An Entropy Weighted Nonnegative Matrix Factorization Algorithm for Feature Representation

Jiao Wei, Can Tong, Bingxue Wu, Qiang He, Yudong Yao, Fellow, IEEE, and Yueyang Teng

Abstract—Nonnegative matrix factorization (NMF) has been widely used to learn low-dimensional representations of data. However, NMF pays the same attention to all attributes of a data point, which inevitably leads to inaccurate representations. For example, in a human-face dataset, if an image contains a hat on a head, the hat should be removed or the importance of its corresponding attributes should be decreased during matrix factorization. This article proposes a new type of NMF called entropy weighted NMF (EWNNMF), which uses an optimizable weight for each attribute of each data point to emphasize their importance. This process is achieved by adding an entropy regularizer to the cost function and then using the Lagrange multiplier method to solve the problem. Experimental results with several datasets demonstrate the feasibility and effectiveness of the proposed method. The code developed in this study is available at https://github.com/Poisson-EM/Entropy-weighted-NMF.

Index Terms—Clustering, entropy regularizer, low-dimensional representation, nonnegative matrix factorization (NMF).

I. INTRODUCTION

WITH the rapid development of data acquisition technology, large amounts of data, such as online documents, medical images, various video series, traffic data, health data, and other high-dimensional data, are accumulating. Therefore, dimensionality reduction [1] has become an essential step in data mining. Vector quantization (VQ) [2], singular value decomposition (SVD) [3], principal component analysis (PCA) [4], independent component analysis (ICA) [5], and concept factorization (CF) [6], [7] are some of the most commonly used dimensionality reduction methods. However, due to negative components, these methods typically cannot be reasonably explained in some practical problems. Therefore, developing a nonnegative factorization method is valuable for research. Many researchers have investigated the nonnegative matrix factorization (NMF) proposed by Lee and Seung [8]. This method represents a nonnegative matrix by a product of two low-rank nonnegative matrices and then learns a part-based representation. NMF has been applied in many fields, including clustering, dimensionality reduction, and blind source separation [9]–[12].

In recent years, many variants of NMF have been proposed to extend its applicable range. For example, Ding et al. [13] learned new low-dimensional features from data with convenient clustering interpretation using Semi-NMF, which allows the data matrix and the base matrix to have mixed signs [14], [15]. They also developed Convex-NMF by restricting the base vectors to a convex combination of data points. Considering that orthogonality constraint leads to sparsity, Ding et al. [16] proposed an orthogonal NMF (ONMF) method, which adds orthogonality constraints to the base and representation matrices. There are two ways to solve ONMF: the Lagrange method [16] and the natural gradient on the Stiefel manifold [17]. To make the NMF method more robust to noise, Gao et al. [18] proposed using the capped norm for the cost function and removing the outliers using a threshold. For the same purpose, Guan et al. [19] proposed a truncated CauchyNMF method to manage outliers by truncating large errors, which learns subspaces on datasets subject to large-scale noise or corruption. Because the cost function contains a nonlinear logarithmic function that is difficult to optimize, the authors use half quadratic (HQ) programming based on convex conjugate theory to optimize the method.

These studies place equal importance on all features of the data points. Therefore, some researchers have addressed the importance of the features of data points. Blondel et al. [20] added predetermined weights to each attribute of each data point, demonstrating that weights can produce important flexibility by better emphasizing certain features in image approximation problems. Then, Kim and Choi [21] proposed a new weighted NMF (WNMF) method to process an incomplete data matrix with missing entries, which combined the binary weights into the NMF multiplication update. Wang et al. [22] proposed a new WNMF method, which sets the weight of a data point as the product of column weights and row weights. These methods are referred to as “hard WNMF”; however, one primary drawback is that the weights that they rely on must be predetermined. In 2021, Chen et al. [23] proposed...
a new NMF that automatically projected the samples into the subspace with a transformation matrix and used a graph regularization term to capture the data manifold. Then, they used the constrained Laplacian rank algorithm to solve the problems. However, two weighted matrices are involved in the cost function, and a complex algorithm is used to obtain the solution of the weights, which leads to an algorithm with high complexity.

This article presents an entropy weighted NMF (EWNMF) method that assigns a weight to indicate the importance of each attribute of one data point in matrix factorizing. Note that the proposed method is markedly different from [23], in which the former independently enhances every image, and the latter extracts the features from the entire space. Then, the entropy of these weights is used to regularize the cost function for obtaining an easily computable solution, which makes the range of weights fall within [0, 1] with a summation of 1; thus, the weights can be explained as the probability of the contribution of an attribute of one data point to NMF. The experimental results with several real datasets show that EWNMF performs better than other NMF variants.

II. METHODOLOGY

A. Related Research

NMF is a matrix factorization method that focuses on data matrices with nonnegative elements, which can reveal hidden structures and patterns from generally redundant data. We will review the standard NMF as follows.

Notations: In this article, matrices are denoted as capital letters. For a matrix $A$, $A_{ai}$, $A_{ia}$, and $A_{ij}$ denote the $i$th column, the $i$th row, and $(i, j)$th element of $A$, respectively; the Frobenius norm is represented as $\|A\|_F$; $\odot$ and $\cdot$ mean the item-by-item multiplication and division of two matrices, respectively; $A^T$ denotes the transpose of $A$; $Tr(\cdot)$ denotes the trace of a matrix; and $A \geq 0$ means that all the elements of $A$ are equal to or larger than 0.

The expression of NMF is

$$X \approx WH$$

(1)

where the matrix $X \subseteq R^{M \times N}$ denotes the given nonnegative matrix in which each column is a data point. The goal of NMF is to find two low-dimensional nonnegative matrices: $W \subseteq R^{M \times K}$ is called the base matrix, and $H \subseteq R^{K \times N}$ is called the representation matrix, whose product can approximate the original matrix [24], [25], where $K \ll \min\{M, N\}$.

There are different standards to measure the quality of decomposition. Lee and Seung proposed using the square of the Euclidean distance and the Kullback-Leibler divergence. In this article, the square of the Euclidean distance is used, and the formula is expressed as

$$\min F_1(W, H) = \|X - WH\|_F^2$$

(2)

s.t. $W \geq 0, \quad H \geq 0.$

To alternatively minimize $W$ and $H$ in (2), the construction of the auxiliary function is important to determine the iterative update rule.

**Definition 1 (Auxiliary Function):** If the function $G(h, h')$ satisfies the following conditions:

$$G(h, h') \geq F(h) \quad \text{and} \quad G(h', h') = F(h')$$

then, $h'$ is a given value. Then, $G(h, h')$ is the auxiliary function of $F(h)$ on $h'$. Then, we can draw the following conclusion.

**Lemma 1:** If $G(h, h')$ is an auxiliary function of $F(h)$, then we have

$$h^* = \arg \min_h G(h, h')$$

(4)

and the function $F(h)$ does not increase.

**Proof:** The conditions satisfied by the auxiliary function make this proof marked because

$$F(h^*) \leq G(h^*, h') \leq G(h', h') \leq F(h').$$

(5)

Then, we construct the update rule for the NMF problem. We consider $W$ first, where $W^T \geq 0$ and $H \geq 0$ are given. Let $\xi_{ijk} = W^T_{ik}H_{kj}/(WH)_{ij}$, of course, $\xi_{ijk} \geq 0$ and $\sum_{k=1}^{K} \xi_{ijk} = 1$. Therefore, the auxiliary function of standard NMF is

$$f_i(W, W') = \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{K} \xi_{ijk} (X_{ij} - W_{ik}H_{kj}^2)$$

(6)

Because the function is separable, it can be easily minimized. We thus take the partial derivative of (6) and set it to zero so that we can obtain the following update rule:

$$W \leftarrow W \odot (XH^T) \cdot (WHH^T).$$

(7)

The method of constructing the auxiliary function of $H$ is similar to that of $W$. Then, the following update to $H$ is obtained:

$$H \leftarrow H \odot (W^TX) \cdot (W^TWH).$$

(8)

B. Proposed Method

Different from previous methods, an optimizable weight matrix is used to measure the importance of the attributes in matrix factorizing. Fig. 1 shows the explanation of the proposed idea that if there is a hat in the only image, the corresponding attributes of the hat must destroy the result of NMF, and we should certainly eliminate or weaken their importance. Thus, it may be necessary to provide different importance to the attributes of each data point based on the following constraint, which will lead to a new optimization model:

$$\min F_2(W, H, T) = \sum_{i=1}^{M} \sum_{j=1}^{N} T_{ij} |X_{ij} - (WH)_{ij}|^2$$

s.t. $W \geq 0, \quad H \geq 0, \quad T \geq 0, \quad \sum_{i=1}^{M} T_{ij} = 1.$

(9)

We first consider the new variable $T$, which can be solved in an alternative optimization manner. For fixed $W$ and $H$, $T_{ij}$ is very easy to solve as $T_{ij} = 1$ if
Fig. 1. Diagram of the proposed method: a hat in the first image destroys the accuracy of feature representation, which should be assigned no or little importance.

$E_{ij} = \min \{|E_{ij}|, |E_{j1}|, \ldots, |E_{Mj}|\}$ or 0 otherwise\(^1\), where $E = X - WH$, and it demonstrates the simple fact that only one element of $T_{ij}$ is 1, and the others are 0, which is incompatible with the real problem.

To address this issue, we apply an entropy regularizer to penalize the cost function of NMF to obtain a weight in the range of $[0, 1]$ instead of 0 or 1, which uses information entropy to calculate the uncertainty of weights. The new optimization problem is rewritten as follows:

$$
\min F_3(W, H, T) = \sum_{i=1}^{M} \sum_{j=1}^{N} T_{ij} |X_{ij} - (WH)_{ij}|^2
$$

$$
+ \gamma \sum_{i=1}^{M} \sum_{j=1}^{N} T_{ij} \ln(T_{ij})
$$

s.t. $W \geq 0$, $H \geq 0$, $T \geq 0$, $\sum_{i=1}^{M} T_{ij} = 1$  \quad (10)

where $\gamma \geq 0$ is a given hyperparameter. The first term in (10) is the sum of errors, and the second term is the negative entropy of the weights. In theory, entropy can be interpreted as the degree of disorder or randomness in the system. In the cost function, the minimization of the second term indicates that the features can be extracted in as many dimensions as possible. The original cost function in (9) results in only one attribute of each data point being involved in feature representation, and the entropy regularizer will stimulate more attributes to help feature representation.

This equation can be solved by a simple algorithm, which is based on the following proposition.

**Proposition 1:** Given the matrices $W$ and $H$, $T_{ij}$ in (10) is minimized when

$$
T_{ij} = \frac{e^{-\frac{|X_{ij} - (WH)_{ij}|^2}{\gamma}}}{\sum_{i=1}^{M} e^{-\frac{|X_{ij} - (WH)_{ij}|^2}{\gamma}}}.
$$  \quad (11)

\(^1\)The update rule can be easily explained by an example as

$$
\min \{3, 1, 2\} = \min 3T_1 + 1T_2 + 2T_3
$$

s.t. $T_1 \geq 0$, $T_2 \geq 0$, $T_3 \geq 0$

$T_1 + T_2 + T_3 = 1.$

The solution is that $T_1 = 0$, $T_2 = 1$, and $T_3 = 0$, in which $T_2$ corresponds to the minimum value of $\{3, 1, 2\}$. This process is similar to the computation of weights in the k-means algorithm.

**Proof:** We construct the Lagrange function of (10) with respect to $T$ as

$$
L(T, \lambda) = \sum_{i=1}^{M} \sum_{j=1}^{N} T_{ij} |X_{ij} - (WH)_{ij}|^2 + \gamma \sum_{i=1}^{M} \sum_{j=1}^{N} T_{ij} \ln(T_{ij})
$$

$$
- \sum_{j=1}^{N} \lambda_{j} \left( \sum_{i=1}^{M} T_{ij} - 1 \right)
$$

where $[\lambda_1, \lambda_2, \ldots, \lambda_N]$ is a vector containing the Lagrange multipliers corresponding to the constraints.

By setting the gradient of (12) with respect to $\lambda_j$ and $T_{ij}$ to zero, we obtain the following equation system

$$
\frac{\partial L}{\partial \lambda_{j}} = \sum_{i=1}^{M} T_{ij} - 1 = 0
$$

$$
\frac{\partial L}{\partial T_{ij}} = |X_{ij} - (WH)_{ij}|^2 + \gamma \ln(T_{ij}) + \gamma - \lambda_{j} = 0.
$$  \quad (14)

From (14), we know that

$$
T_{ij} = e^{-\frac{\gamma - \lambda_{j}}{\gamma}} \cdot e^{-\frac{|X_{ij} - (WH)_{ij}|^2}{\gamma}}.
$$  \quad (15)

Substituting (15) into (13), we have

$$
\sum_{i=1}^{M} T_{ij} = e^{-\frac{\gamma - \lambda_{j}}{\gamma}} \cdot \sum_{i=1}^{M} e^{-\frac{|X_{ij} - (WH)_{ij}|^2}{\gamma}} = 1.
$$  \quad (16)

It follows that:

$$
e^{-\frac{\gamma - \lambda_{j}}{\gamma}} = \frac{1}{\sum_{i=1}^{M} e^{-\frac{|X_{ij} - (WH)_{ij}|^2}{\gamma}}}.
$$  \quad (17)

Substituting this expression into (15), we find that

$$
T_{ij} = e^{-\frac{|X_{ij} - (WH)_{ij}|^2}{\gamma}} \cdot \frac{1}{\sum_{i=1}^{M} e^{-\frac{|X_{ij} - (WH)_{ij}|^2}{\gamma}}}.
$$  \quad (18)

Then, we can solve $W$ and $H$ with fixed $T$, which is similar to the standard NMF method. For example, we can construct the following auxiliary function about $W$:

$$
f_3(W, W') = \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{K} T_{ijk} |X_{ij} - \frac{W_{ik}H_{jk}}{\zeta_{ijk}}|^2
$$

$$
+ \gamma \sum_{i=1}^{M} \sum_{j=1}^{N} T_{ij} \ln(T_{ij}).
$$  \quad (19)

Setting the partial derivative of $f_3(W, W')$ to zero yields the following update rule:

$$
W \leftarrow W \odot (T \odot X)H^T / \{T \odot (WH)H^T \}.
$$  \quad (20)

Similarly, we can also easily obtain the update rule for $H$ as follows:

$$
H \leftarrow H \odot W^T (T \odot X) / \{W^T[T \odot (WH)] \}.
$$  \quad (21)

The update rules to $W$ and $H$ are similar to the existing WNMF methods. Optimizing EWNMF is summarized as follows in Algorithm 1.
Algorithm 1: Entropy Weighted Nonnegative Matrix Factorization (EWNMF)

**Input:** Given the input nonnegative matrix \( X \subseteq R^{M \times N} \), the number \( K \) of reduced dimensions and hyperparameter \( \gamma \).

**Output:** the weight matrix \( T \), the base matrix \( W \), and the representation matrix \( H \).

1. Randomly initialize \( W \subseteq R^{M \times K} > 0 \) and \( H \subseteq R^{K \times N} > 0 \);
2. while not convergence do
   3. Update \( T \) by (18);
   4. Update \( W \) by (20);
   5. Update \( H \) by (21);
3. end while
7. return \( T \), \( W \) and \( H \).

As shown in algorithm derivation, we use a distribution optimization strategy to solve \( T \), \( W \), and \( H \). It is key to solve \( T \), which uses the Lagrange multiplier method to obtain a computable and interpretable solution. When some elements of a matrix can meet factorization, i.e., \([X - (WH)]_{ij}\) approaches zero, more attention should be paid.

C. Extensions

The proposed entropy weighted method can also be applied to NMF methods with different measures instead of the square of the Euclidean distance, such as KL divergence and \( \gamma \) divergence NMF, its cost function can be expressed as

\[
F_4(T, W, H) = \frac{1}{\alpha(\alpha - 1)} \sum_{i=1}^{M} \sum_{j=1}^{N} T_{ij} \times [X_{ij}^{\alpha}(WH)_{ij}^{1-\alpha} - \alpha X_{ij} + (\alpha - 1)(WH)_{ij}] + \gamma \sum_{i=1}^{M} \sum_{j=1}^{N} T_{ij} \ln(T_{ij}) \tag{22}
\]

where \( \alpha \in R \) is a given value. The update method of \( T \) is as follows:

\[
T_{ij} = \frac{e^{\frac{[X_{ij}^{\alpha}(WH)_{ij}^{1-\alpha} - \alpha X_{ij} + (\alpha - 1)(WH)_{ij}]}{\gamma}}}{\sum_{i=1}^{M} e^{\frac{[X_{ij}^{\alpha}(WH)_{ij}^{1-\alpha} - \alpha X_{ij} + (\alpha - 1)(WH)_{ij}]}{\gamma}}}. \tag{23}
\]

Note that \( \alpha \) divergence is not generally a distance between two vectors because it cannot guarantee the symmetry satisfied. \( \alpha \) divergence includes many existing NMF; for example, when \( \alpha = 1 \), it is the well-known KL divergence NMF; when \( \alpha = -1 \), it is \( \chi^2 \) divergence NMF.

Introducing the entropy regularizer technique into GNMF [26], its cost function can be expressed as

\[
F_5(T, W, H) = \sum_{i=1}^{M} \sum_{j=1}^{N} T_{ij} [X_{ij} - (WH)_{ij}]^2 \tag{24}
\]

where \( L = D - V \), which is called the graph Laplacian matrix, \( D \) is a diagonal matrix whose entries are column sums of \( V \) \((D_{jj} = \sum_{i} V_{ij})\), and \( V \) is a weight matrix. Therefore, \( T \) is updated iteratively as

\[
T_{ij} = \frac{e^{\frac{[X_{ij}^{\alpha}(WH)_{ij}^{1-\alpha} - \alpha X_{ij} + (\alpha - 1)(WH)_{ij}]}{\gamma}}}{\sum_{i=1}^{M} e^{\frac{[X_{ij}^{\alpha}(WH)_{ij}^{1-\alpha} - \alpha X_{ij} + (\alpha - 1)(WH)_{ij}]}{\gamma}}} \tag{25}
\]

We omit the update rules to \( W \) and \( H \) for the above two methods because the derivation of them is similar to the proposed algorithm.

III. EXPERIMENTS

A. Experimental Description

The experiments were performed on an HP Compaq PC with a 3.40-GHz Core i7-6700 CPU and 16 GB of RAM, and all the methods were implemented in MATLAB. We compare the performance of the proposed methods with NMF [8], ONMF [16], Semi-NMF [13], Convex-NMF [13], RNMF [18], and CauchyNMF [19] on five public datasets, including the Yale, UMISTFace, Caltech101, GTDF, and TDT2 datasets.

The Yale face dataset [27] contains 165 face images captured from 15 persons. Each person has 11 face images with different expressions, postures, and lighting. The UMISTFace dataset [28] has a total of 1012 images, including 20 people, each with different angles and different poses.

The Caltech101 dataset [29] contains 9144 images split between 101 different object categories, as well as an additional background/clutter category. Most categories have approximately 50 images.

GTDF [30] contains 750 images taken in two different sessions and includes 50 people. The images show faces with different facial expressions and lighting conditions.

The TDT2 Audio Corpus [31] contains six sources, including two news special lines (APW and NYT), two radio programs (VOA and PRI), and two TV programs (CNN and ABC). In this experiment, only the largest 30 categories are used (a total of 9394 documents).

Important details of these datasets are shown in Table I.

| Data set       | Yale | UMISTFace | Caltech101 | GTDF | TDT2 |
|----------------|------|-----------|------------|------|------|
| Points         | 165  | 1012      | 9144       | 750  | 9394 |
| Dimensions     | 1024 | 1024      | 1024       | 1024 | 36771|
| Class          | 15   | 20        | 101        | 50   | 30   |

After obtaining a new feature representation, we use k-means to cluster them and then compare it with the label to evaluate the clustering results. Clustering accuracy (ACC) [32], [33] and normalized mutual information (NMI) [34], [35] are used to evaluate the performance of these clustering results.
Given a set of ground true class labels $y$ and obtained cluster labels $y'$, the clustering accuracy is defined as

$$\text{ACC} = \frac{\sum_{i=1}^{N} \delta(y_i, \text{map}(y'_i))}{N}$$

(26)

where

$$\delta(a, b) = \begin{cases} 1, & a = b \\ 0, & \text{otherwise} \end{cases}$$

and map(·) is a permutation mapping function that maps the obtained cluster labels to real labels. The higher the ACC value, the better the clustering performance.

NMI is used to calculate the agreement between the two data distributions and is defined as follows:

$$\text{NMI}(y, y') = \frac{\text{MI}(y, y')}{\max(H(y), H(y'))}$$

(27)

where $H(y)$ is the entropy of $y$. $\text{MI}(y, y')$ quantifies the amount of information between two random variables (i.e., $y$ and $y'$) and is defined as

$$\text{MI}(y, y') = \sum_{y_i \in y, y'_j \in y'} p(y_i, y'_j) \log \left( \frac{p(y_i, y'_j)}{p(y_i)p(y'_j)} \right)$$

(28)

where $p(y_i)$ and $p(y'_j)$ are the probabilities that a data point selected from the dataset belongs to clusters $y_i$ and $y'_j$, respectively, and $p(y_i, y'_j)$ is the joint probability that an arbitrarily selected data point belongs to clusters $y_i$ and $y'_j$ concurrently. The NMI score ranges from 0 to 1, and the larger the NMI, the better the clustering performance.

Before the experiment, we normalized all the datasets to scale the minimum and maximum values of each data point to 0 and 1, respectively. All the methods used the same random distribution for initialization of $W$ and $H$ to make them uniformly distributed on [0.1 1.1] and performed 300 iterations to ensure sufficient convergence.

B. Experimental Results

1) Signal Unmixing on Synthetic Data: We use a synthetic dataset to simulate a signal-mixture process, which mixes two source signals by a uniformly distributed matrix on [0 1] and then destroys the beginning part of the first one. The destroyed part in the mixed signal hinders signal recovery. Then, we use EWNMF to unmix the signals to demonstrate the usefulness of the proposed entropy weighted method, where the hyperparameter $\gamma$ is set to 0.01. Fig. 2 shows the source
signals, mixed signals including the destroyed signal, and the obtained weights. The weights of the destroyed elements in the first signal are small and can even be considered zero, and the other weights are similar. Thus, we can conclude that EWNMF can provide correct weights to the importance of attributes in the dataset.

2) EWNMF Compared With the Standard NMF: We investigate the ability of the entropy weighted strategy to improve
Fig. 7. Clustering performance versus cluster number on the Caltech101 dataset: (a) AC and (b) NMI.

Fig. 8. Clustering performance versus cluster number on the GTFD dataset: (a) AC and (b) NMI.

Fig. 9. Clustering performance versus cluster number on the TDT2 dataset: (a) AC and (b) NMI.

the performance of the standard NMF. We apply NMF and EWNMF to reduce the dimension number of the Yale and UMISTface datasets and then use k-means to cluster these new representations. The number of reduced dimensions is equal to that of the clusters. Figs. 3 and 4 show the clustering results evaluated by ACC and NMI. EWNMF indeed provides better performance than the standard NMF and can achieve a consistently superior performance to the standard NMF.
Fig. 10. Cost function on the Yale dataset, which is obtained by (a) NMF, (b) ONMF, (c) Semi-NMF, (d) Convex-NMF, (e) RNMF, (f) CauchyNMF, and (g) EWNMF.

Fig. 11. Some base images with respect to the Yale dataset, in which the summation of every column is normalized to 1. They are obtained by (a) NMF, (b) ONMF, (c) Semi-NMF, (d) Convex-NMF, (e) RNMF, (f) CauchyNMF, and (g) EWNMF.

on a wide range of hyperparameters $\gamma$, demonstrating the robustness of EWNMF.

3) Clustering Results With Different Cluster Numbers:
We study the relationship between evaluation standards and
cluster number. The hyperparameter $\gamma$ is selected in $\{10^i, i = -8, -7, \ldots, 7, 8\}$ to obtain the optimal results in a large range.
Because the NMF problem does not have a sole solution,
we randomly initialized ten times to obtain a credible averaged
ACC and NMI. Different cluster numbers ranging from 2 to
10 are selected. In a certain dataset with $k$ clusters, the
experimental details are described as follows.

1) Randomly select $k$ categories as a subset for the follow-
ing experiment.
2) Randomly initialize $W$ and $H$, obtain new represen-
tations, and cluster them by k-means. EWNMF uses
the optimal hyperparameter $\gamma$ according to the above
instruction.
3) Repeat 1) and 2) ten times to obtain an average result.
The clustering results, ACC and NMI versus the number
of clusters, are reported in Figs. 5–9. The proposed method
can generally yield more accurate clustering results, and
only Semi-NMF and CauchyNMF on the GTFD dataset and
CauchyNMF on the Caltech101 dataset occasionally surpass the accuracy of the proposed method. In addition, none of the other methods performs better in all aspects than EWNMF.

The optimal clustering results on the entire Yale and UMISTface datasets are shown in Tables II and III. An interesting observation is highlighted and shows a consistent result with the above figures that EWNMF still provides the best evaluation standards.

4) Convergence Speed and Part-Based Learning: From the theoretical analysis described above, we can conclude that the proposed method is monotonically decreasing, but due to the nonconvexity of the cost function, it cannot be guaranteed to be strictly convergent [36]. Thus, we investigate the convergence speed of the NMF methods.

Fig. 10 shows the cost functions of NMF, ONMF, Semi-NMF, Convex-NMF, RNMF, CauchyNMF, and EWNMF on the Yale dataset. Except for the Convex-NMF method, which converges slowly, the other methods reached a stable point within 200 iterations, demonstrating that they have similar convergence speeds. However, ONMF cannot guarantee the monotonic decrease of the cost function because it uses a proximal Lagrange multiplier method. Fig. 11 shows the base images on the Yale dataset and demonstrates that Semi-NMF, Convex-NMF, and RNMF identify more global faces; thus, they do not have high clustering results on some datasets. EWNMF and CauchyNMF have similar localization.

Also, they produce similar clustering results in the Yale and GTFD datasets; however, because EWNMF introduces a weight matrix, its clustering results are better. In addition, EWNMF has poor locality compared with ONMF; however, the former has a better clustering effect than the latter.

5) Time Complexity: To further verify the performance of the selected methods, we compared the running time of the seven methods, as shown in Table IV. EWNMF outperforms the RNMF and CauchyNMF methods on all the datasets; outperforms the Convex-NMF methods on the UMISTFace, GTFD, and Caltech101 datasets; and outperforms ONMF on the TDT2 dataset. In general, the time complexity of the proposed method is at an intermediate level.

IV. CONCLUSION

This article proposes a new NMF method that adds a weight to each attribute of each data to emphasize their importance. We introduce an entropy regularizer to restrict these weights as the probabilities of importance within [0 1], which mimics the process of human reasoning more accurately. These weights can be solved by the Lagrange multiplier method, and a simple update is achieved. The experimental results show that the proposed method produces performance that is competitive with those of the existing methods.

The entropy regularizer requires an additional hyperparameter to control the certainty of the weights. In the future, we plan to develop an auto-adjustment strategy for this hyperparameter.

ACKNOWLEDGMENT

The authors would like to thank the editor and anonymous reviewers for their constructive comments and suggestions.

REFERENCES

[1] I. K. Fodor, “A survey of dimensionality reduction techniques,” in Proc. Center Appl. Sci. Comput., Lawrence Livermore Nat. Lab., 2002, pp. 1–24.
[2] A. Gersho and R. M. Gray, Vector Quantization and Signal Compression. New York, USA: Springer, 1992.
[3] C. Hu, X. Lu, M. Ye, and W. Zeng, “Singular value decomposition and local near neighbors for face recognition under varying illumination,” Pattern Recognit., vol. 64, pp. 60–83, Apr. 2017.
[4] I. T. Jolliffe, Principal Component Analysis. New York, NY, USA: Springer, 2002.
[5] A. Hyvärinen and E. Oja, “Independent component analysis: Algorithms and applications,” Neural Netw., vol. 13, nos. 4–5, pp. 411–430, Jun. 2000.
[6] S. Peng, W. Ser, B. Chen, L. Sun, and Z. Lin, “Correntropy based graph regularized concept factorization for clustering,” Neurocomputing, vol. 316, pp. 34–48, Nov. 2018.
[7] W. Yan, B. Zhang, S. Ma, and Z. Yang, “A novel regularized concept factorization for document clustering,” Knowl.-Based Syst., vol. 135, pp. 147–158, Nov. 2017.
Shouliang Qi received the doctor’s degree in philosophy from Shanghai Jiao Tong University, Shanghai, China, in 2007, and he joined the GE Global Research Center and was responsible for designing an innovative magnetic resonance imaging (MRI) systems. From 2014 to 2015, he worked as a Visiting Scholar with the Eindhoven University of Technology, Eindhoven, The Netherlands, and the Epilepsy Center Kempenhaeghe, Heeze, The Netherlands. He is currently an Associate Professor with Northeastern University, Shenyang, China. In recent years, he has been conducting productive studies in advanced MRI technology, intelligent medical imaging computing and modeling, and convergence of nano-bio-info-cogn (NBIC). He has published more than 60 papers in peer-reviewed journals and international conferences.

Dr. Qi has won many academic awards, such as the Chinese Excellent Ph.D. Dissertation Nomination Award and the Award for Outstanding Achievement in Scientific Research from Ministry of Education.

Yudong Yao (Fellow, IEEE) received the B.Eng. and M.Eng. degrees from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1982 and 1985, respectively, and the Ph.D. degree from Southeast University, Nanjing, in 1988, all in electrical engineering. He was a Visiting Student with Carleton University, Ottawa, ON, Canada, from 1987 to 1988. He has been with the Stevens Institute of Technology, Hoboken, NJ, USA, since 2000, where he is currently a Professor and the Department Director of electrical and computer engineering. He is also the Director of the Stevens’ Wireless Information Systems Engineering Laboratory (WISELAB). Previously, from 1989 to 2000, he worked with Carleton University, Ottawa; Spar Aerospace Ltd., Montreal, ON, Canada; and Qualcomm Inc., San Diego, CA, USA. He has been active in a nonprofit organization, WOCC, Inc., which promotes wireless and optical communications research and technical exchange. He holds one Chinese patent and 13 U.S. patents. His research interests include wireless communications and networking, cognitive radio, machine learning, and big data analytics.

Dr. Yao was elected an IEEE Fellow in 2011 for his contributions to wireless communications systems. He was an IEEE ComSoc Distinguished Lecturer from 2015 to 2018. In 2015, he was elected a fellow of the National Academy of Inventors. He has served as the WOCC President from 2008 to 2010 and the Chairperson of the board of trustees from 2010 to 2012. He was an Associate Editor of IEEE COMMUNICATIONS LETTERS from 2000 to 2008 and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2001 to 2006 and an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2001 to 2005.

Yueyang Teng received the bachelor’s degree from the Department of Applied Mathematics, Liaoning Normal University, Dalian, China, in 2002, the master’s degree from the Dalian University of Technology, Dalian, in 2005, and the Ph.D. degree in computer software and theory from Northeastern University, Shenyang, China, in 2013. From 2005 to 2013, he was a Software Engineer with Neusoft Positron Medical Systems Company Ltd., Shenyang. Since 2013, he has been a Lecturer with the Sino-Dutch Biomedical and Information Engineering School, Northeastern University. His research interests include image processing and machine learning.