Privacy Preserving Data Classification with Rotation Perturbation

Keke Chen Ling Liu
College of Computing, Georgia Institute of Technology
{kekechen, lingliu}@cc.gatech.edu

1 Introduction

Data perturbation techniques are one of the most popular models for privacy preserving data mining [3, 1]. It is especially convenient for applications where the data owners need to export/publish the privacy-sensitive data. A data perturbation procedure can be simply described as follows. Before the data owner publishes the data, they *randomly* change the data in certain way to disguise the sensitive information while preserving the particular data property that is critical for building the data models. Several perturbation techniques have been proposed recently, among which the most typical ones are randomization approach [3] and condensation approach [1].

Loss of Privacy vs. Loss of Information.
Perturbation techniques are often evaluated with two basic metrics, loss of privacy and loss of information (resulting in loss of accuracy for data classification). An ideal data perturbation algorithm should aim at minimizing both privacy loss and information loss. However, the two metrics are not well-balanced in many existing perturbation techniques [3, 2, 5, 1].

Loss of privacy can be intuitively described as the difficulty level in estimating the original value from the perturbed data. In [3], the variance of the added random noise is used as the level of difficulty for estimating the original values. However, later research [5, 2] reveals that variance is not an effective indicator for randomization approach since the original data distribution has to be known. In addition, paper [8] shows that the loss of privacy is also subject to the special attacks that can reconstruct the original data from the perturbed data.

The loss of information typically refers to the amount of critical information preserved about the datasets after the perturbation. Different data mining tasks, such as classification and association mining, typically utilize different set of properties of a dataset. Thus, the information that is considered critical to data classification may differ from those critical to association rule mining. We argue that the exact information that need to be preserved after perturbation should be "task-specific". Since most classification models typically concern multi-dimensional properties, perturbation techniques for data classification should perturb multiple columns together. To our knowledge, very few perturbation-based privacy protection proposals so far have considered *multi-dimensional perturbation techniques*.

Contribution and Scope of the paper.
Bearing these issues in mind, we have developed a random rotation perturbation approach to privacy preserving data classification. In contrast to other existing privacy preserving classification methods [1, 3, 9], our approach exploits the task-specific information about the datasets to be classified, aiming at producing a robust data perturbation that exhibits a better balance between loss of privacy and loss of information, without performance penalty.

Concretely, we observe that the multi-dimensional geometric properties of datasets are the critical "task-specific information" for many classification algorithms. One intuitive way to preserve the multi-dimensional geometric properties is to perturb the original dataset through rotation transformation. We have identified and proved that kernel methods, SVM classifiers with the three popular kernels, and the hyperplane-based classifiers, are the three categories of classifiers that are rotation-invariant.

Another important challenge for the rotation perturbation approach is the privacy loss measurement (the level of uncertainty) and privacy assurance (the resilience of the rotation transformation against unauthorized disclosure). Given that a random rotation based perturbation is a multi-dimensional perturbation, the privacy guarantee of the multiple dimensions (attributes) should be evaluated collectively to ensure the privacy of all columns involved and the privacy of the multi-column correlations. We design a unified privacy model to tackle the problem of privacy evaluation for multi-dimensional perturbation, which addresses three types of possible attacks: direct estimation, approximate reconstruction, and distribution-based inference attacks. With the unified privacy metric, we present the privacy assurance of the random rotation perturbation as an optimization problem: given that all rotation transformations result in zero-loss of accuracy for the discussed classifiers, we want to pick one rotation matrix that provides higher privacy guarantee and stronger resilience against the three types of inference attacks.

2 Rotation and Classifiers

In this section, we first describe rotation transformation and the set of geometric properties of the datasets significant to most classifiers, and then we define rotation-invariant classifiers.

Notations for Datasets Training dataset is the part of data that has to be exported/published in privacy-preserving data classification. A classifier learns the classification model from the training data and then is applied to classify the unclassified data. Suppose that X is a training dataset consisting of N data rows (records) and d columns (attributes). For the convenience of mathematical manipulation, we use $X_{d\times N}$ to notate the dataset, i.e., $X = [\mathbf{x}_1 \dots \mathbf{x}_N]$, where $\mathbf{x}_i$ is a data tuple, representing a vector in the real space $\mathbb{R}^d$. Each data tuple belongs to a predefined class, which is determined by its class label attribute y_i. The class labels can be nominal (or continuous for regression). The class label attribute of the data tuple is public, i.e., privacy-insensitive. All other attributes containing private information needs to be protected.

Properties of Geometric Rotation Let $R_{d\times d}$ represent the rotation matrix. Geometric rotation of the data X is generally notated as a function $g(X)$, $g(X) = RX$. Note that the transformation will not change the class label of data tuples, i.e., $R\mathbf{x}_i$ still has the label y_i.

A rotation matrix $R_{d\times d}$ is defined as a matrix having the follows properties. Let R^T represent the transpose of the matrix R, r_{ij} represent the (i, j) element of R, and I be the identity matrix. Both the rows and the columns of R are *orthonormal*, i.e., for any column j, $\sum_{i=1}^{d} r_{ij}^2 = 1$, and for any two columns j and k, $\sum_{i=1}^{d} r_{ij}r_{ik} = 0$. The similar property is held for rows. The definition infers that $R^TR = RR^T = I$. It also implies that by changing the order of the rows or columns of rotation matrix, the resulting matrix is still a rotation matrix. A key feature of rotation transformation is preserving length. It follows that rotation also preserves inner product and Euclidean distance between any pair of points. In general, rotation preserves the geometric shapes such as hyperplane and hyper curved surface in the multidimensional space.

Rotation-invariant Classifiers We can treat the classification problem as function approximation problem – the classifiers are the functions learned from the training data. Therefore, we can use functions to represent the classifiers. Let $\hat{f}_X$ represent a classifier $\hat{f}$ trained with dataset X and $\hat{f}_X(Y)$ be the classification result on dataset Y. Let $T(X)$ be any transformation function, which transforms the dataset X to another dataset X'. We use $Err(\hat{f}_X(Y))$ to notate the error rate of classifier $\hat{f}_X$ on testing data Y and let ε be some small real number, $|\varepsilon| < 1$.

Definition 1. *A classifier $\hat{f}$ is invariant to some transformation T if and only if $Err(\hat{f}_X(Y)) = Err(\hat{f}_{T(X)}(T(Y))) + \varepsilon$ for any training dataset X and testing dataset Y.*

It follows that the strict condition $\hat{f}_X(Y) \equiv \hat{f}_{T(X)}(T(Y))$ trivially guarantees the invariance property. If a classifier $\hat{f}$ is invariant to *rotation* transformation, we specifically name it as a *rotation-invariant classifier*.

The initial result shows several popular classifiers dealing with numerical data are rotation-invariant. Due to the space limitation, we will ignore the concrete proofs [4], and summarize that KNN , general Kernel methods, SVM classifiers using polynomial, radial basis, and neural network kernels, and Hyperplane-based classifiers are invariant to rotation.

3 Evaluating Privacy Quality for Random Rotation Perturbation

The goals of rotation based data perturbation are twofold: preserving the accuracy of classifiers, and preserving the privacy of data. The discussion about the rotation-invariant classifiers has proven that the rotation transformation theoretically guarantees zero-loss of accuracy for three popular types of classifiers. We dedicate this section to discuss how good the rotation perturbation approach is in terms of preserving privacy. The critical step to identify the *good* rotation perturbation is to define a multi-column privacy measure for evaluating the privacy quality of any rotation perturbation to a given dataset. With this privacy measure, we can employ some optimization methods to find good rotation perturbations for a given dataset.

For data perturbation approach, the quality of preserved privacy can be understood as the difficulty level of estimating the original data from the perturbed data. Basically, the attacks to the data perturbation techniques can be summarized in three categories: (1)estimating the original data directly from the perturbed data [3, 2], without any other knowledge about the data (naive inference); (2) approximately reconstructing the data from the perturbed data and then estimating the original data from the reconstructed data [8, 6] (approximation-based inference); and (3) if the distributions of the original columns are known, the values or the properties of the values in the particular part of the distribution can be estimated [2, 5] (distribution-based inference). A multi-colum metric should be applicable to all three types of inference attacks to determine the robustness of the perturbation technique. We will focus on the first two attacks in this paper.

3.1 Privacy Model for Multi-column Perturbation

Unlike the existing value randomization methods, where multiple columns are perturbed separately, the random rotation perturbation needs to perturb *all* columns together, where the privacy quality of all columns is correlated under one single transformation.

Since in practice different columns(attributes) may have different privacy concern, we consider that a general-purpose privacy metric Φ for entire dataset is based on **column privacy metric**. An abstract privacy model is defined as follows. Let $\mathbf{p}$ be the column privacy metric vector $\mathbf{p} = (p_1, p_2, \ldots, p_d)$ for d columns, and there are **privacy weights** associated to the columns, respectively, notated as $\mathbf{w} = (w_1, w_2, \ldots, w_d)$. $\Phi = \Phi(\mathbf{p}, \mathbf{w})$ defines the overall privacy guarantee. Basically, the design of privacy model should consider determining the three factors $\mathbf{p}$, $\mathbf{w}$, and function Φ. We summarize our design of privacy metric as follows.

Unified Column Privacy Metrics Below we extend the variance-based privacy metric [3] to the multi-column unified metric. Let $\mathbf{Y}$ be a random variable, representing a column of the dataset, $\mathbf{Y}'$ be the perturbed/reconstructed result of $\mathbf{Y}$, and $\mathbf{D}$ be the difference between $\mathbf{Y}$ and $\mathbf{Y}'$. Let $E[\mathbf{D}]$ and $Var(\mathbf{D})$ denote the mean and the variance of difference (VoD) respectively. $E[\mathbf{D}]$ is not effective in protecting privacy, thus VoD becomes the primary measure in terms of the first level of inferences. Unfortunately, this single-column privacy metric does not work across different columns since it ignores the effect of value range and the mean of the original data column. It is easy to understand that the same amount of VoD is not equally effective for different value ranges. One effective way to unify the different value ranges is via *normalization*.

Let $s_i = 1/(max(\mathbf{Y}_i) - min(\mathbf{Y}_i))$, $t_i = min(\mathbf{Y}_i)/(max(\mathbf{Y}_i) - min(\mathbf{Y}_i))$ denote the constants determined by the value range of the column $\mathbf{Y}_i$. The column $\mathbf{Y}_i$ is scaled to range [0, 1], generating $\mathbf{Y}_{si}$, with the transformation $\mathbf{Y}_{si} = s_i(\mathbf{Y}_i - t_i)$. This allows all columns to be evaluated on the same base, eliminating the effect of diverse value ranges. The normalized data $\mathbf{Y}_{si}$ is then perturbed to $\mathbf{Y}'_{si}$. Let $\mathbf{D}'_i$ =$\mathbf{Y}'_{si} - \mathbf{Y}_{si}$. We use $Var(\mathbf{D}'_i)$, instead of $Var(\mathbf{D}_i)$, as the unified column privacy metric.

Composing the Column Metrics Having the unified column metrics $\mathbf{p}$, we can compose the multiple metrics into one metric for optimization. Let $\mathbf{w}$ denote the importance of columns in terms of preserving privacy. Intuitively, the more important the column is, the higher level of privacy guarantee will be required for the perturbed data column. Therefore, we let $\sum_{i=1}^{d} w_i = 1$ and use p_i/w_i to represent the *weighted column privacy*.

The first composition function is the *minimum privacy guarantee* among all columns. Concretely, when we measure the privacy quality of a multi-column perturbation, we need to pay special attention to the column having the lowest weighted column privacy, because such columns could become the breaking point of privacy. Hence, we design the minimum privacy guarantee $\Phi_1 = \min_{i=1}^{d}\{p_i/w_i\}$. Similarly, the *average privacy guarantee* of the multi-column perturbation, $\Phi_2 = \frac{1}{d}\sum_{i=1}^{d} p_i/w_i$, is another interesting measure. With the definition of privacy guarantee, we can evaluate and optimize the privacy quality of a give perturbation.

Multi-column Privacy Analysis for Random Rotation Perturbation With the variance metric over the normalized data, we can formally analyze the privacy quality of random rotation perturbation. Let X be the normalized dataset, X' be the rotation of X, and I_d be the d-dimensional identity matrix. Thus, VoD can be evaluated based on the difference matrix $X' - X$, and the VoD for column i is the element (i,i) in the covariance matrix of $X' - X$, which is represented as

$$\begin{aligned} & Cov(X' - X)_{(i,i)} = Cov(RX - X)_{(i,i)} \\ = \quad & ((R - I_d)Cov(X)(R - I_d)^T)_{(i,i)} \end{aligned} \qquad (1)$$

Let r_{ij} represent the element (i, j) in the matrix R, and c_{ij} be the element (i, j) in the covariance matrix of X. The VoD for ith column is computed as follows.

$$Cov(X' - X)_{(i,i)} = \sum_{j=1}^{d}\sum_{k=1}^{d} r_{ij}r_{ik}c_{kj} - 2\sum_{j=1}^{d} r_{ij}c_{ij} + c_{ii} \qquad (2)$$

We develop a simple method to implement a fast local optimization. As shown in Equation 2, the privacy metric of column i is only related to the row vectors of rotation. Therefore, swapping the rows of rotation matrix could provide a better rotation that provides higher privacy guarantee. This method can significantly reduce the search space and thus provides better efficiency as we observed in experiments.

ICA-based Attack to Rotation Perturbation Independent Component Analysis (ICA) [7] is the most potential method to breaching the privacy protected by rotation perturbation. However, we argue that ICA is in general not effective in breaking the rotation perturbation, in practice. ICA can be briefly describes as follows. Let matrix X composed by the source signals, where each row vector is a signal, and the observed mixed signals X' be $X' = AX$. ICA model can be applied to estimate the independent components (the row vectors) of the original signals X, from the mixed signals X', if the four conditions are satisfied [7].

Three factors make the ICA attacks are often quite ineffective for rotation perturbation. First, two of the four conditions, although reasonable for signal processing, are not common for data classification: 1) The source signals are independent and 2) All the source signals must be non-Gaussian with possible exception of one signal. In addition, the ordering of the reconstructed signals can not be determined.

In practice, we can evaluate the effectiveness of ICA attacks with difference between the constructed data and the original data. Since the ordering of the reconstructed row

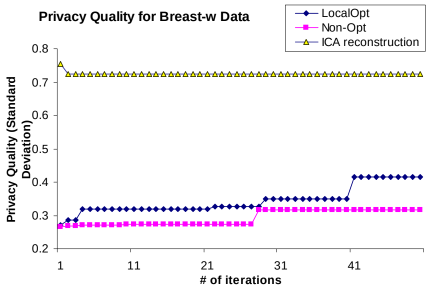

Figure 1. ICA reconstruction has no effect on privacy guarantee.

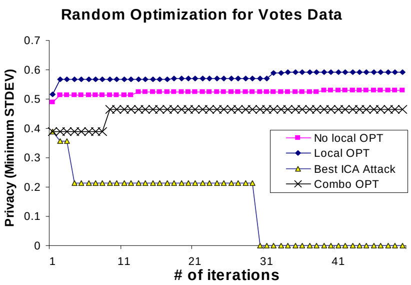

Figure 2. Example that ICA undermines the privacy guarantee.

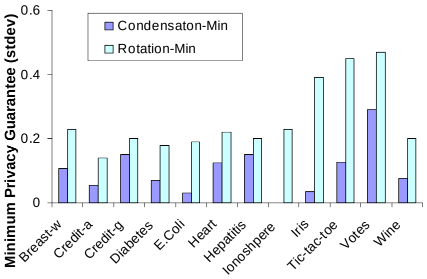

Figure 3. Comparison on minimum privacy level with condensation approach.

vectors is not certain, we estimate the VoDs with the best effort of the $d!$ possible row orderings. Let $\hat{X}_k$ be the ICA reconstructed data with one of the orderings, and P_{ica}^k be the minimum privacy guarantee for $\hat{X}_k$, $k = 1 \ldots d!$. The ordering that gives lowest minimum privacy quality is selected as the most likely ordering and the corresponding privacy quality is the undermined privacy quality.

Algorithm. Combining the local optimization and the test for ICA attacks, we develop a random iterative algorithm to find a better rotation in terms of privacy quality. The algorithm runs in a given number of iterations. In each iteration, it randomly generates a rotation matrix. Local optimization through row-swapping rows is applied to find a better rotation matrix, which is then tested by the ICA reconstruction. We take the combination $P = min\{P_{ica}, P_{opt}\}$ as the final privacy guarantee. The rotation matrix is accepted as the best perturbation yet if it provides highest P among the previous perturbations.

4 Experimental Result

We design three sets of experiments. The first set is used to show that the discussed classifiers are invariant to rotations. The second set shows privacy quality of the good rotation perturbation. The third one compares the privacy quality between the condensation approach and the random rotation approach. Due to the space limitation, we report some results of the later two sets of experiments. The datasets are all from UCI machine learning database. Three results are selected to show the effectiveness of the rotation perturbation approach.

Figure 1 represents a typical scenario that ICA attacks are totally ineffective, while Figure 2 shows, when ICA attacks are substantial, the algorithm can also find a rotation that has the highest combined privacy guarantee in the random rotation matrices. Figure 3 demostrates the rotation approach can provide much higher privacy quality than the condensation approach [1].

5 Conclusion

Loss of privacy and loss of information/accuracy are treated as two conflict factors in privacy preserving data classification. In this paper, we propose a rotation based perturbation technique that guarantees zero loss of accuracy for many classifiers. Meanwhile, we can adjust the rotation to find a locally optimal rotation in terms of basic privacy guarantees and possible attacks, where the optimality is measured by a new multi-column privacy metric. Experiments show that the rotation perturbation can greatly improve the privacy quality without sacrificing accuracy.

References

[1] C. C. Aggarwal and P. S. Yu. A condensation approach to privacy preserving data mining. *Proc. of Intl. Conf. on Extending Database Technology (EDBT)*, 2004.

[2] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. *Proc. of ACM PODS Conference*, 2002.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. *Proc. of ACM SIGMOD Conference*, 2000.

[4] K. Chen and L. Liu. A random rotation perturbation approach to privacy preserving data classification. *Technical Report*, 2005.

[5] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. *Proc. of ACM PODS Conference*, 2003.

[6] Z. Huang, W. Du, and B. Chen. Deriving private information from randomized data. *Proc. of ACM SIGMOD Conference*, 2005.

[7] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley-Interscience, 2001.

[8] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. *Proc. of Intl. Conf. on Data Mining (ICDM)*, 2003.

[9] Y. Lindell and B. Pinkas. Privacy preserving data mining. *Journal of Cryptology*, 15(3), 2000.