2011 Conference on Technologies and Applications of Artificial Intelligence

FISIP: A Distance and Correlation Preserving Transformation for Privacy Preserving Data Mining

Jen-Wei Huang
Dept. of Computer Science and Engineering
Yuan Ze University
135 Yuan-Tung Road, Chung-Li, Taiwan
jwhuang@saturn.yzu.edu.tw

Jun-Wei Su and Ming-Syan Chen
Dept. of Electrical Engineering
National Taiwan University
No. 1, Sec. 4, Roosevelt Road, Taipei, Taiwan
mschen@cc.ee.ntu.edu.tw

***Abstract*—This paper devises a transformation scheme to protect data privacy in the case that data have to be sent to the third party for the analysis purpose. Most conventional transformation schemes suffer from two limits, i.e., the algorithm dependency and the information loss. In this work, we propose a novel privacy preserving transformation scheme without these two limitations. The transformation is referred to as FISIP. Explicitly, by preserving three basic properties, i.e., the first order sum, the second order sum and inner products, of the private data, mining algorithms which depend on these three properties can still be applied to public data. Specifically, any distance-based or correlation-based algorithm has the same performance on the transformed public data as on the original private data. Special perturbation can be added into FISIP transformations to increase the protection level. In the experimental results, FISIP attains data usefulness and data robustness at the same time. In summary, FISIP is able to provide a privacy preserving scheme that preserves the distance and the correlation of the private data after the transformation to the public data.**

***Keywords*-Privacy preserving, data mining, distance, correlation**

I. Introduction

Privacy infringement is an important issue in the data mining. People and organizations usually do not tend to provide their private data or locations to the public because of the privacy concern [1]. Therefore, how to preserve privacy efficiently and effectively during the data mining process has become an emerging research issue.

Note that there are two major limitations in most privacy preserving approaches. The first limitation is the algorithm dependency in that the protection schemes are intrinsically incorporated into certain data mining algorithms. Such protection schemes are not easy to be utilized by other algorithms. The second limitation is the information loss. Most algorithms add some controlled noise in their private data or truncate part of private data to make them public. Consequently, the public data are unrecoverable to the original private ones. Though the privacy is thus protected, the mining results are somewhat altered due to these changes. In this work, we propose a novel privacy preserving scheme to conquer these two limitations. The proposed transformation scheme is referred to as FISIP, FIrst and Second order sum and Inner product Preservation. FISIP transforms the private data to public ones. Explicitly, as will be proved later, by preserving three basic properties, i.e., the first order sum, the second order sum, and inner products, of private data, mining algorithms which depend on these three properties can still be applied to public data. Specifically, distances and correlations can be derived from these three properties. Therefore, any distance-based or correlation-based algorithm has the same performance on the transformed public data as on the original private data. For example, the distance-based clustering algorithm, K-means [2], the density-based clustering algorithm, DBSCAN [3], the correlation clustering method [4], the distance-based and inner-product-based classification, SVM [5], and the neighbor-based classification, kNN [6], are applicable to the transformed data since the relative density can be derived from the transformed data. In addition, the correlation-based feature selection, CFS [7], and distance-based outlier detection [8] are carried out faithfully with FISIP. Moreover, if some ingenious algorithms are invented in the future and their measurements depends only on three properties which FISIP preserves, these algorithms can still work well without additional need to design other special privacy-preserving procedures.

We propose a systematic method to derive a perfect FISIP transformation. Perfect FISIP transformations perfectly preserve three basic properties and provide a certain level of privacy protection without any loss on the mining accuracy. However, if attackers obtain k pairs of independent k-dimensional public and private data, they may be able to estimate other private data using linear combinations of public data. This is called known private data attack. To provide more privacy protection against the known data attack, we also introduce a procedure to generate strong FISIP transformations, which increase attackers' estimation error by sacrificing little mining accuracy.

Evaluation of strong FISIP transformations is done in two folds. We investigate the data usefulness and the data robustness. Do the transformed public data preserve the

978-0-7695-4601-8/11 $26.00 © 2011 IEEE
DOI 10.1109/TAAI.2011.25

relations of original private data? Are the transformed data robust enough against the reconstruction attack? In the experimental results, strong FISIP transformations attain these two goals at the same time.

In all, by preserving three mathematical properties, FISIP is able to provide a transformation that preserves the distances and the correlations of the original private data after the transformation to the public data. As a result, while the privacy is preserved, the data mining quality from the transformed data can remain the same as that from the original private data.

The rest of this work is organized as follows. Preliminaries are given in Section 2, where the problem description is shown and related works are reviewed. Theoretical properties of FISIP are derived and systematic procedures to conduct perfect FISIP transformations are presented in Section 3. To provide more privacy protection, strong FISIP transformations are introduced in Section 4. Finally, experimental studies are shown in Section 5 and this work concludes in Section 6.

II. Preliminaries

A. Problem Description

In this work, we want to devise a transformation that preserves the distance and the correlation for the original private data after the transformation to the public ones. To facilitate our presentation, we here introduce the following definitions.

Definition 1: Distance Preserving Transformation: In k-dimensional vector space, given n points $\{r_1, r_2, ..., r_n\}$, a transformation T_d that can produce n points $\{u_1, u_2, ..., u_n\}$, such that $dist(u_i, u_j) = dist(r_i, r_j)$, $u_i = T_d(r_i)$, $1 \leq i, j \leq n$, is defined as the distance preserving transformation.

The distance metric used in this paper is Euclidean distance. However, the transformation results are of general usefulness since it has been shown that most distance matrices can be reduced to Euclidean form [9].

Definition 2: Correlation Preserving Transformation: In k-dimensional vector space, given n points $\{r_1, r_2, ..., r_n\}$, a transformation T_c that can produce n points $\{u_1, u_2, ..., u_n\}$, such that $corr(u_i, u_j) = corr(r_i, r_j)$, $u_i = T_c(r_i)$, $1 \leq i, j \leq n$, is defined as the correlation preserving transformation, where $corr(x, y)$ denotes the correlation of x and y.

B. Related Works

Most privacy preservation schemes can be roughly categorized into two fields. The first field targets on hiding the data entities of the published database. This field of research becomes popular in recent years because of the threat of quasi-identifiers. It was reported that 87% of US citizens can be uniquely identified by only their zip code, date of birth and gender [10]. Even though their names or social security numbers are truncated from the published database, identities of citizens may still be found via easily obtainable fields by linking attacks. Typical solutions to this concern are based on the k-anonymity model. The published data set is generalized such that there are at least k records in each group of quasi-identifiers. Other well-known models for protecting data entities include t-closeness [11], etc.

The second field of the privacy preservation focuses on hiding data values, instead of entities. Most well-known schemes in this field are based on the data perturbation [12]. Consider we draw n values $x_1, x_2, \ldots, x_n$ from the private data X, and perturb them by adding n independent values $y_1, y_2, \ldots, y_n$ drawn by a random variable Y. As long as the probability distribution function of Y is known, the distribution of X can be reconstructed by $x_1 + y_1$, $x_2 + y_2$, $\ldots$, $x_n + y_n$. However, although the distribution can be reconstructed, it cannot guarantee to have the same mining result from the private database and the reconstructed public database. It also cannot hold analytical properties, such as having the first order sum and the second order sum be the same under such kind of privacy preservation transformation.

In most cases, we want to preserve basic properties that are only critical to algorithms. Some studies have been reported on preserving such basic properties. For instance, condensation approach [13] condenses the data into multiple groups, which have at least k records, say, $\{X_1, ..., X_k\}$ and a record X_i contains d dimensions as $(x_i^1, ..., x_i^d)$. Within each condensed group, the vertically partitioned first order sum $\sum_{t=1}^{k} x_t^j$ and second order sum $\sum_{t=1}^{k} \left(x_t^j\right)^2$ are preserved. [14] preserves distances by Fourier transform. It horizontally transforms each record into frequency domain and vertically truncates small coefficients to strike on a trade-off between mining quality and the privacy preservation. Note that from the point of view of the mining quality, it cannot produce exact mining results between private and public databases. From the point of view of mining versatility, it preserves only distances, which is more limited and less satisfactory than both distance and correlation preservations. However, [13] and [14] cannot preserve inner products, and thus cannot preserve correlation. [15] preserves inner products and correlations, but the first order sums of private records are lost. Relations between vertically-partitioned attributes are also lost.

III. Perfect FISIP Transformation

In this Section, we propose a category of data transformation schemes called FISIP, which stands for FIrst and Second order sum and Inner product Preservation. The transformations belonging to this category preserve the first order sum, the second order sum and inner products. Consequently, the distances and the correlations of original data are also preserved after the transformation.

A. Theoretical Properties of FISIP

Let private vectors be $r_i \in R^{k\times 1}$ public vectors be $u_i \in R^{k\times 1}$, and a linear transformation with its matrix representation be $A \in R^{k\times k}$. The transformation between r_i and u_i via A can be written as $u_i = Ar_i$. As such, the distance between u_i and u_j is $(u_i - u_j)^2 = (u_i - u_j)^{\intercal}(u_i - u_j) = (r_i - r_j)^{\intercal}A^{\intercal}A(r_i - r_j)$. If A is orthogonal, $A^{\intercal}A = I$, where I is an identity matrix. $(u_i - u_j)^2 = (r_i - r_j)^{\intercal}(r_i - r_j) = (r_i - r_j)^2$. Therefore, the distance of vectors after the transformation is preserved. However, this cannot guarantee the preservation of correlations between vectors. A FISIP transformation is defined as follows.

Definition 3: FISIP Transformation: The matrix representation of a linear transformation $A \in R^{k\times k}$ can be written as $A = [A_i] = \left[\begin{array}{cccc} A_1 & A_2 & ... & A_k \end{array}\right]$, $A_i \in R^{k\times 1}$. Additionally, A_i can be written as $A_i = [A_{im}]$, $1 \le m \le k$, $A_{im} \in R$. Then, the transformation is called a FISIP transformation if A has following properties.

(1) $\sum\limits_{m=1}^{k} A_{im} = 1.$

(2) $\sum\limits_{m=1}^{k} A_{im}^2 = 1.$

(3) $\sum\limits_{m=1}^{k} A_{im}A_{jm} = 0,\ for\ i \ne j.$

Note that (1) and (3) implies A is an orthogonal matrix. We then can derive the following three lemmas, which can be proved according to the above properties. For the sake of space, we omit the proof of lemmas and theorems.

Lemma 1 (First Order Sum Preservation): For a FISIP transformation $u_i = Ar_i$, $\sum\limits_{m=1}^{k} u_{im} = \sum\limits_{m=1}^{k} r_{im}$.

Lemma 2 (Second Order Sum Preservation): For a FISIP transformation $u_i = Ar_i$, $\sum\limits_{m=1}^{k} u_{im}^2 = \sum\limits_{m=1}^{k} r_{im}^2$.

Lemma 3 (Inner Product Preservation): For a FISIP transformation $u_i = Ar_i$, $\sum\limits_{m=1}^{k} u_{im}u_{jm} = \sum\limits_{m=1}^{k} r_{im}r_{jm}$.

These three lemmas lead to Theorem 1, which states the distance and correlation preserving property of a FISIP transformation.

Theorem 1 (Property of FISIP transformation): A FISIP transformation is both distance and correlation preserving.

B. General Form Realization

Finding a matrix which corresponds to a FISIP transformation is not easy. We first devise a general procedure to construct FISIP matrices and then a fast computation as follows. Transformations proposed in this section can perfectly preserve the distance and correlation of data, and thus we name the transformations perfect FISIP transformation, or simply perfect FISIP. We firstly define a base matrix, which is referred to as a spreading matrix, and prove that it is a form of FISIP matrices.

Definition 4: Spreading matrix: A k-dimensional spreading matrix, denoted by $A^{[k]}$, is defined as a k by k matrix as constructed by the following formula.

Basic type:

$$A^{[k]} = [a_{ij}] = \begin{bmatrix} \frac{2-k}{k} & \frac{2}{k} & \frac{2}{k} & \cdots & \frac{2}{k} \\ \frac{2}{k} & \frac{2-k}{k} & \frac{2}{k} & \cdots & \frac{2}{k} \\ \frac{2}{k} & \frac{2}{k} & \frac{2-k}{k} & \cdots & \frac{2}{k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{2}{k} & \frac{2}{k} & \frac{2}{k} & \cdots & \frac{2-k}{k} \end{bmatrix}, \text{ where}$$

$1 \le i,\ j \le k$, $a_{ii} = \frac{2-k}{k}$, $a_{ij} = \frac{2}{k}$ for $i \ne j$.

For example, spreading matrices constructed as the basic type are

$$A^{[2]} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\ A^{[3]} = \begin{bmatrix} \frac{-1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{-1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & \frac{-1}{3} \end{bmatrix}$$

Theorem 2 (Property of $A^{[k]}$): A linear transformation with a spreading matrix representation $A^{[k]}$ is a FISIP transformation.

Note that there are other methods conceivable to do the transformation. One can construct an high dimensional FISIP matrix directly using the definition formula or use a low dimensional spreading matrix as a building block to construct a high dimensional FISIP matrix. The advantage of using a low dimensional spreading matrix as the base matrix is that the constructed matrix has more zeros, which can reduce the calculation time of public vectors. Next, we propose a procedure to construct FISIP matrix more efficiently using basic type of spreading matrices.

Definition 5: Derived spreading matrix: A k by k matrix constructed by the following formulas.

Composition type:

$$A_c^{[k]} = \begin{bmatrix} A^{[k_1]} & 0 & \cdots & 0 \\ 0 & A^{[k_2]} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A^{[k_n]} \end{bmatrix}, \text{ where } k = \sum_{i=1}^{n} k_i.$$

Additionally, $A^{[k]}$ can be any row permutation and/or column permutation of the basic type spreading matrix or composition type matrix. For example,

$$A^{[k]} = P_r \begin{bmatrix} \frac{-1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{-1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{-1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{-1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{-1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{-1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{-1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{-1}{2} \end{bmatrix} P_c$$

$$= \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{-1}{2} & 0 & \frac{1}{2} \\ \frac{-1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{-1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{-1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{-1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{-1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{-1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{-1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}, \text{ where}$$

$$P_r = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \text{ and}$$

$$P_c = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

P_r and P_c are the row and column permutation matrices chosen randomly by private data publisher. ■

In general, for $u = A^{[k]}r$ and $u,\ r \in R^{k\times 1}$, each element of u needs k multiplications and $k-1$ additions, i.e., $O(k \times M + (k-1) \times A)$, where M denotes the computation of multiplication and A denotes the addition. To obtain the vector of u, the computation needs $O(k^2 \times M + k \times (k-1) \times A)$. However, each element of u does not necessarily need k multiplications and $k-1$ additions. The computation for each element of u can be reduced to c multiplications and $c-1$ additions, where c is a constant. Thus, the computation of obtaining u can be reduced to $O(k \times c \times M + k \times (c-1) \times A) = O(k \times M + k \times A)$. If the database contains n private records, the computation is $O(n \times k \times M + n \times k \times A)$.

C. Protection Level

We define the protection level as the number of possible variations of private vectors when public vectors are given. Given $[u_i] = P_r A^{[k]} P_c\,[r_i]$, $[r_i] = P_c^{-1}(A^{[k]})^{-1}P_r^{-1}\,[u_i] = P_r(A^{[k]})^T P_c[u_i]$. The possible variations of the transformation matrix, $P_r(A^{[k]})^T P_c$, increases factorially as k increases with a lower bound of $k!$. Therefore, finding real private vectors from possible private vectors is infeasible if a proper number, k, is chosen.

IV. Strong FISIP Transformation

A. Privacy Enhancement via Matrix Perturbation

Assume an attacker has k linearly independent k-dimensional vectors, r_1 to r_k, and their transformed public vectors, u_1 to u_k, at hand. The attacker wants to find the corresponding unknown private vector r_x by the public vector u_x. If the attacker can find a vector $B \in R^{k\times 1}$ such that $u_x = [u_1u_2...u_k]B$, r_x can be computed by $r_x = A^{-1}u_x = A^{-1}[u_1u_2...u_k]B = A^{-1}A[r_1r_2...r_k]B = [r_1r_2...r_k]B$. Unfortunately, such vector B can be easily obtained by the linear combination method. This attack is called known private data attack.

Special protections can be implemented to prevent attackers from doing reverse transform. Defensive manipulations to avoid such kind of attack can be done if we are willing to give up some mining accuracy. If each record is originally multiplied by different A, then corresponding records will not be calculated easily. We propose a data perturbation method to generate different transformation matrices. Since the accuracy of mining results on the transformed data is not exactly the same as the results on the original data. We call this kind of transformation strong FISIP transformation.

Special perturbation can be done on A to preserve first order sum by making the sum of each column's perturbation equal to zero. For example, we denote the perturbed A as A'. For each column i of A, we randomly select a row j and set element of A' as $a'_{ji} = a_{ji} - 2^{pert}$, and for other row $k \neq j$, $a'_{ki} = a_{ki} + \frac{2^{pert}}{d-1}$, where $pert \leq 1$, and d is the total number of rows. As such, the sum of each column of A' is

$$a'_{ji} + \sum_{k\neq j} a'_{ki} = \left(a_{ji} - 2^{pert}\right) + \sum_{k\neq j}\left(a_{ki} + \frac{2^{pert}}{d-1}\right)$$

$$= a_{ji} + \sum_{k\neq j} a_{ki} - 2^{pert} + (d-1) \times \frac{2^{pert}}{d-1} = 1.$$

Therefore, the first order sums of transformed vectors still equal to the sums of private vectors. In this way, attackers are not able to use inverse transform to obtain original private vectors. The estimation error of attackers is analyzed as follows. We assume that $u_x = A_x r_x$, $u_1 = A_1 r_1, \cdots,$ $u_k = A_k r_k$, and b_i represents each element B. Then, the error of attacker's estimated private vector r_{est} is

$$|r_{est} - r_x| = \left|\sum_{i=1}^{k} b_i r_i - A_x^{-1}u_x\right|$$

$$= \left|\sum_{i=1}^{k} b_i r_i - A_x^{-1}(\sum_{i=1}^{k} b_i u_i)\right|$$

$$= \left|\sum_{i=1}^{k} b_i r_i - A_x^{-1}(\sum_{i=1}^{k} b_i A_i r_i)\right| = \left|\sum_{i=1}^{k} b_i (I_k - A_x^{-1}A_i) r_i\right|$$

If each record is transformed by the same A, i.e., $A_x = A_i$, the error is zero. Otherwise, the estimation error will increase as the dimension k increases theoretically.

By using differently perturbed FISIP matrices, we can transform each record in difference way and thus we can prevent attackers from the reverse transform. We will show the effects of the perturbation in the experiments.

V. Experimental Results

Since the perfect FISIP transformations preserve distances and correlations perfectly, we do not show the mining result comparisons before and after the perfect FISIP transformations. In this section, we firstly evaluate data usefulness after strong FISIP transformations. We measure the preservation of basic properties of FISIP under perturbation. Next, data robustness against known data attack is evaluated. In the experiments, we use three real datasets, Iris, Pendigits, and Satlog, from UCI Machine Learning Repository [16]. We appreciate UCI web site for providing testing datasets.

Databases	Iris	Pendigits	Satlog
Number of attributes	4	16	36
Number of records	150	7,494	4,435

A. Strong FISIP Preservation

Let us see how data change by using the first two records in Iris database. The first two private records are

$$r = \begin{bmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3.0 & 1.4 & 0.2 \end{bmatrix},$$

$$u = Ar = \begin{bmatrix} 0.00 & 1.60 & 3.70 & 4.90 \\ -0.15 & 1.75 & 3.35 & 4.55 \end{bmatrix}$$

With the perturbation, $pert = -4$, they are changed to

$$u' = A'r = \begin{bmatrix} 0.20 & 1.81 & 3.80 & 4.40 \\ -0.38 & 1.83 & 3.55 & 4.50 \end{bmatrix}.$$

The distances of r, u and u' are 0.5385, 0.5385, and 0.6398. The correlations of them are 0.9960, 0.9960, and 0.9946. Note that the errors are marginal while the perturbation is introduced.

The first order sums are preserved theoretically and we show the comparisons of second order sums and inner products in Figure 1 and Figure 2. Figure 1 is the percentage of the average of difference while Figure 2 shows the percentage of the standard deviation of difference. The differences of second order sums are defined as $\left|\sum_{i=1}^{d} r_i^2 - \sum_{i=1}^{d} u_i'^2\right| / \sum_{i=1}^{d} r_i^2$ and the differences of inner products are $\left|\mathbf{r}_i \cdot \mathbf{r}_j - \mathbf{u}_i' \cdot \mathbf{u}_j'\right| / \left|\mathbf{r}_i \cdot \mathbf{r}_j\right|$, where $\mathbf{r} = [r_i]$ and $\mathbf{u}' = [u_i']$, $1 \leq i, j \leq$ number of total records. As shown in Figure 1 and Figure 2, we can confirm that the perturbation does not deteriorate our preservation of strong FISIP transformation very much. For the correlation preservation, we randomly select 10,000 pairs of vectors in three databases and measure the differences between private pairs and public pairs, i.e., $\left|corr(\mathbf{r}_i, \mathbf{r}_j) - corr(\mathbf{u}_i, \mathbf{u}_j)\right|$. Figure 3 shows the average of difference and Figure 4 shows the standard deviation of difference. As shown in Figure 3,

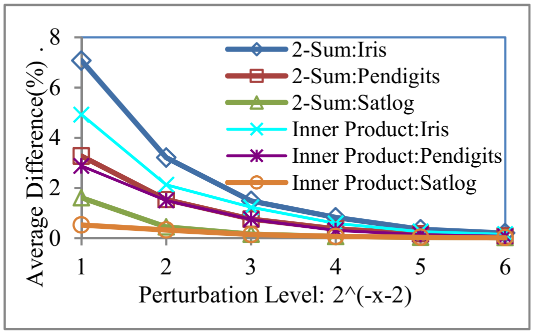

Figure 1. Average difference of distance preservation

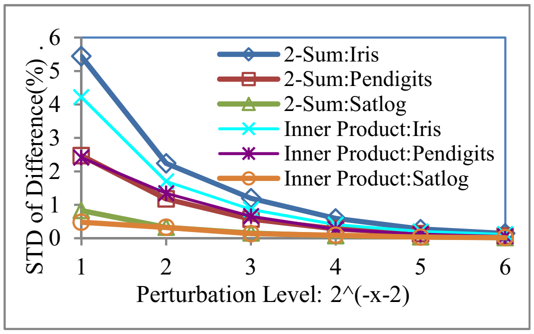

Figure 2. Standard deviation of distance preservation

when the perturbation level decreases, the average difference decreases accordingly and the maximum difference is less than 0.3. Therefore, strong FISIP maintains the relations of correlations very well.

B. Protection against Known Data Attack

The next experiment illustrates how well the protection scheme behaves if different perturbed transformation matrix are adopted. For k-dimensional database, we assume that the attacker has k private-public pairs at hand and uses the techniques we described in previous section to estimate the remaining unknown private records. As stated in the previous section, the estimation error is $\left|\sum_{i=1}^{k} b_i (I_k - A_x^{-1} A_i) r_i\right|$.

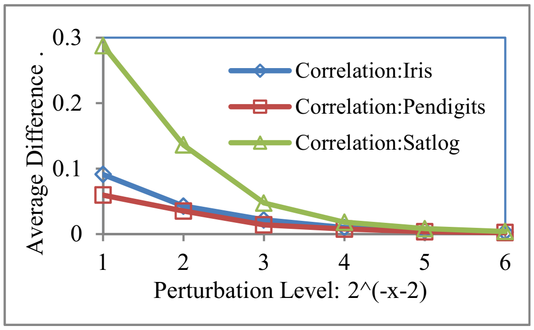

Figure 3. Average difference of correlation preservation

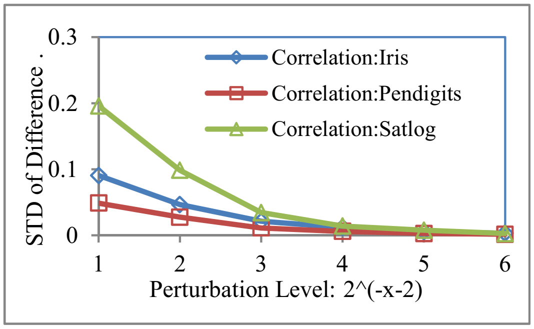

Figure 4. Standard deviation of correlation preservation

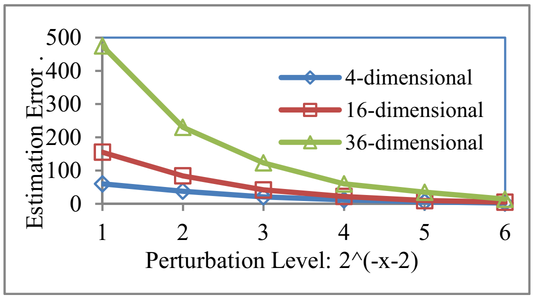

Figure 5. Attacker's Estimation Error

We illustrate the example estimation error by assuming every $b_i = 1$ and $r_i = \left(\frac{1}{k}, \cdots, \frac{1}{k}\right)$. Figure 5 shows that the norm of the estimation error increases as the dimension increases and as the perturbation increases.

VI. Conclusion

Most conventional transformation schemes suffer from the algorithm dependency and the information loss limits. In this work, we proposed a scheme that transforms the private data to the public ones without breaking the relation of distances and correlations between vectors. The mining quality from the transformed public data remain the same as that from the original private data. The privacy of private data can be further enhanced at the expense of slight mining quality by using perturbation of the transformation matrices. The number of different transformations grows in a factorial way as the dimension of data increases, and thus makes attackers hard to recover them back to private counterparts.

Acknowledgment

The authors would like to thank the supports from National Science Council under the projects NSC99-2221-E-155-074, NSC100-2221-E-155-088, and NSC100-2632-S-155-001.

References

[1] D. Lin, E. Bertino, R. Cheng, and S. Prabhakar, "Position transformation: a location privacy protection method for moving objects," *Proc. of Int'l Workshop on Security and Privacy in GIS and LBS*, pp. 62–71, 2008.

[2] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. of Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proc. of Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

[4] E. Achtert, C. Bohm, H.-P. Kriegel, P. Kroger, and A. Zimek, "Robust, complete, and efficient correlation clustering." *Proc. of SIAM Int'l Conf. on Data Mining*, 2007.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[6] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[7] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," *Proc. of Int'l Conf. on Machine Learning*, pp. 359–366, 2000.

[8] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-based detection and prediction of outliers," *IEEE Trans. on Knowl. and Data Eng.*, vol. 18, no. 2, pp. 145–160, 2006.

[9] J. T. li Wang, X. Wang, K. ip Lin, D. Shasha, B. A. Shapiro, and K. Zhang, "Evaluating a class of distance-mapping algorithms for data mining and clustering," *Proc. of Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 307–311, 1999.

[10] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int'l Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

[11] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and l-diversity," *Proc. of Int'l Conf. on Data Engineering*, 2007.

[12] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *Proc. of SIGMOD Int'l Conf. on Management of Data*, vol. 1, no. 33, 2004.

[13] C. Aggarwal, C. C. Aggarwal, and P. S. Yu, "A condensation approach to privacy preserving data mining," *Proc. of Int'l Conf. on Extending Database Technology*, pp. 183–199, 2004.

[14] S. Mukherjee, Z. Chen, and A. Gangopadhyay, "A privacy-preserving technique for euclidean distance-based mining algorithms using fourier-related transforms," *Very Large Data Bases Journal*, vol. 15, no. 4, pp. 293–315, 2006.

[15] K. Liu and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Trans. on Knowl. and Data Eng.*, vol. 18, no. 1, pp. 92–106, 2006.

[16] A. Frank and A. Asuncion, "Uci machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml