Adaptive Manifold Regularized Matrix Factorization for Data Clustering

Lefei Zhang[1], **Qian Zhang**[2], **Bo Du**[1*], **Jane You**[3], **Dacheng Tao**[4]
[1] School of Computer, Wuhan University, [2] Alibaba Group,
[3] Department of Computing, The Hong Kong Polytechnic University,
[4] UBTech Sydney AI Institute, The School of Information Technologies, The University of Sydney,
zhanglefei@whu.edu.cn, qianzhang.zq@alibaba-inc.com, remoteking@whu.edu.cn,
csyjia@comp.polyu.edu.hk, dacheng.tao@sydney.edu.au

Abstract

Data clustering is the task to group the data samples into certain clusters based on the relationships of samples and structures hidden in data, and it is a fundamental and important topic in data mining and machine learning areas. In the literature, the spectral clustering is one of the most popular approaches and has many variants in recent years. However, the performance of spectral clustering is determined by the affinity matrix, which is usually computed by a predefined model (e.g., Gaussian kernel function) with carefully tuned parameters combination, and may not optimal in practice. In this paper, we propose to consider the observed data clustering as a robust matrix factorization point of view, and learn an affinity matrix simultaneously to regularize the proposed matrix factorization. The solution of the proposed adaptive manifold regularized matrix factorization (AMRMF) is reached by a novel Augmented Lagrangian Multiplier (ALM) based algorithm. The experimental results on standard clustering datasets demonstrate the superior performance over the exist alternatives.

1 Introduction

The task of data clustering partitions the input data samples into certain clusters such that the samples in a same group would share high similarity to each other, and it has been widely investigated in the data mining and machine learning areas [Zhang *et al.*, 2012; Wang *et al.*, 2013; Xu *et al.*, 2017; Liu *et al.*, 2017]. Since the data clustering could be regarded as a special case of classification but without any training data, the clustering results always highly depend on the data similarity learning [Nie *et al.*, 2014]. In the past decades, many clustering algorithms have been proposed, e.g., k-means clustering [Hartigan and Wong, 1979], hierarchical clustering [Jain *et al.*, 1999], spectral clustering [von Luxburg, 2007; Shi and Malik, 2000], subspace clustering [Vidal, 2011], and matrix factorization based methods [Ding *et al.*, 2010]. Among which, our proposed method shares both of the advantages of the spectral clustering and matrix factorization, therefore we briefly review the related works along these two directions, respectively.

Spectral clustering is one of the most important clustering techniques in the literature and has demonstrated its strong capability in group objects by analyzing complex data structural information [Yang *et al.*, 2016]. Specifically, it assumes that any two data points in the high density region of the low dimensional data manifold should share the same cluster. In order to capture the nonlinear and low dimensional manifold structure of the input data, an affinity matrix (or data similarity matrix) is required as input [He *et al.*, 2011; Guo, 2015], and then the cluster assignment could be obtained by the spectrum of that affinity matrix. Traditionally, the Gaussian kernel function is usually employed to construct such affinity matrix [Huang *et al.*, 2015]. In recent years, some advanced techniques have been proposed to explore some better affinity matrices or data representations, e.g., constrained Laplacian rank [Nie *et al.*, 2016a], clustering with adaptive neighbors [Nie *et al.*, 2014], low-rank representation [Liu *et al.*, 2013], least squares regression [Lu *et al.*, 2012], and robust subspace segmentation [Guo, 2015]. However, the spectral clustering is basically a two-steps approach, thus usually a simple k-means would be used after the affinity matrix has been learned, which may cause sub-optimal and unscalable clustering results due to the drawbacks inherited by the k-means clustering [Xu *et al.*, 2016].

The non-negative matrix factorization (NMF) aims to find two non-negative matrices whose product provides a good approximation to the observation feature matrix [Liu and Tao, 2016], thus it could be used for data clustering by interpreting the two factor matrices as the cluster indicator and latent feature matrix, respectively [Ding *et al.*, 2010; Cai *et al.*, 2011]. In addition, various of priories have been added into the model to regularize the matrix factorization more fit for the task of data clustering, e.g., the manifold regularization [Huang *et al.*, 2014] and sparsity constraint [Wang *et al.*, 2015]. However, there are still following issues which may easily achieve poor performance in the matrix factorization based clustering. Firstly, as discussed above, most of exist methods directly introduce the Gaussian kernel derived Laplacian matrix for manifold regularization, while ignore to explore a more meaningful affinity matrix to better regularize the model. Secondly, the s-

[*] Corresponding author.

tandard NMF uses the ℓ_2-norm based squared residue minimization to measure the loss, which would be easily effected by the noises and outliers [Meng and la Torre, 2013; Huang *et al.*, 2014].

In this contribution, we propose a novel adaptive manifold regularized matrix factorization (AMRMF) algorithm for data clustering, which avoids the risks mentioned in the above sections. In detail, the major advantages of the proposed algorithm are summarized as follows.

- The AMRMF model regards the clustering in a matrix factorization point of view, thus the desired data labels could be explicitly obtained from one of the factor matrices, i.e., the cluster indicator matrix. By such way, the AMRMF could always arrive at a scalable and reproduceable result in practice, which is superior to the traditional methods such as k-means and spectral clustering.
- The AMRMF model jointly learns an affinity matrix with the matrix factorization, therefore, the ideal data similarity under our assumption has been well uncovered, and the learned affinity matrix may better guide the manifold regularization to fit the clustering task. Compare to the exist graph clustering methods which directly address the Gaussian kernel function to construct the affinity matrix, the proposed AMRMF reveals a more flexible, meaningful but parameter-free way.
- The AMRMF model employs the $\ell_{2,1}$-norm to measure the loss of matrix factorization, therefore, compare to the conventional ℓ_2-norm based matrix factorization, the proposed AMRMF model would not sensitive to the data noises and outliers and could be better applied to practical data mining applications.

However, since the proposed AMRMF model learns the data similarity matrix and cluster indicator simultaneously and introduces a new constraint (i.e., the $\ell_{2,1}$-norm), the conventional auxiliary function optimization method is no longer applicable for our AMRMF problem. Therefore, we have figured out a new Augmented Lagrangian Multiplier (ALM) based procedure to get the solution of our proposed objective function. The rest of the paper is structured as follows: section 2 introduces the detailed objective function of the proposed AMRMF model, section 3 presents an efficient optimization procedure to solve the AMRMF objective function. After that, the experimental results on several real world datasets are reported in section 4, followed by the conclusions in section 5.

2 Adaptive Manifold Regularized Matrix Factorization

In this section, we present the proposed AMRMF model by first formulating the objective function of robust matrix factorization, and then introducing the steps to learn the affinity matrix simultaneously. After that, an efficient algorithm to tackle the AMRMF model is discussed in the next section. Throughout this paper, the matrix is represented as $A \in \mathbb{R}^{p\times q}$, in which the (i,j)-th element of A is denoted by a_{ij}, the i-th column of A is denoted by a vector $a_i \in \mathbb{R}^p$. The trace and transpose of A are denoted by $\mathrm{tr}(A)$ and A^{T}, respectively. The F-norm and $\ell_{2,1}$-norm of A are denoted by $\|A\|_F$ and $\|A\|_{2,1}$, respectively.

Given a data matrix $X \in \mathbb{R}^{l\times n}$, in which l and n are the the feature dimensionality of each sample and number of data samples, respectively. Therefore, each data sample could be denoted by $x_i \in \mathbb{R}^l$. In this paper, we propose to perform the matrix factorization on X and consider one of the factor matrices as the cluster indicator which divides X into k clusters:

$$\arg\min_{U,V} \left\|X - VU^{\mathrm{T}}\right\|_F^2, \text{s.t. } U^{\mathrm{T}}U = I, U \geq 0, \quad (1)$$

where $V \in \mathbb{R}^{l\times k}$ is the latent feature matrix (or the cluster centroid) and $U \in \mathbb{R}^{n\times k}$ is the cluster indicator [Cai *et al.*, 2011; Huang *et al.*, 2014; Trigeorgis *et al.*, 2014]. Note that in the objective function eq. (1), the loss of matrix factorization is measured by the squared residue error in the form of ℓ_2-norm, therefore, the noises and outliers in the dataset with large reconstruction errors will heavily effect the matrix factorization because the errors have been enlarged. In order to relieve this issue, in this paper, we introduce the $\ell_{2,1}$-norm instead of the conventional ℓ_2-norm based matrix factorization, which avoids the samples with large errors (i.e., the noises and outliers) dominate the objective function and thus makes the model more robust. Then we have the following improved matrix factorization optimization:

$$\arg\min_{U,V} \left\|X - VU^{\mathrm{T}}\right\|_{2,1}, \text{s.t. } U^{\mathrm{T}}U = I, U \geq 0. \quad (2)$$

However, the model in eq. (2) would result in unsatisfactory clustering performance since there is no further constraints on the cluster indicator U. In the literature, the manifold regularization is often adopted to let the similar data samples from X share the similar clustering labels [Gong *et al.*, 2015; Zhang *et al.*, 2015], and in most of the cases, the Laplacian regularization is incorporated with the matrix factorization [Cai *et al.*, 2011; Huang *et al.*, 2014]. However, in the exist works, the affinity matrix is usually computed by a predefined model (e.g., Gaussian kernel function) with carefully tuned parameters combination, which may not optimal in practice. Therefore, in this paper, we propose to learn an affinity matrix to better regularize the proposed matrix factorization in eq. (2).

To begin with the data similarity learning, we suppose that each sample x_i could be linked to any other sample x_j with probability s_{ij}, where s_{ij} is an element of the expected similarity matrix S. Obviously, we suggest that the similar sample pair with small distance $\|x_i - x_j\|_2^2$ should be assigned a high probability s_{ij}. Therefore, we have the following objective function to optimize the S which meets our assumption:

$$\begin{aligned} &\arg\min_{S} \sum_{i,j} \|x_i - x_j\|_2^2 s_{ij} + \alpha\|S\|_F^2, \\ &\text{s.t. } \sum\nolimits_j s_{ij} = 1, 1 \geq s_{ij} \geq 0, \end{aligned} \quad (3)$$

where α is the regularization parameter.

In order to further make the model benefit for clustering, we need the similarity matrix contains exact k connected components rather than all the elements have been linked together. Note that we have the equation that $\sum_{ij} \|u_i - u_j\|_2^2 s_{ij} = 2\mathrm{tr}(U^{\mathrm{T}} L_S U)$ in which L_S is the Laplacian matrix computed by the learned S, according to [Chung, 1997], then, the optimization in eq. (3) could achieve an ideal neighbors assignment if we add an additional constraint $\mathrm{rank}(L_S) = n - k$. However, the eq. (3) with this rank constraint is hard to solve. To track this issue, we relax the problem followed by [Nie *et al.*, 2016b]. Let $\sigma_i(L_S)$ denotes the i-th smallest eigenvalue of L_S, since the L_S must be positive semi-definite, then we have $\sigma_i(L_S) \geq 0$. Therefore, the rank constraint $\mathrm{rank}(L_S) = n - k$ equals to $\sum_{i=1}^{k} \sigma_i(L_S) = 0$. On the other hand, we also have the fact that $\sum_{i=1}^{k} \sigma_i(L_S) = \min_U \mathrm{tr}(U^{\mathrm{T}} L_S U)$, thus, we rewrite problem eq. (3) as:

$$\begin{aligned} &\arg\min_{S,U} \sum_{i,j} \|x_i - x_j\|_2^2 s_{ij} + \alpha \|S\|_F^2 + 2\gamma \mathrm{tr}(U^{\mathrm{T}} L_S U), \\ &\quad \text{s.t. } \forall i, \sum\nolimits_j s_{ij} = 1, 1 \geq s_{ij} \geq 0, \\ &\qquad U \in \mathbb{R}^{n \times k}, U^{\mathrm{T}} U = I, U \geq 0. \end{aligned} \tag{4}$$

Note that optimization in eq. (4) is a relaxation of eq. (3) with the rank constraint as long as γ has been set as a large enough value, in such condition, $\mathrm{tr}(U^{\mathrm{T}} L_S U)$ would be forced to close to zero and thus $\sum_{i=1}^{k} \sigma_i(L_S) = 0$ would be satisfied accordingly.

Finally, by combining the eqs. (2) and (4) together with an additional regularization parameter β, we have the objective function of the proposed AMRMF clustering model:

$$\begin{aligned} &\arg\min_{S,U,V} \|X - VU^{\mathrm{T}}\|_{2,1} + 2\gamma \mathrm{tr}(U^{\mathrm{T}} L_S U) \\ &\quad + \beta(\sum_{i,j} \|x_i - x_j\|_2^2 s_{ij} + \alpha \|S\|_F^2), \\ &\text{s.t. } \forall i, \sum\nolimits_j s_{ij} = 1, 1 \geq s_{ij} \geq 0, U^{\mathrm{T}} U = I, U \geq 0. \end{aligned} \tag{5}$$

3 AMRMF Optimization

The objective function in above eq. (5) is not convex in three variables, thus, we consider to use an Augmented Lagrangian Multiplier (ALM) method to optimize them alteratively. By introducing two auxiliary variables $E = X - VU^{\mathrm{T}}$ and $Z = U$. The objective function can be rewritten into the following equivalent problem:

$$\begin{aligned} &\arg\min_{S,U,V,E,Z} \|E\|_{2,1} + 2\gamma \mathrm{tr}(Z^{\mathrm{T}} L_S U) \\ &\quad + \beta(\sum_{ij} \|x_i - x_j\|_2^2 s_{ij} + \alpha \|S\|_F^2), \\ &\text{s.t. } E = X - VU^{\mathrm{T}}, Z = U, U^{\mathrm{T}} U = I, Z \geq 0, \\ &\quad \forall i, \sum\nolimits_j s_{ij} = 1, 1 \geq s_{ij} \geq 0, \end{aligned} \tag{6}$$

which can be solved by the following ALM problem:

$$\begin{aligned} &\arg\min_{S,U,V,E,Z,\lambda_1,\lambda_2,\mu} \|E\|_{2,1} + 2\gamma \mathrm{tr}(Z^{\mathrm{T}} L_S U) \\ &\quad + \beta(\sum_{ij} \|x_i - x_j\|_2^2 s_{ij} + \alpha \|S\|_F^2) \\ &\quad + < \lambda_1, X - VU^{\mathrm{T}} - E > + < \lambda_2, Z - U > \\ &\quad + \frac{\mu}{2}(\|X - VU^{\mathrm{T}} - E\|_F^2 + \|Z - U\|_F^2), \\ &\text{s.t. } U^{\mathrm{T}} U = I, Z \geq 0, \forall i, \sum\nolimits_j s_{ij} = 1, 1 \geq s_{ij} \geq 0, \end{aligned} \tag{7}$$

where λ_1 and λ_2 are the Lagrangian multipliers and μ is a regularity coefficient to control the penalty for the two violation of equality constraints in eq. (7). Since the objective function above carries five variables and additional multipliers, we adopt an alternative optimization method to reduce it to a few manageable subproblems with the closed form solution, each minimizes the objective function with respect to one variable while fixing the other variables.

Update S

To update S, we fix other variables except S and remove terms that are irrelevant to S. Denote $d_{ij}^{x} = \|x_i - x_j\|_2^2$, then eq. (7) becomes:

$$\begin{aligned} &\arg\min_{S} \sum_{i,j} (d_{ij}^{x} s_{ij} + \alpha s_{ij}^2) + \frac{2\gamma}{\beta} \mathrm{tr}(Z^{\mathrm{T}} L_S U), \\ &\quad \text{s.t. } \forall i, \sum\nolimits_j s_{ij} = 1, 1 \geq s_{ij} \geq 0. \end{aligned} \tag{8}$$

Denote $d_{ij}^{uz} = \|z_i - u_j\|_2^2$, note that the problem above is independent between different i, we can deal with following problem individually for each i:

$$\begin{aligned} &\arg\min_{s_i} \sum_{j=1}^{n} (d_{ij}^{x} s_{ij} + \alpha s_{ij}^2 + \frac{\gamma}{\beta} d_{ij}^{uz} s_{ij}), \\ &\quad \text{s.t. } \forall i, \sum\nolimits_j s_{ij} = 1, 1 \geq s_{ij} \geq 0. \end{aligned} \tag{9}$$

Denote $d_i \in \mathbb{R}^n$ is a vector with the j-th element as $d_{ij} = d_{ij}^{x} + \frac{\gamma}{\beta} d_{ij}^{uz}$, then the above problem can be rewritten as follows:

$$\arg\min_{s_i} \|s_i - \frac{1}{2\alpha} d_i\|_2^2, \text{s.t.} \forall i, \sum\nolimits_j s_{ij} = 1, 1 \geq s_{ij} \geq 0. \tag{10}$$

Update U

To update U, we fix other variables except U and remove terms that are irrelevant to U. Then eq. (7) becomes:

$$\begin{aligned} &\arg\min_{U^{\mathrm{T}} U = I} < \lambda_1, X - VU^{\mathrm{T}} - E > + < \lambda_2, Z - U > \\ &+ \frac{\mu}{2}(\|X - VU^{\mathrm{T}} - E\|_F^2 + \|Z - U\|_F^2) + 2\gamma \mathrm{tr}(Z^{\mathrm{T}} L_S U), \end{aligned} \tag{11}$$

which can be further reduced as following:

$$\arg\min_{U^{\mathrm{T}} U = I} \frac{\mu}{2} \|U\|_F^2 - \mu < H, U >, \tag{12}$$

where

$$H = \frac{1}{\mu} \lambda_2 + Z - \frac{2\gamma}{\mu} L_S Z + (X - E + \frac{1}{\mu} * \lambda_1)^{\mathrm{T}} V, \tag{13}$$

Thus, we further arrives:

$$\arg\min_{U^{\mathrm{T}}U=I} \|U - H\|_F^2. \tag{14}$$

Denote:

$$L(U, \Lambda) = \|U - H\|_F^2 + \Lambda(UU^{\mathrm{T}} - I). \tag{15}$$

We then have:

$$U = N_u Q_u^{\mathrm{T}}, \tag{16}$$

where N_u and Q_u are the left and right singular vectors of the economic singular value decomposition of H.

Update V

To update V, we fix other variables except V, then obtain the following objective function:

$$\arg\min_V \frac{\mu}{2}\|X - VU^{\mathrm{T}} - E + \frac{1}{\mu}\lambda_1\|_F^2. \tag{17}$$

Considering that $U^{\mathrm{T}}U = I$, we can rewrite the above objective function as:

$$\arg\min_V \frac{1}{2}\|V - (X - E + \frac{1}{\mu}\lambda_1)U\|_F^2, \tag{18}$$

then we have $V = (X - E + \frac{1}{\mu}\lambda_1)U$.

Update E

To update E, we fixed other variables except E and remove terms that are irrelevant to E. The the objective function becomes:

$$\arg\min_E \frac{1}{2}\|E - (X - VU^{\mathrm{T}} + \frac{1}{\mu}\lambda_1)\|_F^2 + \frac{1}{\mu}\|E\|_{2,1}. \tag{19}$$

Let $B = X - VU^{\mathrm{T}} + \frac{1}{\mu}\lambda_1$, then E can be updated as:

$$e_i = \begin{cases} (1 - \frac{1}{\mu\|b_i\|})b_i, & \text{if} \|b_i\| \geq \frac{1}{\mu}, \\ 0, & \text{otherwise}. \end{cases} \tag{20}$$

Update Z

Optimizing eq. (7) with respect to Z yields the equation:

$$\arg\min_{Z\geq 0} \frac{\mu}{2}\|Z - U\|_F^2 + < \lambda_2, Z - U > + 2\gamma \mathrm{tr}(Z^{\mathrm{T}} L_S U). \tag{21}$$

Then we obtain:

$$\arg\min_{Z\geq 0} \|Z - K\|_F^2, \tag{22}$$

where $K = (U - \frac{1}{\mu}\lambda_2 - \frac{2\gamma}{\mu}L_S U)$.

The above object function can be further decomposed to element-wise optimization problem as:

$$\arg\min_{z_{ij}\geq 0} \|z_{ij} - k_{ij}\|^2. \tag{23}$$

Therefore, the optimal solution of above problems is:

$$z_{ij} = \max(k_{ij}, 0). \tag{24}$$

Update ALM Parameters

Finally we need to update the ALM parameters, i.e., λ_1, λ_2, and μ. According to [Boyd and Vandenberghe, 2004], they should be updated as following:

$$\lambda_1 = \lambda_1 + \mu(X - VU^{\mathrm{T}} - E). \tag{25}$$

$$\lambda_2 = \lambda_2 + \mu(Z - U). \tag{26}$$

$$\mu = \rho\mu. \tag{27}$$

Table 1: Description of datasets for data culstering.

Dataset	Classes	Samples	Features
Caltech101 Silhouettes	101	8461	256
COIL20	20	1440	1024
Control	6	600	60
Dermatology	6	366	34
Ecoli	8	336	343
Movement	15	360	90
MSRA25	12	1799	256
PalmData25	100	2000	256
Seeds	3	210	7
USPS	10	9298	256

4 Experimental Analysis

In this section, we evaluate the performance of the proposed AMRMF method on ten real world benchmark datasets (Table 1). In detail, we firstly introduce the datasets and experimental settings, and then compare the proposed AMRMF with the state-of-the-art clustering algorithms and provide our observations. Finally, the convergence performance of the proposed AMRMF optimization is reported based on all the involved datasets.

The data clustering experiments are conducted on ten public available benchmark datasets, including five image datasets (Caltech101 Silhouettes, COIL20, MSRA25, PalmData25, and USPS) and five non-image datasets from the UCI machine learning repository (Control, Dermatology, Ecoli, Movement, and Seeds). For each dataset, the data is formatted as a feature matrix (i.e., input X of AMRMF algorithm) with the size of number of features and number of samples, and an additional ground truth vector with the size of number of samples. The statistics of the datasets used in the clustering experiments are summarized in Table 1. After the AMRMF algorithm has ended, the predicted label vector is obtained from the cluster indicator (i.e., output U of AMRMF algorithm), and then compared with the ground truth vector for quantitative evaluation.

The following two well accepted measurements have been used as metrics:

- Clustering accuracy (ACC), which discovers the one-to-one relationship between clusters and classes:

  $$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \delta(\mathrm{map}(r_i), l_i)}{n}, \tag{28}$$

  where r_i and l_i are predicted and ground truth label of sample x_i, respectively, $\delta(x, y)$ is the delta function that equals 1 if $x = y$ and equals 0 otherwise, and $\mathrm{map}(r_i)$ is the permutation mapping function that maps each cluster r_i to the equivalent label from the dataset.

- Normalized mutual information (NMI), which measures the the quality of clusters:

  $$\mathrm{NMI} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{i,j} \log\frac{n_{i,j}}{n_i \hat{n}_j}}{\sqrt{(\sum_{i=1}^{k} n_i \log\frac{n_i}{n})(\sum_{j=1}^{k} \hat{n}_j \log\frac{\hat{n}_j}{n})}}, \tag{29}$$

  where n_i denotes the number of data contained in the cluster $C_i (1 \leq i \leq k)$, $\hat{n}_j$ is the number of data belonging to the $L_j (1 \leq j \leq k)$, and $n_{i,j}$ denotes the number

Table 2: Clustering results of different methods by the measurement of ACC.

Dataset	k-means	NCuts	LRR	LSR	RSS	SSC	AMRMF
Caltech101 Silhouettes	0.5457±0.0601	0.5731±0.0683	0.5808	0.6055	0.6526	0.5637	**0.6552**
COIL20	0.5660±0.0560	0.5925±0.0506	0.6007	0.6340	0.5465	0.7521	**0.8576**
Control	0.5823±0.0661	0.6200±0.0466	0.5917	0.6050	0.6983	0.6750	**0.7000**
Dermatology	0.7495±0.1068	0.8235±0.0364	0.9399	0.9426	0.9208	0.9344	**0.9617**
Ecoli	0.5714±0.0526	0.5283±0.0360	0.6696	0.7411	0.7440	0.6190	**0.8214**
Movement	0.4450±0.0228	0.4647±0.0192	0.5000	0.5056	0.5222	**0.5250**	0.5167
MSRA25	0.5252±0.0660	0.5501±0.0395	0.5709	0.5759	0.5675	0.5742	**0.5770**
PalmData25	0.7426±0.0536	0.7900±0.0371	0.7680	0.8740	0.8710	0.8550	**0.8890**
Seeds	0.7086±0.0725	0.8557±0.1020	0.8857	0.9143	0.9000	0.9048	**0.9333**
USPS	0.6322±0.0237	0.6653±0.0372	0.6690	0.7008	0.7094	0.6913	**0.7309**

Table 3: Clustering results of different methods by the measurement of NMI.

Dataset	k-means	NCuts	LRR	LSR	RSS	SSC	AMRMF
Caltech101 Silhouettes	0.5684±0.0021	0.5255±0.0493	0.5416	0.5352	**0.6008**	0.5363	0.5573
COIL20	0.7354±0.0264	0.7322±0.0225	0.7309	0.7325	0.6572	0.8683	**0.9220**
Control	0.6612±0.0527	0.6675±0.0504	0.6068	0.6311	0.6890	0.6971	**0.7959**
Dermatology	0.8462±0.0350	0.8417±0.0373	0.8690	0.8852	0.8292	0.8580	**0.9234**
Ecoli	0.5271±0.0304	0.4495±0.0259	0.5712	0.5399	0.5218	0.4847	**0.6506**
Movement	0.5754±0.0108	0.5993±0.0138	0.6017	0.5868	**0.6423**	0.6303	0.6353
MSRA25	0.5910±0.0480	0.5994±0.0220	0.6312	0.6116	0.6493	0.6683	**0.7122**
PalmData25	0.9026±0.0198	0.9285±0.0135	0.8930	0.9562	**0.9592**	0.9433	0.9538
Seeds	0.4835±0.0437	0.6647±0.1002	0.6576	0.7145	0.6872	0.6990	**0.7658**
USPS	0.5969±0.0058	0.6446±0.0186	0.6491	0.6918	0.7190	0.6817	**0.7350**

of data that are in the intersection between cluster C_i and class L_j.

In the our experiment, we compare the AMRMF algorithm with k-means, Normalized Cuts (NCuts) [Shi and Malik, 2000], low-rank representation (LRR) [Liu *et al.*, 2013], least squares regression (LSR) [Lu *et al.*, 2012], robust subspace segmentation (RSS) [Guo, 2015], and sparse subspace clustering (SSC) [Elhamifar and Vidal, 2013]. Among which, the k-means is executed by the Matlab R2015b statistical toolbox, while the codes of others are downloaded from the authors' webpages. To obtain the best possible performance of the compared methods, the detailed experimental settings are as follows.

For k-means and NCuts, we repeat the experiments ten times and report the average results with standard deviations. In particular, we tune the Gaussian kernel parameter in the range of $10^{[-5:5]}$ for NCuts. For the released codes of LRR, LSR, RSS, and SSC, since the authors have fixed the detailed parameters in the last step of k-means clustering (e.g., initialization, distance measure and number of repeats), the algorithms output relatively stable results so we need not to average their results. In detail, for LRR, we have tried the two versions of LRR uploaded by the authors, with the $\lambda \in [0.001, 0.01, 0.02, 0.05, 0.1, 0.2, 1, 2, 5]$, and report the best results. For LSR, we have also tested two implementations by the authors, with the $\lambda \in 2^{[-5:5]}$ to find the best performance. For RSS, we have tuned the three regularizer weights in the same range of $2^{[-3:3]}$, respectively, which is a much wider range than the author recommended. For SSC, the parameter spaces are $\alpha \in 10^{[-5:5]}$, $\rho \in [1 : 5]$, respectively. Finally, for the proposed AMRMF algorithm, we have tuned the regularization parameters in the same range of $10^{[-5:5]}$. Note that our proposed AMRMF algorithm outputs stable result by the suggested optimization steps, therefore, we also need not to average the reported results.

Tables 2 and 3 summarize the clustering performance for each method on ten datasets. We can see that AMRMF algorithm outperforms other clustering methods in most of the cases. In particular, the Caltech101 Silhouettes is a large dataset which has rarely been considered for clustering experiment. The significant performance on all these datasets, especially the Caltech101 Silhouettes, meets the advantages of our proposed method. To emphasize, the superiority of AMRMF algorithm arises in the following aspects. Firstly, the objective function combines the matrix factorization and similarity learning into a single framework, by which an effective affinity matrix has been optimized and it could be further benefit for the clustering task. Besides, the introduced $\ell_{2,1}$-norm helps to alleviate the data noises and outliers issues that are common among other clustering methods, which brings positive effects into the proposed clustering model.

Also from the experimental results above, we have the following detailed observations.

The basic data clustering algorithms, i.e., k-means and NCuts, present poor clustering rates with large variations (in many times greater than 0.05 in ACC and 0.03 in NMI, respectively) on all the datasets. In detail, the k-means algorithm presents larger variations compares to NCuts on most of the datasets, which makes the clustering results impossible to be reproduced in practice. In fact, the NCuts could be viewed as the k-means clustering but in a new representation rather than the raw input data, since the clustering accuracies

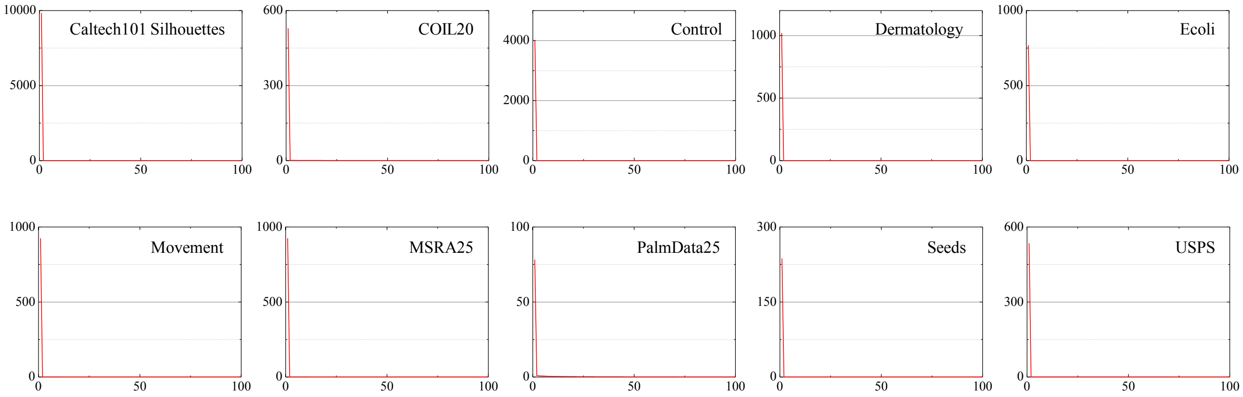

Figure 1: Convergence performance of the proposed algorithm on all the datasets.

of NCuts always outperform the k-means as shown in table 2, we could know that it is crucial to seek a better feature representation (or data similarity).

The advanced algorithms, i.e., LRR, LSR, RSS and SSC, which could be understood as to learn a better affinity matrix for spectral clustering (or consider the learned affinity matrix itself as a better feature representation for clustering), usually show respectable improvements compare to the basic algorithms mentioned above. For those algorithms, we have also experimentally found that the variations on clustering accuracies have been greatly reduced even use the free k-means (without any parameter settings) for final clustering. However, it is also observed that for some datasets, the clustering accuracies are comparable or even lower than the basic algorithms, e.g., in the Caltech101 Silhouettes dataset, the mean ACC of NCuts reaches 0.5731 while the SSC only achieves 0.5637.

The proposed AMRMF algorithm arrives at the best ACCs for nine datasets (only failed on the dataset of Movement with less than 0.01 of ACC), and such results could be accurately reproduced, compared to the basic algorithms (k-means and NCuts, the results of which have large variations) and advanced algorithms (LRR, LSR, RSS and SSC, the results of which are relatively stable but actually the identical reproduction couldn't be guaranteed). Furthermore, the proposed AMRMF algorithm has obtained the best NMI on seven datasets (only failed on the datasets of Movement and PalmData25 with less than 0.01 and lower than RSS with 0.04 on the Caltech101 Silhouettes dataset. However, AMRMF leads its competitors around 0.1 for many times (e.g., the metric of ACC on the datasets of COIL20 and Ecoli), which confirms the superior performance of the proposed algorithm.

Finally, we would like to present the convergence performance of the proposed AMRMF optimization on the ten datasets, as illustrated in Figure 1. In each sub-figure, the transverse axis indicates the number of iterations from 1 to 100, while the longitudinal axis shows the error of objective function value in eq. (5) between iterations. It is clear that the AMRMF optimization often converges at stable values in less than ten iterations, which suggests that the proposed AMRMF optimization is very efficient in practice.

5 Conclusion

In this paper, we propose an adaptive manifold regularized matrix factorization (AMRMF) algorithm for joint learning the data affinity matrix and data clustering. The AMRMF is based on the idea of spectral clustering and low-rank matrix factorization. In order to overcome the point that the affinity matrix is always model based and may not optimal, i.e., computed by a predefined model with carefully tuned parameters combination, the proposed AMRMF aims to learn an affinity matrix jointly with the data clustering framework by considering it as an additional regularization. In this way, the learned affinity matrix could better guide the matrix factorization in a manifold point of view. Furthermore, the $\ell_{2,1}$-norm is applied to the matrix factorization to obtain the robust solution against the noises and outliers. Experimental results on numerous of datasets demonstrate the superior performance of the proposed method in accuracy and stability perspectives. For future work, the proposed method could be further extended to its more generalized version, which could deal with the out-of-sample problem and be employed for big data clustering [Cai and Chen, 2015].

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grants 61401317 and 61471274, the Australian Research Council Projects FT-130101457, DP-140102164, and LP-150100671, and the Hong Kong Scholars Program and Hong Kong Government General Research Fund Scheme under grant 152202/14E.

References

[Boyd and Vandenberghe, 2004] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, U.K., 2004.

[Cai and Chen, 2015] Deng Cai and Xinlei Chen. Large scale spectral clustering via landmark-based sparse representation. *IEEE Trans. Cybern.*, 45(8):1669–1680, 2015.

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

[Chung, 1997] Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, USA, 1997.

[Ding *et al.*, 2010] Chris Ding, Tao Li, and Michael I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(1):45–55, 2010.

[Elhamifar and Vidal, 2013] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.

[Gong *et al.*, 2015] Chen Gong, Tongliang Liu, Dacheng Tao, Keren Fu, Enmei Tu, and Jie Yang. Deformed graph laplacian for semisupervised learning. *IEEE Trans. Neural Netw. Learn. Syst.*, 26(10):2261–2274, 2015.

[Guo, 2015] Xiaojie Guo. Robust subspace segmentation by simultaneously learning data representations and their affinity matrix. In *Proc. IJCAI*, pages 3547–3553, 2015.

[Hartigan and Wong, 1979] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *J. R. Statist. Soc. C*, 28(1):100–108, 1979.

[He *et al.*, 2011] Xiaofei He, Deng Cai, Yuanlong Shao, Hujun Bao, and Jiawei Han. Laplacian regularized gaussian mixture model for data clustering. *IEEE Trans. Knowl. Data Eng.*, 23(9):1406–1418, 2011.

[Huang *et al.*, 2014] Jin Huang, Feiping Nie, Heng Huang, and Chris Ding. Robust manifold nonnegative matrix factorization. *ACM TKDD*, 8(3):11: 1–21, 2014.

[Huang *et al.*, 2015] Jin Huang, Feiping Nie, and Heng Huang. A new simplex sparse learning model to measure data similarity for clustering. In *Proc. IJCAI*, pages 3569–3575, 2015.

[Jain *et al.*, 1999] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[Liu and Tao, 2016] Tongliang Liu and Dacheng Tao. On the performance of manhattan nonnegative matrix factorization. *IEEE Trans. Neural Netw. Learn. Syst.*, 27(9):1851–1863, 2016.

[Liu *et al.*, 2013] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):171–184, 2013.

[Liu *et al.*, 2017] Hongfu Liu, Junjie Wu, Tongliang Liu, Dacheng Tao, and Yun Fu. Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence. *IEEE Trans. Knowl. Data Eng.*, 29:DOI 10.1109/TKDE.2017.2650229, 2017.

[Lu *et al.*, 2012] Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. In *Proc. ECCV*, pages 347–360, 2012.

[Meng and la Torre, 2013] Deyu Meng and Fernando De la Torre. Robust matrix factorization with unknown noise. In *Proc. ICCV*, pages 1337–1344, 2013.

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

[Nie *et al.*, 2016a] Feiping Nie, Xiaoqian Wang, Michael I. Jordan, and Heng Huang. The constrained laplacian rank algorithm for graph-based clustering. In *Proc. AAAI*, pages 1969–1976, 2016.

[Nie *et al.*, 2016b] Feiping Nie, Wei Zhu, and Xuelong Li. Unsupervised feature selection with structured graph optimization. In *Proc. AAAI*, pages 1302–1308, 2016.

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

[Trigeorgis *et al.*, 2014] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Bjorn W. Schuller. A deep semi-nmf model for learning hidden representations. In *Proc. ICML*, pages 1692–1700, 2014.

[Vidal, 2011] Rene Vidal. Subspace clustering. *IEEE Signal Process. Mag.*, 28(2):52–68, 2011.

[von Luxburg, 2007] Ulrike von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, 2007.

[Wang *et al.*, 2013] Can Wang, Zhong She, and Longbing Cao. Coupled clustering ensemble: Incorporating coupling relationships both between base clusterings and objects. In *Proc. ICDE*, pages 374–385, 2013.

[Wang *et al.*, 2015] Suhang Wang, Jiliang Tang, and Huan Liu. Embedded unsupervised feature selection. In *Proc. AAAI*, pages 470–476, 2015.

[Xu *et al.*, 2016] Jinglin Xu, Junwei Han, Kai Xiong, and Feiping Nie. Robust and sparse fuzzy k-means clustering. In *Proc. IJCAI*, pages 2224–2230, 2016.

[Xu *et al.*, 2017] Jinglin Xu, Junwei Han, Feiping Nie, and Xuelong Li. Re-weighted discriminatively embedded k-means for multi-view clustering. *IEEE Trans. Image Process.*, 26(6):3016–3027, 2017.

[Yang *et al.*, 2016] Yang Yang, Fumin Shen, Zi Huang, and Heng Tao Shen. A unified framework for discrete spectral clustering. In *Proc. IJCAI*, pages 2273–2279, 2016.

[Zhang *et al.*, 2012] Lijun Zhang, Chun Chen, Jiajun Bu, Zhengguang Chen, Deng Cai, and Jiawei Han. Locally discriminative coclustering. *IEEE Trans. Knowl. Data Eng.*, 24(6):1025–1035, 2012.

[Zhang *et al.*, 2015] Lefei Zhang, Qian Zhang, Liangpei Zhang, Dacheng Tao, Xin Huang, and Bo Du. Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding. *Pattern Recognit.*, 48(10):3102–3112, 2015.