Pseudo Supervised Matrix Factorization in Discriminative Subspace

Jiaqi Ma[1] , **Yipeng Zhang**[1] , **Lefei Zhang**[1*] , **Bo Du**[1] and **Dapeng Tao**[2]
[1]School of Computer Science, Wuhan University
[2]School of Information Science and Engineering, Yunnan University
{jiaqima, zyp91, zhanglefei, remoteking}@whu.edu.cn, dapeng.tao@gmail.com

Abstract

Non-negative Matrix Factorization (NMF) and spectral clustering have been proved to be efficient and effective for data clustering tasks and have been applied to various real-world scenes. However, there are still some drawbacks in traditional methods: (1) most existing algorithms only consider high-dimensional data directly while neglect the intrinsic data structure in the low-dimensional subspace; (2) the pseudo-information got in the optimization process is not relevant to most spectral clustering and manifold regularization methods. In this paper, a novel unsupervised matrix factorization method, Pseudo Supervised Matrix Factorization (PSMF), is proposed for data clustering. The main contributions are threefold: (1) to cluster in the discriminant subspace, Linear Discriminant Analysis (LDA) combines with NMF to become a unified framework; (2) we propose a pseudo supervised manifold regularization term which utilizes the pseudo-information to instruct the regularization term in order to find subspace that discriminates different classes; (3) an efficient optimization algorithm is designed to solve the proposed problem with proved convergence. Extensive experiments on multiple benchmark datasets illustrate that the proposed model outperforms other state-of-the-art clustering algorithms.

1 Introduction

Data clustering is always a hot research topic that has been widely studied in various areas, such as document clustering [Cai *et al.*, 2011a], gene selection [Jiang *et al.*, 2004] and image segmentation [Shi and Malik, 2000]. Among them, Non-negative Matrix Factorization (NMF) and spectral clustering are two widely-used methods. NMF aims to factorize a matrix into two non-negative matrices whose product reconstructs the original data matrix. According to [Ding *et al.*, 2010], the two matrices correspond to the cluster centroid and indicator, respectively. Thus we obtain the clustering result directly by the cluster indicator matrix without doing extra post-processing. Spectral clustering can adapt to a wider range of geometries and detect non-convex patterns and linearly non-separable clusters [Ng *et al.*, 2002]. The general spectral clustering method needs to construct an adjacency matrix and calculate the Eigen decomposition value of the corresponding Laplacian matrix [Chung and Graham, 1997].

In the past few years, many algorithms and frameworks have been proposed in order to improve the data clustering performance. For K-means methods, Liu [Liu *et al.*, 2017] proposed a compressed K-means method for fast large-scale clustering. Besides, Shen [Shen *et al.*, 2017] tried a sparse embedded algorithm to accelerate K-means clustering. For NMF-based algorithms, Cai [Cai *et al.*, 2011b] kept the local geometry of the data in low dimensional space by proposing Graph Regularized NMF. Wang [Wang *et al.*, 2018] incorporated the ordinal relations and proposed a novel ranking preserving NMF approach. To enhance the robust of matrix factorization, Ke [Ke and Kanade, 2005] replaced l_2-norm with l_1-norm to improve the robustness, but the l_1-norm still cannot guarantee the feature rotation invariance. Jin [Huang *et al.*, 2013] utilized $l_{2,1}$-norm to factorize the data matrix. And then, Zhang [Zhang *et al.*, 2017a] adopted $l_{2,1}$-norm to learn the low-dimensional representation of the original data and showed its effectiveness. Lu [Lu *et al.*, 2017] attempted to establish a connection between Linear Discriminant Analysis (LDA) and NMF in a supervised or semi-supervised way, but it can not be applied to data clustering.

In particular, the manifold regularization term has been combined with clustering models to improve the performance. Cai [Cai *et al.*, 2011a] proposed a graph model capturing the local manifold geometry to address the underlying concepts which are consistent with the intrinsic manifold. To learn a better affinity matrix, Zhang [Zhang *et al.*, 2017b] adopted the adaptive manifold regularization to matrix factorization. Wang [Wang *et al.*, 2017] tried to learn a mapping in a relative low-dimensional with a more discriminative ability by a Grassmann manifold. What's more, Ma [Ma *et al.*, 2018] extended the standard concept factorization model with an adaptive manifold regularizer which can represent of the raw data itself and a graph manifold regularizer which can reveal the local structure information of original data. And Zhang [Zhang *et al.*, 2017a] incorporated the manifold regularization terms on both the low-dimensional feature representation and the cluster labels and get better local geometrical infor-

[*] Corresponding author

mation. All of those models show that the manifold regularization term has an extraordinarily good clustering ability and can be expanded to other frameworks.

However, there are still some drawbacks in those existing methods. First, the framework based on NMF only factorizes matrices on high dimensional data space while ignores the intrinsic data information in the low-dimensional subspace. Thus, ordinary NMF needs more constraints to capture complicated structures. Second, local relationships among all data points got in the optimization process, such as pseudo-information, cannot be used in most spectral clustering and manifold regularization methods. The geometry-metric structure of data distribution lacks an effective way to capture. Third, matrix factorization squares the residue error of each data point with a l_2-norm objective function, so the clustering results are easily affected by the outliers.

In this paper, a novel unsupervised matrix factorization method, Pseudo Supervised Matrix Factorization (PSMF), is proposed for data clustering. The major contributions of this paper are summarized as follows:

1. A unified framework to cluster in the discriminant subspace is proposed, which combines LDA with NMF. So the intrinsic data structure in the low-dimensional subspace can be found and utilized.
2. A pseudo supervised manifold regularization term is proposed in order to find the subspace that discriminates different classes. This regularization term utilizes the pseudo-information to instruct itself and to refine the clustering results.
3. A novel Augmented Lagrangian Method (ALM) based optimization algorithm is designed to effectively and can efficiently seek the optimal solution of the problem.

The reminder of this paper is organized as follows: the derivation of our model is described in detail in Section 2. And then, the optimization algorithm is proposed in Section 3. After that, the experimental results on several benchmark datasets and further study are presented in Section 4, followed by the conclusion in Section 5.

2 Proposed Model

2.1 Non-negative Matrix Factorization

Given the data matrix $X = [x_1, x_2, ..., x_n], x_i \in \mathbb{R}^{d\times 1}$, among which d and n are the dimensionality and sample number, respectively. NMF approximates X with the product of two non-negative matrices:

$$\min_{F\geq 0, G\geq 0} \|X - FG^T\|_F^2, \tag{1}$$

where $F \in \mathbb{R}^{d\times c}$ is the cluster centroid and $G \in \mathbb{R}^{n\times c}$ is the cluster indicator matrix. Note that the product of δF and G^T/δ results in the residue error when the scalar $\delta > 0$. To get the unique solution of problem (1), Huang [Huang *et al.*, 2013] imposed the orthogonal constraint on G. Problem (1) can be reformulated as:

$$\min_{G\geq 0, G^TG=I_c} \|X - FG^T\|_F^2, \tag{2}$$

where $I_c \in \mathbb{R}^{c\times c}$ is the identity matrix. Then the optimal solution maintains its uniqueness.

2.2 Linear Discriminant Analysis

LDA tries to learn a linear projection matrix $W \in \mathbb{R}^{d\times m}$ to project the d-dimensional data into the m-dimensional representation. Given a label matrix $L = [l_1, l_2, ..., l_n]^T \in \{0, 1\}^{n\times c}$, Yang [Yang *et al.*, 2011] defined total-class scatter S_t, between-class scatter S_b and within-class scatter S_w:

$$\begin{aligned} S_t &= \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T = XHHX^T, \\ S_b &= \sum_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^T = XHSS^THX^T, \\ S_w &= S_t - S_b, \end{aligned} \tag{3}$$

where μ is the mean of all data points, μ_i is the mean of points in the i-th class, n_i is the number of points in the i-th class. $S = L(L^TL)^{-1/2}$ is the scaled label matrix and $H \in \mathbb{R}^{n\times n}$ is $I_n - \frac{1}{n}1_n1_n^T$. ($I_n$ is the n-dimensional identity matrix and $1_n \in \mathbb{R}^{n\times 1}$ is a column vector with all its elements as 1)

One of the aims of LDA is to find the optimal W to make points from the same class closer to each other. Thus, the objective function can be written as:

$$\min_{W^TS_tW=I_m} Tr(W^TS_wW), \tag{4}$$

where $I_m \in \mathbb{R}^{m\times m}$ is an identity matrix, and $Tr(\cdot)$ is trace operator. According to Eq.(3), problem (4) is equivalent to:

$$\begin{aligned} &\min_{W^TS_tW=I_m} Tr(W^T(S_t - S_b)W) \\ = &\min_{W^TS_tW=I_m} \|W^TXH(I_n - SS^T)\|_F^2. \end{aligned} \tag{5}$$

2.3 Matrix Factorization in Discriminative Subspace

For most of the applications nowadays, high dimensional data is difficult to capture its intrinsic structure but to count on its low-dimensional subspace. In order to find it, a connection between matrix factorization and LDA is proposed. Considering the framework of NMF, when G is fixed, the optimal F is computed as XG. Replace F with XG, then problem (2) becomes:

$$\min \|X - XGG^T\|_F^2 = \min \|X(I_n - GG^T)\|_F^2, \tag{6}$$

which is similar to problem (5). The difference between problem (5) and problem (6) is that X and G are replaced with W^TXH and S, respectively. What's more, the optimal G for problem (2) is the cluster indicator matrix in the matrix framework while S in problem (5) is the scaled label matrix. In this way, G and S have similar meanings in practice. Therefore, the unsupervised matrix factorization framework is connected to LDA:

$$\begin{aligned} &\min_{F,G,W} \|W^TXH - FG^T\|_F^2, \\ &s.t. \quad F \in \mathbb{R}^{m\times c}, G \in \mathbb{R}^{n\times c}, W \in \mathbb{R}^{d\times m}, \\ &G \geq 0, G^TG = I_c, W^TS_tW = I_m, \end{aligned} \tag{7}$$

where G can be considered as the cluster indicator matrix or the pseudo-information matrix. What's more, W is the projection matrix that can find the discriminant subspace in the matrix factorization framework.

According to [Huang *et al.*, 2013], we replace the Frobenius norm with the $l_{2,1}$-norm to improve the robustness. In this way, we get a robust unsupervised framework:

$$\begin{aligned} &\min_{F,G,W} \|W^T XH - FG^T\|_{2,1}, \\ &s.t. \quad F \in \mathbb{R}^{m\times c}, G \in \mathbb{R}^{n\times c}, W \in \mathbb{R}^{d\times m}, \\ &G \geq 0, G^T G = I_c, W^T S_t W = I_m, \end{aligned} \tag{8}$$

2.4 Pseudo Supervised Manifold Regularization

Considering the original data matrix $X \in \mathbb{R}^{d\times n}$ and its projected low-dimensional data matrix $Y \in \mathbb{R}^{m\times n}$, $W \in \mathbb{R}^{d\times m}$ is the matrix which projects the d-dimensional data into the m-dimensional representation, i.e. $Y = W^T X$.

LDA tries to find the subspace that discriminates different classes by minimizing the trace of the within-class scatter matrix S_W while maximizing the trace of the between-class scatter matrix S_B. According to [Zhang *et al.*, 2009], to minimize S_W, we need to solve the following problem:

$$\begin{aligned} &\min Tr(S_W) \\ &= \min_{\vec{y}_i^{(j)}} Tr\left(\sum_{i=1}^{C}\sum_{j=1}^{N_i}(\vec{y}_i^{(j)} - \vec{y}_i^m)(\vec{y}_i^{(j)} - \vec{y}_i^m)^T\right), \end{aligned} \tag{9}$$

where C is the number of classes; N_i is the number of samples in the i-th class; $\vec{y}_i^{(j)}$ is the j-th sample in the i-th class and $\vec{y}_i^m$ is the centroid of the i-th class. Problem (9) can be reduced into:

$$\min_{Y_i} \sum_{i=1}^{N} Tr(Y_i L_i^W Y_i^T), \tag{10}$$

where

$$L_i^W = \frac{1}{N_i^2}\begin{bmatrix} N_i - 1 \\ -\vec{e}_{N_i-1} \end{bmatrix}[N_i - 1 \quad -\vec{e}_{N_i-1}^T],$$

$Y_i = [\vec{y}_i, \vec{y}_{i_1}, ..., \vec{y}_{i_{N_i-1}}]$ and $\vec{e}_{N_i-1} = [1, ..., 1]^T \in \mathbb{R}^{N_i-1}$.

$L_W = \sum_{i=1}^{N} S_i L_i^W S_i^T$ is the alignment matrix which can be obtained by an iterative procedure [Zhang and Zha, 2004]:

$$L_W(F_i, F_i) \leftarrow L_W(F_i, F_i) + L_i^W, \tag{11}$$

for $i = 1, ..., N$ with the initialization $L_W = 0$.

With L_i^W and Eq.(11) [Zhang *et al.*, 2009], we get:

$$\min_Y Tr(Y L_W Y^T). \tag{12}$$

According to the definition of LDA, we replace Y with the form of $W^T X$. So problem (12) is equivalent to this problem:

$$\begin{aligned} &\min_W Tr(W^T X L_W X^T W), \\ &s.t. \quad WW^T = I_d. \end{aligned} \tag{13}$$

Problem (13) is a manifold regularizer which is developed from LDA, so it needs data structure information to instruct itself.

In addition, we assume that any two points in high-density region of the low-dimensional manifold should share the same cluster in perspective of manifold learning. The low-dimensional data graph $S \in \mathbb{R}^{n\times n}$ is built by $W^T X$ when W is fixed. Naturally, we minimize the following problem to get its geometry structure from low-dimensional subspace:

$$\min_G Tr(G^T L_m G), \tag{14}$$

where L_m is the Laplacian matrix of the data graph S.

2.5 Objective Function

By combing the problem (8), (13) and (14) together, we finally get the objective function of the proposed Pseudo Supervised Matrix Factorization (PSMF) method:

$$\begin{aligned} &\min_{F,G,W} \|W^T XH - FG^T\|_{2,1} + \lambda_1 Tr(G^T L_m G) + \\ &\lambda_2 Tr(W^T X L_W X^T W), \\ &s.t. \quad F \in \mathbb{R}^{m\times c}, G \in \mathbb{R}^{n\times c}, W \in \mathbb{R}^{d\times m}, G \geq 0, \\ &G^T G = I_c, W^T S_t W = I_m, WW^T = I_d, \end{aligned} \tag{15}$$

where λ_1 and λ_2 are two non-negative manifold regularization parameters. Every time when W and G are updated, we reconstruct L_m and L_W which are a Laplacian matrix and an alignment matrix for further updates.

3 Optimization

With the constraint $W^T S_t W = I_m$, problem (15) is difficult to solve. So we first disturb the diagonal elements of S_t by a scalar $\epsilon > 0$ which is small enough, and then S_t is positive definite. We can decompose the constraint with Cholesky decomposition in the form of $S_t = R^T R$. Thus, denoting RW as P, and denoting $(R^{-1})^T XH$ and $(R^{-1})^T X$ as A and B, respectively. Problem (15) is simplified into:

$$\begin{aligned} &\min_{F,G,P} \|P^T A - FG^T\|_{2,1} + \lambda_1 Tr(G^T L_m G) + \\ &\lambda_2 Tr(P^T B L_W B^T P), \\ &s.t. \quad F \in \mathbb{R}^{m\times c}, G \in \mathbb{R}^{n\times c}, P \in \mathbb{R}^{d\times m}, G \geq 0, \\ &G^T G = I_c, P^T P = I_m. \end{aligned} \tag{16}$$

The above problem is not convex with three variables, so we propose to solve it with Augmented Lagrangian Multiplier (ALM) [Nie *et al.*, 2015]. Since problem (16) depend on P and G, we introduce three auxiliary variables $E = P^T A - FG^T$, $Z_1 = G$ and $Z_2 = P$. Then problem (16) is equivalent to the following ALM problem:

$$\begin{aligned} &\min_{E,G,Z_1,P,Z_2,F} \|E\|_{2,1} + \lambda_1 Tr(G^T L_m Z_1) + \\ &\lambda_2 Tr(P^T B L_W B^T Z_2) + \frac{\mu}{2}\|P^T A - FG^T - E + \frac{\Lambda_1}{\mu}\|_F^2 + \\ &\frac{\mu}{2}\|G - Z_1 + \frac{\Lambda_2}{\mu}\|_F^2 + \frac{\mu}{2}\|P - Z_2 + \frac{\Lambda_3}{\mu}\|_F^2, \\ &s.t. \quad Z_1 \geq 0, G \geq 0, G^T G = I_c, P^T P = I_m, \end{aligned} \tag{17}$$

where $\mu \in \mathbb{R}^{1\times 1}$ is the ALM parameter, and Λ_1, Λ_2 and Λ_3 are ALM multipliers. Then we optimize each variable iteratively.

3.1 Update E

When fixing all the variables except E, we have:

$$\min_{E} \|E\|_{2,1} + \frac{\mu}{2}\|E - M\|_F^2, \tag{18}$$

where $M = P^T - FG^T + \frac{\Lambda_1}{\mu}$. According to [Huang *et al.*, 2013], the optimal E is computed as:

$$E_{:,q} = \begin{cases} (1 - \frac{1}{\mu\|M_{:,i}\|_2})M_{:,i}, & if\ \|M_{:,i}\|_2 \geq \frac{1}{\mu}. \\ 0, & \text{else.} \end{cases} \tag{19}$$

3.2 Update Z_1

When updating Z_1, problem (17) becomes:

$$\begin{aligned} &\min_{Z_1} \lambda_1 Tr(G^T L_m Z_1) + \frac{\mu}{2}\|G - Z_1 + \frac{\Lambda_2}{\mu}\|_F^2, \\ &s.t. \quad Z_1 \geq 0, \end{aligned} \tag{20}$$

which can be further reduced into a closed-from problem:

$$\begin{aligned} &\min_{Z_1} \|Z_1 - T\|_F^2, \\ &s.t. \quad Z_1 \geq 0, \end{aligned} \tag{21}$$

where $T = G + \frac{\Lambda_2}{\mu} - \frac{\lambda_1}{\mu} L_m G$. Therefore, the optimal Z_1 is:

$$Z_{1_{ij}} = \max(T_{ij}, 0). \tag{22}$$

3.3 Update Z_2

To update Z_2, problem (17) is reduced to:

$$\begin{aligned} &\min_{Z_2} \lambda_2 Tr(P^T B L_W B^T Z_2) + \frac{\mu}{2}\|P - Z_2 + \frac{\Lambda_3}{\mu}\|_F^2, \\ &s.t. \quad Z_2^T Z_2 = I_m. \end{aligned} \tag{23}$$

By expanding the objective function and removing the irrelevant terms, we get:

$$\begin{aligned} &\min_{Z_2} \|Z_2 - R\|_F^2, \\ &s.t. \quad Z_2^T Z_2 = I_m, \end{aligned} \tag{24}$$

where $R = -\frac{\lambda_2}{\mu} B L_W B^T P + P + \frac{\Lambda_3}{\mu}$. It is equivalent to:

$$\begin{aligned} &\max_{Z_2} Tr(Z_2^T R), \\ &s.t. \quad Z_2^T Z_2 = I_m. \end{aligned} \tag{25}$$

And [Huang *et al.*, 2013] proved that the optimal solution of the above problem is:

$$Z_2 = U_3 V_3^T, \tag{26}$$

where $U_3 \in \mathbb{R}^{d\times m}$ and $V_3 \in \mathbb{R}^{m\times m}$ are the left and right singular vectors of the compact singular value decomposition of R.

3.4 Update G

To update G, problem (17) is reduced to:

$$\begin{aligned} &\min_{G} \lambda_1 Tr(G^T L_m Z_1) + \frac{\mu}{2}\|P^T A - FG^T - E \\ &+ \frac{\Lambda_1}{\mu}\|_F^2 + \frac{\mu}{2}\|G - Z_1 + \frac{\Lambda_2}{\mu}\|_F^2, \\ &s.t. \quad G \geq 0, G^T G = I_c. \end{aligned} \tag{27}$$

Similar to Z_2, the optimal solution of the above problem is:

$$G = U_1 V_1^T, \tag{28}$$

where $U_1 \in \mathbb{R}^{n\times c}$ and $V_1 \in \mathbb{R}^{c\times c}$ are the left and right singular vectors of the compact singular value decomposition of K and $K = (-\frac{\lambda_1}{\mu} L_m Z_1) + (P^T A - E + \frac{\Lambda_1}{\mu})^T F + (Z_1 - \frac{\Lambda_2}{\mu})$.

3.5 Update P

To update P, problem (17) is reduced to:

$$\begin{aligned} &\min_{P} \lambda_2 Tr(P^T B L_W B^T Z_2) + \frac{\mu}{2}\|P^T A - FG^T \\ &- E + \frac{\Lambda_1}{\mu}\|_F^2 + \frac{\mu}{2}\|P - Z_2 + \frac{\Lambda_3}{\mu}\|_F^2, \\ &s.t. \quad P^T P = I_m. \end{aligned} \tag{29}$$

Similar to Z_2, the optimal solution of the above problem is:

$$P = U_2 V_2^T, \tag{30}$$

where $U_2 \in \mathbb{R}^{d\times m}$ and $V_2 \in \mathbb{R}^{m\times m}$ are the left and right singular vectors of the compact singular value decomposition of D and $D = -\frac{\lambda_2}{\mu} B L_W B^T Z_2 + A(FG^T + E - \frac{\Lambda_1}{\mu})^T + (Z_2 - \frac{\Lambda_3}{\mu})$.

3.6 Update F

Optimizing problem (17) with regard to F yields the following sub-problem:

$$\min_{F} \frac{\mu}{2}\|P^T A - FG^T - E + \frac{\Lambda_1}{\mu}\|_F^2. \tag{31}$$

Because $G^T G = I_c$, the above problem is reformulated as:

$$\min_{F} \|F - (P^T A - E + \frac{\Lambda_1}{\mu} G)\|_F^2. \tag{32}$$

Finally, the optimal F can be computed as:

$$F = (P^T A - E + \frac{\Lambda_3}{\mu})G. \tag{33}$$

3.7 Update $\mu, \Lambda_1, \Lambda_2$ and Λ_3

The ALM parameters are updated as follows:

$$\begin{aligned} &\Lambda_1 = \Lambda_1 + \mu(P^T A - FG^T - E), \\ &\Lambda_1 = \Lambda_1 + \mu(G - Z_1), \\ &\Lambda_1 = \Lambda_1 + \mu(P - Z_2), \\ &\mu = \rho\mu, \end{aligned} \tag{34}$$

where the parameter ρ controls the convergence speed.

With the updating rules discussed in this section, the optimization algorithm of problem (17) is summarized in Algorithm 1.

Algorithm 1 Algorithm to solve problem (17)

```
Input:
Original data matrix $X \in \mathbb{R}^{d \times n}$;
Number of clusters $k$;
Regularization parameters $\lambda$ and $\mu$;
Output:
Cluster indicator $G$.
Initialize $G \in \mathbb{R}^{n \times c}$ and $W \in \mathbb{R}^{d \times m}$;
Initialize $F = W^T XG$ and $P = RW$ and compute $H, A, B, S_t, S_b, S_w$;
while not converge do
    Update $E$ by Eq.(19);
    Update $Z_1$ by Eq.(22);
    Update $Z_2$ by Eq.(26);
    Update $G$ by Eq.(28);
    Update $P$ by Eq.(30);
    Update $F$ by Eq.(33);
    Calculate ALM parameters by Eq.(34);
end while
```

4 Experiments

In this section, the effectiveness of the proposed PSMF is demonstrated by nine real-world datasets. Parameter sensitivity of it is also discussed here.

4.1 Performance on Benchmark Datasets

In this part, we evaluate the performance of several methods on benchmark datasets to show the effectiveness of our algorithm on data clustering.

Datasets

There are in total nine datasets used in experiments, including one object image dataset, i.e. COIL20 [Cai *et al.*, 2011b], two face image datasets, i.e. YALE [He *et al.*, 2005] and UMIST [Wechsler *et al.*, 2012], and six datasets from the UCI Machine Learning Repository, i.e. Dermatology, Movement, Scale, Iris, Automobile, and Lung-discrete. Table 1 summarizes the characteristics of the datasets used in our experiments.

Evaluation Metrics

Following [Liu *et al.*, 2018], we adopt two widely used evaluation metrics to quantitatively measure clustering performance of our algorithm.

Clustering Accuracy (Acc) [Cai *et al.*, 2005] discovers the one-to-one relationship between clusters and classes and measures the extent to which each cluster contains data points from the corresponding class. It is defined as follows:

$$Acc = \frac{\sum_{i=1}^{n} \delta(map(r_i), l_i)}{n}, \tag{35}$$

where r_i denotes the cluster label and l_i denotes the true class label, n is the total number of samples, $\delta(x, y)$ is the delta function that equals 1 if $x = y$ and equals 0 otherwise, and $map(r_i)$ is the permutation mapping function that maps each r_i to the equivalent label from the data set.

Dataset	Number of Samples	Dimensions	Classes
COIL20	1440	1024	20
YALE	165	1024	15
UMIST	575	644	20
Dermatology	366	34	6
Movement	360	90	15
Scale	625	4	3
Iris	150	4	3
Automobile	205	25	6
Lung-discrete	73	325	7

Table 1: Description of Datasets

Normalized Mutual Information (NMI) [Estévez *et al.*, 2009] is used for determining the quality of clusters. Given a clustering result, it is estimated by:

$$NMI = \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} n_{i,j} \log \frac{n_{i,j}}{n_i \hat{n}_j}}{\sqrt{(\sum_{i=1}^{c} n_i \log \frac{n_i}{n})(\sum_{j=1}^{c} \hat{n}_j \log \frac{\hat{n}_j}{n})}}, \tag{36}$$

where n_i denotes the number of data contained in cluster C_i, $\hat{n}_j$ is the number of data belonging to class L_j, and $n_{i,j}$ denotes the number of data that is in the intersection between cluster C_i and class L_j.

Compared Algorithms

Six state-of-the-art clustering methods are taken for comparison, including K-means, Normalized Cut (NCut) [Shi and Malik, 2000], NMF [Lee and Seung, 1999], Graph Regularized NMF (GNMF) [Cai *et al.*, 2011b], Robust Manifold NMF (RMNMF) [Huang *et al.*, 2013] and Robust manifold Matrix Factorization (RMMF) [Zhang *et al.*, 2017a]. For all the clustering methods, the number of clusters is known as input.

Initialization and Parameters Setting

For K-means, we use a faster Matlab method [Cai, 2011]. For NCut, the data graph is based on the Euclidean distances between two data points. For GNMF and RMNMF, the graph is constructed by finding the five nearest neighbors, in which every edge is weighted from 0 to 1. For NMF, GNMF will degrade to the ordinary NMF when the regularization parameter is 0. For RMMF and our method, the cluster indicator is initialized with the method in [Nie *et al.*, 2014]. All the initial values of μ and Λ are set empirically since they have little influence on the clustering results. K-means, NCut, NMF, GNMF, and RMNMF are sensitive to the initialization, so we run 10 times and calculate the average results.

To compare these methods fairly, we run them with some selected parameter combinations and choose the best results for comparison. For GNMF and RMNMF, we set the regularization parameters α and μ by searching the grid of $\{10^{-5}, 10^{-4}, ..., 10^{4}, 10^{5}\}$. For RMMF, we set the number of iterations as 200 and choose the regularization parameters α and β by searching the grid of $\{10^{-5}, 10^{-4}, ..., 10^{4}, 10^{5}\}$. Finally, for the proposed PSMF method, we set the number of iterations as 50 and choose the regularization parameters λ_1 and λ_2 by searching the grid of $\{10^{-5}, 10^{-4}, ..., 10^{4}, 10^{5}\}$. Note that there is no parameter selection required for K-means and NMF since the number of clusters is given as input.

Datasets	K-means	NCut	NMF	GNMF	RMNMF	RMMF	PSMF
COIL20	54.48	61.85	47.60	49.65	67.92	67.01	**84.58**
YALE	36.85	41.82	36.47	34.24	34.36	32.73	**42.42**
UMIST	38.89	44.28	35.27	33.50	42.05	56.00	**57.04**
Dermatology	77.76	82.60	71.20	87.16	82.21	85.25	**95.36**
Movement	43.89	45.61	36.50	37.78	48.33	47.22	**50.28**
Scale	51.20	47.31	48.82	48.54	51.39	48.80	**53.12**
Iris	82.07	84.20	70.27	70.60	88.93	86.67	**90.00**
Automobile	35.76	30.44	34.83	31.37	34.63	33.17	**37.07**
Lung-discrete	65.62	72.47	72.19	70.41	68.12	75.34	**84.93**

Table 2: Clustering results of different algorithms by the measurement of ACC in percentage

Datasets	K-means	NCut	NMF	GNMF	RMNMF	RMMF	PSMF
COIL20	70.23	74.78	59.03	60.78	76.67	74.50	**92.00**
YALE	43.15	45.59	50.71	38.87	39.91	37.81	**45.60**
UMIST	57.64	62.41	41.96	47.70	63.37	76.66	**78.02**
Dermatology	84.67	83.18	69.86	84.60	86.52	76.44	**91.18**
Movement	57.25	59.71	42.81	42.13	60.47	61.63	**61.97**
Scale	9.57	5.92	9.19	10.18	9.79	5.01	**17.82**
Iris	70.47	75.64	52.95	54.73	74.24	72.85	**78.69**
Automobile	10.46	3.64	6.91	8.10	9.85	9.23	**11.39**
Lung-discrete	63.45	67.60	66.95	67.38	66.36	64.90	**75.71**

Table 3: Clustering results of different algorithms by the measurement of NMI in percentage

Clustering Results

Table 2 and 3 present the ACC and NMI comparison results of all included clustering algorithms on nine datasets. From two tables, we can easily observe that our method always achieves the highest ACC and NMI. In addition, we can find the following detailed points from the results:

1. Compared with other algorithms employing the idea of manifold regularization, such as GNMF, RMNMF, and RMMF, our algorithm shows a better performance both in the measurement of ACC and NMI, which suggests that our method can preserve the local geometrical structure embedded in the high dimensional space well. In other words, PSMF performs better on discovering the intrinsic geometrical and discriminative data structure for the clustering task.
2. The clustering performance of PSMF in datasets with high dimensionalities, such as COIL20, YALE, and Lung-discrete, is extremely better than other algorithms both in the measurement of ACC and NMI, which indicates that our method can effectively figure out the clustering problem of high dimensional data compared with other methods. What's more, PSMF can project the high-dimensional data into a low-dimensional subspace while maintains a good performance on clustering. It shows that PSMF can find a good intrinsic data representation in the low-dimensional subspace.
3. For datasets with categorical attributes, such as Dermatology, COIL20, and UMIST, the clustering performance of PSMF is also better than other algorithms in both two measurements. The reason for it may be that our method can extract comprehensive information of the original data, no matter what type of data is (numerical, categorical or mixed).

4.2 Parameter Sensitivity

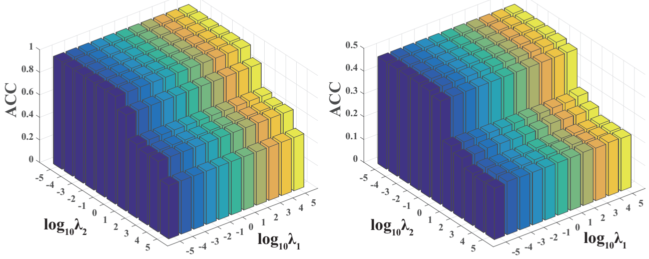
In the proposed PSMF method, there are two manifold regularization parameters, i.e. λ_1 and λ_2, which determine the weight of two manifold regularizations. In order to study the influence of λ_1 and λ_2 on the clustering performance, we tune λ_1 and λ_2 in the same range of $\{10^{-5}, 10^{-4}, ..., 10^4, 10^5\}$ and show the clustering accuracy of PSMF by a 3D visualizable way in Figure 1. According to Figure 1, it is easy to learn that the regularization parameters have much effect on clustering accuracy. Therefore, one of the important future work is to develop a regularization parameter setting rule to get the optimal parameter combination.

5 Conclusion

In this paper, we propose a novel Pseudo Supervised Matrix Factorization (PSMF) in discriminative subspace for data clustering. Different from other manifold regularized clustering methods, our method utilizes the pseudo-information to optimize the objective function iteratively. In addition, the intrinsic geometry structure can be captured in the discriminative subspace by PSMF. Furthermore, the $l_{2,1}$-norm is introduced to enhance the robustness, so that our method is not sensitive to the data outliers. The proposed PSMF can be satisfactorily optimized by the suggested ALM-based method. Extensive experiments on multiple benchmark datasets illustrate that the proposed model outperforms other state-of-the-art clustering algorithms. In future research, we may develop a regularization parameter setting rule for an optimal parameter combination.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 61771349, 61822113 and 61772455, the National Key R&D Program of China under Grant 2018YFA0605501, the Natural Science Foundation of Hubei Province under Grants 2018CFA050 and 2018CFB432 and Yunnan Natural Science Funds under Grant 2018FY001(-013).



(a) Dermatology **(b) Movement**

Figure 1: ACC of PSMF on (a) Dermatology and (b) Movement with varying λ_1 and λ_2

References

[Cai *et al.*, 2005] Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. *IEEE TKDE*, 17(12):1624–1637, 2005.

[Cai *et al.*, 2011a] Deng Cai, Xiaofei He, and Jiawei Han. Locally consistent concept factorization for document clustering. *IEEE TKDE*, 23(6):902–913, 2011.

[Cai *et al.*, 2011b] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE TPAMI*, 33(8):1548–1560, 2011.

[Cai, 2011] Deng Cai. Litekmeans: the fastest matlab implementation of kmeans. *Available at: http://www.zjucadcg.cn/dengcai/Data/Clustering.html*, 2011.

[Chung and Graham, 1997] Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. American Mathematical Soc., 1997.

[Ding *et al.*, 2010] Chris H. Q. Ding, Tao Li, and Michael I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE TPAMI*, 32(1):45–55, 2010.

[Estévez *et al.*, 2009] Pablo A. Estévez, M. Tesmer, Claudio A. Perez, and Jacek M. Zurada. Normalized mutual information feature selection. *IEEE TNN*, 20(2):189–201, 2009.

[He *et al.*, 2005] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and HongJiang Zhang. Face recognition using laplacianfaces. *IEEE TPAMI*, 27(3):328–340, 2005.

[Huang *et al.*, 2013] Jin Huang, Feiping Nie, Heng Huang, and Chris H. Q. Ding. Robust manifold nonnegative matrix factorization. *ACM TKDD*, 8(3):11:1–11:21, 2013.

[Jiang *et al.*, 2004] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: A survey. *IEEE TKDE*, 16(11):1370–1386, 2004.

[Ke and Kanade, 2005] Qifa Ke and Takeo Kanade. Robust l1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *Proc. CVPR*, pages 739–746, 2005.

[Lee and Seung, 1999] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[Liu *et al.*, 2017] Weiwei Liu, Xiao-Bo Shen, and Ivor W. Tsang. Sparse embedded k-means clustering. In *Proc. NIPS*, pages 3321–3329, 2017.

[Liu *et al.*, 2018] Yang Liu, Quanxue Gao, Zhaohua Yang, and Shujian Wang. Learning with adaptive neighbors for image clustering. In *Proc. IJCAI*, pages 2483–2489, 2018.

[Lu *et al.*, 2017] Yuwu Lu, Zhihui Lai, Xu Yong, Xuelong Li, David Zhang, and Chun Yuan. Nonnegative discriminant matrix factorization. *IEEE TCSVT*, 27(7):1392–1405, 2017.

[Ma *et al.*, 2018] Sihan Ma, Lefei Zhang, Wenbin Hu, Yipeng Zhang, Jia Wu, Xuelong Li, et al. Self-representative manifold concept factorization with adaptive neighbors for clustering. In *Proc. IJCAI*, pages 2539–2545, 2018.

[Ng *et al.*, 2002] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proc. NIPS*, pages 849–856, 2002.

[Nie *et al.*, 2014] Feiping Nie, Xiaoqian Wang, and Heng Huang. Clustering and projected clustering with adaptive neighbors. In *Proc. ACM SIGKDD*, pages 977–986, 2014.

[Nie *et al.*, 2015] Feiping Nie, Hua Wang, Heng Huang, and Chris Ding. Joint schatten p-norm and ℓ_p-norm robust matrix completion for missing value recovery. *KAIS*, 42(3):525–544, 2015.

[Shen *et al.*, 2017] Xiao-Bo Shen, Weiwei Liu, Ivor W. Tsang, Fumin Shen, and Quan-Sen Sun. Compressed k-means for large-scale clustering. In *Proc. AAAI*, pages 2527–2533, 2017.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, 2000.

[Wang *et al.*, 2017] Boyue Wang, Yongli Hu, Junbin Gao, Yanfeng Sun, Haoran Chen, Muhammad Ali, and Baocai Yin. Locality preserving projections for grassmann manifold. In *Proc. IJCAI*, pages 2893–2900, 2017.

[Wang *et al.*, 2018] Jing Wang, Feng Tian, Weiwei Liu, Xiao Wang, Wenjie Zhang, and Kenji Yamanishi. Ranking preserving nonnegative matrix factorization. In *Proc. IJCAI*, pages 2776–2782, 2018.

[Wechsler *et al.*, 2012] Harry Wechsler, Jonathon P Phillips, Vicki Bruce, Francoise Fogelman Soulie, and Thomas S Huang. *Face recognition: From theory to applications*, volume 163. 2012.

[Yang *et al.*, 2011] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. L2,1-norm regularized discriminative feature selection for unsupervised learning. In *Proc. IJCAI*, pages 1589–1594, 2011.

[Zhang and Zha, 2004] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *Journal of Shanghai University*, 8(4):406–424, 2004.

[Zhang *et al.*, 2009] Tianhao Zhang, Dacheng Tao, Xuelong Li, and Yang Jie. Patch alignment for dimensionality reduction. *IEEE TKDE*, 21(9):1299–1313, 2009.

[Zhang *et al.*, 2017a] Lefei Zhang, Qian Zhang, Bo Du, Dacheng Tao, and Jane You. Robust manifold matrix factorization for joint clustering and feature extraction. In *Proc. AAAI*, pages 1662–1668, 2017.

[Zhang *et al.*, 2017b] Lefei Zhang, Qian Zhang, Bo Du, Jane You, and Dacheng Tao. Adaptive manifold regularized matrix factorization for data clustering. In *Proc. IJCAI*, pages 3399–3405, 2017.