融合关联矩阵自学习和显式秩约束的数据表示分簇算法

郑建炜 鞠振宇 朱文博 王万良

(浙江工业大学计算机科学与技术学院 杭州 310023)

摘 要 复杂异构分布的高维数据在大数据时代随处可见,高效地挖掘其子空间结构并进行准确的分簇是机器视觉和模式识别领域的研究热点.低秩表示算法(Low-Rank Representation,LRR)因其优越的低维子空间挖掘能力而备受关注,其性能很大程度上取决于关联矩阵的构建,常见的方法都是通过原始输入数据或表示系数直接一次成形.然而,这些方法都采用独立的步骤进行表示系数计算以及关联矩阵学习,无法保证总体算法的最优性.针对该问题,该文提出一种新的 LRR 型数据表示分簇法(Data Representation Clustering,DRC)应用于实际子空间分割问题.首先,为实现模型的快速求解,DRC 保留了基本数据表示框架中的光滑正则项并剔除了非负性、稀疏性等复杂约束;其次,将相似度矩阵的自适应学习策略添加至统一的数据表示框架,联合原始输入数据和表示系数确保目标关联矩阵在无噪环境下具备明确的对角分布结构.最后,对关联矩阵对应的 Laplacian 矩阵添加一种新的秩约束,在含噪环境下引导相似度连接结构与簇目标数的一致性.采用交替更新法对模型进行求解,保证目标函数单变量优化的全局最优性以及整体收敛性.人工合成数据和 8 个公开数据集的实验结果表明,DRC 算法在分簇精度、归一化互信息、参数敏感性等指标上都具有优秀的性能.

关键词 关联矩阵;低秩表示;谱分簇;拉普拉斯正则项;归一化互信息

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.00497

Data Representation Clustering via Simultaneously Affinity Matrix Self-Learning and Explicit Rank Constraint

ZHENG Jian-Wei JU Zhen-Yu ZHU Wen-Bo WANG Wan-Liang

(*College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou* 310023)

Abstract The high dimensional data with complex and heterogeneous distribution can be seen everywhere in the era of big data, and it is a hot topic in the field of machine vision and pattern recognition to efficiently and accurately excavate the spatial structure and segment the subspace cluster. Low Rank Representation (LRR), as one of the most well-known clustering methods, has recently received much more attentions in subspace clustering and visual segmentation due to its high efficacy and strong robustness in exploring low dimensional latent structures of input samples. For a specific set of collected dictionary corrupted with various kinds of errors, the main purpose of LRR is to learn an intrinsic rank representation of all samples jointly. The performance of LRR based clustering approaches heavily depends on learned affinity matrices, which are usually constructed either directly from the raw data or from their computed representation coefficients. However, these methods may not guarantee an overall optimum since data representation and

收稿日期:2016-10-17;在线出版日期:2017-08-13.本课题得到国家自然科学基金(61602413,61873240)、浙江省自然科学基金(LY19F030016)资助.**郑建炜**,男,1982 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为机器学习、模式识别.E-mail:zjw@zjut.edu.cn.**鞠振宇**,男,1993 年生,硕士研究生,主要研究方向为数据挖掘、人工智能.**朱文博**,男,1990 年生,硕士研究生,中国计算机学会(CCF)会员,主要研究方向为数据挖掘、人工智能.**王万良**(通信作者),男,1957 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为人工智能、大数据分析.E-mail:wwl@zjut.edu.cn.

similarity measurement are often conducted in two independent steps. To improve LRR with this regard, in this paper we propose a new LRR based clustering method, namely Data Representation Clustering (DRC), for practical subspace segmentation problems. DRC embraces three characteristics for precisely learning of affinity matrix. First, along with some deep surveys of related literatures, our model relies on the general framework of data representation with smooth regularization such as Frobenius norm retained but with non-smooth regularization such as sparse or non-negative constraints eliminated for fast implementation. Second, different from the existing LRR based methods, both of the raw data and the iteratively updated coefficient vectors are adopted for adaptively learning the similarity measurement (or Laplacian matrix) to form a block diagonal affinity matrix in noiseless data. Moreover, we fit this structure learning into the general low rank representation formulation for simultaneously optimization of coefficients matrix and affinity matrix. Finally, an explicit low rank constraint is further imposed on the Laplacian matrix of the corresponding affinity matrix, which leads to an exact cluster structure of connected components not only in the clear environment but also in the noisy setting. Aiming for the newly and deliberately proposed model, an efficient algorithm is derived to iteratively update the undetermined variables of our proposed method following up the alternate direction multiplier method (ADMM), and then the theoretical analysis of convergence and complexity are also given. We conduct several experiments on synthetic data to reveal the visual quality and anti-noise capability for clustering of our proposed method, and on real data to demonstrate the superior practical performance of our method over the state-of-the-art alternatives. The experimental results demonstrate that DRC not only achieves better performance for clustering in the indicator of accuracy and normalized mutual information, but also obtains strong competitiveness in model selection and computational efficiency. Furthermore, three reduced versions of our proposed method, such as DRC without similarity matrix, DRC without coefficients matrix and DRC with these two terms optimized independently, prove the effectiveness of our jointly update of affinity matrix and coefficients matrix.

Keywords affinity matrix; low-rank representation; spectral clustering; Laplacian regularizer; Normalized mutual information

1 引 言

复杂异构分布的高维数据在大数据时代随处可见,高效地挖掘其子空间结构并进行准确的分簇是机器视觉和模式识别领域的研究热点.依据子空间表示机制的区别,现有的分簇算法可以分为矩阵分解[1]、代数[2]、统计[3]、谱分簇[4-5]四种主流类型.其中,谱分簇算法具备完整的理论支撑和优秀的应用性能,获得了更为广泛的研究关注和应用扩展,本文工作亦隶属于谱分簇类型.

谱分簇算法的关键步骤是关联矩阵(亦被称为相似度矩阵、邻域矩阵、关联图等)构建,在理想状态下,关联矩阵中的簇间相似度严格为零,而簇内相似度则处于(0,1]之间.常见的方法都采用原始数据或表示系数进行一次构建.其中原始数据法如 ε 邻域图、k 近邻图、全连接图等都存在着明显的缺陷,包括参数敏感度强、数据适应性弱、鲁棒性差等.Nie 等人[6]通过局部连通性约束自适应生成相似度矩阵,能够有效抑制奇异点,也提升了关联矩阵构建的数据适应性.Hou 等人[7]则从全局散度矩阵出发,结合指示矩阵构建目标函数,提出鉴别嵌入分簇算法(Discriminative Embedded Clustering,DEC),也获得了良好的性能指标.然而,单纯地对原始数据的局部空间度量操作难以展现数据的全局结构.因此,表示系数法得到了广泛的研究应用.Cheng 等人[8]引入稀疏表示进行相似图构建,具有更高的算法鲁棒性,也规避了高敏感度的人工设定参数 ε 和 k,但其正则项参数的选择较为困难,且存在 l_1 范数问题求解效率低的问题.Huang 等人[9]采用非负和加权

限制替代文献[8]中的 l_1 范数约束，提出了单纯表示型关联矩阵构建方法，不包含任何待定参数，运行效率得以大幅提升. 类似地，Elhamifar 等人[10]对自表示矩阵添加稀疏性约束，提出了稀疏子空间分簇法(Sparse Subspace Clustering，SSC).

上述表示型算法在数据处理过程中都忽略了样本的联合分布结构. 鉴于此，Liu 等人[11]提出了低秩表示算法(Low-Rank Representation，LRR)用于子空间分簇，以数据的多簇子空间流形分布为前提，通过系数矩阵低秩约束和行空间稀疏约束，实现精确的样本簇结构挖掘和噪声点抑制，在异常监测[12]、目标跟踪[13]、人脸识别[14]和超分辨率重建[15]等众多领域得到了广泛的应用. 为进一步精炼数据表示能力，Feng 等人[16]添加块对角先验于 LRR 的系数矩阵，对关联矩阵的簇结构有明显的效果提升. Nie 等人[17]则采用 Schatten p 范数替换 LRR 中的迹范数，并通过分组结构约束提出一种新的低秩子空间分簇算法(Low Rank Subspace clustering，LRS)，将关联矩阵构建和谱簇指示融为一个步骤，具有更好的低秩逼近能力.

结合局部邻域结构和全局分布状态是谱分簇算法的研究趋势，结构自学习的特征选择算法(Feature Selection with Adaptive Structure Learning，FSASL)[18]联合保局部投影[19]和稀疏系数保持[20]两种思想建立目标函数，获得了更好的分簇指标. 更多算法[21-25]则是以 LRR 为基础添加关联矩阵约束项. Peng 等人[21]将特征选择机制引入 LRR，提出特征选择分簇法(Feature Selection Clustering，FSC)，可以有效抑制奇异特征干扰，但算法缺乏局部空间结构考虑. 针对 LRR 存在的稀疏性不足及噪声敏感等问题，Li 等人[22]提出了一种基于局部图拉普拉斯约束的鲁棒低秩表示聚类模型，在保持表示矩阵分块对角的特性下，增强其稀疏性，减弱了表示字典数据之间的线性相关性. Yin 等人[23]则将图拉普拉斯正则项引入隐低秩表示算法[26]，兼顾样本空间和特征空间的子域流形进行分簇应用. 最近，Yin 等人[24]结合文献[22-23]的优势，进一步提出非负稀疏 Laplacian 正则约束的 LRR 模型(Non-negative Sparse Laplacian regularized LRR，NSLLRR)，以非负性、稀疏性为条件，增加超图拉普拉斯约束，其性能报道优于其它分簇算法. 此外，平滑表示分簇算法(Smooth Representation Clustering，SMR)[25]提出强制组效应概念，并通过样本邻域相似度增强分簇算法的组效应，不仅具有优秀的分簇效果，而且其求解过程不需要迭代运算，效率优于现存的其它 LRR 型算法.

上述算法虽然都兼顾了 LRR 的全局表示能力和图 Laplacian 的局部结构挖掘特性，但都存在类似的缺陷，即采用独立的步骤进行图 Laplacian 构建和 LRR 系数计算，无法保证算法的全局最优性. 鲁棒子域分割算法(Robust Subspace Segmentation，RSS)[27]同时学习低秩表示和关联矩阵，以压缩感知理论为基础，通过重构矢量生成相似度矩阵，由于在迭代过程中受初始表示系数次优性的影响，整体算法依然是次优的，且模型待定参数过多，抑制了其扩展应用能力. 针对现存问题，本文提出一种新的数据表示分簇算法(Data Representation Clustering，DRC)，将低秩表示、相似度学习和簇结构约束融入同一框架，在结合现有算法优势的同时，摒弃相应算法的高复杂度成分和无效因子. 本文工作的贡献包含以下几点：

(1) 提出通过原始输入数据自适应学习的邻域流形结构正则化 LRR 算法，兼顾全局分布特性和局部相似度的同时，剔除额外的表示系数非负性和稀疏性约束，在保证性能的前提下提升算法的运行效率；

(2) 在关联矩阵自学习过程中，添加一种新的显式低秩约束，与经典的低秩约束具有等价性，且算法求解过程更为直观易懂；

(3) 设计模型求解策略，算法具备唯一的最优解. 对算法部分人工待设参数进行了理论分析，推导出经验设置方法或通过实验得到快速优选法则.

本文第 2 节描述 LRR 型分簇算法的相关工作，包括表示范数约束、系数正则项、关联矩阵构建等；第 3 节给出 DRC 算法的详细描述，并给出模型求解策略；第 4 节对 DRC 算法的模型参数、收敛性和复杂度进行理论分析；第 5 节分别通过人工合成数据和公开数据集验证 DRC 的实际分簇性能；最后第 6 节为全文总结.

2　相关工作

本节主要回顾以 LRR 为基础的相关谱分簇算法，给定待分簇数据集为 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in R^{m\times n}$，其中 $\boldsymbol{x}_i$ 是第 i 个 m 维输入数据，n 是样本总数. 假设所有输入数据的分布空间由 c 个互相正交的子空间合并而成，如 $\bigcup_{i=1}^{c} S_i$，其中 $S_1,\cdots,S_c$ 是各个

低维子空间.谱分簇的目标是归纳各种性质尽可能准确地把每个样本 $\boldsymbol{x}_i$ 划分到应属的子簇 S_i. LRR 型算法采用自表示思想求解关联矩阵的相似度元素,并添加不同的系数约束项或 Laplacian 正则项,其构建表示矩阵 $\boldsymbol{Z}$ 的基本框架为

$$\min_{\boldsymbol{Z}} \alpha \|\boldsymbol{X}-A(\boldsymbol{X})\boldsymbol{Z}\|_l+\Omega(\boldsymbol{X},\boldsymbol{Z})$$
$$\text{s.t. } \boldsymbol{Z}\in C \tag{1}$$

其中,α 是人工设定的正则项参数,$A(\boldsymbol{X})$ 是预学习得到的字典矩阵,一般都直接选用 $\boldsymbol{X}$ 以减少算法复杂度[28],$\|\cdot\|_l$ 是表示误差的范数模型,$\Omega(\boldsymbol{X},\boldsymbol{Z})$ 和 C 分别是系数正则项和约束集.式(1)计算得到的最优系数矩阵 $\boldsymbol{Z}^*$ 可进一步通过式(2)构建成对称关联矩阵,并由经典的分簇算法如 Kmeans、NCut[29] 或 RatioCut[30] 完成最终的聚类任务.

$$\boldsymbol{W}=(|\boldsymbol{Z}|+|\boldsymbol{Z}^{\mathrm{T}}|)/2 \tag{2}$$

不同的误差范数模型、系数正则项和约束集构成了不同的 LRR 型分簇算法,表 1 列了部分现存算法的 $\|\cdot\|_l$、$\Omega(\boldsymbol{X},\boldsymbol{Z})$ 和 C 各项具体形式,其中 $\|\cdot\|_1$ 是 l_1 范数,表示所有元素的绝对值之和;$\|\cdot\|_{2,1}$ 是 $l_{2,1}$ 范数,表示所有列向量的 l_2 范数之和;$\|\cdot\|_*$ 是核范数,表示奇异值之和;$\|\cdot\|_{\mathrm{F}}$ 是 Frobenius 范数,表示所有元素平方和的平方根.此外,符号 $\varnothing$ 表示空,λ 和 β 是正则项参数,$\mathbf{1}$ 是所有元素都为 1 的向量,$\boldsymbol{S}$ 是由 $\boldsymbol{Z}$ 学习得到的关联矩阵.在表 1 所有算法中,不同的约束决定了算法不同的性能,例如 F 范数具有平滑性,算法运算效率较高,符合高斯噪声分布的输入数据;l_1 范数具备稀疏性,对特征噪声有较强的抑制作用;$l_{2,1}$ 范数具备列稀疏性,对样本奇异点的抑制效果更为明显.虽然误差模型对分簇算法的影响较大,但本文重点关注关联矩阵与系数矩阵的联合构建问题,因此对误差约束简单采用 F 范数.

表 1 现存 LRR 型算法各子项约束形式

算法	$\|\cdot\|_l$	$\Omega(\boldsymbol{X},\boldsymbol{Z})$	C
SSC	$\|\cdot\|_1$	$\|\boldsymbol{Z}\|_1$	$\{\boldsymbol{Z}\mid z_{ii}=0\}$
LRR	$\|\cdot\|_{2,1}$	$\|\boldsymbol{Z}\|_*$	$\varnothing$
Ref.[22]	$\|\cdot\|_{2,1}$	$\|\boldsymbol{Z}\|_*+\lambda\,\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^{\mathrm{T}})$	$\varnothing$
SMR	$\|\cdot\|_{\mathrm{F}}^2$	$\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^{\mathrm{T}})$	$\varnothing$
NSLLRR	$\|\cdot\|_1$	$\|\boldsymbol{Z}\|_*+\lambda\|\boldsymbol{Z}\|_1+\beta\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^{\mathrm{T}})$	$\{\boldsymbol{Z}\mid z_{ij}\geq 0\}$
RSS	$\|\cdot\|_{\mathrm{F}}^2$	$\|\boldsymbol{Z}\|_{\mathrm{F}}^2+\lambda\|\boldsymbol{S}\|_{\mathrm{F}}^2+\beta\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_a\boldsymbol{Z}^{\mathrm{T}})$	$\{\boldsymbol{S}\mid s_{ij}\geq 0\}$

从表 1 可见,经典算法如 SSC、LRR 等都单纯的对 $\boldsymbol{Z}$ 进行范数特性约束,而后续的 SMR 和 NSLLRR 等算法则联合关联矩阵和表示系数优化 $\boldsymbol{Z}$ 矩阵,其报道的性能较 SSC 和 LRR 有明显的提升.然而,SMR 和 NSLLRR 中的拉普拉斯矩阵 $\boldsymbol{L}$ 仅通过原始输入数据进行一次性构建,无论是在抗噪性还是在算法最优性方面都难以保证.RSS 是唯一联合优化拉普拉斯矩阵和系数矩阵的分簇算法,但其 $\boldsymbol{L}$ 采用表示系数进行构建,在迭代初期是次优的,因此 RSS 算法容易陷入局部最优,影响分簇性能.综合以上分析,考虑到联合局部学习和全局表示对分簇效果的重要性,本文将以此为基础构建目标模型,并添加块对角正则项约束关联矩阵的簇结构.

3 DRC 算法描述

本节提出一种新的数据表示分簇算法,融合了考虑关联矩阵的自适应学习和显式的矩阵秩约束,算法简称为 DRC.首先对所提算法的目标函数构建过程进行描述,然后给出模型求解策略.

3.1 目标函数描述

LRR 作为一种全局分簇算法,在求解邻接矩阵分块对角结构的同时,忽略了样本的局部邻域关系,致使簇间相似度包含稠密的非零元素;此外,混杂噪声和离群奇异点对 LRR 算法的影响也较为明显.针对存在的问题,本文在式(1)表示框架下提出一种融合局部图拉普拉斯学习的低秩表示分簇算法,首先以 SMR 模型为基准框架,将所提算法的初始目标函数设为

$$\min_{\boldsymbol{Z}} \alpha \|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{\mathrm{F}}^2+\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_s\boldsymbol{Z}^{\mathrm{T}}) \tag{3}$$

其中 $\mathrm{tr}(\cdot)$ 是矩阵的迹运算符号,$\boldsymbol{L}_s$ 是由相似度元素构建而成的 Laplacian 矩阵.根据上节相关工作描述,现存算法的相似度矩阵都通过原始输入数据直接构建,独立于模型优化求解过程.在给定对称相似度矩阵 $\boldsymbol{S}$ 后,Laplacian 矩阵构建方式不一,经典的谱分簇算法中,RatioCut[30] 的 $\boldsymbol{L}_s$ 矩阵计算方式为 $\boldsymbol{L}_s=\boldsymbol{D}-\boldsymbol{S}$,其中对角矩阵 $\boldsymbol{D}$ 称为度矩阵,对角元素为 $d_i=\sum_{j=1}^{n}s_{ij}$. NCut[29] 将上述 $\boldsymbol{L}_s$ 进行规范化操作,最终的 Laplacian 矩阵分为对称型 $\boldsymbol{L}_s^{s}=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}$ 和非对称型 $\boldsymbol{L}_s^{ns}=\boldsymbol{D}^{-1}\boldsymbol{L}$ 两种.当给定非对称的 $\boldsymbol{S}$ 时,则对应的 $\boldsymbol{L}_s$ 计算为 $\boldsymbol{L}_s=\boldsymbol{D}-(\boldsymbol{S}^{\mathrm{T}}+\boldsymbol{S})/2$[6],其中度矩阵 $\boldsymbol{D}$ 的对角元素为 $d_i=\sum_j(s_{ij}+s_{ji})/2$.本文采用经典的非规范化 $\boldsymbol{L}_s$ 矩阵构建方法,重点研究关联矩阵的计算策略.首先,根据数据对之间距离更小则相似度更高的假设,相似因子的约束正则项为

$$\min_{\forall i\,\boldsymbol{s}_i\geq 0,\boldsymbol{s}_i^{\mathrm{T}}\mathbf{1}=1}\sum_{i,j=1}^{n}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2 s_{ij} \tag{4}$$

其中 s_{ij} 是相似矩阵 $\boldsymbol{S}$ 的元素,表示样本对 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之

间的同簇概率，条件 $s_i \geq 0$ 和 $s_i^{\mathrm{T}}\mathbf{1}=1$ 用于确保相似度权值的概率特性. 然而，直接采用式(4)计算样本相似度的结果是仅最近邻样本对的权值 $s_{ij}=1$，而其余样本对的相似度则都趋于 0. 为避免平凡解，常采用 l_1 范数和 l_2 范数联合约束的方式优化模型结果[31]. 考虑到 l_1 范数与式(4)约束条件的等价性，通过简单的代数运算可将式(4)正则项进一步调整为

$$\min_{\forall i\, s_i\geq 0,\, s_i^{\mathrm{T}}\mathbf{1}=1} \mathrm{tr}(\boldsymbol{X}\boldsymbol{L}_s\boldsymbol{X}^{\mathrm{T}})+\gamma\|\boldsymbol{S}\|_{\mathrm{F}}^2 \tag{5}$$

将式(5)添加至式(3)，则目标函数更新为

$$\min_{\boldsymbol{Z},\boldsymbol{S}} \alpha\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{\mathrm{F}}^2+\mu\,\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_s\boldsymbol{Z}^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{X}\boldsymbol{L}_s\boldsymbol{X}^{\mathrm{T}})+\gamma\|\boldsymbol{S}\|_{\mathrm{F}}^2$$
$$\text{s.t. } \forall i\ \boldsymbol{s}_i\geq 0,\ \boldsymbol{s}_i^{\mathrm{T}}\mathbf{1}=1 \tag{6}$$

其中，μ 是平衡参数，第一项 $\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{\mathrm{F}}^2$ 在 F 范数约束下最小化逼近误差，第二项 $\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_s\boldsymbol{Z}^{\mathrm{T}})$ 通过 Laplacian 矩阵提升 $\boldsymbol{Z}$ 的对角化结构，第三项用于 $\boldsymbol{L}_s$ 矩阵的联合学习，最后一项则可以规避相似矩阵的平凡解. 值得注意的是，式(6)为提升运算效率，剔除了 $\boldsymbol{Z}$ 的非负稀疏性限制，却增加了 $\boldsymbol{S}$ 矩阵的非负加和约束. 从 3.2 节可知，$\boldsymbol{S}$ 在迭代过程中通过闭式解更新，效率明显优于非负限制下 $\boldsymbol{Z}$ 的计算过程[24].

众所周知，Laplacian 矩阵的块对角化结构在分簇算法中至关重要. 在式(6)中，目标函数第二项在用于对角性能优化时需要关联矩阵 $\boldsymbol{S}$ 具备更高的对角特性. Feng 等人[16]直接通过对 Laplacian 矩阵进行秩约束以达到此目的，即条件

$$K=\left\{\boldsymbol{Z}\,|\,\mathrm{rank}(\boldsymbol{L}_{\boldsymbol{S}})=n-c,\boldsymbol{S}=\frac{1}{2}(|\boldsymbol{Z}|+|\boldsymbol{Z}^{\mathrm{T}}|)\right\} \tag{7}$$

其中，c 表示对角块个数，等价于目标待分簇数. Feng 等人[16]将上述约束添加至 LRR 得到块对角的 LRR 聚类算法(LRR with Block-diagonal Constraint，LRRBC). 类似地，本文将之添加至式(6)作为 DRC 算法最终的目标函数，即

$$\min_{\boldsymbol{Z},\boldsymbol{S}} \alpha\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{\mathrm{F}}^2+\mu\,\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_s\boldsymbol{Z}^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{X}\boldsymbol{L}_s\boldsymbol{X}^{\mathrm{T}})+\gamma\|\boldsymbol{S}\|_{\mathrm{F}}^2,$$
$$\text{s.t. } \boldsymbol{s}_i\geq 0,\ \boldsymbol{s}_i^{\mathrm{T}}\mathbf{1}=1,\ \mathrm{rank}(\boldsymbol{L}_s)=n-c \tag{8}$$

3.2　模型求解策略

DRC 目标模型(8)包含相似矩阵 $\boldsymbol{S}$ 和系数矩阵 $\boldsymbol{Z}$ 两个优化变量，因此无法直接求取其闭式最优解. 按照交替方向乘子法的思路[32]，本文通过迭代更新策略，按序优化各个目标变量，其中每一个单变量优化过程都是一个凸求解问题.

假设 $\boldsymbol{Z}$ 已知，则 $\boldsymbol{S}$ 的子目标函数为

$$\min_{\boldsymbol{S}} \mu\,\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_s\boldsymbol{Z}^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{X}\boldsymbol{L}_s\boldsymbol{X}^{\mathrm{T}})+\gamma\|\boldsymbol{S}\|_{\mathrm{F}}^2$$
$$\text{s.t. } \forall i\ \boldsymbol{s}_i\geq 0,\ \boldsymbol{s}_i^{\mathrm{T}}\mathbf{1}=1,\ \mathrm{rank}(\boldsymbol{L}_s)=n-c \tag{9}$$

然而，直接采用式(7)作为约束条件将导致式(9)的求解过程非常复杂[16]. 根据定理 1 和命题 1 可知，约束式(7)等价为 $\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_s\boldsymbol{F})$ 最小化，促使 $\boldsymbol{L}_s$ 矩阵具备清晰的对角结构，因此式(9)可进一步改为

$$\min_{\boldsymbol{S}} \mu\,\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_s\boldsymbol{Z}^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{X}\boldsymbol{L}_s\boldsymbol{X}^{\mathrm{T}})+\gamma\|\boldsymbol{S}\|_{\mathrm{F}}^2+\lambda\,\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_s\boldsymbol{F})$$
$$\text{s.t. } \forall i\ \boldsymbol{s}_i\geq 0,\ \boldsymbol{s}_i^{\mathrm{T}}\mathbf{1}=1,\ \boldsymbol{F}\in R^{n\times c},\ \boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}=\boldsymbol{I} \tag{10}$$

定理 1[16,30]. Laplacian 矩阵 $\boldsymbol{L}_s$ 零特征值的个数与关联矩阵 $\boldsymbol{S}$ 的联接块结构数一致.

命题 1. 当 λ 足够大时，正则项 $\lambda\,\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_s\boldsymbol{F})$ 最小化约束等价于 Laplacian 矩阵秩 $\mathrm{rank}(\boldsymbol{L}_s)=n-c$，其中 $\boldsymbol{F}\in R^{n\times c}$ 是由 $\boldsymbol{L}_s$ 最小 c 个特征值对应的特征向量矩阵，亦称为投影矩阵.

证明. 基于 Laplacian 矩阵的半正定性，有 $\sigma_i(\boldsymbol{L}_s)\geq 0$，其中 $\sigma_i(\boldsymbol{L}_s)$ 是 $\boldsymbol{L}_s$ 第 i 个特征值. 因此，当 λ 足够大时，条件 $\mathrm{rank}(\boldsymbol{L}_s)=n-c$ 等价于正则项 $\lambda\sum_{i=1}^{c}\sigma_i(\boldsymbol{L}_s)$ 最小化，即

$$\min_{\boldsymbol{S}} \mu\,\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_s\boldsymbol{Z}^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{X}\boldsymbol{L}_s\boldsymbol{X}^{\mathrm{T}})+\gamma\|\boldsymbol{S}\|_{\mathrm{F}}^2+\lambda\sum_{i=1}^{c}\sigma_i(\boldsymbol{L}_s),$$
$$\text{s.t. } \forall i\ \boldsymbol{s}_i\geq 0,\ \boldsymbol{s}_i^{\mathrm{T}}\mathbf{1}=1 \tag{11}$$

另外，结合 $\boldsymbol{S}$ 连通性与零特征值重根数的关系(定理 1)，式(11)最优解 $\boldsymbol{S}$ 隐含 $\sum_{i=1}^{c}\sigma_i(\boldsymbol{L}_s)=0$，即满足 $\mathrm{rank}(\boldsymbol{L}_s)=n-c$. 最后，根据 Ky Fan 定理[33]描述

$$\sum_{i=1}^{c}\sigma_i(\boldsymbol{L}_s)=\min\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_s\boldsymbol{F}),$$

其中，$\boldsymbol{F}\in R^{n\times c}$ 是正交投影矩阵. 证毕.

对式(10)中的矩阵形式进行约简可得

$$\min\sum_{i,j=1}^{n}\{\mu\|\boldsymbol{z}_i-\boldsymbol{z}_j\|_2^2 s_{ij}+\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2 s_{ij}+$$
$$\lambda\|\boldsymbol{f}_i-\boldsymbol{f}_j\|_2^2 s_{ij}+\gamma s_{ij}^2\}\Leftrightarrow\min\sum_{i=1}^{n}\left\|\boldsymbol{s}_i+\frac{\boldsymbol{g}_i}{2\gamma}\right\|_2^2$$
$$\text{s.t. } \boldsymbol{s}_i^{\mathrm{T}}\mathbf{1}=1,\ 0\leq s_{ij}\leq 1 \tag{12}$$

其中 $\boldsymbol{z}_i\in n\times 1$ 是 $\boldsymbol{Z}$ 矩阵的第 i 列向量，$\boldsymbol{f}_i\in c\times 1$ 是 $\boldsymbol{F}$ 矩阵的第 i 行向量，辅助变量 $\boldsymbol{g}_i$ 的元素 $g_{ij}=\mu\|\boldsymbol{z}_i-\boldsymbol{z}_j\|_2^2+\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2+\lambda\|\boldsymbol{f}_i-\boldsymbol{f}_j\|_2^2$. 从式(12)可见，不同的 $\boldsymbol{s}_i$ 具有独立性，且每个 $\boldsymbol{s}_i$ 的优化过程是一个典型的二次规划求解问题，可通过数值优化技术如牛顿法、有效集法等[34]进行目标变量优化. 为提高效率，本节寻求闭式求解方法，首先给出 $\boldsymbol{s}_i$ 问题的拉格朗日乘子式

$$L(\boldsymbol{s}_i,\eta,\boldsymbol{\xi}_i)=\frac{1}{2}\left\|\boldsymbol{s}_i+\frac{\boldsymbol{g}_i}{2\gamma_i}\right\|_2^2-\eta(\boldsymbol{s}_i^{\mathrm{T}}\mathbf{1}-1)-\boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{s}_i \tag{13}$$

其中 η 和 $\boldsymbol{\xi}_i>0$ 是拉格朗日因子. 对式(13)中的 $\boldsymbol{s}_i$ 进行微分并赋 0 可得初步解为

$$\boldsymbol{s}_i=\left(-\frac{\boldsymbol{g}_i}{2\gamma_i}+\eta\right)_+ \tag{14}$$

其中符号$(\cdot)_+$表示括号内元素值非负.由于相似矩阵$\boldsymbol{S}$在全联接单簇结构时无法得到投影矩阵$\boldsymbol{F}\in R^{n\times c}$.因此,假设相似向量$\boldsymbol{s}_i$具备稀疏性,其非零元素个数为$k$,且$k\ll n$成立,以此为前提将式(14)代入约束条件$\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{1}=1$,可得$\eta=1/k+\left(\sum_{j=1}^{k}g_{ij}\right)/2k\gamma_i$,即$\boldsymbol{s}_i$的求解式为

$$\boldsymbol{s}_i=\left(\frac{2\gamma_i-k\boldsymbol{g}_i+\sum_{j=1}^{k}g_{ij}}{2k\gamma_i}\right)_+ \tag{15}$$

假设相似矩阵$\boldsymbol{S}$已知,其投影矩阵$\boldsymbol{F}$可直接通过$\boldsymbol{L}_s$的谱分解求取,而未知量$\boldsymbol{Z}$的子目标函数为

$$\min_{\boldsymbol{Z}}\{\alpha\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{\mathrm{F}}^2+\mu\,\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_s\boldsymbol{Z}^{\mathrm{T}})\} \tag{16}$$

根据命题2,对式(16)进行简单的代数调整并依$\boldsymbol{Z}$进行微分取零可得

$$\begin{aligned}&\min\{\alpha\,\mathrm{tr}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z})^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z})+\mu\,\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_s\boldsymbol{Z}^{\mathrm{T}})\}\\&\Rightarrow\mathrm{tr}(\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{Z}+\mu\boldsymbol{Z}\boldsymbol{L}_s-\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})=0\\&\Rightarrow\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{Z}+\mu\boldsymbol{Z}\boldsymbol{L}_s=\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\end{aligned} \tag{17}$$

是一个标准的Sylvester公式[35],具有唯一解.

命题2. 假设$\boldsymbol{A}\in\boldsymbol{S}_+^n$、$\boldsymbol{B}\in\boldsymbol{S}_+^n$为对称正半定矩阵,则$\mathrm{tr}(\boldsymbol{A}\boldsymbol{B})=0$等价于$\boldsymbol{A}\boldsymbol{B}=\boldsymbol{0}$.

证明. 根据对称矩阵$\boldsymbol{A}$、$\boldsymbol{B}$的正半定性,存在矩阵$\boldsymbol{P}\in R^{n\times n}$和$\boldsymbol{Q}\in R^{n\times n}$可分别对$\boldsymbol{A}$和$\boldsymbol{B}$进行Cholesky分解,即

$$\boldsymbol{A}=\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}},\ \boldsymbol{B}=\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}.$$

使得

$$\mathrm{tr}(\boldsymbol{A}\boldsymbol{B})=\mathrm{tr}(\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}})=\mathrm{tr}(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q})$$

进一步由Frobenius范数定义$\|\boldsymbol{A}\|_{\mathrm{F}}^2=\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})$,可得

$$\mathrm{tr}(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q})=\|\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q}\|_{\mathrm{F}}^2,$$

再依据范数定义可知$\|\boldsymbol{A}\|=0$当且仅当$\boldsymbol{A}=\boldsymbol{0}$,因此得$\mathrm{tr}(\boldsymbol{A}\boldsymbol{B})=\|\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q}\|_{\mathrm{F}}^2=0\Leftrightarrow\boldsymbol{A}\boldsymbol{B}=0$. 证毕.

综上所述,DRC的优化求解过程如算法1描述.模型DRC的优化结果$\boldsymbol{Z}^*$和$\boldsymbol{S}^*$可以分别构建独立的谱分簇关联矩阵.从定理1可知,当模型(8)的收敛条件达到$\mathrm{rank}(\boldsymbol{L}_s)=n-c$,则相似度矩阵$\boldsymbol{S}$正好是$c$联接结构,无需后续的Kmeans、NCut等操作;当模型迭代至式(8)目标函数值收敛,则可采用$\boldsymbol{Z}$构建关联矩阵,并通过经典谱分簇算法完成数据聚类;当模型(8)达到最高迭代次数退出,将关联矩阵以$\boldsymbol{Z}.*\boldsymbol{S}$为基础设定,联合次优的表示矩阵和相似度矩阵进入后续步骤.此外,式(8)各正则项都具备光滑性,在模型求解中不需要无偏次梯度运算、对角化投影等耗时子过程,因此更为高效易解.

4 参数设定与算法分析

本节讨论DRC算法在具体实施过程中的实现细节,包括运行过程中部分模型参数的设定和收敛条件补充. 此外,对DRC算法的收敛性和复杂度进行了理论分析.

算法1. 数据表示分簇DRC.

输入:样本集$\boldsymbol{X}$,目标簇类c,迭代最大值t_m,模型参数k、α、μ、λ和γ

1. 设$\mu=\lambda=0$,依式(15)初始化相似矩阵$\boldsymbol{S}^0$,并计算Laplacian矩阵$\boldsymbol{L}_s^0$和投影矩阵$\boldsymbol{F}^0$;
2. 迭代次数$t=1$;
3. 固定$\boldsymbol{S}$,依Sylvester公式(17)计算$\boldsymbol{Z}^t$;
4. 固定$\boldsymbol{Z}$,依次按式(15)求解相似向量$\boldsymbol{s}_i$并组建成$\boldsymbol{S}^t$;
5. 更新Laplacian矩阵$\boldsymbol{L}_s^t$并计算投影矩阵$\boldsymbol{F}^t$;
6. 检查$(J_t-J_{t+1})/J_t<1\mathrm{e}{-3}$或$t\geq t_m$是否满足,其中$J_t$为目标函数值(式(8)),如是则终止算法;反之则令$t=t+1$,算法循环至第3步继续进行.

输出:最优化相似矩阵$\boldsymbol{S}^*$和系数矩阵$\boldsymbol{Z}^*$.

4.1 部分参数设定

从第3节DRC算法描述可见,其实施过程包含多个人工设定参数,即k、α、μ、λ和γ.不同参数的具体取值对算法的分簇性能影响较大,而众多参数的优选过程又需要耗费大量的运行时间.因此,对模型参数进行分析并给出具体实施优选范围能极大地提升算法的应用推广性.本节给出DRC算法中k、λ和γ三个参数的实施方案,在后续第5节则通过实验分析参数α和μ的经验取值.

由式(15)可知,相似向量$\boldsymbol{s}_i$的迭代更新值由非零邻域数$k\in(0,n)$以及模型正则化参数$\gamma_i>0$确定.不失一般性,将$\boldsymbol{s}_i$元素值按从大到小排序,即

$$\left.\begin{matrix}s_{ik}>0\\s_{i,k+1}\leq0\end{matrix}\right\}\Rightarrow\left.\begin{matrix}\left(2\gamma_i-kg_{ik}+\sum_{j=1}^{k}g_{ij}\right)/2k\gamma_i>0\\\left(2\gamma_i-kg_{i,k+1}+\sum_{j=1}^{k}g_{ij}\right)/2k\gamma_i\leq0\end{matrix}\right\} \tag{18}$$

从中可知参数γ_i的取值范围

$$\frac{k}{2}g_{ik}-\frac{1}{2}\sum_{j=1}^{k}g_{ij}<\gamma_i\leq\frac{k}{2}g_{i,k+1}-\frac{1}{2}\sum_{j=1}^{k}g_{ij} \tag{19}$$

为获得非负稀疏的$\boldsymbol{s}_i$,取$\gamma_i=\left(kg_{i,k+1}-\sum_{j=1}^{k}g_{ij}\right)/2$,仅保留$\boldsymbol{s}_i$中最大$k$个邻域值,其余元素都为零,则$\boldsymbol{s}_i$由计算公式(15)调整为

$$\boldsymbol{s}_i=\frac{g_{i,k+1}-\boldsymbol{g}_i}{kg_{i,k+1}-\sum_{j=1}^{k}g_{ij}} \tag{20}$$

即在 $\boldsymbol{s}_i$ 的计算中取消模型参数 γ_i 的设置，采用单纯的邻域参数 k 进行最优值获取，而 k 值具有明确的物理含义，一般取值为 3～18 的整数，简化了相似度矩阵求解过程，比直接设置 γ_i 值更直观便捷.

从 3.1 节可知，$\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_s\boldsymbol{F})$ 等价于 $\mathrm{rank}(\boldsymbol{L}_s)=n-c$，因此在相似矩阵 $\boldsymbol{S}$ 的优化过程中，如能使目标式(9)中的 $\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_s\boldsymbol{F})$ 项趋于 0(对应足够大的 λ)，则可得到 c 联接结构的 $\boldsymbol{S}^*$，直接进行数据分簇. 因此可在算法运行中对参数 λ 的取值进行自调整操作，本文将其初始化为 $\lambda=\sum_{i=1}^{n}\gamma_i$，在每次迭代中分别计算 $\nu_1=\sum_{i=1}^{c}\sigma_i(\boldsymbol{L}_s)$ 和 $\nu_2=\sum_{i=1}^{c+1}\sigma_i(\boldsymbol{L}_s)$，当 ν_1 大于极小值 ε 时(本文设 $\varepsilon=1\mathrm{e}-11$)，则令 $\lambda^{t+1}=2\lambda^{t+1}$，使得 $\mathrm{Tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_s\boldsymbol{F})$ 更趋于 0；反之，当 $\nu_2\leqslant\varepsilon$ 则令 $\lambda^{t+1}=\lambda^{t+1}/2$，以减少 $\boldsymbol{L}_s$ 的簇数. 进一步可将 λ 值调整与 DRC 收敛条件绑定，当满足 $\nu_1<\varepsilon$ 且 $\nu_2>\varepsilon$ 时，说明相似矩阵恰好具有 c 联接结构，算法迭代结束，$\boldsymbol{S}$ 可直接作为关联矩阵.

4.2　算法分析

根据 4.1 节描述，相似矩阵 $\boldsymbol{S}$ 在优化过程中趋于 c 联接状态，具有自然的对角结构. 根据命题 3 可知，系数矩阵 $\boldsymbol{Z}$ 亦具有类似的对角结构，保证 DRC 算法后续的谱分簇操作具有良好的性能.

命题 3.　在数据子空间独立且无噪声情形下，式(16)子目标函数得到的最优化 $\boldsymbol{Z}$ 具有对角结构.

证明.　给定转换矩阵 $\boldsymbol{P}$，输入数据 $\boldsymbol{X}$ 对应的 Laplacian 矩阵满足 $\boldsymbol{L}_s(\boldsymbol{XP})=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{L}_s(\boldsymbol{X})\boldsymbol{P}$，则有 $\mathrm{tr}(\boldsymbol{ZL}_s(\boldsymbol{X})\boldsymbol{Z}^{\mathrm{T}})=\mathrm{tr}(\boldsymbol{ZPL}_s(\boldsymbol{X})\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}})$，即式(1)中的 $\Omega(\boldsymbol{X},\boldsymbol{Z})=\Omega(\boldsymbol{XP},\boldsymbol{ZP})$；设 $\boldsymbol{Z}^d=\begin{bmatrix}\boldsymbol{A}&0\\0&\boldsymbol{D}\end{bmatrix}$，其中 $\boldsymbol{A}$ 和 $\boldsymbol{D}$ 取自于 $\boldsymbol{Z}=\begin{bmatrix}\boldsymbol{A}&\boldsymbol{B}\\\boldsymbol{C}&\boldsymbol{D}\end{bmatrix}$，并将 $\boldsymbol{Z}$ 和 $\boldsymbol{Z}^d$ 代入 $\frac{1}{2}\sum_{i,j=1}^{n}\|\boldsymbol{z}_i-\boldsymbol{z}_j\|_2^2 s_{ij}$ 可得 $\Omega(\boldsymbol{X},\boldsymbol{Z})\geqslant\Omega(\boldsymbol{X},\boldsymbol{Z}^d)$ 以及 $\Omega(\boldsymbol{X},\boldsymbol{Z}^d)=\Omega(\boldsymbol{X},\boldsymbol{A})+\Omega(\boldsymbol{X},\boldsymbol{D})$，其中当且仅当 $\boldsymbol{B}=\boldsymbol{C}=0$ 不等式中等号成立. 因此，子目标函数式(16)中的 $\boldsymbol{Z}$ 正则项满足严格块对角条件[36]，根据文献[36]中的定理 2 可知，在数据子空间独立且无噪声情形下，式(16)子目标函数的最优解具有对角结构.　证毕.

由于目标模型(8)是非凸函数，因此针对 DRC 算法各未知变量的交替更新策略缺乏全局最优解理论支撑[27]. 然而，结合一阶最优性条件[34]和命题 4 可保证 DRC 算法优化至稳定解. 具体地，在每步迭代中，特征向量矩阵 $\boldsymbol{F}$ 仅依赖于相似矩阵 $\boldsymbol{S}$，而式(15)是闭式解，因此 $\boldsymbol{S}$ 和 $\boldsymbol{F}$ 的单变量更新操作是唯一最优解. 此外，由于 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 和 Laplacian 矩阵 $\boldsymbol{L}_s$ 都是正半定的，因此两者的特征值非负，进一步推出两者特征值之和非负，在此条件下 $\boldsymbol{Z}$ 的单变量更新操作也是唯一最优的[35]. 综合各变量优化解的唯一性以及式(8)的非负性，命题 4 说明了 DRC 目标函数值随迭代过程逐渐下降并收敛至稳定解. 第 5 节实验部分也表明 DRC 具有快速稳定的收敛性. 在具体实施过程中，为获得算法的全局最优解，可尝试不同的初始化操作. 除算法 1 所描述的方案外，可依 k 近邻或 ε 邻域计算 $\boldsymbol{S}^0$ 矩阵，也可通过常规表示型算法计算系数矩阵 $\boldsymbol{Z}$ 间接初始化 $\boldsymbol{S}^0$.

命题 4.　DRC 算法的目标函数值随迭代过程逐步下降.

证明.　假设第 t 次迭代时得到 $\boldsymbol{S}^t$，则在第 $t+1$ 次迭代中计算得到 $\boldsymbol{F}^{t+1}$ 和 $\boldsymbol{Z}^{t+1}$，根据单变量更新的最优性，不等式

$$\begin{aligned}&\alpha\|\boldsymbol{X}-\boldsymbol{XZ}^t\|_{\mathrm{F}}^2+\mu\,\mathrm{tr}(\boldsymbol{Z}^t\boldsymbol{L}_s\boldsymbol{Z}^{t\mathrm{T}})+\lambda\,\mathrm{tr}(\boldsymbol{F}^{t\mathrm{T}}\boldsymbol{L}_s\boldsymbol{F}^t)\leqslant\\&\alpha\|\boldsymbol{X}-\boldsymbol{XZ}^{t+1}\|_{\mathrm{F}}^2+\mu\,\mathrm{tr}(\boldsymbol{Z}^{t+1}\boldsymbol{L}_s\boldsymbol{Z}^{(t+1)\mathrm{T}})+\\&\lambda\,\mathrm{tr}(\boldsymbol{F}^{(t+1)\mathrm{T}}\boldsymbol{L}_s\boldsymbol{F}^{t+1})\end{aligned} \tag{21}$$

成立，同样在给出 $\boldsymbol{F}^{t+1}$ 和 $\boldsymbol{Z}^{t+1}$ 后更新相似矩阵 $\boldsymbol{S}^{t+1}$，则不等式

$$\begin{aligned}&\mu\,\mathrm{tr}(\boldsymbol{Z}^{t+1}\boldsymbol{L}_s^t(\boldsymbol{Z}^{t+1})^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{XL}_s^t\boldsymbol{X}^{\mathrm{T}})+\gamma\|\boldsymbol{S}^t\|_{\mathrm{F}}^2+\\&\lambda\,\mathrm{tr}((\boldsymbol{F}^{t+1})^{\mathrm{T}}\boldsymbol{L}_s^t\boldsymbol{F}^{t+1})\leqslant\\&\mu\,\mathrm{tr}(\boldsymbol{Z}^{t+1}\boldsymbol{L}_s^{t+1}(\boldsymbol{Z}^{t+1})^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{XL}_s^{t+1}\boldsymbol{X}^{\mathrm{T}})+\\&\gamma\|\boldsymbol{S}^t\|_{\mathrm{F}}^2+\lambda\,\mathrm{tr}((\boldsymbol{F}^{t+1})^{\mathrm{T}}\boldsymbol{L}_s^{t+1}\boldsymbol{F}^{t+1})\end{aligned} \tag{22}$$

成立. 将式(21)和(22)两者联合，则目标函数(8)随迭代进程逐渐下降.　证毕.

在运算量方面，从算法 1 可知，DRC 的关键操作包括式(17)、式(20)和谱分解三个步骤，分别对应 $\boldsymbol{Z}$、$\boldsymbol{S}$ 和 $\boldsymbol{F}$ 三者的单步优化操作. Sylvester 等式求解的典型算法是 Bartel-Stewart 法，先通过 QR 分解将系数矩阵变换成 Schur 形式，再采用回代法求解三角系统，整体运算复杂度是 $O(n^3)$；相似向量依式(20)求解，其复杂度为 $O(n)$，则 $\boldsymbol{S}$ 的整体复杂度是 $O(n^2)$；加上谱分解的复杂度 $O(n^3)$，则 DRC 算法单次迭代复杂度为 $O(n^3)$. 假设算法收敛时迭代次数为 t，则 DRC 的总复杂度为 $O(n^3t)$.

5　实验结果

分别采用多个人工合成数据和实际公开数据集

验证所提 DRC 的分簇应用能力. 所设计的实验包括关联矩阵对角结构显示、分簇准确度、参数敏感性和算法运行效率 4 种不同类型. 除可视化分簇效果和结构化展示外,实验通过准确性(Accuracy,ACC)和归一化互信息(Normalized Mutual Information,NMI)两种指标对比所有算法的分簇结果. *ACC* 的计算公式为

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

其中 l_i 是给定样本 $\boldsymbol{x}_i$ 的真实簇类别标签,r_i 是算法计算所得簇类别标签;当 $l_i=r_i$ 时,函数 $\delta(l_i,r_i)$ 取值 1,反之取值为 0;$\mathrm{map}(r_i)$ 是置换映射函数,用于将簇标签 r_i 置换为同等的数据集类标签[37]. 此外,互信息 *MI* 的计算公式为

$$MI(C,C')=\sum_{c_i\in C,c_j'\in C'}p(c_i,c_j')\log_2\frac{p(c_i,c_j')}{p(c_i)p(c_j')},$$

其中,C 是所有样本真实的簇标签集,C' 是算法计算得到的簇标签集,$p(c_i)$ 和 $p(c_j')$ 分别表示任选样本分别隶属于 c_i 和 c_j' 的概率,而 $p(c_i,c_j')$ 则是任意样本同时属于 c_i 和 c_j' 的联合概率. 进一步,*NMI* 的定义为

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

其中,$H(C)$ 和 $H(C')$ 分别是集合 C 和 C' 的熵. *NMI* 是一种典型的相似度度量指标,其值为[0,1],当两个簇标签集合一致时,*NMI* 值取 1,反之则取 0.

5.1 合成示例

首先通过人工合成的 5 簇数据验证 DRC 算法的关联矩阵对角化能力以及抗噪效果. 以原点为中心采用 $N(0,1)$ 高斯分布随机生成 5 个 $m=250$ 且 $n=25$ 的数据,以 0.2 为偏移构建 5 簇独立子空间样本,实验过程中随机选择其中的 $p\%$ 个特征样本叠加 $N(0,1)$ 的高斯噪声. 图 1 显示了不同谱分簇算法所生成的关联邻域图,包括 SMR[25]、RSS[27]、LRS[17] 和 DRC,其中图 1(a)至图 1(c)分别是噪声干扰 $p=\{0,30,60\}$ 下的结果,所有算法的参数都通过网格式搜索以最优对角结构为目标确定. 从图 1(a)可见,不添加噪声干扰时,选取的 4 种算法都生成了具有明显对角结构的关联矩阵,表 2 的性能指标也验证了几种算法具有极高的分簇能力. 然而,从图 1(a)仍可以看出不同算法间的差异,SMR 和 RSS 的簇间相似度具有非零值,因此不具备绝对对角结构,在噪声环境中将影响算法性能;LRS 关联矩阵的对角结构非常清晰,然而其簇内相似度具有恒值,说明 LRS 缺乏簇内样本的局部状态描述,在多模态分布的数据中难以呈现较好的效果;经比较

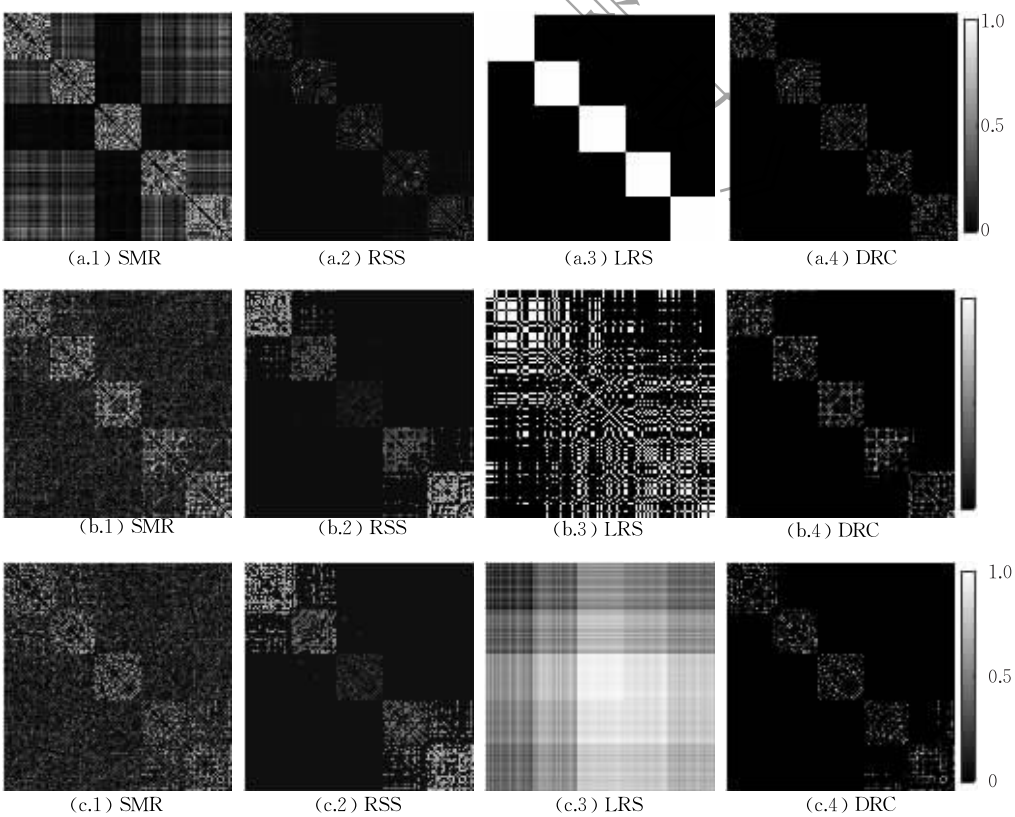


图 1 合成 5 簇数据在不同噪声等级下的相似性结构

可见，DRC 的簇间相似度为绝对零值，而簇内相似度则具有丰富的权值变化，更适合于复杂的应用任务. 从图 1(b)和图 1(c)可见，在噪声环境下，DRC 的关联矩阵具有更为明显的 5 簇对角结构，而 SMR、RSS 和 LRS 的关联矩阵则都受到了严重的影响. 尤其是 LRS，很难发现其相似度矩阵的对角结构. 值得注意的是，RSS 算法的簇结构在噪声环境下又显现出另一缺陷，即不同簇具有不同的相似度均值，不符合一般分簇算法中样本独立同分布的前提条件，将影响其分簇效果. 表 2 的数值结果也验证了 DRC 具有更高的鲁棒性，且 LRS 的分簇性能在噪声环境中下降非常明显.

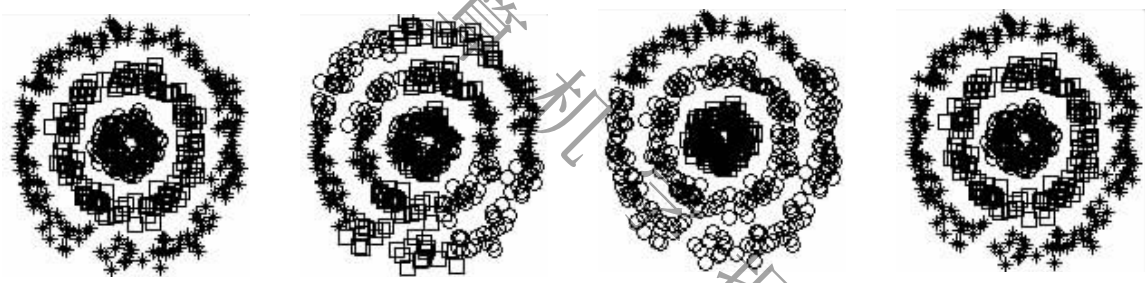
表 2　不同算法在合成 5 簇数据中的性能指标对比

Methods	*ACC*			*NMI*		
	$p=0$	$p=30$	$p=60$	$p=0$	$p=30$	$p=60$
SMR	99.22	97.21	76.05	97.91	94.30	62.10
RSS	99.22	97.60	85.15	96.87	95.02	78.47
LRS	100.00	47.40	40.50	100.00	66.86	55.76
DRC	100.00	98.40	95.20	100.00	96.50	91.41

表 3　不同算法在 3 环数据和双月数据中的分簇性能指标

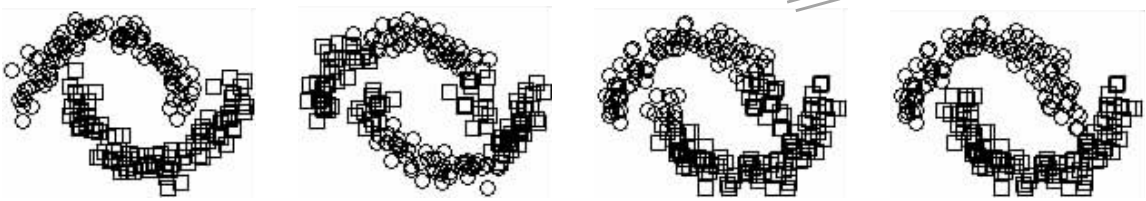
Methods	three-ring		two-moon	
	ACC	*NMI*	*ACC*	*NMI*
SMR	37.60	18.90	51.55	45.68
RSS	71.60	50.64	74.00	51.85
DRC	100.00	100.00	100.00	100.00

进一步采用人工合成的 3 环数据和双月数据验证 DRC 的分簇能力. 3 环数据由图 2(a)所示，随机产生 500 个样本，其中内环和中环都是 150 个，外环 200 个. 双月数据由图 3(a)所示，随机产生 200 个样本，每个半月 100 个，将两个半月对面交叉放置. 图 2 和图 3 给出了 SMR、RSS 和 DRC 的分簇结果，表 3 则从数值上给出了算法的分簇指标. 从中可见，SMR 和 RSS 在这两种非常规分布的数据中表现欠佳，分簇结果与原始数据差异较大，分簇精度和相似度指标较低，而 DRC 算法对两种数据集都能够进行正确分簇，在 3 环数据和双月数据集中的分簇指标都是 100%，展现了优秀的分簇能力.



(a) Original　(b) SMR　(c) RSS　(d) DRC

图 2　SMR、RSS 和 DRC 在三环合成数据中的分簇效果对比



(a) Original　(b) SMR　(c) RSS　(d) DRC

图 3　SMR、RSS 和 DRC 在双半环合成数据中的分簇效果对比

5.2　公开数据集实验

采用 8 个公开数据集验证 DRC 的实际分簇应用性能，包含 JAFFE、ORL、AR 三个人脸数据集，ISOLET 语音数据集，LUNG 生物数据集，COIL 对象数据集以及 2 个手写字体数据集 USPS 和 MNIST. 表 4 列出了实验中各数据集的细节描述，为体现样本的多样性以及便于横向文献对比，将部分数据集的簇数进行修改，例如 USPS 与 MNIST 两个手写字体数据集都是 10 分簇结构，因此筛选 USPS 中的 2、5、8 三个数字作为分簇应用.

表 4　数据集描述汇总

描述特征	样本量	特征维度	目标簇数
USPS	2913	256	3
LUNG	203	3312	5
JAFFE	213	676	10
MNIST	2000	784	10
COIL	1440	1024	20
ISOLET	1559	617	26
ORL	400	1024	40
AR	700	792	100

对比算法选用经典的 Kmeans 以及最新报道的 FSASL[18]，NSLLRR[24]，DEC[7]，SMR[25]，LRS[17]、RSS[27]、LRR[11]、LRRBC[16] 和 FSC[21] 等方法，其中 FSASL 和 DEC 分别通过特征选择和特征提取策略挖掘输入样本的显著性特征，最后再采用 Kmeans 方法进行数据分簇. LRS 引入 Schatten p 范数和分组秩最小化进行最优子空间搜索，并通过簇指示矩阵获得最终的聚类结果. 其余 6 种算法都以重构表示为基础，采用不同的系数约束或重构范数制定分簇目标规则用以构建关联矩阵，最终采用 NCut 或 RatioCut 等算法完成谱分簇任务. 在实验过程中，所有对比算法的待选参数都通过网格式搜索寻优，具体范围如表 5 所示，其中参数符号都与各算法文献保持一致. 针对含义相似的参数，其搜索范围保持一致，而不同含义的参数搜索范围也选用相同的网格个数，便于横向算法比较的公平公正. 表 6 和表 7 分别列出了所有竞争算法在各数据集中的精度 *ACC* 和归一化互信息 *NMI* 两个实验结果，所有数值都通过 20 次随机初始化实验并取平均获得，表中每一数据集的前 2 名算法都进行了排序标注，最优值为黑斜体、次优值为黑体. 从表 6 和表 7 可以得出如下实验结论：(1) 没有一个算法能够在所有数据集中都占据最高的精度和归一化互信息指标，依据输入样本的先验分布信息和目标任务的特殊要求进行具体合理的分簇实现仍是一个值得探索的公知问题；(2) 以低秩表示为基础的几种算法较 Kmeans、FSASL、DEC、LRS 具有更优的分簇结果，几乎占据了所有数据集的最优精度和归一化互信息值，仅 FSASL 在 ISOLET 数据集中获得了 1 个次优 *NMI* 值；(3) DRC 具有明显的性能优势. 在 8 个数据集的 *ACC* 和 *NMI* 对比中，DRC 共获得了 10 个最优值和 5 个次优值. 其余算法中，NSLLRR 有 3 个最优值，RSS 和 LRRBC 分别有 2 个最优值和 2 个次优值，FSC 和 DEC 则分别有 3 和 2 个次优值，综合性

表 5 对比算法中可调参数的搜索范围

Methods	Setting of all tunable parameters
FSASL	$k\in\{3,6,9,12,15\}$; $\alpha,\beta\in\{10^{-2},10^{-1},\cdots,10^{2}\}$; $\gamma\in\{0.001,0.005,0.01,0.05,0.1\}$
NSLLRR	$k\in\{3,6,9,12,15\}$; $\lambda,\gamma,\beta\in\{10^{-2},10^{-1},\cdots,10^{2}\}$
DEC	$k\in\{3,6,9,12,15\}$; $\lambda\in\{10^{-2},10^{-1},\cdots,10^{2}\}$
SMR	$k\in\{3,6,9,12,15\}$; $\alpha,\gamma\in\{10^{-2},10^{-1},\cdots,10^{2}\}$
LRS	$k\in\{3,6,9,12,15\}$; $p\in\{0.2,0.4,0.6,0.8,1\}$
RSS	$k\in\{3,6,9,12,15\}$; $\lambda_1,\lambda_2,\lambda_3\in\{10^{-2},10^{-1},\cdots,10^{2}\}$
LRR	$\lambda\in\{10^{-2},10^{-1},\cdots,10^{2}\}$; $\rho\in\{1.0,1.1,1.2,1.3,1.4\}$
LRRBC	$\lambda,\beta\in\{10^{-2},10^{-1},\cdots,10^{2}\}$; $\rho\in\{1.0,1.1,1.2,1.3,1.4\}$
FSC	$\lambda\in\{10^{-2},10^{-1},\cdots,10^{2}\}$; $M\in\lceil(\{0.2,0.3,0.4,0.5,0.6\}\times m)\rceil$
DRC	$k\in\{3,6,9,12,15\}$; $\alpha,\mu\in\{10^{-2},10^{-1},\cdots,10^{2}\}$

表 6 所有算法在不同数据集下的分簇准确度对比

Method	USPS	JAFFE	ISOLET	MNIST	LUNG	COIL	ORL	AR
Kmeans	91.69	71.57	49.64	54.16	78.38	59.17	51.79	26.86
FSASL	96.68	78.76	57.47	57.46	90.14	66.67	54.75	33.29
NSLLRR	94.79	98.59	***58.94***	51.90	***90.64***	62.01	65.35	51.00
DEC	95.91	95.31	58.56	**62.38**	83.74	73.06	62.25	32.43
SMR	97.05	97.34	52.80	62.35	89.16	70.21	72.05	60.99
LRS	78.72	71.83	26.62	35.43	80.30	49.72	53.25	43.57
RSS	***98.80***	33.33	45.86	51.71	80.30	**82.92**	21.25	18.71
LRR	95.23	97.66	54.78	53.95	73.89	66.94	66.75	60.57
LRRBC	95.26	***99.53***	56.96	54.00	73.89	67.92	70.00	**62.00**
FSC	97.52	96.71	49.74	55.15	87.19	58.96	**72.80**	61.71
DRC	**97.63**	***99.53***	**58.63**	***67.20***	**90.15**	***86.74***	***73.77***	***65.29***

表 7 所有算法在不同数据集下的归一化互信息对比

Method	USPS	JAFFE	ISOLET	MNIST	LUNG	COIL	ORL	AR
Kmeans	80.21	81.52	70.00	50.84	60.37	75.58	74.26	65.37
FSASL	85.88	88.29	***75.64***	52.95	74.76	79.20	75.88	69.93
NSLLRR	87.90	97.81	70.39	52.90	***76.12***	72.47	79.86	76.37
DEC	83.70	94.20	**73.48**	53.84	68.25	81.05	79.55	67.46
SMR	87.24	97.20	68.91	61.80	73.57	79.51	84.91	79.60
LRS	69.72	78.33	38.98	27.59	58.24	56.66	75.12	74.68
RSS	***93.66***	27.20	64.58	55.81	54.27	**94.52**	41.76	53.25
LRR	81.68	97.45	69.86	55.31	50.55	77.28	82.58	80.12
LRRBC	81.76	***99.18***	69.87	55.37	50.55	78.14	82.98	**82.20**
FSC	88.81	97.10	58.87	**63.13**	72.31	68.57	**86.03**	80.18
DRC	**89.19**	***99.18***	73.21	***65.33***	**74.82**	***95.65***	***87.15***	***82.86***

能远远弱于 DRC. 在 8 个数据集的平均 *NMI* 指标中，DRC 达到 83.42%，而次优算法仅为 79.9%；(4) DRC 算法在不同数据集中具有良好的适应性，虽然在 ISOLET 的 *NMI* 指标中没有获得前 2 名的结果，但其值 73.21%较次优值 73.48%仅落后 0.27%. RSS 在 USPS 和 COIL 两个数据集中的性能较为突出，但在 JAFFE 和 ORL 中的 *ACC* 指标分别只有 33.33%和 21.25%，远远低于其余对比算法. NSLLRR 和 FSC 则分别在 MNIST 和 COIL 中的精度指标低于基准模型 Kmeans.

为进一步显示 DRC 各约束项的贡献，并评估联合关联矩阵学习和秩约束的优势，本节进一步将 DRC 的 3 个减化版本，即无 $\boldsymbol{S}$ 版(NS)、无 $\boldsymbol{Z}$ 版(NZ)、以及 $\boldsymbol{S}$ 和 $\boldsymbol{Z}$ 独立优化版 Ind，与原始版 Ori 进行对比，实验结果如图 4 所示. NS 版是指直接以式(16)作为终极目标函数，由最优化 $\boldsymbol{Z}$ 计算关联矩阵；NZ 版以式(10)子目标函数优化 $\boldsymbol{S}$ 并构建关联矩阵；Ind 版则先通过式(10)求解得到 $\boldsymbol{S}$，将相应的 Laplacian 矩阵输入式(16)计算最终的系数矩阵 $\boldsymbol{Z}$. 从图 4 可知，原始 DRC 版本的性能较各减化版有明显的优势，其中 NS 和 NZ 在不同的数据集中表现各有胜负，符合目前主流的表示型和近邻型两类谱分簇算法现状，也为本文联合两者构建新的算法提供了可行性支撑. Ind 版虽然结合了两者的结构化性质，但缺乏学习过程，未能达到整体模型的最优性，其性能弱于 Ori 版. 值得注意的是，在 USPS、MNIST、COIL 三个数据集中，Ind 版与 Ori 版的性能较为接近，通过 5.4 节实验可知，DRC 在这几个数据集中仅通过 2 次迭代即可达到收敛，说明 Ind 版接近于最优解.

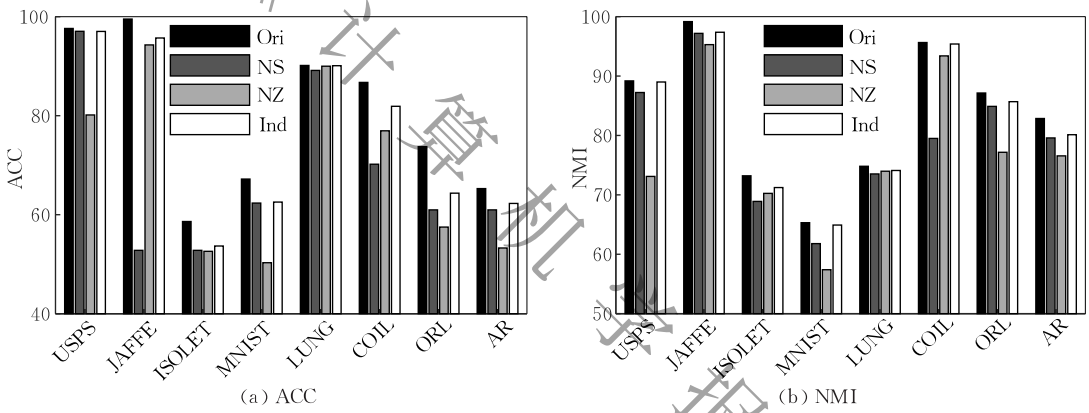


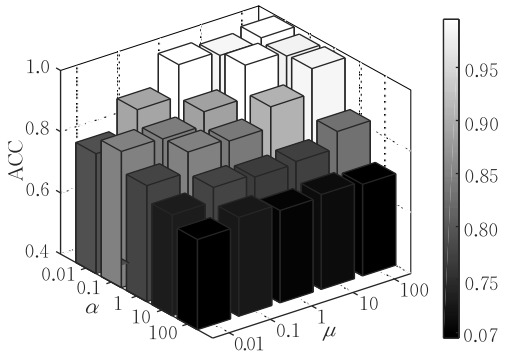
图 4　DRC 各简化版本的聚类性能对比

5.3　参数敏感性分析

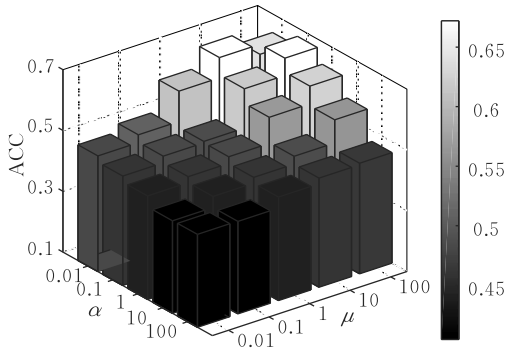
根据表 5 所示，所有分簇方法都有个数不一的待选参数需要人工搜索寻优，其具体选值对分簇效果有一定影响. 因此，所提算法中待设参数的个数以及算法性能对参数具体取值的敏感度是衡量算法应用能力的重要指标. 在表 5 中，FSASL、NSLLRR、RSS 三种算法有 4 个人工待设参数，SMR、LRRBC 和 DRC 有 3 个待设参数，而余下的 DEC、LRS、LRR 和 FSC 则只有 2 个待设参数. 如果能够进一步将 DRC 的未知参数缩减为 2 个，将极大地提升其实施便捷性. 在 3 个未知参数中，邻域数 k 依具体应用任务而定，存在于大部分算法中，如何对其进行优化确定仍是机器学习领域的公知问题. 此外，DRC 还有 α 和 μ 两个正则项参数. 以 JAFFE、MNIST、COIL 和 AR 四个数据集为例，图 5 给出了 DRC 算法在 α 和 μ 参数不同选值下的分簇精度变化. 从图 5 可见，在不同数据集下 DRC 随着 α 或 μ 变化，其分簇精度也具有较大的变化，即 DRC 对单个正则项参数变化的敏感度较高. 然而，深入图 5 各子图 *ACC* 取值可知，各数据集下的最优实验结果都出现在 α 参数值近似于 μ 参数值的 100 倍时. 例如，在 JAFFE 中，当 $\alpha=1$ 而 $\mu=100$ 时，*ACC*=99.53%为最优值；相同的参数选择下，DRC 在 AR 中也达到最优精度值 65.29%. 基于此，可减少 DRC 的待设参数至 2 个，提升其应用效率. 为横向比较，图 6 展示了 FSC 算法在相同数据集中的参数搜索结果，其分簇指标在不同参数值下的变化极为突兀，较 DRC 明显具有更强的敏感性.

5.4　运行效率分析

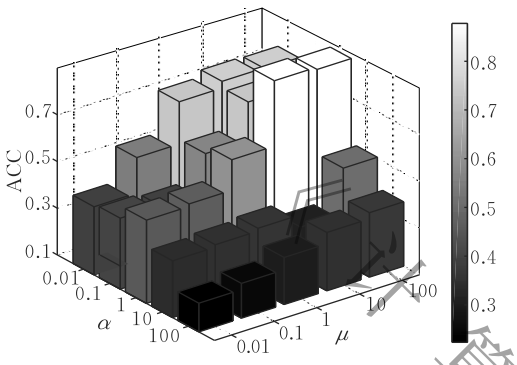
运行效率是除分簇精度、参数敏感性之外的另一重要指标，对分簇算法的用户接纳度至关重要. 表 8 以秒为单位展示了几种算法在各数据集中的单次运



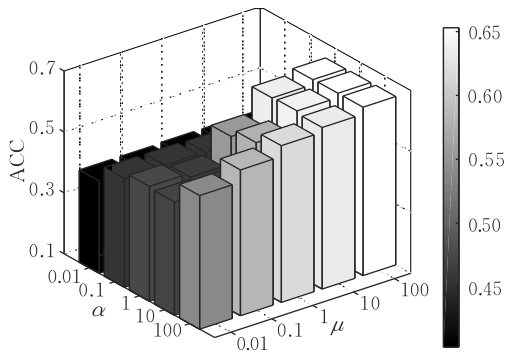
(a) JAFFE



(b) MNIST

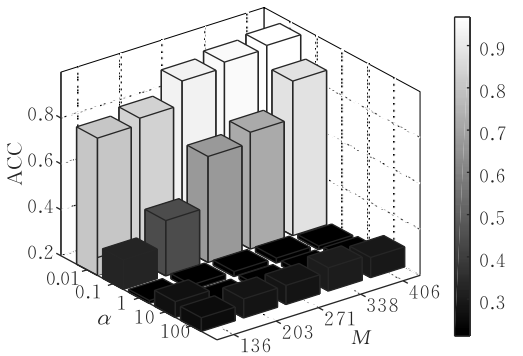


(c) COIL

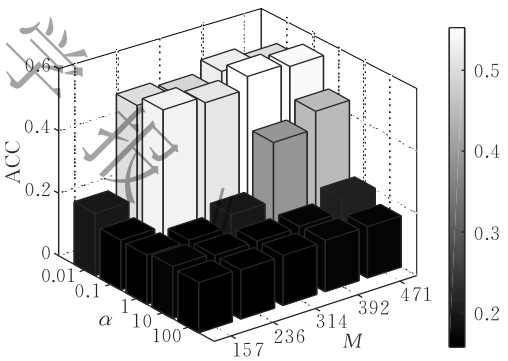


(d) AR

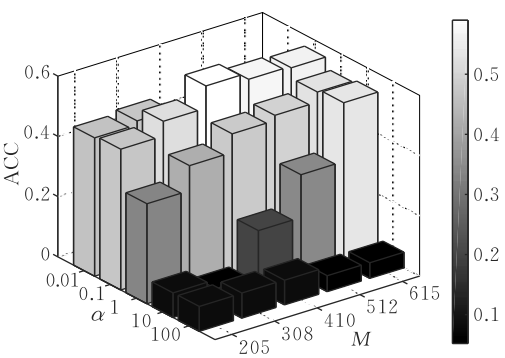
图 5 不同参数选值下的 DRC 分簇精度对比



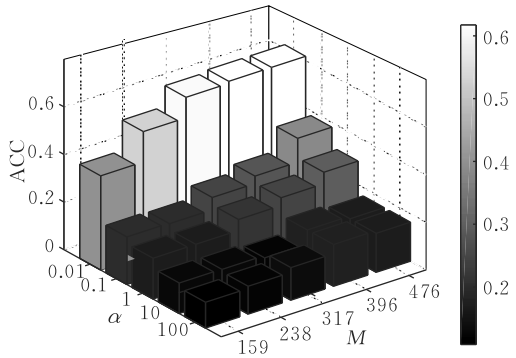
(a) JAFFE



(b) MNIST



(c) COIL



(d) AR

图 6 不同参数选值下的 FSC 分簇精度对比

行时间，所有数值均由 20 次运行结果取平均值得到，实验在 Intel Core i7 CPU，双核主频 2.40 GHz，内存 8 GB，Win7 操作系统以及 Matlab2015b 软件平台上完成. 为综合考虑分簇性能指标，所选择的对比算法都在表 6 或表 7 中获得过黑体值，即在 8 个数据集中有过排名前 2 的性能指标. 从表 8 可见，NSLLRR 和 LRRBC 两者的运行效率明显低于其他算法，尤其是 LRRBC 算法，在大部分数据集中都需要耗费更多的运算时间，虽然其分簇性能优于 LRR，但以运行效率为代价，不宜于实际聚类应用. DEC 的运行效率受输入数据维数的影响非常严重，在 LUNG、ORL 等高 m 低 n 样本集中的效率弱于其它对比算法. RSS 的单步迭代运算量与 DRC 一致，但其收敛速度略慢于 DRC，因此整体运行效率略低. FSASL 算法受样本维数影响较大，在 LUNG 和 ORL 数据集中的运行效率低于 DRC，在余下数据集中则与 DRC 相近. 值得强调的是，结合表 5 可知，FSASL 和 RSS 都需要人工搜索 4 个待定参数，而 DRC 仅 2 个待设参数，因此，DRC 的整体实施效率较 FSASL 和 RSS 更高.

表 8　所有算法在各数据集中的运行效率对比

Method	USPS	JAFFE	ISOLET	MNIST	LUNG	COIL	ORL	AR
FSASL	172.88	3.36	40.35	98.32	93.89	10.38	24.97	19.78
NSLLRR	9.577e3	36.27	2.43e3	7.605e3	101.83	2.64e3	105.32	433.12
DEC	200.41	23.45	288.43	637.37	3.64e3	362.71	237.17	224.79
RSS	251.96	0.76	53.92	100.41	16.53	42.70	4.041	26.98
LRRBC	3.689e4	38.55	5.048e3	9.361e3	56.15	4.59e3	113.45	647.88
FSC	1.022e3	0.74	190.30	867.88	4.725	150.27	6.489	15.67
DRC	235.99	0.358	45.55	89.72	2.69	36.36	4.072	15.46

此外，为验证 DRC 算法的收敛性，图 7 给出了其各个子目标函数值在 8 个数据集中的运行变化情况，其中 F 项是指 $\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_s\boldsymbol{F})$ 正则项，S 项是指相似度矩阵 $\boldsymbol{S}$ 的子目标函数式(10)，Z 项是指系数矩阵 $\boldsymbol{Z}$ 的子目标函数式(16). 为便于显示，其中各曲线值都在 log 域给出. 从图 7 可见，由于各子目标函数都具有唯一闭式解，模型收敛速度很快，其中 USPS、MNIST 和 COIL 三个数据集都只需要 2 次迭代运算即满足了 DRC 的停止条件，收敛速度最慢的 ORL 数据集也只需要 11 次迭代运算. 具体收敛结果中，数据集 LUNG、COIL 和 ORL 在满足 $\mathrm{rank}(\boldsymbol{L}_s)=n-c$ 时结束运算，即达到 $\nu_1<\varepsilon$ 且 $\nu_2>\varepsilon$，而其余算法则在目标函数值收敛时结束，即达到 $(J_t-J_{t+1})/J_t<1\mathrm{e}{-3}$，符合 3.1 节以及 4.1 节收敛条件描述.

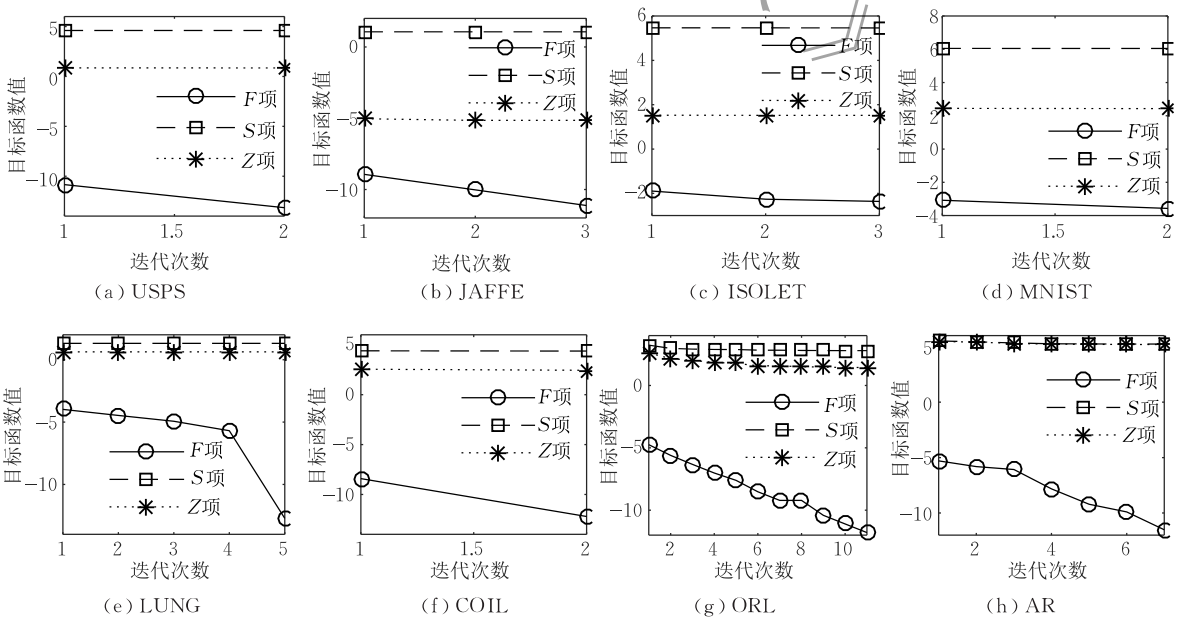


图 7　DRC 目标函数各优化子项在不同数据集中的收敛性

6 总 结

以低秩表示模型为基础，本文提出一种新的数据表示型谱分簇算法 DRC，包括邻域自适应学习和显式矩阵簇结构约束两个关键技术. 首先，目标函数包含基本的数据表示项和相似矩阵联合项，兼顾数据全局结构挖掘以及局部邻域保持特性；其次，在目标函数中添加相似矩阵的自适应学习正则项，在非负性的基础上隐含稀疏特性，符合目标簇联接结构要求；最后，通过特征分解操作增加显式的矩阵秩约束，对结果关联矩阵的块对角结构起关键作用. 此外，通过迭代交替操作给出了模型的优化求解策略，保证各变量优化子步骤都具有全局最优解，并对 DRC 算法的模型参数设定、复杂度和收敛性进行了理论分析. 多个人工合成数据和公开数据集的实验表明，所提的 DRC 分簇算法在正确率、归一化互信息、参数敏感性以及模型实施效率等综合性能上明显优于现存分簇算法.

在算法实施过程中发现，DRC 分簇模型存在参数较多的缺陷，虽然经过参数设定分析给出了部分参数的设置技巧，但仍然会影响具体算法运行过程的效率，并进一步抑制 DRC 算法的应用扩展性. 因此，后续将主要集中于参数优化工作. 此外，数据分簇任务中的目标簇确定问题仍是一个公知问题，有待进一步研究探索.

参 考 文 献

[1] Liang P, Wongthanavasu S. Hybrid linear matrix factorization for topic-coherent terms clustering. Expert Systems with Applications, 2016, 62(11): 358-372

[2] Ma Y, Yang A, Derksen H, et al. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. SIAM Review, 2008, 50(3): 413-458

[3] Ma Y, Derksen H, Hong W, et al. Segmentation of multivariate mixed data via lossy coding and compression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(9): 1546-1562

[4] Ahn I, Kim C. Face and hair region labeling using semi-supervised spectral clustering-based multiple segmentations. IEEE Transactions on Multimedia, 2016, 18(7): 1414-1421

[5] Luo J J, Jiao L C, Lozano J A. A sparse spectral clustering framework via multi-objective evolutionary algorithm. IEEE Transactions on Evolutionary Computation, 2016, 20(3): 418-433

[6] Nie F P, Wang X Q, Huang H. Clustering and projected clustering with adaptive neighbors//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 977-986

[7] Hou C P, Nie F P, Yi D Y, et al. Discriminative embedded clustering: A framework for grouping high-dimensional data. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(6): 1287-1299

[8] Cheng B, Yang J C, Yan S C, Fu Y, et al. Learning with-graph for image analysis. IEEE Transactions on Image Processing, 2010, 19(4): 858-866

[9] Huang J, Nie F P, Huang H. A new simplex sparse learning model to measure data similarity for clustering//Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 3569-3575

[10] Elhamifar E, Vidal R. Sparse subspace clustering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 2790-2797

[11] Liu G C, Lin Z C, Yan S C, et al. Robust recovery of subspace structures by low-rank representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 171-184

[12] Xu Y, Wu Z B, Li J, et al. Anomaly detection in hyperspectral images based on low-rank and sparse representation. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(4): 1990-2000

[13] He Y J, Li M, Zhang J L, et al. Infrared target tracking based on robust low-rank sparse learning. IEEE Geoscience and Remote Sensing Letters, 2016, 13(2): 232-236

[14] Xiao S J, Xu D, Wu J X. Automatic face naming by learning discriminative affinity matrices from weakly labeled images. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(10): 2440-2452

[15] Chen Xiao-Xuan, Qi Chun. Single-image super-resolution via low-rank matrix recovery and joint learning. Chinese Journal of Computers, 2014, 37(6): 1372-1379(in Chinese)
(陈晓璇，齐春. 基于低秩矩阵恢复和联合学习的图像超分辨率重建. 计算机学报，2014，37(6)：1372-1379)

[16] Feng J S, Lin Z C, Xu H, et al. Robust subspace segmentation with block-diagonal prior//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 3818-3825

[17] Nie F P, Huang H. Subspace clustering via new low-rank model with discrete group structure constraint//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA, 2016: 1874-1880

[18] Du L, Shen Y D. Unsupervised feature selection with adaptive structure learning//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015: 209-218

[19] Shikkenawis G, Mitra S K. 2D orthogonal locality preserving projection for image denoising. IEEE Transactions on Image Processing, 2016, 25(1): 262-273

[20] Weng L, Dornaika F, Jin Z. Flexible constrained sparsity preserving embedding. Pattern Recognition, 2016, 60(12): 813-823

[21] Peng C, Kang Z, Yang M, et al. Feature selection embedded subspace clustering. IEEE Signal Processing Letters, 2016, 23(7): 1018-1022

[22] Li Bo, Lu Chun-Yuan, Leng Cheng-Cai, et al. Robust low rank subspace clustering based on local graph laplace constraint. Acta Automatica Sinica, 2015, 41(11): 1971-1980(in Chinese)
(李波，卢春园，冷成财等. 基于局部图拉普拉斯约束的鲁棒低秩表示聚类方法. 自动化学报，2015，41(11)：1971-1980)

[23] Yin M, Gao J B, Lin Z C, et al. Dual graph regularized latent low-rank representation for subspace clustering. IEEE Transactions on Image Processing, 2015, 24(12): 4918-4933

[24] Yin M, Gao J B, Lin Z C. Laplacian regularized low-rank representation and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(3): 504-517

[25] Hu H, Lin Z C, Feng J J, et al. Smooth representation clustering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 3818-3825

[26] Liu G C, Yan S C. Latent low-rank representation for subspace segmentation and feature extraction//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 1615-1622

[27] Guo X J. Robust subspace segmentation by simultaneously learning data representations and their affinity matrix//Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 3547-3553

[28] Zheng J W, Yang P, Chen S Y, et al. Iterative re-constrained group sparse face recognition with adaptive weights learning. IEEE Transactions on Image Processing, 2017, 26(5): 2408-2423

[29] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905

[30] Luxburg U V. A tutorial on spectral clustering. Statistics and Computing, 2007, 17(4): 395-416

[31] Algamal Z Y, Lee M H. Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. Computers in Biology and Medicine, 2015, 67: 136-145

[32] Lin Z C, Huang Y M. Fast multidimensional ellipsoid-specific fitting by alternating direction method of multipliers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(5): 1021-1026

[33] Fan K. On a theorem of Weyl concerning eigenvalues of linear transformations I. Proceedings of the National Academy of Sciences of the United States of America, 1949, 35(11): 652-655

[34] Nocedal J, Wright S. Numerical Optimization. Series in Operations Research and Financial Engineering, New York, USA: Springer, 2006

[35] Lancaster P. Explicit solutions of linear matrix equations. SIAM Review, 1970, 12(4): 544-566

[36] Lu C Y, Min H, Zhao Z Q, et al. Robust and efficient subspace segmentation via least squares regression//Proceedings of the 12th European Conference on Computer Vision (ECCV). Heidelberg, German, 2012: 347-360

[37] Zhao Z, He X, Cai D, et al. Graph regularized feature selection with data reconstruction. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(3): 689-700

ZHENG Jian-Wei, born in 1982, Ph. D., associate professor. His research interests include machine learning and pattern recognition.

JU Zhen-Yu, born in 1993, M. S. His research interests include data mining, artificial intelligence.

ZHU Wen-Bo, born in 1990, M. S. His research interests include data mining, artificial intelligence.

WANG Wan-Liang, born in 1957, Ph. D., professor. His research interests include artificial intelligence, big data analysis.

Background

Low-Rank Representation (LRR), as a promising method to capture the underlying low-dimensional structures of data, has attracted great interest in the pattern analysis and signal processing communities. Specifically, the problems involving the estimation of low-rank matrices have drawn considerable attention in recent years. LRR has been widely used in subspace segmentation, image destriping, image clustering and video background/foreground separation.

The LRR method focuses on low rank data representation based on the hypothesis that the data is approximately jointly spanned by several low-dimensional subspaces and it also takes care of largely contaminated outliers by incorporating l_1 noise models, thus LRR can accurately recover the row space of the original data and detect outliers under mild conditions. In general, the resulting problem becomes a convex optimization problem aiming at the optimal solution of minimizing a combination of the nuclear norm and the l_1-norm in polynomial time.

The performance of LRR based clustering approaches heavily depends on learned data affinity matrices, which are usually constructed either directly from the raw data or from their computed representations. However, these methods may not guarantee an overall optimum since data representation and similarity measurement are often conducted in two independent steps. To improve LRR in this regard, this research proposes a new LRR based Data Representation Clustering (DRC) method with a block diagonal affinity matrix in noiseless data. In our model, we adaptively learn the similarity measurement and fit it into the general low-rank representation framework. Meanwhile, an explicit rank constraint is imposed on the Laplacian matrix of the affinity matrix, leading to an exact cluster structure of connected components. We derive an efficient algorithm to optimize the proposed problem and show the theoretical analysis of convergence and complexity. Experiments are conducted on synthetic data and eight publicly available datasets to demonstrate that DRC outperforms the state-of-the-art approaches in clustering.

This research is supported by the National Natural Science Foundation of China with Grant No. 61873240, the Natural Science Foundation of Zhejiang Province under Grant No. LY19F030016, particularly the National Natural Science Foundation of China with Grant No. 61602413 named as "Robust Nonconvex Nonsmooth Sparse Representation Classification via Adaptive Feature Learning". The main purpose of this project is to develop a general framework of robust data representation with adaptive feature weights learning and nonconvex nonsmooth prior constraints. The DRC model plays an important role in spectral clustering system, which belongs to the general representation framework of this project.