Computational Statistics and Data Analysis 79 (2014) 1–13

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Data privacy using an evolutionary algorithm for invariant PRAM matrices

Jordi Marés [a,*], Natalie Shlomo [b]

[a] Artificial Intelligence Research Institute (IIIA), Spanish Council of Scientific Research (CSIC), Universitat Autònoma de Barcelona (UAB), Bellaterra, Catalonia, 08193, Spain

[b] Social Statistics and the Cathie Marsh Centre for Census and Survey Research (CCSR), The University of Manchester, Humanities Bridgeford Street, Manchester, M13 9PL, UK

ARTICLE INFO

Article history:
Received 22 May 2013
Received in revised form 6 April 2014
Accepted 2 May 2014
Available online 13 May 2014

Keywords:
Probability transition matrices
Genetic operators
Fitness function
Disclosure risk
Data utility

ABSTRACT

Dissemination of data with sensitive information has an implicit risk of unauthorized disclosure. Several masking methods have been developed in order to protect the data without the loss of too much information. One such method is the Post Randomization Method (PRAM) based on perturbations of a categorical variable according to a Markov probability transition matrix. The method has the drawback that it is difficult to find an optimal transition matrix to perform perturbations and maximize data utility. An evolutionary algorithm which generates an optimal probability transition matrix is proposed. Optimality is with respect to a pre-defined fitness function dependent on the aspects of the data that need to be preserved following perturbation. The algorithm embeds two properties: the invariance of the transition matrix to preserve marginal totals in expectation, and the control of diagonal probabilities which determine the amount of perturbation. Experimental results using a real data set are presented in order to illustrate and empirically evaluate the application of this algorithm.

1. Introduction

With growing demands for more data to carry out research and inform policies, privacy and the prevention of disclosure of sensitive information about persons or organizations held in released data sets have become of increasing importance in recent decades. Statistical agencies and data providers are under legal and ethical obligations to preserve the privacy of responding units included in these data. On the other hand, technological advancements and the wide use of the internet have made it easier for potential 'intruders' to gather information for initiating attacks on the data. For these reasons, statistical agencies and data providers must carefully consider the release of data where there are generally two options: restricted data and restricted access, or a combination of both. Under restricted access, special licenses may be required to use the data which is only made available in on-site secure locations. Under restricted data, the data is coarsened or altered prior to its release. The need to preserve privacy in restricted data has led to much research and development on statistical disclosure control techniques which aim to construct a masked data set that maintains the privacy of the data respondents whilst minimizing the loss of information (Fienberg, 1994).

The data included in these data sets can be of multiple types but the two most common are continuous and categorical variables. Continuous variables, such as age and income, are numerical and can take any value within an interval. Categorical

* Corresponding author. Tel.: +34 93 580 95 70; fax: +34 93 580 96 61.
E-mail addresses: jmares@iiia.csic.es, jordimares@gmail.com (J. Marés), natalie.shlomo@manchester.ac.uk (N. Shlomo).

http://dx.doi.org/10.1016/j.csda.2014.05.002
0167-9473/

variables, such as place of residence and ethnicity, have a finite range of values and are more difficult to protect because of their limited range of values. The indirect identifying variables in a data set are typically categorical variables and must be protected since they can be used to identify a responding unit in the data leading to the disclosure of attributes.

The protection of categorical variables can be carried out using a statistical disclosure control method called Post Randomization (PRAM). PRAM was introduced in Gouweleeuw et al. (1998) and Kooiman et al. (1997) as a method for masking categorical variables in microdata files based on a Markov probability transition matrix. The difficulty in finding an appropriate transition matrix and the fact that analyses have to be adjusted to obtain valid results from a perturbed data set are the main reasons why it is not widely used.

There have been several approaches to obtain optimal PRAM transition matrices in the literature. In Rebollo-Monedero et al. (2010), the authors propose an analytical approach to compute PRAM transition matrices based on information theory measures. In Cator et al. (2005), the authors provide a similar analytical approach which depends on minimizing information loss measures using Maximum Likelihood Estimation (MLE) and information theory measures.

A technique to boost the usability of PRAM and the performance of probability transition matrices is to include the property of invariance. This technique is also discussed in Van den Hout (2004) and Van den Hout and Elamir (2006) as a corrective method for analyzing perturbed data sets. The invariance property ensures that sufficient statistics of the protected variables in the perturbed data are preserved in the expectation and that the perturbed data is an unbiased moment estimator of the original data. Another technique to boost the usability of PRAM is to control the diagonal probabilities of the transition matrices. The diagonal probabilities determine the desired level of perturbation which is generally set by the standards and policies within the governing board of the agency releasing the data. In addition, dominant diagonals guarantee that transition matrices can be inverted.

The proposed approach for obtaining optimal probability transition matrices for use in PRAM is through an evolutionary algorithm. The proposed evolutionary algorithm will embed the property of invariance and include controls on diagonal probabilities of the generated transition matrices. Evolutionary algorithms are optimization algorithms that search for a better solution of the transition matrix at each iteration (generation) until a stopping condition is reached. Optimality is with respect to a pre-defined fitness function dependent on the aspects of the data that need to be preserved following perturbation. Under the proposed evolutionary algorithm, probability transition matrices show better performance with respect to selected quantifying measures of disclosure risk and data utility. The previous work was shown in Marés and Torra (2010). However, in that paper, the transition matrices obtained were not statistically valid and led to final distributions of perturbed variables very different from their original distributions, producing a useless data set for statistical analysis.

There exist other types of optimization techniques which may be used to optimize PRAM transition matrices such as MCMC algorithms (Andrieu et al., 2003) and the simulated annealing optimization approach (Press et al., 2007). MCMC algorithms require prior probabilities to initiate the algorithm and can be difficult to compute with sufficient accuracy. In the evolutionary algorithm approach, no prior information is needed to explore the search space of transition matrices. In simulated annealing, the algorithm only moves to close neighbors which can lead to large computation time to reach a further good transition matrix. In the evolutionary algorithm approach, the operators allow big 'jumps' in the search space and hence avoid a large number of intermediate possible solutions.

The remainder of this paper is organized as follows. Section 2 describes the Post Randomization Method. The description of our proposed algorithm is provided in Section 3. Experimental results with a real data set are shown in Section 4. Finally, Section 5 contains some concluding remarks.

2. The post randomization method (PRAM)

The Post Randomization method (PRAM) was introduced in Gouweleeuw et al. (1998) and Kooiman et al. (1997) as a method for masking categorical variables in microdata files. In De Wolf and Van Gelder (2004), De Wolf et al. (1998) and Van den Hout (2004) the method and some of its implications are discussed in more detail. However, the PRAM method is still one of the least used for protecting microdata because of the difficulty in obtaining an optimal transition matrix for perturbing the data whilst maintaining data utility. This was demonstrated in experiments carried out in Domingo-Ferrer and Torra (2001) where the PRAM method based on a maximum entropy transition matrix (every category has an equal probability to switch to another category), was shown to have a low overall score calculated as an average of data utility and disclosure risk measures compared to other methods.

The PRAM method is as follows: let t be the vector of frequencies and t/N the vector of relative frequencies of a categorical variable having L categories and N is the number of records in the microdata. Let X be a $L \times L$ probability transition matrix containing conditional probabilities: $x_{ij} = p(value_{perturbed} = j | value_{original} = i)$. In each record of the data, the category of the variable is changed or not changed according to the prescribed transition probabilities in the matrix X and the result of a draw of a random multinomial variate u with parameters x_{ij} $(j = 1, \ldots, L)$. If the j-th category is selected, category i is moved to category j. When $i = j$, no change occurs.

Let t^* be the vector of perturbed frequencies. t^* is a random variable and $E(t^*|t) = tX$. Assuming that the transition matrix X has an inverse X^{-1}, this can be used to obtain an unbiased moment estimator of the original data: $\hat{t} = t^*X^{-1}$. Statistical analyses can be carried out on $\hat{t}$. In order to ensure that the transition matrix has an inverse and to control the amount of perturbation, the matrix X is dominant on the main diagonal with entries over 0.5.