTION

In today's era, data can be easily collected from various sources and stored in various types of databases. The collection of data in databases is meaningless until database owners conduct certain data analysis to excavate valuable information from the databases. In general, data analysis is carried out to extract useful information from databases, more specifically when it is used to find hidden knowledge in the databases then it is so called data mining. Data mining plays an important role in many applications such as business management, marketing analysis, and science exploration [1]. The true value of data mining techniques does not reside in a set of complex algorithms; instead it resides in the practical problems that it can help to solve [2]. There are two categories of data mining models such as predictive and descriptive. The predictive model aims to picture some predictions of a certain trend or correlation between one variable with other variables in a database such as regression, classification and time series analysis. On the other hand, descriptive model focuses on exploring knowledge from databases. Several data mining tasks that included in these models such as clustering, summarization, association rules and sequence discovery.

Nowadays, various data mining software have been developed and published in software market. However, not all people or institutions have ability to utilize the software appropriately due to the lack of resources and limited knowledge in the institutions. Recent trend shows that institutions prefer to hire or use services from data mining company to mine their data. Handling a raw data to other institutions is not encouraged since there might be some sensitive information related to the institutions, their people or customers.

Table 1. Patients data table

Name	Birth date	Post code	Occupation	Disease
John	1975/12/10	71794	Engineer	Tuberculosis
Monna	1980/3/15	71780	Accountant	Dengue
Jane	1984/5/10	71794	Teacher	Pneumonia
Matip	1977/2/12	71793	Engineer	HIV
Mark	1978/8/16	71790	Programmer	Pneunomia
Hardy	1981/11/1	71790	IT specialist	Tuberculosis

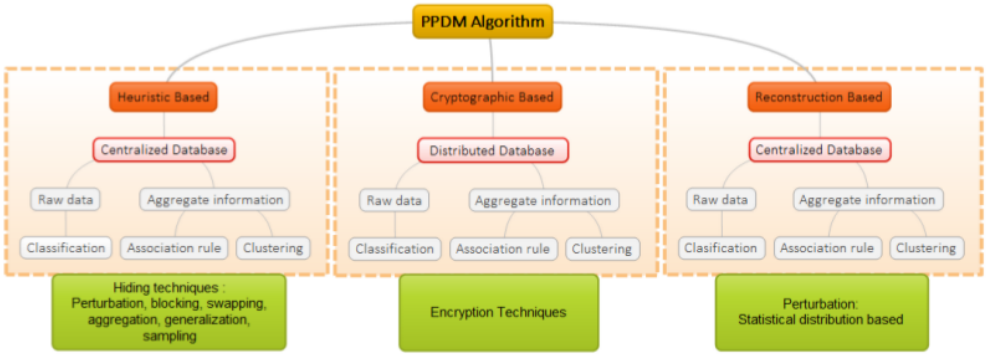

Figure 1. Taxonomy of PPDM techniques

In reality, a number of companies do not really pay attention about the privacy issues residing in their database and that results in serious privacy violation [3]. Therefore, in this situation the database owner should have to be careful in handling the database to other companies for mining process due to some data mining tools may causes sensitive information breach [4]. In another case, data recipients may also act as adversarial parties who might unfairly use the database to disclose sensitive information of individuals [5].

To mitigate the possibility of such breaches a solution called privacy preserving data mining (PPDM) has been developed. Since the last decade, researchers have been developed various privacy protection methods to limit sensitive information leakage by performing data sanitizing into databases. Therefore, prior to sharing or handling a database to other parties the data owners are encouraged to apply privacy preserving data mining (PPDM) algorithms with respect to balance the trade-off between privacy and data utility. In the context of PPDM, database owner knows in advance the types of data mining tasks performed by data miner.

There are three requirements that should be satisfied to design PPDM algorithms. The first and the most important requirement is the entire sensitive values or sensitive itemsets cannot be mined in the sanitized database. While the second requirement is non-sensitive values or itemsets in the original database should also can be mined from the sanitized database. The last, the difference between original database and the sanitized database should be minimized. Obtaining a sanitized database which achieves those three requirements is a very difficult problem, and actually it has been proved that the problem is NP-Hard [6]. Therefore, various techniques with various settings have been proposed to balance the trade-off and satisfy the requirements of database owners. Since the pioneering work in [7], [8] and [9], several approaches have been proposed in the PPDM area to deal with privacy in data mining.

II. Privacy Preserving Problem

Prior to describing the classification of the PPDM techniques, it is important to highlight the intuition why the PPDM techniques need to be developed. In general, databases have several attributes that can be distinguished into three different types such as key attribute, public attribute and private or sensitive attribute [10]. The key attribute contains information which can be used to identify individuals, for example user id, customer id or individuals name. The second attribute holds information that accessible to authorized people. In addition, this attribute may lead to individual's privacy breach if not adequately preserved. The last attribute is the attribute which conserves sensitive information and it should be well protected.

Let us consider a tabular database such as in Table 1 which contains several records. The database consists of several attributes that can be categorized into those three. The key attribute of the database is name in which this value directly refers to individuals, while birth date, zip code and occupation are the public attributes. The disease attribute is a sensitive attribute of the data table and thus it should be protected.

The sensitive information breach is possible if one who holds public attributes has a function for constructing logical information to infer sensitive information of an individual through data mining tools. Therefore, PPDM investigates the side effects of data mining methods that originate from the penetration into the privacy of individuals and organizations[11]. Accodingly, to design a PPDM method which can modify databases in such a way data miner could not infer individual sensitive information one should consider various a trade-off between privacy and utility. The problem which occurs due to the leak of confidential information is refered as database inference [12]. Additionally, the PPDM should also be able to preserve similar data utility in the sanitized database like that of the original one.

III. Classification of PPDM Algorithm

Currently, various PPDM techniques have been proposed. The proposed algorithms can be categories into three different groups based on the taxonomy techniques namely reconstruction technique, cryptographic based technique and heuristic based technique. The taxonomy techniques is represented in Figure 1 while the strategy of the techniques is described in Table 2.

Table 2. PPDM Methods and Their Properties

Classification	Method	Strategy
Reconstruction	Additive noise	Modelling noise addition
	Microaggregation	Replacing original values with aggregate value
	Swapping	Swapping values among records
Randomization	Random noise	Generating random value as a noise
Cryptographic	Secure Multiparty Computing	Semi-honest protocol
	Homomorphic encryption	Encryption
Heuristic	Hiding sensitive items	Replacing sensitive items with non-sensitive items
	Item grouping	Generating identical cluster with the same sub-itemset

3.1. RECONSTRUCTION BASED TECHNIQUES

Reconstruction based techniques relies on perturbing original values such that an adversarial data miner could not find the original values and the perturbed database maintains its statistical properties. The method perturbs databases and reconstructs their data distribution in aggregate level to estimate probability distribution of original values as a result the databases statistical properties does not deviate drastically from that of the original one.

A. Perturbation Techniques

The main idea of data perturbation is delivering modified database or sanitized database with additional noise that does not result in significant different from an original data mining results. This method achieves privacy protection by modifying attributes value from a database, such that private value cannot be reconstructed or disclosed. A simple illustration of perturbation is for example a database owner consider an attribute says disease is sensitive, then he can decide how much noise to add to the real value such that the real value cannot be revealed. The amount of noise totally depends on the data owner view, it can be generated randomly under certain probability distribution.

There are three types of perturbation technique in PPDM such as additive noise, microaggregation and rank swapping.

1) *Additive Noise*

As it is indicated by the name, additive noise hides sensitive information by adding some values in a data record or adding artificial records in a database. The idea of using additive noise to sanitize a database in privacy preserving sensitive frequent itemset mining for transaction data have been proposed in [13]. The proposed method appends some artificial transactions into original database. Initially, the method is calculating the number of transaction that should be added in the database, this called maximum safety bound (MSB). Equation 1 represents the computation of MSB.

$$\max(SB_i) = \left\lceil \frac{|si_i|}{\sigma} - m \right\rceil + 1 \qquad (1)$$

The notation $\max(SB_i)$ refers to the safety bound of each sensitive itemset, while m is the number of records contained in database D and $|si_i|$ represents the count of sensitive itemset in the database D. Once the $\max(SB_i)$ is determined, the next step is counting the number of items for each additional transaction p_n, based on standard normal distribution.

Another method for hiding frequent sensitive itemset proposed in [14]. The proposed method protects sensitive frequent itemset by inserting noisy items in certain transactions. The noisy items are selected based on queue and random number generator. Moreover, if a transaction has more items then the more noisy items are generated and added in the transaction. Adding some itemsets or artificial transactions data in a database successfully protect frequent sensitive items since it cannot be mined under the same defined support.

Aiming to protect individual privacy in numerical data [15] proposed *individually adaptable two-phase perturbation method*. In this methods individuals are granted a permission to choose their privacy level. The method firstly perturb original database values using random values generated from independent identically distributed random variable. Following that, it splits perturbed data in to several predetermined intervals. User then chose any part of the split and choose a privacy level e.g. top, high, medium, low. After selecting the options, it adopts an interval length which correspond to the selected privacy level. Finally generating the perturbed values by sampling the interval uniformly. These values are then dispatched to the data miner.

An idea called *select-a-size* has been proposed in [16] to sanitize a database in privacy preserving association rules mining for transaction data. The proposed method employs uniform randomization to generate random itemset of a transaction. The modified itemset is sent to server and the server collect statistical value of the modified transaction. Even tough the algorithm is effective in protecting sensitive information, it takes significant computation cost since the method uses per-transaction strategy which recursively computes the random value.

Adding noise into database is an effective way to guarantee privacy protection in data mining process. However, we should be carefully deciding the amount of noise and the strategy to generate the noise since quality of the sanitized database depends on it.

2) *Microaggregation*

The underlying concept of microaggretation is releasing a database with continuous values for a data mining task i.e. clustering, where the original values are

replaced with values that generated from small original values aggregates.

To generate microaggregation from a database we firstly define a number of groups g, each group contains at least k records. The next phase is calculating the average value for each attribute data for each group and then replace its original averaged values with the average value from each group. The challenge in microagregation is finding optimal k-partition. It should maximizes homogeneity values within a group to reduce information loss.

In the case when a database contains several attributes, the microaggregation can be performed to aggregate all the data in all attributes or it can also be performed by dividing the attributes into several groups. One of some microaggregation methods that commonly used is called Distance to Average Vector (MDAV) [17]. Described in [18] MDAV consists of six sequential processes to generate microaggregate database. The steps are described as follows.

1) Calculate the average record $\bar{x}$ of all records in a database. Select the most distant record x_r from $\bar{x}$ based on distant measurement e.g. Euclidean distance.
2) Determine another most distant record x_s from the previous $\bar{x}$.
3) Generate two groups around x_r and x_s where each group has k-1 closest records.
4) If there are at least 3 k records not belong to any of those two groups, back to step 1-3 by taking the rest of the non-grouped records as a new database.
5) If there are between 3 k-1 and 2 k records do not belong to any of those two groups, do:
   a. Calculate the average record $\bar{x}$ of the remaining records.
   b. Find the most distant record x_r from $\bar{x}$
   c. Generate a group containing x_r and k-1 closest records to x_r.
   d. Create another group consisting the rest of records and end the algorithm
6) If there are less than 2 k records not belonging to any group from the step 3, construct new group containing those records and finish the algorithm.

Microaggregation has been successfully implemented for protecting privacy in query logs which satisfies k-anonymity concept [19]. Since the query logs contain several attributes such as query terms, timestamps, domain name, some distance measurements (Euclidean distance, Levenshtein distance and Hausdorff distance) were used to calculate aggregation of those attributes values.

3) Swapping

The main idea of data swapping is exchanging sensitive values of a record to another record while at the same time maintaining frequency counts. Originally data swapping is developed to protect continuous and categorical values. Data swapping firstly introduced in [20] to protect a database from statistical disclosure. One variant of data swapping called *rank swapping*, hides sensitive values in categorical data and it is successfully implemented for numerical data [21]. The process of *rank swapping* can be briefly described as follows:

1) Determine a parameter value, p.
2) An attribute values of a database is ranked in ascending order.
3) Randomly select value of the attribute in a record and swap it with another attribute value from another record.
4) The rank of those two swapped values should not differ by more than p% of the total number of records in a database.

Several empirical studies such as in [4, 22] argues that rank swapping results in balance trade-off between information loss and disclosure risk.

Perturbation based techniques are promising in preserving original values from adversarial data miner. However, some side effects occur when such techniques are applied, for example the data truthfulness is no longer hold. In some critical database such as health record database this kind of side effect is not tolerate since it may danger people's life.

3.2. RANDOMIZATION

Randomization is closely related to perturbation-based technique since most of the perturbation techniques especially for numerical database generates additional noise using some randomization methods. To achieve a sanitized database which protects privacy in the database, the additional noise should be defined carefully to preserve probability distribution of the database. If we consider X is an original database, Y is noise and Z is a sanitized database, then to generate Z we straightforwardly compute $Z = X + Y$.

The general concept of randomization can be reflected into several points. Initially we can assume a database D contains a set of records X, where $X=\{x_1,..x_N\}$, N equals |D|. For each record x_i a noise y_i is added. The added noise $(y_1, \ldots, y_N)$ is generated independently to result distorted database. Thus, the database D contains values $Z = x_1+y_1,\ldots, x_N + y_n$.

In fact, other than additive strategy randomization there is another variant called multiplicative strategy. Random values for multiplicative strategy also can be generated randomly. Another interesting part is randomization can be applied in data-collection process so that it is not necessary to use trusted machine for performing data transformations. One important distinct point between additive and multiplicative randomization is that, in additive randomization the original aggregate distribution can be reconstructed while in multiplicative randomization not only aggregate distribution that can be reconstructed but also more specific information such as distant between original value and modified value can be preserved.

Performing randomization strategy should also consider balancing trade-off between privacy and utility of a database. Therefore, an alternative solution to achieve good quality sanitized database is generating *conditional* noise which fits to values in the database. The term *conditional* refers to a flexibility in modeling

randomization process. Thus, it indicates that there is still challenge to design randomization technique which preserves privacy and retains useful information.

3.3. CRYPTOGRAPHIC BASED TECHNIQUES

Different from the previous one cryptographic based technique takes part in securing sensitive information from a data mining task under distributed computi5 system. In general criptographic techniques such as homomorphic encryption and secure multiparty computation are used in this technique.

Cryptographic is a very dynamic research field in computer science and mathematics. Presentluy, a lot of cryptographic techniques have been developed and successfully implemented in various areas of computer science including PPDM. Therefore, it is not surprising that this field attracts many researchers to utilize the cryptographic based techniques to design PPDM methods for preserving privacy in databases.

A. Secure Multiparty Computation

One of the mostly used techniques in PPDM which considers distributed system scenario is *Secure Multiparty Computation* (SMC). In this scenario, a data owner wants other partner perform computation over a database without revealing any private data in the database. Generally, SMC-based approaches consider semi-honest model where all involved parties permanently obey the protocol.

Pioneering work which employed cryptographic technique in set-valued database or transactional database for association rule mining which horizontally partitioned proposed in [23]. The proposed technique consists of five steps. First, all the involved parties should encrypt their itemsets using commutative encryption schemes. Second, each party exchanges its encrypted itemsets to another party. The party whose receipt the encrypted itemsets should re-encrypt it. Third, one party sends a token to another p15y. The token contains item frequency count and a random variable to its neighbor. Forth, the neighbor then adds its item frequency count and sends back the token to its party. The last, comparing betw2en the initiating party and final one to know whether the final result is higher than the defined threshold and its random value.

Another research in the same task also proposed in [24]4The technique is implemented to preserve privacy for *k*-means clustering task over vertically partitioned database. In every procedure of the clustering process, each data point successfully and securely computed to finds the smallest distant to its cluster center and mean value.

B. Homomorphic Encryption

Another research for protecting private information when linear regression model is performed in a database has been proposed in [25]. The method used fully homomorphic encryption schemes and assumed that all involved partners are semi-honest. Each independent attribute in the database is hold by different individual.

To encrypt plaintext x from attribute values, the method needs to perform a function $f : f(x) = (x + rp)c^r \% n$. While to decrypt the chipper-text y, it needs to perform $x = f^{-1}(y) = (b^{tr})^{-1}\, y \,\%\, p$, where $n = p \times q$, both p and q are prime, wh6 b is a primitive root mod $n * c$. In addition, r are positive integers less than n, while t is a discrete logarithm, $b^t \equiv c \pmod n$.

The following steps are the way of the proposed algorithm achieve privacy preserving linear regression model:

6

1) A key generator (KG) sends public encryption key and different private encryption keys to each partner. 6
2) Each partner performs encryption to their original data and sent the ciphertext result to data miner. Since data miner has not private decryption key, plaintext cannot be obtained.
3) Data miner perform calculation to generate encrypted coefficient correlation value E_β.
4) KG decrypts the coefficient correlation value to obtain the regression coefficient correlation value β.

Homomorphic encryption can also be implemented in set-valued database to find frequent association rule [26]. The scenario in that research assumes that there are two parties who own horizontally partitioned 26abase D_A and D_B, they want to determine the interesting association rules from the combination of their database $D = \{D_A \cup D_B\}$ without compromising individual sensitive condition. All the parties have their own secret number, later it will be used to encrypt the number of their frequent itemset. The first stage is each party A and B determine their global frequent itemset L_a based on the given minimum support s. To determine whether itemset from A and that from B are frequent, each party have to send the item counts to another party in the following way:

1) A sends its itemset count c_A and $|D_A|$ to B.
2) B also sends its itemset count c_B and $|D_B|$ to A.
3) Both A and B privately compute whether the itemsets are frequent using the following equation (Eq. 2).

$$\frac{c_A + c_B}{|D_A + D_B|} \geq \frac{s}{100} \qquad (2)$$

One each party determined its frequent itemset, then both parties should generate their global frequent itemset L_a. The last step is generating association rule from L_a using the given minimum confidence threshold c. Both parties should sp25each of their frequent itemset into two parts to generate all possible combinations of association rules from each itemset.

Even though the cryptographic schemes are quite promising for guaranteeing privacy protection in PPDM, it is still challenging to be implemented in a real situation since databases for data mining process usually have very large size which may lead to result in high computation costs and time consuming.

C. Heuristic Based Techniques

Achieving sanitized database with respect to preserve maximum data utility and maximum privacy protection is hard problem [6]. Therefore, a lot of heuristic approaches have been proposed to generate sanitized databases with acceptable privacy protection as well as retain enough data utility. An initial work called *Sanit* which follows heuristic algorithm has been proposed in [6]. The proposed method is specifically designed to protect sensitive items in transactional database.

Initially, *Sanit* generates a sorted graph of frequent itemset in descending order based on the items support value. Some sensitive items are then omitted from several records while minimizing a side effect such as frequent itemset lost. Aiming to hiding frequent sensitive itemset in transactional database several methods [27], [28]proposed heuristic method.

Different from the previous one, [27]proposed heuristic method called Item Grouping Algorithm (IGA). IGA groups itemsets into several identical cluster where in each cluster itemsets share the same sub-itemset. By performing such grouping IGA could assign a victim item in each group. If there are overlapping items among the groups, all the itemset in overlapping area will be removed, as a result each group only hold their distinct itemsets. Experimental results show that IGA successfully reduce misses cost which means sensitive itemset cannot be mined from sanitized database while non-sensitive itemset can still be mined in sanitized database.

Another heuristic technique for hiding sensitive frequent itemset also has been proposed in [29] namely Maximum Item Conflict First (MICF). The method can achieve sanitized database by removing sensitive items so that it reduces the support value of the sensitive items. There are several main steps in MICF such as identifying sensitive transactions or records (transaction containing sensitive itemset) from a database and determine a part of the transactions to be sanitized. For each sensitive transaction it decides an item to be removed, called victim item and perform data modification. The data modification result is re-written in a memory as a sanitized database.

The proposed method in [28] assumes that data owner has an ability to determine sensitive items in a database and define a support threshold for frequent itemset mining task. The hiding strategy firstly scan all records in the database then records that contain sensitive items are subjected to be modified while other records without any sensitive items are kept as is. To determine which sensitive items in a record that should be removed, they propose *degree sensitivity*, *si* as a boundary. Thus, any sensitive items which occurs more than the *si* value will be omitted from transaction.

In a situation where achieving exact result is NP-hard, heuristic approach is an alternative solution to generate a database which protect privacy and maintain data utility. Although the results might not be optimal, it usually applicable in real situation.

IV. Measurement

Generating sanitized database which achieves maximum privacy protecting and maintains data utility for knowledge discovery is NP-hard problem. Therefore, various techniques have been proposed since the last decades in which it results in various measurement techniques to evaluate the result of those methods.

4.1. Privacy Protection Measurement

To measure the privacy protection over a sanitized database [7] proposed a quantification metric for perturbation based technique. that is if a perturbed value can be estimated under a confidence level c % which belong to an interval $[x_{1.2}]$, then we can estimate the privacy by subtracting x_1 to x_2 with the confidence c %.

Measuring privacy in multiplicative random noise has also been described in [30]. In this measurement, it assumes that if x_1 is a original attribute value and x_2 is the distorted value of the x_1, we can estimates the original value using the following equation (Eq. 3).

$$\sqrt{\frac{\mathrm{Var}(x_1 - x_2)}{\mathrm{Var}(x_1)}} \quad (3)$$

Another important privacy measurement is called *hiding failure* (*HF*) which firstly introduced in [27]. This measurement plays an important role to quantify the balance between privacy and knowledge discovery in a database. The hiding failure calculates ratio between the number of sensitive frequent pattern in sanitized database that still can be mined $\#P(D')$ and the number of that in the original database $\#P(D)$, the formula of calculating *HF* is described in the following equation. A good data sanitization method would result in minimum percentage of *HF*. Therefore, since there is a trade-off between privacy and data utility designing a data sanitization method which can minimize *HF* or even zero *HF* is still a challenge. To compute *HF*, one can use the formula in Eq. 4.

$$HF = \frac{\#P(D')}{\#P(D)} \quad (4)$$

4.2. Utility Measurement

Measuring data utility in PPDM should also be taken into account since it represents the quality of a sanitized database. It is further acknowledged in [31] that there are no generic measurements to evaluate utility in sanitized database. Therefore, various data utility measurements have been proposed.

There are two important measurements to quantify data utility in PPDM, the first is called *Misses Cost* (MC) the second is *Artificial Pattern* (AP). MC refers to the number of non-sensitive patterns that are accidentally hidden due to performing PPDM algorithm. The formula to compute MC is stated in Eq. 5, where notations $\#P_{ns}(D)$ and $\#P_{ns}(D')$ denote the number of non-sensitive patterns in an original database and that in a sanitized database, respectively.

$$MC = \frac{\#P_{ns}(D) - \#P_{ns}(D')}{\#P_{ns}(D)} \quad (5)$$

Mean while, AP represents the number of artificial pattern that generated in sanitized database. Artificial pattern refers to an occurrence of patterns that previously does not exist in original database but it becomes exist in the sanitized database.

$$AP = \frac{|P(D)| - |P(D) \cap P(D')|}{|P(D)|} \quad (6)$$

Achieving lowest value or even zero of MC and AP is desirable designing PPDM algorithm. However, we should not that there is always trade-off between data utility and privacy.

A measurement called missclassification error (M_E) has also been proposed in [30] to evaluate the quality of sanitized database's for clustering task. M_E is basically measuring the number of information loss resulted by clustering algorithms. Misscalssification error can be computed using the following equation in Eq. 7.

$$M_E = \frac{1}{N}\sum_{i=1}^{k}(|C_i(D)| - |C_i(D')|) \quad (7)$$

The notation N refers to the number of points in the original database while k is the number of clusters. $|C_i(D)|$ and $|C_i(D')|$ represent the number of data point in cluster ith from original database and that in sanitized database, respectively. Since data sanitation somehow changes the values inside the database, it is important to maintain the consistency of the clustering results.

4.3. Simmilarity Measurement

Measuring similarity of a sanitized database should also be taken into account since it represents the closeness between an original database and sanitized database. It is further believed that by knowing the similarity between those to databases, data owner can avoid disbelief from the database recipients [32].

To measure similarity in transactional database, [33] proposed dissimilarity measurement ($Diss$). The underlying idea of such measurement is comparing the histogram frequency of items in an original database with that of the sanitized one.

$$Diss = \frac{\sum_{i=1}^{n} |fD(i) - fD^i(i)}{\sum_{i=1}^{n} fD(i)} \quad (8)$$

As described in Eq. 8, $fD(i)$ represents frequency item i in original database, whereas $fD^i(i)$ refers to frequency of item i in the sanitized database. It is obvious that dissimilarity between original and sanitized database should be minimized to provide acceptable data similarity in knowledge discovery process.

V. Research Opportunities

In today's era where most people are connected to the Internet and do their activity on-line in many different ways individual privacy protection is an interesting issue to be explored. It is generally known that each individual may have different concern about privacy, for example one may think that his political view is sensitive information while some other do not think so. Thus, developing some ideas to guarantee individual privacy while does not change the general data pattern is an interesting issue.

Looking from the fact that generated data in this era e.g. mobile technology and IOT technology results in various types and abundant amount of data size, designing distributed PPDM algorithms that resilient to handle very large database with ensuring its communication security and data integrity will be very prominent in the future.

VI. Conclusion

PPDM is one of the field study in data mining area which aims to protect private information in a database that might be leaked during knowledge discovery process. Various PPDM algorithms have been proposed to not only ensuring privacy protection but also maintaining data usefulness from a modified database. However, there are still many areas to be explored.

Since each algorithm has its own design purpose, none of the proposed algorithm can fit to protect privacy from different mining tasks. The implementation of PPDM algorithms should also consider the type of databases that are used whether it is statistical database, categorical database, or transactional database since different types of databases need different treatments.

Even tough, some PPDM algorithms seems very promising in protecting privacy and data utility based on its empirical studies, we still need to ensure their applicability and effectiveness with respect the performance and computation costs due to data mining tasks usually involves very large database.

Ensuring PPDM algorithms results is also another important thing. Thus, various measurement tools have also been suggested to evaluate performance of the PPDM algorithms. However, utilizing one metric is not adequate since there might be multiple parameters in a database that should be evaluated. Moreover, the proposed measurements tools are application specific as a result, it is difficult to compare between the existing PPDM techniques.

References

[1] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Advances in Knowledge Discovery and Data Mining," U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996, pp. 1–34.

[2] H. Mark, M. Erik, and V. Sunil, *Java Data Mining: Strategy, Standard, and Practice*, 1st Editio. Morgan Kaufmann, 2006.

[3] Merdeka.com, "Mafia Jual Beli Data Pribadi." 2020.

[4] S. Xingzhi and P. S. Yu, "A border-based approach for hiding sensitive frequent itemsets," 2005, doi: 10.1109/ICDM.2005.2.

[5] X. Sun and P. S. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," in *Proceedings of the Fifth IEEE International Conference on Data Mining*, 2005, pp. 426–433, doi: 10.1109/ICDM.2005.2.

[6] M. Atallah, A. Elmagarmid, M. Ibrahim, E. Bertino, and V. Verykios, "Disclosure Limitation of Sensitive Rules," in *Proceedings of the 1999 Workshop on Knowledge and Data*

Engineering Exchange, 1999, pp. 45--, [Online]. Available: http://dl.acm.org/citation.cfm?id=519168.788219.

[7] R. Agrawal and R. Srikant, "Privacy-preserving Data Mining," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, pp. 439–450, doi: 10.1145/342009.335438.

[8] B. Pinkas, "Cryptographic techniques for privacy-preserving data mining," *ACM SIGKDD Explorations Newsletter*, 2002, doi: 10.1145/772862.772865.

[9] Y. Lindell and B. Pinkas, "Privacy preserving data mining," *Journal of Cryptology*, 2003, doi: 10.1007/s00145-001-0019-2.

[10] T. Hsu, C.-J. Liau, and D.-W. Wang, "A Logical Model for Privacy Protection," in *Information Security*, 2001, pp. 110–124.

[11] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association Rule Hiding," *IEEE Transactions on Knowledge and Data Engineering*, 2004, doi: 10.1109/TKDE.2004.1269668.

[12] V. S. Verykios *et al.*, "State-of-the-art in Privacy Preserving Data Mining Classification of Privacy Pre," *ACM SIGMOD Record*, vol. 33, no. 1, pp. 50–57, 2004, doi: 10.1145/974121.974131.

[13] L. Chun-Wei, H. Tzung-Pei, C. Chia-Ching, and W. Shyue-Liang, "A Greedy-based Approach for Hiding Sensitive Itemsets by Transaction Insertion," *Journal of Information Hiding and Multimedia Signal Processing.*, vol. 4, no. 4, pp. 201–2014, 2013.

[14] J.-L. Lin and Y.-W. Cheng, "Privacy Preserving Itemset Mining Through Noisy Items," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5711–5717, Apr. 2009, doi: 10.1016/j.eswa.2008.06.052.

[15] L. Liu, M. Kantarcioglu, and B. Thuraisingham, "The Applicability of the Perturbation Based Privacy Preserving Data Mining for Real-world Data," *Data Knowl. Eng.*, vol. 65, no. 1, pp. 5–21, Apr. 2008, doi: 10.1016/j.datak.2007.06.011.

[16] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 217–228, doi: 10.1145/775047.775080.

[17] J. Domingo-Ferrer and V. Torra, "Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation," *Data Min. Knowl. Discov.*, vol. 11, no. 2, pp. 195–212, Sep. 2005, doi: 10.1007/s10618-005-0007-5.

[18] C. C. Aggarwal and P. S. Yu, *Privacy-Preserving Data Mining: Models and Algorithms*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[19] G. Navarro-Arribas, V. Torra, A. Erola, and J. Castellà-Roca, "User k-anonymity for privacy preserving data mining of query logs," *Information Processing & Management*, vol. 48, no. 3, pp. 476–487, 2012, doi: https://doi.org/10.1016/j.ipm.2011.01.004.

[20] T. Dalenius and S. P. Reiss, "Data-swapping: A technique for disclosure control," *Journal of Statistical Planning and Inference*, vol. 6, no. 1, pp. 73–85, 1982, doi: https://doi.org/10.1016/0378-3758(82)90058-1.

[21] W. E. Winkler, "Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems," in *Privacy in Statistical Databases*, 2004, pp. 231–246.

[22] D. P, lane J.i, T. J.J.M, and Z. L.V, *Confidentiality, disclosure, and data acces: theory and practical applications for statistical agencies*. Amsterdam: Elsevier Science, 2001.

[23] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," *IEEE Trans. on Knowl. and Data Eng.*, vol. 16, no. 9, pp. 1026–1037, Sep. 2004, doi: 10.1109/TKDE.2004.45.

[24] J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 639–644, doi: 10.1145/775047.775142.

[25] W. Fang, C. Zhou, and B. Yang, "Privacy preserving linear regression modeling of distributed databases," *Optimization Letters*, vol. 7, no. 4, pp. 807–818, Apr. 2013, doi: 10.1007/s11590-012-0482-8.

[26] M. G. Kaosar, R. Paulet, and X. Yi, "Fully Homomorphic Encryption Based Two-party Association Rule Mining," *Data Knowl. Eng.*, vol. 76–78, pp. 1–15, Jun. 2012, doi: 10.1016/j.datak.2012.03.003.

[27] S. R. M. Oliveira and O. R. Zaane, "Privacy Preserving Frequent Itemset Mining," *Proceedings of the IEEE international conference on Privacy, security and data mining*, 2002.

[28] D. Gunawan and G. Lee, "Heuristic Approach on Protecting Sensitive Frequent Itemsets in Parallel Computing Environment," in *The 1ST UMM International Conference on Pure and Applied Research (UMM-ICOPAR 2015)*, 2015, pp. 41–49.

[29] Y.-C. Li, J.-S. Yeh, and C.-C. Chang, "MICF: An effective sanitization algorithm for hiding sensitive patterns on data mining," *Advanced Engineering Informatics*, vol. 21, no. 3, pp. 269–280, 2007, doi: 10.1016/j.aei.2006.12.003.

[30] S. R. M. Oliveira and O. R. Zaïane, "Privacy Preserving Clustering By Data Transformation," *Proc. of the 18th Brazilian Symposium on Databases*, pp. 304–318, 2003, doi: 10.1.1.2.42.

[31] J. Domingo-Ferrer and V. Torra, "Disclosure risk assessment in statistical data protection," *Journal of Computational and Applied Mathematics*, 2004, doi: 10.1016/S0377-0427(03)00643-5.

[32] D. Gunawan and M. Mambo, "Set-valued data anonymization maintaining data utility and data property," Jan. 2018, doi: 10.1145/3164541.3164583.

[33] S. R. M. Oliveira and O. R. Zaiane, "Privacy preserving frequent itemset mining," *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*, vol. 14, pp. 43–54, 2002, [Online]. Available: http://portal.acm.org/citation.cfm?id=850782.850789.

Classification of Privacy Preserving Data Mining Algorithms: A review

ORIGINALITY REPORT

15%

SIMILARITY INDEX

PRIMARY SOURCES

1	publikasiilmiah.ums.ac.id Internet	129 words — 2%
2	www.yumpu.com Internet	112 words — 2%
3	"Encyclopedia of Database Systems", Springer Science and Business Media LLC, 2018 Crossref	60 words — 1%
4	link.springer.com Internet	33 words — 1%
5	mafiadoc.com Internet	28 words — 1%
6	Weiwei Fang, Changsheng Zhou, Bingru Yang. "Privacy preserving linear regression modeling of distributed databases", Optimization Letters, 2012 Crossref	27 words — 1%
7	Tzung-Pei Hong. "Hiding sensitive itemsets by inserting dummy transactions", 2011 IEEE International Conference on Granular Computing, 11/2011 Crossref	24 words — < 1%
8	M HORNICK. "Overview of Data Mining", Java Data Mining, 2007 Crossref	23 words — < 1%
9	En Tzu Wang, Guanling Lee. "An efficient sanitization algorithm for balancing information	20 words — < 1%

privacy and knowledge discovery in association patterns mining", Data & Knowledge Engineering, 2008
Crossref

10 vneumann.etse.urv.es
Internet
18 words — < 1%

11 www.jatit.org
Internet
17 words — < 1%

12 eden.dei.uc.pt
Internet
14 words — < 1%

13 Jerry Chun-Wei Lin, Qiankun Liu, Philippe Fournier-Viger, Tzung-Pei Hong, Miroslav Voznak, Justin Zhan. "A sanitization approach for hiding sensitive itemsets based on particle swarm optimization", Engineering Applications of Artificial Intelligence, 2016
Crossref
14 words — < 1%

14 Solikin, Mochamad, and Budi Setiawan. "Mechanical Properties of Class C High Volume Fly Ash Concrete with Lime Water as Mixing Water", Applied Mechanics and Materials, 2014.
Crossref
14 words — < 1%

15 I. Saleh, A. Mokhtar, A. Shoukry, M. Eltoweissy. "P3ARM: Privacy-Preserving Protocol for Association Rule Mining", 2006 IEEE Information Assurance Workshop, 2006
Crossref
13 words — < 1%

16 Chirag Modi. "A Survey on Preserving Privacy for Sensitive Association Rules in Databases", Communications in Computer and Information Science, 2010
Crossref
12 words — < 1%

17 Ahmet Cumhur Öztürk, Belgin Ergenç. "Dynamic Itemset Hiding Algorithm for Multiple Sensitive Support Thresholds", International Journal of Data Warehousing and Mining, 2018
Crossref
11 words — < 1%

18	research.sabanciuniv.edu Internet	11 words — < 1%
19	Josep Domingo-Ferrer. "Statistical Databases", Wiley, 2008 Crossref	11 words — < 1%
20	ddd.uab.cat Internet	11 words — < 1%
21	www.iaeng.org Internet	10 words — < 1%
22	www.ijicic.org Internet	9 words — < 1%
23	Amiri, A.. "Dare to share: Protecting sensitive knowledge with data sanitization", Decision Support Systems, 200702 Crossref	9 words — < 1%
24	Cheng, Peng, Jeng-Shyang Pan, and Chun-Wei Lin Harbin. "Use EMO to protect sensitive knowledge in association rule mining by removing items", 2014 IEEE Congress on Evolutionary Computation (CEC), 2014. Crossref	9 words — < 1%
25	vuir.vu.edu.au Internet	9 words — < 1%
26	Kaosar, M.G.. "Fully homomorphic encryption based two-party association rule mining", Data & Knowledge Engineering, 201206/08 Crossref	9 words — < 1%
27	Janakiramaiah Bonam, Ramamohan Reddy. "Balanced Approach for Hiding Sensitive Association Rules in Data Sharing Environment", International Journal of Information Security and Privacy, 2014 Crossref	9 words — < 1%
	media.proquest.com	

28	Internet	9 words — < 1%
29	ir.nuk.edu.tw:8080 Internet	8 words — < 1%
30	www.iis.sinica.edu.tw Internet	8 words — < 1%
31	Yi-hung Wu. "<![CDATA[Hiding Sensitive Association Rules with Limited Side Effects]]>", IEEE Transactions on Knowledge and Data Engineering, 1/2007 Crossref	8 words — < 1%
32	Mortazavi, Reza, and Saeed Jalili. "Fast data-oriented microaggregation algorithm for large numerical datasets", Knowledge-Based Systems, 2014. Crossref	8 words — < 1%
33	www.ijert.org Internet	8 words — < 1%
34	theses.gla.ac.uk Internet	8 words — < 1%
35	Navarro-Arribas, G.. "Information fusion in data privacy: A survey", Information Fusion, 201210 Crossref	7 words — < 1%
36	J. Domingo-Ferrer, A. Solanas, A. Martinez-Balleste. "Privacy in Statistical Databases: k-Anonymity Through Microaggregation", 2006 IEEE International Conference on Granular Computing, 2006 Crossref	7 words — < 1%
37	Chang, C.C.. "TFRP: An efficient microaggregation algorithm for statistical disclosure control", The Journal of Systems & Software, 200711 Crossref	7 words — < 1%
38	Chai Wah Wu. "Privacy preserving data mining with unidirectional interaction", 2005 IEEE International	7 words — < 1%

Symposium on Circuits and Systems, 2005
Crossref

39 Shuo Qiu, Boyang Wang, Ming Li, Jiqiang Liu, Yanfeng Shi. "Toward Practical Privacy-Preserving Frequent Itemset Mining on Encrypted Cloud Data", IEEE Transactions on Cloud Computing, 2020 — 7 words — < 1%
Crossref

40 "Encyclopedia of Database Systems", Springer Science and Business Media LLC, 2009 — 7 words — < 1%
Crossref

41 "Privacy-Preserving Data Mining", Springer Science and Business Media LLC, 2008 — 7 words — < 1%
Crossref

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		