Information Fusion 25 (2015) 85–104

Contents lists available at ScienceDirect

Information Fusion

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

Preference-based anonymization of numerical datasets by multi-objective microaggregation

Reza Mortazavi, Saeed Jalili *

Computer Engineering Department, Electrical and Computer Engineering Faculty, Tarbiat Modares University, Tehran, Iran

ARTICLE INFO

Article history:
Received 11 June 2014
Received in revised form 19 September 2014
Accepted 19 October 2014
Available online 30 October 2014

Keywords:
k-anonymity
Microaggregation
Multi-objective optimization
Privacy

ABSTRACT

Microaggregation is a statistical disclosure control mechanism to realize *k*-anonymity as a basic privacy model. The method first *partitions* the dataset into groups of at least *k* records and then *aggregates* the group members. Generally, larger values of *k* provide lower Disclosure Risk (*DR*) at the expense of increasing Information Loss (*IL*). Therefore, the data publisher has to set appropriate microaggregation parameters to produce a protected and useful anonymized data. Unfortunately, in the most of the conventional microaggregation methods, the only available parameter of the algorithm, i.e., *k* does not enable the data publisher to effectively control the trade-off problem between *DR* and *IL*. This paper proposes a novel microaggregation method to optimize information loss and disclosure risk, simultaneously. The trade-off problem is expressed and solved within a multi-objective optimization framework. The data publisher can choose a more preferred protected dataset from a set of non-dominated candidate solutions, or even direct the method toward a desired point. Experimental results show that for a fixed value of *k*, the proposed method can usually produce more protected and useful datasets in comparison with the conventional methods.

1. Introduction

The growing demand for data sharing in the form of microdata has made the protection of personal data a major society concern. To achieve the privacy, some models such as *k*-anonymity [1], *l*-diversity [2], and *p*-sensitivity [3] are introduced. In a *k*-anonymous dataset, all records are partitioned into groups of at least *k* members, where *k* is the aggregation level defined by the data publisher. In order to produce a protected dataset, all group members are aggregated, and then the aggregated values are published.

There are multiple protection methods introduced in the Statistical Disclosure Control (SDC) literature such as noise addition [4], synthetic microdata generation [5], microaggregation [6], and hybrid methods [7]. The methods are classified as perturbative and non-perturbative based on the effect on the original data [8,9]. Additionally, depending on the underlying data type, they are divided into methods for continuous and categorical data [10]. More comprehensive reviews about the methods can be found in previous studies [10,11].

Microaggregation is a mechanism to realize *k*-anonymity [12], and often is used by statistical agencies [13]. The method is originally proposed for continuous numerical attributes, while there are extensions for other data types [14,15]. Moreover, other complex privacy models may be implemented by variations of microaggregation [16,17]. The method is categorized as a perturbative approach to sanitize a database, where the original values in the dataset are modified to satisfy *k*-anonymity requirements. In an operational point of view, microaggregation is defined in terms of partition and aggregation [11]. In the partition phase, records are clustered into groups of at least *k* members, while aggregation involves a fusing method which produces centroids to represent group members. Finally, all records are replaced by their associated centroids to hide the original values.

There is usually a tension between respondent privacy and data utility, i.e., more privacy generally results in a degraded data quality and vice versa. Finding the optimal combination of these two measures is a difficult and challenging task [18]. The data publisher has to execute a protection algorithm (or a set of different methods) with different parameters to attain a desired *trade-off* between privacy and utility. Domingo-Ferrer and Torra have shown that microaggregation algorithms produce promising results in terms of both the privacy and the utility preserved after protection [19]. Different microaggregation methods have similar characteristics in general, i.e., as the aggregation level *k* of a microaggregation algorithm increases, more privacy is achieved and it is less probable for an intruder to associate a record to its

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

http://dx.doi.org/10.1016/j.inffus.2014.10.003
1566-2535/

corresponding data subject (record owner). This type of Disclosure Risk (*DR*) is called linkage disclosure [20]. Linkage disclosure is not the only measure to quantify *DR* of a protected dataset. In fact, even if an intruder cannot find the exact record of a data subject, she may still be able to limit the value of a confidential attribute of an individual to a narrow range. Interval disclosure quantifies this kind of risk after microaggregation for numerical datasets.[1]

Generally, in almost all conventional microaggregation methods, k is the only available parameter for the data publisher, which limits the way to control *DR* and *IL* of a protected data. In order to attain a maximum allowed *DR*, the data publisher has to increase k, which usually results in an undesirable quality degradation of published data.

The main contribution of this paper is to propose a new preference-based microaggregation algorithm to address the mentioned drawbacks. The method enables the data publisher to effectively control the trade-off between *DR* and *IL* for a fixed value of k. In other words, the proposed method enables the data publisher to anonymize a dataset with a larger value of k without a significant loss of its utility. The optimization problem between *DR* and *IL* of the protected dataset is stated as a multi-objective optimization problem that is solved by an evolutionary algorithm. Our novel encoding for both the partition and aggregation is based on the optimal univariate microaggregation algorithm [21]. It is *effective*, i.e., explores the whole space of valid partitions, and *efficient*. Furthermore, the proposed method is *flexible* and can be applied to anonymize large scale datasets for different performance indices. The solution consists of a number of non-dominated points that each of them represents a candidate protected dataset for a public privacy preserving data release. The data publisher can choose a more desirable solution based on the specific application constraints or even state a more preferred trade-off point to direct the method toward it during the algorithm execution. Experimental results show that our method is effective to decrease both the conflicting measures, *DR* and *IL*.

The remainder of the paper is organized as follows: Section 2 reviews some background concepts. This includes a formal definition of microaggregation, clarification of some performance measures to assess a protected dataset, and some definitions about multi-objective optimization. Section 3 is devoted to review some previous microaggregation algorithms and preference-based anonymization methods. Section 4 introduces the partition and aggregation phases of the proposed method and presents a way to incorporate the data publisher preferences during the anonymization process. Section 5 reports evaluation results and compares them with the results of some conventional microaggregation algorithms. The comparison of the proposed method with respect to previous ones is mentioned in Section 6. Finally, Section 7 summarizes and concludes the paper.

2. Background

In this section, we describe the classic[2] version of microaggregation problem for continuous datasets, introduce the performance measures to assess a protected dataset, and review some definitions about multi-objective problems.

2.1. Microaggregation problem

Suppose a numerical dataset T of n records $x_i, i \in \{1,\dots,n\}$ in a d dimensional space is given. The data publisher selects an aggregation level k, which is usually lower than 10 for practical SDC applications [13]. However, the value may be increased up to hundreds in some real-world applications, such as Location-Based Services (LBS) [22,23]. An *eligible partitioning* algorithm of a *valid* microaggregation method clusters the dataset into c partitions such that the following two constraints are satisfied.

1. Group size limits are satisfied, i.e., $k \leqslant |G_p| < 2k$, $p \in \{1,\dots,c\}$, where $|G_p|$ denotes the number of records in G_p.
2. The whole dataset is partitioned into c non-overlapping groups, i.e., $\cup_{p=1}^{c} G_p = T$, and $G_p \bigcap G_q = \emptyset,\ \forall p,\ q \in \{1,\dots,c\},\ p \neq q$.

The main aim of a classic microaggregation algorithm is to produce a protected dataset which is as similar as possible to the original dataset for a given aggregation level enforced by the parameter k. Therefore, it attempts to aggregate similar records and calculates the centroids that best represent all group members. To this end, the sum of within-group squared error (*SSE*) is used for optimization. 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. The measure is minimized for $\bar{x}_p = \frac{1}{|G_p|}\sum_{j=1}^{|G_p|} x_{pj}$. 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 $L = SSE/SST$ is always between 0 and 1. Lower values of L indicate more similar centroids to original records and less quality degradation due to the perturbation.

2.2. Performance measures of microdata protection methods

The data publishers must be able to assess different mechanisms with different parameters to find the one that best suits their requirements. The assessment of a microdata protection method is based on *DR* and *IL* measures.[3] Given an original dataset $T = \{x_i\},\ i \in \{1,\dots,n\}$, an anonymization mechanism $\mathcal{F}$ produces $T' = \{x'_i\},\ i \in \{1,\dots,n\}$, where $x'_i = \mathcal{F}(x_i)$. [4] There are two general risk types that are averaged as *DR* [10]:

1. Linkage Disclosure[5](LD): This is the standard mechanism used to measure disclosure risk of a protection method [24]. One kind of record linkage is called Distance-based Linkage Disclosure (*DLD*) [25] that calculates the distances between original and protected records. In this paper, similar to [26], we consider a scenario in which the intruder tries to link a protected record to its corresponding original one.[6]

[1] See Section 2.2 for a formal definition.

[2] We used the term classic to notify that the main objective of the methods is to reduce *SSE*, in contrast with our proposed method to reduce *DR* and *IL*.

[3] We review some general purpose *DR* and *IL* measures only for continuous data type, which is addressed in this paper. More details of the measures for other data types can be found in [10].

[4] We assume that the mechanism $\mathcal{F}$ produces the same number of protected records and in the same order of the original records in T.

[5] It is also known as identity disclosure or re-identification risk.

[6] In other words, the search is conducted on the original dataset for each protected record. This definition enables us to accelerate our computations where *DLD* is to be computed for a large number of different protected datasets, since the search data structure is constructed only once on the original dataset. However, our experiments on more than 500 different protected datasets show that the computed index in this approach is highly correlated to the *DLD* value of the case where the search is conducted on the protected dataset.