

Generalized Random Rotation Perturbation for Vertically Partitioned Data Sets *

Zhenmin Lin, Jie Wang, Lian Liu, Jun Zhang[†]

Laboratory for High Performance Scientific Computing and Computer Simulation,
Department of Computer Science, University of Kentucky,
Lexington, KY 40506-0046, USA

July 7, 2008

Abstract

Random rotation is one of the common perturbation approaches for privacy preserving data classification, in which the data matrix is multiplied by a random rotation matrix before publishing in order to preserve data privacy. One distinct advantage of this approach is that it can maintain the geometric properties of the data matrix, so several categories of classifiers that are based on the geometric properties of the data can achieve similar accuracy on the transformed data as that on the original data. In this paper, we generalize this idea to the situation where the data matrix is assumed to be vertically partitioned into several submatrices and held by different owners. Each data holder can choose a rotation matrix randomly and independently to perturb their individual data. Then they all send the transformed data to a third party, who collects all of them and forms a whole data set for data mining or other analysis purposes. We show that under such a scheme the geometric properties of the data set is also preserved and thus it can maintain the accuracy of many classifiers and clustering techniques applied on the transformed data as on the original data. Experiments on two real data sets show that such generalization is effective for vertically partitioned data sets.

keyword: Privacy preserving, matrix rotation, data perturbation, data mining

*Technical Report CMIDA-HiPSCCS 008-08, Department of Computer Science, University of Kentucky, KY, 2008. The research work of J. Zhang was supported in part by NSF under grants CCF-0527967 and CCF-0727600, in part by NIH under grant 1R01HL086644-01, in part by Alzheimer's Association under grant NIGR-06-25460, and in part by KSEF under grant KSEF-148-502-06-186.

[†]E-mail: jzhang@cs.uky.edu. URL: <http://www.cs.uky.edu/~jzhang>.

1 Introduction

Data mining has been a powerful technique in analyzing and utilizing data in today’s information-rich society. However, privacy is nowadays a major concern in data mining applications, which has led to a new research area, privacy preserving data mining. A large amount of research work has been devoted to this area, and resulted in such techniques as k -anonymity [1], data perturbation [2, 3, 4, 5], and privacy preserving distributed data mining [6, 7].

Data perturbation is one of the commonly used models for privacy preserving data mining. The data owners change the data values in some way to hide the sensitive information while try to maintain the utility of the data. They publish the distorted data instead of the original one. Several techniques have been proposed for data perturbation, such as [2, 3, 4, 5].

Paper [2] proposes a method which adds random noise to the data to preserve the sensitive information while maintaining the distribution of the underlying data and hence the utility of data. This method is vulnerable to breach. Several techniques have been proposed to reconstruct the original data from the perturbed one, such as spectral filtering [8], Principal Component Analysis [9]. Generally, if we try to preserve more privacy, we may have to loss more information. It seems contradictory to achieve the goal of preserving privacy and maintaining the utility of the data.

As pointed out in [3], the information we try to maintain is specific to the data mining task and the particular model. In [3], a random rotation technique is proposed to perturb data, in which the data matrix will be rotated randomly. This method maintains the geometric property of the data, and thus will not affect the performances of many classifiers, known as the rotation-invariant classifiers, which utilize the geometric properties of the underlying data for classification. So this method can preserve the data privacy without any loss of information for specific data mining tasks. We will discuss this method in more detail in Section 2.

In this paper, we generalize the random rotation techniques for vertically partitioned data set. The whole data set is assumed to be vertically partitioned into several submatrices and held by different data owners. Each owner can choose a rotation matrix randomly and perturb their data set independently. Then they all send the transformed data to a third party, which assembles them into a whole data matrix for data mining. We show that such method will still maintain the geometric properties of the data and thus will not affect the performance of rotation-invariant classifiers and clustering techniques. We discuss the generalization of random rotation transformation in Section 3. Experimental

results are presented in Section 4 and the concluding remarks are included in Section 5.

2 Random rotation perturbation

Random rotation perturbation is proposed in [3] for privacy preserving data classification. The basic ideas are as following. We represent n objects with m attributes with an $n * m$ matrix M , whose entries are assumed to be real numbers. We can view the matrix M as n points in an m dimensional space and call each object a point correspondingly. We generate an $m*m$ rotation matrix R randomly and multiply M with R to get the perturbed matrix $P = M * R$. This perturbed matrix P is then published for the task of classification. The privacy is preserved as the data values of the published matrix P is quite different from those of the original matrix M .

A rotation matrix R is a matrix which satisfies the following property:

$$R * R^T = R^T * R = I.$$

Here R^T denotes the transpose of R and I is the identity matrix. This property implies that both the rows and columns of the matrix are orthonormal, that is, for any row i ,

$$\sum_{l=1}^m r_{il}^2 = 1$$

and for any two different rows i, j ,

$$\sum_{l=1}^m r_{il} * r_{jl} = 0.$$

Furthermore, for any column i ,

$$\sum_{l=1}^m r_{li}^2 = 1$$

and for any two different columns i, j ,

$$\sum_{l=1}^m r_{li} * r_{lj} = 0.$$

One important feature of the rotation transformation is that it maintains length. Let $|x| = \sqrt{x * x^T}$ represent the length of a row vector x . Since R is a rotation matrix,

$$|x * R|^2 = (x * R) * (x * R)^T = x * R * R^T * x^T = x * x^T = |x|^2,$$

so $|x * R| = |x|$. Thus for any pair of points x and y , $|(x - y) * R| = |x - y|$. It follows that the Euclidean distance between any two points is maintained. Similarly, the inner product of any two points is also maintained. This can be shown below. Let $\langle x, y \rangle = x * y^T$, then

$$\langle x * R, y * R \rangle = x * R * (y * R)^T = x * R * R^T * y = x * y^T = \langle x, y \rangle .$$

Intuitively, rotation transformation will maintain the geometric shape of the data points in the multi-dimensional space.

In [3], the authors defined the concept of “transformation invariant classifiers”. Informally, if a classifier trained on the transformed data has the same accuracy as that trained on the original data, we say that this classification is invariant to the transformation. In particular, if a classifier is invariant to rotation transformation, we say that it is rotation invariant. Based on the observation that rotation transformation will not change the geometric shape of the data points, paper [3] identifies three categories of rotation-invariant classifier: kernel methods (including KNN classifier), SVM and perceptrons. Experiments in Section 5 of [3] verify the invariance properties of KNN classifier, SVM classifier with RBF kernel and perceptrons.

Although not mentioned in paper [3], it is clear that clustering techniques which are based on Euclidean distance are also invariant to rotation transformation. They can cluster transformed data with similar accuracy as they cluster the original data. One such example is the k -means clustering.

3 Generalized random rotation perturbation for vertically partitioned data sets

We consider the situation where the $n * m$ data matrix M is vertically partitioned into several disjoint submatrices, $M = (M_1, M_2, \dots, M_r)$, where M_i is an $n * m_i$ submatrix, and $\sum_{i=1}^r m_i = m$. Each M_i is held by a different owner i , and all the data holders will send their individual data to a third party, who collects all of them to form a whole matrix M and then do data mining task on it. To preserve privacy, each owner will perturb their own data independently before publishing. We show that each owner can use random rotation transformation to perturb their data independently while at the same time maintain the Euclidean distances and inner products of data points represented in the whole data matrix. That is, the data owner i can choose an $m_i * m_i$ rotation matrix R_i randomly and independently and applies rotation transformation to his/her data M_i and

then release the transformed data $P_i = M_i * R_i$ to the third party. The third party collects all the transformed data and forms the whole matrix $P = (P_1, P_2, \dots, P_r)$. We prove that the Euclidean distances and inner products of the points in P are the same as those of the corresponding points in M . For simplicity, we consider that the data matrix is vertically partitioned into two parts, that is, $r = 2$. The results can be easily generalized to $r > 2$.

Let x denote the m dimensional row vector of an object in the matrix M . Let $x = (x_1, x_2)$, where x_1 is an m_1 -dimensional row vector in the submatrix M_1 held by owner 1 and x_2 is an m_2 -dimensional row vector in M_2 held by owner 2. Suppose owner 1 chooses a random rotation matrix R_1 independently and owner 2 chooses a random rotation matrix R_2 independently. Now the submatrix $M_i (i = 1, 2)$ is transformed into $P_i = M_i * R_i$, respectively. Correspondingly, x_1 is transformed into $x_1 * R_1$, and x_2 into $x_2 * R_2$. Now the third party has the perturbed matrix $(M_1 * R_1, M_2 * R_2)$ and the transformed vector corresponding to x is $(x_1 * R_1, x_2 * R_2)$.

We first show that the inner product of any two points is preserved. Let $x = (x_1, x_2), y = (y_1, y_2)$ be any two row vectors. They are transformed into $(x_1 * R_1, x_2 * R_2)$ and $(y_1 * R_1, y_2 * R_2)$, respectively. Then

$$\begin{aligned}
& \langle (x_1 * R_1, x_2 * R_2), (y_1 * R_1, y_2 * R_2) \rangle \\
&= (x_1 * R_1, x_2 * R_2) * (y_1 * R_1, y_2 * R_2)^T \\
&= (x_1 * R_1) * (y_1 * R_1)^T + (x_2 * R_2) * (y_2 * R_2)^T \\
&= x_1 * R_1 * R_1^T * y_1^T + x_2 * R_2 * R_2^T * y_2^T \\
&= x_1 * y_1^T + x_2 * y_2^T \\
&= (x_1, x_2) * (y_1, y_2)^T \\
&= \langle (x_1, x_2), (y_1, y_2) \rangle \\
&= \langle x, y \rangle
\end{aligned}$$

Using the above property, we can show that the length of a vector is preserved.

$$\begin{aligned}
& |(x_1 * R_1, x_2 * R_2)|^2 \\
&= \langle (x_1 * R_1, x_2 * R_2), (x_1 * R_1, x_2 * R_2) \rangle \\
&= \langle (x_1, x_2), (x_1, x_2) \rangle \\
&= |(x_1, x_2)|^2 \\
&= |x|^2.
\end{aligned}$$

So we have $|(x_1 * R_1, x_2 * R_2)| = |(x_1, x_2)|$. The distance between any two points is also

preserved as

$$\begin{aligned}
& |(x_1 * R_1, x_2 * R_2) - (y_1 * R_1, y_2 * R_2)| \\
&= |(x_1 * R_1 - y_1 * R_1, x_2 * R_2 - y_2 * R_2)| \\
&= |(x_1 - y_1) * R_1, (x_2 - y_2) * R_2| \\
&= |(x_1 - y_1, x_2 - y_2)| \\
&= |x - y|.
\end{aligned}$$

As with the case where we transform the whole data matrix using only one rotation matrix, when we transform the submatrix using different rotation matrix independently, both the distance and the inner product of any two points are maintained. Hence the geometric properties of data sets are also preserved. So if we use rotation-invariant classifiers and clustering techniques on the transformed data sets, we can get similar accuracies as those on the original data sets.

4 Experiments

4.1 Privacy Metrics

We first discuss briefly the privacy metrics used in the experiments for evaluating the privacy quality of the data perturbation approaches [3]. Let Y be a random variable, which represents a column of the data matrix, Y' be the perturbed result of Y . Let D be the difference between Y' and Y , i.e., $D = Y' - Y$, and we denote its expectation and standard deviation by $E[D]$ and σ , respectively. Let c be some constant. Using Chebyshev Inequality, we know that for a perturbed value y' , the original value y is located in the range $[y' - E[D] - c, y' - E[D] + c]$ with probability at least $(1 - 1/c^2)$ [3, 10]. If the length of the interval $2c\sigma$ is big, it is more difficult to estimate the original value. So in [3], the standard deviation of D , σ , is used as the privacy metric for a single column perturbation.

To evaluate the privacy quality of multi-column perturbation, [3] first normalizes each column Y_i by

$$Y_{si} = \frac{Y_i - \min(Y_i)}{\max(Y_i) - \min(Y_i)}$$

so that Y_i is scaled to the range $[0, 1]$. Then the normalized data Y_{si} is perturbed to Y'_{si} . Let $D'_i = Y'_{si} - Y_{si}$. The standard deviation of D'_i , instead of D_i , is used as privacy metric. Now we have one privacy level p_i for each column. Suppose that a weight of importance w_i is assigned to each column i such that $\sum_{i=1}^m w_i = 1$. The minimum privacy guarantee

is defined as

$$\phi_1 = \min_i \left\{ \frac{p_i}{w_i} \right\}$$

and the average privacy guarantee is defined as

$$\phi_2 = \frac{1}{m} \sum_{i=1}^m \frac{p_i}{w_i}.$$

If we treat the privacy of each column equally important, we may exclude the weights in the definition and define the minimum privacy guarantee as

$$\phi_1 = \min_i \{p_i\}$$

and the average privacy guarantee as

$$\phi_2 = \frac{1}{m} \sum_{i=1}^m p_i.$$

4.2 Experimental Results

We test the generalized random rotation perturbation approaches using k -means clustering and SVM on two data sets from UCI repository [11]. We use Kmeans function in MATLAB software for clustering. The SVM software is downloaded from <http://svmlight.joachims.org/>. We use ten folds for classification, and run the SVM using the default setting. In the experiments, whenever we try to perturb a (sub)matrix, we choose 100 rotation matrices randomly and select the one with the largest minimum privacy guarantee to perturb the data (sub)matrix. We denote minimum privacy guarantee by L_{min} and the average privacy guarantee by L_{ave} . We treat the privacy of each column equally important and use

$$\phi_1 = \min_i \{p_i\}$$

to calculate L_{min} , and

$$\phi_2 = \frac{1}{m} \sum_{i=1}^m p_i$$

for L_{ave} .

The first data set is Wisconsin Breast Cancer, denoted by WBC below, which has 9 attributes. We vertically partition this data set into two submatrices, $WBC = (WBC-1, WBC-2)$, where WBC-1 consists of the first 5 attributes and WBC-2 of the remaining 4 attributes. We perturb each submatrix using different random rotation matrices indepen-

Privacy Level	WBC-1	WBC-2	WBC
L_{min}	0.40	0.39	0.39
L_{ave}	0.53	0.54	0.53

Table 1: Privacy level of WBC dataset

Privacy Level	IONO-1	IONO-2	IONO-3	IONO
L_{min}	0.30	0.34	0.33	0.30
L_{ave}	0.38	0.45	0.41	0.42

Table 2: Privacy level of Ionosphere dataset

dently and then assemble these two transformed submatrices into one matrix. The privacy qualities of the perturbed matrices WBC-1, WBC-2 and WBC are presented in Table 1. We classify and cluster the transformed matrix using SVM and k -means and compare the results with the ones we get on the original data set. The accuracies of classification and clustering are presented in Table 3.

The other data set is Ionosphere, denoted by IONO below, which has 34 attributes. We vertically partition it into 3 submatrices, IONO=(IONO-1, IONO-2, IONO-3), which have 12, 11 and 11 attributes respectively. We randomly rotate each submatrix independently and then assemble them together. We present the privacy qualities of the perturbed submatrices and the whole matrix in Table 2. The comparison of classification and clustering on the transformed data set with those on the original one are presented in Table 3.

As discussed in section 4.1, the privacy level is calculated over the normalized matrix. Note that the privacy level is the standard deviation of $D = Y - Y'$ and that the values of the entries in the normalized matrix are within the range $[0,1]$. So the results in Tables 1 and 2 indicates that the privacy of the assembled data and the partitioned data is well preserved. Also note that the minimum privacy guarantee of the assembled matrix is the minimum of the minimum privacy guarantees of its submatrices as the minimum privacy guarantee of a perturbed matrix is defined as the minimum of all its columns. And if we assign the number of attributes to each submatrix as its weight, the average privacy level of the whole matrix is the weighted average of the average privacy level of its submatrices. The results in Table 3 show that there is no change in the accuracy of the classification and clustering results with respect to both data sets.

Note that, although in Table 3 the accuracies of classification and clustering on the transformed data set are exactly the same as those on the original one, they may be slightly different in practice due to some other factors, such as the different initial guess of the clustering centers in the k -means clustering, and the numerical rounding errors caused by the rotation matrix multiplication.

Data Sets	SVM on Original Matrix	SVM on Transformed Matrix	k -means on Original Matrix	k -means on Transformed Matrix
WBC	96.7%	96.7%	96.0%	96.0%
IONO	84.5%	84.5%	71.2%	71.2%

Table 3: Accuracy of classification and clustering on data sets WBC and Ionosphere

5 Conclusion

Random rotation is one of the popular approaches for data perturbation. It can preserve the data privacy without affecting the accuracy for rotation-invariant classifiers and clustering. We generalize this idea for vertically partitioned data sets. We rotate each submatrix randomly and independently and prove that it will preserve the geometric properties of the data matrix and thus the rotation-invariant classifier and clustering techniques will achieve similar accuracy on the transformed data as on the original data. Experiments on two real data sets show that this generalization is effective for vertically partitioned data sets. We note that such generalization only works for vertically partitioned data sets, but not for horizontally partitioned data sets.

References

- [1] Sweeney, L. k -anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557-570, 2002.
- [2] Agrawal, R. and Srikand R. Privacy preserving data mining. In *Proc. Of ACM SIGMOD Conference*, pp. 439-450, 2000.
- [3] Chen, K. and Liu. L. A random rotation perturbation approach to privacy data classification. In *Proc of IEEE Intl. Conf. on Data Mining (ICDM)*, pp. 589-592, 2005.
- [4] Xu, S., Zhang, J., Han, D. and Wang J. Singular value decomposition based data distortion strategy for privacy distortion. *Knowledge and Information System*, 10(3):383-397, 2006.
- [5] Mukherjee, S., Chen, Z. and Gangopadhyay, A. A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier related transforms. *Journal of VLDB*, 15(4):293-315, 2006.
- [6] Vaidya, J. and Clifton, C. Privacy preserving k -means clustering over vertically partitioned data. In *Prof. of ACM SIGKDD Conference*, pp. 206-215, 2003.

- [7] Vaidya, J., Yu, H. and Jiang, X. Privacy preserving SVM classification. *Knowledge and Information Systems*, 14:161-178, 2007.
- [8] Kargupta, H., Datta, S., Wang, Q. and Sivakumar, K. Random data perturbation techniques and privacy preserving data mining. *Knowledge and Information Systems*, 7:387-414, 2005.
- [9] Huang, Z., Du, W. and Chen, B. Deriving private information from randomized data. In *Proc. of ACM SIGMOD Conference*, pp. 37-48, 2005.
- [10] Ross, S.M. A first course in probability, Macimilan publishing company, 1998.
- [11] Asuncion, A. and Newman, D.J. UCI Machine Learning Repository [<http://www.ics.uci.edu/~mlern/MLRepository.html>]. 2007.