EXPLOITING DISCRIMINANT AND SVM CONSTRAINTS IN NMF

Olga Zoidi, Anastasios Tefas, Ioannis Pitas

Department of Informatics
Aristotle University of Thessaloniki
Box 451, Thessaloniki 54124, GREECE
{ozoidi,tefas,pitas}@aiia.csd.auth.gr

ABSTRACT

A novel method is introduced for exploiting the support vector machine and additional discriminant constraints in nonnegative matrix factorization. The notion of the proposed method is to find the projection matrix that projects the data to a low-dimensional space so that the data projections have minimum within-class variance, maximum between-class variance and the data projections between the two classes are separated by a hyperplane with maximum margin. Experiments were performed on several two-class UCI data sets, as well as on the Cohn-Kanade database for facial expression recognition. Experimental results showed that the proposed method achieves better classification performance than the state of the art nonnegative matrix factorization and discriminant nonnegative matrix factorization followed by support vector machines classification.

Index Terms— Non-negative Matrix Factorization, Support Vector Machines, Joint Optimization, Maximum Margin Classification

1. INTRODUCTION

Given the nonnegative matrix $\mathbf{X} \in \Re^{N \times M}$, nonnegative matrix factorization (NMF) [1] searches for a pair of nonnegative matrices $\mathbf{Z} \in \Re^{N \times L}$, $\mathbf{H} \in \Re^{L \times M}$ whose product approximates $\mathbf{X}$, i.e., the objective of NMF is the minimization of the reconstruction error:

$$\arg \min_{\mathbf{Z},\mathbf{H}} \{D(\mathbf{X} \| \mathbf{ZH})\} \tag{1}$$

$$\text{subject to } z_{il} \geq 0, h_{lj} \geq 0. \tag{2}$$

$\mathbf{X}$ is the data matrix, whose column $\mathbf{x}_j$, $j = 1 \ldots M$, represents the j-th element vector of dimension N. $\mathbf{Z}$ is a basis matrix, that projects the data to a space with dimensionality L. By setting $L << N$ data dimensionality reduction is achieved. Finally, $\mathbf{H}$ is the matrix of the data projections. In discriminant NMF (DNMF) [2] additional discriminant constraints to the cost function of NMF are incorporated, by minimizing the Fisher criterion:

$$\arg \min_{\mathbf{Z},\mathbf{H}} \{D(\mathbf{X} \| \mathbf{ZH}) + \gamma \text{tr}[\mathbf{S}_w] - \delta \text{tr}[\mathbf{S}_b]\} \tag{3}$$

$$\text{subject to } z_{il} \geq 0, h_{lj} \geq 0, \tag{4}$$

where $\mathbf{S}_w$ and $\mathbf{S}_b$ are the within-class and between-class scatter matrices of the projected data, respectively,

$$\mathbf{S}_w = \sum_{c=1}^{C} \sum_{j=1}^{M_c} (\mathbf{h}_j^c - \bar{\mathbf{h}}_c)(\mathbf{h}_j^c - \bar{\mathbf{h}}_c)^T \tag{5}$$

$$\mathbf{S}_b = \sum_{c=1}^{C} M_c (\bar{\mathbf{h}}_c - \bar{h})(\bar{\mathbf{h}}_c - \bar{\mathbf{h}})^T, \tag{6}$$

where C is the number of classes, M_c is the cardinality of class c, $\mathbf{h}_j^c$ denotes the projected data of class c, $\bar{\mathbf{h}}_c$ is the mean vector of class c, $\bar{\mathbf{h}}$ is the mean vector of all classes and tr$[\cdot]$ denotes the trace operator.

After NMF (or DNMF) is performed, support vector machines (SVMs) [3] are employed on the projected data for classification. The objective of SVM is to find the maximum-margin hyperplane, i.e., the hyperplane whose distance from the nearest data of each class is maximal. The elements, whose removal from the training data set change the maximum-margin hyperplane are called support vectors.

Apart from DNMF, several other modifications of NMF and SVM exist, that impose additional constraints to the objective functions of NMF and SVM, respectively, for enhanced discrimination ability of the data projections, such as the principal components analysis NMF (PCA-NMF) [4], which maximizes the coefficient matrix covariance, the spatially localized NMF (LNMF) [5], which imposes sparseness constraints to the basis matrix and the use of the separable case approximation (SCA) algorithm [6], which computes the SVM on the modified separable training data. In these methods, data representation (NMF) and classification (SVM) occur in independent steps.

In this paper, data representation through DNMF and classification through SVMs are formulated into a single objec-

The research leading to these results has received funding from the Collaborative European Project MOBISERV FP7-248434 (http://www. mobiserv.eu), An Integrated Intelligent Home Environment for the Provision of Health, Nutrition and Mobility Services to the Elderly.

tive function, whose optimization aims the projection of the data in a space with reduced dimensions, ensuring that the data projections have minimum within-class variance, maximum between-class variance and the data projections between the two classes are separated with maximum margin. More precisely, the maximum-margin hyperplane is defined as a linear combination of the data projections $\mathbf{h}_j, j = 1, \ldots, M$ and it is incorporated in the objective function of DNMF.

2. DISCRIMINANT NMF WITH SVM CONSTRAINTS

Let $\mathcal{D} = \{\{\mathbf{x}_j, y_j\}, j = 1, ..., M, \mathbf{x}_j \in \Re^N, y_j \in \{-1, 1\}\}$ be the set of M training data, where $\mathbf{x}_j$ denote the data points and y_j are the corresponding labels. In the standard approach, first NMF (or DNMF) is employed in order to find two non-negative matrices $\mathbf{Z}$, $\mathbf{H}$, that minimize the reconstruction error of the data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$ according to (1) (or (3)), where $D(\mathbf{X}\|\mathbf{ZH})$ denotes the Frobenius norm:

$$D(\mathbf{X}\|\mathbf{ZH}) = \|\mathbf{X} - \mathbf{ZH}\|_2 \tag{7}$$

or the Kullback-Leibler divergence:

$$D(\mathbf{X}\|\mathbf{ZH}) = \sum_{i,j} x_{ij} \ln \left(\frac{x_{ij}}{\sum_{l=1}^{L} z_{il} h_{lj}} \right) + \sum_{l=1}^{L} z_{il} h_{lj} - x_{ij} \tag{8}$$

between $\mathbf{X}$ and $\mathbf{ZH}$. Then, SVM is performed on the projected data $\mathbf{h}_j = \mathbf{Z}^T \mathbf{x}_j$ in order to find the hyperplane that maximizes the margin $\frac{2}{\|\mathbf{w}\|}$ between the two classes or, equivalently,

$$\arg\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|_2^2 \tag{9}$$

$$\text{subject to } y_j(\mathbf{w}^T \mathbf{h}_j + b) - 1 \geq 0, \forall j = 1, \ldots, M, \tag{10}$$

where $\mathbf{w}$ is the normal vector to the hyperplane and b is the bias. Taking into account the Lagrangian multipliers method and the KKT conditions, the objective of SVM can be written in the form:

$$\min_{a_j} \left\{ \frac{1}{2} \sum_{j=1}^{M} \sum_{k=1}^{M} a_j a_k y_j y_k \mathbf{h}_j^T \mathbf{h}_k - \sum_{j=1}^{M} a_j \right\} \tag{11}$$

$$\text{subject to } a_j \geq 0, \forall j = 1, \ldots, M, \tag{12}$$

where a_j are the Lagrange multipliers.

In this paper, we exploit the SVM constraints in the optimization framework of DNMF, i.e., we want to find a nonnegative base matrix $\mathbf{Z}$ so that, the data projections $\mathbf{h}_j$ minimize the reconstruction error (8), minimize the within-class variance $\mathbf{S}_w$, maximize the between-class variance $\mathbf{S}_b$ and they are separated with maximum margin by the hyperplane $\mathbf{w}$, which, according to the representer theorem [7], lies in the span of the data projections $\mathbf{w} = \sum_{j=1}^{M} a_j y_j \mathbf{h}_j$. This is accomplished by combining the cost functions (3) and (11) into a single objective function:

$$\begin{aligned} F(z_{il}, h_{lj}, a_j) &= \lambda D(\mathbf{X}\|\mathbf{ZH}) + \gamma \text{tr}[\mathbf{S}_w] - \delta \text{tr}[\mathbf{S}_b] \\ &+ \frac{1}{2} \sum_{jk}^{M} a_k a_j y_k y_j \sum_{l}^{L} h_{lj} h_{lk} - \sum_{j}^{M} a_j, \end{aligned} \tag{13}$$

where $D(\mathbf{X}\|\mathbf{ZH})$ is given by (8), which we want to minimize with respect to z_{il}, h_{lj}, a_j, subject to the constraints:

$$z_{il} \geq 0, \;\; h_{lj} \geq 0, \;\; a_j \geq 0, \;\; \text{and} \;\; \sum_{i=1}^{N} z_{il} = 1, \;\; \forall l = 1, \ldots, L. \tag{14}$$

The direct minimization of (13) is difficult. However, a local minimum of (13) can be found, by performing the EM algorithm, since the objective function (13), subject to the constraints (14), is convex with respect to either z_{il}, h_{lj} or a_j. This can be proved by showing that:

$$\frac{\partial^2}{\partial z_{il}^2} F(z_{il}) = \lambda \sum_{j} \frac{x_{ij} h_{lj}^2}{\left(\sum_k z_{ik} h_{kj}\right)^2} \geq 0 \tag{15}$$

$$\frac{\partial^2}{\partial h_{lj}^2} F(h_{lj}) = \sum_i \frac{x_{ij}}{h_{lj}^2} + 2\gamma \left(1 - \frac{1}{M_c}\right) - 2\delta \left(\frac{1}{M_c} - \frac{1}{M}\right) + a_j^2 y_j^2 \geq 0 \tag{16}$$

$$\frac{\partial^2}{\partial a_j^2} F(a_j) = \sum_{l} h_{lj}^2 \geq 0 \tag{17}$$

$\forall z_{il}, h_{lj}, a_j \geq 0$ and $M_c \geq 1 + \delta/\gamma$, where, for simplicity in notation, we defined:

$$F(z_{il}) = F(z_{il}, h_{lj}, a_j)|_{h_{lj}, a_j = \text{constant}} \tag{18}$$

$$F(h_{lj}) = F(z_{il}, h_{lj}, a_j)|_{z_{il}, a_j = \text{constant}} \tag{19}$$

$$F(a_j) = F(z_{il}, h_{lj}, a_j)|_{z_{il}, h_{lj} = \text{constant}}. \tag{20}$$

By choosing $\gamma \geq \delta$ the condition $M_c \geq 2$ is obtained, which means that the convexity holds when each class has at least two samples. This is a very loose restriction, which is satisfied in all the conducted experiments. Therefore, a local minimum of (13) can be found by minimizing three auxiliary functions $G(z_{il}, z_{il}^{(t)})$, $G(h_{lj}, h_{lj}^{(t)})$ and $G(a_j, a_j^{(t)})$ for the functions $F(z_{il})$, $F(h_{lj})$ and $F(a_j)$, respectively. The function $G(x, x^{(t)})$ is defined to be an auxiliary function for $F(x)$ if $G(x, x^{(t)}) \geq F(x)$ and $G(x, x) = F(x)$. It is proven in [1] that if $G(x, x^{(t)})$ is an auxiliary function for $F(x)$, then the minimization of $G(x, x^{(t)})$ with respect to x leads to minimization of $F(x)$. As a consequence, $F(x)$ is monotonically decreasing under the update rule:

$$x^{(t+1)} = \arg\min_{x} \{G(x, x^{(t)})\}. \tag{21}$$

2.1. Minimization of $F(z_{il}, h_{lj}, a_j)$ w.r.t. z_{il}

The function

$$G(z_{il}, z_{il}^{(t)}) = \lambda \left[\sum_{ij} (x_{ij} \ln x_{ij} - x_{ij}) - \sum_{ijl} x_{ij} \frac{z_{il}^{(t)} h_{lj}}{\sum_m z_{im}^{(t)} h_{mj}} \right.$$

$$\times \left(\ln z_{il} h_{lj} - \ln \frac{z_{il}^{(t)} h_{lj}}{\sum_m z_{im}^{(t)} h_{mj}} \right) + \sum_{ijl} z_{il} h_{lj} \Big] + \gamma \text{tr}[\mathbf{S}_w] - \delta \text{tr}[\mathbf{S}_b]$$
$$+ \tfrac{1}{2} \sum_{jk}^{M} a_k a_j y_k y_j \sum_{l}^{M} h_{lj} h_{lk} - \sum_{j}^{M} a_j \qquad (22)$$

is an auxiliary function for the cost function $F(z_{il})$. The derivation of (22) is straightforward from the derivation of the update rule of z_{il} in NMF [8]. The minimization of (22) is performed by setting the partial derivative of $G(z_{il}, z_{il}^{(t)})$ with respect to z_{il} to zero. As a result, $F(z_{il})$ subject to the constraints $z_{il} \geq 0$ and $\sum_{l=1}^{L} z_{il} = 1$ is non-increasing under the following update rules:

$$z_{il}^{'(t+1)} = \sum_j x_{ij} \frac{h_{lj}}{\sum_m z_{im}^{'(t)} h_{mj}} \frac{1}{\sum_j h_{lj}} z_{il}^{'(t)} \qquad (23)$$

$$z_{il}^{(t+1)} = \frac{z_{il}^{'(t+1)}}{\sum_{i=1}^{N} z_{il}^{'(t+1)}}. \qquad (24)$$

The constraint $\sum_{l=1}^{L} z_{il} = 1$ ensures that the nonnegative basis matrix $\mathbf{Z}$ is sparse.

2.2. Minimization of $F(z_{il}, h_{lj}, a_j)$ w.r.t. a_j

The function

$$G(a_j, a_j^{(t)}) = \lambda D(\mathbf{X} \| \mathbf{ZH}) + \gamma \text{tr}[\mathbf{S}_w] - \delta \text{tr}[\mathbf{S}_b]$$
$$+ \frac{1}{2} \sum_{jk} \frac{A_{jk}^{+} a_k^{(t)}}{a_j^{(t)}} a_j^2 - \frac{1}{2} \sum_{jk} A_{jk}^{-} a_j^{(t)} a_k^{(t)}$$
$$\times \left(1 + \ln \frac{a_j a_k}{a_j^{(t)} a_k^{(t)}} \right) - \sum_j a_j, \qquad (25)$$

where $A_{jk} = y_j y_k \sum_l h_{lj} h_{lk}$, $A_{jk}^{+} = \max(A_{jk}, 0)$ and $A_{jk}^{-} = \max(-A_{jk}, 0)$, is an auxiliary function for the cost function $F(a_j)$. The derivation of the auxiliary function (25) is straightforward from the derivation of the update rules of a_j in SVM [9]. By setting the partial derivative of $G(a_j, a_j^{(t)})$ with respect to a_j to zero, the following update rule for a_j is derived:

$$a_j^{(t+1)} = \frac{1 + \sqrt{1 + 4 \sum_k A_{jk}^{+} a_k^{(t)} \sum_k A_{jk}^{-} a_k^{(t)}}}{2 \sum_k A_{jk}^{+} a_k^{(t)}} a_j^{(t)}. \qquad (26)$$

2.3. Minimization of $F(z_{il}, h_{lj}, a_j)$ w.r.t. h_{lj}

The function

$$G(h_{lj}, h_{lj}^{(t)}) = G_1(h_{lj}, h_{lj}^{(t)}) + G_2(h_{lj}, h_{lj}^{(t)}), \qquad (27)$$

where

$$G_1(h_{lj}, h_{lj}^{(t)}) = \lambda \Bigg[\sum_{ij} (x_{ij} \ln x_{ij} - x_{ij}) - \sum_{ijl} x_{ij} \frac{z_{ij} h_{lj}^{(t)}}{\sum_m z_{im} h_{mj}^{(t)}}$$
$$\times \left(\ln z_{il} h_{lj} - \ln \frac{z_{il} h_{lj}^{(t)}}{\sum_m z_{im} h_{mj}^{(t)}} \right) + \sum_{ijl} z_{il} h_{lj} \Bigg] + \gamma \text{tr}[\mathbf{S}_w] - \delta \text{tr}[\mathbf{S}_b], \qquad (28)$$

and

$$G_2(h_{lj}, h_{lj}^{(t)}) = \tfrac{1}{2} \sum_{ljk} \frac{B_{jk}^{+} h_{lk}^{(t)}}{h_{lj}^{(t)}} h_{lj}^2 - \tfrac{1}{2} \sum_{ljk} B_{jk}^{-} h_{lj}^{(t)} h_{lk}^{(t)}$$
$$\times \left(1 + \ln \frac{h_{lj} h_{lk}}{h_{lj}^{(t)} h_{lk}^{(t)}} \right) - \sum_j a_j \, , \qquad (29)$$

$B_{jk} = a_j a_k y_j y_k$, $B_{jk}^{+} = \max(B_{jk}, 0)$ and $B_{jk}^{-} = \max(-B_{jk}, 0)$ is an auxiliary function for the cost function $F(h_{lj})$. The derivation of the auxiliary function (27)-(29) is straightforward from the derivation of the update rules of h_{lj} in NMF [8] and a_j in SVM [9]. By setting the partial derivative of $G(h_{lj}, h_{lj}^{(t)})$ with respect to h_{lj} to zero, the following update rule for h_{lj} is derived:

$$h_{lj} = \frac{-T_1 + \sqrt{T_1^2 + 4T_2 \left[\lambda \sum_i x_{ij} \frac{z_{il} h_{lj}^t}{\sum_m z_{im} h_{mj}^t} + \sum_k B_{jk}^{-} h_{jl}^t h_{lk}^t \right]}}{2T_2}, \qquad (30)$$

where:

$$T_1 = \lambda - 2\gamma \tfrac{1}{M_r} \sum_{k=1, k \neq j}^{M_r} h_{lk} - 2\delta \tfrac{1}{M_r} \sum_{k=1, k \neq j}^{M_r} h_{lk} + 2\delta \tfrac{1}{M} \sum_{k=1, k \neq j}^{M} h_{lk} \qquad (31)$$

and:

$$T_2 = \sum_k \frac{B_{jk}^{+} h_{lk}^t}{h_{lj}^t} + 2\gamma \left(1 - \frac{1}{M_r} \right) - 2\delta \left(\frac{1}{M_r} - \frac{1}{M} \right), \qquad (32)$$

2.4. Minimization of $F(z_{il}, h_{lj}, a_j)$ w.r.t. z_{il}, h_{lj} and a_j

Based on the analysis in subsections 2.1-2.3, the cost function $F(z_{il}, h_{lj}, a_j)$ (13) is non-increasing under the iterative update rules (23), (24), (26) and (30). In each iteration, the update rules are computed sequentially, until the convergence of the cost function to a local minimum, i.e, when the change in the value of $F(z_{il}, h_{lj}, a_j)$ in two successive iterations is very small. Experimental results showed that $F(z_{il}, h_{lj}, a_j)$ converges to a local minimum in approximately 1000 iterations. In equation (13), the parameter λ regulates the significance of the NMF part in the objective function. During the first iterations λ takes large values, increasing the significance of the correct data representation. λ decreases exponentially with the number of iterations t, according to $\lambda_0/(1+e)^t$, where the parameter $e << 1$ regulates the decrease rate. λ plays an important role in the classification decision. Experimental results showed that typical values for λ_0 are $\lambda_0 = 100$ or $\lambda_0 = 1000$, while the decrease rate e takes values in the range from 10^{-3} to 10^{-2}. Moreover, the proper values for the weights γ and δ were experimentally found to be 0.1 and 0.05, respectively.

When the algorithm converges, the train data are projected to the reduced dimensional space using the transpose base matrix $\mathbf{Z}^T$. Alternatively, the data projections $\mathbf{h}_j$ can be estimated using the pseudo-inverse $\mathbf{Z}^{\dagger} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T$, or by the

multiplicative update rule (30). The test data are projected to the reduced dimensional space accordingly. The maximum margin hyperplane of SVM is computed by:

$$\mathbf{w} = \sum_{j=1}^{M} a_j y_j \mathbf{h}_j \tag{33}$$

$$b = \frac{1}{|\mathcal{M}_{\mathcal{SV}}|} \sum_{j \in \mathcal{M}_{\mathcal{SV}}} \left(\mathbf{w}^T \mathbf{h}_j - y_j \right) \tag{34}$$

where $\mathcal{M}_{\mathcal{SV}}$ denotes the set of support vectors and, finally, the classification decision on the train and test data is taken according to

$$y_j = \text{sign}\left(\mathbf{w}^T \mathbf{h}_j + b \right). \tag{35}$$

3. EXPERIMENTAL RESULTS

In this section, an experimental evaluation of the proposed DNMF with SVM constraints optimization framework (DNMFSVM) is presented. In Subsection 3.1, DNMFSVM was employed on seven two-class UCI data sets and in Subsection 3.2, DNMFSVM was performed on the Cohn-Kanade database for facial expression recognition. In all experiments, the classification accuracy of DNMFSVM was compared to the accuracy of the state of the art first NMF then SVM (NMF+SVM) and first DNMF then SVM (DNMF+SVM) approaches.

3.1. UCI Databases

In the first experiment, we tested the performance of DNMFSVM for varying values of the reduced dimensionality L on the liver disorders data set [10]. The classification error for the DNMFSVM, NMF+SVM and DNMF+SVM algorithms was computed by using the ten-fold-cross-validation method. The results are depicted in Table 1. We notice that the classification error of DNMFSVM decreases with the increase of dimensions L. Such behavior is not observed in the state of the art NMF+SVM and DNM+SVM methods. Moreover, the classification error of DNMFSVM is smaller than the classification error of NMF+SVM and DNMF+SVM when $L \geq 2$.

In the next experiment, the performance of DNMFSVM was tested in six more two-class data sets from the UCI repository [10]: the ionoshpere data set, the hill/valley and hill/valley with noise data sets, the pima Indians diabetes data set and the Wisconsin breast cancer (prognostic and diagnostic) data sets. The classification errors of the proposed DNMFSVM and the state of the art NMF+SVM and DNMF+SVM methods are shown in Table 2. The original and projected data dimensions are shown in the second and third columns of Table 2, respectively. Except from the hill/valley and hill/valey with noise data sets (third and fourth rows of Table 2), where standard training and testing sets are provided, in the rest data sets the ten fold cross validation method was employed. We notice that in five out of six cases the proposed DNMFSVM method achieves the minimum classification error.

Table 1. Classification error (%) of NMF+SVM, DNMF+SVM and DNMFSVM algorithms for variable L for the liver disorders data set.

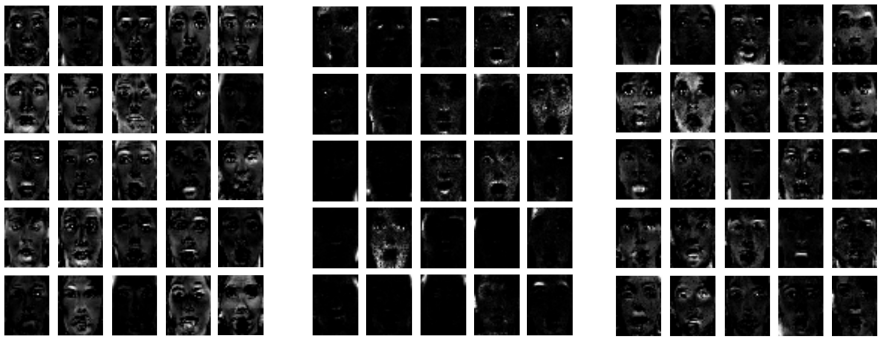
L	NMF+SVM	DNMF+SVM	DNMFSVM
1	38.53	**38.24**	43.82
2	39.41	37.35	**32.35**
3	40.88	38.82	**32.06**
4	37.65	39.41	**30.59**
5	38.82	38.24	**28.82**
6	42.94	40.88	**29.12**

Table 2. Classification error (%) of NMF+SVM, DNMF+SVM and DNMFSVM algorithms for six UCI data sets

database	N	L	NMF+SVM	DNMF+SVM	DNMFSVM
ionosphere	34	2	45.14±3.65	42.86±5.08	**29.14±3.22**
hill/valley	100	10	29.70±0.5	7.05±0.22	**7.00±0.26**
h/v noise	100	10	38.94±0.5	**9.08±0.1**	**9.08±0.1**
pima	8	2	19.47±8.41	**14.74±3.91**	28.42±8.58
wdbc	30	2	35.96±1.58	33.16±3.57	**13.51±2.5**
wpbc	30	2	5.26±0.94	**2.11±0.7**	**2.11±0.7**

3.2. Cohn-Kanade Database

In this Subsection we test the performance of DNMFSVM in facial expression recognition. The algorithm was employed on the Cohn-Kanade database [11], which consists of 486 images from 97 persons performing six expressions: anger, disgust, fear, happiness, sadness and surprise. The size of the facial images is 40×30, thus the original dimension of the data is $N = 1200$. In the conducted experiment, the dimensionality of the data is reduced to $L = 100$. The task of facial expression recognition is divided into two-class subtasks, by selecting 2-combinations from the six classes and performing five fold cross validation. Figure 1 depicts the first 25 basis images of the proposed DNMFSVM algorithm and the state of the art NMF+SVM and DNMF+SVM methods. From Figure 1 we notice that, the sparseness of the basis images produced by DNMFSVM is greater than NMF+SVM and lower than DNMF+SVM method. The average classification errors for the proposed DNMFSVM and the state of the art NMF+SVM and DNMF+SVM methods are shown in Table 3. We notice that DNMFSVM achieves the lowest classification error (21.90%) followed by the state of the art DNMF+SVM (24.90%).



(a) (b) (c)

Fig. 1. A set of 25 basis images for (a) NMF+SVM, (b) DNMF+SVM, (c) DNMFSVM.

Table 3. Classification error (%) of NMF+SVM, DNMF+SVM and DNMFSVM algorithms for the Cohn-Kanade Database

NMF+SVM	DNMF+SVM	DNMFSVM
25.52	24.29	**21.90**

4. CONCLUSIONS

In this paper a novel method was introduced that incorporates discriminative constraints and the maximum margin constraints of SVMs in the objective function of NMF. The intuition behind the proposed framework was to find the projection matrix that projects the data to a low-dimensional space so that the data projections have minimum within-class variance, maximum between-class variance and the data projections between the two classes are separated by a hyperplane with maximum margin. Experimental results on several data sets showed the supremacy of the proposed method with respect to the state of the art NMF and DNMF followed by SVM classification.

5. REFERENCES

[1] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[2] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Transactions on Neural Networks*, vol. 17, no. 3, pp. 683 –695, may 2006.

[3] Christopher J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, vol. 2, pp. 121–167, June 1998.

[4] Yuan Wang, Yunde Jia, Changbo Hu, and Matthew Turk, "Non-negative matrix factorization framework for face recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, pp. 495–511, 2005.

[5] S.Z. Li, Xin Wen Hou, Hong Jiang Zhang, and Qian Sheng Cheng, "Learning spatially localized, parts-based representation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. 207 – 212.

[6] Dries Geebelen, Johan A. K. Suykens, and Joos Vandewalle, "Reducing the number of support vectors of svm classifiers using the smoothed separable case approximation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 4, pp. 682–688, april 2012.

[7] Thomas Hofmann, Bernhard Schlkopf, and Alexander J. Smola, "Kernel methods in machine learning," *Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[8] Daniel D. Lee and H. Sebastian Seung, "Algorithms for Non-negative Matrix Factorization," in *Advances in Neural Information Processing Systems 13*. Apr. 2001, pp. 556–562, MIT Press.

[9] Fei Sha, Yuanqing Lin, Lawrence K. Saul, and Daniel D. Lee, "Multiplicative updates for nonnegative quadratic programming," *Neural Comput.*, vol. 19, pp. 2004–2031, August 2007.

[10] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.

[11] T. Kanade, J.F. Cohn, and Yingli T., "Comprehensive database for facial expression analysis," in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 46 –53.