A COMPARATIVE STUDY ON MICROAGGREGATION TECHNIQUES FOR MICRODATA PROTECTION

Sarat Kumar Chettri[1], Bonani Paul[2] and Ajoy Krishna Dutta[2]

[1]Department of Computer Science, Saint Mary's College, Shillong, Meghalaya, India
s.chettri@smcs.ac.in
[2]Department of Computer Science, Saint Mary's College, Shillong, Meghalaya, India
b.paul@smcs.ac.in
[3]Department of Computer Science, Saint Mary's College, Shillong, Meghalaya, India
a.dutta@smcs.ac.in

ABSTRACT

Microaggregation is an efficient Statistical Disclosure Control (SDC) perturbative technique for microdata protection. It is a unified approach and naturally satisfies k-Anonymity without generalization or suppression of data. Various microaggregation techniques: fixed-size and data-oriented for univariate and multivariate data exists in the literature. These methods have been evaluated using the standard measures: Disclosure Risk (DR) and Information Loss (IL). Every time a new microaggregation technique was proposed, a better trade-off between risk of disclosing data and data utility was achieved. Though there exists an optimal univariate microaggregation method but unfortunately an optimal multivariate microaggregation method is an NP hard problem. Consequently, several heuristics have been proposed but no such method outperforms the other in all the possible criteria. In this paper we have performed a study of the various microaggregation techniques so that we get a detailed insight on how to design an efficient microaggregation method which satisfies all the criteria.

KEYWORDS

Statiscal Disclosure Control, Information Loss, Disclosure Risk, microdata, anonymity, microaggregation

1. INTRODUCTION

With the advent of various data mining techniques, the process of knowledge discovery from large databases or data sets has improved considerably. The discovered knowledge which was previously unknown facilitates the decision making processes in different areas of marketing and supply-chain management, medical and health care, deciding policies and planning strategies etc. Let us see some of the scenario under where data mining techniques plays an important role for data analyses and knowledge discovery. If a government of a country decides for implementing various social welfare schemes for its people, then detailed study is needed to be done on the demography of the region, population etc. For a company to launch any new product in a market, it first needs to study the market such as consumption trend, buying habits of people etc. For such research analysis and planning, large amount of databases are being shared and published, which in turn increases the risk of breaching the privacy of individuals associated with the database. An efficient technique was needed to effectively analyze data without hampering the sentiments of individuals. It was observed that simple de-identification realized by removing the identifying attributes from the database was not effective to protect individuals' privacy. One better way to protect the individuals' privacy is to modify data prior to its release. But the challenge here is how to modify the data in such a way that data utility lies above certain threshold limit even after its modification. This challenge is the raison d̀être of Statistical Disclosure Control as stated in

DOI : 10.5121/ijdkp.2012.2603

[1]. The Statistical Disclosure Control (SDC) 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. The definition of privacy has been formally stated in [2] as *"The right of an entity to be secure from unauthorized disclosure of sensible information that are contained in an electronic repository or that can be derived as aggregate and complex information from data stored in an electronic repository"*.

Traditionally, SDC methods have been devised to protect respondent privacy by entailing some degrees of data modification. Microaggregation is an efficient Statistical Disclosure Control perturbative technique for microdata protection i.e. protection of individual data. Unlike *k*-Anonymity, microaggregation method modifies data without suppressing or generalizing it. It was first proposed in the year 1995 by Defays and Anwar as a special clustering problem where a data set is partitioned into small homogenous groups. Each group contains at least *k* records and instead of releasing the raw microdata values, the mean of the group they belong to is reported in their place prior to their publication or release. Thus, we can say that microaggregation naturally satisfies *k*-Anonymity. But microaggregation is not about simple clustering or partitioning a data set into homogenous groups where each group consists of at least *k* records. It is very crucial to group records in such a way that the data disclosure risk is kept at the minimal level while keeping the data utility high. In other words we can say that a better trade-off is required between the risks of disclosing the sensitive data and the loss of information occurred due to data modification. The microaggregation method was originally defined for continuous data by Defays and Nanopoulos [3] and also in other works as can be seen in [4, 5]. It was then extended for categorical data [7] and later for heterogeneous data [6]. The optimal microaggregation method partitions a data set into groups of size lying between *k* and *2k*-1. The user defined parameter *k* decides the degree of perturbation, large value of *k* may ensure higher data privacy but the data may not be useful for statistical analyses as information loss may be higher. Normally, the *k* value is taken as 3, 4, 5 or 10 in any microaggregation method.

The rest of the paper is organized as follows. Section 2 gives the concepts of microaggregation method and microdata protection. Section 3 reviews the various approaches and to microaggregation method and the works related to it. Section 4 lists the comparison of various fixed-sizes and data-oriented microaggregation methods based on their complexities and comparison is also done on real dataset with respect to information loss (IL) measure. Finally, in Section 5 conclusion is drawn with possible future research directions.

2. MICROAGGREGATION AND MICRODATA PROTECTION

2.1. *Microdata Concepts*

Microdata are information about respondent individual for e.g. company data, data related to a person etc. It can be also viewed as a file which consists of *n* individual records with *m* attributes. The microdata attributes can be classified into following categories –

- Identifiers – These attributes can be used to identify individual records uniquely, for e.g. Employee ID, patient code etc.

- Quasi-identifiers – These attributes can be used to identify individual records, but not uniquely, as the records which are identified may be ambiguous. For e.g. person's age, name etc.

- Confidential attributes – These attributes contains some individual respondent information which is sensitive in nature to some extent. For e.g. patient's diagnosis report, person's community etc.

- Non-confidential attributes – The attributes which do not fall in any of the categories as mentioned above belong to this category. For e.g. person's hobbies, language skills etc. These kinds of attributes cannot be neglected as they can be a part of quasi-identifier.

The microdata file is shared among users/analysts for various research analyses which increase the risk of disclosing some sensitive information about the individuals concerning the data. There are various techniques available for protecting microdata from individual identification. It can be performed either by data modification/data masking or by generating synthetic data [12]. In both the techniques, the main aim is to get new microdata set V' from its original counterpart V. Irrespective of the techniques applied to obtain V', it should serve the primary goal of low risk of disclosing data keeping its statistical information content high. The data masking technique can be broadly classified into two categories as shown in [9, 10] –

- Perturbative method – Here the data is distorted prior to its release. Rank swapping [14], resampling [16], microaggregation [4], additive noise [15] etc are some of the techniques under this category.

- Non Perturbative method – Here the data is not distorted as in the case with perturbative method, but it would be generalized or suppressed prior to its release. Some of the techniques which falls under this category are global recoding or generalization [10], local suppression [11], top and bottom coding [13] etc.

Table 1. Perturbative methods and data types

Method	**Categorical Data**	**Continuous Data**
Rank Swapping	✓	✓
Resampling		✓
Microaggregation	✓	✓
Additive Noise		✓

The data masking technique have been found to be more effective than synthesizing data in terms of data utility and data disclosure risk. As seen in [8] due to over fitting with the original data, the synthetic data has the tendency of being re-identified.

Table 2. Non- Perturbative methods and data types

Method	**Categorical Data**	**Continuous Data**
Generalization	✓	✓
Local Suppression	✓	
Top and bottom coding	✓	✓

Tables 1 and 2 shows the various perturbative and non-perturbative methods and the data types where the respective methods can be applied.

Different methods may be applied for preserving the privacy of data. In a survey paper [2] Elisa Bertino et.al have stated that any privacy preserving data mining algorithm can be evaluated based on following set of criteria –

- Privacy level – It determines how closely the sensitive data in the data set can be identified after any privacy preserving technique has been applied to it.
- Data quality – It indicates whether the analyses on the data set obtained by applying perturbative or non-perturbative method in it are similar to the original data.
- Hiding failure – It indicates the failure of the privacy preserving technique to hide the portion of the sensitive data of the data set.
- Complexity – This criterion indicates about the performance of the privacy preserving method in terms of computational time and resources consumed.

2.2. *Microaggregation method*

Microaggregation is a Statistical Disclosure Control (SDC) method which is perturbative in nature. It is an efficient method for microdata protection and was first proposed by Defays and Anwar [4] in the year 1995. It was originally defined for continuous data by Defays and Nanopoulos [3] and also in other works as can be seen in [4, 5] and was then extended for categorical data [7] and later for heterogeneous data [6]. The microaggregation method follows mainly two steps; first it partitions the dataset into homogenous groups where each group consists of at least *k* records (where *k* is a user defined parameter) and then every record of a group is substituted with the corresponding group's mean value. There is no constraint in the number of groups that can be formed but group size should lie between *k* and 2*k*-1. Microaggregation automatically satisfies *k*-Anonymity [17] without generalizing or suppressing data. In *k*-Anonymity, every record is indistinguishable from at least (*k*-1) other records. Usually the distance measure used to determine the similarity of records in microaggregation method is Euclidean distance. To be more specific let us consider a microdata set *R* with *d*-dimensional variables on *n* individuals. Now, when microaggregation method is applied on the microdata set then *m* groups are formed with at least *k* records in each group. The optimal partition of the microdata set is measured in terms of within group sum of squares (SSW) (1) or alternatively by the between-group sum of squares (SSB) (2). The SSW value should be least, while SSB value should be as high as possible.

$$SSW = \sum_{i=1}^{m} \sum_{j=1}^{k_i} ||x_{ij} - \bar{x}_i||^2 \qquad (1)$$

$$SSB = \sum_{i=1}^{m} k_i \, ||\bar{x}_i - \bar{x}||^2 \qquad (2)$$

Where,

$\bar{x}_i$ Average data vector over the *i*-th group.

x_{ij} *j*-th record in the *i*-th group.

$\bar{x}$ Total mean of the whole dataset.

k_i k_i ƒrecords in the *i*-th group.

The total sum of squares is computed as

$$SST = SSW + SSB = \sum_{l=1}^{n} ||\bar{x}_l - \bar{x}||^2 \qquad (3)$$

The information loss (IL) caused due to microaggregation is measured as –

$$IL = \frac{SSW}{SST}.100 \qquad (4)$$

According to the dimensionality of data in the microdata set, microaggregation method can be divided into two categories –

- Univariate micoaggregation – It is applied to each variable of a microdata set in an independent manner. The problem becomes easier, as only single variable is involved, where the idea of individual ranking can be applied as can be seen in [5]. Furthermore, in [18] we can see that there exists a polynomial-time optimal algorithm for univariate microaggregation method.

- Multivariate microaggregation – Here, the grouping process is applied to sets of variables of the microdata set. In this case, when all the variables are microaggregated together, *k*-Anonymity is automatically satisfied thereby reducing the risk of data disclosure. Thus, one can concentrate in maximising data utility. A polynomial time optimal multivariate microaggregation method is an NP hard problem as stated in [19]. Consequently, several heuristics have been proposed under this category.

Irrespective of the data dimensionality of the microdata set, the microaggregation method applied can be of fixed-size or data-oriented (variable size). The fixed-size method partitions a microdata set into groups of size *k* where each group contains *k* records except one which may contain more than *k* records when the number of records in the microdata set is not a multiple of *k,* whereas the data-oriented microaggregation method produces groups of variable sizes. The group size lies between *k* and 2*k*-1. Though fixed-size microaggregation method takes less computation time in partitioning the dataset by reducing the search space but variable size method tends be more flexible in grouping records as it can adapt to various data distribution, thus increasing within group homogeneity and incurring lesser information loss.

3. VARIOUS APPROACHES TO MICROAGGREGATION METHOD

Various univariate microaggregation methods have been proposed and there even exist an optimal univariate microaggregation method [18] in the literature, but an optimal multivariate microaggregation method is an NP hard problem as stated in [19].

Consequently several microaggregation heuristics have been proposed in the literature. In this section we give a brief overview of the various approaches to microaggregation method which exists in the literature.

3.1. Genetic-Algorithm-Based Microaggregation

The Genetic-Algorithm (*G-A*) based microaggregation algorithm was proposed in [27]. Genetic Algorithm is a method for moving from one population of chromosomes to a new population by using a kind of natural selection together with the genetics-inspired operators of crossover, mutation, and inversion.

In [27] a Genetic Algorithm has been modified to address the issues of microaggregation where N-array coding is used. Here, each chromosome has been considered to be having a fixed number of genes equalling the number of records in the data set. Thus the value of the i-th gene in a chromosome defines the group of the k-partition which the i-th record in the data set belongs to. Though the sum squared errors (SSE) (5) of *G-A* which gives the within group homogeneity was found to be better than *MDAV* but there its efficiency decreases in case of large data set. So, a hybrid method was also proposed in [27] which take the advantage of both *MDAV* and classic Genetic Algorithm and produces better result in terms of SSE while dealing with large multivariate data set. The SSE is computed as follows –

$$SSE = \sum_{i=1}^{c} \sum_{x_{ij} \in C_i} (x_{ij} - \bar{x_\iota})'(x_{ij} - \bar{x}_i) \qquad (5)$$

Where,

c is the total number of clusters or groups.

C_i is the i-th cluster and $\bar{x}_i$ is the centroid of C_i.

The hybrid method can be summarised as follows –

1. A small value of k is taken (e.g. $k = 3$).
2. Let K be larger than k and divisible by k, small enough to be suitable for the modified Genetic Algorithm (e.g. $K = 21$).
3. Use any fixed-size heuristic (e.g. *MDAV*) to build a k-partition of the data set.
4. Taking as input records, the average vectors obtained in the previous step, the fixed-size heuristic to build *macro-groups* (i.e. sets of average vectors) of size K/k is applied.
5. For each given *macro-group*, the average vectors by the k original records are replaced to obtain a K-partition.

Finally, the *G-A* is applied to each macro-group in the K- partition in order to generate an optimal or near optimal k-partition of the macro-group. The composition of the k-partitions of all macro-groups yields a k-partition for the entire data set.

3.2. Hybrid Microdata Using Microaggregation

A new method called *mycrohybrid* has been proposed in [28]. It has been shown that hybrid data can be obtained using microaggregation with any synthetic data generator. Here, a method has been devised that produces hybrid data without following the conventional method of combining the original and the synthetic data to produce hybrid data. Let V be the original data set consisting of n records. Now, using the technique [28] a hybrid data set V' is produced with $k \in [1, n]$. The hybrid data V' so produced preserves the means and co- variances of the original data set V. The procedure calls two algorithms –

- A synthetic data generator which generates a synthetic data set.
- A microaggregation heuristic, which partitions a data set into groups of sizes between k and $2k$-1.

The *microhybrid* method can be summarized as follows –

1. Partition the dataset V into clusters containing k and ($2k$-1) records.
2. Apply a synthetic data generator algorithm to obtain a synthetic version of each cluster.
3. Replace the original records in each cluster by the records in the corresponding synthetic cluster.

In step 3, the conventional way of microaggregation method is not followed where the records are replaced with the mean value of the cluster to which they belong. The *microhybrid* procedure is a simple approach to preserve privacy of data. It can be applied to any data type and can produce groups of variable size.

3.3. Density-Based Algorithm (DBA) for Microaggregation

A *Density-Based Algorithm* (*DBA*) for microaggregation has been proposed in [29]. The *DBA* follows two phase method; firstly *DBA*-1 partitions a data set into groups where each group contains at least k records. To partition the data set, *DBA*-1 uses k-neighborhood of the record with the highest k-density among all the records that are not assigned to any group. The grouping process continues till k records remain unassigned. These remaining k records are then assigned to its nearest groups.

The second phase of *DBA* known as *DBA*-2 is then applied to further tune the partition in order to achieve low information loss and high data utility. *DBA*-2 may decompose the formed groups or may merge its records to other groups. The criterion for decomposing is the information loss. Let IL_{bmerge} be the information loss of a group G before any group G_i is merged into it and IL_{amerge} be the information loss incurred after merging the group G_i into G. If $IL_{bmerge} > IL_{\mathrm{amerge}}$ then split the merged group and merge each record of G_i to its nearest group. If at the end of *DBA*-2 method, few group ends up having more than ($2k$-1) records then in that case *MDAV*-1 algorithm [29] is applied to each group whose size is above ($2k$-1). In this way the information loss is minimized. This state is finally called *MDAV*-2. The *DBA*-2 method is similar to *TFRP*-2 but *TFRP*-2 does not allow a record to merge into a group of size ($4k$-1).

3.4. Maximum Distance to Average Vector (MDAV)

MDAV is one of the best heuristic methods for multivariate microaggregation. It is a fixed-size method and was first proposed in [13] as a part of a multivariate microaggregation method implemented in the μ-Argus package for statistical disclosure control. Later several variant of this method were proposed in the literature with minor modifications made to it. The algorithm is as follows-

Algorithm: *MDAV*

```
1. Compute centroid C of dataset D.
2. Find the most distant record $x$ from the centroid C.
3. Build group $g_i$ with ($k$-1) closest records to $x$.
4. Find the most distant record $x_s$ from $x$.
5. Build group $g_{i+1}$ with ($k$-1) closest records to $x_s$.
```

6. Repeat the steps 1 to 5 till there are more than (2*k*-1) records left to be assigned to any group.
7. If there remains more than (*k*-1) records to be assigned then form a new group with the remaining records.
8. Assign the remaining records to the closest group.
9. Build a microaggregated data set D′ by replacing the records with its mean value of the group to which it belongs.

In step 8, if less than *k* records remain then all the records of this subgroup are assigned to its closest group determined by computing distance between centroids of the groups. *MDAV* finally ends up forming groups of the same size *k* except only one.

3.5. Two Fixed Reference Points (TFRP)

Two Fixed Refernce Points (*TFRP*) as proposed by Chang et al. in [22] is a two stage method for microaggregation and its two stages are denoted as *TFRP*-1 and *TFRP*-2 respectively. *TFRP* has a computation time of $O(n^2/k)$ and low information loss particularly in sparse data sets with large value of *k*. The *TFRP* algorithm is as follows-

Algorithm: *TFRP*-1 (First Phase)

1. Compute the two reference points *R1* and *R2*. All vectors are assigned to a set (SET).
2. Select a reference point.
3. Select an initial point x^i from the reference point.
4. Calculate the distance of each vector to x^i.
5. Select (*k*-1) closest vectors together with x^i to form a group, and remove the *k* vectors from SET.
6. Select another reference point, then go to Step 2 until |SET| < *k*.
7. Assign each remaining vector of SET to its closest group.

In step 1 the two reference points *R1* and *R2* are two extreme points in the microdata set. In step 7, each remaining vector is assigned to its closest group. It has been found that the within group sum of squares (SSE) (5) of the formed group is high, thus to reduce the information loss the *TFRP* algorithm goes through second phase (*TFRP*-2).

Algorithm: *TFRP*-2 (Second Phase)

1. Compute SSE of each group, and sort them in decreasing order.
2. Select a group G_i in order and compute the current total sum of the within-group squared errors (SSE_1).
3. Calculate the distance of each vector of G_i to any other group.
4. Assign each vector of G_i to its closest group provisionally, and compute the current total sum of the within-group squared errors (SSE_2).
5. If $SSE_1 > SSE_2$, then assign each vector of G_i to its closest group; otherwise, regain G_i.
6. Return to Step 2 and repeat until each group is checked.

After applying the *TFRP*-2, if several groups contains greater than or equal to 2*k* records then the groups are broken down using any fixed-size microaggregation method. And in case the size of the closest group to which a vector x^i has to be assigned is (4*k*-1) records then the vector x^i is assigned to its second closest group.

3.6. Variable-size Maximum Distance to Average Vector (V-MDAV)

Variable-size MDAV or *V-MDAV* [23] in contrast with fixed-size *MDAV*, yields k-partitions with group sizes varying between *k* and 2*k*-1. Such flexibility can be exploited to achieve higher within-group homogeneity and optimal partition of data. *V-MDAV* method was proposed by Agusti Solanas and Antoni Martínez-Ballesté. It is a heuristic approach to multivariate microaggregation, which provides variable size groups and thus higher in within-group homogeneity measured by SSE, with a computational cost similar to the one of fixed-size microaggregation heuristics. Moreover, the way in which *V-MDAV* expands the groups can be tuned by using the gain factor γ. The value of γ was set as 0.2 for scattered datasets and γ =1.1 for clustered data set. The procedure for *V-MDAV* method is as follows-

Algorithm: *V-MDAV*

1. Compute distance matrix of the dataset D.
2. Compute centroid C of dataset D.
3. Select the most distant record *x* from the centroid C.
4. Build group g_i with (*k*-1) closest records to *x*.
5. Extend the group $g_{i.}$
6. Repeat the steps 3 to 5 till there are (*k*-1) records left to be assigned to any group.
7. Assign the remaining unassigned records to its closest group.
8. Build a microaggregated data set D′ by replacing the records with its mean value of the group to which it belongs.

In step 5, group g_i extension is done by checking if the distance d_{in} between the nearest unassigned record *x* to a group g_i is less than γ multiplied by the minimum distance d_{out} from the selected record *x* to any of the remaining unassigned record i.e. if $d_{in} < \gamma\ d_{out}$ then add *x* to g_i else do not add. At the end of the algorithm if there still exist few unassigned records then they are added to its nearest groups. Though the gain factor γ can be tuned for efficiently partitioning the data set depending on its data distribution but determining the optimal value for γ is not a straight forward task.

4. COMPARISON OF VARIOUS MICROAGGREGATION METHODS

Microaggregation being a perturbative statistical disclosure control method modifies the data to some extent to preserve its confidentiality. As a result there exists some loss of information due to data modification. To prove the efficiency of such method it has to be evaluated based on certain standard measures. Also to show the efficiency of one microaggregation method over another, various microaggregation methods; fixed-size or data-oriented methods, univariate or multivariate methods are compared based on its complexities, SSE, information loss (IL) or finding trade-off between data disclosure risk and information loss by using score method as can be seen in [30].

Table 3 lists some of the fixed-size microaggregation methods and their respective complexities. *TFRP*-1 is the first phase of the *TFRP* method [22].The complexities of methods *MD, MDAV* and *CBFS* has been quoted from [21], *TFRP*-1 from [22], *M-d* from [20] and *IP* from [24]. Table 4 lists some of the data-oriented microaggregation methods and their respective complexities. The complexities of methods *MD-MHM, MDAV-MHM* and *CBFS-MHM* has been quoted from [21], *V-MDAV* from [23].

Table 3. Fixed-Size Multivariate Microaggregation Methods

Fixed-size Microaggregation Method	**Complexity (with n records)**
Maximum Distance (MD)	$O\left(\frac{n^3}{12k}\right)$
Maximum Distance to Average Vector (MDAV)	$O\left(\frac{n^2}{2k}\right)$
Centroid-Based Fixed-Size microaggregation (CBFS)	$O\left(\frac{n^2}{2k}\right)$
*Two Fixed Reference Points Phase I (TFRP-*1*)*	$O\left(\frac{n^2}{k}\right)$
Minimum Spanning Tree based method (M-d)	$O\left(\frac{n^2}{2k}\right)$
Importance Partitioning (IP)	$O(n^2)$

Table 4. Data-Oriented Multivariate Microaggregation Methods

Data-oriented Microaggregation Method	**Complexity (with n records)**
Maximum Distance – Multivariate Hansen-Mukherjee (MD–MHM)	$O\left(\frac{n^3}{12k}\right)$
MDAV-MHM	$O\left(\frac{n^2}{2k}\right)$
CBFS-MHM	$O\left(\frac{n^2}{2k}\right)$
Variable-size Maximum Distance to Average Vector (V-MDAV)	$O(n^2)$

Here, in this paper for comparison of different microaggregation methods we have considered the three referenced data sets [25]; “Tarragona” data set that contains 834 records and 13 numerical attributes, the “Census” data set which contains 1080 records and 13 numerical attributes and the “EIA” data set which contains 4092 records and 11 numerical attributes. These three referenced data sets are the benchmarks used to evaluate various microaggregation methods. We have used the information loss (IL) (4) as a measure of comparison with different values of k = 3, 4, 5, and 10. Table 5, 6 and 7 give the resulting information losses in each case. The information losses of methods *MDAV-MHM, MD-MHM, CBFS-MHM, NPN-MHM* and *M-d* (for k = 3*,* 5*,* 10) are quoted from [21]; the information losses of methods *μ-Approx* and *M-d* (for k = 4) are quoted from [26], the information loss (IL) of IP (for k = 3,5) are taken from [24], *TFRP*-1 (for k = 3,4,5,10) are taken from [22], and *MDAV*, *V-MDAV* (for k = 3,4,5,10) are quoted from [23], *DBA*-1 and *DBA*-2 (for k = 3,4,5,10) are quoted from [29]. Comparable results of *M-d* and *CBFS-MHM* are not available for “EIA” data set. The “EIA” data set is a non-homogenous data set with clustered records. In such cases, heuristics which partitions the data set into variable size groups is more appropriate.

Table5. Information loss (IL) comparison using Tarragona data set

Method	***k* = 3**	***k* = 4**	***k* = 5**	***k* = 10**
M-d	16.630	19.66	24.5000	38.5800
TFRP-1	17.228	19.396	22.110	33.186
MDAV	16.96	19.70	22.88	33.26
MDAV-MHM	16.932	-	22.462	33.192
MD-MHM	16.983	-	22.527	33.183
CBFS-MHM	16.971	-	22.823	33.219
IP	15.61	-	22.45	-
V-MDAV	16.96	19.70	22.88	33.26
μ-Approx	17.10	20.51	26.04	38.80
NPN-MHM	17.395	-	27.0213	40.183
DBA-1	20.699	23.828	26.001	35.393
DBA-2	16.153	22.671	25.450	34.807

Table 6. Information loss (IL) comparison using Census data set

Method	***k* = 3**	***k* = 4**	***k* = 5**	***k* = 10**
M-d	6.110	8.24	10.300	17.170
TFRP-1	5.931	7.880	9.357	14.442
MDAV	5.66	7.51	9.01	14.07
MDAV-MHM	5.652	-	9.087	14.224
MD-MHM	5.697	-	8.986	14.397
CBFS-MHM	5.673	-	8.894	13.893
IP	5.34	-	8.68	-
V-MDAV	5.69	7.52	8.98	14.07
μ-Approx	6.25	8.47	10.78	17.01
NPN-MHM	6.349	-	11.344	18.734
DBA-1	6.145	9.128	10.842	15.785
DBA-2	5.582	7.591	9.046	13.521

Table 7. Information loss (IL) comparison using EIA data set

Method	***k* = 3**	***k* = 4**	***k* = 5**	***k* = 10**
M-d	-	-	-	-
TFRP-1	0.530	0.661	1.651	3.242
MDAV	0.49	0.67	1.78	3.54
MDAV-MHM	0.408	-	1.256	3.773
MD-MHM	0.442	-	1.263	3.637
CBFS-MHM	-	-	-	-
IP	0.47		1.53	-
V-MDAV	0.53	0.75	1.30	2.82
μ-Approx	0.43	0.59	0.83	2.26
NPN-MHM	0.553	-	0.960	2.319
DBA-1	1.09	0.843	1.896	4.266
DBA-2	0.421	0.559	0.818	2.081

5. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we have seen the different approaches to microaggregation methods for microdata protection and also the various criteria on which any privacy preserving data mining algorithm can be evaluated. Tables 5 through 7 give comparative results of various fixed-size and data–oriented microaggregation methods, where the methods are compared based on information loss (IL) measure. The benchmark data sets used for comparing the various microaggregation methods are "Tarragona", "Census" and "EIA" data sets. Comparison tables shows that the information loss (IL) measure of most of the microaggregation methods differ in their fractional parts only. But comparatively the information loss (IL) of *IP* method (for k = 3) is found to be slightly lower than its counterparts in case of "Census" and "Tarragona" data sets. While the information loss (IL) of methods *MDAV-MHM* (for k = 3) and *DBA*-2 (for k = 5) is found to be little lower than other methods in case of "EIA" data set.

For future research it would be interesting to repeat the study performed in this paper by comparing the various microaggregation methods for a very large data set. Also the comparison can be based on the trade-off between data disclosure risk and information loss. Other interesting line for future research includes development of efficient multivariate microaggregation heuristics which can deal with time series and mixed type of data.

ACKNOWLEDGEMENTS

This work is partly supported by the UGC, India under Minor Research Project grant (Vide No. F.5-313/2011-12/MRP/NERO/10898 Dated 5th Dec 2011)

REFERENCES

[1] L. Willenborg and T. DeWaal, "Elements of Statistical Disclosure Control", Lecture Notes in Statistics, Springer-Verlag, New York, 2001.

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

[3] D. Defays and P. Nanopoulos, "Panels of enterprises and confidentiality: The small aggregates method", in 92 Symposium on Design and Analysis of Longitudinal Surveys, Canada, Ottawa, 1993, pp. 195–204.

[4] D. Defays and N. Anwar, " Micro-aggregation: A generic method, in 2nd International Symposium on Statistical Confidentiality", Eurostat, Luxembourg,1995, pp. 69–78.

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

[6] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogenerous k-anonymity through microaggregation", Data Mining and Knowledge Discovery, Vol. 11, No.2, 2005, pp. 195–212.

[7] V. Torra, " Microaggregation for categorical variables: A median based approach", in J. Domingo-Ferrer and V. Torra Eds. Lecture Notes in Computer Science, Vol. 3050, Springer, Berlin, Heidelberg, pp. 162–174.

[8] W. E. Winkler, "Re-identification methods for masked microdata", in J. Domingo-Ferrer and V. Torra Eds. Lecture Notes in Computer Science, Privacy in Statistical Databases, Vol. 3050, Springer, Berlin, Heidelberg, 2004, pp. 216–230.

[9] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogenous k-anonymity through microaggregation", Data Mining and Knowledge Discovery, Vol. 11, No. 2, 2005, pp. 195–212.

[10] P. Samarati, "Protecting respondents' identities in microdata release", IEEE Trans. Knowledge and Data Engineering, Vol. 13, No. 6, 2001, pp. 1010–1027.

[11] A.G. DeWaal and L. Willenborg, "Global recodings and local suppressions in microdata sets", in Statistics Canada Symposium'95, Statistics Canada, Ottawa, 1995, pp. 121–132.

[12] T. J. Raghunathan, J. P. Reiter, and D. Rubin, "Multiple imputation for statistical disclosure limitation", Journal of Official Statistics,Vol. 19, No. 1, 2003, pp. 1–16.

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

[14] B. Greenberg, "Rank swapping for ordinal data", Washington, DC, U.S. Bureau of the Census 1987, unpublished.

[15] R. Brand, "Microdata protection through noise addition", in J. Domingo-Ferrer, Eds. Lecture Notes in Computer Science, Inference Control in Statistical Databases, Vol. 2316, Springer, Berlin, Heidelberg, 2002, pp. 97–116.

[16] J. Domingo-Ferrer and J. M. Mateo-Sanz., "On resampling for statistical confidentiality in contingency tables", Computers & Mathematics with Applications, Vol. 38, 1999, pp.13–32.

[17] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression", International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, Vol. 10, No. 5, 2002, pp. 571–588.

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

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

[20] M. Laszlo and S. Mukherjee, "Minimum spanning tree partitioning algorithm for microaggregation", IEEE Trans. Knowledge and Data Engineering, Vol. 17, No. 7, 2005, pp. 902–911.

[21] J. Domingo-Ferrer, A. Martnez-Ballest, J.M. Mateo-Sanz, and F. Sebe´, "Efficient multivariate data-oriented microaggregation", The VLDB Journal, Vol. 15, No. 4, 2006, pp. 355-369.

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

[23] A. Solanas and A. Marti´nez-Balleste´, "V-MDAV: A multivariate microaggregation with variable group size", in Computational Statistics COMPSTAT 2006, Springer's Physica Verlag, 2006, pp. 917–925.

[24] D. Fouskakis and G. Kokolakis, “Importance patitioning in microaggregation”, Computational Statistics and Data Analysis 53, Elsevier, 2009, pp. 2439-2445.

[25] R. Brand, J. Domingo-Ferrer and J.M. Mateo-Sanz, Reference data sets to test and compare SDC methods for protection of numerical microdata, European Project IST-2000-25069 CASC, 2002, http://neon.vb.cbs.nl/casc.

[26] J. Domingo-Ferrer, F. Sebe´ and A. Solanas, “A polynomial-time approximation to optimal multivariate microaggregation”, Computer and Mathematics with Applications, Vol. 55, No. 4, 2008, pp. 714–732.

[27] A. Solanas, “Privacy Protection with Genetic Algorithms”, in Success in evolutionary computation, A. Yang, Y.Shan and L.T. Bui, Eds. Springer, Heidelberg, Studies in Computational Intelligence, Vol 92, pp. 215–237.

[28] J.Domingo-Ferrer and U´rsula Gonza´lez-Nicola´s, “Hybrid microdata using microaggregation”, Information Sciences, Vol 180, No. 15, 2010. pp. 2834–2844.

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

[30] J.Domingo-Ferrer and V. Torra, A quantitative comparison of disclosure control methods for microdata”, in P. Doyle, J.I. Lane, J.J.M. Theeuwes, and L. Zayatz, Eds. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland,2001,pp.111–134.

AUTHORS

Sarat Kumar Chettri is an Assistant Professor in the Department of Computer Science, Saint Mary's College, Shillong, Meghalaya, India. He is currently pursuing Ph.D. degree in computer science from the Department of Computer Science and Engineering, Tezpur University, Tezpur, India. His research interests include database technology, data mining and knowledge discovery, machine learning and data privacy preservation.

Bonani Paul is an Assistant Professor in the Department of Computer Science, Saint Mary’s College, Shillong 793003, Meghalaya. Her research interest includes Database Technology, Security, Data Privacy and Preservation.

Ajoy Krishna Dutta is an Assistant Professor in the Department of Computer Science, Saint Mary’s College, Shillong 793003, Meghalaya. He is interested in various computer languages and research and Developments. His research interest is in the field of Network and Security, Data Privacy and Preservation.