ROBUST DISCRIMINATIVE NON-NEGATIVE MATRIX FACTORIZATION WITH MAXIMUM CORRENTROPY CRITERION

Hang Cheng, Shixiong Wang and Naiyang Guan

National Innovation Institute of Defense Technology,
Academy of Military Sciences, Beijing, China

ABSTRACT

Non-negative matrix factorization (NMF) is an effective dimension reduction tool widely used in pattern recognition and computer vision. However, conventional NMF models are neither robust enough, as their objective functions are sensitive to outliers, nor discriminative enough, as they completely ignore the discriminative information in data. In this paper, we proposed a robust discriminative NMF model (RDNMF) for learning an effective discriminative subspace from noisy dataset. In particular, RDNMF approximates observations by their reconstructions in the subspace via maximum correntropy criterion to prohibit outliers from influencing the subspace. To incorporate the discriminative information, RDNMF builds adjacent graphs by using maximum correntropy criterion based robust representation, and regularizes the model by margin maximization criterion. We developed a multiplicative update rule to optimize RDNMF and theoretically proved its convergence. Experimental results on popular datasets verify the effectiveness of RDNMF comparing with conventional NMF models, discriminative NMF models, and robust NMF models.

KEYWORDS

Dimension reduction, non-negative matrix factorization, maximum correntropy criterion, supervised learning, margin maximization.

1. INTRODUCTION

Dimension reduction plays an important role in pattern recognition, computer vision and information retrieval. It projects samples from high-dimensional space onto a low-dimensional space, and thus reveals the intrinsic structure of a dataset to boost the subsequent processing. Recently, non-negative matrix factorization (NMF, [1]) has been proven to be a powerful dimension reduction method which approximates a non-negative data matrix by the product of two lower dimensional non-negative matrices. Since NMF learns a natural parts-based representation, it has been widely used in many tasks such as data mining [2], pattern recognition [3,4], and computer vision [5].

Since traditional NMF methods cannot take advantage of the labels of a dataset, they usually perform unsatisfactorily in classification tasks. To overcome this deficiency, Zaferiou et al. [6] proposed discriminant NMF (DNMF) to incorporate the Fisher's criterion into NMF. However, DNMF intrinsically assumes that samples obey Gaussian distribution, and this assumption is sometimes improper because NMF itself does not assume samples are Gaussian distributed. To overcome this problem, Guan et al. [7] proposed manifold regularized discriminative NMF (MD-NMF) to retain discriminative information for subsequent classification by marginal

David C. Wyld et al. (Eds): AI, AIMLNET, BIOS, BINLP, CSTY, MaVaS, SIGI - 2022
pp. 37-49, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121804

maximization. Neither DNMF nor MD-NMF can perform well on some seriously noisy datasets because their Frobenius norm based [8] or Kullback-Leiblur (KL) divergence based [9] loss functions are sensitive to outliers.

In this paper, we propose a correntropy supervised NMF (CSNMF) to overcome this deficiency. In particular, inspired by [12], CSNMF measures the loss of NMF by the well-known correntropy induced metric (CIM, [10]) instead of Frobenius-norm and KL-divergence. CIM is controlled by a kernel size and approximates L_0-norm when the loss is relatively large, and thus it is robust to noise of large magnitudes or outliers. Assuming even noisy samples have correct labels, to utilize the labels of the dataset, CSNMF narrows the distance between any samples of the same class in the lower dimensional space. Since this discriminative information is noise-free and the utilized CIM-based loss function filters out any noise of large magnitude in the dataset, CSNMF can boost subsequent classification performance on the noisy datasets. In addition, we developed a multiplicative update rule to optimize CSNMF and theoretically proved its convergence. The experimental results on several popular face image datasets confirm the effectiveness of our CSNMF comparing with the supervised NMF variants and the robustified NMF variants.

The rest of this paper is organized as follows. Section 2 briefly reviews NMF and its variants. Section 3 presents the proposed CSNMF, the multiplicative update rule and its convergence. Section 4 shows the experimental results on popular face image datasets. We conclude this paper in Section 5.

2. RELATED WORKS

2.1. NMF

Given any non-negative matrix, i.e., $X \in \mathbb{R}_+^{m\times n}$, non-negative matrix factorization (NMF, [1]) aims at finding two lower dimensional non-negative matrices, i.e., $U \in \mathbb{R}_+^{m\times r}$ and $V \in \mathbb{R}_+^{r\times n}$, by minimizing the distance between X and UV. Conventional NMF methods measure the distance by using either Kullback-Leibler (KL) divergence [9] or squared Frobenius-norm [8], and thus they are not robust enough because their underlying distributions cannot model outliers. In addition, traditional NMF methods are not supervised because they completely ignore labels of a dataset.

2.2. NMF's Variants

From the seminal work of Lee and Seung [1] until now on, many NMF variants have been developed to deal with various practical tasks. Guan *et al.* [11] proposed a non-negative patch alignment framework (NPAF) to unify the popular NMF-related dimension reduction methods. The objective function of NPAF is

$$\min_{U\geq 0,V\geq 0} \frac{\gamma}{2} tr\left(VLV^T\right) + D\left(X,UV\right), \tag{1}$$

where $D(\cdot,\cdot)$ measures the loss of such factorization, γ is a positive tradeoff parameter, and L is an alignment matrix that encodes the statistical information of the datasets.

Based on NPAF, one can easily develop novel NMF-related dimension reduction method. For example, Guan *et al.* [11] developed a manifold regularized discriminative non-negative matrix factorization (MD-NMF) method to preserve the local geometric structure and incorporate the discriminative information of the dataset. The objective function of MD-NMF is

$$\min_{U\geq 0,V\geq 0}\frac{\gamma}{2}tr\left(VLV^T\right)+\sum_{ij}\left(X_{ij}\log\frac{X_{ij}}{\left(UV\right)_{ij}}-X_{ij}+\left(UV\right)_{ij}\right). \tag{2}$$

However, like NMF, MD-NMF is not robust enough because its loss function is sensitive to outliers.

To enhance the robustness of NMF, Du *et al.* [12] proposed a correntropy induced metric (CIM) based NMF (CIMNMF), which introduce CIM to measure the loss of the factorization. The objective function of CIMNMF is

$$\min_{U\geq 0,V\geq 0} CIM^2(X,UV), \tag{3}$$

where $CIM(X,UV)=\left(k_\sigma(0)-\frac{1}{mn}\sum_{ij}k_\sigma(X_{ij}-(UV)_{ij})\right)^{1/2}$ and $k_\sigma(z)=e^{-z^2/2\sigma^2}$ is a Gaussian kernel function and σ is the kernel size.

Li *et al.* [13] proposed a graph regularized nonnegative matrix factorization method by maximizing the correntropy criterion (MCCGR) to incorporate the local geometric structure into CIMNMF, i.e.,

$$\max_{U\geq 0,V\geq 0}\sum_{i=1}^{m}k_\sigma\left(\sqrt{\sum_{j=1}^{n}\left(X_{ij}-\sum_{k=1}^{r}U_{ik}V_{kj}\right)^2}\right)-\frac{1}{2}tr\left(VLV^T\right), \tag{4}$$

where L is the graph Laplacian of the constructed adjacent graph.

Huang *et al.* [14] proposed robust manifold NMF (RMNMF) to preserve the local geometric structure in their previously developed robust NMF (RNMF) with $L_{2,1}$-norm, i.e.,

$$\min_{U\geq 0,V\geq 0,V^TV=I}\|X-UV\|_{2,1}+\gamma tr\left(VLV^T\right), \tag{5}$$

where $\|Y\|_{2,1}=\sum_j\|Y_{\cdot j}\|_2$ signifies the $L_{2,1}$-norm. Both MCCGR and RMNMF can be easily unified under NPAF.

Although both MCCGR and RMNMF show promises by the authors, they still have flawless. Since outliers might violate the intrinsic geometric structure of the clean dataset without outliers, the alignment matrix constructed on the noisy observations might be inaccurate from the view point of NPAF. Therefore, there are still some space to develop a novel robust NMF variant by simultaneously considering the robustness of loss function and alignment matrix.

3. CORRENTROPY SUPERVISED NMF

In this section, we first described a novel correntropy supervised NMF (CSNMF) based on NPAF to overcome the deficiencies of NMF and its variants. Then we developed a multiplicative update rule (MUR) to solve CSNMF. At last, we theoretically proved that the objective function of CSNMF is non-increasing under MUR.

3.1. Correntropy Induced Metric

In information-theoretic learning (ITL), one often uses correntropy to process noise [10]. Correntropy is defined as a generalized similarity between two variables

$$C_\sigma(x,y) = E\left(k_\sigma(x-y)\right), \tag{6}$$

where k_σ is the kernel function, and both x and y represent random variables. Given n samples, the estimator of correntropy is

$$\hat{C}_\sigma(x,y) = \frac{1}{n}\sum_{i=1}^{n} k_\sigma(x_i, y_i). \tag{7}$$

Based on ITL, Liu *et al.* [10] proposed the correntropy induced metric (CIM)

$$CIM(x,y) = \left(k_\sigma(0) - \frac{1}{n}\sum\nolimits_{i=1}^{n} k_\sigma(e_i)\right)^{1/2}, \tag{8}$$

where $e_i = x_i - y_i$ represents the reconstruction error. The CIM value of large error in (8) is upper bounded by 1 regardless the scale of error. Therefore, CIM is less influenced by outliers. Due to its robustness, CIM has been widely used in many signal processing [10] and face recognition [15-16] tasks.

3.2. The CSNMF Model

Given n samples arranged in a non-negative matrix, i.e., $X \in \mathbb{R}_{+}^{m\times n}$, correntropy supervised NMF (CSNMF) decomposes it into the product of two lower-rank matrices, i.e., $X \approx UV$, by minimizing the CIM between X and UV, i.e.,

$$\min_{U\geq 0, V\geq 0} CIM^2(X, UV), \tag{9}$$

where $U \in \mathbb{R}_{+}^{m\times r}$ signifies the bases, and $V \in \mathbb{R}_{+}^{r\times n}$ signifies the coefficients. Based on ITL [10][12], CSNMF succeeds to filter out outliers.

For supervised learning, we assume that all samples including noisy samples are correctly labeled. Such assumption makes sense in some situations. Taking face recognition system for example, training images might be corrupted by illuminations but the subjects' name are known and correct. CSNMF expect to dig the discriminative information from labels of the dataset by incorporating the labels of a dataset, i.e., it narrows the distance between samples of the same class, i.e.,

$$\min_{V\geq 0} \sum_{i\neq j} \left\| v_i - v_j \right\|_2^2 S_{ij}, \tag{10}$$

where $\|\cdot\|_2$ signifies the L_2-norm, and S_{ij} reflects the similarity between x_i and x_j, i.e.,

$$S_{ij} = \begin{cases} 1, & l(x_i) = l(x_j), i \neq j \\ 0, & otherwise \end{cases}, \tag{11}$$

where $l(\cdot)$ means the label of a sample. Although samples X might be corrupted, considering the robustness of CIM, it is reasonable to trust that the coefficients V is much less sensitive to outliers than X. Therefore, we can measure the distance between coefficients of two samples by the L_2-norm in (10) to benefit from its nice mathematic property.

By simple algebra, we can rewrite the objective function in (10) as $\sum_{i \neq j} \left\| v_i - v_j \right\|_2^2 S_{ij} = tr(\mathrm{VLV}^T)$, where $L = D - S$ and D is a diagonal matrix with $D_{ii} = \sum_{ij} S_{ij}$. By combing (9) and (10), we obtain the objective function of CSNMF as follows:

$$\min_{U \geq 0, V \geq 0} CIM^2(X, UV) + \frac{\gamma}{2} \mathrm{tr}(\mathrm{VLV}^T), \tag{12}$$

where γ is a positive trade-off parameter. Obviously, CSNMF can be unified by NPAF and L is considered an alignment matrix. CSNMF is jointly non-convex, and thus it is impossible to find its global optimum in polynomial time. In the following section, we developed a multiplicative update rule (MUR) to find its local minimum.

3.3. MUR for Optimizing CSNMF

For solving the constrained optimization problem (12), by using the Lagrangian multiplier method [8], we can obtain the Lagrangian function as follows:

$$\mathrm{L} = CIM^2(X, UV) + \frac{\gamma}{2} tr\left(VLV^T\right) + tr(\alpha U) + tr(\beta V), \tag{13}$$

where α and β are the Lagrangian multipliers for the constraints $U \geq 0$ and $V \geq 0$, respectively.

To solve (12), we firstly calculate the first-order derivatives of L with respect to U and V, i.e.,

$$\begin{aligned} \frac{\partial \mathrm{L}}{\partial v_{ab}} &= -\frac{1}{mn\sigma^2} \sum_k u_{ka} \left(X_{kb} - \sum_r u_{kr} v_{rb} \right) e^{-\frac{\left(X_{kb} - \sum_r u_{kr} v_{rb}\right)^2}{2\sigma^2}} + \gamma \sum_r v_{ar} L_{rb} + \beta_{ab} \\ &= -\frac{1}{mn\sigma^2} \left[U^T \left((X - UV) \otimes W \right) \right]_{ab} + \gamma (VL)_{ab} + \beta_{ab}. \end{aligned} \tag{14}$$

And

$$\begin{aligned} \frac{\partial \mathrm{L}}{\partial u_{ij}} &= -\frac{1}{mn\sigma^2} \sum_k v_{jk} \left(X_{ik} - \sum_r u_{ir} v_{rk} \right) e^{-\frac{\left(X_{ik} - \sum_r u_{ir} v_{rk}\right)^2}{2\sigma^2}} + \alpha_{ij} \\ &= -\frac{1}{mn\sigma^2} \left[W \otimes (X - UV) V^T \right]_{ij} + \alpha_{ij}, \end{aligned} \tag{15}$$

The above two equations can be further written as

$$\frac{\partial \mathrm{L}}{\partial v_{ab}} = -\frac{1}{mn\sigma^2} \sum_k u_{ka} X_{kb} e^{-\frac{\left(X_{kb} - \sum_r u_{kr} v_{rb}\right)^2}{2\sigma^2}} + \frac{1}{mn\sigma^2} \sum_k u_{ka} e^{-\frac{\left(X_{kb} - \sum_r u_{kr} v_{rb}\right)^2}{2\sigma^2}} \sum_r u_{kr} v_{rb} + \gamma \left(VL^+\right)_{ab} - \gamma \left(VL^-\right)_{ab} + \beta_{ab}, \tag{16}$$

and

$$\frac{\partial \mathrm{L}}{\partial u_{ij}} = -\frac{1}{mn\sigma^2}\sum_k v_{jk} X_{ik} e^{-\frac{\left(X_{ik}-\sum_r u_{ir}v_{rk}\right)^2}{2\sigma^2}} + \frac{1}{mn\sigma^2}\sum_k v_{jk} e^{-\frac{\left(X_{ik}-\sum_r u_{ir}v_{rk}\right)^2}{2\sigma^2}} \sum_r u_{ir}v_{rk} + \alpha_{ij}, \tag{17}$$

where $L = L^{+} - L^{-} = D - S$, and $D \geq 0$ and $S \geq 0$ construct the positive and negative components of L , respectively.

By using the K.K.T. conditions [17], any stationary point of (12) satisfies the following conditions

$$\begin{cases} \frac{\partial \mathrm{L}}{\partial u_{ij}} = 0, \ \frac{\partial \mathrm{L}}{\partial v_{ab}} = 0 \\ \alpha_{ij} u_{ij} = 0, \ \beta_{ab} v_{ab} = 0. \\ u_{ij} \geq 0, \ v_{ab} \geq 0 \\ \alpha_{ij} \geq 0, \ \beta_{ab} \geq 0 \end{cases} \tag{18}$$

By combing the first two conditions in (18), we can obtain the following equations

$$(-\frac{1}{mn\sigma^2}\sum_k v_{jk} X_{ik} e^{-\frac{\left(X_{ik}-\sum_r u_{ir}v_{rk}\right)^2}{2\sigma^2}} + \frac{1}{mn\sigma^2}\sum_k v_{jk} e^{-\frac{\left(X_{ik}-\sum_r u_{ir}v_{rk}\right)^2}{2\sigma^2}} \sum_r u_{ir}v_{rk})u_{ij} = 0, \tag{19}$$

$$(-\frac{1}{mn\sigma^2}\sum_k u_{ka} X_{kb} e^{-\frac{\left(X_{kb}-\sum_r u_{kr}v_{rb}\right)^2}{2\sigma^2}} + \frac{1}{mn\sigma^2}\sum_k u_{ka} e^{-\frac{\left(X_{kb}-\sum_r u_{kr}v_{rb}\right)^2}{2\sigma^2}} \sum_r u_{kr}v_{rb} + \gamma\left(VL^{+}\right)_{ab} \text{-}\gamma\left(VL^{-}\right)_{ab})v_{ab} = 0. \tag{20}$$

From (19-20), we can obtain the following multiplicative update rules:

$$V^{t+1} \leftarrow V^t \otimes \frac{\gamma V^t S + \left(U^{tT}(W^t \otimes X)\right)/mn\sigma^2}{\gamma V^t D + \left(U^{tT}\left(W^t \otimes \left(U^t V^t\right)\right)\right)/mn\sigma^2}, \tag{21}$$

where $W_{ij}^t = e^{-\left(X_{ij}-\left(U^t V^t\right)_{ij}\right)^2/2\sigma^2}$, and $\otimes$ signifies element-wise multiplication. Then

$$U^{t+1} \leftarrow U^t \otimes \frac{\left(W^{t+1} \otimes X\right)V^{t+1T}}{\left(W^{t+1} \otimes U^t V^{t+1}\right)V^{t+1T}}, \tag{22}$$

where $W_{ij}^{t+1} = e^{-\left(X_{ij}-\left(U^t V^{t+1}\right)_{ij}\right)^2/2\sigma^2}$. According to [12], the kernel size can be adaptively updated by

$$\sigma^2 = \frac{1}{2mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(X_{ij} - \left(U^t V^t\right)_{ij}\right)^2. \tag{23}$$

We summarized the total procedure of MUR in **Algorithm 1**. The stopping condition is written as follow: $\left|\frac{F(U^{t+1},V^{t+1}) - F\left(U^t,V^t\right)}{F(U^{t+1},V^{t+1}) - F\left(U^0,V^0\right)}\right| < tol$, where $F(U,V)$ is the objective value in (12). It accepts the non-negative samples and outputs its factorization results. Although the MUR is derived from the Lagrangian multiplier method, we can theoretically analyze its convergence by using the auxiliary function in the following section.

```
Algorithm 1: The MUR for Optimizing CSNMF
Input : $X \in \mathbb{R}_{+}^{m\times n}$, $r \ll \min\{m,n\}$.
Output: $U \in \mathbb{R}_{+}^{m\times r}$, $V \in \mathbb{R}_{+}^{r\times n}$.
1. Calculate the alignment matrix $L$ according to (12).
2. Initialize $U^0$, $V^0$, and calculate $\left(W_{ij}^{0}\right)_U = \exp\left(-\frac{\left(X_{ij}-\left(U^0V^0\right)_{ij}\right)^2}{2\sigma^2}\right)$.
3. Repeat
   $V^{t+1} \leftarrow V^t \otimes \frac{\gamma V^t S + \left(U^{tT}(W^t \otimes X)\right)/mn\sigma^2}{\gamma V^t D + \left(U^{tT}\left(W^t \otimes \left(U^tV^t\right)\right)\right)/mn\sigma^2}$, .
   $W_{ij}^{t+1} = \exp\left(-\frac{\left(X_{ij}-\left(U^tV^{t+1}\right)_{ij}\right)^2}{2\sigma^2}\right)$.
   $U^{t+1} = U^t \otimes \frac{\left(W^{t+1}\otimes X\right)\left(V^{t+1}\right)^T}{\left(W^{t+1}\otimes\left(U^tV^{t+1}\right)\right)\left(V^{t+1}\right)^T}$ .
   $W_{ij}^{t+1} = \exp\left(-\frac{\left(X_{ij}-\left(U^{t+1}V^{t+1}\right)_{ij}\right)^2}{2\sigma^2}\right)$.
   $\sigma^2 = \frac{1}{2mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(X_{ij}-\left(U^{t+1}V^{t+1}\right)_{ij}\right)^2$.
   $t \leftarrow t+1$.
4. Until {The stopping condition is satisfied.}
5. $U = U^t$, $V = V^t$ .
```

The computational complexity of **Algorithm 1** is dominated by two parts: the updating statements and the construction of the alignment matrix. The complexity of the first part is $\#iter \times O\left(mnr+n^2r\right)$, where the $\#iter$ is the iteration of the algorithm and $O\left(mnr+n^2r\right)$ is the time cost of each iteration. The complexity of the second part is $O(n^2)$. Therefore, the total time complexity of **Algorithm 1** is $\#iter \times O\left(mnr+n^2r\right)+o(n^2)$.

3.4. Theoretical Analysis

Here we use the auxiliary function technique to prove the convergence of **Algorithm 1**. Let $F(U,V) = CIM^2(X,UV)+\frac{\gamma}{2}tr\left(VLV^T\right)$ denote the objective function in (12).

Lamma 1. If there exists an function G for $F(x)$ which satisfies $G(x,x') \geq F(x)$ and $G(x,x)=F(x)$, then we call it auxiliary function, and F is non-increasing under the following update rule:

$$x^{t+1} = \arg\min_{x} G\left(x,x'\right). \tag{24}$$

Let $J(V)=F_V(U,V)$ denotes the function with respect to V when U is fixed, and $J(U)=F_U(U,V)$ denotes the function with respect to U with V fixed.

Lamma 2. Given U^t, the following function

$$
\begin{aligned}
G\left(u,u_{ij}^t\right) &= J_{u_{ij}}\left(u_{ij}^t\right)+J'_{u_{ij}}\left(u_{ij}^t\right)\left(u-u_{ij}^t\right)+\frac{1}{2mn\sigma^2}\frac{\sum_k v_{jk}w_{ik}\sum_r u_{ir}v_{rk}}{u_{ij}^t}\left(u-u_{ij}^t\right)^2 \\
&= J_{u_{ij}}\left(u_{ij}^t\right)+J'_{u_{ij}}\left(u_{ij}^t\right)\left(u-u_{ij}^t\right)+\frac{1}{2mn\sigma^2}\frac{\left\{\left[W\otimes(UV)\right]V^T\right\}_{ij}}{u_{ij}^t}\left(u-u_{ij}^t\right)^2,
\end{aligned}
\tag{25}
$$

where $J'_{u_{ij}}$ is the first order derivative with respect to U, is an auxiliary function of $J_{u_{ij}}$.

Lamma 3. Given V^t, the following function

$$
\begin{aligned}
G\left(v,v_{ab}^t\right) &= J_{v_{ab}}\left(v_{ab}^t\right)+J'_{v_{ab}}\left(v_{ab}^t\right)\left(v-v_{ab}^t\right)+\frac{1}{2}\frac{\sum_k u_{ka}w_{kb}\sum_r u_{kr}v_{rb}\Big/mn\sigma^2+\gamma\sum_k v_{ak}L_{kb}}{v_{ab}^t}\left(v-v_{ab}^t\right)^2 \\
&= J_{v_{ab}}\left(v_{ab}^t\right)+J'_{v_{ab}}\left(v_{ab}^t\right)\left(v-v_{ab}^t\right)+\frac{1}{2}\frac{\left\{U^T\left[(X-UV)\otimes W\right]\right\}_{ab}\Big/mn\sigma^2+\gamma(VL)_{ab}}{v_{ab}^t}\left(v-v_{ab}^t\right)^2,
\end{aligned}
\tag{26}
$$

where $J'_{v_{ab}}$ is the first order derivative with respect to V, is an auxiliary function of $J_{v_{ab}}$.

The proofs of both ***Lemma 2*** and ***Lemma 3*** are deduced in **Appendix A** and **Appendix B** for the smoothness of logic.

According to ***Lamma 1*** and ***Lemma 2***, $F_U\left(\arg\min_U G(U,U'),V\right)\le F_U(U,V)$. Let $\frac{\partial G(U,U')}{\partial u_{ij}}=0$, we have the multiplicative update rule (22). According to ***Lamma 1*** and ***Lemma 3***, $F_V\left(U,\arg\min_V G(V,V')\right)\le F_V(U,V)$. Let $\frac{\partial G(V,V')}{\partial v_{ab}}=0$, we have the multiplicative update rule (21).

4. EXPERIMENTS

In this section, we evaluate the effectiveness of the proposed CSNMF on several popular face datasets by comparing with CIMNMF [12] and MCCGR [13]. We also compared CSNMF with a supervised NMF (SNMF) to demonstrate the robustness of CSNMF. SNMF incorporates the identical discriminative information like CSNMF, i.e.,

$$
\min_{U\ge 0,V\ge 0}\frac{1}{2}\|X-UV\|_F^2+\frac{\gamma}{2}tr(\mathrm{VLV}^T), \tag{27}
$$

where $\|\cdot\|_F$ signifies the Frobenius norm, and L is defined as the same alignment matrix as (12).

4.1. Experimental Setting

Our experiments are conducted on the Yale [18], UMIST [19] and ORL [20] datasets. Each experiment was test on the dataset which was divided to training set and test set. To sufficiently compare the performance, we use different sizes of training sets to learn the lower dimensional space. And in order to obtain better performance, it is important to choose a proper tradeoff parameter γ in (12), here we set it to be 10^{-5} in all experiments. In the classification stage, we applied the nearest neighbor (NN) rule as a classifier to determine the labels of test samples.

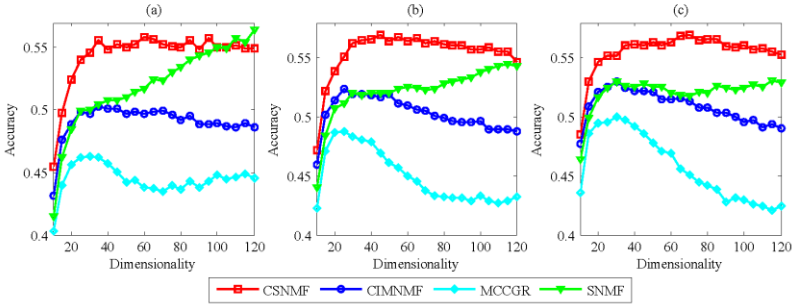

Figure 1. Face recognition accuracy on the Yale dataset. We randomly selected (a) 4, (b) 5, and (c) 6 images from each subject to learn the lower dimensional space and evaluate on the remaining images. The average results of 10 runs are reported.

4.2. Face Recognition

Yale dataset: The Yale [18] dataset contains 165 frontal view images from 15 subjects. Each one was taken 11 photos with varying facial expressions. And each image was normalized to 32 x 32 pixel array then reformulated to a vector form. Figure 1 shows the average accuracy of CSNMF, MCCGR, SNMF and CIMNMF. Their dimensions ranged from 10 to 120, and we random selected 4, 5 and 6 images for each individual for training. Table 1 records the highest average classification accuracy and their corresponding dimension. The experimental results show that our CSNMF is significantly superior to other algorithms in most cases. The comparison between CSNMF and SNMF shows that the utilized CIM successfully filters out outliers. The comparison between CSNMF and CIMNMF shows that the incorporated labels enhances the discriminative ability of the learned subspace. The comparison between CSNMF and MCCGR reflects that the adjacent graph constructed purely on labels are more robust than that constructed on features. In summary, CSNMF can fully takes advantage of the labels of dataset meanwhile take off the influence of the noises from both samples and constructed adjacent graphs.

Table 1. The highest average face recognition accuracies on the Yale dataset.

Algorithm	**6**	**7**	**9**
CSNMF	0.5705(35)	**0.5994(30)**	**0.6453(40)**
CIMNMF	0.5229(30)	0.5694(30)	0.6060(35)
MCCGR	0.4719(30)	0.5400(15)	0.5820(20)
SNMF	**0.5852(115)**	0.5661(120)	0.6033(65)

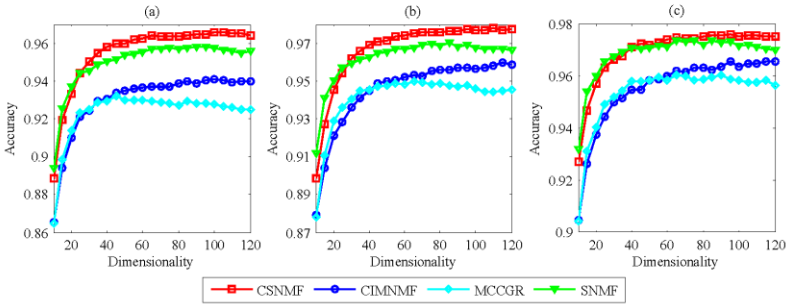

Figure 2. Face recognition accuracy on the UMIST dataset. We randomly selected (a) 6, (b) 7, and (c) 9 images from each subject to learn the lower dimensional space and evaluate on the remaining images. The average results of 10 runs are reported.

UMIST dataset: The UMIST [19] dataset contains 575 images from 20 subjects. Each one holds 41 to 82 images which varying in poses from profile to frontal views， and each image was normalized to 40 x 40 pixel array then reformulated to a vector form. Fig.2 shows the average accuracy of CSNMF, SNMF, MCCGR and CIM-NMF. Their dimensions range from 10 to 120, and we random selected 6, 7 and 9 images for each subject to comprise the training set. Table 3 records the highest average face recognition accuracy and their corresponding dimension. The experimental results show that our CSNMF is superior to other two algorithms mostly because it simultaneously takes advantages of the robustness of CIM and the discriminative information of labels of the dataset.

Table 2. The highest average face recognition accuracies on the UMIST dataset.

Algorithm	**6**	**7**	**9**
CSNMF	**0.9839(85)**	**0.9972(75)**	**1.0000(75)**
CIMNMF	0.9750(115)	0.9869(95)	0.9942(110)
MCCGR	0.9697(45)	0.9800(105)	0.9812(45)
SNMF	0.9769(75)	0.9931(110)	0.9963(60)

ORL dataset: The Cambridge ORL dataset consists of 400 images from 40 subjects, and each subject hold 10 images with varying lighting, facial expressions, and facial details. Each image was normalized to 32 x 32 pixel array then reformulated to a vector form. Fig.3 shows the average accuracy of CSNMF, SNMF, MCCGR and CIMNMF. Their dimensions range from 10 to 120, and we random selected 4, 5 and 6 images for each subject in the learning procedure. Table 4 records the highest average face recognition accuracy and their corresponding dimension. The experimental results show that our CSNMF is superior to SNMF, MCCGR and CIMNMF, and it confirms that CSNMF simultaneously takes their advantages without introducing their disadvantages.

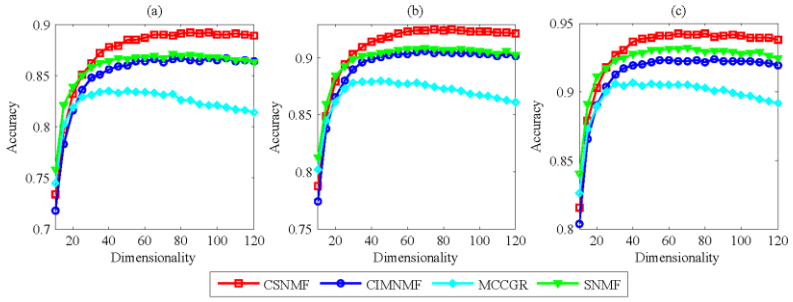

Figure 3. Face recognition accuracy on the ORL dataset. We randomly selected (a) 4, (b) 5, and (c) 6 images from each subject to learn the lower dimensional space and evaluate on the remaining images. The average results of 10 runs are reported.

Table 3. The highest average face recognition accuracy on the ORL dataset.

Algorithm	**6**	**7**	**9**
CSNMF	**0.8900(110)**	**0.9552(110)**	**0.9544(90)**
CIMNMF	0.8640(105)	0.9320(100)	0.9422(90)
MCCGR	0.8363(45)	0.9137(70)	0.9284(45)
SNMF	0.8869(80)	0.9400(75)	0.9478(55)

5. CONCLUSIONS

In this paper, we proposed a correntropy supervised non-negative matrix factorization (CSNMF) method for learning the discriminative lower dimensional space from noisy datasets. Since CSNMF can take advantages of the robustness of maximum correntropy criterion and discriminant power of the labels of dataset under the non-negative patch alignment framework, it outperforms both robust variant and supervised variant of NMF. We developed a multiplicative update rule to solve CSNMF and proved that it can monotonically decreases the objective function of CSNMF.

REFERENCES

[1] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization.” Nature, vol. 401, no. 6755, pp. 781-791, 1999.

[2] C. Liu, H. Yang, J. Fan, L. He, and Y. Wang, “Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce,” Procedings of the 19th international conference on World wide web 10, 2010.

[3] S. Li, X. W. H. X. W. Hou, H. J. Z. H. J. Zhang, and Q. S. C. Q. S. Cheng, “Learning spatially localized, parts-based representation,” Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition., vol. 1, 2001.

[4] A. Pascual-Montano, J. M. Carazo, K. Kochi, D. Lehmann, and R. D. Pascual-Marqui, “Nonsmooth nonnegative matrix factorization (nsNMF),” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, pp. 403-415, 2006.

[5] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Hunag, “Constrained nonnegative matrix factorization for image representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence., vol. 34, pp. 1299-1311, 2012.

[6] S. Zafeiriou, “Discriminant nonnegative tensor factorization algorithms,” IEEE Transactions on Neural Networks, vol. 20, pp. 217-235, 2009.

[7] N. Guan, D. Tao, Z. Luo, and B. Yuan, “Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent,” IEEE Transactions on Image Processing, vol. 20, no. 7, pp. 2030-2048, 2011.

[8] D. D. Lee and H. S. Seung, “Algorithms for nonnegative matrix factorization,” Advances in neural information processing systems, vol. 13, pp. 556-562, 2001.

[9] Z. Yang, H. Zhang, Z. Yuan, and E. Oja, “Kullback-Leibler divergence for nonnegative matrix factorization,” in Lecture Notes in Conputer Science (includeing subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6791 LNCS, 2011, pp. 250-257.

[10] W. Liu, P. P. Pokharel, and J. C. Principe, “Correntropy: Properties and applications in non-Gaussian signal processing,” IEEE Transactions on Signal Processing, vol. 55, no. 11, pp. 5286-5298, 2007.

[11] N. Guan, D. Tao, Z. Luo, and B. Yuan, “Non-negative patch alignment framework,” IEEE Transactions on Neural Networks, vol. 22, no. 8, pp. 1218-1230, Aug 2011.

[12] L. Du, X. Li, and Y. D. Shen, “Robust nonnegative matrix factorization via half-quadratic minimization,” in Proceedings of IEEE International Conference on Data Mining, ICDM, 2012, pp. 201-210.

[13] L. Le, Y. J, Z. K, and et al, “Graph Regularized Non-negative Matrix Factorization By Maximizing Correntropy.” Ar Xiv preprint arXiv, vol. 1405, no. 2246, 2014.

[14] H. Jin, N. Feiping, H. Heng, and C. Ding, “Robust Manifold Nonnegative Matrix Factorization,” ACM Transations Knowledge Discovery Data, vol. 8, no.3, 2014.

[15] X. Yuan and B. Hu, “Robust Feature Extraction via Information Theoretic Learning,” in Proceedings of the 26the International Conference on Machine Learning, 2009, pp. 1193-1200.

[16] R. He, W. S. Zheng, and B. G. Hu, “Maximum correntropy criterion for robust face recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, pp. 1561-1576, 2011.

[17] D. P Bertsekas, Nonlinear Programming, 2nd ed. MA : Athena Scientic: Belmount, 1999.

[18] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711-720, 1997.

[19] D. B. Graham and N. M. Allinson, “Characterizing virtual eigensignatures for general purpose face recognition,” Proceedings of Face Recognition: Theory Application., vol. 163, pp. 446-456, 1998.

[20] F.S. Samaria and A. C. Harter, “Parameterisation of a stochastic model for human face identification,” Proceedings of 1994 IEEE Workshop on Applicaiotns of Computer Vision, vol. 13, pp. 138-142, Dec 1994.

APPENDICES

Appendix A: Proof of Lemma 2

It is obvious that $G(u,u)=J_{u_{ij}}(u)$. With the Taylor series expansion of $J_{u_{ij}}(u)$, we can obtain

$$J_{u_{ij}}(u)=J_{u_{ij}}\left(u_{ij}^{t}\right)+J'_{u_{ij}}\left(u_{ij}^{t}\right)\left(u-u_{ij}^{t}\right)+\frac{1}{2}J''_{u_{ij}}\left(u-u_{ij}^{t}\right)^{2} \tag{28}$$

And

$$\begin{aligned}
J''_{u_{ij}} &= \frac{\partial^{2}\mathrm{L}}{\partial u_{ij}^{2}} \\
&= \frac{1}{mn\sigma^{2}}\sum_{k}\left\{e^{-\frac{\left(X_{ik}-\sum_{r}u_{ir}v_{rk}\right)^{2}}{2\sigma^{2}}}\left[1-\frac{1}{\sigma^{2}}\left(X_{ik}-\sum_{r}u_{ir}v_{rk}\right)^{2}\right]\right\}v_{jk}^{2} \quad . \\
&= \frac{1}{mn\sigma^{2}}\left\{\left[W\otimes\left(I_{W}-\frac{1}{\sigma^{2}}(X-UV)\otimes(X-UV)\right)\right](V\otimes V)\right\}_{ij}
\end{aligned}$$

Then we can see that,

$$\begin{aligned}
\left[W\otimes(UV)V^{T}\right]_{ij} &= \sum_{k}w_{ik}v_{jk}\sum_{r}u_{ir}v_{rk} \geq \sum_{k}w_{ik}v_{jk}v_{jk}u_{ij} = \left[W(V\otimes V)\right]_{ij}u_{ij} \\
&\geq \left\{\left[W\otimes\left(I_{W}\text{-}\frac{1}{\sigma^{2}}(X-UV)\otimes(X-UV)\right)\right](V\otimes V)\right\}_{ij}u_{ij}
\end{aligned} \tag{27}$$

So we have $G\left(u,u_{ij}^{t}\right)\geq J_{u_{ij}}(u)$. This completes the proof. □

Appendix B: Proof of Lemma 3

It is obvious that $G(v,v)=J_{v_{ab}}(v)$. With the Taylor series expansion of $J_{v_{ab}}(v)$, we have

$$J_{v_{ab}}(v)=J_{v_{ab}}\left(v_{ab}^{t}\right)+J'_{v_{ab}}\left(v_{ab}^{t}\right)\left(v-v_{ab}^{t}\right)+\frac{1}{2}J''_{v_{ab}}\left(v-v_{ab}^{t}\right)^{2}, \tag{29}$$

And

$$\begin{aligned}
J''_{v_{ab}} &= \frac{\partial^{2}\mathrm{L}}{\partial v_{ab}^{2}} \\
&= \frac{1}{mn\sigma^{2}}\sum_{k}u_{ka}^{2}e^{-\frac{\left(X_{ik}-\sum_{r}u_{ir}v_{rk}\right)^{2}}{2\sigma^{2}}}\left[1-\frac{1}{\sigma^{2}}\left(X_{kb}-\sum_{r}u_{kr}v_{rb}\right)^{2}\right]+\gamma L_{bb} \quad . \\
&= \frac{1}{mn\sigma^{2}}\left\{(U\otimes U)^{T}\left[\left(I_{W}-\frac{1}{\sigma^{2}}(X-UV)\otimes(X-UV)\right)\otimes W\right]\right\}_{ab}+\gamma L_{bb}
\end{aligned}$$

Then we can see that

$$\begin{aligned}&\left\{U^T\left[\left(X-UV\right)\otimes W\right]\right\}_{ab}\Big/mn\sigma^2+\gamma\left(VL\right)_{ab}=\sum_k u_{ka}w_{kb}\sum_r u_{kr}v_{rb}\Big/mn\sigma^2+\gamma\sum_k v_{ak}L_{kb}\\&\geq\sum_k u_{ka}w_{kb}u_{ka}v_{ab}\Big/mn\sigma^2+\gamma v_{ab}L_{bb}=\left[\left(U\otimes U\right)^T W\right]_{ab}v_{ab}\Big/mn\sigma^2+\gamma L_{bb}v_{ab}\\&\geq\left\{\left[\left(I_W-\frac{1}{\sigma^2}\left(X-UV\right)\otimes\left(X-UV\right)\right)\otimes W\right]\right\}_{ab}v_{ab}\Big/mn\sigma^2+\gamma L_{bb}v_{ab}.\end{aligned}\tag{30}$$

So we have $G\left(v,v_{ab}^t\right)\geq J_{v_{ab}}\left(v\right)$. This completes the proof. □

AUTHORS

Hang Cheng: He is currently studying for a master's degree at the Research Centre for Artificial Intelligence, National Innovation Institute of Defense Technology. His research interests include computer vision, few-shot object detection and machine learning.

Shixiong Wang: received the B.S. degree in 2013 from Tsinghua University, Beijing, China, and the Ph.D. degree in 2019 from National University of Defense Technology, Changsha, China. He is now an assistant professor with National Innovation Institute of Defense Technology, Beijing, China. His research fields include computer science and machine learning.

Naiyang Guan: received the B.S., M.S., and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 2004, 2006, and 2011, respectively. He is currently an Associate Professor with the Defense Innovation Institute (DII), Academy of Military Sciences, Beijing, China. He has authored or coauthored over 60 research articles on top-tier journals, including the IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), IEEE Transactions on Neural Networks and Learning Systems (T-NNLS), IEEE Transactions on Image Processing (T-IP), and IEEE Transactions on Signal Processing (T-SP), and top-tier conferences, including the IEEE International Conference on Data Mining (ICDM), International Joint Conference on Artificial Intelligence (IJCAI), European Conference on Computer Vision (ECCV), and International Joint Conference on Neural Networks (IJCNN). His research interests include machine learning, computer vision, and data mining.