

NNMF-Based Factorization Techniques for High-Accuracy Privacy Protection on Non-negative-valued Datasets

Jie Wang and Jun Zhang

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

Abstract— Powerful modern access to a huge amount of various data having high or low level of privacy brings out a concurrent increasing demand for preserving data privacy. The challenge is how to protect attribute values without jeopardizing the similarity between data objects under analysis. In this paper, we further our previous work on applying matrix techniques to protect privacy and present a novel algebraic technique based on iterative methods for non-negative-valued data distortion. As an unsupervised learning method for uncovering latent features in high-dimensional data, a low rank nonnegative matrix factorization (NNMF) is used to preserve natural data non-negativity and avoid subtractive basis vector and encoding interactions present in techniques such as principal component analysis. It is the first in privacy preserving data mining in our paper that combining non-negative matrix decomposition with distortion processing. Two iterative methods to solve bound-constrained optimization problem in NMF are compared by experiments on Wisconsin Breast Cancer Dataset. The overall performance of NMF on distortion level and data utility is compared to our previously-proposed SVD-based distortion strategies and other existing popular data perturbation methods. Data utility is examined by cross validation of a binary classification using the support vector machine. Our experimental results on data mining benchmark datasets indicate that, in comparison with standard data distortion techniques, the proposed NMF-based method are very efficient in balancing data privacy and data utility, and it affords a feasible solution with a good promise on high-accuracy privacy preserving data mining.

Index Terms— non-negative matrix factorization, privacy, iterative method

I. INTRODUCTION

THE availability of large scale computing platforms and instrumentation for data collection have created extremely large data repositories that can be utilized through data mining

Technical Report No. 464-06, Department of Computer Science, University of Kentucky, Lexington, KY, 2006.

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, USA (phone: 859-257-9348; e-mail: jwanga@csr.uky.edu).

tools. A trade-off between sharing confidential information for analysis and keeping individual, corporate and countries privacy is a growing challenge. It motivated a great deal of research aimed to answer the following questions: how can data be exchanged securely for cooperative analysis or outsourcing analysis? How can important structure and underlying patterns be found within a large data set? How and when can this hidden structure be used to help predict missing data or to clean data that is imprecise or partially incorrect [1]? The increasing concern on privacy and related research brings out a new branch, known as privacy preserving data mining (PPDM).

The general goal of our work is defined as to hide to the outside world sensitive individual data, and simultaneously preserve the underlying data pattern and semantics so that the construction of a decision model on distorted data is enabled and it is equivalent to or even better than the model using the original data from the viewpoint of decision accuracy [2]. A desirable solution must consider not only privacy safeguards, but also accurate mining results.

The input in a data mining task can be represented by a vector space model, where a collection of records or objects is encoded as a $n \times m$ object-attribute matrix. Datasets of interest often lead to a very high dimensional matrix representation [3]. And it is observable that many real-world datasets have non-negative values for attributes. In fact, any of existing data distortion methods inevitably falls into the context of matrix computation. For instance, having the longest history in privacy protection area and by adding random noise to the data, additive noise methods can be viewed as a random matrix and therefore its properties can be understood by studying the properties of random matrices [4] [5].

Matrix decomposition in numerical linear algebra typically serves the purpose of finding a computationally convenient means for obtaining a solution to the original linear system. In contrast to its usage as a mechanism for obtaining another end, within the field of data mining, matrix decomposition also plays a very major role but usually not just for the purpose of solving systems of equations. In this context its major purpose is to obtain some form of simplified low-rank approximation to original dataset for understanding the structure of data, particularly the relationship

TABLE I SUMMARY OF NOTATION

Symbol	Meaning	Symbol	Meaning	SYMBOL	Meaning
A	Original matrix	$A(i,j)$	i,j^{th} entry of A	$A(i)$	i th row of matrix A
n	Number of rows	m	Number of columns	A_i	Submatrix of A
\tilde{A}	Approximation matrix of A	$A^{(k)}$	Rank- k approximation of matrix A	A^T	Transpose of A
$\ \cdot \ _F$	Frobenius-norm of a matrix	$\ \cdot \ _2$	2-norm of a vector	σ	Eigenvalue / singular value

within the objects and within the attributes and how the objects relate to the attributes [10]. Our work is the first to begin the study of matrix decomposition techniques on privacy-preserving data mining. A unique characteristic of matrix decomposition, a compact representation with reduced-rank while preserving dominant data patterns, stimulates our attempt on utilizing it to realize a two-win task both on high privacy and high accuracy.

In our previous paper [2][11], a set of hybrid methods that combines Singular Value decomposition (SVD) and sparsification strategies [12] was proposed. It has been experimentally proved that application of matrix decomposition techniques is one of feasible channels to better results on privacy protection and higher accuracy than additive noise methods for high accuracy privacy preservation classification. With our previous work on matrix decomposition in [6][7][8][9], we recently investigated on using matrix decomposition techniques to achieve the general goal claimed above: high-accuracy privacy preservation.

Our current study is carried out to continue previous research in [2][11], focusing on the context of classifying objects from large non-negative-valued datasets. For this framework, taking advantage of matrix theory and powerful computing capability of iterative methods, the main objective on target is to provide an efficient and flexible technique for an error-bounded approximation of non-negative-valued datasets. Our proposed method has two important aspects: (i) non-negative matrix factorization (NMF) is adapted to provide a least-square compression version of original datasets. (ii) By using iterative methods to solve the least-square optimization problem is provided an attractive flexibility for data administrators to tailor our solution according to their specific requirement.

The remaining part of the paper is organized as follows. Section II offers an overview of the related work and the application of matrix decomposition techniques in privacy preserving data mining. Iterative implementation of non-negative matrix factorization and description of the proposed algorithm are presented in Section III. Measurement on data privacy and data utility is defined in Section IV. Section V discusses performance of our method in terms of data privacy and data utility and makes a comparison with other privacy protection methods. Finally, we conclude our work with a summary of results and outlines of ongoing research in Section VI.

II. BACKGROUND AND RELATED WORK

Intuitively there are three ideas on disguising sensitive data. One is to transform original data into protected, publishable data

by data perturbation. An alternative to data perturbation is to generate a new dataset (synthetic dataset), not from the original data, but from random values that are adjusted in order to have the same feature pattern as the original data. A third possibility is to build a hybrid dataset as a mixture of a distorted one and a synthetic one [13]. Most methods in literature are based on element-wise random perturbation.

A. Matrix Approximation

In practical, problems often involve fifty or a hundred attributes. We might typically believe that each feature is useful for at least some of the discriminations; while we may doubt that each feature provides independent information, intentionally superfluous features have not been included. The most important is how classification accuracy depends upon the dimensionality and amount of training data and the second is the computational complexity of designing the classifier [14].

Therefore, finding an error-bounded low-rank approximation of real dataset is always an essential subtask. It can be very useful in many problems where distance computations and comparisons are needed, because in high dimensions distance computations are very slow and moreover it is known that the distance between almost all pairs of points is the same with high probability and almost all pairs of points are orthogonal [46].

The Johnson-Lindenstrauss lemma [16] 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 $(1 \pm \epsilon)$. In other words, a data set of n points can be embedded in a subspace of dimension $O(\log n)$ with little distortion on the pair-wise distances. Random projection and Singular Value Decomposition (SVD) are two recently popular dimensionality reduction techniques. Non-negative matrix factorization is used in our paper.

B. Non-negative matrix factorization (NNMF)

The idea of positive matrix factorization is developed by P. Paatero at the University of Helsinki, and to be popular in the computational science community [17]. Interest in positive matrix factorization increased when a fast algorithm for Non-negative Matrix Factorization (NNMF), based on iterative update, was developed by Lee and Seung [18], particularly as they were able to show that it produced intuitively reasonable factorizations for a face recognition problem, and NNMF facilitates the analysis and classification of data from image or sensor articulation databases made up of images showing a composite object in many

articulations, poses, or observation views. They also found NMF to be a useful tool in text data mining [19]. In the past few years, several papers have discussed NNMF techniques and successful applications to various databases where the data values are non-negative [20]. NNMF has recently been shown to be very useful technique in approximating high dimensional data where the data are comprised of non-negative components. [21] [22] [23] [24] [25] [26].

III. ITERATIVE DATA DISTORTION STRATEGY

Truncated singular value decomposition can be viewed as a weighted summation of rank-one approximations to a sequence of matrices. The associated weights are the corresponding singular values.

$$A^{(k)} = \sum_{i \leq k} \sigma_i U(i) V(i)^T \quad (1)$$

Our previous work shows that SVD is a good solution for data protection of high accuracy classification. However, a drawback is associated with extraction of singular vector of orthogonal decompositions. If the underlying data only consists of nonoverlapping, i.e. orthogonal patterns, SVD performs successfully. If patterns with similar strengths overlap, attributes contained in some of the previously discovered patterns are extracted from each pattern. In orthogonalizing the second vector with respect to the first vector, SVD introduces negative values into the second vector. There is no easy interpretation of these negative values in the context of most data mining tasks, and negative components contradict physical realities.

Considering the non-negative-valued characteristic of most datasets, a nonorthogonal decomposition that does not introduce negative values into the component vectors is required. In the paper, non-negative matrix factorization is used to distort original dataset.

NNMF is a vector space method to obtain a representation of data using non-negative constraints. These constraints can lead to a parts-based representation because they allow only additive, not subtractive, combinations of the original data. This is in contrast to techniques for finding a reduced dimensional representation based on SVD. [20]

A. Non-negative Matrix Factorization

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

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

The usual way to find W and H is by the following least-square optimization, which minimized the difference between A and WH :

$$\begin{aligned} \min_{W, H} \quad & f(W, H) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m (A(i, j) - (WH)(i, j))^2 \\ \text{subject to} \quad & W_{ia} \geq 0, H_{bj} \geq 0, \forall i, a, b, j. \end{aligned} \quad (3)$$

In optimization, inequalities upper- and lower-bounding variables are referred to as bound constraints. Hence (3) is a standard bound-constrained optimization problem. There are several methods to solve (3) in literature. Algorithms designed to approximate A generally begin by initial estimates of the matrices W and H , followed by alternating iterations to improve these estimates.

B. First Order Optimality Condition

The gradient of $f(W, H)$ at (W, H) can be expressed as two partial derivatives to elements in W and H respectively.

$$\begin{aligned} \nabla f(W, H) &= (\nabla_W f(W, H), \nabla_H f(W, H)) \\ \nabla_W f(W, H) &= (WH - A)H^T \\ \nabla_H f(W, H) &= W^T(WH - A) \end{aligned} \quad (4)$$

By Corollary in [27], at a solution (W, H) of the non-negative matrix factorization problem, it is necessary that

$$(A - WH)H^T \leq 0 \quad W^T(A - WH) \leq 0 \quad (5)$$

Therefore, (W, H) is a stationary point if and only if

$$\begin{aligned} W_{ia} &\geq 0, H_{bj} \geq 0, \\ \nabla_W f(W, H)_{ia} &\geq 0, \nabla_H f(W, H)_{bj} \geq 0, \\ W_{ia} \nabla_W f(W, H)_{ia} &= 0, H_{bj} \nabla_H f(W, H)_{bj} = 0, \forall i, a, b, j. \end{aligned} \quad (6)$$

C. Multiplicative update algorithm

Multiplicative update algorithm is the most popular approach by Lee and Seung [18]. The iterative algorithm updates each entry of W and H on each iterative step.

Algorithm 1 Multiplicative Update

1. Initialize W and H and scale the columns of W to unit norm. $W_{ia}^1 > 0, H_{bj}^1 > 0, \forall i, a, b, j$.
2. In each step, $k=1, 2, \dots$

$$W_{ia}^{k+1} = W_{ia}^k \frac{(A(H^k)^T)_{ia}}{(W^k H^k (H^k)^T)_{ia}}, \forall i, a. \quad (7)$$

$$H_{bj}^{k+1} = H_{bj}^k \frac{((W^{k+1})^T A)_{bj}}{((W^{k+1})^T W^{k+1} H^k)_{bj}}, \forall b, j. \quad (8)$$

It is proved that under the multiplicative update rules the distance $\|A - WH\|_F^2$ is monotonically non-increasing. This algorithm is a fixed-point type method. The overall cost of Algorithm 1 is $\#iterations \times O(nmk)$

D. Block Coordinate Descent Method

By block coordinate descent method in bound-constrained optimization by Bertsekas [28], we can update W^{k+1} on (W^k, H^k) and H^{k+1} on (W^{k+1}, H^k) alternatively.

Algorithm 2 Alternating Non-negative Least Squares

1. Initialize $W_{ia}^1 > 0, H_{bj}^1 > 0, \forall i, a, b, j$.
2. For $k=1, 2, \dots$

$$W^{k+1} \in \arg \min_w f(W, H^k) \quad (9)$$

$$H^{k+1} \in \arg \min_H f(W^{k+1}, H) \quad (10)$$

(9) and (10) are sub-problems. When on block of variables is fixed, each sub-problem is indeed the collection of several non-negative least square problems. In (10), the j th column of H^{k+1} is an optimal solution of

$$\begin{aligned} \min_{W, H} \quad & \|A^T(j) - W^{k+1}h\|^2 \\ \text{subject to} \quad & H_b \geq 0, b=1, \dots, k, \end{aligned} \quad (11)$$

Projected Newton's methods are suggested to solve each problem in (11). Solving (9) and (10) per iteration could cost a lot more than the simple update in Algorithm 1. Thus efficient methods to solve sub-problems are essential.

E. Alternating Non-negative Least Squares Using Projected Gradients [29]

Lin proposes two methods for NNMF by applying projected gradient method to solve non-negative least square problem in Algorithm 2 or directly minimize (3). (10) consists of m independent non-negative least square problems (11). In this method, they are solved together rather than separately in Algorithm 2 and (10) is rewritten as

$$\begin{aligned} \min_H \quad & \bar{f}(H) \equiv \frac{1}{2} \|A - WH\|_F^2 \\ \text{subject to} \quad & H_{bj} \geq 0, \forall b, j \end{aligned} \quad (12)$$

and an improved projected gradient method is used to solve (12).

This method leads to faster convergence than the popular multiplicative update method, and the overall cost is $\#iterations \times (O(nmr) + \#sub-iterations \times O(tmr^2 + tnr^2))$

Algorithm 3 An improved projected gradient method

1. Given $0 < \beta < 1, 0 < \sigma < 1$
2. Initialize any feasible X^1 and set $\alpha_0 = 1$.
3. For $k=1, 2, \dots$
 - a) Assign $\alpha_k \leftarrow \alpha_{k-1}$
 - b) If α_k satisfies

$$X^{k+1} = P[X^k - \alpha_k \nabla f(X^k)] \text{ where } \alpha_k = \beta^{t_k}, \text{ and } t_k$$

is the first non-negative integer t for which $f(X^{k+1}) - f(X^k) \leq \sigma \nabla f(X^k)(X^{k+1} - X^k)$

Then repeatedly increase it by $\alpha_k \leftarrow \alpha_k \cdot \beta$ until α_k satisfies.

- c) Set $X^{k+1} = X(\alpha_k)$
-

F. Iteration Stopping Conditions

Without a predefined stopping condition, iterative steps will continue forever. A limitation on running time or the number of iteration is often used to interrupt the infinite loop. The difference between two recent iterations can also be a stop condition. If the difference is small enough, then the loop stops. Such a stopping condition does not reveal whether a solution is close to a stationary point or not. However, in the context of data distortion, we do not need an accurate factorization. We only require a sparse low-rank non-negative approximation of the original matrix. In our method, a requirement on privacy level is integrated as a stopping condition to the iterative procedure. In practical, with a comprehensible consideration on data dimension, privacy level and accuracy level, and computation cost, determining a simple limitation on norm, running time or number of iteration by trial-and-error is an economical way.

G. Iterative NNMF Data Distortion Method

Our proposed method consists of three parts: Initialization, Iterative Loop and Distortion. Each part includes several steps detailed in Algorithm 4.

IV. EVALUATION MEASURES

The issue here to address is to hide to the outside world sensitive individual data while retaining the underlying data pattern and semantics to enable a construction of an accurate decision model on the distorted data. Therefore two categories of evaluation metrics are suggested here to perform evaluation of the new method.

A. Data distortion measures

The privacy protection measure is used to evaluate dissimilarity between the original and distorted data. It should indicate how closely the original value of an item can be estimated from the distorted data. Some privacy metrics have been proposed in the literature. Some data distortion measures defined in [2] are used here to assess the level of data distortion which only depends on the original matrix A and its distorted counterpart \tilde{A} .

1) Value Difference (VD)

After a data matrix is distorted, the value of its elements changes. The value difference (VD) of the datasets is represented by the relative value difference in the Frobenius norm. Thus VD is the ratio of the Frobenius norm of the difference of \tilde{A} from A to the Frobenius norm of A :

$$VD = \frac{\|A - \tilde{A}\|_F}{\|A\|_F} \quad (13)$$

2) Position Difference

After a data distortion, the order of the value of the data elements changes, too. We use several metrics to measure the position difference of the data elements. Rank Position (RP), Rank Maintenance (RM), Change of Rank of Attributes (CP),

Algorithm 4 Iterative NNMF Data Distortion Method

Input : $A_{n \times m}$: non-negative matrix,

k : size of dimension

$tol(i)$: limit values of errors and stopping conditions

Output : W , H : two factor matrices.

r : the final reduced dimension.

$\tilde{A}^{(r)}$: the final distorted dataset.

Initialization:

1. Preprocessing the original dataset $A_{n \times m}$
2. Examine its non-negative property;
3. Set up stopping condition: S
4. Set up dimension value $k < \min\{n, m\}$
5. Randomly generate initial estimate of non-negative matrices $(W_{n \times k}^{(0)}, H_{k \times m}^{(0)})$.

Iterative Loop:

6. Compute initial value of stopping condition, $S^{(0)}$
7. For each iteration $i=0,1,\dots$ until stopping condition satisfied, Do
8. Compute $(W_{n \times k}^{(k+1)}, H_{k \times m}^{(k+1)}) = NNMF_algorithm(W_{n \times k}^{(k)}, H_{k \times m}^{(k)})$
9. Compute $S^{(k+1)}$
10. If $S^{(k+1)}$ satisfies stopping condition,
11. Output W and H ;
12. Stop;
13. EndIf
14. EndDo

Distortion :

15. Compute approximation $\tilde{A} = WH$
 16. Choose an integer $r < k$
 17. For $r = k, k-1, \dots, 1$, Do
 18. Do further distortion: $\tilde{A}^{(r)} = W_{n \times r} H_{r \times n}$
 19. Compute privacy metrics on $\tilde{A}^{(r)}$
 20. Train classifier on $\tilde{A}^{(r)}$ and compute classification accuracy.
 21. EndDo
 22. Choose one $\tilde{A}^{(r)}$ with satisfied privacy level and accuracy
 23. Publish the final distorted dataset $\tilde{A}^{(r)}$
-

and Maintenance of Rank of Attributes (CK) are used in our experiments. Detailed definition and calculation are described in [2].

B. Data Utility Measure

Data utility measures indicate the accuracy of data mining algorithms on distorted data after the manipulation of certain perturbation. In this paper, Support Vector Machine (SVM) classification is chosen as the data utility measure by building a classifier on distorted dataset and applying five-fold cross validation method to compute classification accuracy as a reasonable data utility measure. [30]

V. EXPERIMENTS AND RESULTS

The experiments here are designed in three steps: dataset creation, data distortion and measurement calculation. A real non-negative-value dataset is used in our experiments to examine the performance of the proposed new data distortion strategies and compare with our previous proposed strategies. All implementations of NNMF are available at <http://www.csie.ntu.edu.tw/~cjlin/nmf>. Experiments are conducted on a Sunblade 150 workstation.

A. Real Dataset Descriptions

A large number of datasets from different application domains (such as financial, medical, scientific, demographic, military environments) can be used to identify the performance of our proposed approach. An example of such dataset is the Wisconsin Breast Cancer (WBC) dataset available from the UCI Machine Learning Repository at <http://www.ics.uci.edu/~mllearn/MLRepository.html>.

TABLE II

ATTRIBUTE INFORMATION OF WBC DATASET		
Number	Attribute	Domain
1	Sample code number	Id number
2	Clump Thickness	1 - 10
3	Uniformity of Cell Size	1 - 10
4	Uniformity of Cell Shape	1 - 10
5	Marginal Adhesion	1 - 10
6	Single Epithelial Cell Size	1 - 10
7	Bare Nuclei	1 - 10
8	Bland Chromatin	1 - 10
9	Normal Nucleoli	1 - 10
10	Mitoses	1 - 10
11	Class :	2 for benign, 4 for malignant
Class distribution:		Benign: 458 (65.5%)

The original version is used here, which consists of 699 instances, 10 integer-valued attributes and one class attribute. And there are 16 missing attribute values for Bare Nuclei. Table II is a description on WBC original version. Some modifications on the original WBC dataset are performed to make it suitable for our tests, which include:

1. Remove Sample code number attribute and select the other 9 attributes.
2. Class label: replace 2 with 1, and 4 with -1

- Fill in the missing values of Bare Nuclei using the following rule:

$$\text{The missing vale of Bare Nuclei} = \begin{cases} 1, & \text{class label is benign} \\ 8, & \text{classlabel} = \text{malignant} \end{cases}$$

The target WBC dataset is a 699 by 10 matrix with the 10th column representing class label.

B. Notation Description

Notations in experiments are described in Table III.

C. Default Value of Experimental Parameters

For all the experiments here, the default values of some

TABLE III
EXPERIMENT NOTATION SUMMARY

Notation	Description
WBC	Wisconsin Breast Cancer Dataset: [699*9]
UD	Uniformly-noise-added method
ND	Normally-noise-added method
SVD	Singular-value-decomposition method
NMF	Non-negative matrix factorization using Alternating non-negative least squares by projected gradients
NMFM	Non-negative matrix factorization by multiplicative update
SSVD	Sparsified SVD method using STS strategy
CSVD	Sparsified SVD method using CTS strategy
ESVD	Sparsified SVD method using ETS strategy
SNMF	Sparsified NMF method using STS strategy
CNMF	Sparsified NMF method using CTS strategy
ENMF	Sparsified NMF method using ETS strategy

parameters in distortion methods are listed as follows:

- NMF: tolerance for stopping condition $\text{tol}=1e-4$, time limit = 4000, iteration number limit = 20000.
- ND: the normally distributed noise is generated with $u = 0$ and $\sigma = 0.46$, see [2] for the meaning of these two parameters.
- UD: the uniformly distributed noise is generated from the interval $[0, 0.8]$.
- STS sparsification: the threshold value $\varepsilon = 0.001$
- CTS sparsification: $\varepsilon = 0.2$
- ETS sparsification: $\varepsilon = 0.01$, $\alpha = 0.2$.
- SVM classification: radial base function (RBF) is chosen as the kernel function and $\gamma = 0.001$.

D. Comparison of two iterative NMF algorithms: Experiment 1

The two NMF algorithms are implemented on WBC to compare the performance. One is multiplicative update in Algorithm 1, denoted by NMFM. The other is alternating projected gradients for each sub-problem, denoted by NMF. The problem size $(n,k,m)=(699,7,9)$. All the tests are share the same initial estimate of $(W_{699 \times 7}^{(0)}, H_{7 \times 9}^{(0)})$. The tolerance is set to be 10^{-3} , 10^{-4} , 10^{-5} and 10^{-6} in order to examine convergence speed. We also impose a time limit of 4000 seconds and a maximal number of 50000 iterations on each method.

Table IV shows that when tolerance is 10^{-5} , NMFM often exceed the iteration limit of 50000. Obviously NMF is superior to

NMFM. The data in the succeeding experiments are collected by using NMF algorithm in iterations.

E. Performance of NMF Algorithm using Projected Gradients: Experiment 2

An initial random guess on W and H is the first step in the beginning of iteration. Different start value leads to different initial gradient norm. Therefore, the result and iteration time are dependent on the initial guess. The computation cost are roughly examined on dimension value from 9 to 1 under the tolerance is $1e-4$.

TABLE V PERFORMANCE OF NMF ALGORITHM

dimension	Initial Gradient Norm	Iteration Times	Iteration Time(seconds)
9	16525	83	12.41
8	11584	94	7.44
7	10648	80	7.38
6	7499	109	8.84
5	4816	117	7.85
4	5196	128	9.2
3	3265	76	4.65
2	4312	20	0.52

F. Sparseness Level of W and H: Experiment 3

NNMF factorization makes two submatrices with higher sparseness than those by singular value decomposition. In the experiment, sparseness of a vector x of length n is defined as

$$\text{sparseness}(x) = \frac{\sqrt{n} - \|x\|_1 / \|x\|_2}{\sqrt{n} - 1}$$

To measure sparseness of a matrix, we stack columns of the matrix to form a vector. The maximal of sparseness of x is 1 if containing $n - 1$ zeros, and it reaches zero if the absolute value of all coefficient of x coincide.

Fig.1. and Fig.2. illustrate the bar plots of W and H created by NMF algorithm WBC with $k=7$ and tolerance= 10^{-4} . The sparseness of W and H are 0.34 and 0.64 respectively. More than 50% of entries in H are zeros. The algorithms to solve W and H used in our method make H sparser in preference to W . Hence, in the natural interpretation of the factorization, H is the basis or factor vectors and it tends to be sparse. Implicitly this suggests that the basis will involve only some of the original attributes. While that W is denser than H implies the objects are combinations of all of basis.

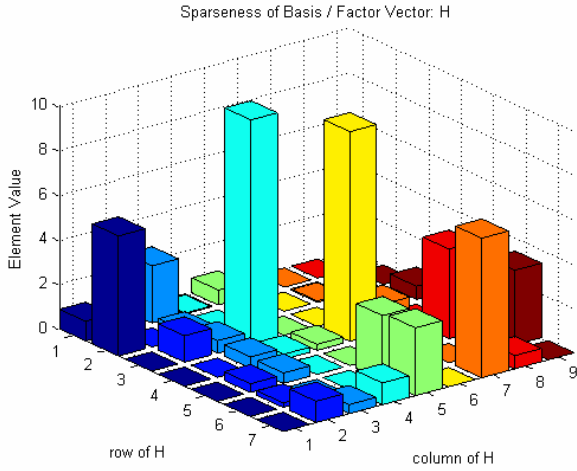


Fig. 1. Sparseness of Basis / Factor Vectors H, created by NMF algorithm on WBC with $k=7$ and tolerance= 10^{-4} . Sparseness of H is 0.64

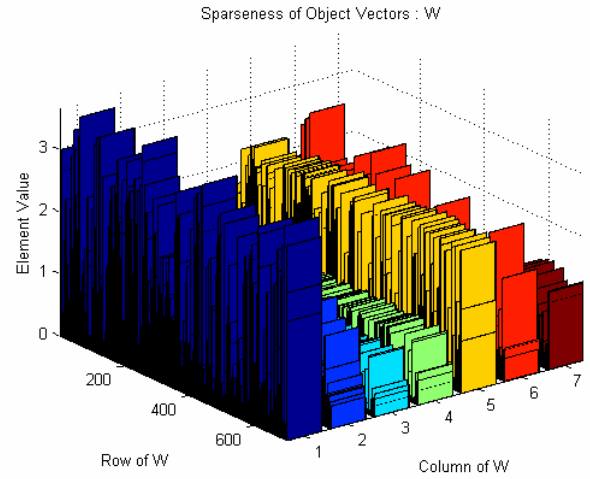


Fig. 2. Sparseness of Object Vectors W, created by NMF algorithm on WBC with $k=7$ and tolerance= 10^{-4} . Sparseness of W is 0.34.

TABLE VI
COMPARISON OF DIFFERENT DISTORTION STRATEGIES ON WBC

Methods	Level of Distortion					Accuracy (%) (classification)
	VD	RP	RK	CP	CK	
WBC	-	-	-	-	-	96.4
UD	0.1085	219.6993	0.0130	0	1	96.4
ND	0.1098	224.8148	0.0084	0	1	96.3
SVD	0.1222	228.8972	0.0114	0.2222	0.7778	96.4
NMF	0.1228	228.4295	0.0100	0.2222	0.7778	96.7
SSVD	1.2662	228.1370	0.0013	3.3333	0	96.6
CSVD	1.2702	230.1561	0.0021	3.3333	0	96.4
ESVD	1.2704	228.0744	0.0014	3.3333	0	96.4
SNMF	0.1228	228.4362	0.0076	0.2222	0.7778	96.4
CNMF	0.1297	226.5042	0.0081	0.2222	0.7778	96.5
ENMF	0.1234	228.2035	0.0089	1.1111	0.5556	96.5

Note: Parameters: SVD: $\kappa=7$, NMF: $\kappa=7$ and $r=7$. The value of VD is adjusted as close as possible for UD, ND, SVD and NMF, in order to make a fair comparison.

G. Comparison of Iterative NNMf Data Distortion Strategies with SVD, UD and ND on WBC: Experiment 4

The ten distortion methods, NNMf-based, uniformly distributed noise (UD), normally distributed noise (ND), SVD, SSVD, SSVD with matrix partition, are implemented on WBC to compare the performance. In order to be fair in comparing the privacy metrics, parameters are set to such certain values as to make VD values of UD, ND, SVD and NMF as close as possible. Rank κ of SVD is 7. Dimension size in NMF is 7 and final

dimension is also 7. The results of performance evaluation on ten methods are provided in Table VI and Fig.3.

Under the premise on the same level of value dissimilarity, the fact that CP value of UD and ND is 0 and CK value be 1 indicate that additive noise methods are worse than matrix-decomposition-based methods.

Experimental data in Table VI supports the following conclusions

1. NNMf-based distortion strategies achieve a comparable performance with SVD-based strategies. In particular, it achieves the highest classification accuracy.

TABLE IV
PERFORMANCE COMPARISON OF TWO NMF ALGORITHM

Tolerance	Number of Iteration		Iteration Time (seconds)		Final Gradient Norm		Objective Values	
	NMF	NMFM	NMF	NMFM	NMF	NMFM	NMFM	NMF
1e-3	17	3060	0.8	2.6	1.04	7.11	41.4	41.5
1e-4	94	20000	3.6	23.1	0.09	1.54	41.3	41.4
1e-5	386	50000	9.8	49.7	0.01	0.84	41.4	41.5
1e-6	2382	-	63.3	-	0.001	-	41.4	-

Initial objective value: 276.2; Initial Gradient Norm: 7609.7; dimension:7;. When tolerance is greater than 1e-5, number of iteration of NMFM exceeds the prescribed limit.

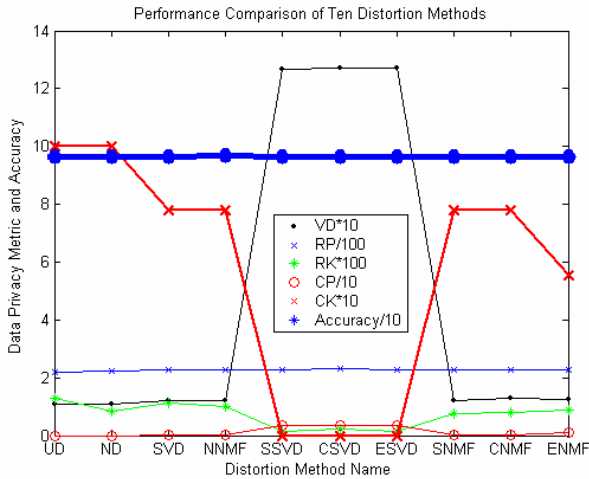


Fig. 3. Comparison of Overall Performance of Ten Distortion Methods. The uppermost blue bold line shows classification accuracy.

2. No improvement on performance of NNMF-based methods by applying sparsification strategies. It is reasonable under the condition that NNMF is a sparse factorization and two factors, W and H , has a deep level of sparseness. Thus, further sparseification does not provide any improvement.
3. Sparsified SVD performs best on privacy level without any degradation on data mining accuracy. It is obvious that sparsification has a strong effect on data privacy level of SVD-based methods by making all the attributes change their rank in average value because CK value is 0.
4. As to data utility, all the ten methods achieve a level at least not worse than the original dataset.

H. Sensibility of Performance on Dimension of NNMF: Experiment 5

To examine the effect of dimension size on privacy level and data utility level in the approximation, we conduct the experiment on WBC and Fig.2 illustrates the influence of dimension size on privacy level and classification accuracy. Here W and H are solved under a dimension of 7. Then the final compressed approximation of WBC is computed by setting up r from 6 to 2.

Dimension size is a key element both for dimension reduction and privacy level. The less dimension size, the higher privacy level of the method is. However, clearly, dimension size is inversely proportional to data utility level. Fig.4. illustrates the above relationship. How to choose dimension size in the proposed method is an empirical problem. For WBC, our experiments imply one possible good choice for our distortion method both considering data utility and data privacy. When the initial dimension size is 7, we can choose 4 as a reasonable size.

VI. SUMMARY

Experiment results indicate that by a careful choice of iterative parameter settings, two sparse non-negative factors can be solved by some efficient iterative algorithms. Alternating least square using projected gradients in computing NNMF converges faster

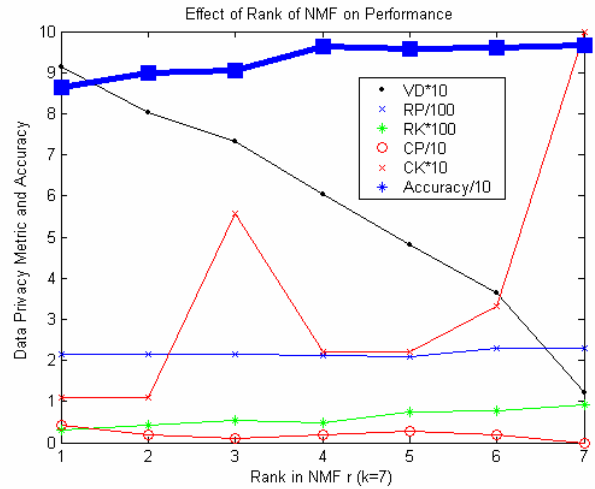


Fig. 4. Effect of Dimension Size on Performance of Privacy and Accuracy. The dimension size in NMF is 7. The uppermost line shows classification accuracy with respect to dimension size

than multiplicative update methods. The value of these two matrices is not unique because it is dependent on initial estimates in the beginning of iterative procedure. This dependency provides our method both with uncertainty and flexibility. For non-negative-valued datasets, our proposed method provides a possibility of simultaneously achieve satisfactory privacy, accuracy and efficiency. With the same level of privacy as other data distortion methods, the method demonstrates the highest classification accuracy. In particular, we foresee that using iterative factorization of original datasets, there is an opportunity where all these triple goals can reach an above-average point.

VII. FUTURE WORK

In this paper, for the first time, we have considered high accuracy privacy preserving of non-negative-valued datasets. The important property of non-negative matrix factorization, non-negative and sparseness, makes it not only a good dimension reduction technique but also an efficient privacy preserving tool. The satisfied performance of the proposed method on our global target further inspires our future work emphasizing on matrix techniques.

With the powerful underpinning of deeply-rooted matrix decomposition theory and well-developed iterative computing methods, it is possible to do some modification on traditional NNMF. Our plans are thus to continue the study of NNMF on privacy preserving data mining. By matrix multiplication, we know that any matrix decomposition remains unchanged if the factor matrix is multiplied by an arbitrary invertible matrix, and the coordinate matrix is multiplied by its inverse. This corresponds to a rotation of the axes of the new space. It is useful to apply such a rotation after the decomposition has been computed in order to make the low-dimension approximation sparser. Currently, modifying NNMF convergence conditions and embedding special requirements into the iterative loop of NNMF are of our interests. In the iterative process, initialization of the factors and uncertainty on final value of factors also imply a wide space to delve into the more potential of non-negative matrix

factorization on high accuracy privacy distortion.

Our study is one part of a broader effort to explore far-reaching significance of matrix techniques, particularly matrix decomposition, on high accuracy privacy distortion. Iterative NMF-based distortion method is a good solution for data mining problems on the basis of discriminant functions. However it seems not preserve dissimilarity among objects. Therefore, it might not be applicable for distance-based mining problems, such as k-means clustering and k-nearest-neighbor classification. The most recent task of our work is going to conduct extensive experiments of NMF-based and SVD-based distortion method on high dimension real datasets and different data mining algorithms. The exposure of any poor performance or disadvantage would be highly valuable on directing our future research.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for their valuable comments and suggestions on the original manuscript.

REFERENCES

- [1] Frank McSherry, "Spectral Methods for Data Analysis", Ph.D dissertation, University of Washington, 2004.
- [2] J. Wang, W.J. Zhong, J. Zhang and S.T. Xu, "Selective Data Distortion via Structural Partition and SSVD for Privacy Preservation," In Proceedings of the 2006 International conference on Information & Knowledge Engineering, pp: 114 - 120, CSREA Press, Las Vegas, Nevada, USA, June 26-29, 2006
- [3] D. Achlioptas, "Random matrices in data analysis," [online]. Available: <http://www.cs.ucsc.edu/~optas/papers/matrices.pdf>
- [4] H. Kargupta, K. Sivakumar and S. Ghosh, "Dependency detection in mobimine and random matrices," In Proceedings of the 6th Europe Conference on Principles and Practices of Knowledge Discovery in Databases, pp.250-262, 2002
- [5] ML. Mahta, Random matrices, 2nd edition. Academic, London, 1991
- [6] J. Gao and J. Zhang, "Clustered SVD strategies in latent semantic indexing," Information Processing and Management, vol.41, no.5, pp:1051-1063, 2005.
- [7] S. Xu and J. Zhang, "A new data mining approach to predicting matrix condition numbers," Commun. Inform. Systems, vol.4, no.4, pp:325-340, 2004
- [8] S. Xu and J. Zhang, "A data mining approach to matrix preconditioning problem," in Proceedings of the 8th Workshop on Mining Scientific and Engineering Databases (MSD05), in conjunction with the 5th SIAM International Conference on Data Mining, Newport Beach, CA, Apr.2005
- [9] S. Xu and J. Zhang, "SVM classification for predicting sparse matrix solvability with parameterized matrix preconditioners," Technical report No. 458-06, Department of Computer Science, University of Kentucky, 2006
- [10] L. Hubert, J. Meulman and W. Heiser, "Two purposes for matrix factorization: a historical appraisal," SIAM Review, vol.42, no.4, pp:68-82, 2000
- [11] S. T. Xu, J. Zhang, D. Han and J. Wang, "A singular Value Decomposition Based Data Distortion Strategy for Privacy Protection," Accepted for publish and in press. Knowledge and Information Systems (KAIS) journal, 2006
- [12] J. Gao J. Zhang, "Sparsification strategies in latent semantic indexing," In Proceedings of the 2003 Text Mining Workshop, M.W. Berry and W.M. Pottenger, editor. San Francisco, CA, pp:93-103, 2003
- [13] J. M. Mateo-Sanz, A. M. Balleste and J. D. Ferrer, "Fast generation of accurate synthetic microdata," In J. Domingo-Ferrer and V. Torra, editors, Privacy in Statistical Databases, vol.3050 of LNCS, pp:298-306, Berlin Heidelberg, 2004.Springer.
- [14] R. O. Duda and P. E. Hart, D. G.Stork, Pattern Classification, 2nd edition, John Wiley&Sons. 2001
- [15] F. N. Afrati, "On approximation algorithms for data mining applications," National Technical University of Athens, Greece. Jun.2004
- [16] W. Johnson and J. Lindenstrauss, "Extensions of lipschitz mapping into Hilbert space," Contemporary Mathematics, vol.26, pp:189-206, 1984
- [17] M. Juvela, K. Lehtinen and P. Paatero. "The use of positive matrix factorization in the analysis of molecular line spectra from the thumbprint nebula," In Proceedings of the Fourth Haystack Conference "Clouds; cores and low mass stars", Astronomical Society of the Pacific Conference Series, vol.65, pp:176-180, 1994
- [18] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," In NIPS, Neural Information Processing Systems, pp:556-562, 2000
- [19] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol.401, pp:788-791, 1999
- [20] .V. P. Pauca, F. Shahnaz, M. W. Berry and R. J.Plemmons, "Text mining using non-negative matrix factorizations," In Proceedings of the 4th SIAM International Conference on Data Mining, pp:452-456, 2004
- [21] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," preprint, Department of Statistics, Stanford University, 2003.
- [22] J. T.Giles, L. Wo and M. W. Berry. "GTP(general text parser) software for text mining," in Statistical Data Mining and Knowledge Discovery, H.Bozdogan(Ed.), CRC Press, Boca Raton, pp:455-471, 2003
- [23] D. Guillaumet and J. Vitria, "Determining a suitable metric when using non-negative matrix factorization," 16th International Conference on Pattern Recognition (ICPR'02), vol.2, Quebec City, QC, Canada, 2002
- [24] P. Hoyer, "Non-negative sparse coding," Neural Networks for Signal Processing XII (Proc. IEEE Workshop on Neural Networks for Signal Processing), Martigny, Switzerland, 2002
- [25] W. Liu and J. Yi, "Existing and new algorithms for non-negative matrix factorization," preprint, Computer Sciences Dept., UT Austin, 2003
- [26] S. Wild, J. Curry and A. Dougherty, "Motivating non-negative matrix factorizations," In Proceedings of the 8th SIAM Conference on Applied Linear Algebra, Williamsburg, VA, Jul.15-17, 2003. Online Available, <http://www.siam.org/meetings/la03/proceedings/>
- [27] M. Chu, F. Fiele et.al, "Optimality, computation and interpretation of nonnegative matrix factorizations," Available online, Oct. 2004
- [28] D. P. Bertsekas, "Nonlinear Programming," Athena Scientific, Belmont, MA 02178-9998, second edition, 1999
- [29] C. J. Lin, "Projected gradient methods for non-negative matrix factorization," Available online, <http://>
- [30] T. Joachims, Making large-scale SVN learning practical. Advances in Kernel Methods – Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999