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

Jie Wang[1, 2] Weijun Zhong[1] Jun Zhang[2]
[1]Department of Management Sciences and Engineering, Southeast University, Nanjing, 210096, P.R. China, jwanga@csr.uky.edu, zhongweijun@seu.edu.cn
[2]Laboratory for High Performance Scientific Computing and Computer Simulation, Department of Computer Sciences, University of Kentucky, USA, jzhang@cs.uky.edu

***Abstract*— The challenge in preserving data privacy 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

A trade-off between sharing confidential information for analysis and keeping individual, corporate and countries privacy motivated a great deal of research aimed to 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 [1].

Many real-world datasets have non-negative values for attributes. In our previous paper [1][2], a set of hybrid methods that combines Singular Value decomposition (SVD) and sparsification strategies [3] 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.

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.

With our previous work on matrix decomposition in [4][5][6][7], our current study is carried out to continue previous research in [1][2], 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.

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 [8]. The idea of positive matrix factorization is developed by P. Paatero at the University of Helsinki, and to be

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 4056, USA (phone: 859-257-9348; e-mail: jwanga@csr.uky.edu).

popular in the computational science community [9]. 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 [10], particularly as they were able to show that it produced intuitively reasonable factorizations for a face recognition problem. NNMF has recently been shown to be very useful technique in approximating high dimensional data where the data are comprised of non-negative components. [11] [12] [13] [14] [15] [16]. 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. [17]

III. ITERATIVE DATA DISTORTION STRATEGY

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

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

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

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

Multiplicative update algorithm is the most popular approach by Lee and Seung [10]. The iterative algorithm updates each entry of W and H on each iterative step. This algorithm is a fixed-point type method. The overall cost of Algorithm 1 is $\#iterations \times O(nmk)$. By block coordinate descent method in bound-constrained optimization by Bertsekas [18], we can update W^{k+1} on (W^k, H^k) and H^{k+1} on (W^{k+1}, H^k) alternatively.

Algorithm 1 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,…

$$W^{k+1} \in \arg\min_{w} f(W, H^k) \quad (3)$$

$$H^{k+1} \in \arg\min_{H} f(W^{k+1}, H) \quad (4)$$

$$\min_{W,H} \quad \left\|A^T(j) - W^{k+1}h\right\|^2 \quad (5)$$
$$\text{subject to} \quad H_b \geq 0, b = 1,....k,$$

B. Alternating Non-negative Least Squares Using Projected Gradients [19]

Lin proposes two methods for NNMF by applying projected gradient method to solve non-negative least square problem in Algorithm 2 or directly minimize (2). (4) consists of m independent non-negative least square problems (5).

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

Algorithm 2 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,…
   a) Assign $\alpha_k \leftarrow \alpha_{k-1}$
   b) If α_k satisfies
      $X^{k+1} = P[X^k - \alpha_k \nabla f(X^k)]$ where $\alpha_k = \beta^{t_k}$, 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 $\alpha_k \leftarrow \alpha_k.\beta$ until α_k satisfies.
   c) Set $X^{k+1} = X(\alpha_k)$

C. 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 3. 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..

Algorithm 3 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 $i = 0,1,.....$ until stopping condition satisfied, Do
8. Compute $(W_{n\times k}^{(k+1)}, H_{k\times m}^{(k+1)}) = NNMF\_algoithm(W_{n\times k}^{(k)}, H_{k\times m}^{(k)})$
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 $r = k, k-1,.....1,$ 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 $\tilde{A}^{(r)}$

IV. 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. Wisconsin Breast Cancer (WBC) dataset is used here. http://www.ics.uci.edu/~mlearn/MLRepository.html. 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. 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
3. Fill in the missing values of Bare Nuclei using the following rule:

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

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

Notations in experiments are described in Table I.

TABLE I
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

A. Default Value of Experimental Parameters

For all the experiments here, the default values of some parameters in distortion methods are listed as follows:

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

B. Evaluation measures

Some data distortion measures defined in [1] are used here to assess the level of data distortion which only depends on Value Difference (VD). Rank Position (RP), Rank Maintenance (RM), Change of Rank of Attributes (CP), and Maintenace of Rank of Attributes (CK) are used in our experiments. Detailed definition and calculation are described in [2]. 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]

C. Comparison of two iterative NMF algorithms: Experiment 1

The two NMF algorithms are implemented on WBC to compare the performance. One is multiplicative update 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.

D. 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 II 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

E. 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

$$sparseness(x) = \frac{\sqrt{n} - \|x\|_1 / \|x\|_2}{\sqrt{n} - 1} \quad (20)$$

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.

By applying NMF algorithm to 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.

F. 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 IV and Fig.1.

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

TABLE III
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.

TABLE IV
COMPARISON OF DIFFERENT DISTORTION STRATEGIES ON WBC

Methods	*Level of Distortion*					*Accuracy (%)*
	VD	**RP**	**RK**	**CP**	**CK**	(classification)
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: κ=7, NMF: κ=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.

performance with SVD-based strategies. In particular, it has the highest 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.

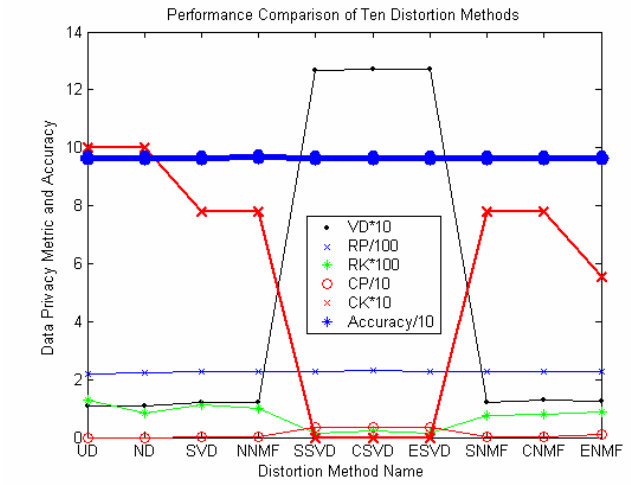

Fig. 1. Comparison of Overall Performance of Ten Distortion Methods. The uppermost blue bold line shows classification accuracy.

V. 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 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.

Iterative NMF-based distortion method is a good solution for data mining problems on the basis of discriminant functions. 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.

REFERENCES

[1] 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

[2] 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

[3] 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

[4] J. Gao and J. Zhang, "Clustered SVD strategies in latent semantic indexing," Information Processing and Management, vol.41, no.5, pp:1051-1063, 2005.

[5] 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

[6] 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

[7] 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

[8] 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.

[9] 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 Fouth Haystack Conference "Clouds; cores and low mass stars", Astronomical Society of the Pacific Conference Series, vol.65, pp:176-180, 1994

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," In NIPS, Neural Information Processing Systems, pp:556-562, 2000

[11] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," preprint, Department of Statistics, Stanford University, 2003.

[12] 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

[13] D. Guillamet 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

[14] P. Hoyer, "Non-negative sparse coding," Neural Networks for Signal Processing XII (Proc. IEEE Workshop on Neural Networks for Signal Processing), Martigny, Switzerland, 2002

[15] W. Liu and J. Yi, "Existing and new algorithms for non-negative matrix factorization," preprint, Computer Sciences Dept., UT Austin, 2003

[16] 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/

[17] .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

[18] D. P. Bertsekas, "Nonlinear Programming," Athena Scientific, Belmont, MA 02178-9998, second edition, 1999

[19] C. J. Lin, "Projected gradient methods for non-negative matrix factorization," Available online

[20] 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