

Data Pattern Maintenance by Matrix Approximation: An Application to Information Security

Jie Wang and Jun Zhang

Laboratory for High Performance Scientific Computing and Computer Simulation, Department of Computer Sciences, University of Kentucky,
Lexington, KY 40506-0046, USA. E-mail: jwanga@csr.uky.edu , jzhang@cs.uky.edu

Abstract— Maintaining data mining accuracy on distorted datasets is an important issue in privacy preserving data mining. Using matrix approximation, we propose several efficient and flexible techniques to address this issue, and utilize some statistical metrics to analyse change of data pattern. We use the K-nearest neighbour classification to compare accuracy maintenance after data distortion by different methods. With better performance than some classical data perturbation approaches, nonnegative matrix factorization and singular value decomposition are considered to be promising techniques for privacy preserving data mining. Experimental results demonstrate that mining accuracy on the distorted data used these methods is almost as good as that on the original data, with added property of privacy preservation. It indicates that our matrix factorization-based data distortion schemes perturb only confidential attributes to meet privacy requirements while preserving general data pattern for knowledge extraction.

Index Terms — matrix factorization, nonnegative matrix factorization, privacy, data mining

I. INTRODUCTION

Recently, there has been increasing interest in developing new techniques specifically to address the issues relevant to minimize disclosure level of confidential information. Protection of privacy from unauthorized access is one of the primary concerns in data use, from national security to business transactions. It brings out a new branch of data mining, known as privacy preserving data mining (PPDM). For numerical data, data perturbation is one of the most popular models which can be classified into two categories: the independent attribute perturbation approach, wherein the value of each attribute is distorted independently of the rest; and the dependent attribute perturbation approach, where the distortion of attributes may be correlated [1]. Before data owners publish the data, they modify the data values by adding random noise to numerical attributes. Many of them use randomized data distortion techniques to mask the data by randomly modifying the data values. Additive noise and multiplicative noise are most commonly used [2]. However, a potential problem is that since each data element is perturbed independently, the pair-wise similarity of objects may not be maintained.

Jie Wang and Jun Zhang are with the Laboratory for High Performance Scientific Computing and Computer Simulation, Department of Computer Sciences, University of Kentucky, Lexington, KY 40506-0046, USA (*corresponding author: E-mail: jzhang@cs.uky.edu phone: 859-257-9348). Technical Report No. 477-07, Department of Computer Science, University of Kentucky, Lexington, KY, 2007.

Furthermore, privacy breach has recently been identified as one of its major problems. It has been discovered that some privacy intrusion techniques can be used to reconstruct private data from the randomized data. The spectral properties of the randomized data could help the data miner to separate noise from private data. In particular, a filtering method based on random matrix theory is proposed to reconstruct private data from the randomized data set. It shows that randomization preserves little privacy in many cases [3]. Two other data reconstruction methods, Principal Component Analysis-based and Bayes Estimate-based, are proposed in [4] to restore original data from disturbed data. It is suggested that the amount of private information that can be disclosed is related to data correlation, and the more the correlation of noises resembles that of the original data, the better privacy preservation can be achieved [4].

A key lemma of matrix approximation or dimensionality reduction, the Johnson-Lindenstrauss lemma, shows that a set of n points in high dimensional Euclidean space can be mapped down into an $O(\log n/\epsilon^2)$ dimensional Euclidean space such that the distance between any two points changes by only a factor of $(1\pm\epsilon)$. Within the context of data mining, matrix factorization plays a major role to obtain some version of simplified low-rank approximation to the original datasets. This characteristic stimulates our interest in utilizing it to achieve both high level privacy preservation and high degree data mining accuracy.

In this paper, we build a matrix decomposition framework to address the accuracy issues in privacy preserving classification without the loss of satisfied data distortion level. Our targeted database is defined as a centralized one with numerical values. The accuracy evaluation is conducted using a distance-based classification.

II. MATRIX APPROXIMATION FOR DATA DISTORTION

We assume that the original dataset consists of n objects with each object having m numerical attributes. It is encoded by a vector space model as an $n\times m$ object-attribute matrix \mathbf{A} . We use $\tilde{\mathbf{A}}$ to denote the distorted counterpart of \mathbf{A} .

A. Two Popular Data Distortion Techniques

Additive noise approach can be denoted by the addition of a perturbation matrix \mathbf{M}_P to the original data as $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{M}_P$, which is of the same size as \mathbf{A} . In our experiments, we use Uniformly Distributed Noise (UD) and Normally Distributed

Noise (ND) methods. In the UD, the original data \mathbf{A} is reconstructed by adding a uniformly distributed noise matrix \mathbf{N}_u [5]. The distorted matrix is $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{N}_u$. The entries of \mathbf{N}_u are random numbers drawn from the continuous uniform distribution on the interval from C_1 to C_2 . Similar to the UD, ND uses a noise matrix \mathbf{N}_n generated from a normal distribution with some mean and standard deviation.

Another distortion approach, a random projection based on multiplicative noise, can be denoted by a matrix multiplication. The original m -dimensional data is projected into a k -dimensional subspace using a random matrix \mathbf{R} whose columns have unit lengths. $\tilde{\mathbf{A}}' = \mathbf{R} * \mathbf{A}'$.

B. Matrix Factorization Techniques

1) Singular Value Decomposition (SVD)

The SVD of the matrix \mathbf{A} can be written as $\mathbf{A}_{n \times m} = \mathbf{U}_{n \times n} \mathbf{\Sigma}_{n \times m} \mathbf{V}'_{m \times m}$. \mathbf{U} and \mathbf{V} contain the left and right singular vectors of \mathbf{A} , respectively, and the diagonal of $\mathbf{\Sigma}$ is the singular values in descending order. These three matrices reflect a breakdown of the original relationship into linearly independent vectors. Using a truncated SVD, $\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k'$, the dimensionality of the data can be reduced by projecting the m column vectors onto a k dimensional space corresponding to the k largest singular values. The truncation also removes noise by ignoring less significant data structures. Therefore, it is possible to achieve higher level data mining accuracy by performing the truncated SVD operation on the original data.

2) Nonnegative Matrix Factorization (NMF)

NMF is a vector space method to obtain a representation of data using nonnegative constraints. These constraints can lead to a part-based representation because they allow only additive combinations of the original data [6]. It has been broadly applied to a variety of research areas.

Given a nonnegative matrix $\mathbf{A} \in R^{n \times k}$ with $\mathbf{A}(i, j) \geq 0$ and a pre-specified positive integer $k < \min(n, m)$, NMF finds two nonnegative matrices $\mathbf{W} \in R^{n \times k}$ with $\mathbf{W}(i, j) \geq 0$ and $\mathbf{H} \in R^{k \times m}$ with $\mathbf{H}(i, j) \geq 0$ so that $\mathbf{A} \approx \mathbf{WH}$ that minimizes the objective function

$$f(\mathbf{W}, \mathbf{H}) = \frac{1}{2} \|\mathbf{A} - \mathbf{WH}\|_F^2 \quad (1)$$

A standard way to find \mathbf{W} and \mathbf{H} is by the following least-square optimization, which minimizes the difference between \mathbf{A} and \mathbf{WH} :

$$\min_{\mathbf{W}, \mathbf{H}} f(\mathbf{W}, \mathbf{H}) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m (\mathbf{A}(i, j) - (\mathbf{WH})(i, j))^2 \quad (2)$$

subject to $W_{ia} \geq 0, H_{bj} \geq 0, \forall i, a, b, j$.

We use the alternating nonnegative least-squares by projected gradients, proposed by Lin [6]. This method leads to a faster convergence than the popular multiplicative update method.

C. Accuracy Metrics

For distance-based mining algorithms, each object that is mapped to the same class may be thought of as more similar to the objects in that class than to the objects in other classes. Distance measure is mostly used to identify the ‘‘alikeeness’’ of different objects in the datasets. K-nearest neighbors (KNN) and k-means clustering are two popular data mining algorithms based on distances. Therefore, their mining accuracy on the distorted datasets depends on the maintenance level of dissimilarity or distance before and after the data distortion.

1) Pair-wise Object Dissimilarity Analysis

We define a symmetric matrix $\mathbf{D}_{n \times n}$ as a dissimilarity matrix that stores a collection of pair-wise distances between every pair of objects in a data set. Each element $d(i, j)$ corresponds to the distance or dissimilarity between objects i and j . In general, $d(i, j)$ is a nonnegative value that is close to zero when the objects i and j are very similar to each other, and becomes larger the more they differ. We use the most popular distance measure, the Euclidean distance, to calculate $\mathbf{D}_{n \times n}$.

$$d(i, j) = \left[\sum_{k=1}^m (x_{ik} - x_{jk})^2 \right]^{1/2} \quad (3)$$

where X_i and X_j are m -dimensional data objects.

We define SM as the average percentage of distances that maintain their ranks in all the pair-wise distances of every object after the distortion. It is computed as:

$$SM = \frac{\sum_{i=1}^{n-1} [(\sum_{i+1}^n Dk_j^i) / (n-i)]}{(n-1)} \quad (4)$$

where Dk_j^i means whether a distance keeps its rank in a single column:

$$Dk_j^i = \begin{cases} 1, & \text{if } DRank_j^i = (DRank_j^i)^* \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

$DRank_j^i$ is the rank of $d(i, j)$ in the i th column of $\mathbf{D}_{n \times n}$, and

$(DRank_j^i)^*$ denotes the rank of $d(i, j)$ in the i th column of $\tilde{\mathbf{D}}_{n \times n}$,

where $\mathbf{D}_{n \times n}$ is the similarity matrix of \mathbf{A} and $\tilde{\mathbf{D}}_{n \times n}$ is the similarity matrix of $\tilde{\mathbf{A}}$. The larger the value of SM is, the better the pair-wise distances is kept in the distortion strategy. The distortion strategies with better maintenance of similarity are able to achieve higher accuracy in distance-based mining.

2) Pair-wise Attribute Independency Analysis

Covariance is the measure of how much two attributes vary together. If two attributes tend to vary together, then the covariance between the two attributes is positive. The zero value of covariance means an orthogonal relation. Covariance matrix $\mathbf{C}_{m \times m}$ is a normalized measure of linear relationship strength between variables. For attributes by column vectors, we use correlation coefficients as a measure of the dependency between two attributes. Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect negative correlation while a value of +1.00 represents a perfect

positive correlation. A value of 0.00 represents a lack of correlation. We define CM as the average percentage of attributes that maintain their ranks in all the pair-wise correlations of every attribute. CM is computed in the similar way as SM . Ck_j^i represents whether a pair of attribute keep the rank in $C_{m \times m}$, defined as in (5).

$$CM = \frac{\sum_{i=1}^{m-1} [(\sum_{i+1}^m Ck_j^i) / (m-i)]}{(m-1)} Ck_j^i \quad (6)$$

3) Multi-attribute Correlation Analysis

Correlation maintenance among the attributes over the whole feature space is measured by the cophenetic correlation coefficient, defined as:

$$C = \frac{\sum_{i < j} (\mathbf{X}_{ij} - x)(\mathbf{Y}_{ij} - y)}{\sqrt{\sum_{i < j} (\mathbf{X}_{ij} - x)^2 \sum_{i < j} (\mathbf{Y}_{ij} - y)^2}} \quad (7)$$

D. Implementation procedure description

Given the original confidential dataset \mathbf{A} and a dimension k

1. Normalize \mathbf{A}
2. Compute approximation $\tilde{\mathbf{A}}$ using SVD or NMF
3. Choose an integer $r < k$
4. For $r = k, k-1, \dots, 1$, Do
5. Do further distortion: $\tilde{\mathbf{A}}^{(r)} = \mathbf{W}_{n \times r} \mathbf{H}_{r \times n}$ or $\tilde{\mathbf{A}}^{(r)} = \mathbf{U}_{n \times r} \Sigma_{r \times r} \mathbf{H}_{m \times r}$
6. Compute data distortion metrics on $\tilde{\mathbf{A}}^{(r)}$
7. Train classifier on $\tilde{\mathbf{A}}^{(r)}$ and compute classification accuracy.
8. EndDo
9. Choose one $\tilde{\mathbf{A}}^{(r)}$ with satisfied data distortion level and accuracy.
10. Publish the final distorted dataset $\tilde{\mathbf{A}}^{(r)}$

III. EXPERIMENTS AND RESULTS

The experiments are carried out on two real world datasets, LBin and Wbc. LBin is a terrorist dataset creating by collecting information of 100 terrorists from a terrorist analysis website [7]. We selected 41 attributes ($m=41$), such as their nationality, different sibling relationships, pilot training, locations of temporary residency, wedding attendance, meeting attendance, etc. The original matrix is of dimension 100 by 41. WBC is a nonnegative-value dataset from the Wisconsin Breast Cancer dataset at the UCI Machine Learning Repository. WBC is a 699 by 10 matrix with the 10th column representing class label. Experiments were conducted on a Sunblade 150 workstation.

A. Experimental parameter settings and notations

We use NMF to represent the distortion method using NMF. The same meaning goes with SVD, UD and ND. For all the experiments here, the default values of some parameters are listed as follows:

1. NMF: tolerance for stopping condition=1e-4, time limit = 4000, iteration number limit = 20000.

2. ND: the normally distributed noise is generated with mean $\mu = 0$ and deviation $\sigma = 0.46$.
3. UD: the uniformly distributed noise is generated from the interval $[0, 0.8]$.
4. K-nearest neighbour classification: $k=5$. Euclidean distance and five-fold cross validation are used.
5. For LBin, size of initial estimates on \mathbf{W} and \mathbf{H} is 100 by 30 and 30 by 41. For WBC, initial dimension is set to be 7.

B. Experimental Results of WBC

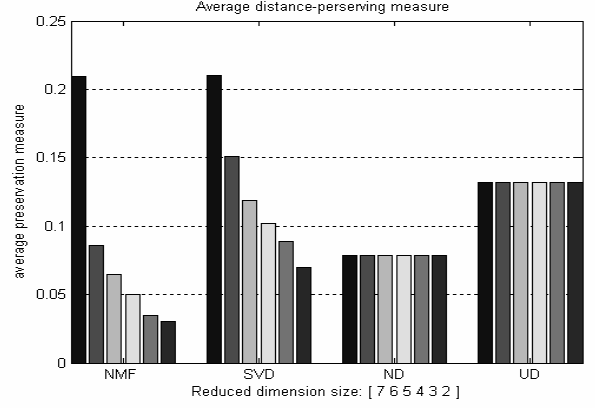


Fig. 1. Average distance-preserving comparison on NMF, SVD, ND and UD, to original data. For each group under 7 different reduced dimension from 7 to 2.

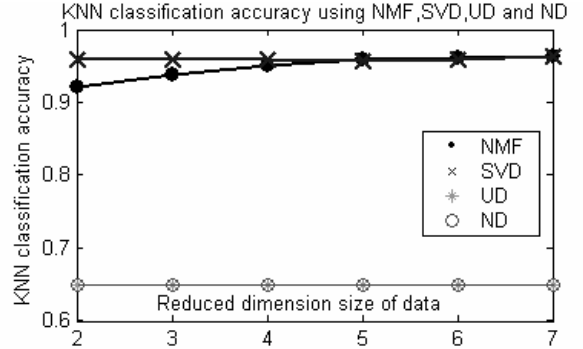


Fig. 2. Classification accuracy comparison using NMF, SVD, ND and UD. For each method, the reduced dimension size is changed from 2 to 7. The accuracy on Wbc is 96%.

Fig. 1 and Fig. 2 shows that NMF and SVD have the similar pattern on the relationship of distance preserving level and reduced dimensionality size, in which both the preserving and accuracy increase with the size. Fig. 2 shows that NMF and SVD achieve a very competitive accuracy almost as good as that of original Wbc. UD and ND only have an accuracy of 65%. The dependency of KNN accuracy on the reduced dimensionality size is slightly more in NMF than in SVD.

C. Experimental Results of LBin

From Fig. 3 and Fig. 4, it is observable that for LBin dataset, distance preserving and correlation preserving demonstrate the same behavior for the same method. SVD has the greater change on these two metrics, while in NMF, dimensionality size does not have much influence on distance and correlation

preserving. It implies that NMF be able to keep data pattern by a much smaller dimensionality size than SVD. UD performs better than ND in keeping distances and correlations, it follows that UD has better accuracy than ND as shown in Fig. 5. As we can see in Fig. 5, the original data accuracy is 75%. Except for the accuracy of ND is 74%, all the other methods have higher accuracy than the accuracy on the original data. Here NMF can achieve a better accuracy even when we reduce its dimensionality size to 1. In some ranges of dimensionality size, the accuracy of SVD and NMF is inversely related to data dimension.

IV. CONCLUSION

Experimental results indicate that for centralized datasets with numerical attributes, matrix factorization-based distortion strategies achieve a satisfactory performance far better than independent data perturbation distortion methods, such as noise additive approaches. In particular, they even achieve better classification accuracy than that of the original datasets since a good dimensionality reduced approximation improves data quality on knowledge mining by removing noise from data. Experiments indicate an optimal dimensionality size exist when the best mining accuracy and the best approximation meet at the same point. For nonnegative valued datasets, NMF-based distortion methods is a better choice among the matrix approximation techniques due to the attractive characteristic that the value of two factor matrices of NMF is not unique, because it is dependent on initial estimates in the beginning of iterative factorization procedure. This kind of dependency provides both uncertainty and flexibility on the data distortion. Therefore, matrix factorization-based approach provides a possibility of simultaneously achieve satisfactory privacy, accuracy and efficiency.

REFERENCES

- [1] S. Agrawal, J.R. Haritsa, "A framework for high-accuracy privacy-preserving mining," *Proceedings of the 21st Int. Conf. on Data Engineering*, 2005
- [2] H. Kargupta, S. Datta, Q. Wang, "Random-data perturbation techniques and privacy-preserving data mining," *Knowledge and Information Systems*, Vol. 7, pp:387-414, 2005
- [3] H. Kargupta, K. Sivakumar, S. Ghosh. "Dependency detection in mobimine and random matrices," *Proceedings of the 6th Europe Conference on Principles and Practices of Knowledge Discovery in Databases*, pp:250-262, 2002
- [4] A. Evfimievski, J. Gehrke, R. Srikant. "Limiting privacy breaches in privacy preserving data mining," *Proceedings of ACM PODS Conference*, 2003
- [5] J. Wang, W.J. Zhong, J. Zhang and S.T. Xu, "Selective data distortion via structural partition and SSVD for privacy preservation," *Proceedings of the 2006 International Conference on Information & Knowledge Engineering*, pp:114-120, CSREA Press, Las Vegas, Nevada, USA, June 26-29, 2006
- [6] C. J. Lin, "Projected gradient methods for non-negative matrix factorization," <http://www.csie.ntu.edu.tw/~cjlin/papers/pgradnmf.pdf>
- [7] www.trackingthethreat.com

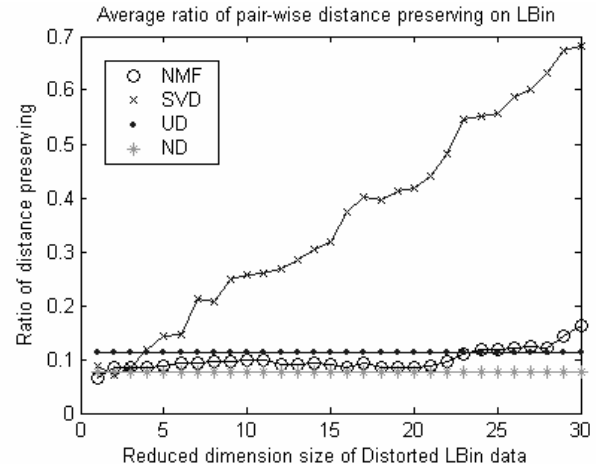


Fig. 3. Average distance-preserving comparison on NMF, SVD, ND and UD, to original data.

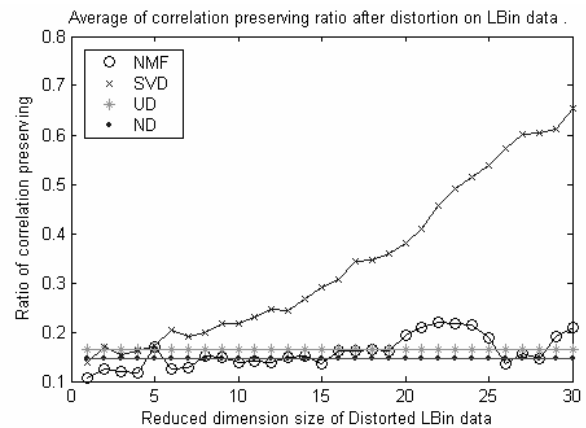


Fig. 4. Attribute correlation comparison on NMF, SVD, ND and UD, to the original data. For each method, the reduced dimensionality size is changed from 2 to 7.

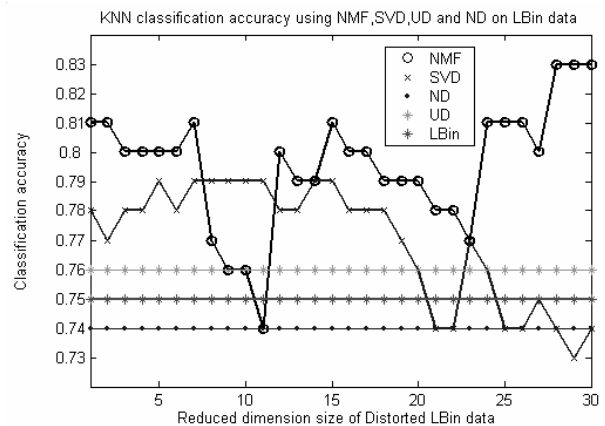


Fig. 5. Mining accuracy comparison on NMF, SVD, ND and UD. For each method, the reduced dimensionality size is changed from 2 to 7.