$K-$means Clustering Microaggregation for Statistical Disclosure Control

Md Enamul Kabir, Abdun Naser Mahmood and Abdul K Mustafa

Abstract This paper presents a K-means clustering technique that satisfies the bi-objective function to minimize the information loss and maintain k-anonymity. The proposed technique starts with one cluster and subsequently partitions the dataset into two or more clusters such that the total information loss across all clusters is the least, while satisfying the k-anonymity requirement. The structure of $K-$ means clustering problem is defined and investigated and an algorithm of the proposed problem is developed. The performance of the $K-$ means clustering algorithm is compared against the most recent microaggregation methods. Experimental results show that $K-$ means clustering algorithm incurs less information loss than the latest microaggregation methods for all of the test situations.

1 Introduction

Microaggregation is a family of Statistical Disclosure Control (SDC) methods for protecting microdata sets that have been extensively studied recently [1, 2, 7, 9]. The basic idea of microaggregation is to partition a dataset into mutually exclusive groups of at least k records prior to publication, and then publish the centroid over each group instead of individual records. The resulting anonymized dataset satisfies k-anonymity [6], requiring each record in a dataset to be identical to at least $(k-1)$ other records in the same dataset.

Md Enamul Kabir
The University of New South Wales, Canberra, e-mail: m.kabir@adfa.edu.au

Abdun Naser Mahmood
The University of New South Wales, Canberra, e-mail: Abdun.Mahmood@unsw.edu.au

Abdul K Mustafa
University of Canberra e-mail: abdul.mustafa@jcu.edu.au

The effectiveness of a microaggregation method is measured by calculating its information loss. *k*-anonymity [5, 6, 8] provides sufficient protection of personal confidentiality of microdata, while ensuring the quality of the anonymized dataset, an effective microaggregation method should incur as little information loss as possible. To minimize the information loss due to microaggregation, all records are partitioned into several groups such that each group contains at least *k* similar records, and then the records in each group are replaced by their corresponding mean such that the values of each variable are the same. Such similar groups are known as clusters.

The reminder of this paper is organized as follows. We introduce a problem of microaggregation in Section 2. Section 3 introduces the basic concept of microaggregation. We present a brief description of our proposed microaggregation method in Section 4. Section 5 shows experimental results of the proposed method. Finally, concluding remarks are included in Section 6.

2 Problem Statement

The algorithms for microaggregation works by partitioning the microdata into homogeneous groups so that information loss is low. The level of privacy required is controlled by a security parameter *k*, the minimum number of records in a cluster. This work presents a new clustering-based method for microaggregation which finds the minimal information loss clustering for an increasing number of clusters. The method works by calculating the maximum number of clusters by $K = \text{int}\lfloor \frac{n}{k} \rfloor$, where n is the total number of records in the dataset and k is the anonymity parameter for *k*-anonymization. Recall that in a *k*-anonymous clustering each cluster must have *k* or more instances. It is easy to prove that a clustering which satisfies *k*+1 anonymity also satisfies *k* anonymity. Therefore, the premise of this work is to find a *k*+*i* anonymous clustering which has the lowest information loss. The trivial solution is to form a single cluster with all the records in the dataset and calculate the information loss. Clearly, this cluster is *k*-anonymous (assuming $k << n$), however, the information loss may be high. Observe that in the rare case where every instance in the dataset is identical, this method can find the *k*-anonymous clustering in the quickest possible manner. For the general case, total information loss would decrease as the number of cluster increases. Note, in the rare case that all the instances are completely different such that they belong to their own clusters, total information loss would be zero since there is no information loss due to each cluster represented by one instance. However, this would certainly breach *k*-anonymity requirement since *k* must be greater than 1. Consequently, the problem is to design a technique that can take advantage of this *k*-anonymity property by checking fewer clusters first, which is a different approach taken from existing methods. The proposed method is explained in Section 4 and compared against the most recent widely used microaggregation methods in Section 5. The experimental results demonstrate that the proposed microaggregation technique outperforms all of the compared techniques for at least one of the benchmark datasets and has comparable results with these techniques for the other dataset.

3 Background

Consider a microdata set T with p numeric attributes and n records, where each record is represented as a vector in a p-dimensional space. For a given positive integer $k \leq n$, a microaggregation method partitions T into K clusters, where each cluster contains at least k records (to satisfy k-anonymity), and then replaces the records in each cluster with the centroid of the cluster. Let n_i denote the number of records in the ith cluster, and $x_{ij}, 1 \leq j \leq n_i$, denote the jth record in the ith cluster. Then, $n_i \geq k$ for $i = 1$ to K, and $\sum_{i=1}^{K} n_i = n$. The centroid of the ith cluster, denoted by $\bar{x}_i$ is calculated as the average vector of all the records in the ith cluster.

In the same way, the centroid of T, denoted by $\bar{x}$, is the average vector of all the records in T. Information loss is used to quantify the amount of information of a dataset that is lost after applying a microaggregation method. In this paper we use the most common definition of information loss by Domingo-Ferrer and Mateo-Sanz [1] as follows:

$$IL = \frac{SSE}{SST} \tag{1}$$

where SSE is the within-cluster squared error, calculated by summing the Euclidean distance of each record x_{ij} to the average value $\bar{x}_i$ as follows:

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)' (x_{ij} - \bar{x}_i) \tag{2}$$

and SST is the sum of squared error within the entire dataset T, calculated by summing the Euclidean distance of each record x_{ij} to the average value $\bar{x}$ as follows:

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

4 The Proposed Approach

This section presents the proposed K-means based anonymization technique to solve the dual objective of minimum information loss and k-anonymity. The proposed approach builds one cluster at the first instance and subsequently adding more clusters such that k-anonymity requirements and information losses are guaranteed.

4.1 Clustering Technique

One of the most widely used clustering algorithms is Lloyd's K-means algorithm [12]. Given a set of N records $(n_1, n_2, \cdots n_n)$, where each record is a d-dimensional vector, the K-means clustering partitions the N records into K clusters $(K < N)$ $S = (S_1, S_2, \cdots, S_k)$ such that intra cluster distance is minimized and inter cluster distance is maximized. The number of clusters to be fixed in K-means clustering. Let the initial centroids be $(w_1, w_2, \cdots, w_k)$ be initialized to one of the N input patterns. The quality of the clustering is determined by the following error function.

$$E = \sum_{i=1}^{k} \sum_{n_l \varepsilon C_j} \parallel n_l - w_j \parallel^2 \tag{4}$$

where C$_j$ is the jth cluster whose value is a disjoint subset of input patterns.

K means algorithm works iteratively on a given set of K clusters. Each iteration consists of two steps:

- Each data item is compared with the K centroids and associated with the closest centroid creating K clusters.
- The new sets of centroids are determined as the mean of the points in the cluster created in the previous step.

The algorithm repeats until the centroids do not change or when the error reaches a threshold value. The computational complexity of algorithm is $O(NKd)$.

4.2 $K-$means anonymization technique

Based on the clustering technique and the definition of the microaggregation problem, next we discuss the $k-$means clustering microaggregation algorithm.

The algorithm first identifies the maximum number of clusters by, $K = \frac{n}{k}$, where k is the anonymity parameter for k-anonymization and round this as integer. Form a cluster with all the n records in the dataset. It will then form two clusters (see step 3 of Table 1) that causes least information loss and satisfy the k-anonymity requirement. The algorithm compares the information loss with the previous step and selects clusters that satisfy both the requirements of data quality and the anonymity parameter (see step 4 of Table 1. The algorithm then continues to build clusters (see step 5 of Table 1) up to K (maximum number of clusters) and finally selects the optimum number of clusters where both the least information loss and the k-anonymity requirements are satisfied.

Table 1 $K-$means clustering algorithm

```
Input: a dataset $T$ of $n$ records and a positive integer $k$.

Output: a partitioning $G = \{G_1, G_2, ..., G_K\}$ of $T$, where $K = |G|$
    and $G_i \geq k$ for $i = 1$ to $K$.

1. Let $K = \mathrm{int}\lfloor \frac{n}{k} \rfloor$;
2. Form a cluster with all records in $T$ and calculate the information;
    loss. Obviously the information loss would be 1;
3. Form one more cluster that causes least information loss among all;
    possible combination of such clusters. Check each cluster satisfy;
    $k-$anonymity requirement;
4. Choose clusters that cause least information loss and satisfy the;
    $k$-anonymity requirement;
5. Repeat steps 3-4 for up to $K$ clusters and finally select clusters;
    where least information loss and $k-$anonymity are guaranteed;
```

5 Experimental Results

The objective of our experiment is to investigate the performance of our approach in terms of data quality. We demonstrate the effectiveness of the proposed approach by comparing it against a basket of well-known techniques.

5.1 Comparison against existing techniques

This section experimentally evaluates the effectiveness of the K-means clustering mcroaggregation algorithm. The following two datasets [3], which have been used as benchmarks in previous studies to evaluate various microaggregation methods, were adopted in our experiments.

1. The "Tarragona" dataset contains 834 records with 13 numerical attributes.
2. The "Census" dataset contains 1,080 records with 13 numerical attributes.

To accurately evaluate our approach, the performance of the proposed K-means clustering mcroaggregation algorithm is compared in this section with various microaggregation methods.
Tables 2-3 show the information loss for several values of k for the *Census* and for the *Tarragona* datasets respectively. The information loss is compared with the K-means clustering microaggregation algorithm among the latest microaggregation methods listed above. Information loss is measured as $\frac{\mathrm{SSE}}{\mathrm{SST}} \times 100$, where SST is the total sum of the squares of the dataset. Note that the within-groups sum of squares SSE is never greater than SST so that the reported information loss measure takes

Table 2 Information loss comparison using Census dataset

Method	$k=3$	$k=4$	$k=5$	$k=10$
MDAV-MHM	5.6523		9.0870	14.2239
MD-MHM	5.69724		8.98594	14.3965
CBFS-MHM	5.6734		8.8942	13.8925
NPN-MHM	6.3498		11.3443	18.7335
M-d	6.1100	8.24	10.3000	17.1700
μ-Approx	6.25	8.47	10.78	17.01
TFRP-1	5.931	7.880	9.357	14.442
TFRP-2	5.803	7.638	8.980	13.959
MDAV-1	5.692186279	7.494699833	9.088435498	14.15593043
MDAV-2	5.656049371	7.409645342	9.012389597	13.94411775
DBA-1	6.144855154	9.127883805	10.84218735	15.78549732
DBA-2	5.581605762	7.591307664	9.046162117	13.52140518
K-C	**3.575**	**3.9561**	**4.532**	**6.8419**

Table 3 Information loss comparison using Tarragona dataset

Method	$k=3$	$k=4$	$k=5$	$k=10$
MDAV-MHM	16.9326		22.4617	33.1923
MD-MHM	16.9829		22.5269	33.1834
CBFS-MHM	16.9714		22.8227	33.2188
NPN-MHM	17.3949		27.0213	40.1831
M-d	16.6300	19.66	24.5000	38.5800
μ-Approx	17.10	20.51	26.04	38.80
TFRP-1	17.228	19.396	22.110	33.186
TFRP-2	16.881	19.181	21.847	33.088
MDAV-1	16.93258762	19.54578612	22.46128236	33.19235838
MDAV-2	16.38261429	**19.01314997**	22.07965363	33.17932950
DBA-1	20.69948803	23.82761456	26.00129826	35.39295837
DBA-2	**16.15265063**	22.67107728	25.45039236	34.80675148
K-C	20.2425	20.2425	**20.2425**	**23.9761**

values in the range [0,100].

Tables 2-3 show the lowest information losses obtained by applying all the microaggregation methods. The information loss of the proposed algorithm (**K-C**) is at the last row of each table. The lowest information loss for each dataset and each k value is shown in bold face. Note that the proposed algorithm has the best performance among all the techniques for the *Census* dataset. For the *Tarragona* dataset **K-C** has the lowest information for $k=5$ and $k=10$, but DBA-2 and MDAV-2 have the lowest values for $k=3$ and $k=4$, respectively. The information losses of methods DBA-1, DBA-2, MDAV-1 and MDAV-2 are quoted from [11]; the information losses of methods MDAV-MHM, MD-MHM, CBFS-MHM, NPN-MHM and M-d (for $k=3,5,10$) are quoted from [3]; the information losses of methods μ-Approx and M-d (for $k=4$) are quoted from [4], and the information losses of methods TFRP-1 and TFRP-2 are quoted from [10]. TFRP is a two-stage method and its two stages are denoted as TRFP-1 and TRFP-2 respectively. The TFRP-2 is similar to the DBA-2 but disallows merging a record to a group of size over $(4k-1)$. The experimental results illustrate that in all of the test situations, the K- means algorithm incurs significantly less information loss than any of the microaggregation methods listed in the table.

6 Conclusion

Microaggregation is an effective method in SDC to protect privacy in microdata and has been extensively used world-wide. This work has presented a new K-means clustering mcroaggregation method for numerical attributes that works by partitioning the dataset into as few clusters as possible with the lowest information loss. A comparison has been made of the proposed algorithm with the most widely used microaggregation methods using two benchmark datasets (Census and Tarragona). The experimental results show that the proposed algorithm has a significant dominance over the recent microaggregation methods with respect to information loss. is very effective microaggregation method in preserving the privacy of data.

References

1. J. Domingo-Ferrer and J. Mateo-Sanz, “Practical data-oriented microaggregation for statistical disclosure control,” *IEEE Transactions on Knowledge and Data Engineering,* vol. 14, no. 1, pp. 189–201, 2002.
2. J. Domingo-Ferrer and V. Torra, “Ordinal, continuous and heterogeneous kanonymity through microaggregation,” *Data Mining and Knowledge Discovery,* vol. 11, no. 2, pp. 195–212, 2005.
3. J. Domingo-Ferrer, A. Martinez-Balleste, J.M. Mateo-Sanz and F. Sebe, “Efficient multivariate data-oriented microaggregation,” *The VLDB Journal,* vol. 15, no. 4, pp. 355–369, 2006.
4. 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, pp. 714–732, 2008.
5. P. Samarati, “Protecting respondent’s privacy in microdata release,” *IEEE Transactions on Knowledge and Data Engineering,* vol. 13, no. 6, pp. 1010–1027, 2001.
6. L. Sweeney, “k-Anonymity: A model for protecting privacy,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* vol. 10, no. 5, pp. 557–570, 2002.
7. M.E. Kabir and H. Wang, “Systematic Clustering-based Microaggregation for Statistical Disclosure Control,” in *Proc. IEEE International Conference on Network and System Security, Melbourne, Sep. 2010,* pp. 435–441.
8. M.E. Kabir, H. Wang, E. Bertino and Y. Chi, “Systematic Clustering Method for l-diversity Model,” in *Proc. Australasian Database Conference, Brisbane, Jan. 2010,* pp. 93–102.
9. M.E. Kabir and H. Wang, “Microdata Protection Method Through Microaggragation: A Median Based Approach,” *Information Security Journal: A Global Perspective,* (in Press).
10. 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, pp. 1866–1878, 2007.
11. 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, pp. 3256–3263, 2010.
12. S. Lloyd, S., “Least squares quantization in PCM,” *Information Theory, IEEE Transactions on*, vol.28, no.2, pp. 129- 137, Mar 1982