Multivariate Equi-width Data Swapping for Private Data Publication

Yidong Li and Hong Shen

School of Computer Science, University of Adelaide, SA 5005, Australia
{yidong,hong}@cs.adelaide.edu.au

Abstract. In many privacy preserving applications, specific variables are required to be disturbed simultaneously in order to guarantee correlations among them. Multivariate Equi-Depth Swapping (MEDS) is a natural solution in such cases, since it provides uniform privacy protection for each data tuple. However, this approach performs ineffectively not only in computational complexity (basically $O(n^3)$ for n data tuples), but in data utility for distance-based data analysis. This paper discusses the utilisation of Multivariate Equi-Width Swapping (MEWS) to enhance the utility preservation for such cases. With extensive theoretical analysis and experimental results, we show that, MEWS can achieve a similar performance in privacy preservation to that of MEDS and has only $O(n)$ computational complexity.

Keywords: Private data publication, data swapping, equi-width partitioning, multivariate data perturbation.

1 Introduction

Private data publication has been widely studied in statistical disclosure control and privacy preserving data mining areas. The core task for a qualified publication is to disturb data in a way that does not lead to disclosure of sensitive information, but maintains data utility as much as possible. Many existing studies [1,10,7,11] are focused on developing algorithms performing univariate data perturbation, which operates each variable in a dataset individually. However, many real-world applications with multivariate data are required to guarantee data utility of several variables (or attributes), which have closer correlations than with others. This paper discusses the utilisation of *data swapping* for multivariate data perturbation.

As the name implies, data swapping is to disturb a dataset by exchanging values of sensitive variables among data tuples. This method is a natural solution to protect confidential information from identity disclosure [10,7], while maintains lower-order statistics of a dataset with its value-invariant property. Equi-depth swapping is a widely used solution to guarantee each bin containing roughly the same number or frequency of data tuples, so as to provide the same level of privacy protection for all entities. However, this approach introduces the computational complexity as $O(n^3)$, where n is the size of dataset in

M.J. Zaki et al. (Eds.): PAKDD 2010, Part I, LNAI 6118, pp. 208–215, 2010.

multivariate scenarios. In addition, it performs ineffectively on preserving data utility for distance-based applications, such as data mining on interval data and multivariate density estimation with histogram, where bins are determined by their relative distances rather than relative orders.

This paper considers the use of data swapping with *equi-width* partitioning, which ensures the width of each bin approximately the same, to overcome the drawbacks above. However, Equi-Width Swapping (EWS) can preserve data privacy at the similar level as EDS for large datasets. It is motivated by the idea that, the value-invariant method hides the detailed partitioning information such as swapping distance. Our analysis shows that, this approach can achieve good tradeoff between data utility and privacy.

The remainder of this paper is organized as follows. Section 2 presents a brief overview of the related literature. In Section 3, we provide algorithms for both MEDS and MEWS. Section 4 presents an exclusive analysis on privacy for the methods above. Section 5 discusses experimental results to justify the effectiveness of MEWS. Finally, we conclude this paper in Section 6.

2 Related Work

Data swapping was first introduced in [4] as an efficient value-invariant approach for statistical disclosure control, and the following work [9] extends the idea to preserve numerical data. Moore et al. [8] described a popular local swapping method based on equi-depth partitioning with univariate ranking. In this method, a term called *swapping distance* defines the depth of each swapping domain, i.e., the number of tuples in each interval for swapping. Then it ranks and localizes data tuples in the light of a specified variable in each iteration, and then swaps candidates in each interval randomly. However, this study is limited with its assumption that the data is uniformly distributed.

The work in [3] follows the idea of ranking but performs random sampling as localization. Although it provides good maintenance of correlations among variables, this technique has been proved highly inefficient [6] and provides practically no protection from attacks. A recent work [11] proposes a swapping-like method named data shuffling, which is based on joint and/or conditional distribution of variables in the original dataset, in order to minimize disclosure risk of sensitive data. A comprehensive study of data swapping and its applications can be found in [5].

3 Multivariate Data Swapping

In this section we first introduce some basic notation that will be used in the remainder of this paper. Let $X = \{x_1,\ x_2,\ \ldots,\ x_n\}$ be a dataset. Let A denote the set of all attributes $\{A_1,\ A_2,\ \ldots,\ A_m\}$ and $x[A_i]$ denote the value of attribute A_i for a tuple x. We define *swapping set* as a set of attributes $\{A_i,\ \ldots,\ A_j\} \subset A$ $(i < j)$ for simultaneous perturbation, denoted by S, and let $\overline{S}$ represent the set of all other attributes. Then we use $x[S]$ to denote a

Algorithm 1. Multivariate EDS

Input: The input dataset X, the swapping set S, and swapping distance k
Output: A perturbed dataset Y
1: $x_r = x_s = \mathbf{0}$, $Y = \phi$;
2: while $|X| \geq 2k$ do
3: select x_r, x_s, where $\|x_r[S] - x_s[S]\|$ is the largest;
4: form two bins b_r and b_s containing x_r and x_s with their $k-1$ nearest neighbours;
5: for each group do
6: select pairs of elements randomly;
7: swap swapping set values for each pair;
8: $Y \leftarrow b_r \cup b_s$ and remove them from X;

partial tuple $(x[S_1 = A_i],\ \ldots,\ x[S_p = A_j])$ $(p < m)$, which is the projection of x onto the attributes in S. Without loss of generality, we assume there is only one swapping set containing the first p attributes $A_1,\ A_2,\ \ldots,\ A_p$, and it is not a unique set to identify individuals.

3.1 Multivariate EDS

This section presents a heuristic method for multivariate equi-depth swapping. The idea is to cluster tuples based on Euclidean distance between each other and then swap attribute values in the swapping set simultaneously. We define $\mathbf{k} = (k[A_1],\ k[A_2],\ \ldots,\ k[A_m])$ as a vector of swapping distance for each attribute. As the assumption that the swapping set consists of the first p attributes, the swapping distances for these attributes are equal, i.e., $k[A_1] = k[A_2] = \ldots = k[A_p] = k$. Algorithm 1 describes the process of perturbing a dataset according to its swapping set. The perturbation for attributes in $\overline{S}$ is omitted here since it is exactly the same as in univariate EDS [7].

In Algorithm 1, the computational complexity of computing the most distant tuples x_r and x_s is $O(n^2)$. The swapping process only costs $O(\frac{k}{2})$. There are $\lfloor \frac{n}{2k} \rfloor$ iterations. Therefore, the total computational complexity of Multivariate EDS is $O(\frac{n^3}{k})$, where k is the swapping distance. In real-world cases, $k \ll n$, thus the complexity is $O(n^3)$ finally. We can improve this algorithm by calculating a symmetrical distance matrix in advance. This will reduce the time complexity to $O(n^2)$. However, such matrix introduces the storage complexity as $O(\frac{n(n-1)}{2})$, while the original method only has $O(n)$. As the size of data set growth, the distance matrix becomes impractical to be stored in memory.

Since MEDS only considers the relative ordering among tuples for clustering, it inherently leads to errors in many distance-based applications, such as data mining on interval data and density estimation with histogram. These require a perturbation method to maintain not only the ordering but also quantitative properties of the bins including bin width, distance between bins and relative distance between tuples. Moreover, this method only guarantees data utility in bins formed with the p variables in swapping set. That is, if a data user is willing to explore the published data in lower-dimension (e.g., a subset of swapping set),

Algorithm 2. Multivariate EWS

Input: The input dataset X, the swapping set S, and swapping distance t
Output: A perturbed dataset Y
1: partition domain of X according to S and t;
2: allocate each tuple into its corresponding bin;
3: for each non-empty bin do
4: select pairs of elements randomly;
5: swap swapping set values for each pair;
6: end for
7: $Y \leftarrow X$.

it may cause unacceptable errors. In the following section, we shall discuss a multivariate swapping technique to improve data utility in the cases above while preserving data privacy.

3.2 Multivariate EWS

To get around of the deficiencies of MEDS, we propose Algorithm 2 by redefining swapping distance as bin width, denoted by $\mathbf{t} = (t[A_1],\ t[A_2],\ \ldots,\ t[A_m])$, where $t[A_i]$ is the width on A_i. It implies that the value change for each swapping candidate is at most t. Moreover, to split a dataset with continuous attributes, every bin resulted by univariate partitioning will not be empty if partition degree is larger than a threshold. However, for a multivariate partitioning, many bins may not hold any tuple even with a very small threshold. These empty bins have no use for analysis of utility and privacy. Therefore, we modify the definition of *partition degree* [7] for MEWS as follows:

Definition 1 *The* partition degree *is the number of non-empty bins formed by splitting the original dataset, denoted by* $d \in \mathbb{N}$.

Let d_{total} denote the total number of bins and for a partitioning PT_i, we have $d^{(i)} \leq d^{(i)}_{total}$. Then, the complexity of partitioning dataset is $O(d_{total})$, and that of allocating process is $O(n)$. The worst case of the iteration step is when $d-1$ bins contain one tuple each, and the other bin contains the rest $n-d+1$ tuples. Thus, the computational complexity of Algorithm 2 is $O(dn)$. While in real-world applications, $d \ll n$, we have the complexity as $O(n)$.

The Pruning Scheme. Although the idea of Algorithm 2 is straightforward, it is even challenged by selecting swapping distance t. Theoretically, there exist infinite possibilities for t within the range $(0,\ \frac{domain(x)}{2}]$ due to $t \in \mathbb{R}^+$. However, many partitionings with small differences are duplicate and meaningless. We introduce the Pruning scheme that determines a proper domain of swapping distance efficiently.

Step 1: Calculate or assign a upper bound t_{ub} for t;
Step 2: Compute a lower bound t_{lb} for t;

Step 3: Reduce the number of partitionings to a finite number.

The first step guarantees data utility of a published dataset, since a too-large t can make perturbed data useless due to its unacceptable bias. Step 2 considers data privacy of a published dataset. The idea behind is to ensure that each non-empty bin must hold at least two tuples for data exchanging. The final step is to further remove the redundant candidates. we define *Minimal Depth Bin (MDB)* to represent a bin that contains the least number of tuples corresponding to t and let N^* represent the number of an MDB. Then, we say two partitionings with $t \in [t_r,\ t_s]$ are similar if $N^*_{r-1} < N^*_r = \ldots = N^*_s < N^*_{s+1}$. Since the similar partitionings lead to the same swapping result in MDB and the number of sets of similar partitionings are limited in a finite dataset, the Pruning scheme finally reduces the number of partitionings to a finite number. (The details of selections can be found in the extended version due to space limits.)

4 Privacy Analysis

MEWS theoretically guarantees zero-loss of utility among variables in swapping set, which makes this approach applicable to a large category of applications of multivariate data analysis. We dedicate this section to discuss how good MEWS is in terms of preserving privacy.

As the concept of privacy may have various concerns, we describe the privacy, P, for the entire dataset as *the probability of revealing a swapping pair who has the highest disclosure risk in the database*, which is $P = max(Pr(X = x_i))\quad (1 \leq i \leq n)$. The disclosure risk for a tuple can be computed as,

$$Pr(X = x_j[S]) = \sum_{l=1}^{q} Pr(d_l) Pr(X = x_j[S] \mid d_l) \tag{1}$$

where $(x_i[S],\ x_j[S])$ is a swapping pair, q indicates the number of valid partitionings according to the Pruning scheme. Then, we can compute the privacy for MEDS and MEWS with the following lemmas.

Lemma 1. *Given n data tuples and swapping distance k, the privacy for MEDS is*

$$P_{ed} = \frac{1}{q_{ed}} \times \Big(\sum_{i=1}^{q_{ed}/2} \frac{1}{i} + \frac{q_{ed}}{2(q_{ed}+2)}\Big) \simeq \frac{ln(q_{ed}) + c_1}{q_{ed}} \tag{2}$$

where q_{ed} is the number of possible partitionings and c_1 is a small constant in the range of $[\gamma + 0.2,\ 1.2]$.

The proof of Lemma 1 is omitted here and a similar process can be found in [7] if interest. The above Lemma shows that, if a dataset is large, the privacy provided by MEDS is only relative to the number of possible partitionings rather than the number of tuples in a bin.

Lemma 2. *Given a dataset X and swapping distance t, the privacy for MEWS is,*

$$P_{ew} = \frac{1}{q_{ew}} \times \sum_{i=1}^{q_{ew}} \frac{1}{N_i^* - 1} \simeq \frac{ln(q_{ew}) + c_2}{q_{ew}}, \tag{3}$$

where q_{ew} is the number of possible partitionings and c_2 is a small constant within the range $[\gamma,\ 1]$.

Although the proof of the lemma is omitted due to space constraints, one can see the privacy provided by MEWS is only relative to the number of possible partitionings which is the same as MEDS.

Based on Lemma 1 and 2, let us consider perturbing a large database, which is quite common in data mining. Generally speaking, MEDS provides more possible partitionings than MEWS does, i.e., $q_{ed} \geq q_{ew}$. But for a dataset with large size of tuples, the domain of swapping distance will be large even with the Pruning scheme. That is, q_{ew} can still be large enough for protecting data. Moreover, with the same size of MDB N^*, the two approaches result in different partition degree d_{ed} and d_{ew}, where $d_{ed} \geq d_{ew}$ in usual. However, degree is not a deterministic parameter for privacy measuring. It also implies that MEWS can provide good performance on preserving privacy as MEDS. We will show further analysis on privacy based on specific distributed datasets in the experimental part.

5 Experiments

5.1 Privacy on Binormal Distribution

As the privacy for MEWS is sensitive to data distribution, it is deserved to explore relations between privacy and distribution parameters by generating a dataset with normalized bivariate normal distribution $f_{XY}(x, y; 1, 1, 0, 0, \rho_{XY})$, where ρ is the correlation coefficient of attributes X and Y. In addition, a data suppression is adopted to restrict domains of attributes within the range $[-2,\ 2]$, which is reasonable because most of sparse data will be cleaned before data analysis.

The results in Table 1 show that the privacy P for MEWS has direct relationships with three factors: ρ, t_{ub} and n. For a large dataset (e.g., $n = 100,000$), the disclosure probability is very small ($P \leq 7\%$) even the attributes are not highly related, and turns to large ($P \geq 30\%$) while the correlations of attributes are small and the size of data is not large (e.g., the up-left corner of the table).

Table 1. The impact of distribution on privacy P for MEWS

	$n = 10,000$			$n = 100,000$		
t_{ub}	$\rho = 0.1$	0.5	0.9	0.1	0.5	0.9
0.2	0.3466	0.2986	0.1450	0.1324	0.0635	0.0250
0.4	0.1973	0.1060	0.0461	0.0359	0.0172	0.0069
0.8	0.0443	0.0254	0.0125	0.0066	0.0036	0.0017

Since most datasets for data mining are very large, and attributes in a swapping set generally have very close relations (otherwise the swapping set will make no sense for data analysis), we can observe that the MEWS method can provide good data protection. This is not intuitive but reasonable because constraints on equi-width partition is much more relaxed than that on equi-depth partition. This property is held only in the context of local data swapping rather than other local perturbation approaches.

5.2 Comparison of Data Utility

Since swapping distance has various definitions for MEDS and MEWS, we consider the impact of partition degree on the correlation matrix here. Given a multivariate dataset X, let ρ_{ij} be correlation coefficient of X_i and X_j and $\rho_{ij}^{'}$ be the corresponding coefficient once data have been swapped using a specified method. Then, the bias of the correlation is $\Delta\rho_{ij} = \left|\rho_{ij} - \rho_{ij}^{'}\right|$. We compute the average $\overline{\Delta\rho}$ and the standard deviation $\sigma_{\Delta\rho}$ as utility metrics. The experiments are run over one synthetic and two real datasets: 1) *Syn100k*, contains 100,000 data tuples and four variables, which the first two variables are generated with the standard binormal distribution with $\rho = 0.5$ and others are generated individually with the standard normal distribution; and 2) *Abalone* and *MagicTele*, are both formed by their continuous variables in [2].

Table 2 shows that, the $\overline{\Delta\rho}$ decreases as the partition degree increases, which is intuitive and reasonable. In most cases, MEWS performs better than MEDS, especially when the partition degree is not large. For example, the $\overline{\Delta\rho}$ resulted by MEDS is twice more than that by MEWS in both sythetic datasets and much higher in the real datasets. In addition, during the experiment, we find that MEWS provides more steady performance than MEDS. It implies that the variance of $\sigma_{\Delta\rho}$ for MEDS is larger than that for MEWS.

It should notice that, even the experimental results show that MEWS can achieve a better parametric utility in our tested data, this can not be guaranteed in other cases. We can easily construct some datasets that fit better to MEDS. However, the MEWS method is still a good alternative for commonly used data distributions.

Table 2. Impact of Partition Degree on Correlations

		$d = 100$		$d = 500$		$d = 1000$	
Dataset	**Method**	$\overline{\Delta\rho}$	$\sigma_{\Delta\rho}$	$\overline{\Delta\rho}$	$\sigma_{\Delta\rho}$	$\overline{\Delta\rho}$	$\sigma_{\Delta\rho}$
Syn100k	MEWS	0.40	0.16	0.20	0.08	0.11	0.08
	MEDS	0.61	0.25	0.37	0.18	0.18	0.11
Abalone	MEWS	0.20	0.06	0.15	0.05	0.10	0.05
	MEDS	0.51	0.10	0.30	0.08	0.18	0.07
MagicTele	MEWS	0.30	0.09	0.21	0.10	0.12	0.08
	MEDS	0.62	0.20	0.41	0.18	0.20	0.10

6 Conclusion

This paper explores the use of multivariate equi-width swapping as a tool for private data publication. It makes primary benefits on two different grounds. First, the time complexity of the algorithm is linearly proportional to the size of a dataset, which makes this approach quite efficient especially for applying on very large datasets. Then, it provides excellent preservation on distance-based data utilities, which has great potential for use in the real-world data analysis. Compared to the multivariate equi-depth swapping which provides uniform privacy protection for each tuple, the proposed method still performs quite reasonably on privacy protection. In our future work, the basic MEWS method can be optimized to achieve more efficient variants for specified applications. Applying this approach in distributed private data publication offers another interesting direction.

References

1. Aggarwal, C.C., Yu, P.S.: A condensation approach to privacy preserving data mining. In: EDBT, pp. 183–199 (2004)
2. Asuncion, A., Newman, D.: Uci machine learning repository. University of California, Irvine (2007)
3. Carlson, M., Salabasis, M.: A data-swapping technique for generating synthetic samples; a method for disclosure control. Research in Official Statistics 5, 35–64 (2002)
4. Dalenius, T., Reiss, S.P.: Data-swapping: A technique for disclosure control (extened abstract). In: The Section on Survey Research Methods, Washington, DC, pp. 191–194 (1978)
5. Fienberg, S.E., McIntyre, J.: Data swapping: Variations on a theme by ddalenius and reiss. J. Official Statist. 21, 209–323 (2005)
6. Fienberg, S.E.: Comment on a paper by m. carlson and m. salabasis: a data-swapping technique using ranks - a method for disclosure control. Research in Official Statistics 5, 65–70 (2002)
7. Li, Y., Shen, H.: Equi-width data swapping for private data publication. In: PDCAT 2009: The Tenth International Conference on Parallel and Distributed Computing, Applications and Technologies, Hiroshima, Japan (2009)
8. Moore, R.A.: Controlled data-swapping techniques for masking public use microdata sets. Research report RR96/04, U.S. Bureau of the Census, Statistical Research Division, Washington D.C (1996)
9. Reiss, S.P., Post, M.J., Dalenius, T.: Non-reversible privacy transformations. In: PODS 1982: Proceedings of the 1st ACM SIGACT-SIGMOD symposium on Principles of database systems, pp. 139–146. ACM, New York (1982)
10. Mukherjee, S., Chen, Z., Gangopadhyay, A.: A privacy-preserving technique for euclidean distance-based mining algorithms using fourier-related transforms. The VLDB Journal 15, 293–315 (2006)
11. Muralidhar, K., Sarathy, R.: Data shuffling - a new masking approach for numerical data. Management Science 52(5), 658–670 (2006)