Partial-Label and Structure-constrained Deep Coupled Factorization Network

Yan Zhang,[1] **Zhao Zhang,** [1,2] **Yang Wang,** [2] **Zheng Zhang,**[3] **Li Zhang,** [1] **Shuicheng Yan,** [4] **Meng Wang** [2]

[1] School of Computer Science and Technology, Soochow University, Suzhou 215006, China
[2] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
[3] Harbin Institute of Technology & Peng Cheng Laboratory, Shenzhen, China
[4] YITU Technology
{zhangyan0712suda, cszzhang, yeungwangresearch, darrenzz219, eric.mengwang}@gmail.com,
zhangliml@suda.edu.cn,shuicheng.yan@yitu-inc.com

Abstract

In this paper, we technically propose an enriched prior guided framework, called Dual-constrained Deep Semi-Supervised Coupled Factorization Network (DS^{2}CF-Net), for discovering hierarchical coupled data representation. To extract hidden deep features, DS^{2}CF-Net is formulated as a partial-label and geometrical structure-constrained framework. Specifically, DS^{2}CF-Net designs a deep factorization architecture using multilayers of linear transformations, which can coupled update both the basis vectors and new representations in each layer. To enable learned deep representations and coefficients to be discriminative, we also consider enriching the supervised prior by joint deep coefficients-based label prediction and then incorporate the enriched prior information as additional label and structure constraints. The label constraint can enable the intra-class samples to have same coordinate in feature space, and the structure constraint forces the coefficients in each layer to be block-diagonal so that the enriched prior using the self-expressive label propagation are more accurate. Our network also integrates the adaptive dual-graph learning to retain the local structures of both data and feature manifolds in each layer. Extensive experiments on image datasets demonstrate the effectiveness of DS^{2}CF-Net for representation learning and clustering.

Introduction

Learning compact representation of high-dimensional data is one of core topics in artificial intelligence research. To learn effective representations, Matrix Factorization (MF) is one of widely-used methods (Zhang *et al*. 2019b; Zhang *et al*. 2019c; Ma *et al*. 2019; Lin *et al*. 2020). Classical MF methods include Singular Value Decomposition (SVD) (Golub *et al*. 1970), Vector Quantization (VQ) (Gray 1984), Nonnegative Matrix Factorization (NMF) (Lee 1999) and Concept Factorization (CF) (Wei *et al*. 2004), etc. NMF and CF use the nonnegative constraints on factorization matrices to learn parts-based representations that are distinguishing for subsequent high-level tasks. Specifically, they decompose the data matrix into two/three factors, where one factor is basis vectors capturing high-level features so each sample can be reconstructed by a linear combination over them, and the other one corresponds to the new representation.

Note that CF offers an obvious advantage over NMF, i.e., it can be kernelized easily, but they both cannot encode the local geometry of features and also fail to use any label information even if available. To retain the local information, some graph regularized methods have been proposed, e.g., Graph-Regularized CF with Local Coordinate (LGCF) (Li *et al*. 2017a), Graph Regularized NMF (GNMF) (Cai *et al*. 2011a), Graph-Regularized LCF (GRLCF) (Ye *et al*. 2017), Locally Consistent CF (LCCF) (Cai *et al*. 2011b), Dual Regularization NMF (DNMF) (Shang *et al*. 2012) and Dual-graph regularized CF (GCF) (Ye *et al*. 2014). Note that these methods usually use the graph Laplacian to smooth the representation and encode the geometrical information of data space. Different from GNMF and LCCF, both DNMF and GCF can not only preserve the geometrical structures of data manifold but also the feature manifold using the dual-graph regularization (Shang *et al*. 2012; Ye *et al*. 2014). Although these algorithms have obtained en-couraging clustering abilities, they still suffer from certain shortcomings: 1) High sensitivity and tricky optimal determination of the number k of nearest neighbors (Roweis *et al*. 2000); 2) Separating the graph construction from the factorization process by two independent steps cannot ensure the pre-encoded weights to be optimal for subsequent data representation; 3) They cannot use the label information to improve the representation and clustering tasks due to unsupervised nature, similarly as NMF and CF. For the discriminative MF to use label information, some semi-supervised algorithms were proposed, e.g., Constrained Nonnegative Matrix Factorization (CNMF) (Liu *et al*. 2012), Semi-supervised GNMF (SemiGNMF) (Cai *et al*. 2011a) and Constrained Concept Factorization (CCF) (Liu *et al*. 2014). Although CNMF, SemiGNMF and CCF can use label information clearly, they fail to fully utilize unlabeled data, as they do not consider learning an explicit label indicator matrix and predicting the labels of unlabeled data, and mapping them into respective subspaces in feature space as well. In addition, CNMF, SemiGNMF and CCF also cannot self-express data in a recovered clean space. Although preserving local information or incorporating supervised prior can enhance NMF and CF,

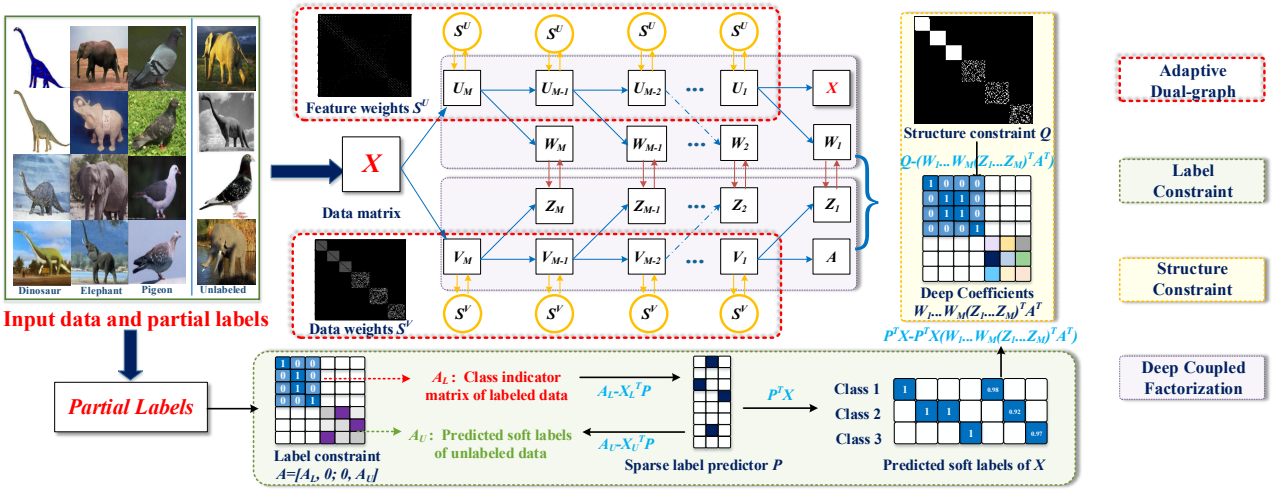

Figure 1: The flowchart and learning principle of our proposed DS^{2}CF-Net framework.

however, all above mentioned algorithms are single-layer models that can only discover shallow features from input data, but cannot obtain deep hidden features and hierarchical information.

In this paper, we propose a novel deep semi-supervised self-expressive coupled MF strategy that can represent data more appropriately by using partial labeled data and a deep structure. The main contributions are summarized as:

(1) Technically, a new supervised prior enrichment guided Dual-constrained Deep Semi-Supervised Coupled Factorization Network (DS^{2}CF-Net) is proposed. To discover and encode hidden deep features accurately, DS^{2}CF-Net designs a novel updating strategy for the deep concept factorization, i.e., it coupled optimizes the basis vectors and representation matrix in each layer, learning with partial labeled data. Fig.1 illustrates the flowchart of our DS^{2}CF-Net clearly.

(2) For discriminant representations, the innovations of our DS^{2}CF-Net are twofold: 1) enriching the supervised prior clearly by joint label prediction; 2) incorporating the enriched supervised prior as additional label and structure constraints. To enrich the prior, DS^{2}CF-Net fully utilizes unlabeled data by propagating and predicting their labels using a robust label predictor learned from labeled data. Dual-constraints are also included to improve and enhance the discriminating ability of the learned representation.

(3) To achieve locality-preserving higher-level representation, DS^{2}CF-Net uses a self-weighted dual-graph learning strategy in each layer, i.e., optimizing the weights jointly with MF. Specifically, in each layer, DS^{2}CF-Net performs the adaptive weighting based on both the deep basis vector graph and deep feature graph at the same time. Note that the self-weighted dual-graph learning can avoid the tricky issue of determining nearest neighbors, which is suffered in most existing locality preserving models. Such an operation can also obtain the adaptive neighborhood preserving deep basis vectors and deep features to enhance the performance.

Related Work

Concept Factorization

Given a data matrix $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{D\times N}$, where x_i is a sample vector, N is the number of samples and D is the original dimensionality. Let $U \in \mathbb{R}^{D\times r}$ and $V \in \mathbb{R}^{N\times r}$ be two nonnegative matrices whose product $UV^T \in \mathbb{R}^{D\times N}$ is the approximation to X, where r is rank. By representing each basis by a linear combination of x_i, i.e., $\sum_{i=1}^{N} w_{ij}x_i$, where $w_{ij} \geq 0$, CF solves:

$$O = \left\|X - XWV^T\right\|_F^2, \quad \text{s.t. } W, V \geq 0, \tag{1}$$

where $W = [w_{ij}] \in \mathbb{R}^{N\times r}$, XW denotes the bases, V^T is the learned representation of X, and ^{T} is matrix transpose.

Constrained Concept Factorization

CCF extends CF to semi-supervised scenario by using label information as an additional constraint. If X contains a labeled set $X_L \in \mathbb{R}^{D\times l}$ and an unlabeled set $X_U \in \mathbb{R}^{D\times u}$, i.e., $l + u = N$, where l and u are the numbers of labeled and unlabeled data respectively, then CCF defines a label constraint matrix A. Let $A_L \in \mathbb{R}^{l\times c}$ be the class indicator matrix over X_L, where c is class number. The element $(A_L)_{ij}$ is defined as 1 if x_i is labeled as the j-th class, and 0 otherwise. Note that CCF did not define a class indicator for X_U and simply used an identity matrix $I_{u\times u}$ for X_U. As such, the overall label constraint matrix A is defined as

$$A = \begin{bmatrix} (A_L)_{l\times c} & 0 \\ 0 & I_{u\times u} \end{bmatrix} \in \mathbb{R}^{(l+u)\times(c+u)}. \tag{2}$$

To ensure the samples of the same label to be mapped into the same v_i, CCF imposes a label constraint by an auxiliary matrix Z, i.e., $V = AZ$. Finally, CCF computes $W \in \mathbb{R}^{N\times r}$ and $Z \in \mathbb{R}^{(c+u)\times r}$ from the following objective function:

$$O = \left\|X - XWZ^TA^T\right\|_F^2, \text{ s.t. } W, Z \geq 0. \tag{3}$$

Next, we briefly review several related deep MF models.

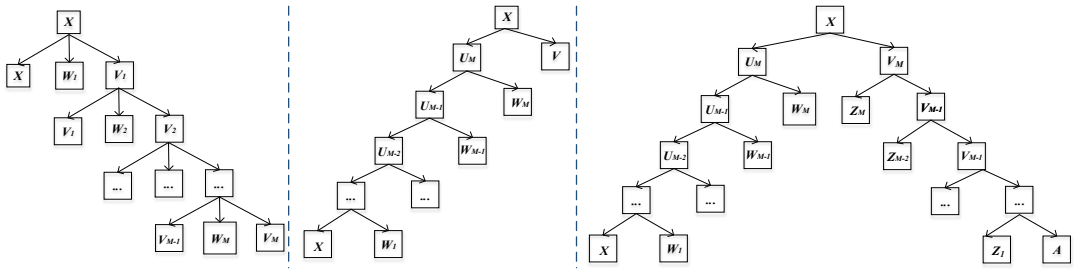

Figure 2: Architecture comparison of existing multilayer MF frameworks, including traditional multilayer CF model (e.g., MNMF, MCF and GMCF) *(left)*, optimized multilayer CF model (e.g., DSCF-Net) *(middle)*, and our DS^{2}CF-Net *(right)*.

Traditional Multilayer MF

The methods of this category usually directly use the output of previous layer (i.e., intermediate representation V) as the input of subsequent layer, without considering optimizing the representation or basis vectors in each layer. As such, as they cannot ensure the intermediate representation to be good for subsequent layers, the performance may be degraded. Examples of traditional multilayer methods include MNMF (Cichocki *et al*. 2006), MCF (Li *et al*. 2015) and GMCF (Li *et al*. 2017b), etc. We show the deep structure of this category in Fig.2 (*left*).

Optimized Deep MF Models

Optimized models aim to learn deep features by multilayer of linear transformations and updating the basis vectors or representation in each layer, e.g., Weakly-supervised Deep MF (WDMF) (Li *et al*. 2017c), Deep Semi-NMF (DSNMF) (Trigeorgis *et al*. 2015) and Deep Self-representative CF Network (DSCF-Net) (Zhang *et al*. 2019a). We show the structure of DSCF-Net in Fig.2(*middle*) and ours in Fig.2(*right*). We see that ours coupled optimizes the basis vectors and representation in each layer.

Proposed Formulation

Given $X = [X_L, X_U]$, to enhance the representation ability, we design a hierarchical and coupled factorization network of M layers. DS^{2}CF-Net is modeled as the one of learning updated pairs of representation matrices and basis vectors $XW_1 \dots W_M$, and M updated label constraint matrices A. That is, A is optimized and moreover enriched in our model.

Factorization Model

We firstly describe the initial problem of DS^{2}CF-Net as

$$\begin{aligned} O = &\left\| X - XW_0 \dots W_M \left(Z_0 \dots Z_M\right)^T A^T \right\|_F^2 \\ &+ \alpha J_2 + \beta J_3 + \gamma J_1 \\ \text{s.t.} &\forall_{i \in \{1,2\dots,M\}} W_i \geq 0, Z_i \geq 0 \end{aligned}, \quad (4)$$

where $XW_0 \dots W_M$ is deep basis vector, $\left(Z_0 \dots Z_M\right)^T A^T$ denotes the deep representation, the first term is the deep reconstruction error, J_1, J_2 and J_3 will be described shortly. W_0 and Z_0 are added to facilitate the descriptions, and both are fixed to be an identity matrix. Different from CCF, we define the label constraint matrix A as follows:

$$\begin{aligned} &A = \begin{bmatrix} A_L & 0 \\ 0 & A_U \end{bmatrix} \in \mathbb{R}^{(l \times u) \times (c+c)}, \\ &A_L \in \mathbb{R}^{l \times c}, A_U \in \mathbb{R}^{u \times c} \end{aligned} \quad (5)$$

where A_L is the class indicator for X_L. Note that we also learn an explicit class indicator A_U for X_U to enrich the supervised prior rather than fixing it to be an identity matrix as CCF, which can better group the representation of both labeled and unlabeled data using dual constraints. According to the self-expressive property on coefficients (Ma *et al*. 2018), the reconstruction error can be rewritten as

$$\begin{aligned} &\left\| X - XR_M \right\|_F^2, \\ &where R_M = W_0 \dots W_M \left(Z_0 \dots Z_M\right)^T A^T, \end{aligned} \quad (6)$$

where R_M is a meaningful coefficient matrix self-expressing X. Then, the factorization model can be presented as

$$\begin{aligned} X &\leftarrow U_M V_M^T \\ U_M = U_{M-1} W_M &\quad V_M = V_{M-1} Z_M \\ \vdots &\quad \vdots \\ U_2 = U_1 W_2 &\quad V_2 = V_1 Z_2 \\ U_1 = X W_1 &\quad V_1 = A Z_1 \end{aligned} \quad (7)$$

where U_m is the set of basis vectors of the m-th layer, V_m^T is the new representation, W_m is the intermediate matrix for updating the basis vectors and Z_m is the intermediate auxiliary matrix for updating the representations.

Enriched Prior Based Dual-constraints

DS^{2}CF-Net learns a robust label predictor $P \in \mathbb{R}^{D \times c}$ over the labeled data by minimizing a label fitness error $\left\| A_L - X_L^T P \right\|_F^2$, which can map each x_i into a label space in terms of $P^T x_i$. In addition, DS^{2}CF-Net also considers preserving the neighborhood information of the predicted soft labels $P^T X$ by self-expressing it using R_M. To be specific, the problem for learning the label predictor P is defined as follows:

$$\begin{aligned} J_1 &= \left\| A_L - X_L^T P \right\|_F^2 + \left\| P^T X - P^T X R_M \right\|_F^2 + \|P\|_{2,1} \\ &= \left\| A_L - X_L^T P \right\|_F^2 + \|P\|_{2,1} \\ &+ \left\| P^T X - P^T X \left(W_0 \dots W_M \left(Z_0 \dots Z_M\right)^T A^T \right) \right\|_F^2 \end{aligned}, \quad (8)$$

where the $L_{2,1}$-norm can further enable the learned label predictor to be robust against noise.

Enriched prior based label constraint A. After P is obtained, we can predict the soft label of each unlabeled sample $x_i \in X_U$ as $x_i^T P$. Then, we can obtain A_U for unlabeled data by using the normalized soft labels as follows:

$$(A_U)_{ij} = (X_U^T P)_{ij} / \sum_{j=1}^{c} (X_U^T P)_{ij}. \tag{9}$$

Clearly, the normalized soft labels meet the definition of probability, i.e., column-sum-to-one.

Enriched prior based structure constraint Q. We add Q to constrain the coefficients by minimizing the approximation error between Q and $W_0 \dots W_M (Z_0 \dots Z_M)^T A^T$:

$$\begin{aligned} J_2 = &\left\| Q - W_0 \dots W_M (Z_0 \dots Z_M)^T A^T \right\|_F^2 \\ &+ \left\| W_0 \dots W_M (Z_0 \dots Z_M)^T A^T \right\|_F^2, \end{aligned} \tag{10}$$

where the structure constraint matrix Q is defined As

$$Q = \begin{bmatrix} Q_L & 0 \\ 0 & Q_U \end{bmatrix}, Q_L = \begin{bmatrix} Q_1 & 0 & 0 & 0 \\ 0 & Q_2 & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & Q_c \end{bmatrix} \tag{11}$$

where Q_L and Q_U are the sub-matrices over X_L and X_U. As X_L is labeled, Q_L is strict block-diagonal, where each block Q_i, $i = 1, 2, \dots c$ is an $l_i \times l_i$ matrix of all ones, defined according to the labels, and l_i is the number of samples in class i in X_L. We initiate Q_U by the cosine similarities over X_U and update Q_U in m-th ($m > 1$) layer using the cosine similarities defined on the new representation of X_U.

Self-weighted Dual-graph Learning

We also incorporate the self-weighted dual-graph learning to retain the neighborhood information of both deep basis vectors $XW_0 \dots W_M$ and deep representation $(Z_0 \dots Z_M)^T A^T$ in an adaptive manner. Specifically, we obtain the data weight matrix $S^V \in \mathbb{R}^{N \times N}$ and the feature weight matrix $S^U \in \mathbb{R}^{D \times D}$ by minimizing:

$$\begin{aligned} J_3 = &\left\| (XW_0 \dots W_M)^T - (XW_0 \dots W_M)^T S^U \right\|_F^2 \\ &+ \left\| \left((Z_0 \dots Z_M)^T A^T \right) - \left((Z_0 \dots Z_M)^T A^T \right) S^V \right\|_F^2 \\ &\text{s.t. } S^U \geq 0, S^V \geq 0 \end{aligned} \tag{12}$$

By substituting J_1, J_2 and J_3 back into Eq.(4), the final objective function of DS^{2}CF-Net can be defined as

$$\begin{aligned} O = &\min_{\substack{W_1 \dots W_M, S^V, \\ Z_1 \dots Z_M, S^U, P}} \left\| X - XW_0 \dots W_M (Z_0 \dots Z_M)^T A^T \right\|_F^2 \\ &+ \alpha \left[\|Q - R_M\|_F^2 + \|R_M\|_F^2 \right] \\ &+ \beta \left[\left\| U_M^T - U_M^T S^U \right\|_F^2 + \left\| V_M^T - V_M^T S^V \right\|_F^2 \right] \\ &+ \gamma \left[\left\| A_L - X_L^T P \right\|_F^2 + \left\| P^T X - P^T X R_M \right\|_F^2 + \|P\|_{2,1} \right] \\ &\text{s.t. } \forall_{m \in \{1,2,\dots,M\}} W_m \geq 0, Z_m \geq 0, S^U \geq 0, S^V \geq 0 \end{aligned} \tag{13}$$

where $U_M = XW_0 \dots X_M$, $V_M = A(Z_0 \dots Z_M)$ and $R_M = W_0 \dots W_M V_M^T$.

Optimization

(1) Fix others, update the factors W_m and Z_m: By defining $\Pi_{m-1} = W_0 \dots W_{m-1}$ and $\Lambda_{m-1} = Z_0 \dots Z_{m-1}$, W_m and Z_m can be obtained from the reduced problem. After simple computations, the updating rules of W_m and Z_m are obtained as follows

$$\begin{aligned} (W_m)_{ik} &\leftarrow (W_m)_{ik} \cdot \\ &\frac{\left(2\Pi_{m-1}^T K_X V_M + 2\alpha \Pi_{m-1}^T Q V_M + \Omega_W\right)_{ik}}{\left(2\Pi_{m-1}^T K_X \Pi_{m-1} W_m V_M^T V_M + \Phi_W\right)_{ik}}, \end{aligned} \tag{14}$$

$$\begin{aligned} (Z_m)_{ik} &\leftarrow (Z_m)_{ik} \cdot \\ &\frac{\left(2\Lambda_{m-1}^T A^T K_X \Pi_m + 2\alpha \Lambda_{m-1}^T A^T Q^T \Pi_m + \Omega_Z\right)_{ik}}{\left(2\Lambda_{m-1}^T K_A \Lambda_{m-1} Z_m U_M^T U_M + \Phi_Z\right)_{ik}} \end{aligned} \tag{15}$$

where
$H_u = \left(I - S^U\right)\left(I - S^U\right)^T$, $H_v = \left(I - S^V\right)\left(I - S^V\right)^T$, $\Pi_{m-1} = W_0 \dots W_{m-1}$, and $\Lambda_{m-1} = Z_0 \dots Z_{m-1}$, I is an identity matrix, $K_X = X^T X$, $K_A = A^T A$, $K_P = X^T P P^T X$. $\Pi_m = \Pi_{m-1} W_m$ and Π_m is known when updating Z_m. $\Phi_W = 4\alpha Q^T \Pi_{m-1} W_m V_M^T V_M + \beta \Pi_{m-1}^T X^T (H_u + H_u^T) X \Pi_{m-1} W + 2\gamma \Pi_{m-1}^T K_P \Pi_{m-1} W_m V_M^T V_M$, $\Phi_Z = 4\alpha \Lambda_{m-1}^T K_A \Lambda_{m-1} Z_m W_m^T Q^T \Pi_m + \beta \Lambda_{m-1}^T (H_v + H_v^T) \Lambda_{m-1} Z_m K_A^T + 2\gamma \Lambda_{m-1}^T K_A \Lambda_{m-1} Z_m U_M^T P P^T U_M$, $\Omega_W = 2\gamma \Pi_{m-1}^T K_P V_M$ and $\Omega_Z = 2\gamma \Lambda_{m-1}^T A^T K_P \Pi_m$ are auxiliary matrices to simplify descriptions.

(2) Fix others, update S^U and S^V: Let $U_M = X\Pi_{m-1} W_m$ and $V_M = A\Lambda_{m-1} Z_m$, we can obtain the updating rules for S^U and S^V as follows:

$$\begin{aligned} \left(S^U\right)_{ik} &\leftarrow \left(S^U\right)_{ik} \cdot \\ &\frac{\left(\left(X\Pi_{m-1} W_m\right) \left(X\Pi_{m-1} W_m\right)^T \right)_{ik}}{\left(\left(X\Pi_{m-1} W_m\right) \left(X\Pi_{m-1} W_m\right)^T S^U \right)_{ik}}, \end{aligned} \tag{16}$$

$$\begin{aligned} \left(S^V\right)_{ik} &\leftarrow \left(S^V\right)_{ik} \cdot \\ &\frac{\left(\left(A\Lambda_{m-1} Z_m\right) \left(A\Lambda_{m-1} Z_m\right)^T \right)_{ik}}{\left(\left(A\Lambda_{m-1} Z_m\right) \left(A\Lambda_{m-1} Z_m\right)^T S^V \right)_{ik}}. \end{aligned} \tag{17}$$

(3) Fix others, update P: By the properties of $L_{2,1}$-norm (Yang *et al*. 2011), we have $\|P\|_{2,1} = 2\,\mathrm{tr}\left(P^T B P\right)$, where B is a diagonal matrix with entries $b_{ii} = 1/\left(2\left\|p^i\right\|_2\right)$, where p_i is the i-th row of P. Finally, we can infer P in each layer as follows

$$P = \left(X_L X_L^T + X_L H_M X_L^T + B\right)^{-1} X_L A_L, \tag{18}$$

where $H_M = (I - R_M)(I - R_M)^T$. After P is obtained, we can use it to update B and predict the labels of unlabeled data. After that, we can use the normalized soft labels to optimize the label constraint matrix A for representation. For complete presentation, we summarize the procedures in

Algorithm 1. Optimization procedures of DS^{2}CF-Net

Inputs: Partially labeled data matrix $X = [X_L, X_U]$, the constant r and tunable parameters α, β and γ.

Initialization:

$t = 0$;

Initialize W and Z to be random matrices;

Initialize P and A by labeled data;

Initialize Q_U by the cosine similarities over X_U;

Initialize S^U by the cosine similarities over X;

Initialize S^V using semi-supervised weights, that is, supervised ones for X_L and cosine similarities for X_U.

For each fixed number *m* of layers:

While not converged do

1. Update W_m^{t+1} and Z_m^{t+1} by Eqs.(14-15), and then we can obtain $V_m^{t+1} = AZ_0 \dots Z_m^{t+1}$;
2. Update $(S^U)^{t+1}$ and $(S^V)^{t+1}$ by Eqs.(16-17);
3. Update P^{t+1} by Eq.(18), update the estimated soft labels of X_U as $X_U^T P^{t+1}$, and then update A_U by Eq.(9);
4. Update the label constraint matrix A by Eq.(5);
5. Update Q_U using cosine similarities based on $(V_m^{t+1})_i$ $i \in \{l+1, \dots, N\}$, and update matrix Q;
6. Check the convergence conditions:

if $\left\|W_m^{t+1} - W_m^t\right\|_F^2 \leq \mathcal{E}$ and $\left\|V_m^{t+1} - V_m^t\right\|_F^2 \leq \mathcal{E}$, stop; else $t = t + 1$.

End While

End for

Outputs: Deep low-dimensional representation V_m^*.

Algorithm 1, where the diagonal matrix B is initialized as an identity matrix. We initialize the linear label predictor as $P = \left(X_L X_L^T + I\right)^{-1} X_L A_L$ (Zhang *et al*. 2020) and predict the soft labels of X_U as $X_U^T P$, and normalize the soft labels by Eq.(9). Based on the normalized soft labels of unlabeled data, we can initialize the label constraint matrix A.

Simulation Results and Analysis

The experimental results of DS^{2}CF-Net are compared with 5 deep MF models (i.e., MNMF, MCF, GMCF, DSNMF and DSCF-Net), 3 single-layer MF models (i.e., DNMF, GCF and SRMCF (Ma *et al*. 2018)), and 4 semi-supervised MF models (i.e., SemiGNMF, CNMF, CCF and RS^{2}ACF). In this study, 4 public databases are involved, i.e., AR (Bergstra *et al*. 2013), ETH80 (Leibe *et al*. 2003), USPS (Hull 1994) and Fashion MNIST (Xiao *et al*. 2017). Detailed information of the used databases is described in Table 1. We normalize each column of input data matrix to have unit norm.

Visual Image Analysis by Visualization

Since the representation $V_M = A(Z_0 \dots Z_M)$ is the final output of model, we evaluate its representation ability by visualizing the adaptive weights S^V on V_M. AR database is used, and for clear observation we only choose two categories to construct , with 10 labeled images per class. The matrix S^V is visualized in Fig.3, where we show the adaptive weights obtained in the first 4 layers. We see that the weights have approximate block-diagonal structures in each layer. Specifically, the structures of weights get clearer with less noise and inter-class connections as the number of layers increases, which means the new representation V_M has a strong representation ability and moreover our deep model can potentially improve the similarity measure.

#Name	#sample	#class	#dim
AR face database	2600	100	1024
USPS digits database	9298	10	256
ETH80 object database	3280	80	1024
Fashion MNIST database	70000	10	784

Table 1: List of evaluated databases.

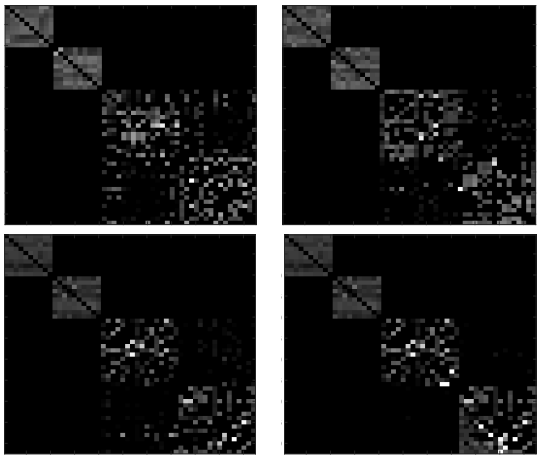

Figure 3: Visualization of the data weight matrix S^V obtained by DS^{2}CF-Net. (Top-left) 1-st layer, (Top-right) 2-nd layer, (Bottom-left) 3-rd layer, (Bottom-right) 4-th layer.

Convergence Analysis

We show the convergence results of our DS^{2}CF-Net in the third layer on AR database, with 40% labeled per class, in Fig.4. We can see that our DS^{2}CF-Net converges rapidly and usually converges within 5 times iterations in the 3rd layer.

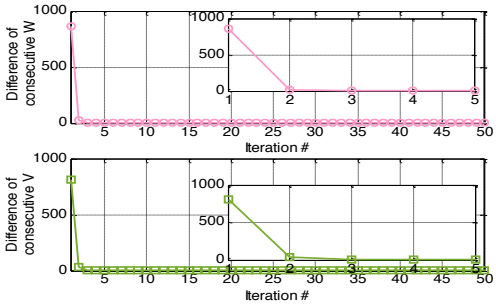

Figure 4: Convergence analysis of DS^{2}CF-Net on AR.

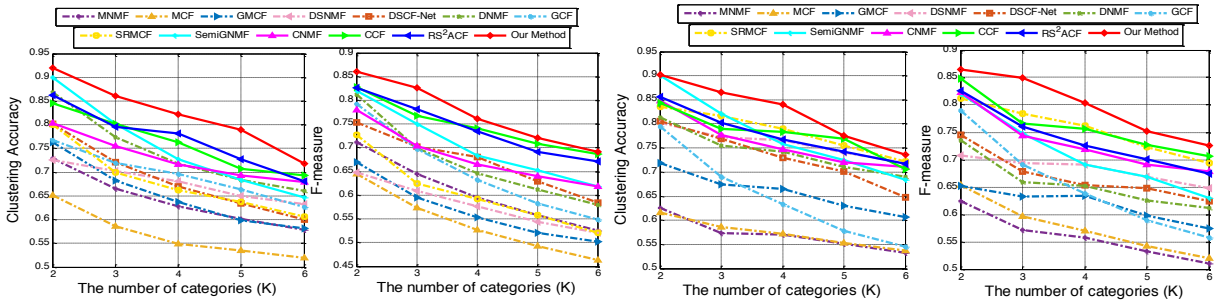

Figure 5: Clustering performance over varied K values. (Left) USPS, (Right) Fashion MNIST.

Methods	AC		F-measure	
	USPS digits	Fashion MNIST	USPS digits	Fashion MNIST
MNMF	0.6406±0.0592	0.5701±0.0348	0.6070±0.0735	0.5592±0.0433
MCF	0.5682±0.0525	0.5727±0.0302	0.5400±0.0712	0.5765±0.0517
GMCF	0.6524±0.0729	0.6590±0.0428	0.5683±0.0665	0.6180±0.0311
DSNMF	0.6786±0.0379	0.7601±0.0314	0.5799±0.0506	0.6817±0.0236
DSCF-Net	0.6853±0.0792	0.7307±0.0609	0.6700±0.0654	0.6696±0.0463
DNMF	0.7410±0.0830	0.7426±0.0472	0.6695±0.0908	0.6569±0.0476
GCF	0.6949±0.0540	0.6484±0.0986	0.6503±0.0972	0.6531±0.0914
SRMCF	0.6811±0.0746	0.7841±0.0460	0.6040±0.0785	0.7551±0.0470
SemiGNMF	0.7520±0.1010	0.7779±0.0847	0.7050±0.0805	0.7117±0.0741
CNMF	0.7293±0.0503	0.7605±0.0551	0.6814±0.0636	0.7308±0.0575
CCF	0.7621±0.0642	0.7782±0.0492	0.7461±0.0553	0.7607±0.0539
RS^{2}ACF	0.7697±0.0690	0.7775±0.0545	0.7412±0.0637	0.7373±0.0590
Our method	**0.8219±0.0757**	**0.8236±0.0676**	**0.7722±0.0708**	**0.7991±0.0600**

Table 2: Averaged clustering accuracies (AC) and F-scores (Mean±std) based on the evaluated real image databases.

Quantitative Clustering Evaluations

(1) Clustering evaluation process. For quantitative evaluations, we perform the K-means algorithm with cosine distance on the learned representation by each model. Following the procedures in (Liu *et al*. 2014; Sugiyama 2007), for each number K of clusters, we choose K categories from each database randomly to form the data matrix X. The value of K is tuned from 2 to 6. The rank of the representation is set to K+1 for clustering as (Liu *et al*. 2014; Zhang *et al*.2019a). The final results are averaged over 10 random selections of K categories. For fair comparison, we randomly choose 40% labeled samples per class for semi-supervised models and set the number of layers to 3 for deep models.

(2) Evaluation metric. We use the Accuracy (AC) and F-measure (Cai *et al*. 2005) as evaluation metrics in this work.

(3) Evaluation results. The clustering curves on USPS and Fashion MNIST databases are shown in Fig.5, and the according averaged AC and F-scores are described in Table 2. We see that: (1) the AC and F-measure of each method go down as the number of categories is increased, which is easy to understand, since clustering data of less categories is easier than clustering more; (2) DS^{2}CF-Net delivers better results than other related methods in most cases.

Ablation Study

(1) Clustering with different labeled proportions. First, we evaluate each semi-supervised MF models by using different numbers of labeled data in each class. For each database, the labeled proportion varies from 10% to 90% and we randomly choose 3 categories. The averaged clustering results are reported in Fig.6. We can see that: 1) the increasing number of labeled samples can greatly improve the clustering performance of each method; It can also be found that the improvement by our DS^{2}CF-Net over other compared methods is more obvious, especially when the proportion of labeled data is relatively small; 2) our DS^{2}CF-Net delivers better results across different labeled proportions.

(2) Clustering with different numbers of layers. In this study, we vary the number of layers from 1 to 10 with step 1. For each database, we choose three categories for evaluation. The averaged ACs are shown in Fig.7. We see that: 1) DS^{2}CF-Net delivers higher accuracies than other methods in most cases; 2) the increase of the number of layers can generally improve the performance, implying that discovering hidden deep features can improve the representations.

(3) Parameter sensitivity analysis. Finally, we explore the effects of parameters in the objective function on the representation ability. Following common procedures, we use the grid search and linear strategy (Zhang *et al*. 2016; Ren *et al*. 2020; Zhang *et al*. 2020) in our experiments. Specifically, we first fix $\gamma = 1$ to tune α and β from $\{10^{-5}, 10^{-4},..., 10^{5}\}$. Then, we use selected α and β to tune γ. We choose three categories to train our model and the number of layers is set to 3. The selection results on ETH80 are shown in Fig.8 as an example, where the results are averaged based on 5 random initializations of the cluster centers of K-means.

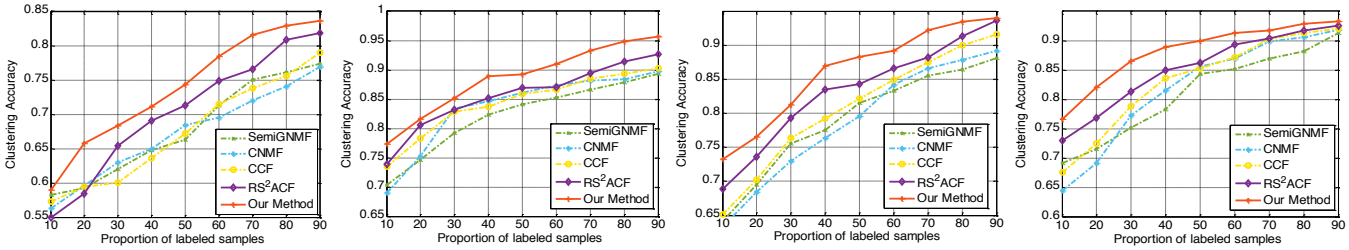

Figure 6: ACs vs. varied proportions of labeled samples over (L1) AR, (L2) ETH80, (R2) USPS, (R1) Fashion MNIST.

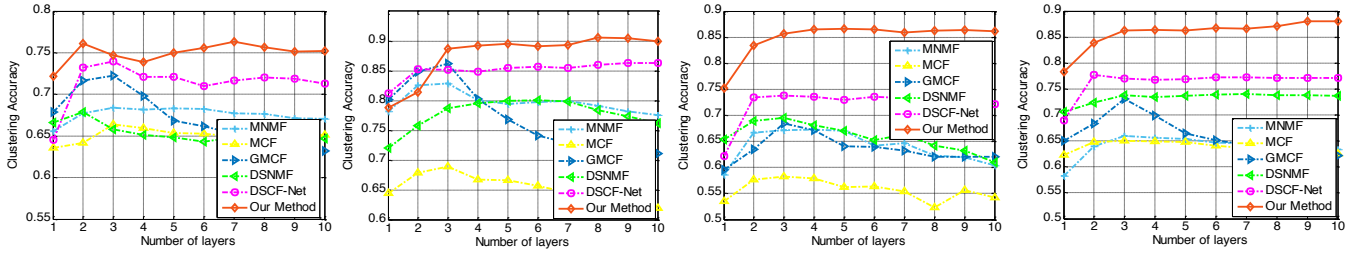

Figure 7: ACs vs. varied number of layers over (L1) AR, (L2) ETH80, (R2) USPS, (R1) Fashion MNIST.

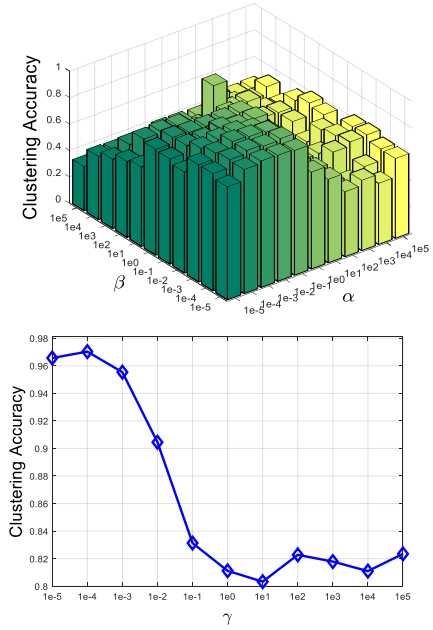

Figure 8: Clustering accuracies vs. varied model parameters of our DS^{2}CF-Net on the ETH80 database.

Conclusion

We proposed an enriched prior based dual-constrained deep semi-supervised coupled factorization network to discover deep hierarchical features. DS^{2}CF-Net designs a coupled hierarchical deep and geometry structures-constrained factorization model using multiple layers of linear transformations of basis vectors and representation. To improve the discriminating deep representations, DS^{2}CF-Net clearly considers enriching the supervised prior by the joint deep coefficients-regularized label prediction, and incorporates the enriched prior information as additional dual constraints. Extensive visual and quantitative clustering evaluations demonstrated the effectiveness of DS^{2}CF-Net. In future, more effective coupled factorization strategy will be investigated.

Acknowledgments

This work is partially supported by National Key R&D Program of China (2018YFB0804202), National Natural Science Foundation of China (61672365, 62072151, 61806035, 62002085 and U1936217), Anhui Provincial Natural Science Fund for Distinguished Young Scholars (2008085J30), and the Fundamental Research Funds for the Central Universities of China (JZ2019HGPA0102). Zhao Zhang, Yang Wang and Zheng Zhang are the co-corresponding authors.

References

Bergstra, J.; Yamins, D.; and Cox, D. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of International Conference on Machine Learning*, 28(1): 115-123. Atlanta, USA.

Cai, D.; He, X. F.; Han, J.; and Huang, T. 2011a. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(8): 1548-1560.

Cai, D.; He, X.; and Han, J. 2005. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering* 17(12): 1624-1637.

Cai, D.; He, X.; and Han, J. 2011b. Locally consistent concept factorization for document clustering. *IEEE Transactions on Knowledge and Data Engineering* 23(6): 902-913.

Cichocki, A., and Zdunek, R. 2006. Multilayer nonnegative matrix factorization. *Electronics Letters* 42(16): 947-948.

Golub, G. H., and Reinsch, C. 1970. Singular value decomposition and least squares solutions. *Numerische mathematik* 14(5): 403-420.

Gray, R. 1984.Vector quantization. *IEEE Assp Magazine* 1 (2): 4-29.

Hull, J. 1994. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16(5): 550-554.

Lee, D., and Seung, H. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401: 788-791.

Leibe, B., and Schiele, B. 2003. Analyzing appearance and contour based methods for object categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 409-415.

Li, H.; Zhang, J.; and Liu, J. 2017a. Graph-regularized CF with local coordinate for image representation. *Journal of Visual Comm. and Image Representation* 49: 392-400.

Li, X.; Shen, X.; Shu, Z.; Ye, Q.; and Zhao, C. 2017b. Graph regularized multilayer concept factorization for data representation. *Neurocomputing* 238: 139-151.

Li, X.; Zhao, C. X.; Shu, Z.; and Wang, Q. 2015. Multilayer Concept Factorization for Data Representation. In *Proceedings of International Conference on Computer Science & Education*, 486-491. Cambridge, UK.

Li, Z., and Tang, J. 2017c. Weakly-supervised Deep Matrix Factorization for Social Image Understanding. *IEEE Transactions on Image Processing* 26(1): 276-288.

Lin, B. H.; Tao, X. M.; and Lu, J. H. 2020. Hyperspectral Image Denoising via Matrix Factorization and Deep Prior Regularization. *IEEE Trans. on Image Proc.* 29: 565-578.

Liu, H.; Wu, Z.; and Li, X. 2012. Constrained nonnegative matrix factorization for image representation. *IEEE Trans. on Pattern Analysis and Machine Intell.* 34(7): 1299-1311.

Liu, H.; Yang, G.; and Wu, Z. 2014. Constrained concept factorization for image representation. *IEEE Transactions on Cybernetics* 44(7): 1214-1224.

Ma, S.; Zhang, L.; Hu, W.; Zhang, Y.; Wu, J.; and Li, X. 2018. Self-Representative Manifold Concept Factorization with Adaptive Neighbors for Clustering. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*, 2539-2545. Stockholm, Sweden.

Ma, X. K.; Dong, D.; and Wang, Q. 2019. Community Detection in Multi-Layer Networks Using Joint Nonnegative Matrix Factorization. *IEEE Transactions on Knowledge and Data Engineering* 31(2): 273-286.

Ren, J.; Zhang, Z.; Li, S.; Wang, Y.; Liu, G.; Yan, S; and Wang, M. 2020. Learning Hybrid Representation by Robust Dictionary Learning in Factorized Compressed Space. *IEEE Transactions on Image Processing* 29: 3941-3956.

Roweis, S., and Saul, L. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290: 2323-2326.

Shang, F.; Jiao, L.; and Wang, F. 2012. Graph dual regularization non-negative matrix factorization for co-clustering. *Pattern Recognition* 45(6): 2237-2250.

Sugiyama, M. 2007. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research* 8: 1027-1061.

Trigeorgis, G.; Bousmalis, K.; Zafeiriou, S.; and Schuller, B. W. 2015. A deep matrix factorization method for learning attribute representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(3): 417-429.

Wei, X., and Gong, Y. 2004. Document clustering by concept factorization. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Sheffield, UK.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv: 178.07747v2.

Yang, Y.; Shen, H.; Ma, Z.; Huang, Z.; and Zhou, X. 2011. L2, 1-Norm Regularized Discriminative Feature Selection for Unsupervised Learning. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*, 1589-1594. Barcelona, Spain.

Ye, J., and Jin, Z. 2014. Dual-graph regularized concept factorization for clustering. *Neurocomputing* 138(11): 120-130.

Ye, J., and Jin, Z. 2017. Graph-Regularized Local Coordinate Concept Factorization for Image Representation. *Neural Processing Letters* 46(2): 427-449.

Zhang, Y.; Zhang, Z.; Zhang, Z.; Zhao, M.; Zhang, L.; Zha, Z.; and Wang, M. 2019a. Deep Self-representative Concept Factorization Network for Representation Learning. In *Proceedings of SIAM International Conference on Data Mining*. Cincinnati, USA.

Zhang, Z.; Li, F.; Zhao, M.; Zhang, L.; and Yan, S. 2016. Joint Low-Rank and Sparse Principal Feature Coding for Enhanced Robust Representation and Visual Classification. *IEEE Transactions on Image Processing* 25(6): 2429-2443.

Zhang, Z.; Zhang, Y.; Li, S.; Liu, G.; Wang, M.; and Yan, S. 2019b. Robust Unsupervised Flexible Auto-weighted Local-Coordinate Concept Factorization for Image Clustering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2092-2096. Brighton, UK.

Zhang, Z.; Zhang, Y.; Li, S.; Liu, G.; Zeng, D.; Yan, S.; and Wang, M. 2019c. Flexible Auto-weighted Local-coordinate Concept Factorization: A Robust Framework for Unsupervised Clustering. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, Z.; Zhang, Y.; Liu, G.; Tang, J.; Yan, S.; and Wang, M. 2020. Joint Label Prediction based Semi-Supervised Adaptive Concept Factorization for Robust Data Representation. *IEEE Transactions on Knowledge and Data Engineering* 32: 952-970.