PMA for Privacy Preservation in Data Mining

V. Jane Varamani Sulekha
Assistant Professor
Department of Information Technology
Fatima College
Mary land, Madurai, Tamilnadu, India

Dr. G. Arumugam
Senior Professor and Head, Department of Computer Science
Madurai Kamaraj University
Madurai, India

***Abstract*—Privacy is becoming a progressively important issue in many data mining applications. This has initiated the development of many privacy preserving data mining techniques. In recent years, various data mining algorithms combining privacy preserving techniques have been established that hide sensitive identifiers or patterns. When applying privacy preservation techniques, importance is given to the utility and information loss. In this paper we propose Statistical Disclosure Control (SDC) based Perturbed Micro Aggregation (PMA) for anonymizing the individual records. Through the experimental results, the proposed technique is validated to prevent the disclosure of sensitive data without degradation of data utilization. Our work highlights some discussions about future work and promising directions in the perspective of privacy preservation in data mining.**

Keywords—PPDM; privacy; microaggregation; microdata; anonymization; data mining

I. INTRODUCTION

Recent advances in data collection, data distribution and related technologies have initiated a new area of research in which existing data mining algorithms should be reevaluated from a different point of view, the privacy preservation. Different communication channels through internet services such as electronic commerce, online-banking, research, social media and online trade, have stretched to a level where threats against the confidentiality are very common on a daily basis and they need serious thinking about privacy. The definition of privacy has been properly stated in [1] as “The right of an individual to be protected from unauthorized disclosure of sensitive information that are confined in an electronic repository or that can be inferred as aggregate and complex information from data stored in an electronic repository”. In other words Privacy relates to specific information that a person would not wish others to know without authorization, and to a person’s right to be free from the attention of others (UN Declaration of Human Rights, 1948).

The term, Privacy Preserving Data Mining (PPDM) was first introduced by Lindell, Y., & Pinkas, B. [2]. Various solutions have been proposed by researchers. Noise addition, Perturbation, Blocking, Anonymization, Aggregation, Swapping, Sampling, Sanitization, Differential privacy, Condensation, Cryptography and Evolutionary algorithms based transformation are some of the Privacy preservation techniques. These techniques fall into two main categories. The first category is Data Modification and the second one is Secure Multiparty Computation (SMC). The first category of the data modification approach trades privacy with improved performance. These techniques allow a data owner to transform its data in different ways to hide the sensitive attributes of the original data but at the same time still allow useful mining operations over the transformed data. SMC method provides robust level of privacy. Any data mining algorithm can be executed by using generic algorithms of SMC [3]. But, these algorithms are extremely expensive in practice, and impractical for real use. Our paper concentrates on the Statistical Disclosure Control (SDC) based microaggregation method. The advantage of this method is the minimization of the information loss. This classification is shown in Figure. 1.

Identifying sensitive attribute and modifying that attribute is an emerging technique in PPDM. Data distortion method and Probability distribution in the form of randomization method attempt to hide the sensitive attribute. Modification of sensitive attribute can also be done using noise obfuscation, amplification and substitution. Often sensitive attributes are manipulated using noise. We can group Data Distortion, Data Randomization, Noise Obfuscation, Amplification and Random Substitutions under Perturbation based PPDM. There are two main categories in data perturbation, one based on probability distribution and another one fixed data perturbation. In the probability distribution, sensitive value is replaced with some distribution sample. Fixed data perturbation methods are used for numerical, categorical data. In perturbation methods, a sensitive attribute is perturbed by addition of a noise term e, to get a perturbed attribute Y = X + e. This Method is known as Additive Data Perturbation (ADP). Likewise in Multiplicative Data Perturbation (MDP), perturbed attribute Y = Xe. Data distortion by probability distribution is proposed in [4]. Original dataset is replaced with distorted dataset generated using probability distribution. Gaussian perturbation or Uniform Perturbation based randomizing function [5] is used to perturb the sensitive values. Amplification for limiting privacy is proposed in [6]. If all the sensitive values x are reasonably randomized into a given y, then randomized value R(x) =y does not reveal anything about x. Here the probability is amplified. A random matrix-based spectral filtering technique [7] is used for perturbation. It can also be used to reconstruct original data from the distorted data. Random rotation perturbation method [8] for privacy preserving data classification, without breaching privacy and without loss of information, is used. Multi-dimensional perturbation techniques are addressed through rotation transformation. Multiplicative random projection matrices [9] are used for privacy preserving distributed data mining. It is based on the Johnson-Lindenstrauss lemma.

Geometric Perturbation Technique [10] is a combination of Rotation perturbation, Translation perturbation and Noise addition perturbation. Gaussian random vector is used. It shows that the enhancement in geometric perturbation can provide satisfactory results.

SDC is a technique used in data-driven research to ensure no person or organization is identifiable from the results of the analysis of survey or administrative data, to protect the confidentiality of the respondents and subjects of the research [11]. This technique attempts to have a balance between a person's right to privacy and the right of a society to know about the data for analyses. SDC tries to protect statistical data so that they can be publicly released and mined without giving away secret information that can be related to specific people or entities. Microaggregation is an efficient Inference Control randomization technique for microdata protection, i.e. protection of sensitive information.

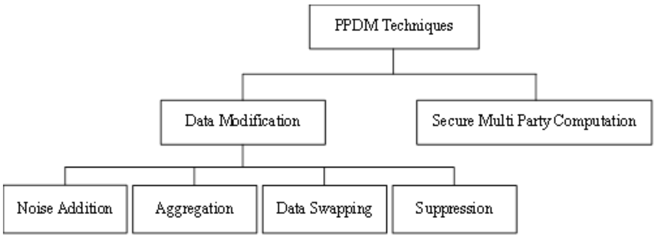


Fig. 1. Taxonomy of PPDM Techniques

Microaggregation technique modifies data without suppressing or generalizing it. Microaggregation [12] is addressed as a clustering problem, where large dataset is partitioned into small and similar groups. Each group comprises at least k records and instead of publishing the raw microdata values, the mean of the group they belong to is replaced in their original place prior to their publishing or release. K-anonymity is one of the most classic models designed to protect individual privacy [13]. In the *k*-anonymous tables, a dataset is *k*-anonymous ($k \geq 1$) if each record in the dataset is indistinguishable from at least (k-1) other records within the same dataset. The larger the value of *k*, the better the privacy is protected. We can say that microaggregation satisfies k-Anonymity. Simply Partitioning or clustering a dataset into homogenous groups is not considered as microaggregation. Strictly each group consists of at least k records. It is important to group the records with minimal disclosure risk and high data utility. In other words, we can state that a better trade-off is required between the risks of revealing the sensitive information and the information loss occurred due to data modification.

The optimum microaggregation technique partitions a dataset into groups of size lying between k and 2k-1. The user specific variable k decides the degree of randomization, a great value of k may ensure the highest data privacy, but the data may not be useful for data mining analyses as information loss may be higher. Usually, for adequate size dataset the k value can be 3, 4, 5 or 10 in any microaggregation technique. The remainder of this paper is organized as follows. Section 2 provides some significant concepts about microaggregation. Section 3 reviews different Microaggregation techniques. Section 4 introduces our proposed method Perturbed Micro Aggregation (PMA). Section 5 presents experimental results and Section 6 describes considerations about future extensions and promising directions in the perspective of privacy preserving data mining.

II. MICROAGGREGATION

Microaggregation was first proposed by Defays, D. and Anwar, N. [12]. In Microaggregation the individual values are replaced by values computed on small aggregates prior to releasing. In other words, instead of releasing the actual values of the individual records, the system releases the mean of the group (or median, mode, weighted average) to which the observation belongs. Microaggregation technique has two phases. Partitioning, in this phase the original micro dataset is partitioned into several disjointed clusters/groups so that all records in the same group are very much related to each other and, simultaneously, dissimilar to the records in other groups. Additionally, each group is forced to contain at least *k* records. Next Phase aggregation, computes aggregated value for each cluster/ group, and it replaces the original values in the micro dataset by the computed aggregated value. This phase usually depends on the type of the variable concerned. Microaggregation technique requires a clustering method and an aggregation method. Microaggregation methods were originally used for numerical data types. Figure 2, Shows examples of micro aggregated data where the original values are replaced by mean.

A micro dataset is a file or a table with n records and m attributes. The attributes can be classified into four categories, generally they are not disjoint. They are Identifiers, Quasi-Identifiers, Confidential Outcome attributes and Non Confidential attributes. Identifiers are used to identify the individual person. Name, Passport number and social security number are examples of identifiers. A combination of Quasi-identifiers can be used to identify the individual person. Address, gender, age, telephone and pincode are examples of Quasi-Identifiers. Confidential outcome attributes describe the individual person. Salary, Religion and health condition are few examples of confidential outcome attributes. Non confidential attributes will not reveal any sensitive information about the person. A micro dataset with n records can be micro aggregated by forming different groups with size at least k. Each attribute is replaced with the average of the group that the attribute belongs. Usually groups are formed with maximal similarities. After updating the original value, the resulting records can be released for mining. The ideal k-partition with minimum information loss is defined to be the one that maximizes the group similarity. The higher the group similarity, the lower the information loss.

Microaggregation replaces values in a group by the group mean. The sum of squares criterion is used to measure the similarity in clusters. Within the group sum of squares SSE is stated as

$$SSE = \sum_{i-1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)' (x_{ij} - \bar{x}_i) \qquad (1)$$

The lower SSE, the similarity is higher in the cluster. The between group sum of squares SSA is stated as

$$SSA = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{x})'(\bar{x}_i - \bar{x}) \qquad (2)$$

The total sum of squares SST is stated as

$$SST = \sum_{i-1}^{g} \quad \sum_{j=1}^{n_i}(x_{ij} - \bar{x})'(x_{ij} - \bar{x}) \qquad (3)$$

Information Loss (IL) is standardized between 0 to 1 and defined as

$$IL = \frac{SSE}{SST}$$

According to the dimensionality of the data, Microaggregation can be classified into univariate and multivariate microaggregation. Univariate microaggregation is applied to each variable of a micro dataset. Complexity is lesser when single variable is involved, at the same time utility should be considered. In Individual ranking based technique of univariate microaggregation [14], data vectors are ranked by the first variable and then these variables are micro aggregated. Variables are grouped together, and microaggregation is applied in Multivariate aggregation. Multivariate fixed-size microaggregation and Multivariate data-oriented microaggregation using multivariate distance is proposed by Mateo Sanz, J. M., & Domingo Ferrer, J. [15]. Without considering the dimensionality, partitioning of micro dataset can be classified as fixed size and variable size partition. In fixed size partition micro dataset is partitioned into groups of size k, except one group which may have more than k records when the number of records in the micro dataset is not a multiple of k. Group size between k and 2k-1. In variable size partition groups have variable sizes. Fixed size microaggregation takes less computation time in partitioning the dataset, but the variable size partition method is more flexible.

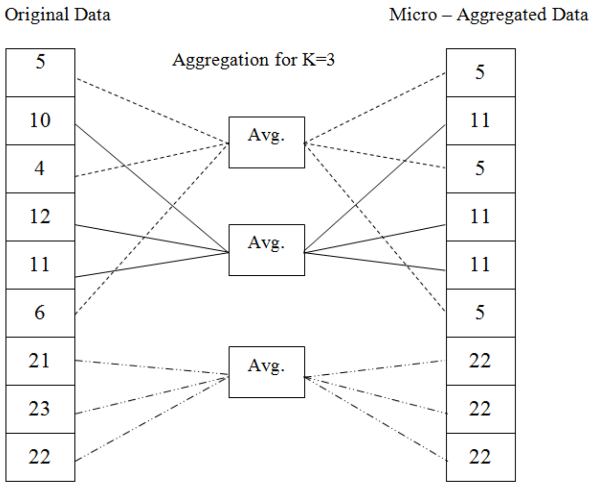


Fig. 2. Example of Microaggregation using Mean

III. MICROAGGREGATION METHODS

Microaggregation methods have been divided into two categories, namely Fixed Size [14] and Data Oriented microaggregation [16]. For Fixed Size microaggregation, the grouping is done by dividing a dataset into clusters that have size k, but one cluster may have a size between k and 2k-1. It depends on the value k and total numbers of records n. Fixed Size methods reduce space complexity, and thus are more efficient than Data Oriented methods. Data oriented methods may achieve lower information loss than Fixed Size methods. Computational complexity of optimal microaggregation [17], with minimal information loss for a fixed security level, is proposed.

The Maximum Distance (MD) Method [14] is proposed as a multivariate microaggregation method. The advantage of this method is its simplicity and performance. The main shortcoming of this method is its computational complexity, i.e. O(n^3). Microaggregation problem is formulated as a shortest path problem on a graph. First graph is constructed, then each arc of the graph corresponds to a possible group may be considered as an optimal partition. Each arc is labeled by the error so that it will restrict the group to be included in the partition. This method is known as optimal microaggregation method [18].

Minimum Spanning Tree Partitioning (MSTP) for microaggregation [19] is proposed as a variable size multivariate microaggregation method. This method first builds Minimum Spanning Tree (MST) using Prim Method. But the standard MST partitioning algorithm does not give solution to the microaggregation problem as the group size is not considered in the algorithm. When the oversized clusters are further divided into small clusters, the MSTP algorithm works for the micro aggregated problem with this small modification. When data points are distributed in a scattered way, MSTP performance will decrease.

Maximum Distance to Average Vector Method (MDAV) [20], is a Multivariate Fixed size microaggregation method employed in the μ-Argus package for statistical disclosure control. It is based on forming groups, with the distance between centroid and distinct data. The disadvantage of MDAV is it's not flexible. It only generates subsets of fixed cardinality k. Performance degradation will occur if the data points are scattered in the clusters. Variable Size MDAV or V-MDAV [21] in contrast with fixed size MDAV, produces k partitions with group sizes varying between *k* and 2*k*-1. This flexibility can be used to achieve similarity within the group and optimal partition of data.

Micro aggregation based p-sensitive k-anonymity [22] is proposed. Its idea is that with the same grouping of key attribute values, the number of different values for each confidential attribute is at least p within the same group. Two Fixed Reference Points (TFRP) [23], is proposed. TFRP has two stages, denoted as TFRP-1 and TFRP-2. In the first phase, TFRP uses a fixed size algorithm to partition the dataset. In the second phase, TFRP reduces the number of partitions produced by the first phase to improve the data quality. For sparse datasets with large k value TFRP produces a very low information loss. A new method called micro hybrid [24] is proposed. This method first partitions the dataset into clusters containing k and 2k-1 records. By applying the synthetic data generator algorithm, synthetic version of each cluster is obtained. Then the original records are replaced in each cluster by the records in the equivalent synthetic cluster. The micro hybrid method is a simple approach to preserve privacy of data. It can be applied to any data type and can yield groups of variable size. The means and covariance of the confidential attributes in original dataset and synthetic dataset exactly the same. Thus utility is preserved.

A Density Based Algorithm (DBA) [25] for microaggregation is proposed. The DBA has two phases. First Phase (DBA-1), partitions the dataset into groups in which each group contains at least *k* records. To partition the dataset, it uses k nearest neighborhood of the record with the maximum *k*-density among all the records that are not allocated to any group. The grouping procedure continues till *k* records remain unassigned. These remaining *k* records are then assigned to its nearest groups. The second phase (DBA-2) is then applied to further tune the partition in order to achieve small information loss and maximum data utility. *DBA*-2 may decompose the formed groups or may merge its records to other groups. Microdata Protection Method through Microaggregation based on Median [26], is proposed. It divides the whole micro dataset into a number of exhaustive and mutually exclusive groups before publication. After grouping it publishes the median instead of individual records. It promises that the modification does not affect the result. Modified data and the original data are similar in this method.

T-Closeness through Microaggregation [27] primarily generates a cluster of size k based on the quasi-identifier attributes. Then the cluster is iteratively refined until t-closeness is satisfied. In the refinement, the algorithm checks whether t-closeness is satisfied and, if it is not, it selects the closest record not in the cluster based on the quasi-identifiers and swaps it with a record in the cluster selected. It takes the t-closeness requirement into account at the moment of cluster formation during microaggregation and this provides best results.

IV. PROPOSED METHOD

A. *Privacy Preservation*

Privacy preservation can be best illustrated through the following example. Consider a hospital that collects a database of disease information which could be useful for research purpose. At the same time, it is important for the hospital to take safety measures by protecting the privacy of patients. For example, hiding the identities of individual persons and protecting other sensitive information such as Cancer Disease. Here, according to the utility based pattern, selection of data can be done. A typical data mining relies on data owner to define what kind of pattern they are going to mine. Each data mining application may need a unique kind of data, instead of releasing the whole dataset, utility based on the preferences in the parameters of the dataset can be released for data mining. This will improve the computation time and storage space. For example, preference could be disease in the age group between 30 to 50, Raised cholesterol and obesity level in males over 40, buying pattern of the metropolitan population or climatic disease in a particular area. This method also reduces the risk of individual disclosure and data mining algorithm complexity.

As Han, J., and Kamber, M. [28] state, a data mining system has the capability to generate thousands or even millions of patterns. But a pattern is interesting if it is potentially useful. Though objective measures help to identify interesting patterns, they are often insufficient. It should be combined with subjective measures that reflect a particular user’s interests and needs. For instance, the hospital data is released to data miner for modeling causes of diseases. The patterns describing the disease among patients of a hospital would be interesting to the hospital administration, but of slight interest to other analysts reviewing the same database. Normally, it is not the responsibility for a data owner to build models but it is the responsibility for a data owner to preserve data privacy when the data is released for data mining. The data owner has to execute a protection technique with different preferred utility based parameters to attain a desired trade-off between privacy and utility. The data owner can choose a more preferred utility based dataset from a set of non-dominated dataset. Also, it is necessary for data mining systems to generate interesting patterns, as one need not examine the pattern generated to identify the really interesting ones. Considering this a novel privacy preservation technique is proposed in this work and it is assumed that the preferred utility based dataset contains quasi identifiers.

B. *Perturbed Micro Aggregation (PMA)*

Existing microaggregation techniques replace the original values with computed aggregates like mean, median, mode and centroid. These aggregated values can be reconstructed and may violate privacy. To overcome this problem we develop a new algorithm called Perturbed Micro Aggregation (PMA), which assures privacy and utility. In addition, preference based dataset can also be obtained by this method. We present an approach that combines microaggregation and ε differential privacy based perturbation which ensure low information loss and guarantees privacy. Figure 3 describes the perturbation model.

Perturbed Micro Aggregation can be divided into two major parts Microaggregation and Noise Addition. In Microaggregation phase, K ward hierarchical clustering algorithm is used to partition the dataset. In our work we are taking the variable age as Preference Based Variable (PBV) which is used for single dimension partition. By using K-ward algorithm, dataset is grouped into n partitions based on the PBV. Then groups of k successive values of the PBV are formed and, inside each group, values are replaced by the group mean. After the microaggregation, Laplace noise is added to each micro aggregated value and this perturbed dataset is released for mining. For numerical attributes noise is usually added using a random number. This random number is generally derived from a normal distribution with small standard deviation and zero mean. Noise is added in a controlled way so that it won’t affect the mining result.

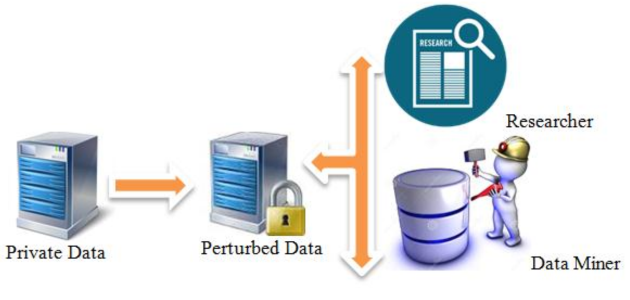


Fig. 3. Perturbation Model

A randomized function K gives ε-differential privacy [29] if for all datasets D1 and D2 differing on at most one element, and for all S ⊆ Range(K), Pr[K(D1) ∈ S] ≤ exp(ε) × Pr[K(D2) ∈ S]. Laplace noise addition is the primary method that has been advanced for satisfying the differential privacy. The main idea of our proposed method is to form groups using *K* ward hierarchical clustering algorithm [30]. Dataset is grouped into n partitions. Original sensitive attribute in each cluster is replaced with its micro aggregated value "mean". Then Laplace noise is added to each micro aggregated value and this perturbed dataset is released for mining. Figure 4, describes the proposed perturbation Framework. For numerical attributes noise is usually added using a random number. This random number is generally derived from a normal distribution with small standard deviation and zero mean. Noise is added in a controlled way so that it won't affect the mining result. From the survey we found that, Laplace noise addition satisfies the differential privacy.

X denotes all the attributes of the original dataset. X′ denotes the perturbed dataset. When the original data is replaced with the cluster mean, the sensitivity of the dataset will be represented as $\Delta x/k$. where Δx is the distance between the most distant records in the cluster. The sensitivity of the whole dataset is $n/k \times \Delta x/k$. To obtain differential privacy, Laplace noise (n/k × Δx/k)/ε is added to the numerical data.

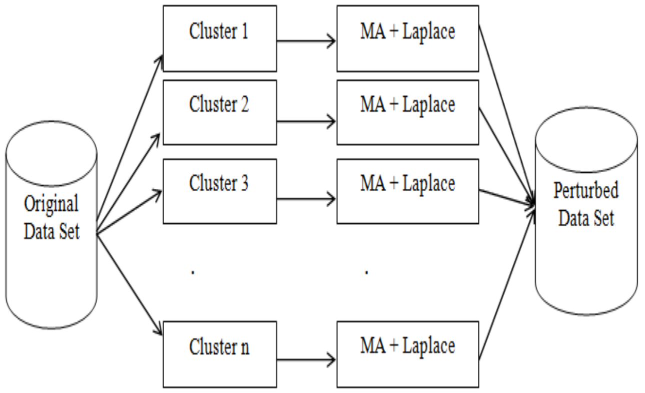


Fig. 4. Proposed Framework

Pseudocode of our proposed work:

Step1. Form a cluster with the first k elements of the original dataset and another group with the last k elements of the original dataset.

Step2. Use Wards method until all elements in the original dataset belong to a group containing k or more data elements. In this process of forming groups by Wards method, never join two groups which have both a size greater than or equal to k.

Step3. For each group in the final partition that contains 2k or more data elements, apply this algorithm recursively. Within each cluster, the entire attribute values are replaced by the cluster mean, so each micro aggregated cluster consists of k repeated mean values.

Step4. Add Laplace noise (n/k *Δx/k)/ε to each attribute in the clusters.

The first step ensures that in each recursive step the dataset is split into at least 2 groups. The second step ensures that the formed groups are never combined because of their size. Third step guarantees *k* anonymity, with 2k or more elements. The last step ensures privacy of individual record. In case the clusters formation is difficult, data can be projected on to a single axis can solve the problem or any one of the distance measures can be applied to find the most distant elements and the groups around the distant measure can be combined to form clusters.

We combine microaggregation and ε differential privacy. This combination gives better performance. Low information loss and privacy guarantee will be obtained by this method. This algorithm can be applied to the whole dataset or else preference based dataset. The main difference between our proposed method with the previous microaggregation algorithm is that, the given method can produce multiple protected univariate numerical dataset, which can be either used as a whole dataset or else preference based dataset mentioned in our earlier work [31]. In each partition, the perturbation method applied is different, so it may restrict the reconstruction problem. The perturbed dataset obtained from original dataset will give the same mining result while applying classification or clustering algorithm. This method reduces the risk of individual disclosure.

C. Chronic Kidney Disease Analysis

We consider the single dimension partition based on age. Age is the preference based variable and partition is done on age and the preference based dataset based on age is released for mining. If a hospital wishes to know the Chronic Kidney Disease (CKD) among the age group 40 to 50, the preference based utility dataset between age group 40 to 50 is released for mining. To ensure the individual's privacy, the preference based dataset is micro aggregated and added with Laplace Noise, before releasing it for mining. The dataset is clustered with Preference based Variable (PBV). In our work, partitions are done using n=3. And the clusters are named as c_1 , c_2 , c_3 . The cluster c_1 has values between 1 to 39, c_2 has values between 40 to 50 and c_3 has values between 51 to 90. Table I shows Sample patient dataset.

TABLE I. SAMPLE PATIENT DATASET

Age	**BP**	**al**	**Rbcc**	**Alb**	**Class**
39	100	3	2.8	1	ckd
68	80	0	4.5	2	notckd
41	100	3	2.8	0	ckd
20	90	0	4.0	2	notckd
33	100	3	2.0	2	ckd
80	100	3	2.5	2	ckd
75	100	3	2.5	2	ckd
44	80	0	4.5	2	notckd
49	100	3	2.8	1	ckd

The proposed algorithm is applied to the sample dataset and the intermediate results of the clusters are shown in Table 2. Original dataset is partitioned into 3 groups. Each group cluster values are replaced with mean of that group and Laplace noise is added to the mean value and this perturbed cluster is released for mining. In the final phase, preference based clusters are released for mining. Here our preference is between 40 to 50. So the second partition alone can be released for mining.

TABLE II. CLUSTERS C_1 , C_2 AND C_3

Age	BP	al	Rbcc	Alb	Class
20	90	0	4.0	2	notckd
33	100	3	2.0	2	ckd
39	100	3	2.8	1	ckd

Age	BP	al	Rbcc	Alb	Class
41	100	3	2.8	0	ckd
44	80	0	4.5	2	notckd
49	100	3	2.8	1	ckd

Age	BP	al	Rbcc	Alb	Class
68	80	0	4.5	2	notckd
75	100	3	2.5	2	ckd
80	100	3	2.5	2	ckd

Table 3 shows the privacy preserved sample data without any sensitive attributes.

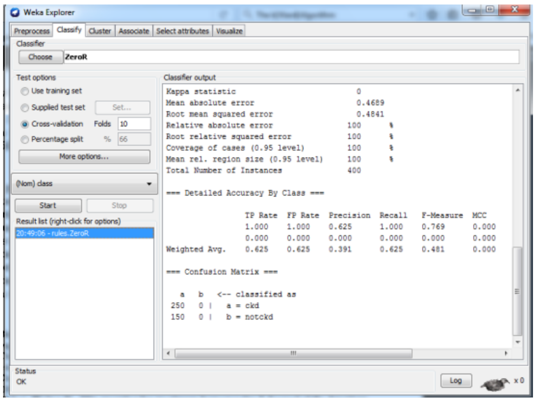
TABLE III. PRIVACY PRESERVED PATIENT DATASET

Age	**BP**	**al**	**Rbcc**	**Alb**	**Class**
46	100	3	2.8	0	ckd
46	80	0	4.5	2	notckd
46	100	3	2.8	1	ckd

V. EXPERIMENTAL EVALUATION

CKD dataset obtained from Bethel hospital, Madurai, Tamilnadu, India is utilized. CKD dataset consists of 400 records with 20 attributes. Our experiments reveal that our framework is effective, meets privacy requirements, and guarantees valid data mining results while protecting sensitive information. We used ZeroR classifier in WEKA tool to classify the CKD dataset. Taking the ZeroR Classifier, we show that our algorithm can be effectively tailored for preserving information in data mining tasks.

We compared the mining result of the original dataset with the privacy preserved dataset using WEKA tool. Our proposed method performed well and produced valid data mining results. Figure4 shows the mining result of the original dataset. Figure 5, shows the mining result of the privacy preserved dataset. Table 4 describes the classification results.



Classifier: ZeroR

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) class

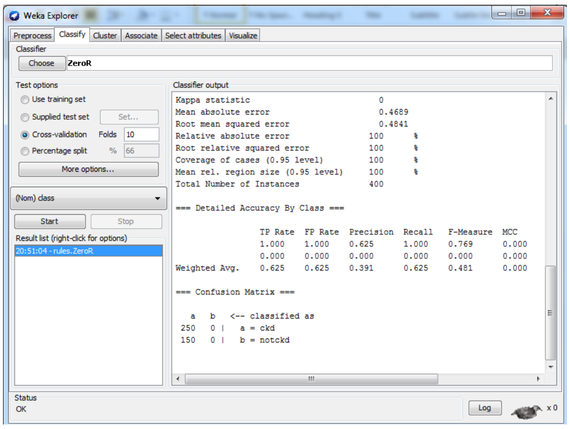
Result list (right-click for options): 20:49:06 - rules.ZeroR

Classifier output

```
Kappa statistic                          0
Mean absolute error                      0.4689
Root mean squared error                  0.4841
Relative absolute error                100      %
Root relative squared error            100      %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 1.000    1.000    0.625      1.000    0.769      0.000
                 0.000    0.000    0.000      0.000    0.000      0.000
Weighted Avg.    0.625    0.625    0.391      0.625    0.481      0.000

=== Confusion Matrix ===

   a   b   <-- classified as
 250   0 |   a = ckd
 150   0 |   b = notckd
```

Status: OK

Fig. 5. Mining Result of the Original Dataset



Classifier: ZeroR

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) class

Result list (right-click for options): 20:51:04 - rules.ZeroR

Classifier output

```
Kappa statistic                          0
Mean absolute error                      0.4689
Root mean squared error                  0.4841
Relative absolute error                100      %
Root relative squared error            100      %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 1.000    1.000    0.625      1.000    0.769      0.000
                 0.000    0.000    0.000      0.000    0.000      0.000
Weighted Avg.    0.625    0.625    0.391      0.625    0.481      0.000

=== Confusion Matrix ===

   a   b   <-- classified as
 250   0 |   a = ckd
 150   0 |   b = notckd
```

Status: OK

Fig. 6. Mining Result of the Perturbed Dataset

TABLE IV. COMPARISON OF CLASSIFICATION RESULTS

Original dataset	Perturbed Dataset
Scheme: Scheme: weka.classifiers.rules.ZeroR Relation: fullset Instances: 400 Attributes: 20 Time taken to build model: 0.01 seconds === Confusion Matrix === a b <-- classified as 250 0 | a = ckd 150 0 | b = notckd	Scheme: Scheme: weka.classifiers.rules.ZeroR Relation: Perturbedset Instances: 400 Attributes: 20 Time taken to build model: 0.01 seconds === Confusion Matrix === a b <-- classified as 250 0 | a = ckd 150 0 | b = notckd

Information loss is the major research issue in privacy preservation approaches. Generally, the information loss should be lesser to attain higher data utility. On the other hand, higher the information loss, lesser would be the data utility. We ran our approach on various k values such as 40, 80, 120, 160, 200, 240, 280, 320, 360and 400. The total information loss was calculated during each run of the experiment. In Figure 6, we show the information loss of original dataset, additive perturbation imposed dataset and Perturbed Microaggregation imposed dataset. We observe that proposed method outperforms the other two existing methods.

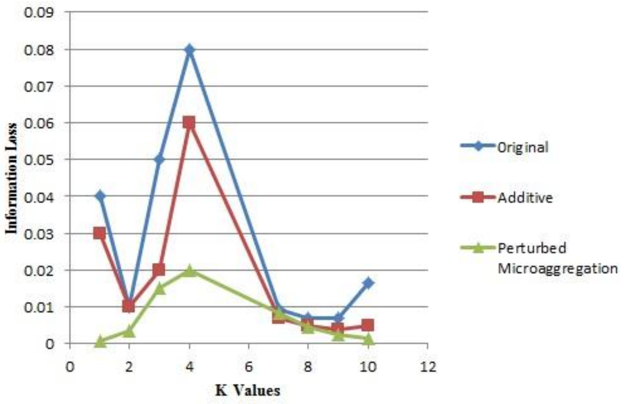


Fig. 7. Information Loss

VI. CONCLUSION AND FUTURE WORK

Data mining is an evolving technology that can be useful in sales forecast, customer behavior prediction and future trends which support administrations to make useful and knowledge driven decisions. Privacy has become a crucial issue in data mining. Numerous privacy preservation techniques are available. In this paper, we have proposed PMA based privacy Preservation in Data Mining which satisfies data utility and minimum information loss. Experiments show that the proposed method reduces information loss and maintain data utility. There won't be any single techniques, which satisfy performance, utility, cost, complexity and tolerance. One technique may perform better than another on one particular criterion. PPDM techniques applied may consider the factors such as Privacy loss, Information loss, Data mining task, Data dimension and Volume, Data Type, Resistant to various data mining algorithms, Complexity and cost, etc.,

Many challenges still remain in PPDM. These challenges will be an active and significant research area. We conclude with some fascinating directions for future research. Privacy in mobile data mining, Privacy in data stream mining, Efficiency and minimum computation cost in distributed PPDM, Privacy and accuracy with minimal loss, Anonymization without attacks and loss, Differential privacy with real and large dataset and Developing graph mining algorithms for complex, dynamic networks with multiple node and edge types are some of the areas where future research can be undertaken.

REFERENCES

[1] Bertino, E., Lin, D. and Jiang, W.(2008) "A Survey of Quantification of Privacy Preserving Data Mining Algorithms", *in Privacy Preserving Data Mining, Springer*, US.

[2] Lindell, Y., & Pinkas, B. (2000). Privacy preserving data mining. In *Annual International Cryptology Conference* (pp. 36-54). Springer Berlin Heidelberg.

[3] Goldreich, O. (2002). Secure multi-party computation. *Final (incomplete) draft, version 1.4*.

[4] Liew, C. K., Choi, U. J., & Liew, C. J. (1985). A data distortion by probability distribution. ACM Transactions on Database Systems (TODS), 10(3), 395-411.

[5] Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. In ACM Sigmod Record (Vol. 29, No. 2, pp. 439-450). ACM.

[6] Evfimievski, A., Gehrke, J., & Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 211-222). ACM.

[7] Kargupta, H., Datta, S., Wang, Q., & Sivakumar, K. (2003). On the privacy preserving properties of random data perturbation techniques. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 99-106). IEEE.

[8] Chen, K., & Liu, L. (2005). Privacy preserving data classification with rotation perturbation. In Fifth IEEE International Conference on Data Mining (ICDM'05) (pp. 4-pp). IEEE.

[9] Chen, K., & Liu, L. (2005). Privacy preserving data classification with rotation perturbation. In Fifth IEEE International Conference on Data Mining (ICDM'05) (pp. 4-pp). IEEE.

[10] Chen, K., Sun, G., & Liu, L. (2007). Towards Attack-Resilient Geometric Data Perturbation. In SDM (pp. 78-89).

[11] Skinner, C. (2009). Statistical disclosure control for survey data. *Handbook of statistics*, *29*, 381-396.

[12] Defays, D. and Anwar, N. (1995) "Microaggregation: A generic method", *in 2nd International Symposium on Statistical Confidentiality*, Eurostat, Luxembourg, 69–78.

[13] Samarati, P., & Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information. In *PODS* (Vol. 98, p. 188).

[14] Domingo Ferrer, J. and Mateo Sanz, J.M. (2002) "Practical data oriented microaggregation for statistical disclosure control", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 14, No. 1, 189–201.

[15] Mateo Sanz, J. M., & Domingo Ferrer, J. (1998). A comparative study of microaggregation methods.

[16] Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J. M., & Sebé, F. (2006). Efficient multivariate data-oriented microaggregation. *The VLDB Journal*, *15*(4), 355-369.

[17] Oganian, A. and Domingo-Ferrer, J. (2001) "On the complexity of optimal microaggregation for statistical disclosure control", *Statistical Journal of the United Nations Economic Comission for Europe*, Vol. 18, No. 4, 345–354.

[18] Hansen, S.L. and Mukherjee, S. (2003) "A polynomial algorithm for optimal univariate microaggregation", *IEEE Transactions on Knowledge and Data Engineering*, Vol.15, No. 4, pp. 1043–1044.

[19] Laszlo, M. and Mukherjee, S.(2005) "Minimum spanning tree partitioning algorithm for microaggregation". *IEEE Transactions on Knowledge and Data Engineering*,17(7):902–911.

[20] Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P., Domingo-Ferrer, J., Torra, V., Brand, R., Giessing, S. (2005): *μ-ARGUS version 4.0 Software and User's Manual*, Statistics Netherlands, Voorburg NL, http://neon.vb.cbs.nl/casc.

[21] Solanas, A. and Martinez-Balleste, A. (2006) "V-MDAV: A multivariate microaggregation with variable group size", *in Computational Statistics COMPSTAT 2006*, Springer's Physica Verlag, pp. 917–925.

[22] Solanas, A., Sebe, F. and Domingo-Ferrer, J. (2008) "Micro-aggregation-based heuristics for p-sensitive k-anonymity: one step beyond". *In Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society*, pp. 61–69, New York, NY, USA, ACM.

[23] Chang, C. C., Li, Y. C. and Huang, W. H. (2007) "TFRP : An efficient microaggregation algorithm for statistical disclosure control", *Journal of Systems and Software*,Vol. 80, No. 11, pp. 1866–1878.

[24] Domingo-Ferrer, J. and Ursula Gonzalez-Nicolas (2010) "Hybrid microdata using microaggregation", *Information Sciences*, Vol 180, No. 15, pp. 2834–2844.

[25] Lin, J. L., Wen, T. H., Hsieh, J. C. and Chang, P. C. (2010) "Density-based microaggregation for statistical disclosure control", *Expert Systems with Applications*, Vol. 37, No. 4, pp. 3256–3263.

[26] Kabir, Md Enamul, and Hua Wang (2011) "Microdata protection method through microaggregation: A median-based approach." *Information Security Journal: A Global Perspective* 20.1: 1-8.

[27] Soria-Comas, J., Domingo-Ferrer, J., Sanchez, D., & Martinez, S. (2015). t-closeness through microaggregation: Strict privacy with enhanced utility preservation. *IEEE Transactions on Knowledge and Data Engineering*, *27*(11), 3098-3110.

[28] Han, J., and Kamber, M. (2000) : Data Mining: Concepts and Techniques. *Morgan Kaufmann Publishers*.

[29] Dwork, C. (2008) Differential privacy: a survey of results. *In Theory and Applications of Models of Computation*—TAMC, pp. 1–19.

[30] WISHART, D. (1969), "An Algorithm for Hierachical Classifications",*Biometrics* 25,165–170.

[31] Arumugam, G., & Sulekha, V. (2016). IMR based Anonymization for Privacy Preservation in Data Mining. In *Proceedings of the The 11th International Knowledge Management in Organizations Conference on The changing face of Knowledge Management Impacting Society* (p. 18). ACM.