

Quantification of a Privacy Preserving Data Mining Transformation

Mohammed Ketel
School of Information Technology
University of Baltimore
1420 North Charles Street
Baltimore, MD 21201
mketel@ubalt.edu

Abstract—Data mining, with its promise to extract valuable, previously unknown and potentially useful patterns or knowledge from large data sets that contain private information is vulnerable to misuse. To protect the private or sensitive information, many privacy-preserving data mining (PPDM) techniques have emerged. A large fraction of these techniques use randomized data distortion by adding noise from a known distribution function (e.g., uniform, normal) to the sensitive data. However, non-careful noise addition may introduce biases to the statistical parameters of these data. To preserve the statistical properties and meet privacy requirements of the sensitive data, we use a data transformation technique called Rotation-Based Transformation (RBT). This method distorts only private numerical attributes and preserves the statistical properties of the data.

Keywords- Data mining, Privacy, Data transformation.

I. INTRODUCTION

Advances in database technologies and computer networking have enabled the collection and storage of vast quantities of data. The attempt to extract valuable knowledge and trends from this huge amount of data has led to the challenging field of data mining. In many contexts, data are distributed across different sites. And organizations have realized that they can often obtain better results by pooling their data together in the same data warehouse. Often organizations are extremely dependent on data mining in their every day activities and paybacks include better decision-making, providing better service, and achieving greater profit. However, the collected data may contain sensitive or private information about the organizations or their customers, and privacy concerns becomes a serious issue if data is shared between multiple organizations.

Data mining in such privacy-sensitive domains is facing growing concerns. Therefore, we need to develop data mining techniques that are sensitive to the privacy issue. This has led the development of a class of data mining

algorithms [3, 4, 6, 7] that try to extract the data patterns without directly accessing the original data and guarantees that the mining process does not get sufficient information to reconstruct the original data. This paper considers a data technique namely Rotation-Based Transformation (RBT) introduced by Oliveira and Zaiane [1, 2]. This method uses the RBT on the sensitive data attributes while preserving their underlying statistical properties.

An obvious step to protect the privacy of the individuals (or entities) is to replace any explicitly identifying information by some randomized placeholder. For example, a randomized token could replace the uniquely identifying social security number of a person. However, this is not sufficient since the released data contains other information which, when linked with other data sets, can identify or narrow down the individuals or entities [8, 9]. An example in [9] illustrates the identification by linking a medical data set using fields like zip code, date of birth and gender. In addition to the identity disclosure problem discussed above, attribute disclosure occurs when something about an individual is learnt from the released data. Attribute disclosure can happen even without identity disclosure. Also, attribute disclosure in the broad sense can include inferential disclosure in which some characteristic of the individual can be inferred more accurately because of the data release [8]. Attributes whose disclosure needs to be protected in the strictest sense are denoted to be sensitive. One approach to solving the identity disclosure problem is to randomize the sensitive data (using techniques like adding noise) [3, 4]. Also, privacy-preserving data mining using randomization, introduced in [3], arose as a solution by allowing parties to cooperate in the extraction of knowledge without any of the cooperating parties having to reveal their individual data items to each other or any other parties. A review of randomization approach is found in section II. Another basic approach to modify the sensitive attributes is by using the value-class membership method. In this approach, the values for an attribute are partitioned into a set of disjoint, mutually-exclusive classes. Instead of a true attribute value, the user provides the interval in which the value lies. Data discretization is

the method used most often for hiding individual values. In this paper, we consider the approach of transforming the data using a geometric transformation (rotation) to satisfy privacy.

II. RANDOMIZATION APPROACH

In this section, we review the randomization technique that has been proposed in [3, 4]. Based on this approach, the entire privacy-preserving process can be considered in two steps. The first step is for data users/clients to randomize their data, and transmit the perturbed data to the data miner. Several randomization methods have been proposed including the random perturbation operator [3]. For the concept of random perturbation to be useful, we need to be able to reconstruct the original data distribution from the randomized data. In the second step, the data miner employs a distribution reconstruction algorithm that intends to reconstruct the original data distribution. Several distribution reconstruction algorithms have been proposed [3, 4, 12]. For example, the expectation maximization (EM) algorithm [4] reconstructs a distribution which converges to the maximum likelihood estimate of the original distribution.

The problem can be formulated as follow: Consider a set of n original numerical data values (attributes) $x_i, i = 1, 2, \dots, n$, each considered as an instance of a random variable X_i . Assume that all X_i are independent and identically distributed with probability density function (pdf) denoted by f_X . Each user/client randomizes its value $x_i, i = 1, 2, \dots, n$ by adding to it a random perturbation $y_i, i = 1, 2, \dots, n$. The random values y_i are independent and identically distributed with pdf f_Y . The pdf f_Y is chosen in advance and is known to the data miner. Each provider sends a randomized value $z_i = x_i + y_i$ to the data miner. Given f_Y and the perturbed values $z_i, i = 1, 2, \dots, n$, the goal is to estimate the function f_X . Also, it is necessary to understand how to choose f_Y so that:

- The data miner can approximate f_X reasonably well. Hence, we assume that enough data is available to make the statistical approximations and have the computing facilities required for data processing.
- The perturbed value z_i does not disclose too much about x_i .

The details of EM reconstruction algorithm for f_X and its convergence properties can be found in [4]. The algorithm proposed in [4] for the reconstruction of $f_X(x)$ is as follow:

Assume that the data domain, D_X , of $f_X(x)$ is partitioned into L fixed non-overlapping intervals D_1, D_2, \dots, D_L such that $D_X = \bigcup_{l=1}^{l=L} D_l$ and the length of each D_l is m_l . Assume that $f_X(x)$ is constant over each interval D_l with a constant value $\theta_l (l = 1, 2, \dots, L)$. The estimated pdf is a piecewise constant function of the form $\sum_{l=1}^{l=L} \theta_l I_{D_l}(x)$ where $I_{D_l}(x)$ is an indicator function ($I_{D_l}(x) = 1$ if $x \in D_l$ and 0 otherwise). The parameters $\theta_l (l = 1, 2, \dots, L)$ are computed via the iterative reconstruction algorithm:

1. Initialize $\theta_l^0 = \frac{1}{L}, l = 1, 2, \dots, L$
2. Update

$$\theta_l^{(j+1)} = \frac{\theta_l^j}{m_l n} \sum_{i=1}^{i=n} \frac{\Pr(Y \in z_i - D_l)}{\sum_{l=1}^{l=L} \theta_l^j \Pr(Y \in z_i - D_l)}$$
3. $j = j + 1$
4. Return to step 2 if the termination criterion is not met.

While the randomization approach is intuitive, recently some researchers have identified privacy breaches as one of the major problems with this technique. It has been shown [13] that the spectral properties of randomized data could help the data miner to separate noise from the private data. In particular, a filtering method has been proposed to reconstruct private data from the randomized data set [13].

III. DATA MATRIX

In this paper, the data is assumed to be a matrix (table) D_{mn} where each of the m rows is an observation, $O_i (i = 1, 2, \dots, m)$ and each observation contains values for each of the n attributes, $A_j (j = 1, 2, \dots, n)$.

$$D_{mn} = \begin{bmatrix} a_{11} & \cdot & \cdot & \cdot & a_{1n} \\ a_{21} & \cdot & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & a_{ij} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{m1} & \cdot & \cdot & \cdot & a_{mn} \end{bmatrix} \quad (1)$$

The matrix D_{mn} may contain categorical and numerical attributes. RBT rely on d numerical sensitive attributes A_j ($j= 1, 2, \dots, d$). Thus, the new D_{md} matrix, which is subject to transformation has m observations and d attributes [1, 2]. The attributes in the data matrix are sometimes normalized before being used. There are many methods for data normalization, see for instance reference [5]. Min-max normalization performs a linear transformation on the original data. Each attribute is normalized by scaling its values so that they fall within a small specific range, such as 0.0 and 1.0. Min-max normalization maps a value v of an attribute A to v' as follows:

$$v' = \frac{v - \min A}{\max A - \min A} * (\text{new_max } A - \text{new_min } A) + \text{new_min } A \quad (2)$$

where $\min A$ and $\max A$ represent the minimum and maximum values of an attribute A , respectively, while $\text{new_min } A$ and $\text{new_max } A$ are the new range in which the normalized data will fall.

When the actual minimum and maximum of an attribute are unknown, or when there are outliers that dominate the min-max normalization, z-score normalization (also called zero-mean normalization) should be used. In z-score normalization, the values for an attribute A are normalized based on the mean and the standard deviation of A . A value v is mapped to v' as follows:

$$v' = \frac{v - \bar{A}}{\sigma_A} \quad (3)$$

where \bar{A} and σ_A are the mean and the standard deviation of the attribute A , respectively.

IV. ROTATION-BASED TRANSFORMATION METHOD

In this work, the focus is primarily on rotations. For the sake of simplicity, the basics of such a transformation are described in a 2-D discrete space [1, 2]. In its simplest form, this transformation is for the rotation of a point about the coordinate axes. Rotation of a point in a 2-D discrete space by an angle θ is achieved by using the transformation matrix in Equation (4). The rotation angle

θ is measured clockwise and this transformation affects the values of X and Y coordinates. Thus, the rotation of a point in a 2-D discrete space could be seen as a matrix representation $v' = Rv$, where R is a 2×2 rotation matrix, v is the vector column containing the original coordinates, and v' is a column vector whose coordinates are the rotated coordinates.

$$R(\theta) = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \quad (4)$$

R is applied to the observations of the confidential attributes in a pair wise manner to preserve privacy. RBT may be applied more than once to some confidential attributes.

Example:

Let's consider a sample from a relational database in Table 1. Assume that the attributes age and salary are confidential. The corresponding normalized and distorted database for confidential data (age, salary, $\theta = 13.7$) are shown in Table 2 and Table 3 respectively.

Table 1. Sample relational database

O#	Occupation	Age	Sex	Salary
1	engineer	36	M	72,000
2	cashier	21	F	19,000
3	technician	32	M	36,000
4	teacher	45	F	38,000
5	nurse	33	F	42,000
6	teller	41	M	35,000

Table 2. Corresponding Normalized Database

O #	Occupation	Age	Salary
1	Engineer	0.1757	1.9939
2	Cashier	-1.8014	-1.3433
3	Technician	-0.3515	-0.2729
4	Teacher	1.3621	-0.1469
5	Nurse	-0.2197	0.1049
6	Teller	0.8348	-0.3358

Table 3. Rotation Data Perturbation

O #	Occupation	Age	Salary
1	Engineer	1.8808	0.6850
2	Cashier	-1.9796	1.0633
3	technician	-0.3961	0.2029
4	teacher	0.4436	-1.2962
5	nurse	0.0020	0.2435
6	teller	0.0492	-0.8985

Note that $R(0) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ corresponds to the unity matrix,

and therefore there is no data perturbation.

V. PRIVACY QUANTIFICATION

The rotation based transformation transform the numerical confidential attributes x and y into $z_{1\theta} = x \cos \theta + y \sin \theta$ (correspondent to attribute x after RBT) and $z_{2\theta} = -x \sin \theta + y \cos \theta$ (correspondent to attribute y after RBT) respectively. Assume in the following analysis that x and y are samples from normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ which is realistic assumption in many practical situations. Due to the rotational symmetry of the joint distribution, it is clear that without calculation that the probability distribution (density) function of $z_{1\theta}$ ($z_{2\theta}$) must be the same as x and y [10]. Let's concentrate on only $z_{1\theta} = x \cos \theta + y \sin \theta$ similar results can be derived for $z_{2\theta}$. In the following analysis the subscripts 1 and θ in $z_{1\theta}$ and use only $z = x \cos \theta + y \sin \theta$ which is basically the attribute x after the RBT transformation. Recall that X has normal (μ, σ^2) distribution, then $(X - \mu)/\sigma$ has normal $(0,1)$ distribution.

$$U = (X - \mu)/\sigma \text{ is } N(0,1)$$

In the following notation, let the capital letters (X, Y, Z) be the normalized random variables corresponding to (x, y, z) and assume that X and Y are independent.

Then for the random variable $Z = X \cos \theta + Y \sin \theta$ we have the following facts.

By rotational symmetry of the joint distribution of X and Y , the distribution of Z is standard normal. Thus

$$\begin{aligned} E(X) &= E(Y) = E(Z) = 0 \\ SD(X) &= SD(Y) = SD(Z) = 1 \\ \rho(X, Z) &= E(XZ) = E[X(X \cos \theta + Y \sin \theta)] = \cos \theta \end{aligned}$$

Since $E(X^2) = 1$, and $E(XY) = E(X)E(Y) = 0$ by independence of X and Y . To summarize, X and Z are standard normal variables with correlation $\rho = \cos \theta$. Note the special cases:

$$\begin{aligned} \theta = 0, \rho = 1 \text{ and } Z &= X, \\ \theta = \Pi/2, \rho = 0 \text{ and } Z &\text{ is independent of } X \\ \theta = \Pi, \rho = -1 \text{ and } Z &= -X \end{aligned}$$

In this paper, we use privacy measure proposed in [4]. This measure is based on the differential entropy of a random variable. The differential entropy $h(X)$ of a random variable X is defined as follows:

$$h(X) = E(-\log_2 f_X(x))$$

$$h(X) = -\int_{\Omega_X} f_X(x) \log_2 f_X(x) dx \quad (5)$$

where Ω_X is the domain of X . It is well-known that $h(X)$ is a measure of uncertainty inherent in the value of X [11]. The entropy is always positive since the logarithm of probabilities is always negative. It has been shown that among all the laws of probability distributions for a continuous random variable with a given mean and variance, the normal law has the maximum entropy. It can be shown that the entropy for a normal variable X :

$$\begin{aligned} x &\sim \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ h(x) &= \log_2(\sqrt{2\pi e\sigma^2}) \quad (6) \end{aligned}$$

Let $2^{h(X)}$ be the measure of privacy inherent in the random variable X and denote it by $\Pi(X)$ [4]. Thus, the normal random variable has privacy:

$$\Pi(x) = \sqrt{2\pi e\sigma^2} \quad (7)$$

Now, we will introduce the notion of conditional privacy which takes into account the additional information available in the perturbed values. Given a random variable Y , the conditional differential entropy of X is defined as follows:

$$h(X | Z) = -\int_{\Omega_{X,Z}} f_{X,Z}(x, z) \log_2 f_{X|Z=z}(x) dx dz \quad (8)$$

Let $P(X/Z)$ be the metric for the conditional loss of X given Z ,

$$P(X|Z) = 1 - \Pi(X|Z)/\Pi(X) = 1 - 2^{h(X|Z)}/2^{h(X)} = 1 - 2^{-I(X;Z)} \quad (9)$$

$I(X; Z)$ is known as the mutual information between the random variables X and Z where

$$I(X; Z) = h(X) - h(X/Z) = h(Z) - h(Z/X).$$

For the normalized random variable $Z = X \cos \theta + Y \sin \theta$ we have the following: the privacy for the normalized random variable X before the RBT transformation is

$\Pi(X) = \sqrt{2\pi e}$ and it can be shown after some mathematical manipulations that $I(X; Z)$ is

$$I(X, Z) = \log_2 \frac{1}{\sqrt{1-\rho^2}} \quad (10)$$

Thus the fraction of privacy loss in this case is

$$P(X | Z) = 1 - 2^{-I(X;Z)} = 1 - \sqrt{1 - \rho^2} \quad (11)$$

After revealing Z, X has privacy $\Pi(X|Z) = \Pi(X)[1 - P(X|Z)]$

$$\Pi(X|Z) = \sqrt{2\pi e(1 - \rho^2)} \quad (12)$$

Where $\rho(X, Z) = E(X, Z) = E[X(X \cos\theta + Y \sin\theta)] = E(X^2) \cos\theta + E(XY) \sin\theta = \cos\theta$ as found before. Therefore this privacy is θ dependent. For $\theta = \pi/4$, $P(X | Z) = 0.293$

And $\Pi(X|Z) = \sqrt{\pi e} = 2.921$.

Figure 1 and 2 show the privacy loss and privacy versus the correlation ρ respectively. As shown, for low correlation, we have complete good privacy. For $\rho = 1$ ($\theta = 0$), we have no privacy ($Z = X$).

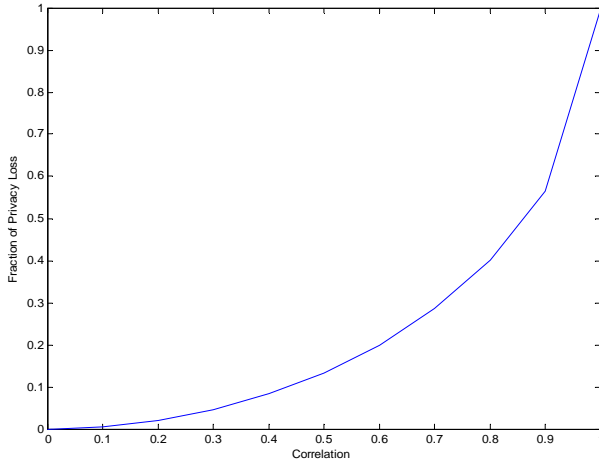


Figure 1. Privacy Loss vs. Correlation ρ

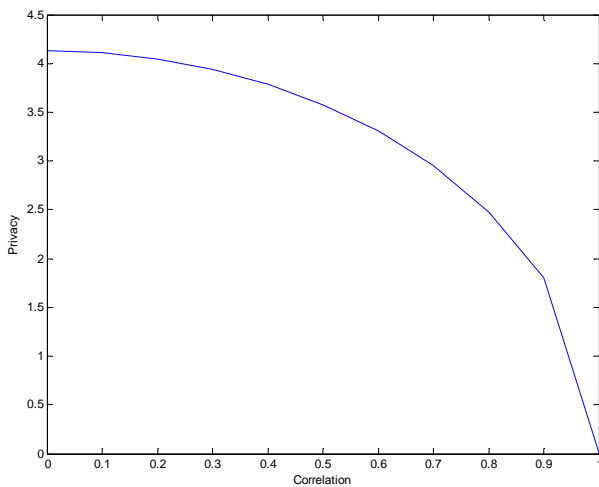


Figure 2. Privacy vs. Correlation ρ

VI. CONCLUSION

The primary objective of PPDM is to have a balance between the extraction of interesting patterns or knowledge from a huge amount of data and the responsibility to protect the privacy of certain attributes of these data. A common approach to handle these situations is to release the perturbed or transformed data. In this paper we considered a data transformation technique namely a rotation-based transformation (RBT) for privacy-preserving. This rotational transformation affects only the confidential attributes while preserving their underlying statistical properties. Privacy quantification where derived for the normalized random variables.

VII. REFERENCES

- [1] Stanley Oliveira and Osmar R. Zaiane, "Privacy preserving clustering by data transformation," in Proc. of the 18th Brazilian Symposium on Databases (SBB D 2003), pp. 304-318, Manaus, Brazil, 6-8 October 2003.
- [2] Stanley Oliveira and Osmar R. Zaiane, "Achieving Privacy Preservation When Sharing Data for Clustering," in VLDB'2004, Springer Verlag LNCS 3178, pp. 67-82, Toronto, Canada, August 30, 2004.
- [3] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in Proceeding of the ACM SIGMOD Conference on Management of Data, pp. 439-450, Dallas, Texas, May 2000. ACM Press.
- [4] D. Agrawal and C. C. Aggawal, "On the design and quantification of privacy preserving data mining algorithms," in Proceedings of the 20th ACM SIMOD Symposium on Principles of Database Systems, pp. 247-255, Santa Barbara, May 2001.
- [5] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, CA, 2001.
- [6] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," in SIGMOD Workshop on DMKD, Madison, WI, June 2002.
- [7] A. Evfimevski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in Proceedings of the ACM SIMOD/PODS Conference, San Diego, CA, June 2003.
- [8] P. Samarati, "Protecting respondents' identities in micro-data release," IEEE Transactions on Knowledge Engineering, 13(6), pp. 1010-1027, 2001.

- [9] L. Sweeney, Datafly, "A system for providing anonymity in medical data," in Proceedings of Eleventh International Conference on Database Security, pp. 356-381. Database Security, Status and Prospects, 1998.
- [10] A. Papoulis, Probability, random variables and stochastic processes, Mc Graw Hill Publishers, 1984.
- [11] R. Hamming, Coding and information theory, Prentice Hall, 1986.
- [12] W. Du and Z. Zhan, "Using randomized response techniques for privacy-preserving data mining," in Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 505–510. ACM Press, 2003.
- [13] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in Proceedings of the 3rd IEEE International Conference on Data Mining, pp. 99–106. IEEE Press, 2003.