ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Optimal Univariate Microaggregation for Privacy Preservation in Data Mining

V. Jane Varamani Sulekha
Assistant Professor
Fatima College
Madurai, Tamilnadu, India

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

Abstract: In recent years, with the massive development in Internet, data collection and data warehousing technologies, privacy preservation has become one of the greater concerns in data mining. For this reason, several data mining algorithms integrating privacy preserving techniques have been developed in order to prevent the disclosure of sensitive information during the knowledge discovery. A number of effective methods for Privacy Preserving Data Mining (PPDM) have been proposed in the literature. In this paper, we present a brief introduction of different kinds of Microaggregation techniques with their merits and demerits and propose Optimal noise addition based Univariate Microaggregation for anonymizing the individual records. Through the experimental results, our 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; Optimal Noise Addition; Data Mining;

I. INTRODUCTION

In recent years, data mining has been viewed as a risk to privacy because of the extensive growth of electronic data maintained by corporations. This has directed to increase the concerns about the privacy of the individual's data. PPDM technique gives unique way to solve this problem. In recent years, a number of techniques have been proposed for transforming or modifying the original data in such a way so as to preserve privacy. PPDM has become an important issue in recent years, because of the large amount of consumer data traced by automated systems on the internet. The increase of Social Networks and E-Commerce on the World Wide Web has resulted in the storage of huge amounts of personal and transactional information about users. In addition, advances in hardware machinery have also made it possible to find information about individuals from transactions in everyday life. For example, a simple purchasing such as buying products online results in automated storage of data about user buying behavior. In many cases, users are not ready to give such personal information unless its privacy is guaranteed. Therefore, in order to ensure effective data collection, it is important to develop techniques which can mine the data with an assurance of privacy. This has resulted to a significant amount of focus on PPDM methods in recent years. PPDM may also express as "obtaining valid data mining results without learning the basic data values" [1]. PPDM contains the dual goal of meeting privacy requirements and ensuring valid data mining results [2].

PPDM consists of two parts. First, sensitive raw data (sensitive attributes, identifiers, quasi identifiers,) such as age, phone number, name, income, disease, address, SIN (Social Insurance Number), SSN (Social Security Number) should be removed or anonymized from the original database, so that the data miner or third party do not interfere into another person's privacy. Next, sensitive information mined from a dataset by using conventional data mining algorithms should also be preserved because that too may compromise data privacy.

This paper reviews Microaggregation, the challenges in privacy pre-serving data mining and proposes a novel Optimal Noise addition based Univariate Microaggregation for privacy preservation. The remainder of this paper is organized as follows. Section 2 analyses microaggregation based PPDM methods. Section 3 introduces our proposed Optimal noise addition based Univariate Microaggregation. Section 4 presents experimental results and Section 5 describes considerations about future extensions and promising directions in the perspective of privacy preserving data mining.

II. MICROAGGREGATION

Microaggregation is a perturbative data preserving method. 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 and aggregation. In partitioning, the original micro data set 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 and in this process cohesion and coupling is introduced among the data. Additionally, each group is forced to contain at least k records. Aggregation, computes aggregated value for each cluster/ group, original values in the micro data set are replaced by the computed aggregated value. This phase usually depends on the type of the variable concerned. Microaggregation methods were originally used for numerical data types. The larger the k, the larger the information loss and the lesser the disclosure risk. Different methods exist in microaggregation. In Univariate microaggregation, microaggregation is applied to every individual variable. In contrast, multivariate microaggregation applied to all variables (or subset) in the cluster. Microaggregation methods can be classified into two types, namely fixed size and data oriented microaggregation. For fixed size microaggregation, the partition is done by dividing a dataset into clusters that have fixed size k, except one cluster which has a size be-tween k and 2k−1, it depends on the total number of records n and the anonymity parameter k. For the data oriented

microaggregation, the partition is based on the data with cluster sizes between k and 2k- 1. Fixed size methods reduce the search space, and thus are more computationally efficient than data oriented methods. Data oriented methods can adapt to dissimilar values of k and to various data dispersals and therefore may attain lower information loss than fixed-size methods.

A. Fixed Size Microaggregationing

For Fixed Size microaggregation [3], 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. Data Oriented univariate microaggregation based on Wards hierarchical algorithm, Genetic algorithm and Fixed size Multivariate microaggregation based on Wards algorithm are discussed. They also derived, if the data set is very large or if microaggregation is to be done on-line then genetic method is good at speed and minimum information loss. While data disclosure is considered, then k -ward based fixed size microaggregation is safe.

Merits/Demerits: Fixed Size methods reduce space complexity, and thus are more efficient than Data Oriented methods.

B. Data Oriented Microaggregation

Data oriented methods [4] may achieve lower information loss than Fixed Size methods. The basic idea is to use fixed size heuristics or other algorithms such as nearest point next (NPN) to construct a path traversing all points in a multivariate dataset. Then the multivariate adaptation of Hansen–Mukherjee's algorithm (MHM) is used on that path. The result is a data oriented k-partition. The NPN selects the first record by computing the record utmost away from the centroid of the entire dataset. The record closest to the first record is selected as the second record. The third record is closest to the second record. This process continues until all of the records have been added to the tour. MHM constructs a graph based on an ordered list of records, and finding the shortest path in the graph. The arcs in the shortest path correspond to a partition of the records that is guaranteed to be the lowest cost partition consistent with the specified ordering.

Merits/Demerits: Data oriented microaggregation with fixed size k are more efficient than the data oriented variable group size. But variable size microaggregation minimizes the information loss.

C. Optimal Microaggregation

Computational complexity of optimal microaggregation [5] with minimal information loss for a fixed security level is proposed. They have shown that the problem of optimal microaggregation is NP-hard.

D. Maximum Distance based Microaggregation (MD)

The Maximum Distance (MD) Method [12] is proposed with univariate and multivariate microaggregation method. The advantage of this method is its simplicity and performance. The MD algorithm builds a k-partition as follows. Using the Euclidean distance two distant records r, s are identified. After that two groups are formed with the first group with r and the k-1 records closest to r and second group with s and the k-1 records closest to s. If there are at least 2k records which do not belong to any of the groups, adopt the same strategy to form new groups. Repeat the step iteratively. Finally, we will get a k-partition of the data set. After the partition, micro aggregated data are computed by replacing each record by the centroid of the group to which it belongs.

Merits/Demerits: Effective portioning is possible by using MD based microaggregation but its computational complexity is higher.

E. Maximum Distance to Average Vector based Microaggregation (MDAV)

Maximum Distance to Average Vector Method (MDAV) [5], is a Multivariate Fixed size microaggregation method employed in the μ-Argus package for statistical disclosure control. It is based on forming groups based on the distance between centroid and distinct data. In MDAV, a square matrix of distances between all records is calculated. Two main approaches can be implemented to perform these distance calculations. The first approach calculates and stores the distances at the beginning of the microaggregation process. This approach is computationally cheaper, but it requires too much memory when the number of records in the data set is large. The second approach calculates the distances dynamically when they are needed. After calculating the matrix of distances, MDAV iterates and builds two groups, at each iteration. In order to build these groups, the centroid C, that is the average vector of the remaining records those are not assigned to any group, is calculated at the beginning of each iteration. Then the most distant record R from C is taken and a group of k records is built around R. The group of k records around R is formed by R and the k- 1 closest records to R. Next, the most distant record S from R is taken and a group of k records is built around S. The generation of groups continues until the number of remaining records (NRR) is less than 2k. When this condition is met, two cases are possible, namely NRR < k or NRR $\geqslant$ k. In the first case, the remaining records are assigned to their closest group. In the second case, a new group is built with all the remaining records.

Merits/Demerits: MDAV is better than MD in terms of computational complexity while maintaining the performance in terms of resulting SSE. The disadvantage of MDAV is it's not flexible. Performance degradation will occur if the data points are scattered in the clusters.

F. Variable - MDAV

Variable Size MDAV or V-MDAV [6] in contrast with fixed size MDAV, produces k partitions with group sizes varying between k and 2k-1. It produces variable size partition. This flexibility can be used to achieve similarity within the group and optimal partition of data. Compute the distances between the records and store them in a distance matrix. Compute the centroid C of the data set. Select the most distant record R from the centroid C. Build group gi with (k-1) closest records to R. Extend the group gi. Repeat the above steps till there are (k-1) records left to be assigned to any group. Assign the remaining unassigned records to its closest group. Build a micro aggregated data set D. Extending the group is determined by the following formula, Unassigned record < γ (Shortest distance from unassigned record to another unassigned record). γ is a gain factor that has to be tuned according to the data set. For γ= 0, V-MDAV is equivalent to MDAV. On the contrary, when the data set is clustered the best values for γ are usually

close to one. The authors selected γ= 0.2 for scattered data sets and γ= 1.1 for clustered data sets.

Merits/Demerits: MDAV generates groups with fixed size. It lacks flexibility for adapting the group size to the distribution of the records in the data set, which may result in poor within-group homogeneity. V-MDAV overcomes the limitation of MDAV with the same computational cost. Determining the optimal value of γ, selecting different values for clustered and scattered data sets are to be researched further.

G. Shortest path algorithm based Microaggregation

Microaggregation problem is formulated as a shortest path problem on a graph. First graph is constructed, and then each arc of the graph corresponds to a possible group that 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 [7].

Merits/Demerits: Minimizes information loss. It can be used on large data sets. Shortest path algorithm based on multivariate data should be researched further.

H. Minimum Spanning Tree Partitioning based Microaggregation

Minimum Spanning Tree Partitioning (MSTP) for microaggregation [8] is pro-posed 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 consider the group size, so that it cannot solve the microaggregation problem. To address this problem, a small modification is made in the MSTP algorithm. That is oversized clusters are further divided into small clusters.

Merits/Demerits: MSTP is efficient. But when data points are distributed in a scattered way, MSTP performance will decrease.

I. Microaggregation based heuristics for p-sensitive k-anonymity

Micro aggregation based p-sensitive k-anonymity [9] is proposed. Its idea is that there are at least p different values for each sensitive attribute within the records sharing a combination of key attributes. This method, initially builds p-sensitive k-anonymity clusters. Then the original data are replaced with its centroid. The authors explained two methods, one is p -sensitive k-anonymity with MDAV and another one is p-sensitive k-anonymity with random seeds.

Merits/Demerits: k-anonymity property is mainly based on suppression and generalization. Here the shortcomings related to generalization and suppressions are eliminated. It minimizes information loss also.

J. Two Fixed Reference Points based Microaggregation

Two Fixed Reference Points (TFRP) based microaggregation [10] is proposed. TFRP has two stages and its two stages are denoted as TFRP-1 and TFRP-2. In the first phase, TFRP uses a fixed size algorithm to partition the data set. In the second phase, TFRP reduces the number of partitions produced by the first phase to improve the data quality.

Merits/Demerits: For sparse data sets and with large k value TFRP produces a very low information loss.

K. Microaggregation based Hybrid data

A new method called microaggregation based hybrid data [11] 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 pre-serve privacy of data. It can be applied to any data type and can yield groups of variable size.

Merits/Demerits: The means and covariance of the sensitive attributes in original data set and synthetic data set are exactly the same. Thus utility is preserved.

L. Density based Microaggregation

A Density Based Microaggregation Algorithm (DBA) [12] is proposed. The DBA has two phases. First Phase (DBA-1), partitions the data set into groups in which each group contains at least k records. To partition the data set, it uses K nearest neighbor-hood 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 unas-signed. 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.

Merits/Demerits: Minimizes information loss. This method works well with univariate numerical value. Multivariate Categorical and mixed data values should be researched further.

M. Median based Microaggregation

Microdata Protection Method through Microaggregation based on Median [13] is proposed. It divides the whole microdata set 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.

N. T-Closeness through Microaggregation

T-Closeness through Microaggregation [14] 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.

O. Individual Ranking based Microaggregation

In order to reduce the amount of noise needed to satisfy differential privacy, Utility Preserving Differentially Private

Data Releases via Individual Ranking Microaggregation [15] is proposed. By using this method, we can improve the utility of differentially private data releases. This can be possible by Individual Ranking. In individual ranking, each variable is treated independently. Data vectors are sorted by the first variable, then groups of k successive values of the first variable are formed and, inside each group, values are replaced by the group average. A similar procedure is repeated for the rest of variables. Microaggregation is done for each variable in turn so that a different partition is obtained for each variable in the microdata set.

Merits/Demerits: Individual ranking owes its popularity to its simplicity and to the fact that it usually preserves more information than one-dimensional projection.

P. Data Recipient centered Microaggregation

A data recipient centered de-identification method to retain statistical attributes [16] is proposed. Based on the input from the recipient (the researcher) de-identification can be done because the researchers have a plan of how to use the data. Using Microaggregation synthetic data are generated.

In our work, we are combining perturbation and microaggregation technique. All the existing PPDM techniques including Microaggregation are applied to the whole data set. In our work, the proposed Optimal Noise addition based Univariate Microaggregation technique is applied to the data set with some utility based preferences imposed on certain parameters in the data set. Preference may be of any kind and different attributes may have different utility. The following are some example where utility based preference can be applied.

- Disease between age group 30 to 50.
- Raised cholesterol and obesity level in males over 40.
- Buying pattern, of the metropolitan population.
- Buying pattern, of a particular age group.
- Climatic disease, in a particular area.
- Depression, Transportation accidents, Respiratory conditions and Drug use disorder among the young age 10 to 19.
- Stress, depression, metabolism and bone problem in females over 40.
- Mobile phone usage among age groups of 15-19 yrs., 20-24 yrs., and 25-34 yrs.
- Phone credit renewal, among age groups of 15-24 yrs., 25-35 yrs., and 36-59 yrs.

Data mining is the process of evaluating data from different perceptions and summarizing it into useful information. A typical data mining process depends on data owner to define what kind of pattern they are going to mine or interested in. According to the utility based pattern, selection of data can be done. Instead of releasing the whole data set, the utility based on the preferences in the parameters of the data set can be released to improve computing time and storage space. This method also reduces the risk of individual disclosure and data mining algorithm complexity.

III. OPTIMAL NOISE ADDITION BASED UNIVARIATE MICROAGGREGATION

As Han and Kamber [17] 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 identify interesting patterns, they are often insufficient. It should be combined with subjective measures that reflect a particular user's interests and needs. For example, patterns describing the disease among patients of a hospital should be interesting to the hospital administration, but may be of little interest to other analysts studying the same database. It is very necessary for data mining systems to generate only interesting and useful patterns. This would be effective for users and data mining systems because neither would have to examine through the patterns generated to identify the really interesting ones. While considering the Electronic Health Records (EHR), dataset might be useful for one purpose but useless for another. User provided constraints and interestingness measures should be added with data mining process to obtain completeness of mining. Generally, it is not the responsibility for a data owner to build models, but it is the responsibility for a data owner to keep privacy when the data are released. The data owner has to execute a privacy protection technique with different preference based parameters to attain a desired trade-off between privacy and utility.

Considering this in our mind we propose a novel privacy preserving technique Optimal Noise addition based Univariate Microaggregation. Optimal Noise addition based Univariate Microaggregation Combines preference based Microaggregation by Individual Ranking and optimal ε differential privacy based perturbation which ensures low information loss and guarantees privacy and utility. 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. Reconstruction won't be possible in Optimal Noise addition based Univariate Microaggregation. The data owner can also choose a preference based dataset [18] from a set of non-dominated dataset.

Optimal Noise addition based Univariate Microaggregation technique can be divided into two major parts Microaggregation and Optimal Noise Addition. In Microaggregation phase, K ward hierarchical clustering algorithm [19] is used to partition the dataset. Individual ranking is a popular microaggregation method. In individual ranking, each variable is treated independently. In our work we are taking the variable age as preference based variable (PBV) and individual ranking is done using age. By using K-ward algorithm, data set 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. A similar technique is repeated for the rest of the variables if we want to use this method for multivariate microaggregation. Individual sorting usually preserves more personal information. After the microaggregation, optimal noise is added to each micro aggregated value and this perturbed data set 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. X denotes all the attributes of the original data set. X' denotes the perturbed data set. When the original data is replaced with the cluster mean, the sensitivity of the data set 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 data set is n/k*Δx/k. To obtain differential privacy, Laplace noise (n/k *Δx/k)/ ε is added to the numerical data. Laplace noise is not optimal.

Let N1 and N2 be two random noise distributions. If N1 can be constructed from N2 by moving some of the probability mass towards zero, then N1 must always be preferred to N2. The reason is that the probability mass of N1 is more concentrated around zero, and thus the distortion introduced by N1 is smaller. A rational user al-ways prefers less distortion and, therefore, prefers N1 to N2. A random noise [20] distribution N1 is optimal within a class C of random noise distributions if N1 is minimal within C; in other words, there is no other random N2 ∈ C such that N2 < N1.

Pseudocode of our proposed work.

Step1. Form a cluster using individual ranking imposed Preference Based Variable (age) with the first k elements of the original data set and another group with the last k elements of the original data set

Step2. Use Wards method until all elements in the original data set 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 Optimal Noise (ON), (n/k *Δx/k)/ε to each attribute in the clusters.

The first step ensures that in each recursive step the data set 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.

We combine Individual ranking based microaggregation and optimal ε differential privacy. This combination gives better performance, low information loss and ensures privacy. The main difference between our proposed technique with the previous microaggregation algorithm is that, the given method can get *privacy preserved multi partitioned univariate (singe attribute based) numerical dataset*. In each partition, the perturbation method applied is different (different noise addition for each partition), so it may restrict the reconstruction problem. The perturbed data set obtained from original data set will give the same mining result while applying classification or clustering algorithm. This method reduces the risk of individual disclosure.

Chronic kidney disease (CKD) is now-a-days common among the middle age. For example, if a hospital wishes to know the CKD among the age group 40 to 50, here the preferred utility based pattern is CKD among the age group 40 to 50. In this work, age is individual ranking imposed PBV. Partition is done on age and the preference based perturbed dataset between age group 40 to 50 is released for mining. To ensure the individual's privacy, the preference based data set is micro aggregated and added with optimal Laplace Noise, before releasing it for mining. Considering the partition as n=3and the clusters are named as c1 , c2 , c3 . The cluster c1 has values between 1 to 39, c2 has values between 40 to 50 and c3 has values between 51 to 90. Table I shows Sample data set.

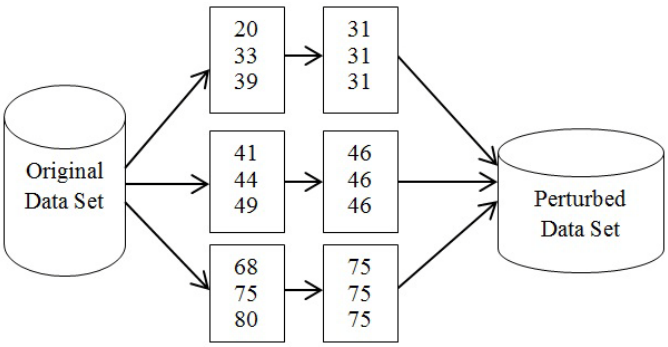

Figure 1. Multiple Protected Dataset.

Table I. Sample Patient Data

Sr. No.	*Age*	*BP*	*al*	*Rbcc*	*Alb*	*Class*
1.	39	100	3	2.8	1	ckd
2.	68	80	0	4.5	2	notckd
3.	41	100	3	2.8	0	ckd
4.	20	90	0	4.0	2	notckd
5.	33	100	3	2.0	2	ckd
6.	80	100	3	2.5	2	ckd
7.	75	100	3	2.5	2	ckd
8.	44	80	0	4.5	2	notckd
9.	49	100	3	2.8	1	ckd

The proposed algorithm is applied to the sample patient dataset and the intermediate results of the clusters are shown in Table 2. Original data set is partitioned into 3 groups. Each group cluster values are replaced with mean of that group and Optimal Noise is added to the mean value. In the final phase, preference based clusters are released for mining.

Table II. Clusters C1, C2 and C3

Sr. No.	*Age*	*BP*	*al*	*Rbcc*	*Alb*	*Class*
1.	20	90	0	4.0	2	notckd
2.	33	100	3	2.0	2	ckd
3.	39	100	3	2.8	1	ckd

Sr. No.	*Age*	*BP*	*al*	*Rbcc*	*Alb*	*Class*
1.	41	100	3	2.8	0	ckd
2.	44	80	0	4.5	2	notckd
3.	49	100	3	2.8	1	ckd

Sr. No.	*Age*	*BP*	*al*	*Rbcc*	*Alb*	*Class*
1.	68	80	0	4.5	2	notckd
2.	75	100	3	2.5	2	ckd
3.	80	100	3	2.5	2	ckd

Table 3 shows the privacy preserved patient data. For the first partition the ON=1.05, the second partition ON=1.4 and for the third partition ON=1. Here we are having 3 Partitons, our preference is age group between 40 to 50. So the second partion alone can be released to dataminers for analysis.

Table III. Privacy Preserved Patient Dataset

Sr. No.	*Age*	*BP*	*al*	*Rbcc*	*Alb*	*Class*
1.	31	90	0	4.0	2	notckd
2.	31	100	3	2.0	2	ckd
3.	31	100	3	2.8	1	ckd

Sr. No.	*Age*	*BP*	*al*	*Rbcc*	*Alb*	*Class*
1.	46	100	3	2.8	0	ckd
2.	46	80	0	4.5	2	notckd
3.	46	100	3	2.8	1	ckd

Sr. No.	*Age*	*BP*	*al*	*Rbcc*	*Alb*	*Class*
1.	75	80	0	4.5	2	notckd
2.	75	100	3	2.5	2	ckd
3.	75	100	3	2.5	2	ckd

Indeed, with individual ranking any intruder knows that the real value of an element in the i^{th} group is between the average of the $i\text{-}1^{th}$ group and the average of the $i+1^{th}$ group. If these two averages are very close to each other, then a very narrow interval for the real value being searched has been determined. Individual ranking is less vulnerable to inference attack.

IV. EXPERIMENTAL RESULTS

CKD data set obtained from Bethel hospital, Madurai, Tamilnadu, India is utilized. Original CKD dataset consists of 1200 records with 20 attributes. For building the Preference based privacy preserved dataset the computing time is lesser. Then after applying Optimal Noise addition based Univariate Microaggregation technique, the mining process also takes less time while using the preference based dataset. First we compared the time taken to mine the original data set with the preference based dataset (2nd partition alone) using WEKA tool. We used ZeroR classifier in WEKA tool to classify the CKD data set. The synthetic data set is generated from the original CKD data set. The synthetic data set consists of 11, 40, 243 records and storage space is 147 MB. After applying the Optimal Noise addition based Univariate Microaggregation technique taking PBV as age, the preference based dataset consists of 5, 36, 346 records and storage space is 54 MB. Table 4, shows the time taken to mine the original dataset and preference based dataset.

Table IV. Time Taken for the Original and Preference based dataset

Sr. No.	*Original Dataset*	*Preference based Dataset*
1.	Scheme: weka.classifiers.rules.ZeroR Relation: full set Instances: 1140243 Attributes: 20 Time taken to build model: 0.42 seconds === Confusion Matrix === a b <-- classified as 716231 0 | a = ckd 424012 0 | b = notckd	Scheme: weka.classifiers.rules.ZeroR Relation: preference based Instances: 536346 Attributes: 20 Time taken to build model: 0.145 seconds == Confusion Matrix === a b <-- classified as 369296 0 | a = ckd 167050 0 | b = notckd

Next we compared the mining result of the original data set with the privacy preserved full data set using WEKA tool. Table 5, shows the classification results of the original and Optimal Noise addition based Univariate Microaggregation technique imposed dataset. Our experiments reveal that our framework is effective, meets privacy requirements, and guarantees valid data mining results while protecting sensitive information. Our proposed method performed well and produced valid data mining results.

Table V. Mining Results of Original and Optimal Microaggregation applied dataset

Sr. No.	*Original Dataset*	*Preference based Dataset*
1.	Scheme: weka.classifiers.rules.ZeroR Relation: full set Instances: 1140243 Attributes: 20 Time taken to build model: 0.42 seconds === Confusion Matrix === a b <-- classified as 716231 0 | a = ckd 424012 0 | b = notckd	Scheme: weka.classifiers.rules.ZeroR Relation: Optimal Microaggregation Instances: 1140243 Attributes: 20 Time taken to build model: 0.42 seconds === Confusion Matrix === a b <-- classified as 716231 0 | a = ckd 424012 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. The sum of squares criterion is used to measure the similarity in clusters. Within the group sum of squares SSE is stated as

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

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

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

The total sum of squares SST is stated as

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

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

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

We used k=120 and ran the algorithm. The total information loss was calculated during each run of the experiment. In Figure 4, we show the information loss of original data set, additive perturbation based dataset, Perturbation based Microaggregation Technique (PMAT) imposed dataset and Optimal Noise addition based Univariate Microaggregation imposed dataset. We observe that proposed method outperforms the other existing methods.

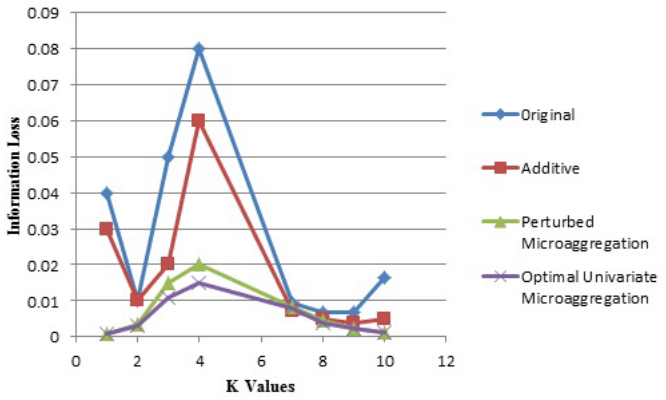

Figure 2. Information Loss.

V. 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 Optimal Noise addition based Univariate Microaggregation 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.

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. Multivariate Individual ranking on numeric data, Univariate and Multivariate Contiguous data based Microaggregation can be researched further.

VI. References

[1] Dehkordi, Mohammad Naderi, Kambiz Badie, and Ahmad Khadem Zadeh. "A Novel Method for Privacy Preserving in Association Rule Mining Based on Genetic Algorithms." *JSW* 4.6 (2009): 555-562.

[2] Lindell, Yehuda, and Benny Pinkas. "Privacy preserving data mining." Annual International Cryptology Conference. Springer Berlin Heidelberg, 2000.

[3] Domingo-Ferrer, Josep, and Josep Maria Mateo-Sanz. "Practical data-oriented microaggregation for statistical disclosure control." *IEEE Transactions on Knowledge and data Engineering* 14.1 (2002): 189-201.

[4] Domingo-Ferrer, Josep, et al. "Efficient multivariate data-oriented microaggregation." *The VLDB Journal—The International Journal on Very Large Data Bases* 15.4 (2006): 355-369.

[5] Hundepool, A., et al. "μ-ARGUS version 4.0 Software and User's Manual." *Statistics Netherlands, Voorburg NL* (2005).

[6] Solanas, Agusti, Antoni Martinez-Balleste, and J. Domingo-Ferrer. "V-MDAV: a multivariate microaggregation with variable group size." *17th COMPSTAT Symposium of the IASC, Rome*. 2006.

[7] Hansen, Stephen Lee, and Sumitra Mukherjee. "A polynomial algorithm for optimal univariate microaggregation." *IEEE Transactions on Knowledge and Data Engineering* 15.4 (2003): 1043-1044.

[8] Laszlo, Michael, and Sumitra Mukherjee. "Minimum spanning tree partitioning algorithm for microaggregation." *IEEE Transactions on Knowledge and Data Engineering* 17.7 (2005): 902-911.

[9] Solanas, Agusti, Francesc Sebé, and Josep Domingo-Ferrer. "Micro-aggregation-based heuristics for p-sensitive k-anonymity: one step beyond." *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*. ACM, 2008.

[10] Chang, Chin-Chen, Yu-Chiang Li, and Wen-Hung Huang. "TFRP: An efficient microaggregation algorithm for statistical disclosure control." *Journal of Systems and Software* 80.11 (2007): 1866-1878.

[11] Domingo-Ferrer, Josep, and Úrsula González-Nicolás. "Hybrid microdata using microaggregation." *Information Sciences* 180.15 (2010): 2834-2844.

[12] Lin, Jun-Lin, et al. "Density-based microaggregation for statistical disclosure control." *Expert Systems with Applications* 37.4 (2010): 3256-3263.

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

[14] Soria-Comas, Jordi, et al. "t-closeness through microaggregation: Strict privacy with enhanced utility preservation." *IEEE Transactions on Knowledge and Data Engineering* 27.11 (2015): 3098-3110.

[15] Sánchez, David, et al. "Utility-preserving differentially private data releases via individual ranking microaggregation." *Information Fusion* 30 (2016): 1-14.

[16] Gal, Tamas S., et al. "A data recipient centered de-identification method to retain statistical attributes." *Journal of biomedical informatics* 50 (2014): 32-45.

[17] Traub, Joseph F., Yechiam Yemini, and H. Woźniakowski. "The statistical security of a statistical database." *ACM Transactions on Database Systems (TODS)* 9.4 (1984): 672-679.

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

[19] Wishart, David. "256. Note: An algorithm for hierarchical classifications." *Biometrics* (1969): 165-170.

[20] Soria-Comas, Jordi, and Josep Domingo-Ferrer. "Optimal data-independent noise for differential privacy." *Information Sciences* 250 (2013): 200-214.