Applied Intelligence (2019) 49:2151–2168
https://doi.org/10.1007/s10489-018-1380-2

Adaptive local learning regularized nonnegative matrix factorization for data clustering

Yongpan Sheng[1] · Meng Wang[2] · Tianxing Wu[3] · Han Xu[1]

Published online: 3 January 2019

Abstract

Data clustering aims to group the input data instances into certain clusters according to the high similarity to each other, and it could be regarded as a fundamental and essential immediate or intermediate task that appears in areas of machine learning, pattern recognition, and information retrieval. Clustering algorithms based on graph regularized extensions have accumulated much interest for a couple of decades, and the performance of this category of approaches is largely determined by the data similarity matrix, which is usually calculated by the predefined model with carefully tuned parameters combination. However, they may lack a more flexible ability and not be optimal in practice. In this paper, we consider both discriminative information as well as the data manifold in a matrix factorization point of view, and propose an adaptive local learning regularized nonnegative matrix factorization (ALLRNMF) approach for data clustering, which assumes that similar instance pairs with a smaller distance should have a larger probability to be assigned to the probabilistic neighbors. ALLRNMF simultaneously learns the data similarity matrix under the assumption and performs the nonnegative matrix factorization. The constraint of the similarity matrix encodes both the discriminative information as well as the learned adaptive local structure and benefits the data clustering on manifold. In order to solve the optimization problem of our approach, an effective alternative optimization algorithm is proposed such that our objective function could be decomposed into several subproblems that each has an optimal solution, and its convergence is theoretically guaranteed. Experiments on real-world benchmark datasets demonstrate the superior performance of our approach against the existing clustering approaches.

Keywords Adaptive local structure learning · Manifold regularization · Nonnegative matrix factorization · Data clustering

1 Introduction

The task of data clustering aims to partition the input data instances into certain clusters such that the instances together with a same group could desire the high similarities with each other. Concretely, clustering could be taken as a special case of classification problems without depending on any training data. The performance of the clustering approaches usually are heavily associated with exploring the local structure of the data, i.e., data similarity learning [25]. For a couple of decades, clustering algorithms have attracted much attention and a number of approaches have been proposed, e.g., k-means clustering [24], hierarchical clustering [19], information-theoretic clustering [10], spectral clustering [23], multi-view clustering [5, 28], and matrix factorization based on methods [7]. Nonnegative matrix factorization (NMF) could be regarded as a category of representative approach and could be used for data clustering by interpreting two decomposed matrices as the cluster indicator and latent feature matrices, respectively. Several studies show that the data instances in real world are often retrieved from a nonlinear low-dimensional manifold [1, 29, 33, 37], which corresponds to the embeddings in the high-dimensional feature space. However, NMF is based on the assumption that the data instances could be sampled from a Euclidean space so that it neglects to the intrinsic geometrical structure of the data. Therefore, various NMF and its graph regularized extensions are proposed [8, 22, 34, 38] to regularize the matrix factorization more fit for the

✉ Meng Wang
wangmengsd@outlook.com

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

[2] School of Computer Science and Engineering, Southeast University, Nanjing, China

[3] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

task of data clustering, e.g., graph-based regularization [4], the manifold regularization [16], the local learning regularization [12], and sparse learning [35]. Although many graph-based regularized algorithms could be superior to quite a bit of prior clustering methods [14, 31] and have been successfully applied in the task of data clustering, they are usually sensitive to the input similarity matrix and likely to mislead the matrix factorization due to the drawbacks that the nearest neighbors in the graph may belong to the different clusters. Apart from this, these approaches may still suffer from the following potential limitations and easily lead to unsatisfactory results: (1) Most of the above mentioned methods employ the fixed Laplacian matrix for data manifold regularization rather than a more effective similarity matrix to better regularize the model; (2) Recently, some advanced approaches have been proposed to explore some better similarity matrices as well as data representations, especially for constructing the affinity matrix based on predefined model (e.g., K-Nearest Neighbors model, Gaussian kernel function, the constrained Laplacian rank [26] and low-rank representation [21]) with carefully tuned parameters combination, they may not be optimal in practice.

To relieve the above issues, in this paper, we propose a novel adaptive local learning regularized nonnegative matrix factorization (ALLRNMF) approach for data clustering. ALLRNMF models the data similarity matrix depending on the local structure learning. It is based on the assumption that the data instances with a smaller distance should have a larger probability to be grouped in the same cluster. Furthermore, to achieve the ideal neighborhood's assignment, we constrain the data similarity matrix based on the above assumption that the neighborhood's assignment becomes an adaptive process, thus an ideal neighbor's assignment could be expected. The constraint of similarity matrix not only encodes the discriminative information and adaptive local structure, but also benefits the data clustering on manifold. For the optimization problem of our approach, an effective alternative optimization algorithm is proposed such that the objective function can be decomposed into several subproblems that each has an optimal solution, and its convergence is theoretically guaranteed. Experiments on real-world datasets commonly regarded as benchmarks show that our approach achieves improvements when comparing with different kinds of NMF and its variants as well as the advanced clustering methods, which verifies the effectiveness of our approach. We summarized the main contributions of this paper as follows:

- We propose a novel adaptive local learning regularized nonnegative matrix factorization (ALLRNMF) approach for data clustering. Moreover, we come up with an effective alternative optimization algorithm to solve the optimization problem related to our approach, which is easy-to-iterate and its convergence is also guaranteed theoretically.
- ALLRNMF has the capacity of jointly learning the data similarity matrix as well as performing matrix factorization. Therefore, the local structure of input data can be well uncovered and benefits the data clustering on manifold. Comparing to the conventional graph-based clustering approaches which learn the affinity matrix based on predefined model with carefully tuned parameters combination. The proposed ALLRNMF approach not only achieves the adaptive learning and the learned data similarity matrix could be better to guide the graph regularization to fit the task of data clustering, but also reveals a more flexible ability.
- We conduct extensive experiments on eight different real-world benchmark datasets in clustering literature. Experimental results demonstrate the superior performance of our approach against the existing clustering approaches.

The remainder of this paper is organized as follows. Section 2 introduces the previous research on NMF and relevant regularized extensions. Section 3 presents our ALLRNMF approach in detail. In Section 4, we describe details about the setup of experiments and report experimental results. Conclusion and future work are summarized at the end of the paper.

2 Related work

In this section, we briefly review the advances of the approaches for NMF as well as its graph regularized extensions related to our work.

Standard NMF tries to find a compressed approximation of the original nonnegative data matrix via two matrices. However, since NMF could not well employ the geometric structure of data on manifold, Cai et al. [4] proposed a graph-based regularized NMF (GRNMF) to explore the intrinsic geometrical structure of data, which is derived from the cluster assumption that adjacent data instances have high probability to be in the same cluster, Huang et al. [16] proposed a robust manifold NMF (RMNMF) method using $\ell_{2,1}$-norm (i.e., the manifold regularization), and [12] proposed a local learning regularized NMF (LLRNMF), which took into account that an additional constraint (i.e., the local learning regularization) is added to NMF to predict the cluster label of each data instance via its neighborhoods. It could overcome the shortages of NMF, e.g., the data instances that achieved sampling using a Euclidean space. This could not encode the discriminative information as well as limits data clustering lying on manifold, Wang et al. [35] directly embedded feature selection into data

clustering via sparse learning without the transformation. The formulations of these approaches, which are introduced as follows.

1. **Standard NMF** NMF (i.e., F-norm NMF) [20] aims to discovery two nonnegative matrices $\mathbf{U} \in \mathbb{R}_+^{m\times c}$ and $\mathbf{V} \in \mathbb{R}_+^{n\times c}$ of lower dimensionality, in which their product provides a good approximation to the original data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}_+^{m\times n}$. The optimization objective function can be written as follows:

$$\begin{aligned}\mathcal{O} &= arg\min_{\mathbf{U},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2 \\ &= arg\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\mathbf{X}_{ij} - (\mathbf{U}\mathbf{V}^T)_{ij}\right)^2 \\ &\quad \mathbf{s.t.} \quad \mathbf{U} \geq 0, \mathbf{V} \geq 0, \end{aligned} \tag{1}$$

where $||\cdot||_F$ is Frobenius norm (denoted as F-norm), e.g., the F-norm of matrix $\mathbf{X}$ can be defined as $||\mathbf{X}||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{X}_{ij}^2} = \sqrt{\sum_{i=1}^{n}||\mathbf{x}_i||_2^2}$. n is the number of data instances and m is the number of data features. It is easy to prove that the objective function in (1) is a convex function when either $\mathbf{U}$ or $\mathbf{V}$ remains, but it is not convex in both $\mathbf{U}$ and $\mathbf{V}$ together. Hence, it is unrealistic to search the global minimum solution. Lee and Seung [20] proposed an effective iterative updating algorithm with multipliers for optimizing the objective function as follows:

$$\begin{aligned}\mathbf{U}_{ij} &\leftarrow \mathbf{U}_{ij}\frac{(\mathbf{X}\mathbf{V})_{ij}}{(\mathbf{U}\mathbf{V}^T\mathbf{V})_{ij}}, \\ \mathbf{V}_{ij} &\leftarrow \mathbf{V}_{ij}\frac{(\mathbf{X}^T\mathbf{U})_{ij}}{(\mathbf{V}\mathbf{U}^T\mathbf{U})_{ij}}. \end{aligned} \tag{2}$$

In the clustering setting of NMF [36], the low-dimensional representation of j-th data instance in the clustering assignment matrix $\mathbf{V}$ could be denoted as $\mathbf{z}_i^T = [\mathrm{v}_{i1}, \mathrm{v}_{i2}, \cdots, \mathrm{v}_{ic}]^T \in \mathbb{R}_+^c$, where c (c $\ll$ m, c $\ll$ n) is the number of clusters.

2. **Graph Regularized NMF (GRNMF)** GRNMF approach incorporates the intrinsic geometrical structure of data manifold and minimize the following objective function:

$$\begin{aligned}\mathcal{O} &= arg\min_{\mathbf{U},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||^2 + \mu\sum_{i=1}^{k}\mathcal{R}_k \\ &= arg\min_{\mathbf{U},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||^2 + \mu\mathrm{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V}) \\ &\quad \mathbf{s.t.} \quad \mathbf{U} \geq 0, \mathbf{V} \geq 0, \end{aligned} \tag{3}$$

where $\mu \geq 0$ is the regularizaion parameter, and $\mathbf{R}_k$ in the second term is used to measure the smoothness of the geodesics in the intrinsic geometry of the data.

3. **Robust NMF (RNMF)** RNMF approach employs the $l_{2,1}$-norm to measure the loss of matrix factorization, and the improved optimization objective function can be written as follows:

$$\begin{aligned}\mathcal{O} &= arg\min_{\mathbf{U},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_{2,1} \\ &\quad \mathbf{s.t.} \quad \mathbf{U} \geq 0, \mathbf{V} \geq 0. \end{aligned} \tag{4}$$

However, the model in (4) may lead to unsatisfactory results since there is no constraint on the latent feature matrix $\mathbf{U}$ and cluster assignment matrix $\mathbf{V}$.

4. **Robust Manifold NMF (RMNMF)** The optimization objective function of RMNMF can be written as follows:

$$\begin{aligned}\mathcal{O} &= arg\min_{\mathbf{U},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_{2,1} + \mu\mathrm{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V}) \\ &\quad \mathbf{s.t.} \quad \mathbf{U} \geq 0, \mathbf{V} \geq 0, \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{5}$$

where $\mu \geq 0$ is a regularization parameter, and the second term is the additional constraint $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. Here, it not only guarantees the uniqueness of the model, but also has a significant purpose for reducing the computation cost for the optimization algorithm.

5. **Local Learning Regularized NMF (LLRNMF)** The optimization objective function of LLRNMF can be written as follows:

$$\begin{aligned}\mathcal{O} &= arg\min_{\mathbf{U},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2 + \mu\mathrm{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V}) \\ &\quad \mathbf{s.t.} \quad \mathbf{U} \geq 0, \mathbf{V} \geq 0, \end{aligned} \tag{6}$$

where $\mu \geq 0$ is a positive parameter that is used for controlling the contribution of the second term. Letting $\mu = 0$, Eq. 6 degenerates to the standard NMF.

Besides, on the basis of GRNMF, Shang et al. [30] proposed a graph dual regularized NMF (GDRNMF) which aimed to explore the intrinsic geometrical structure by considering the data manifold and feature manifold simultaneously. Later, once again the extension version of GDRNMF is presented for further improving the performance of GDRNMF approach. In order to better depict the similarity matrices as well as data representations, Zhang et al. [39] proposed a framework that could learn the affinity matrix simultaneously to regularized matrix factorization, which is called adaptive manifold regularized matrix factorization (AMRMF), Peng et al. [27] proposed a robust graph regularized NMF approach (RGNMF), which could capture the corrupted data as well as in favor of alleviating the impact of noise and outliers due to the spare corruption. Although these graph-based clustering algorithms have been successfully applied in the task of data clustering. However, they are usually sensitive to the input similarity matrix and may mislead the matrix factorization due to the drawbacks that the nearest neighbors in the fixed graph may belong to the

different clusters. Moreover, they still have certain space to achieve improvement in terms of the performance and robustness of data clustering task. In summary, the representative formulations of NMF and its graph regularized variants are summarized in Table 1.

3 The proposed approach

In this section, we first present the significant notations used throughout this paper in Table 2. After that, the learning steps of our supposed ALLRNMF approach are introduced in detail. For the optimization problem of our approach, an effective alternative optimization algorithm is proposed. We decompose the problem into several subproblems that each has an optimal solution.

3.1 Adaptive local structure learning

The clustering results highly depend on the local structure learning in most of the cases since the data cluster could be grouped based on the data similarity learning [25] (i.e., similarity matrix of input data). However, in the existing works, the affinity matrix is usually computed by the predefined model with a careful parameter tuning combination, which may result in unsatisfactory clustering performance in practice. Hence, we consider to learn the data similarity matrix by exploring the local connectivity of given input data, i.e., assigning the adaptive and optimal neighborhood for each data instance. In the following, we will introduce how to achieve the adaptive neighborhood assignment.

Suppose there is an original input nonnegative matrix is represented as $\mathbf{X} \in \mathbb{R}_{+}^{m\times n}$, in which the i-th column could be regarded as an instance denoted by $\mathbf{x}_i \in \mathbb{R}^m$, we suppose that each instance $\mathbf{x}_i$ could be connected to any other instance (e.g., $[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$ in the dataset) with the probability s_{ij}, where s_{ij} is an element of the expected similarity matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$. $\mathbf{S}$ can be obtained by solving the following problem:

$$\min_{\mathbf{s}_i^T\mathbf{1}=1, 0\le s_{ij}\le 1} \sum_{j=1}^{n} s_{ij}^2. \tag{7}$$

It is clear that all the data instances can be defined as the neighborhood of $\mathbf{x}_i$ with the same probability $\frac{1}{n}$. However, our assumption could not meet the actual circumstances. Only the nearest data instances could be defined as the neighborhoods of $\mathbf{x}_i$ with the sum of probability to 1 instead of any other data instance. We suggest that the similar instance pairs with a smaller distance (e.g., Euclidean distance) should be assigned to the probabilistic neighbors, i.e., the smaller the distance measure $||\mathbf{x}_i - \mathbf{x}_j||_2^2$ is, the larger probability s_{ij} will be. Therefore, the objective function can be measured by solving the problem as follows:

$$\min_{\mathbf{s}_i^T\mathbf{1}=1, 0\le s_{ij}\le 1} \sum_{j=1}^{n} ||\mathbf{x}_i - \mathbf{x}_j||_2^2 s_{ij}, \tag{8}$$

where $\mathbf{s}_i \in \mathbb{R}^n$ is a column vector with of $\mathbf{S}$. Note that the probability $s_{ij} \in \mathbf{S}$ in (8) should be constrained such that the neighborhood assignment becomes an adaptive process to assign the optimal neighbors for clustering task.

Combining (7) and (8), the optimization problem can be solved by minimizing the objective as follows:

$$\min_{\mathbf{s}_i^T\mathbf{1}=1, 0\le s_{ij}\le 1} \sum_{j=1}^{n} \left(||\mathbf{x}_i - \mathbf{x}_j||_2^2 s_{ij} + \gamma s_{ij}^2 \right), \tag{9}$$

where $\gamma \ge 0$ is a positive parameter controlling the contribution of the second term in (9). We can assign the neighborhood for data instance $\mathbf{x}_i$ according to (9). Considering for all the data instances in the data set, we have

$$\begin{aligned} &arg\min_{\mathbf{S}} \sum_{i,j=1}^{n} \left(||\mathbf{x}_i - \mathbf{x}_j||_2^2 s_{ij} + \gamma s_{ij}^2 \right) \\ &\textbf{s.t.} \quad \forall i, \Sigma_j s_{ij} = 1, 0 \le s_{ij} \le 1. \end{aligned} \tag{10}$$

Table 1 NMF approach and its variants

Algorithms	Formulations
NMF [20]	$arg\ \min_{\mathbf{U},\mathbf{V}\ge 0} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||^2$
Graph Regularized NMF (GRNMF) [4]	$arg\ \min_{\mathbf{U},\mathbf{V}\ge 0} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||^2 + \mu \mathrm{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V})$
Local Learning Regularized NMF (LLRNMF) [12]	$arg\ \min_{\mathbf{U},\mathbf{V}\ge 0} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2 + \mu \mathrm{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V})$
Convex NMF (Ding et al., 2010)	$arg\ \min_{\mathbf{W},\mathbf{V}\ge 0} ||\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{V}^T||^2$
Robust NMF (Kong et al., 2011)	$arg\ \min_{\mathbf{U},\mathbf{V}\ge 0} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_{2,1}$
Dual Regularized NMF (DRNMF) [30]	$arg\ \min_{\mathbf{U},\mathbf{V}\ge 0} || + \lambda J(\mathbf{V}) + \mu J(\mathbf{U})$
Robust Manifold NMF [16]	$arg\ \min_{\mathbf{U},\mathbf{V}\ge 0} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_{2,1} + \mu \mathrm{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V})$
Adaptive Manifold Regularized Matrix Factorization [39]	$arg\ \min_{\mathbf{U},\mathbf{V}\ge 0} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_{2,1} + 2\gamma Tr(\mathbf{U}^T\mathbf{L}\mathbf{U}) + \beta \left(\sum_{i,j} ||\mathbf{x}_i - \mathbf{x}_j||_2^2 s_{ij} + \alpha ||\mathbf{S}||_F^2 \right)$
Robust Graph Regularized NMF (RGRNMF) [17]	$arg\ \min_{\mathbf{U},\mathbf{V}\ge 0} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T - \mathbf{S}||_F^2 + \alpha ||\mathbf{S}||_1 + \beta J_1(\mathbf{V}) + \lambda J_2(\mathbf{U})$

Table 2 Important Notations Annotated in This Paper

Notations	Description
X	Input Nonnegative data matrix of size m × n
n	The number of data instances (or denoted as data points)
m	The number of data features (or denoted as data dimension)
c	The number of data clusters
k	The number of nearest data instances (or denoted as the neighborhood size)
x_i	i-th column vector of X (i.e, i-th data instance)
X_{ij}	(i, j)-th element of $\mathbf{X}$
U	The latent feature matrix of m × c (or denoted as cluster centroid)
u_i	i-th row of $\mathbf{U}$
V	The cluster assignment matrix of n × c (or denoted as cluster indicator)
v_i	i-th row of $\mathbf{V}$
S	Similarity matrix of data instances (or denoted as affinity matrix)
s_i	i-th row of $\mathbf{S}$
s_{ij}	j-th element of $\mathbf{s}_i$
L_S	Laplacian matrix
W_S	Symmetric similarity matrix
D_S	Graph degree matrix
d_{ij}^x	Data partition matrix of Euclidean distance of data instances in the form of ℓ_2-norm
d_{ij}^v	Data partition matrix of cluster assignment in the form of ℓ_2-norm
d_{ij}	The (i, j)-th element of the matrix that is composed of d_{ij}^x and d_{ij}^v
$Tr(X)$	The trace of $\mathbf{X}$
1	The column vector with the whole elements equals to 1

We constrain the data similarity matrix $\mathbf{S}$ such that the neighborhood assignment becomes an adaptive process. Furthermore, an ideal neighbors assignment can be expected. From another point of view, the similarity matrix $\mathbf{S}$ obtained in the neighbors assignment can be seen as a sparse data graph. As we mentioned before, $\mathbf{z}_i^T = [\mathrm{v}_{i1}, \mathrm{v}_{i2}, \cdots, \mathrm{v}_{ic}]^T \in \mathbb{R}_+^c$ denoted as the i-th row of $\mathbf{V}$, which could be regarded as the low-dimensional representation of j-th data instance. We can measure the smoothness of the low-dimensional representation:

$$\begin{aligned}\Re &= \frac{1}{2}\sum_{i,j=1}^{n}||\mathbf{v}_i - \mathbf{v}_j||_2^2 s_{ij} \\ &= \sum_{i=1}^{n}\mathbf{z}_i^T\mathbf{z}_i(\mathbf{D}_S)_{ii} - \sum_{i,j=1}^{n}\mathbf{z}_j^T\mathbf{z}_j(\mathbf{W}_S)_{ij} \\ &= \mathrm{Tr}(\mathbf{V}^T\mathbf{D}_S\mathbf{V}) - \mathrm{Tr}(\mathbf{V}^T\mathbf{W}_S\mathbf{V}) = \mathrm{Tr}(\mathbf{V}^T\mathbf{L}_S\mathbf{V}),\end{aligned} \tag{11}$$

where $\mathbf{L}_S = \mathbf{D}_S - \mathbf{W}_S$ is called Laplacian matrix computed by the learned $\mathbf{S}$. According to the definition in graph theory [6], $\mathbf{W}_S$ is a symmetric similarity matrix and $\mathbf{W}_S = \frac{\mathbf{S}+\mathbf{S}^T}{2}$, the degree matrix $\mathbf{D}_S \in \mathbb{R}^{n\times n}$ is defined as a diagonal matrix where the i-th diagonal element is $(\mathbf{D}_S)_{ii} = \sum_j(\mathbf{W}_S)_{ij}$. Equation 11 can be called as *local learning regularization*. The better predicted cluster label of each data instance by its neighborhood, the smaller the local learning regularizer will be.

Finally, by combining the (1), Eq. 10 and (11) together with the additional positive regularization parameters μ and λ, we proposed the objective function of ALLRNMF approach as follows:

$$\begin{aligned}\mathcal{O} = arg\min_{\mathbf{S},\mathbf{U},\mathbf{V}} &\mu\sum_{i,j=1}^{n}\left(||\mathbf{x}_i - \mathbf{x}_j||_2^2 s_{ij} + \gamma s_{ij}^2\right) \\ &+\lambda\mathrm{Tr}(\mathbf{V}^T\mathbf{L}_S\mathbf{V}) + ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2 \\ &\textbf{s.t.} \quad \forall i, \Sigma_j s_{ij} = 1, 0 \le s_{ij} \le 1, \mathbf{U} \\ &\ge 0, \mathbf{V} \ge 0,\end{aligned} \tag{12}$$

where $\mu \ge 0$ and $\lambda \ge 0$ are positive regularization parameters balancing the smoothness of data instances and the outliers in the data space during adaptive local structure learning in the first and second terms. We call our approach in (12) as an *adaptive local learning regularized nonnegative matrix factorization* (ALLRNMF). Besides, in order better to estimate a lower bound of the objective function as well as the relieve scale transfer problem [11], we consider L_2 normalization (denoted as ℓ_2-norm) on $\mathbf{S}$, $\mathbf{U}$ and $\mathbf{V}$ for iteratively optimization.

3.2 Alternative optimization algorithm

Since the objective function of ALLRNMF in (12) is not convex and carries three variables and additional multipliers, we decompose the problem into a few subproblems in the form of minimizing solutions, in each of which minimizes the objective function by updating one variable while fixing the other variables. Essentially, this kind of alternative solving method is easy-to-iterate and its convergence is also guaranteed theoretically.

Updating S with fixed U and V Optimizing (12) with respect to $\mathbf{S}$, we fixed other variables expect $\mathbf{S}$ and dropped the terms that are independent of $\mathbf{S}$. Then (12) can be reduced as follows:

$$\begin{aligned} arg \min_{\mathbf{S}} & \sum_{i,j=1}^{n} \left(||\mathbf{x}_i - \mathbf{x}_j||_2^2 s_{ij} + \gamma s_{ij}^2 \right) + \bar{\lambda} \mathrm{Tr}(\mathbf{V}^T \mathbf{L}_S \mathbf{V}) \\ \textbf{s.t.} \quad & \forall i, \Sigma_j s_{ij} = 1, 0 \le s_{ij} \le 1, \end{aligned} \tag{13}$$

where $\bar{\lambda} = \frac{\lambda}{\mu}$. According to (11), the problem (13) can be rewritten in the following form:

$$\begin{aligned} arg \min_{\mathbf{S}} & \sum_{i,j=1}^{n} \left(||\mathbf{x}_i - \mathbf{x}_j||_2^2 s_{ij} + \gamma s_{ij}^2 + \frac{1}{2}\bar{\lambda} ||\mathbf{v}_i - \mathbf{v}_j||_2^2 s_{ij} \right) \\ \textbf{s.t.} \quad & \forall i, \Sigma_j s_{ij} = 1, 0 \le s_{ij} \le 1. \end{aligned} \tag{14}$$

Since problem (14) is independent between different i, the problem can be solved individually for each i:

$$\begin{aligned} arg \min_{\mathbf{s}_i} & \sum_{j=1}^{n} \left(||\mathbf{x}_i - \mathbf{x}_j||_2^2 s_{ij} + \gamma s_{ij}^2 + \frac{1}{2}\bar{\lambda} ||\mathbf{v}_i - \mathbf{v}_j||_2^2 s_{ij} \right) \\ \textbf{s.t.} \quad & \forall i, \Sigma_j s_{ij} = 1, 0 \le s_{ij} \le 1. \end{aligned} \tag{15}$$

Let $d_{ij}^x = ||\mathbf{x}_i - \mathbf{x}_j||_2^2$, $d_{ij}^{\mathrm{v}} = \frac{1}{2}||\mathbf{v}_i - \mathbf{v}_j||_2^2$, and denote $\mathbf{d}_i \in \mathbb{R}^n$ as a vector with the j-th element as $d_{ij} = d_{ij}^x + \bar{\lambda} d_{ij}^{\mathrm{v}}$, then (15) can be further reduced as follows:

$$\begin{aligned} & arg \min_{\mathbf{s}_i} \|\mathbf{s}_i + \frac{1}{2\gamma}\mathbf{d}_i\|_2^2 \\ & \textbf{s.t.} \quad \forall i, \Sigma_j s_{ij} = 1, 0 \le s_{ij} \le 1. \end{aligned} \tag{16}$$

In the next Section 3.3, we will prove that (16) can be solved with a closed form solution.

Updating U with fixed S and V Optimizing (12) with respect to $\mathbf{U}$, we fixed other variables expect $\mathbf{U}$ and dropped the terms that are independent of $\mathbf{U}$. Then (12) can be reduced as follows:

$$\begin{aligned} arg \min_{\mathbf{U}} & ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2 \\ \textbf{s.t.} \quad & \mathbf{U} \ge 0. \end{aligned} \tag{17}$$

Note that (17) has the similar form in comparison with the Standard NMF in (1). Therefore, we can update $\mathbf{U}$ according to the update rule in (2), denote:

$$\mathbf{U}_{ij} \leftarrow \mathbf{U}_{ij} \sqrt{\frac{(\mathbf{X}\mathbf{V})_{ij}}{(\mathbf{U}\mathbf{V}^T\mathbf{V})_{ij}}}. \tag{18}$$

Updating V with fixed S and U Optimizing (12) with respect to $\mathbf{V}$, we fixed other variables expect $\mathbf{V}$ and dropped the terms that are independent of $\mathbf{V}$. Then (12) can be reduced as follows:

$$\begin{aligned} J_{ALLRNMF} = arg \min_{\mathbf{V}} & ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2 + \lambda \mathrm{Tr}(\mathbf{V}^T \mathbf{L}_S \mathbf{V}) \\ \textbf{s.t.} \quad & \mathbf{V} \ge 0. \end{aligned} \tag{19}$$

With a similar procedure of the computation of $\mathbf{U}$, let $\boldsymbol{\beta} \in \mathbb{R}^{n \times c}$ be the Lagrangian multiplier for constraint $\mathbf{V} \ge 0$, and the Lagrangian function can be defined as:

$$L(\mathbf{V}) = ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2 + \lambda \mathrm{Tr}(\mathbf{V}^T \mathbf{L}_S \mathbf{V}) - \mathrm{Tr}(\beta \mathbf{V}^T). \tag{20}$$

Setting $\frac{\partial L(\mathbf{V})}{\partial \mathbf{V}} = 0$, we obtain

$$\boldsymbol{\beta} = -2\mathbf{X}^T\mathbf{U} + 2\mathbf{V}\mathbf{U}^T\mathbf{U} + 2\lambda \mathbf{L}_S \mathbf{V}. \tag{21}$$

According to the Karush-Kuhn-Tucker (KKT) condition [2], $\boldsymbol{\beta}_{ij}\boldsymbol{V}_{ij} = 0$, then we have

$$(-\mathbf{X}^T\mathbf{U} + \mathbf{V}\mathbf{U}^T\mathbf{U} + \lambda \mathbf{L}_S \mathbf{V})_{ij} \mathbf{V}_{ij} = 0. \tag{22}$$

Introduce

$$\mathbf{L}_S = \mathbf{L}_S^+ - \mathbf{L}_S^-, \tag{23}$$

where $(\mathbf{L}_S^+)_{ij} = (|(\mathbf{L}_S)_{ij}| + (\mathbf{L}_S)_{ij})/2$ and $(\mathbf{L}_S^-)_{ij} = (|(\mathbf{L}_S)_{ij}| - (\mathbf{L}_S)_{ij})/2$.

Substitute (23) back into (22), we obtain

$$(\lambda(\mathbf{L}_S^+ - \mathbf{L}_S^-)\mathbf{V} - \mathbf{X}^T\mathbf{U} + \mathbf{V}\mathbf{U}^T\mathbf{U})_{ij}\mathbf{V}_{ij} = 0. \tag{24}$$

With a little algebra, Eq. 24 leads to the following update rule:

$$\mathbf{V}_{ij} \leftarrow \mathbf{V}_{ij} \sqrt{\frac{\mathbf{X}^T\mathbf{U} + \lambda \mathbf{L}_S^- \mathbf{V}}{\mathbf{V}\mathbf{U}^T\mathbf{U} + \lambda \mathbf{L}_S^+ \mathbf{V}}}. \tag{25}$$

3.3 The discussion of parameter γ

In this subsection, an effective approach is presented to determine the most appropriate γ in our objective function in (15). In terms of individual i, the search of parameter γ in the problem (15) is equivalent to the same one in the problem (16). Letting both ζ and $\eta \ge 0$ are the Lagrangian

multipliers, the problem (16) can be rewritten into the following equivalent problem:

$$L(\boldsymbol{s}_i, \eta, \boldsymbol{\zeta}_i) = \frac{1}{2}\|\mathbf{s}_i + \frac{1}{2\boldsymbol{\gamma}_i}\mathbf{d}_i^x\|_2^2 - \eta(\mathbf{s}_i^T\mathbf{1} - 1) - \boldsymbol{\zeta}_i^T\mathbf{s}_i. \quad (26)$$

According to the KKT condition [2], the optimal solution of $\mathbf{s}_i$ can be defined as follows:

$$s_{ij} = \left(-\frac{1}{2\gamma}d_{ij}^x + \eta\right)_+. \quad (27)$$

In general, a better performance of data clustering can be achieved by considering the locality and connectivity of given input data. That is, it is preferred to selecting the k-Nearest of $\mathbf{x}_j (0 \le j \le n)$ that could be connected to $\mathbf{x}_i$ as neighborhoods. Meanwhile, the computation burden could be largely alleviated for the subsequent processing.

Without the loss of generality, we suppose that $d_{i1}^x, d_{i2}^x, \cdots, d_{in}^x$ are ordered from small to large. If the optimal $\mathbf{s}_i$ just has k nonzero elements. According to (27), then it is easy to see that $s_{ik} > 0$ and $s_{i,k+1} = 0$. Thus we obtain

$$s_{ij} = \begin{cases} -\frac{1}{2\gamma}d_{ik}^x + \eta > 0, \\ -\frac{1}{2\gamma}d_{i,k+1}^x + \eta \le 0. \end{cases} \quad (28)$$

Using (27) and the constraint $\Sigma_j s_{ij} = 1$, we have

$$\sum_{j=1}^{k}\left(-\frac{1}{2\gamma_i}d_{ij}^x + \eta\right) = 1 \Rightarrow \eta = \frac{1}{k} + \frac{1}{2k\gamma_i}\sum_{j=1}^{k}d_{ij}^x, \quad (29)$$

according to (28) and (29), we have the inequality relation as follows:

$$\frac{k}{2}d_{ik}^x - \frac{1}{2}\sum_{j=1}^{k}d_{ij}^x < \gamma_i \le \frac{k}{2}d_{i,k+1}^x - \frac{1}{2}\sum_{j=1}^{k}d_{ij}^x, \quad (30)$$

Equation 30 could achieve an ideal and optimal neighborhood assignment solution by adding the constraint of k nonzero values as mentioned in (28), then γ_i could be set

$$\gamma_i = \frac{k}{2}d_{i,k+1}^x - \frac{1}{2}\sum_{j=1}^{k}d_{ij}^x. \quad (31)$$

The overall value of γ could be set to the mean of $\gamma_1, \gamma_2, \cdots, \gamma_n$, denote as:

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{k}{2}d_{i,k+1}^x - \frac{1}{2}\sum_{j=1}^{k}d_{ij}^x\right), \quad (32)$$

since parameter k is an integer with the explicit meaning, it is easier to tune than parameter γ.

To this end, we summarize the proposed algorithm in Algorithm 1.

Algorithm 1 Adaptive local learning regularized nonnegative matrix factorization (ALLRNMF) approach

Require: Input nonnegative data matrix $\mathbf{X}$, initial values of the latent matrix $\mathbf{U}$, the cluster assignment matrix $\mathbf{V}$, the number of data clusters c, the number of nearest data instances k, regularization parameters λ, μ;

Ensure: The cluster assignment matrix $\mathbf{V}$;

Initialize similarity matrix $\mathbf{S}$ by the optimal solution to the problem (10);

repeat

1. Update $\mathbf{U}_{ij} \leftarrow \mathbf{U}_{ij}\sqrt{\frac{(\mathbf{XV})_{ij}}{(\mathbf{UV}^T\mathbf{V})_{ij}}}$;
2. Update $\mathbf{V}_{ij} \leftarrow \mathbf{V}_{ij}\sqrt{\frac{\mathbf{X}^T\mathbf{U}+\lambda\mathbf{L}_S^-\mathbf{V}}{\mathbf{VU}^T\mathbf{U}+\lambda\mathbf{L}_S^+\mathbf{V}}}$;
3. For each i, update the i-th row of $\mathbf{S}$ by solving the problem (16), where $\mathbf{d}_i \in \mathbb{R}^n$ with the j-th element denoted as $d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||_2^2 + \frac{\lambda}{2\mu}||\mathbf{v}_i - \mathbf{v}_j||_2^2$;

until Converges

3.4 Convergence analysis

In this subsection, the convergence of the alternative optimization algorithm is investigated. It is clear that (16) can be solved with a closed form solution which is introduced in detail in Section 3.3, thus we only need to prove Theorem 1 discussed here:

Theorem 1 *The objective function in* (12) *is nonincreasing under the update rules in* (18) *and* (25).

The auxiliary function is used in the expectation-maximization algorithm [20] to prove the convergence of objective function. The definition of the auxiliary function is given by Definition 1.

Definition 1 $g(h, h')$ can be defined as an auxiliary function for $f(h)$ if the conditions

$$g(h, h') \ge f(h), g(h, h) = f(h)$$

are satisfied.

Lemma 1 *If $g(h, h^t)$ is an auxiliary function for $f(h)$, then $f(h)$ is nonincreasing under the update formula as follows:*

$$h^{t+1} = \arg\min_h g(h, h^t),$$

where t is the number of iterations.

Proof $f(h^{t+1}) \le g(h^{t+1}, h^t) \le g(h^t, h^t) = f(h^t)$ □

Thus, by iterating the update in Lemma 1, we obtain a sequence of estimates that converge to a local minimum $h_{min} = argmin_h f(h)$ of the objective function $f(h)$.

Lemma 2 *For any nonnegative matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{k\times k}$, $\boldsymbol{S} \in \mathbb{R}^{n\times k}$, $\boldsymbol{S}^{'} \in \mathbb{R}^{n\times k}$, and $\boldsymbol{A}$, $\boldsymbol{B}$ are symmetric, then the following inequality holds*

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\frac{(\boldsymbol{AS}^{'}\boldsymbol{B})_{ij}\boldsymbol{S}_{ij}^{2}}{\boldsymbol{S}_{ij}^{'}}\right) \geq tr(\boldsymbol{S}^{T}\boldsymbol{ASB}).$$

Theorem 2 *Let*

$$J(\boldsymbol{U}) = tr(-2\boldsymbol{X}^{T}\boldsymbol{UV} + \boldsymbol{V}^{T}\boldsymbol{U}^{T}\boldsymbol{UV}), \tag{33}$$

Then the following function

$$g(\boldsymbol{U},\boldsymbol{U}') = -2\sum_{ij}(\boldsymbol{XV}^{T})_{ij}\boldsymbol{U}'_{ij}\left(1+log\frac{\boldsymbol{U}_{ij}}{\boldsymbol{U}'_{ij}}\right) + \sum_{ij}\frac{(\boldsymbol{U}'\boldsymbol{VV}^{T})_{ij}\boldsymbol{U}_{ij}^{2}}{\boldsymbol{U}'_{ij}}$$

is an auxiliary function for $J(\boldsymbol{U})$. Furthermore, it is a convex function in $\boldsymbol{U}$ and its global minimum is

$$\boldsymbol{U}_{ij} = \boldsymbol{U}_{ij}\sqrt{\frac{(\boldsymbol{XV})_{ij}}{(\boldsymbol{UV}^{T}\boldsymbol{V})_{ij}}}. \tag{34}$$

Proof See Appendix A □

Theorem 3 *Updating $\boldsymbol{U}$ using (18) will monotonically decrease the value of the objective function in (12), hence it converges.*

Proof By Lemma 1 and Theorem 2, we can get that $J(\mathbf{U}^0) = g(\mathbf{U}^0,\mathbf{U}^0) \geq g(\mathbf{U}^1,\mathbf{U}^0) \geq J(\mathbf{U}^1) \geq \ldots$ So $J(\mathbf{U})$ is monotonically decreasing. Since $J(\mathbf{U})$ is obviously bounded below, we prove this theorem. □

Theorem 4 *Let*

$$J(\boldsymbol{V}) = tr(-2\boldsymbol{X}^{T}\boldsymbol{UV} + \boldsymbol{V}^{T}\boldsymbol{U}^{T}\boldsymbol{UV} - \lambda\boldsymbol{VL}_{S}\boldsymbol{V}^{T}), \tag{35}$$

Then the following function

$$g(\boldsymbol{V},\boldsymbol{V}') = \sum_{ij}\frac{(\boldsymbol{U}^{T}\boldsymbol{UV}')_{ij}\boldsymbol{V}_{ij}^{2}}{\boldsymbol{V}'_{ij}} + \lambda\sum_{ij}\frac{\left(\boldsymbol{V}'\boldsymbol{L}_{S}^{-}\right)_{ij}\boldsymbol{V}_{ij}^{2}}{\boldsymbol{V}'_{ij}} - \sum_{ij}(\boldsymbol{U}^{T}\boldsymbol{X})_{ij}\boldsymbol{V}'_{ij}\left(1+log\frac{\boldsymbol{V}_{ij}}{\boldsymbol{V}'_{ij}}\right) - \lambda\sum_{ijk}\left(\boldsymbol{L}_{S}^{+}\right)_{jk}\boldsymbol{V}'_{ij}\boldsymbol{V}'_{ik}\left(1+log\frac{\boldsymbol{V}_{ij}\boldsymbol{V}_{ik}}{\boldsymbol{V}'_{ij}\boldsymbol{V}'_{ik}}\right)$$

is an auxiliary function for $J(\boldsymbol{V})$. Furthermore, it is a convex function in $\boldsymbol{V}$ and its global minimum is

$$\boldsymbol{V}_{ij} = \boldsymbol{V}_{ij}\sqrt{\frac{\boldsymbol{X}^{T}\boldsymbol{U} + \lambda\boldsymbol{L}_{S}^{-}\boldsymbol{V}}{\boldsymbol{VU}^{T}\boldsymbol{U} + \lambda\boldsymbol{L}_{S}^{+}\boldsymbol{V}}}. \tag{36}$$

Proof See Appendix B □

Theorem 5 *Updating $\boldsymbol{V}$ using (25) will monotonically decrease the value of the objective function in (12), hence it converges.*

Proof By Lemma 1 and Theorem 4, we can get that $J(\mathbf{V}^0) = g(\mathbf{V}^0,\mathbf{V}^0) \geq g(\mathbf{V}^1,\mathbf{V}^0) \geq J(\mathbf{V}^1) \geq \ldots$ So $J(\mathbf{V})$ is monotonically decreasing. Since $J(\mathbf{V})$ is obviously bounded below, we prove this theorem. □

4 Experiment

In this section, we conduct experiments to evaluate our approach. The detailed experimental steps and results are reported from Section 4.1 to Section 4.5. We further study the performance on parameter tuning in Section 4.6.

4.1 Datasets

The experiments are performed on eight different real-world benchmark datasets in clustering literature. Table 3 summarizes the statistic characteristics of the datasets as follows:

– Congressional Voting Records (denoted as Vote) are derived from vote collections for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. Abalone dataset corresponds to

Table 3 The description of datasets used for our experiments

Dataset	Number of samples	Number of samples	Number of samples
Vote	435	16	2
TDT2-20	1938	1000	20
Wap	1560	1240	20
Abalone	2282	8	4
Krvs	3196	36	2
La1	3204	1944	6
tr12	313	5186	8
Diag-Bcw	569	30	2
Digit	1797	64	10
Caltech101 Silhouettes	8461	256	101

the physical measurements of abalones was collected from Marine Resources Division. Chess (King-Rook vs. King-Pawn) (denoted as Krvs) dataset was supplied to Holte by Peter Clark of the Turing Institute in Glasgow.

- Wap, La1 and tr12 are textual datasets, and the first one is from the WebACE project [15]. The remain two are derived from LA Times (TREC) which was in TREC-5. These document datasets concern computer, medical, health, research, etc. The number of terms is more than 710 for each document.
- TDT2-20 dataset is the subset of NIST Topic Detection and Tracking (TDT2) corpus, which mixed news, radio programs as well as television programs drawn from six sources (i.e., APW, NYT, VOA, PRI, CNN and ABC).
- Breast Cancer Wisconsin (denoted as Diag-Bcw) dataset, whereas the features are computed from a digitized image of a fine needle aspirate of a breast mass. The hand-written digits image dataset is also used in our experiment, it contains 1797 images cover different hand-written number 0-9 and all of images are resized to 8×8 pixels. Caltech101 Silhouettes dataset is a large dataset based on the Caltech 101 image annotations, where each image includes a high-quality polygon outline of the primary object in the scene, and rendered as 28×28 images as well as 16×16 pixels.

In total, the number of classes in our experimental dataset ranges from 2 to 101, the number of features ranges from 8 to 5186, and the number of samples away from 313 to 8461. These datasets are = widely available to evaluate clustering tasks. Wap and La1 datasets, which could be downloaded in the toolkit.[1] Vote, Diag-Bcw, Abalone, Krvs and Caltech101 Silhouettes can be achived from UCI machine learning repository.[2] tr12 can be derived from TREC collections,[3] and TDT2-20 can be obtained from TDT2 Corpus.[4]

4.2 Baseline methods

To evaluate the performance of the proposed ALLRNMF approach on the benchmark datasets, ALLRNMF is compared with the basic clustering algorithms and state-of-the-art approaches, including k-means, NMF [20], PCA [32], GRNMF [3], GDRNMF [30], LLRNMF [12], LPFNMTF [34], RSS [13], SSC [9], AMRMF [39], ALSLDC [18] and RGRNMF [17], to demonstrate its effectiveness.

We compare ALLRNMF with the following approaches:

(1) k-means. This is a simple iterative method to partition the given dataset into a user-specified number of clusters, k.
(2) NMF [20]. This is a low-rank matrix approximation method for finding two low-rank non-negative matrices whose product provides a good approximation to the original non-negative matrix.
(3) PCA [32]. This aims to combine k-means with principal components analysis (PCA) for data clustering. PCA transform is used for data compression, i.e., by reducing the number of dimensions, without much loss of information.
(4) GRNMF [3]. This aims to explore the intrinsic geometrical structure of data, which is derived from the cluster assumption that adjacent data instances have high probability to be in the same cluster.
(5) GDRNMF [30]. This is a graph-based clustering approach that lies on a nonlinear low dimensional manifold. It simultaneously considers the geometric structures of both the data manifold and the feature manifold. Moreover, two iterative updating optimization schemes for non-negative matrix factorization and tri-factorization are proposed, respectively, and provide the convergence proofs of these two optimization schemes.
(6) LLRNMF [12]. This is a clustering approach based on a local learning regularized nonnegative matrix factorization (NMF). It adds an additional constraint on NMF that the cluster label of each data instance can be predicted by the data instances in its neighborhood. It can be optimized via an iterative multiplicative updating algorithm and its convergence is theoretically guaranteed.
(7) LPFNMTF [34]. This aims to simultaneously cluster both data side and feature side of an input data matrix by considering the incorporated manifold information. It constrains the factor matrices from nonnegative matrix factorization to be cluster indicator matrices, the resulted factor matrices can directly assign cluster labels to data instances and features due to the nature of indicator matrices. An efficient updating algorithm to optimize the objective with quick convergence is presented.
(8) RSS [13]. This is a spectral clustering approach using affinity matrix. It simultaneously learns the representations of data and the affinity matrix of representation in a unified optimization framework. An augmented lagrangian multiplier (ALM) based

[1] Wap, La1 and tr12 are publicly available from http://archive.ics.uci.edu/ml/datasets.html.

[2] Vote, Diag-Bcw, Abalone, Krvs and Caltech101 Silhouettes, which are publicly available from http://glaros.dtc.umn.edu/gkhome/views/cluto/.

[3] tr12 is publicly available from http://trec.nist.gov.

[4] TDT2-20 is publicly available from http://www.nist.gov/speech/tests/tdt/tdt98/index.htm.

algorithm is designed to effectively and efficiently achieve the optimal solution of the problem.

(9) SSC [9]. This is a subspace clustering approach based on sparse representation techniques. A key advantage of the algorithm is its ability to directly deal with data nuisances, such as noise, sparse outlying entries, and missing entries as well as the more general class of affine subspaces by incorporating the corresponding models into the sparse optimization problem.

(10) AMRMF [39]. This aims to jointly learn an affinity matrix with matrix factorization, $\ell_{2,1}$-norm is also employed to measure the loss of matrix factorization. The solution of the proposed approach is reached by a novel augmented lagrangian multiplier (ALM) based algorithm.

(11) ALSLDC [18]. This is a document co-clustering approach with adaptive local structure learning based on nonnegative matrix tri-factorization (NMTF). It performs intrinsic local structure learning and tri-factorization simultaneously, and then document co-clustering is conducted. Besides, an efficient iterative updating algorithm is proposed with guaranteed convergence.

(12) RGRNMF [17]. This is a robust graph-based approach to approximate the cleaned data recovered from sparse outliers for clustering. Under the assumption that there may exist a few entries of the data corrupted arbitrarily, but the corruption is sparse. To address this problem, An error matrix is introduced that could capture the corrupted data as well as the spare corruption. ℓ_1-norm is also employed to alleviate the impact of the unreliable as well as noise and prove the robustness. An efficient iterative updating algorithm is proposed with guaranteed convergence.

4.3 Parameter setting

As each selected baseline method has one or more parameters to be tuned, in order to compare fairly and achieve the best results as reported by the respective authors. Without loss of generality, we independently run each algorithm 10 times with different parameter settings and report the mean performance. In each experiment, we run k-means clustering processing 30 times and obtain the best result to reduce the randomness of k-means. We empirically set the number of data clusters equal to the true number of classes for all clustering algorithms on all benchmark datasets.

For PCA algorithm, we set the dimension of the principle component subspace as the same number of data clusters c, and the final clustering result is obtained by performing k-means on the subspace.

For RSS algorithm, we turn the three regularized weight parameters in the same range of $[2^{-3}, 2^{-2}, \ldots, 2^3]$, respectively. For SSC algorithm, the regularization parameter α is set in the range of $[10^{-5}, 10^{-4}, \ldots, 10^5]$, and the regularization parameter ρ is set in the range of $[1, 2, \ldots, 5]$.

There are several graph regularized matrix factorization algorithms, i.e., GNMF, GDRNMF, LLRNMF, LPFNMTF, AMRMF, ALSLDC and RGRNMF. For GRNMF algorithm, the binary weighting scheme is used for constructing the nearest-neighbor graph and the neighborhood size k is set in the range of $[1, 2, \ldots, 10]$ following [3], and the regularization parameter μ is searched in the range of [0.1, 1, 10, 100, 500, 1000]. For simplicity, GDRNMF algorithm sets the value of regularization parameter on data graph is the same as that of the feature graph, i.e., $\mu = \lambda$. For LLRNMF algorithm, the neighborhood size k for computing the local learning regularization is set in the range of [5, 10, 20, 30, 40, 50, 80], and the regularization parameter μ is searched in the range of [0.1, 1, 10, 100, 500, 1000]. For the image datasets (e.g., Diag-Bcw, Digit and Caltech101 Silhouettes), Gaussian kernel is utilized with the scale parameter tuned, while for the textual datasets (e.g., Wap, La1 and tr12), the cosine kernel is exploited. For LPFNMTF algorithm, we construct the nearest-neighbor graph following [12], where the neighborhood size k for constructing graph is set in the range of $[1, 2, \ldots, 10]$, and the regularization parameters (i.e., α and β) are set in the range of [0.1, 1, 10, 100, 500, 1000]. Kernel selection is the same as that in LLRNMF. For AMRMF algorithm, we turn two regularization parameters (i.e., γ and β) in the same range of $[10^{-5}, 10^{-4}, \ldots, 10^5]$. For ALSLDC algorithm, the neighborhood size k for both constructing data graph and feature graph is set in the range of $[1, 2, \ldots, 10]$, the remain regularization parameters are set in the same value (i.e., $\alpha = \lambda = \mu = \beta$), and are searched in the range of [0.1, 1, 10, 100, 500, 1000] for simplicity. For RGRNMF algorithm, the neighborhood size k for exploiting the intrinsic geometrical structure is set in the range of $[1, 2, \ldots, 5]$, the regularization parameter λ on the feature graph is set in the range of [100, 500, 1000] and the regularization parameter γ is set 1 or 2, which are reported by the published paper at the best performance.

Finally, our proposed ALLRNMF approach has two essential parameters k and λ, the neighborhood size k as described in Section 3.3 is determined by the grid $[1, 2, \ldots, 10]$, the regularization parameter λ is set in the range of [0.1, 1, 10, 100, 500, 1000].

Note that no parameter selection is needed for k-means and NMF, given the number of clusters.

4.4 Evaluation measures

We rely on two standard metrics, clustering accuracy (ACC) and purity to evaluate the performance of our proposed

approach. They are commonly used in the measurement of clustering results [16], and allow us to compare our results against previous works, these two metrics are all generally positive correlated and a larger score indicates a better clustering approach. The accurate definitions can be summarized as follows.

Clustering Accuracy (ACC) ACC is used for discovering the one-to-one relationship between classes and clusters along with measure the extent to which each cluster contained data instances from the corresponding class. ACC is defined as follows:

$$ACC = \frac{\sum_{i=1}^{n} \delta(map(r_i), l_i)}{n}, \tag{37}$$

where r_i and l_i are predicted label and corresponding ground truth label of data instance x_i, respectively. $\delta(x, y)$ is the delta function that should be equivalent to 1 if $x = y$ and equals 0 otherwise. $map(r_i)$ is the permutation mapping function that maps each cluster label r_i to the equivalent label from the data set, and n denotes the total number of the data instances.

Purity is applied to measure the extent to which each cluster contained data instances from primarily one class. The purity of a clustering result is computed by the weighted sum of each cluster purity values and can be defined as follows:

$$Purity = \sum_{i=1}^{c} \frac{n_i}{n} P(S_i), \quad P(S_i) = \frac{1}{n_i} \max_j (n_i^j), \tag{38}$$

where S_i is the particular cluster of size n_i, n_i^j is the number of data instances of the i-th input class that was assigned to the cluster $C_j (1 \leq j \leq c)$, n is the number of data instances as well as c is the number of clusters.

4.5 Clustering results

After comparing the proposed ALLRNMF approach with other baseline algorithms, we show the clustering results on eight benchmark datasets in terms of ACC and Purity in Tables 4 and 5, respectively. We mark the top-2 clustering results in bold face.

The advanced clustering algorithms (i.e., RSS and SSC), which are leveraged to learn the representations of data and the affinity matrix of representation for spectral clustering, show the certain improvements at the accuracy metric comparing to the basic methods such as k-means, NMF and PCA. However, it is also observed that the purity metrics even lower than that of the PCA algorithm as mentioned above. The main cause is that these two advanced algorithms present unsatisfied results for Krvs,

Table 4 Clustering results of different methods in terms of accuracy

Dataset	k-means	NMF	PCA	GRNMF	GDRNMF	LLRNMF	LPFNMTF	RSS	SSC	AMRMF	ALSLDC	RGRNMF	ALLRNMF
Vote	0.8690	0.8667	0.6667	0.8669	**0.8777**	0.8754	0.8543	0.6869	0.6526	0.8609	0.8770	0.8776	**0.8782**
TDT2-20	0.2316	0.3587	0.4922	0.4540	0.2492	0.2745	**0.6233**	0.5277	0.5314	0.5292	0.4995	0.6007	**0.6648**
Wap	0.7966	0.7754	0.7913	0.8337	0.8224	0.8701	**0.8922**	0.8752	0.8602	0.8903	0.8907	0.7323	**0.9270**
Abalone	0.3589	0.3741	0.3537	0.3589	**0.3838**	0.3688	0.3686	0.2008	0.2363	0.3990	0.3735	0.3218	**0.3851**
Krvs	0.7437	0.6191	0.7439	0.6046	0.6043	0.3740	0.7625	0.6772	0.6857	0.7522	**0.7670**	0.6147	**0.7679**
La1	0.6768	0.6193	0.4440	0.6294	0.6327	0.4827	0.6436	0.7892	0.7604	0.7720	**0.8186**	0.7924	**0.8305**
tr12	0.2389	0.2174	0.2596	0.3371	0.3074	0.3073	0.3792	0.3408	0.3379	0.2622	**0.4170**	0.3917	**0.4025**
Dig-Bcw	0.6899	0.7272	0.7420	0.7889	**0.8999**	0.8545	0.8998	0.7303	0.8562	0.7320	0.8677	0.8998	**0.9308**
Digit	0.6193	0.6614	0.6583	0.7819	0.7117	0.7245	0.6338	0.6568	0.6901	0.5922	0.7770	**0.8044**	**0.8125**
Caltech101 Silhouettes	0.5414	0.5902	0.6644	0.7855	**0.7981**	0.7410	0.7292	0.6720	0.6819	0.7825	0.7305	0.7892	**0.8009**
Average	0.5766	0.5810	0.5816	0.6441	0.6287	0.5873	0.6787	0.6157	0.6293	0.6573	**0.7019**	0.6825	**0.7400**

Table 5 Clustering results of different methods in terms of purity

Dataset	k-means	NMF	PCA	GRNMF	GDRNMF	LLRNMF	LPFNMTF	RSS	SSC	AMRMF	ALSLDC	RGRNMF	ALLRNMF
Vote	0.7022	0.6410	0.6978	0.6989	0.7045	0.6966	0.7023	0.7002	**0.7212**	0.7064	0.7033	0.6799	**0.7303**
TDT2-20	0.5575	0.6513	0.7908	0.6554	0.7085	0.5322	**0.8020**	0.6874	0.6533	0.8007	0.5551	0.5312	**0.8206**
Wap	0.8541	0.8525	0.8487	0.8541	0.8478	0.8875	0.8487	0.8506	0.8423	0.9208	**0.9377**	0.7992	**0.9374**
Abalone	0.3176	0.3103	0.3215	0.3179	0.3095	0.2271	0.3244	0.2569	0.2566	**0.4278**	0.4077	0.3004	**0.4506**
Krvs	0.7437	0.6191	0.7585	0.6046	0.6043	0.5751	**0.7626**	0.6503	0.6622	0.7591	0.7588	0.6122	**0.7873**
La1	0.7123	0.6527	0.6981	0.6610	0.6652	0.6175	0.7145	0.8070	0.7293	0.8255	**0.8709**	0.6702	**0.8669**
tr12	0.3030	0.3077	0.3165	0.3371	0.3373	0.3272	0.3068	0.3002	0.3047	0.3168	**0.3296**	0.2989	**0.3298**
Diag-Bcw	0.5874	0.3911	0.5164	0.5720	0.5355	0.4645	0.4701	0.6090	0.6031	**0.6160**	0.5706	0.5902	**0.6168**
Digit	0.7425	0.6904	0.6555	**0.7986**	0.7446	0.7262	0.6639	0.6008	0.6363	0.5352	0.7477	0.7801	**0.8156**
Caltech101 Silhouettes	0.5116	0.6614	0.6583	0.7819	**0.7907**	0.7245	0.7038	0.6224	0.6305	0.7302	0.7244	0.7897	**0.8025**
Average	0.6032	0.5778	0.6262	0.6282	0.6248	0.5778	0.6299	0.6085	0.6040	**0.6639**	0.6606	0.6052	**0.7158**

TDT2-20 datasets, e.g., the purity metrics of SSC only achieves 0.6622 while the PCA reaches 0.7585 in Krvs dataset.

The graph-based clustering algorithms, e.g., GRNMF, GDRNMF, LLRNMF, LP-FNMTF, AMRMF, and ALSLDC, which learn the affinity matrix to model the intrinsic structure of data. RGRNMF algorithm achieves robust NMF by approximating the cleaned data recovered from sparse outliers for clustering. The aforementioned approaches show the respectable improvements comparing to the basic clustering algorithms (i.e., k-means, NMF and PCA) in overall performance for eight benchmark datasets. We find that it is crucial to explore a better data feature representation as well as learning the intrinsic local structure of data for a better clustering result. In general, our proposed ALLRNMF algorithm consistently outperforms the other compared algorithms in terms of ACC and purity metrics. For the advanced GDRNMF and LPFNMTF algorithms, they still produce competitive results at ACC as well as purity metrics for some datasets (e.g., TDT2-20 and Caltech101 Silhouettes). Moreover, ALSLDC algorithm presents the significant performance on all textual datasets and even exceeds the proposed ALLRNMF in both Wap and La1 datasets at purity metric (i.e., on the dataset of Wap and La1 with more than 0.0003 and 0.004 of purity). Compared with the second-best performance, ALLRNMF improves the clustering accuracies by around 0.02 in most of the cases (e.g., for TDT2-20, Wap, La1, Dig-Bcw, etc.). In addition, we calculate the mean performance of the different baseline algorithms on all datasets, shown in the last row of each table. An interesting point is that ALSLDC and AMRMF are then demonstrated to be the second-best approaches in terms of ACC and purity metrics, respectively. While ALLRNMF achieves the best overall performance. The quantitative results fully demonstrate the effectiveness of our proposed ALLRNMF approach.

To emphasize, the superiority of ALLRNMF may arise in the following two aspects: (1) The construction of the data graph by searching the adaptive and optimal neighbors for each data instance is better than just considering the k-Nearest neighbors. Since the intrinsic local structure learning and matrix factorization are performed simultaneously, i.e., intrinsic local structure is adaptively learned from the results of factorization, and the decomposable factors are reformulated to preserve the refined local structure of data, and the intrinsic local structure of the data space can be better captured. (2) The nonnegativity, ALLRNMF inherits the nonnegative matrix factorization (NMF), which is suitable for nonnegative data, e.g., image datasets and textual datasets.

We also investigate the computational cost of five representative baseline algorithms on four complicated

benchmark datasets, as shown in Table 6. The experiments are conducted on an eight-core Windows Server with each core CPU at 2.5GHz, and the total memory is 16G. From Table 6, it is easily observed that graph-based clustering algorithms, including LLRNMF, ALSLDC, GDRNMF, in general have slower computation time than using k-means algorithm alone due to their relatively complicated computations. Besides, among the aforementioned algorithms, LLRNMF algorithm is the most efficient approach which takes around 323.67s, 428.50s, 766.70s to finish one experiment process on TDT2-20, tr12 and Caltech 101 Silhouettes datasets, respectively. However, its accuracies are general as mentioned above. Although ALSLDC algorithm presents the significant performance on the textual datasets such as Wap, its computational performance is worse than that of our proposed ALLRNMF approach on Wap and tr12 datasets, indicating that there exists a trade-off between the accuracy and computational cost so as to make it more feasible to learn local structure of input data for data clustering. The comparison also reveals that the good computational performance of our approach is due to an appropriate design of model to learn the data similarity matrix according to the data, rather than depending on other predefined model, which is comparable with conventional approaches. Therefore, we can further optimize the computational cost of the proposed approach in future research efforts.

4.6 Study on parameter tuning

In order to verify different experimental parameter settings, we select four clustering algorithms for the better accuracies (i.e., k-means, GDRNMF, LLRNMF and ALLRNMF) as described above, and investigate their sensitivity with respect to the neighborhood size k and the regularization parameter λ on six benchmark datasets. We vary the value of one parameter and keep the other parameters fixed at the optimal value. The number of iterations is empirically fixed at 30 times.

The clustering results with respect to the nearest neighbors k varies from 1 to 10 are shown in Fig. 1. It is obvious that ALLRNMF is a little sensitive to the choice of k values comparing to the other algorithms on the six datasets, thus we could know that the suitable value of k is critical to our method. Besides, we find that the overall performance of ALLRNMF decreases slightly as k increases, this might be because the assumption (i.e., similar instance pairs with a smaller distance should have a larger probability to be assigned to the probabilistic neighbors) that ALLRNMF depends on could fail due to the increasement of k value. Generally speaking, ALLRNMF approach arrives at the good ACCs for these six benchmark datasets when k varies from 2 to 8.

The clustering results with respect to the regularization parameter λ varies in the range of [0.1, 1, 10, 100, 500,

Table 6 Running time of different baseline methods

Dataset	Baseline methods	Mean time taken for a single iteration (sec)	Mean time for one experiment (sec)	Mean time for k-means (sec)
TDT2-20	k-means	5.36	192.05	
	GDRNMF	10.65	404.85	39.92
	LLRNMF	8.90	323.67	
	ALSLDC	10.31	346.60	
	ALLRNMF	11.45	378.91	
Wap	k-means	5.29	165.70	
	GDRNMF	11.76	435.91	40.98
	LLRNMF	10.95	378.75	
	ALSLDC	10.32	428.15	
	ALLRNMF	10.80	397.40	
tr12	k-means	7.23	225.92	
	GDRNMF	10.85	446.81	75.08
	LLRNMF	10.80	428.50	
	ALSLDC	11.67	597.40	
	ALLRNMF	11.82	486.05	
Caltech 101 Silhouettes	k-means	11.51	546.27	
	GDRNMF	18.12	828.85	93.40
	LLRNMF	15.95	766.70	
	ALSLDC	22.65	1036.47	
	ALLRNMF	20.85	939.50	

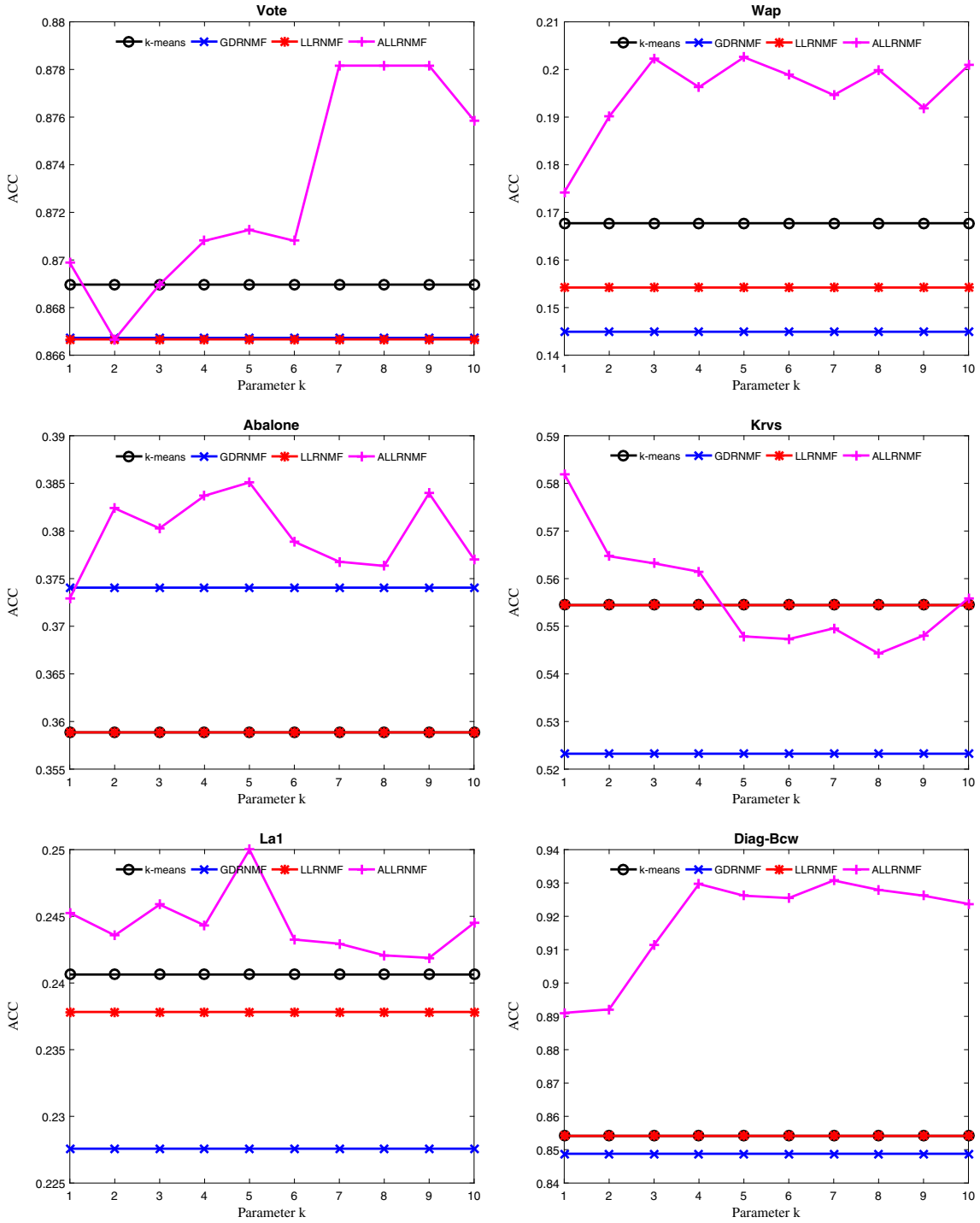

Fig. 1 The computated results of accuracy on our data sets with respect to the neighborhood size k

1000] are shown in Fig. 2. It is obvious that ALLRNMF is a little sensitive to the λ comparing to the other algorithms on the six datasets, thus we could know that the suitable value of λ is critical to our method. We observe that the lower accuracy results with the variations when λ varies from 0 to 10. Specially, the ACC arrives at the lowermost rates when λ = 10 in the Vote, Abalone dataset, respectively. The variations reduce and then present the linear increase on clustering accuracies in almost all datasets when λ values exceeded 10. This might be since the adaptive and optimal neighborhood is not well fully employed and the learned sparse similarity matrix when the choice of the small λ

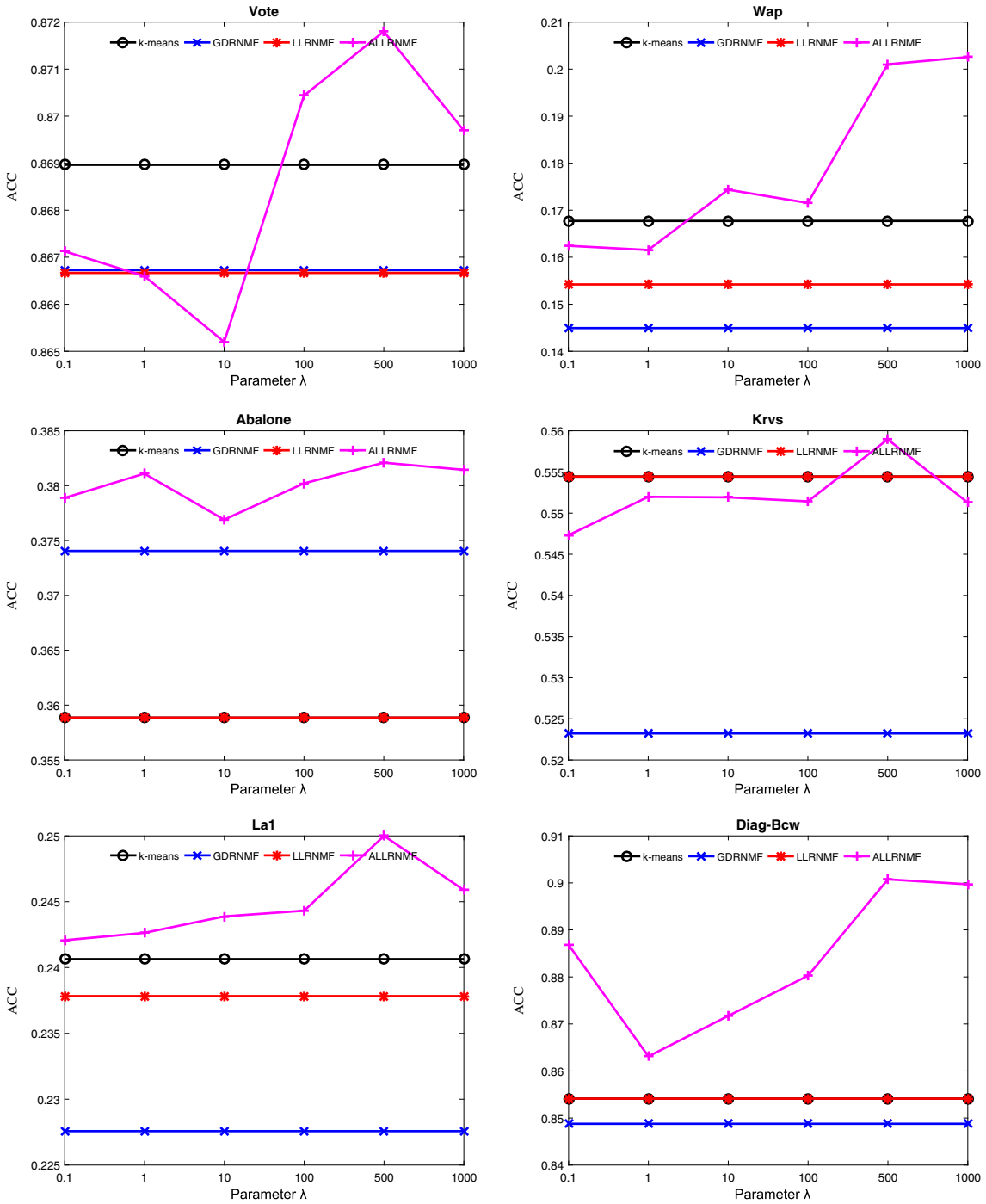

Fig. 2 The computated results of accuracy on our data sets with respect to the regularization parameter λ

values. Generally speaking, ALLRNMF approach arrives at the good ACCs for these six benchmark datasets when λ varies from 100 to 1000.

In summary, for the sake of the overall performance of our proposed ALLRNMF approach, we may vary the value k from 2 to 8, and λ from 100 to 1000 in practical clustering circumstances.

5 Conclusion

In this paper, we proposed an adaptive local learning regularized nonnegative matrix factorization (ALLRNMF) approach for learning a data similarity matrix jointly with the clustering framework. ALLRNMF regarded the matrix as an additional regularization and performs the

nonnegative matrix factorization. It is significant for the constraint of the similarity matrix, because it could encode both the discriminative information as well as the learned adaptive local structure, and benefits the data clustering on manifold. While we also proposed an effective alternative optimization algorithm to solve the optimization problem related to our approach. Experimental results on numerous of real-world datasets demonstrated the effectiveness of our approach.

In our future work, we will investigate the following questions.

- The proposed ALLRNMF approach can be extended to a more generalized version. We have experimentally found that the noises and outliers could affect the experimental results due to the sensibility of ALLRNMF to input data in practice. Hence, we could improve the adaptive capacity of ALLRNMF, in particular to dealing with large-scale data from real world.
- In a matrix factorization point of view, adaptive and optimal neighborhood information have been employed for data similarity matrix learning. Besides, other knowledge (e.g., label[–] information, specified manifold information) could also be used to the local structure learning. In this way, the ALLRNMF approach might arrive at a scalable result in practical clustering circumstances.

Appendix A: Proof of Theorem 2

Proof We rewrite (33) as

$$L(\mathbf{U}) = tr(-2\mathbf{V}\mathbf{X}^T\mathbf{U} + \mathbf{U}\mathbf{V}\mathbf{V}^T\mathbf{U}^T). \tag{39}$$

By applying Lemma 2, we have

$$tr(\mathbf{U}\mathbf{V}\mathbf{V}^T\mathbf{U}^T) \leq \sum_{ij} \frac{(\mathbf{U}'\mathbf{V}\mathbf{V}^T)_{ij}\mathbf{U}_{ij}^2}{\mathbf{U}'_{ij}}.$$

To obatin the lower bound for the remaining terms, we use the inequality that

$$z \geq 1 + logz, \forall z \geq 0. \tag{40}$$

Then

$$tr(\mathbf{V}\mathbf{X}^T\mathbf{U}) \geq \sum_{ij} (\mathbf{X}\mathbf{V}^T)_{ij}\mathbf{U}'_{ij}\left(1 + log\frac{\mathbf{U}_{ij}}{\mathbf{U}'_{ij}}\right).$$

By summing over all the bounds, we can get $g(\mathbf{U}, \mathbf{U}')$, which obviously satisfies: (1) $g(\mathbf{U}, \mathbf{U}') \geq J_{ALLRNMF}(\mathbf{U})$; (2) $g(\mathbf{U}, \mathbf{U}) = J_{ALLRNMF}(\mathbf{U})$.

To find the minimum of $g(\mathbf{U}, \mathbf{U}')$, we take the Hessian matrix of $g(\mathbf{U}, \mathbf{U}')$

$$\frac{\partial^2 g(\mathbf{U}, \mathbf{U}')}{\partial \mathbf{U}_{ij}\partial \mathbf{U}_{kl}} = \delta_{ik}\delta_{jl}\left(\frac{2(\mathbf{U}'\mathbf{V}\mathbf{V}^T)_{ij}}{\mathbf{U}'_{ij}} + 2(\mathbf{X}\mathbf{V}^T)_{ij}\frac{\mathbf{U}'_{ij}}{\mathbf{U}_{ij}^2}\right)$$

which is a diagonal matrix with positive diagonal elements. So $g(\mathbf{U}, \mathbf{U}')$ is a convex function of $\mathbf{U}$, and we can obtain the global minimum of $g(\mathbf{U}, \mathbf{U}')$ by setting $\frac{\partial g(\mathbf{U}, \mathbf{U}')}{\partial \mathbf{U}_{ij}} = 0$ and solving for $\mathbf{U}$, from which we can get (34). □

Appendix B: Proof of Theorem 4

Proof We rewrite (35) as

$$L(\mathbf{V}) = tr\left(-2\mathbf{X}^T\mathbf{U}\mathbf{V} + \mathbf{V}^T\mathbf{U}^T\mathbf{U}\mathbf{V} - \lambda\mathbf{V}\mathbf{L}_S^+\mathbf{V}^T + \lambda\mathbf{V}\mathbf{L}_S^-\mathbf{V}^T\right). \tag{41}$$

By applying Lemma 2, we have

$$tr(\mathbf{V}^T\mathbf{U}^T\mathbf{U}\mathbf{V}) \leq \sum_{ij} \frac{(\mathbf{U}^T\mathbf{U}\mathbf{V}')_{ij}\mathbf{V}_{ij}^2}{\mathbf{V}'_{ij}},$$

$$tr(\mathbf{V}\mathbf{L}_S^-\mathbf{V}^T) \leq \sum_{ij} \frac{(\mathbf{V}'\mathbf{L}_S^-)_{ij}\mathbf{V}_{ij}^2}{\mathbf{V}'_{ij}}.$$

To obatin the lower bound for the remaining terms, we use the inequality in (40), then

$$tr(\mathbf{X}^T\mathbf{U}\mathbf{V}) \geq \sum_{ij} (\mathbf{U}^T\mathbf{X})_{ij}\mathbf{V}'_{ij}\left(1 + log\frac{\mathbf{V}_{ij}}{\mathbf{V}'_{ij}}\right),$$

$$tr(\mathbf{V}\mathbf{L}_S^+\mathbf{V}^T) \geq \sum_{ijk} (\mathbf{L}_S^+)_{jk}\mathbf{V}'_{ij}\mathbf{V}'_{ik}\left(1 + log\frac{\mathbf{V}_{ij}\mathbf{V}_{ik}}{\mathbf{V}'_{ij}\mathbf{V}'_{ik}}\right),$$

By summing over all the bounds, we can get $g(\mathbf{V}, \mathbf{V}')$, which obviously satisfies: (1) $g(\mathbf{V}, \mathbf{V}') \geq J_{ALLRNMF}(\mathbf{V})$; (2) $g(\mathbf{V}, \mathbf{V}) = J_{ALLRNMF}(\mathbf{V})$.

To find the minimum of $g(\mathbf{V}, \mathbf{V}')$, we take the Hessian matrix of $g(\mathbf{V}, \mathbf{V}')$

$$\frac{\partial^2 g(\mathbf{V}, \mathbf{V}')}{\partial \mathbf{V}_{ij}\partial \mathbf{V}_{kl}} = \delta_{ik}\delta_{jl}\left(\frac{2(\mathbf{U}^T\mathbf{X} + 2\lambda\mathbf{L}_S^+)_{ij}\mathbf{V}'_{ij}}{\mathbf{V}_{ij}^2} + \frac{2(\mathbf{U}^T\mathbf{U}\mathbf{V}' + \lambda\mathbf{V}'\mathbf{L}_S^-)_{ij}}{\mathbf{V}'_{ij}}\right)$$

which is a diagonal matrix with positive diagonal elements. So $g(\mathbf{V}, \mathbf{V}')$ is a convex function of $\mathbf{V}$, and we can obtain the global minimum of $g(\mathbf{V}, \mathbf{V}')$ by setting $\frac{\partial g(\mathbf{V}, \mathbf{V}')}{\partial \mathbf{V}_{ij}} = 0$ and solving for $\mathbf{V}$, from which we can get (36). □

References

1. Belkin M, Niyogi P, Sindhwani V (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res 7:2399–2434
2. Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, Cambridge
3. Cai D, He X, Han J, Huang TS (2011) Graph regularized nonnegative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Intell 33(8):1548–1560
4. Cai D, He X, Wu X, Han J (2008) Non-negative matrix factorization on manifold. In: Proceedings of the 8th international conference on data mining. IEEE, Piscataway, pp 63–72
5. Cai X, Nie F, Huang H (2013) Multi-view k-means clustering on big data. In: Proceedings of the 25th international joint conference on artificial intelligence. AAAI, Cambridge, pp 2598–2604
6. Chung FR (1997) Spectral graph theory, vol 92 American Mathematical Soc
7. Ding C, Li T, Jordan MI (2009) Convex and semi-nonnegative matrix factorizations. IEEE Trans Pattern Anal Mach Intell 32(1):45–55
8. Ding C, Li T, Peng W, Park H (2006) Orthogonal nonnegative matrix t-factorizations for clustering. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 126–135
9. Elhamifar E, Vidal R (2015) Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans Pattern Anal Mach Intell 35(11):2765–2781
10. Gokcay E, Principe JC (2002) Information theoretic clustering. IEEE Trans Pattern Anal Mach Intell 24(2):158–171
11. Gu Q, Ding C, Han J (2011) On trivial solution and scale transfer problems in graph regularized nmf. In: Proceedings of the 23rd international joint conference on artificial intelligence, vol 22. AAAI, Cambridge, pp 1288–1295
12. Gu Q, Zhou J (2009) Local learning regularized nonnegative matrix factorization. In: Proceedings of the 21st international joint conference on artificial intelligence. AAAI, Cambridge, pp 1046–1051
13. Guo X (2015) Robust subspace segmentation by simultaneously learning data representations and their affinity matrix Proceeding of the 24nd international joint conference on artificial intelligence. AAAI, Cambridge, pp 3547–3553
14. Hagen L, Kahng AB (2006) New spectral methods for ratio cut partitioning and clustering. IEEE Trans Comput Aided Des Integr Circuits Syst 11(9):1074–1085
15. Han EH, Boley D, Gini M, Gross R, Hastings K, Karypis G, Kumar V, Mobasher B, Moore J (1998) Webace:a web agent for document categorization and exploration. In: Proceedings of the 2nd international conference on autonomous agents, pp 408–415
16. Huang J, Nie F, Huang H, Ding C (2014) Robust manifold nonnegative matrix factorization. ACM Trans Knowl Discov Data 8(3):11
17. Huang S, Wang H, Li T, Li T, Xu Z (2018) Robust graph regularized nonnegative matrix factorization for clustering. Data Min Knowl Disc 32(2):483–503
18. Huang S, Xu Z, Lv J (2018) Adaptive local structure learning for document co-clustering. Knowl-Based Syst 148:74–84
19. Jain AK (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
20. Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: Proceedings of the 14th advances in neural information processing systems. MIT Press, Cambridge, pp 556–562
21. Liu G, Lin Z, Yan S, Sun J, Yu Y, Ma Y (2013) Robust recovery of subspace structures by low-rank representation. IEEE Trans Pattern Anal Mach Intell 35(1):171–184
22. Liu Y, Jiao L, Shang F (2013) A fast tri-factorization method for low-rank matrix recovery and completion. Pattern Recogn 46(1):163–173
23. Luxburg UV (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416
24. MacQueen J, et al. (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 50th Berkeley symposium on mathematical statistics and probability, vol 1, pp 281-297, Oakland, USA
25. Nie F, Wang X, Huang H (2014) Clustering and projected clustering with adaptive neighbors. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 977–986
26. Nie F, Wang X, Jordan MI, Huang H (2016) The constrained laplacian rank algorithm for graph-based clustering. In: Proceedings of the 30th AAAI conference on artificial intelligence. AAAI, Cambridge, pp 1969–1976
27. Peng C, Kang Z, Hu Y, Cheng J, Cheng Q (2017) Robust graph regularized nonnegative matrix factorization for clustering. ACM Trans Knowl Discov Data (TKDD) 11(3):33
28. Rai N, Negi S, Chaudhury S, Deshmukh O (2016) Partial multi-view clustering using graph regularized nmf. In: Proceeding of 23rd international conference on pattern recognition (ICPR). IEEE, Piscataway, pp 2192–2197
29. Seung HS, Lee DD (2000) The manifold ways of perception. Science 290(5500):2268–2269
30. Shang F, Jiao L, Wang F (2012) Graph dual regularization non-negative matrix factorization for co-clustering. Pattern Recogn 45(6):2237–2250
31. Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905
32. Shlens J (2014) A tutorial on principal component analysis. arXiv:1404.1100
33. Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323
34. Wang H, Nie F, Huang H, Makedon F (2011) Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In: Proceedings of the 22nd international joint conference on artificial intelligence. AAAI, Cambridge, pp 1553–1558
35. Wang S, Tang J, Liu H (2015) Embedded unsupervised feature selection. In: Proceedings of the 29th AAAI conference on artificial intelligence. AAAI, Cambridge, pp 470–476
36. Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 267–273
37. Xu Z, King I, Lyu MRT, Jin R (2010) Discriminative semi-supervised feature selection via manifold regularization. IEEE Trans Neural Netw 21(7):1033–1047
38. Yoo J, Choi S (2010) Orthogonal nonnegative matrix tri-factorization for co-clustering: multiplicative updates on stiefel manifolds. Inf Process Manag 46(5):559–570
39. Zhang L, Zhang Q, Du B, You J, Tao D (2017) Adaptive manifold regularized matrix factorization for data clustering. AAAI, Cambridge

Yongpan Sheng is currently a Ph.D. candidate at School of Computer Science and Engineering, University of Electronic Science and Technology of China. He received the bachelor degree in network engineering from the Chengdu University of Information Technology, and master degree in software engineering from University of Electronic Science and Technology of China in 2011 and 2014, respectively. His research interests include knowledge graphs and natural language processing.

Meng Wang is working as an assistant professor in the Knowledge Graph & AI Research Group, School of Computer Science and Engineering, Southeast University, China. He obtained the doctoral degree from the Department of Computer Science and Technology, Xi'an Jiaotong University in 2018. His research areas are knowledge graph (KG), semantic search, NLP, and cross-modal data.

Tianxing Wu is a post-doctoral researcher in School of Computer Science and Engineering, Nanyang Technological University, Singapore. He received the Ph.D. degree from Southeast University (China) in 2018. His research interests include Knowledge Graph, Semantic Web and Data Mining. He has published several papers in proceedings of major conferences or journals, such as AAAI, ECAI, ISWC, Journal of Web Semantics, International Journal on Semantic Web and Information Systems, etc. He is also the editorial board member of International Journal on Semantic Web and Information Systems.

Han Xu is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of Electronic Science and Technology of China. His current research interests include cloud computing, and performance modeling and optimization.