Knowledge-Based Systems 67 (2014) 195–205

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Fast data-oriented microaggregation algorithm for large numerical datasets

Reza Mortazavi, Saeed Jalili *

Computer Engineering Department, Tarbiat Modares University, Tehran, Iran

ARTICLE INFO

Article history:
Received 17 February 2014
Received in revised form 5 May 2014
Accepted 6 May 2014
Available online 21 May 2014

Keywords:
Privacy
Microaggregation
k-Anonymity
Statistical Disclosure Control
Savings heuristic

ABSTRACT

Microaggregation is a successful mechanism to solve the tension between respondent privacy and data quality in the context of Statistical Disclosure Control. Microaggregation, for numerical datasets, is defined as a clustering problem with the constraint of having at least *k* records in each group, such that the sum of the within-group squared error (*SSE*) is minimized. Unfortunately, the data publisher has to execute an algorithm iteratively for different values of *k* to investigate a good trade-off between privacy and utility. Multiple execution of an algorithm on large numerical datasets is resource wasting, since most of the computations are repetitive. In this paper, we propose a Fast Data-oriented Microaggregation algorithm (FDM) that efficiently anonymizes large multivariate numerical datasets for multiple successive values of *k*. Experimental results on real world datasets demonstrate the superiority of the method in terms of both the data quality and time complexity. Moreover, the method usually achieves a better trade-off between disclosure risk and information loss of the protected dataset in comparison with previous techniques.

1. Introduction

Nowadays, there is an increasing attention towards the confidentiality of individuals or enterprises in microdata publishing. The main concern of Statistical Disclosure Control (SDC) [1] is to balance the confidentiality and utility of published data. There are two main approaches in the SDC for microdata protection [2,3]; (1) Generating synthetic data [4], and (2) Masking original data with a modified version of them. Masking methods are also classified into perturbative and non-perturbative techniques based on their effect on the original data [5]. Perturbative methods such as microaggregation [6,7] and noise addition [8] distort the microdata before publishing, while non-perturbative methods utilize generalization and suppression [9]. Microaggregation is known as a mechanism to realize *k*-anonymity [10] in the Privacy Preserving Data Publishing (PPDP) literature [11]. The method is currently being used by many statistical agencies [12], and implemented in a number of SDC software packages such as μ-argus [13] and sdcMicro [14].

Microaggregation is defined as a constraint-clustering problem, where the number of final clusters is not known *a priori*, but the minimum group size constraint is enforced to achieve *k*-anonymity. In microaggregation, each group should contain at least *k* records. After partitioning records into clusters, they are replaced by their associated cluster centers, and these centroids are published. This replacement masks the detailed information of records behind the groups to achieve privacy, so the adversary cannot isolate a record using overlapping queries. Generally, larger values of *k* produce more anonymity, while degrading the quality of the published dataset. The effectiveness of a microaggregation algorithm for a given protection level (in terms of *k*) is measured by incurred Information Loss (*IL*). Lower values of *IL* indicate less information distortion after microaggregation, so the protected dataset becomes more useful. The measure reduces, if similar records are aggregated in a group.

The optimal microaggregation problem can be solved for the univariate case in a polynomial time [15], while it has been proven to be NP-Hard for multivariate datasets [16]. Many heuristic algorithms with different characteristics and complexities have been introduced in the literature,[1] while most of them have a common restriction for use in practice: the data publisher has to execute one (or more) microaggregation algorithm(s) with different privacy parameters *k*, to establish a desired balance between anonymity and utility. This is a time consuming task and impractical for large

* Corresponding author. Tel.: +98 21 82883374.
E-mail addresses: r.mortazavi@modares.ac.ir (R. Mortazavi), sjalili@modares.ac.ir (S. Jalili).

[1] See also Section 3.

http://dx.doi.org/10.1016/j.knosys.2014.05.011
0950-7051/© 2014 Elsevier B.V. All rights reserved.

multivariate numerical datasets. Traditional microaggregation methods resort to pre-partitioning a large dataset to speed up the anonymization process at the expense of increasing information loss. In contrast, the proposed method considers large numerical datasets as a whole, so it preserves more data utility. Moreover, regarding the fact that k is small in practical SDC applications [12], the proposed method can produce multiple protected datasets for $3 \leqslant k \leqslant 10$, in a single run, without reloading the original dataset from disk or cloning it within the main memory. It is designed such that many computations for different consecutive values of k are saved. The amortized cost in terms of both information loss and time complexity is reasonably acceptable in comparison with the most recent techniques in the literature. Multiple experiments on real-life datasets confirm the effectiveness of the proposed method to produce more useful protected data in a reasonable time.

Our main contribution in the present study is to propose a novel microaggregation algorithm to anonymize large multivariate numerical datasets. Additionally, we have extended the optimal univariate microaggregation algorithm [15] to optimally and efficiently partition a sequence of multivariate records in a tour. We have shown experimentally that savings heuristic is useful to produce the sequence which can be passed to the partitioning algorithm to anonymize a dataset for multiple successive privacy parameters. Finally, we have proposed an approximate version of our algorithm that efficiently anonymizes large multivariate numerical datasets and usually attains a better trade-off between disclosure risk and information loss in comparison with the most recent microaggregation algorithms.

The remaining of this paper is organized as follows. Section 2 formalizes the microaggregation problem, and Section 3 reviews some related algorithms to solve the problem. Section 4 describes the proposed method in detail. Experimental results are given in Section 5, and finally, Section 6 concludes the paper, and discusses some future work.

2. Problem definition

In this section, microaggregation problem is formulated. Suppose a numerical dataset T of n records $x_i,\ i \in \{1,\ldots,n\}$ in a $d \geqslant 1$ dimensional space is given. The data publisher provides an input value k as the privacy parameter. This value is usually lower than 10 in practical Statistical Disclosure Control (SDC) usages [12]. A microaggregation algorithm partitions the dataset into c groups such that all records in the same group are assigned the same number, different from the numbers of other groups. The partitioning is accomplished in such a way that the two constraints below are satisfied:

1. The group size constraints[2] are satisfied, i.e., $k \leqslant |G_p| < 2k$, $\forall p \in \{1,\ldots,c\}$, where $|G_p|$ denotes the number of records in G_p.
2. The whole dataset must be partitioned into c non-overlapping groups, i.e., $\cup_{p=1}^{c} G_p = T$, and $G_p \bigcap G_q = \varnothing,\ \forall\ p,q \in \{1,\ldots,c\}$, $p \neq q$.

The objective of microaggregation algorithms is to minimize the sum of within-group squared error (SSE), to obtain groups of similar records. The measure is formulated in Eq. (1).

$$SSE = \sum_{p=1}^{c}\sum_{j=1}^{|G_p|}(x_{pj} - \bar{x}_p)^T(x_{pj} - \bar{x}_p), \tag{1}$$

where x_{pj} is the j-th record of G_p and $\bar{x}_p$ denotes the centroid of G_p, $\bar{x}_p = \frac{1}{|G_p|}\sum_{j=1}^{|G_p|} x_{pj}$. The value of SSE is usually normalized by SST, calculated in Eq. (2), where $\bar{x}$ is the average of the whole dataset.

$$SST = \sum_{p=1}^{c}\sum_{j=1}^{|G_p|}(x_{pj} - \bar{x})^T(x_{pj} - \bar{x}). \tag{2}$$

The normalized measure $IL = SSE/SST * 100\%$ is always between 0% and 100%. Lower values of IL indicate more similar centroids to original records and less quality degradation due to the perturbation. Another information loss measure is the Global Certainty Penalty (GCP) [17]. In order to capture the range of the values in the groups, the GCP utilizes the Normalized Certainty Penalty (NCP) [18]. For a numerical protected dataset T' of n records in a d dimensional space, the NCP quantifies information loss of a single group $G \in T'$, and is defined in Eq. (3).

$$NCP(G) = \sum_{i=1}^{d}\frac{\max_{A_i}^{G} - \min_{A_i}^{G}}{\max_{A_i} - \min_{A_i}}. \tag{3}$$

In Eq. (3), the numerator and denominator represent the ranges of attribute A_i in the group G and in the whole dataset, respectively.[3] The GCP measures the total information loss of T' by Eq. (4).

$$GCP(T') = \frac{\sum_{G \in T'}|G| \cdot NCP(G)}{d \cdot n}. \tag{4}$$

The GCP is between 0 and 1, where 0 means no information loss ($T' = T$), and 1 signifies total information loss, i.e., all records are assigned to the same group.

3. Related works

Microaggregation is designed originally for numerical datasets. However, there are extensions for other data types [2,19,20]. Univariate microaggregation can be solved optimally, in a polynomial time [15], while the multivariate microaggregation is shown to be NP-hard [16]. There are multiple heuristics in the literature for solving multivariate microaggregation [2,21–24]. These methods are classified into fixed size and data-oriented algorithms. In a fixed size algorithm, the number of clusters is fixed to $\lfloor n/k \rfloor$ and exactly k records are aggregated within a cluster (except possibly a few clusters when k does not divide n). Generally, fixed size algorithms are more efficient, but usually result in more distorted protected data. In contrast, data-oriented methods do not specify the exact number of group members beforehand, and produce more useful protected data at the expense of increased time complexity [25].

There are some extensions of optimal univariate microaggregation that is introduced by Hansen and Mukherjee (called MHM in this paper) [15] for multivariate datasets. Domingo-Ferrer et al. [22] proposed a method to order records in the multi-dimensional domain space, and form a path based on some heuristics such as Nearest Point Next (NPN-MHM), MDAV-MHM, and CBFS-MHM with quadratic time complexity. Then, MHM is applied to the records on the path. Moreover, Heaton [26] reported some successful experiments in reducing IL after arranging records in a TSP cycle, breaking the cycle into a path, and then applying MHM on the sequence of records in the path. However, the method is not useful to protect large datasets, because it relies on another NP-hard problem.

Domingo-Ferrer and Torra [2] introduced a well-known fixed size method called Maximum Distance to Average Vector (MDAV).

[2] In addition to the minimum group size constraint, it is proved that for a given privacy requirement, k, a group with more than $2k - 1$ records can be split into more groups in order to decrease information loss [7].

[3] For simplicity, we assumed the same importance weight for all attributes.