doi：10.13413/j.cnki.jdxblxb.2025013

基于自动权重的主动块对角子空间聚类

李向利[1,2,3]，谢腾翅[1]，韦嘉逢[1]

（1. 桂林电子科技大学 数学与计算科学学院，广西 桂林 541004；
2. 广西高校数据分析与计算重点实验室，广西 桂林 541004；3. 广西应用数学中心，广西 桂林 541004）

摘要：针对传统基于谱聚类的子空间聚类方法在高维数据存在离群点时，易受离群点干扰而导致聚类性能下降的问题，提出一种基于自动权重的主动块对角子空间聚类方法. 该方法先为每个数据点赋予相应权重，通过权重差异识别数据中的离群点. 在确定离群点后，主动降低其在表示矩阵中的贡献度，进而构建更优的表示矩阵以提升模型的聚类性能. 在 10 个数据集上与 8 种对比算法的实验结果表明：在含 10%，20%离群点的数据集上，该方法的平均聚类准确率、归一化互信息、调整 Rand 指数等指标普遍优于对比算法；在一般聚类任务中，其在超过半数数据集上性能最优或位居前三. 因此该方法既能高效处理含离群点的高维数据聚类，又能在通用聚类任务中保持竞争力，为提高高维数据聚类的鲁棒性提供了有效方案，有较高的实际应用价值.

关键词：子空间聚类；离群点；自动权重；块对角方法

中图分类号：TP181 **文献标志码**：A **文章编号**：1671-5489(2025)06-1673-12

Active Block Diagonal Subspace Clustering Based on Automatic Weighting

LI Xiangli[1,2,3], XIE Tengchi[1], WEI Jiafeng[1]

(1. *School of Mathematics & Computing Science, Guilin University of Electronic Techology, Guilin* 541004, *Guangxi Zhuang Autonomous Region, China*; 2. *Guangxi University Key Laboratory of Data Analysis and Calculation, Guilin* 541004, *Guangxi Zhuang Autonomous Region, China*;
3. *Guangxi Applied Mathematics Center, Guilin* 541004, *Guangxi Zhuang Autonomous Region, China*)

Abstract: Aiming at the problem that traditional spectral clustering-based subspace clustering methods were prone to outlier interference and thus show degraded clustering performance when there were outliers in high-dimensional data, we proposed an active block diagonal subspace clustering method based on automatic weighting. The method first assigned a corresponding weight to each data point, identified outliers in the data through weight differences, then actively reduced its contribution in the representation matrix to construct a better representation matrix and improved the clustering performance of the model. Experimental results on 10 datasets compared with 8 algorithms show that the average clustering accuracy, normalized mutual information, and adjusted Rand index of the proposed method are generally better than the comparison algorithms on datasets with 10% or 20% outliers. It performs the best or ranks in the top three on more than half of the datasets in general clustering tasks. Therefore, the method can not only efficiently handle high-dimensional data

收稿日期：2025-01-10.

第一作者简介：李向利(1977—)，女，汉族，博士，教授，从事优化和模式识别的研究，E-mail：lixiangli@guet.edu.cn.

基金项目：国家自然科学基金(批准号：11961010).

clustering with outliers, but also maintain competitiveness in general clustering tasks, providing an effective solution to enhance the robustness of high-dimensional data clustering and having high practical application value.

Keywords: subspace clustering; outliers; automatic weighting; block diagonal method

高维数据在机器学习、信号与图像处理、自然语言处理、计算机视觉、生物信息学等领域中普遍存在[1]. 随着科学技术的发展,数据量和数据维度增长呈爆炸性趋势,维度灾难问题也随之产生. 为解决数据维度增加带来的维度灾难问题,需对高维数据进行降维处理,目前常用的降维技术包括主成分分析[2]、线性判别分析[3]和局部线性嵌入[4]等,但数据结构较复杂时应用这些方法应用可能受限. 子空间聚类可处理较复杂的数据结构,相比于传统降维技术,它没有过强约束,数据的每行和每列均可以属于一个或多个簇,也可以不在簇中[5]. 现有的子空间聚类方法主要分为4类:代数方法、迭代方法、统计方法和基于谱聚类的方法[6]. 相较于其他方法,基于谱聚类的方法能有效解决聚类结果重叠问题,且对噪声和初值点的敏感程度较低. 基于谱聚类的子空间聚类方法主要分为两步:先通过子空间模型学习一个表示系数矩阵,用于构建亲和矩阵;再对得到的亲和矩阵进行谱聚类得到最终的聚类结果. 第一步学习到的表示系数对聚类效果至关重要,Elhamifar 等[7]利用压缩感知技术,提出了将每个数据点表示为其他数据的稀疏线性组合得到表示矩阵的稀疏子空间聚类方法(SSC). Liu 等[8]为更好地捕获数据的全局结构,提出了一种基于低秩表示的子空间聚类方法(LRR). Luo 等[9]结合 LRR 和 SSC 的特点提出了多子空间表示(MSR),以更好地学习表示系数矩阵. 事实上,这些方法在学习表示矩阵过程中都忽略了数据的局部结构,为解决该问题,Lu 等[10]提出了一种图正则化低秩表示(GLRR)子空间聚类方法.

一个好的表示矩阵应具有块对角性质,在一定假设下,由上述模型学习到的表示矩阵可能符合块对角结构,但在实际应用中块对角性质不一定能得到保证. Feng 等[11]对模型施加了块对角先验,提出了一个基于图 Laplace 约束的公式,直接追求表示矩阵的块对角结构. 该模型保证了表示矩阵的块对角性质,但它假设数据是无噪声的. 为处理数据噪声问题,Wang 等[12]使用一个惩罚矩阵自适应地对重建误差进行加权,提出了一种可在没有先验知识情况下处理噪声的鲁棒块对角子空间聚类方法. 但这些方法都不针对数据中有离群点的情况,当数据中离群点数量较多时,可能会影响上述方法的聚类效果. 一个常规的数据集可能包含约 1%~10%或更多的离群点,而对于高维数据,这个比例可能更大[13]. 在数据处理过程中,离群点的检测尤为重要,因为它们可能代表了数据质量问题、异常事件或其他值得关注的问题. 此外,数据离群点对模型的性能影响较大,会影响模型的泛化能力,甚至导致模型无法提取有效特征. 因此,识别和分离数据异常值至关重要. 一般常采用鲁棒回归[14]或均值漂移模型[15]解决该问题. 但在子空间聚类问题中,如何利用数据自身特点自动识别离群点并降低其影响仍是一个亟待解决的问题. 因此,将子空间聚类与数据离群点检测相结合有一定意义.

针对数据中含有离群点的情况,本文提出一种基于自动权重的主动块对角子空间聚类方法,用于自动识别离群点并主动降低离群点对子空间聚类过程的影响. 该方法的创新点主要包括:在一般的块对角子空间聚类模型中引入了自动权重,使模型可自动识别数据中的离群点,并给其一个较小的权重;利用第一步中的权重,选择权重显著小的数据,主动减小这些数据在表示其他点时的贡献度,以降低离群点对聚类过程的影响.

1 预备知识

1.1 谱子空间聚类

在众多子空间聚类方法中,基于谱聚类的子空间聚类方法可处理稀疏数据,有良好的理论基础和解释性,且其聚类结果较稳定,近年来得到了广泛研究和应用. 基于谱聚类的子空间聚类一般流程如图1所示. 谱子空间聚类利用自表示性质,通过合适的子空间模型学习一个关于原数据的表示系数矩阵 $\boldsymbol{Z}$,学习到的表示矩阵一般不能直接用于谱聚类,需先构建亲和矩阵 $\boldsymbol{A}$,常用的公式是

$\boldsymbol{A}=(\boldsymbol{Z}+\boldsymbol{Z}^{\mathrm{T}})/2$，最后对亲和矩阵进行谱聚类获得最终的聚类结果.

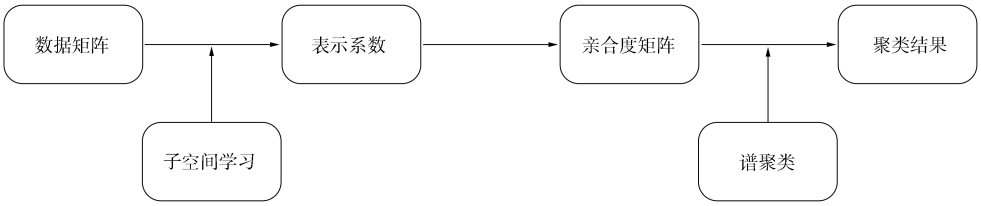


图 1　谱聚类流程

Fig. 1　Flow chart of spectral clustering

定义 1(自表示性质)[16]　给定一个数据集 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)\in\mathbb{R}^{m\times n}$，其中每个样本可表示为数据集中所有样本的线性或仿射组合，即 $\boldsymbol{X}=\boldsymbol{X}\boldsymbol{Z}$.

对于干净的数据，由自表示性质得到的表示矩阵可视为原数据集数据点之间的相似度矩阵. 实际上，数据中常包含噪声和异常值，直接由自表示性质得到的表示系数矩阵不能很好地反映数据之间的相似度，因此数据表示为 $\boldsymbol{X}=\boldsymbol{X}\boldsymbol{Z}+\boldsymbol{E}$，其中 $\boldsymbol{E}$ 为误差矩阵. 为减小表示误差，得到更好的表示矩阵，一般对表示矩阵和误差矩阵施加不同的先验约束，以获得理想的表示矩阵. 基于自表示性质的子空间聚类模型可统一描述为以下优化问题[17]：

$$\min_{\boldsymbol{Z},\boldsymbol{E}} F(\boldsymbol{Z})+\lambda R(\boldsymbol{E}),$$

$$\text{s. t. } \boldsymbol{X}=\boldsymbol{X}\boldsymbol{Z}+\boldsymbol{E},\quad \boldsymbol{Z}\in\Omega,$$

其中：$F(\boldsymbol{Z})$为正则项，用来约束表示矩阵使其保持特定结构，如对表示矩阵施加核范数使其保持低秩结构，或施加 l_1 范数使其保持稀疏结构等；Ω 为 $\boldsymbol{Z}$ 的约束集；$\lambda>0$ 为一个权衡参数；$R(\boldsymbol{E})$为描述真实数据和表示数据之间的误差项，根据不同的噪声分布选择不同范数衡量误差，通常需要一定的先验知识或假设.

1.2　块对角正则项

理想情况下，由子空间聚类模型学习得到的表示系数矩阵 $\boldsymbol{Z}$ 应是块对角的. 若数据集 $\boldsymbol{X}$ 被聚类为 k 个簇，则

$$\boldsymbol{Z}=\begin{pmatrix}\boldsymbol{Z}_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \boldsymbol{Z}_n\end{pmatrix}.$$

但在实际应用中，数据常包含噪声和异常值，直接利用子空间模型通常无法学习到具有块对角结构的表示矩阵. 因此，Feng 等[11]提出了块对角正则项，以保证表示系数具有理想的块对角结构.

定义 2(k 块对角正则项)[11]　对于给定的非负矩阵 $\boldsymbol{Z}\in\mathbb{R}^{m\times n}$，$\boldsymbol{Z}\geqslant 0$ 且 $\boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}}$，$k$ 块对角正则项定义为

$$\|\boldsymbol{Z}\|_k=\sum_{i=n-k+1}^{n}\lambda_i(\boldsymbol{L}_{\boldsymbol{Z}}),\tag{1}$$

其中：$\lambda_i(\boldsymbol{L}_{\boldsymbol{Z}})$为 $\boldsymbol{L}_{\boldsymbol{Z}}$ 的第 i 个按降序排列的特征值；$\boldsymbol{L}_{\boldsymbol{Z}}$ 为 $\boldsymbol{Z}$ 的 Laplace 矩阵，定义为 $\boldsymbol{L}_{\boldsymbol{Z}}=\boldsymbol{D}_{\boldsymbol{Z}}-(\boldsymbol{Z}+\boldsymbol{Z}^{\mathrm{T}})/2$，$\boldsymbol{D}_{\boldsymbol{Z}}$ 为对角矩阵，称为度矩阵，其第 i 个对角元素为 $\frac{1}{2}\sum_j(\boldsymbol{Z}_{i,j}+\boldsymbol{Z}_{j,i})$.

2　自动权重模型

2.1　模型构建

受各种因素的影响，高维数据在收集和处理过程中可能会产生离群点. 数据中的离群点对模型性能影响较大，会使模型的泛化能力下降，甚至导致模型无法提取有效特征. 因此，为识别离群点并降低其对子空间聚类过程的影响，本文提出基于自动权重的主动块对角子空间聚类方法(ABDR).

在子空间聚类过程中，可利用自表示性质对原数据进行线性表示. 子空间并集中的每个样本可有

效地表示为其他样本的线性或仿射组合，即 $\boldsymbol{X}=\boldsymbol{XZ}$，因此可以将矩阵 $\boldsymbol{X}-\boldsymbol{XZ}$ 视为原始数据和表示数据之间的误差项. 若数据中存在离群点，则误差项对应的列数值可能会异常，基于此，本文在子空间模型中加入权重项，自动识别数据中的离群点. 观察 $\boldsymbol{X}=\boldsymbol{XZ}$，对于 $\boldsymbol{X}$ 中的第 i 个数据，有

$$\begin{pmatrix}\boldsymbol{X}_{1,i}\\ \boldsymbol{X}_{2,i}\\ \vdots\\ \boldsymbol{X}_{d,i}\end{pmatrix}=\begin{pmatrix}\boldsymbol{X}_{1,1} & \boldsymbol{X}_{1,2} & \cdots & \boldsymbol{X}_{1,n}\\ \boldsymbol{X}_{2,1} & \boldsymbol{X}_{2,2} & \cdots & \boldsymbol{X}_{2,n}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{X}_{d,1} & \boldsymbol{X}_{d,2} & \cdots & \boldsymbol{X}_{d,n}\end{pmatrix}\begin{pmatrix}\boldsymbol{Z}_{1,i}\\ \boldsymbol{Z}_{2,i}\\ \vdots\\ \boldsymbol{Z}_{d,i}\end{pmatrix}, \tag{2}$$

即表示矩阵 $\boldsymbol{Z}$ 的第 i 列是第 i 个数据被所有数据线性表示时所有数据的表示系数，理想情况下，每个数据应该由同一个子空间中的所有数据线性表出，而离群点由于不属于任何一个子空间，其对应的表示矩阵的列 $\boldsymbol{Z}_{:,i}$ 可能会有异常，进而 $\|\boldsymbol{X}_{:,i}-\boldsymbol{XZ}_{:,i}\|_2^2$ 的值可能会偏大，若在其中加入一个权重，则当 $\|\boldsymbol{X}_{:,i}-\boldsymbol{XZ}_{:,i}\|_2^2$ 值偏大时权重有变化，通过此过程即可识别数据中的离群点. 受上述思想的启发，设 $\boldsymbol{w}$ 为 n 维权重向量，$\boldsymbol{w}$ 的第 i 个元素为 w^i，本文在 $\|\boldsymbol{X}_{:,i}-\boldsymbol{XZ}_{:,i}\|_2^2$ 中加入权重 w^i 得到 $\|w^i(\boldsymbol{X}_{:,i}-\boldsymbol{XZ}_{:,i})\|_2^2$，并对 $w^i,\boldsymbol{Z}_{:,i}$ 极小化有 $\min\limits_{w^i,\boldsymbol{Z}_{:,i}}\sum\limits_{i=1}^{n}\|w^i(\boldsymbol{X}_{:,i}-\boldsymbol{XZ}_{:,i})\|_2^2$，对每个数据加入权重则有 $\min\limits_{w^i,\boldsymbol{Z}}\sum\limits_{i=1}^{n}\|w^i(\boldsymbol{X}_{:,i}-\boldsymbol{XZ}_{:,i})\|_2^2$. 当数据集中含有离群点时，该权重项可以自动识别并赋予离群点较小的权重.

为限制表示矩阵 $\boldsymbol{Z}$ 的结构，本文在模型中加入块对角约束 $\|\boldsymbol{Z}\|_k$，模型可表示为

$$\begin{aligned}&\min_{w^i,\boldsymbol{Z}}\left\{\sum_{i=1}^{n}\|w^i(\boldsymbol{X}_{:,i}-\boldsymbol{XZ}_{:,i})\|_2^2+\frac{\alpha}{2}\|\boldsymbol{w}\|_2^2+\beta\|\boldsymbol{Z}\|_k\right\},\\ &\text{s.t.}\ \sum_{i=1}^{n}w^i=1,\quad w^i\geqslant 0,\quad \boldsymbol{Z}\geqslant 0,\quad \boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}},\quad \operatorname{diag}(\boldsymbol{Z})=0,\end{aligned} \tag{3}$$

其中 α 和 β 是平衡参数，正则项 $\|\boldsymbol{w}\|_2^2$ 及约束条件 $\operatorname{diag}(\boldsymbol{Z})=0$ 的作用是避免平凡解，$\operatorname{diag}(\boldsymbol{Z})=0$ 保证矩阵 $\boldsymbol{Z}$ 的对角线上元素全为 0，块对角正则项要求 $\boldsymbol{Z}\geqslant 0$ 且 $\boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}}$.

$\boldsymbol{Z}$ 的块对角约束会限制其表示能力，本文通过引入一个中间项缓解该问题[18]，则模型(3)转换为

$$\begin{aligned}&\min_{w^i,\boldsymbol{Z},\boldsymbol{B}}\sum_{i=1}^{n}\|w^i(\boldsymbol{X}_{:,i}-\boldsymbol{XB}_{:,i})\|_2^2+\frac{\alpha}{2}\|\boldsymbol{w}\|_2^2+\beta\|\boldsymbol{Z}\|_k+\frac{\gamma}{2}\|\boldsymbol{Z}-\boldsymbol{B}\|_{\mathrm{F}}^2,\\ &\text{s.t.}\ \sum_{i=1}^{n}w^i=1,\quad w^i\geqslant 0,\quad \boldsymbol{Z}\geqslant 0,\quad \boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}},\quad \operatorname{diag}(\boldsymbol{Z})=0,\end{aligned} \tag{4}$$

其中 γ 为平衡参数.

对于离群点或异常数据，该模型会自动分配一个较小的权重，从而可识别出离群点，但权重对表示系数的影响有限，还需对表示矩阵再进行处理以降低异常值的影响. 类似于对式(2)的观察，表示矩阵的第 j 行表示的是第 j 个样本在其他样本表示过程中的表示系数，若权重项识别出离群点，则可凭借此权重确定离群点的位置，再主动降低离群点在其他样本表示过程中的贡献度，即主动缩小表示矩阵第 j 行 $\boldsymbol{Z}_{j,:}$ 的值. 因此，本文提出以下处理过程：若

$$w^j<\frac{1}{n}\times\frac{1}{2}, \tag{5}$$

则令

$$\boldsymbol{Z}_{j,:}=w^j\boldsymbol{Z}_{j,:}. \tag{6}$$

通过该过程主动降低了离群点对表示系数矩阵的影响.

2.2 优化求解

本文使用交替方向乘子法(ADMM)[19]迭代地求解模型，即优化一个变量时固定其他变量，交替求解所有变量. 为方便求解，整理式(4)得

$$\begin{aligned}&\min_{\boldsymbol{D},\boldsymbol{Z},\boldsymbol{B}}\left\{\frac{1}{2}\|(\boldsymbol{X}-\boldsymbol{XB})\boldsymbol{D}\|_2^2+\frac{\alpha}{2}\|\boldsymbol{D}\|_2^2+\beta\|\boldsymbol{Z}\|_k+\frac{\gamma}{2}\|\boldsymbol{Z}-\boldsymbol{B}\|_{\mathrm{F}}^2\right\},\\ &\text{s.t.}\ \boldsymbol{D}=\operatorname{diag}(\boldsymbol{w}),\quad \boldsymbol{w}^{\mathrm{T}}\boldsymbol{1}=1,\quad w^i\geqslant 0,\quad \boldsymbol{Z}\geqslant 0,\quad \boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}},\quad \operatorname{diag}(\boldsymbol{Z})=0,\end{aligned} \tag{7}$$

其中 $\mathbf{1}$ 是元素全为 1 的 n 维向量，diag($\boldsymbol{w}$)表示以向量 $\boldsymbol{w}$ 的元素为对角线元素构成的对角矩阵.

1) 固定 $\boldsymbol{D}$ 和 $\boldsymbol{Z}$，更新 $\boldsymbol{B}$，则式(7)等价于

$$\min_{\boldsymbol{B}}\left\{\frac{1}{2}\|(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{D}\|_2^2+\frac{\gamma}{2}\|\boldsymbol{Z}-\boldsymbol{B}\|_{\mathrm{F}}^2\right\},$$

通过对 $\boldsymbol{B}$ 进行求导且令求导后的式子等于 0，有

$$\boldsymbol{B}=\boldsymbol{Z}+\frac{1}{\gamma}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{B}\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}). \tag{8}$$

2) 固定 $\boldsymbol{D}$ 和 $\boldsymbol{B}$，更新 $\boldsymbol{Z}$，则问题(7)转化为

$$\begin{aligned}&\min_{\boldsymbol{Z}}\left\{\beta\|\boldsymbol{Z}\|_k+\frac{\gamma}{2}\|\boldsymbol{Z}-\boldsymbol{B}\|_{\mathrm{F}}^2\right\},\\&\text{s.t. }\boldsymbol{Z}\geqslant 0,\quad \boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}},\quad \mathrm{diag}(\boldsymbol{Z})=0.\end{aligned} \tag{9}$$

由块对角正则化的定义，根据式(1)可知下式成立[20]：

$$\begin{aligned}&\min_{\boldsymbol{Z}}\|\boldsymbol{Z}\|_k=\min_{\boldsymbol{Z},\boldsymbol{Y}}\{\boldsymbol{L}_{\boldsymbol{Z}},\boldsymbol{Y}\},\\&\text{s.t. }0\leqslant\boldsymbol{Y}\leqslant\boldsymbol{I},\quad \mathrm{rank}(\boldsymbol{Y})=k,\end{aligned}$$

其中 $\boldsymbol{I}$ 是单位矩阵，$\boldsymbol{Y}$ 可由 $\boldsymbol{Y}=\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}$ 更新，$\boldsymbol{F}\in\mathbb{R}^{m\times n}$ 由与 $\boldsymbol{L}_{\boldsymbol{Z}}$ 的最小 k 个特征值相关的 k 个特征向量组成，$\boldsymbol{L}_{\boldsymbol{Z}}$ 是矩阵 $\boldsymbol{Z}$ 的 Laplace 矩阵，等价于 diag($\boldsymbol{Z}\mathbf{1}$)$-\boldsymbol{Z}$. 固定 $\boldsymbol{Y}$，则问题(9)等价于

$$\begin{aligned}&\min_{\boldsymbol{Z}}\left\{\beta\{\mathrm{diag}(\boldsymbol{Z}\mathbf{1})-\boldsymbol{Z},\boldsymbol{Y}\}+\frac{\gamma}{2}\|\boldsymbol{Z}-\boldsymbol{B}\|_{\mathrm{F}}^2\right\},\\&\text{s.t. }\boldsymbol{Z}\geqslant 0,\quad \boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}},\quad \mathrm{diag}(\boldsymbol{Z})=0.\end{aligned} \tag{10}$$

根据 Wang 等[12]的工作，$\boldsymbol{Z}$ 可解为

$$\boldsymbol{Z}=\frac{1}{2}(\hat{\boldsymbol{A}}+\hat{\boldsymbol{A}}^{\mathrm{T}})_+, \tag{11}$$

其中$(\ )_+$约束 $\hat{\boldsymbol{A}}+\hat{\boldsymbol{A}}^{\mathrm{T}}$ 的元素为正值，$\hat{\boldsymbol{A}}=\boldsymbol{A}-\mathrm{diag}(\mathrm{diag}(\boldsymbol{A}))$，$\boldsymbol{A}=\boldsymbol{B}-\dfrac{\gamma(\mathrm{diag}(\boldsymbol{Y})\mathbf{1}^{\mathrm{T}}-\boldsymbol{Y})}{\lambda}$.

3) 固定 $\boldsymbol{Z}$ 和 $\boldsymbol{B}$，更新 $\boldsymbol{D}$，则问题(7)等价于

$$\begin{aligned}&\min_{\boldsymbol{D}}\frac{1}{2}\|(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{D}\|_2^2+\frac{\alpha}{2}\|\boldsymbol{D}\|_2^2,\\&\text{s.t. }\boldsymbol{D}=\mathrm{diag}(\boldsymbol{w}),\quad \boldsymbol{w}^{\mathrm{T}}\mathbf{1}=1,\quad w^i\geqslant 0,\end{aligned} \tag{12}$$

为求解问题(12)，令 $\boldsymbol{E}=\boldsymbol{X}-\boldsymbol{X}\boldsymbol{B}$，则式(12)转化为

$$\begin{aligned}&\min_{\boldsymbol{D}}\left\{\frac{1}{2}\|\boldsymbol{E}\boldsymbol{D}\|_2^2+\frac{\alpha}{2}\|\boldsymbol{D}\|_2^2\right\},\\&\text{s.t. }\boldsymbol{D}=\mathrm{diag}(\boldsymbol{w}),\quad \boldsymbol{w}^{\mathrm{T}}\mathbf{1}=1,\quad w^i\geqslant 0,\end{aligned} \tag{13}$$

问题(13)可视为等式约束优化问题，使用 Lagrange 乘子法求解，其 Lagrange 函数为

$$L(\boldsymbol{w},\theta)=\frac{1}{2}\sum_{i=1}^{n}\Big(\sum_{j=1}^{m}E_{i,j}^2\Big)w^{i2}+\frac{\alpha}{2}\sum_{i=1}^{n}w^{i^2}+\theta\Big(\sum_{i=1}^{n}w^i-1\Big),$$

其中 θ 是 Lagrange 乘子. 令$\dfrac{\partial L(\boldsymbol{w},\theta)}{\partial w^i}=0$，则有

$$w^i=\frac{-\theta}{\sum_{j=1}^{m}E_{i,j}^2+\alpha}, \tag{14}$$

又由约束条件 $\boldsymbol{w}^{\mathrm{T}}\mathbf{1}=1$ 有 $\sum_{i=1}^{n}w^i=1$，进一步有

$$\sum_{i=1}^{n}w^i=\sum_{i=1}^{n}\frac{-\theta}{\sum_{j=1}^{m}E_{i,j}^2+\alpha}=1,$$

求解得

$$\theta=-\frac{1}{\sum_{i=1}^{n}\left(\sum_{j=1}^{m}E_{i,j}^{2}+\alpha\right)^{-1}},\tag{15}$$

将式(15)代入式(14)得

$$w^{i}=\frac{1}{\left[\sum_{i=1}^{n}\left(\sum_{j=1}^{m}E_{i,j}^{2}+\alpha\right)^{-1}\right]\left(\sum_{j=1}^{m}E_{i,j}^{2}+\alpha\right)},\tag{16}$$

显然，最终得到的 w^{i} 满足约束条件 $w^{i}\geqslant 0$.

2.3 算法流程

设上一次迭代中得到的结果为 $\boldsymbol{Z}$ 和 $\boldsymbol{B}$，当前迭代得到的结果为 $\bar{\boldsymbol{Z}}$ 和 $\bar{\boldsymbol{B}}$，则整个优化过程的停止条件为 $\mathrm{STOP}<1\times10^{-5}$，其中 $\mathrm{STOP}=\max\{\bar{\boldsymbol{Z}}-\boldsymbol{Z},\bar{\boldsymbol{B}}-\boldsymbol{B}\}$.

下列算法给出了本文算法的整个优化过程.

算法 1 ABDR.

输入：$\boldsymbol{X},\alpha,\beta,\gamma,k$；

输出：聚类结果；

步骤 1) 固定 $\boldsymbol{Z}$ 和 $\boldsymbol{D}$，通过式(8)更新 $\boldsymbol{B}$；

步骤 2) 固定 $\boldsymbol{D}$ 和 $\boldsymbol{B}$，通过式(11)更新 $\boldsymbol{Z}$；

步骤 3) 固定 $\boldsymbol{Z}$ 和 $\boldsymbol{B}$，通过式(16)更新 $\boldsymbol{D}$；

步骤 4) iter=iter+1；

End

对得到的 $\boldsymbol{Z}$ 进行谱聚类得到最终的聚类结果.

3 实 验

为验证本文模型的有效性及性能，下面进行对比实验. 对比算法选取经典算法和近5年的新算法，包括低秩表示方法(LRR)[8]、稀疏子空间聚类方法(SSC)[7]、基于最小二乘法的子空间聚类方法(LSR)[21]、秩最小化鲁棒谱集成聚类(RSEC)[22]、块对角表示方法(BDR)[18]、层次化加权低秩表示方法(HWLRR)[23]、尺度化单纯形表示子空间聚类方法(SSRSC)[24]、全局和局部结构保持的非负子空间聚类算法(NSC)[25]. 本文用 MATLAB 实现了所有方法. 硬件环境配置：CPU 为 2.30 GHz 12th Gen Intel(R) Core(TM) i7-12700H，内存为 16.0 GB. 为确保公平，在对比实验中，使用原文献源代码中默认或建议参数或在推荐范围内进行调整选取的参数值. 采用的聚类性能评估指标包括平均聚类准确率(ACC)[26]、归一化互信息(NMI)[27]及调整 Rand 指数(ARI)[27]. 为降低随机因素的影响，所有算法的实验结果均进行 20 次独立重复，并取其平均值以获得更稳定的结果. 本文使用多个数据集[28-33]评估所有方法的性能，实验数据集信息列于表 1.

表 1 实验数据集信息

Table 1 Information of experimental dataset

数据集	样本数	维数	类别数	数据集	样本数	维数	类别数
S1500	1 595	20	12	Jaffe	213	676	10
D50	1 595	50	13	Yale	165	1 024	15
Iris	150	4	3	ORL	400	1 024	20
Yeast	1 484	1 470	10	COIL20	1 440	1 024	20
Dig	1 797	64	10	Isolet5	1 559	617	26

3.1 权重有效性验证

下面验证权重项能有效识别数据中的离群点，且 ABDR 模型可有效处理数据中的离群点，得到较好的聚类结果.

数据集 ORL 用于评估 ABDR 模型中权重项对异常值的识别能力. 理论上，ABDR 模型会自动赋

予离群点一个较小的权重，若权重项可有效识别离群点，则权重向量中对应离群点位置的元素值应小于对应正常数据的元素值. 因此，在数据集 ORL 中分别加入 1 个及 1%，10%，20%的离群点，可视化加入异常值前后的权重向量如图 2 所示，其中数据点被赋予的权重越大越接近黄色，被赋予的权重越小越接近红色. 数据集 ORL 中有 400 个样本. 由图 2 可见：当数据集中没有离群点时，每个数据点的权重约为 $1/n$（n 为数据点总数）；数据集中存在 1 个离群点时，该离群点被赋予的权重远小于正常数据的权重；随着离群点数量增加至 10%，20%，离群点所对应的权重仍然较小. 即数据集中离群点数量对权重项的影响较小，对不同比例的离群点，ABDR 模型都能准确识别并给予其较小的权重.

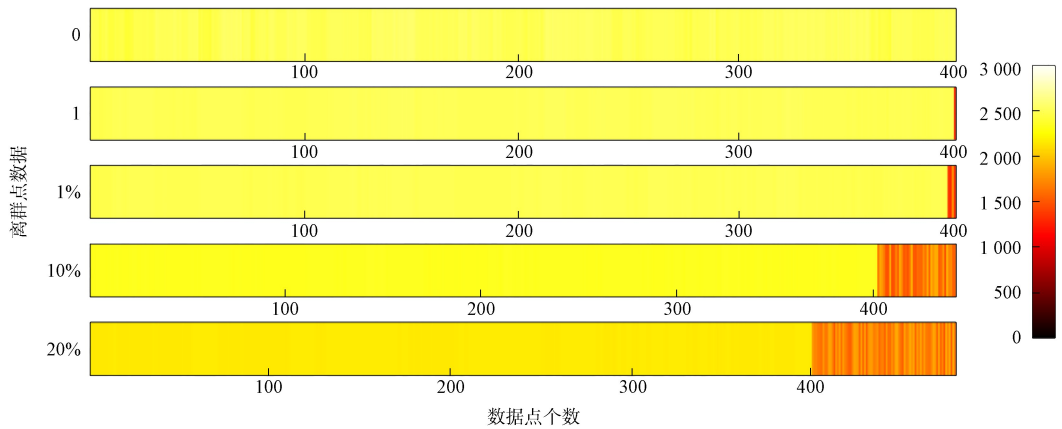


图 2　数据集 ORL 加入不同比例离群点后的权重

Fig. 2　Weights of ORL dataset after adding different proportions of outliers

为评估 ABDR 算法在含有离群点的数据集上的聚类性能，本文在数据集 ORL，Yale 和 S1500 中分别引入了 10%和 20%的离群点，以此比较 ABDR 与其他 8 种算法在处理离群点数据方面的能力. 表 2～表 4 分别列出了 ABDR 算法及对比算法在 3 个评价指标 ACC，NMI，ARI 上的实验结果.

表 2　不同算法对含离群点聚类任务的 ACC 结果

Table 2　ACC results for clustering tasks with outliers by different algorithms　%

数据集	离群点比例	算法								
		SSC	LRR	LSR	BDR	RSEC	HWLRR	SSRSC	NSC	ABDR
Dig	10	9.33	8.64	11.46	8.86	11.31	8.86	9.87	8.40	14.34
	20	4.86	3.77	13.37	7.14	12.65	7.14	5.45	9.52	15.52
ORL	10	31.88	48.39	48.91	4.09	19.49	22.65	45.31	26.52	57.82
	20	23.07	34.80	35.24	4.58	16.82	21.23	35.57	25.43	42.68
Yale	10	7.73	24.25	31.49	7.73	34.56	39.86	38.48	39.59	44.45
	20	8.08	16.74	23.16	8.08	31.72	34.37	27.85	30.76	42.95
S1500	10	10.06	8.39	12.80	8.84	11.48	12.07	9.30	11.59	14.56
	20	5.56	7.48	13.54	8.15	7.37	9.93	5.75	7.00	15.45
Jaffe	10	49.10	45.17	45.02	10.26	44.98	65.75	42.48	43.76	70.69
	20	27.65	26.98	25.98	9.80	26.12	61.43	31.53	26.86	68.24
D50	10	8.83	12.84	14.55	8.83	16.27	8.83	10.16	12.41	18.77
	20	8.15	8.88	14.71	8.15	17.81	8.15	6.04	7.72	18.66
Iris	10	27.82	18.06	16.85	30.91	23.33	22.21	30.91	19.45	68.33
	20	16.28	12.92	16.28	28.33	24.78	17.00	28.33	17.69	65.72
Yeast	10	25.67	12.59	9.82	25.67	13.45	14.60	9.30	5.10	25.67
	20	21.18	9.46	7.34	21.18	12.28	21.18	7.06	3.39	21.18
Isolet5	10	8.78	7.91	7.99	3.53	7.87	3.53	12.71	10.52	23.06
	20	8.72	7.89	7.85	3.53	7.98	3.53	12.02	10.47	22.94
COIL20	10	4.73	16.91	17.35	4.73	16.88	4.73	15.87	18.34	34.87
	20	4.46	9.59	9.84	4.46	9.59	4.46	9.69	13.84	33.10

表3 不同算法对含离群点聚类任务的NMI结果

Table 3 NMI results for clustering tasks with outliers by different algorithms %

数据集	离群点比例	算法								
		SSC	LRR	LSR	BDR	RSEC	HWLRR	SSRSC	NSC	ABDR
Dig	10	59.47	47.77	60.68	12.43	60.73	12.43	48.75	38.10	63.63
	20	53.67	43.44	62.23	17.76	62.87	17.76	45.95	43.21	65.00
ORL	10	64.39	80.21	80.58	31.88	47.93	59.85	76.63	54.60	83.02
	20	64.87	76.17	76.13	38.97	56.22	64.66	73.50	56.82	77.72
Yale	10	29.50	47.76	53.36	29.50	58.95	62.24	60.43	61.68	62.22
	20	37.03	41.01	55.04	37.03	66.09	65.69	61.95	65.30	67.24
S1500	10	59.72	42.88	43.00	16.79	57.44	63.87	59.64	64.05	64.13
	20	55.04	45.18	46.88	24.57	44.04	65.49	59.54	62.71	65.82
Jaffe	10	79.68	77.11	78.16	22.57	76.69	81.09	63.63	76.52	85.60
	20	70.71	70.03	69.82	30.17	69.89	82.39	67.91	71.16	85.75
D50	10	16.78	55.77	56.54	16.78	64.64	16.78	55.29	57.77	68.40
	20	23.65	51.25	57.10	23.65	65.75	23.65	52.04	53.81	68.73
Iris	10	53.06	36.47	32.88	12.65	43.19	56.67	12.65	25.10	70.26
	20	50.72	41.71	36.74	18.99	46.70	58.25	18.99	35.59	75.06
Yeast	10	10.83	30.50	30.34	10.83	27.08	33.34	29.84	28.90	35.97
	20	15.63	42.96	40.62	15.63	37.35	15.63	31.91	36.36	46.05
Isolet5	10	56.53	52.19	52.86	27.27	52.80	27.27	61.00	57.61	67.75
	20	56.34	52.74	52.82	27.27	52.69	27.27	60.25	57.69	67.70
COIL20	10	18.69	66.56	67.80	18.69	67.18	18.69	63.49	69.05	76.67
	20	26.64	62.23	62.18	26.64	61.95	26.64	59.94	67.03	78.10

表4 不同算法对含离群点聚类任务的ARI结果

Table 4 ARI results for clustering tasks with outliers by different algorithms %

数据集	离群点比例	算法								
		SSC	LRR	LSR	BDR	RSEC	HWLRR	SSRSC	NSC	ABDR
Dig	10	87.46	67.19	83.17	17.91	82.77	17.91	69.81	49.84	90.05
	20	79.94	47.94	80.35	23.79	80.42	23.79	59.44	39.82	88.30
ORL	10	49.51	74.20	74.92	20.23	30.41	32.47	67.33	34.77	75.75
	20	47.94	71.33	71.24	26.88	36.11	33.02	66.73	37.09	68.20
Yale	10	22.65	35.99	43.34	22.65	53.26	52.54	55.03	54.59	56.33
	20	29.29	35.71	44.57	29.29	60.23	57.32	55.83	53.48	59.34
S1500	10	83.28	52.55	51.27	18.42	78.01	82.85	83.59	85.24	84.86
	20	76.94	41.44	47.07	25.24	55.10	79.48	78.29	80.06	83.87
Jaffe	10	89.96	86.67	88.44	22.22	86.37	85.68	71.32	85.90	90.72
	20	82.45	81.73	81.45	28.63	81.61	83.96	76.84	79.78	89.86
D50	10	18.46	72.73	72.88	18.46	85.86	18.46	74.68	72.99	91.07
	20	25.27	53.59	64.57	25.27	81.61	25.27	65.47	56.42	88.92
Iris	10	80.94	62.67	61.15	40.61	70.06	86.88	40.61	55.24	72.06
	20	76.67	62.70	54.39	45.56	65.83	82.33	45.56	53.42	82.33
Yeast	10	34.80	51.87	48.68	34.80	38.68	46.08	55.08	51.61	58.44
	20	37.81	54.31	51.34	37.81	42.84	37.81	53.66	49.56	61.26
Isolet5	10	56.42	51.72	52.08	21.18	51.82	21.18	63.36	44.31	67.60
	20	56.40	51.77	52.07	21.18	51.65	21.18	62.64	44.40	67.50
COIL20	10	14.84	78.62	80.79	14.84	79.82	14.84	73.48	77.28	88.07
	20	21.93	76.39	76.50	21.93	76.17	21.93	71.42	69.60	89.33

由表2～表4可见，由于离群点的影响，传统的子空间聚类方法在聚类性能上有衰退. 随着离群点比例从10%增加到20%，ACC,NMI,ARI评价指标均出现了不同程度的下降，表明离群点的增加对

模型的训练过程产生了干扰，进而影响了模型对数据的拟合能力. 在所有算法中，ABDR 算法在 3 个数据集上的 ACC,NMI,ARI 指标普遍优于其他算法，且差异较大. 充分说明了 ABDR 算法在处理数据中的离群点方面效果较好，能获得更高质量的聚类结果.

3.2　一般聚类任务性能验证

对于一般聚类任务，本文算法仍有良好的聚类性能. 表 5 列出了 ABDR 与 8 种对比算法在 Dig，COIL20,Jaffe,Isolet5 等 10 个数据集上的 ACC,NMI,ARI 实验结果. 由表 5 可见，在超过 50%的数据集上，ABDR 算法优于对比算法的性能，尽管 ABDR 算法在部分数据集上的性能排名并非第一，但其性能仍位居前三. 表明作为一种子空间算法，ABDR 算法不仅能有效处理含有离群点的数据，而且在一般聚类任务中，ABDR 算法与现有的无监督子空间算法相比仍具有竞争力.

表 5　不同算法对一般聚类任务的 ACC,NMI,ARI 实验结果

Table 5　Experimental results of ACC, NMI and ARI for general clustering tasks by different algorithms　%

数据集	聚类指标	算法								
		SSC	LRR	LSR	BDR	RSEC	HWLRR	SSRSC	NSC	ABDR
Dig	ACC	70.84	63.21	76.39	73.50	71.13	80.02	62.74	77.38	81.02
	NMI	70.32	69.00	70.62	65.91	68.50	83.46	66.21	82.85	74.28
	ARI	75.20	75.92	76.39	74.28	72.07	80.02	64.72	83.81	81.02
ORL	ACC	71.16	69.71	68.53	59.93	76.13	71.88	81.18	64.28	81.64
	NMI	87.46	83.69	83.36	75.36	46.99	89.29	90.69	71.67	90.72
	ARI	75.99	72.05	71.41	63.25	69.03	76.08	83.25	78.05	83.34
Yale	ACC	43.67	25.67	37.55	45.79	44.76	55.76	58.97	55.79	59.73
	NMI	50.27	33.66	42.42	51.07	49.40	58.46	60.35	59.39	58.57
	ARI	46.73	26.88	38.61	46.97	46.12	56.97	59.58	56.67	59.85
S1500	ACC	60.97	58.23	72.56	71.09	31.06	67.13	78.73	88.70	78.14
	NMI	78.90	59.89	77.44	75.09	77.04	77.72	82.90	88.96	82.94
	ARI	81.94	66.30	79.76	74.54	77.40	77.35	85.18	89.38	82.22
Jaffe	ACC	90.18	91.03	90.18	90.48	90.12	91.46	93.59	95.80	96.71
	NMI	94.35	85.45	95.59	91.14	91.70	90.84	92.68	90.11	95.64
	ARI	90.18	86.92	90.18	92.30	91.15	91.78	96.41	90.80	96.71
D50	ACC	74.71	57.27	86.42	65.77	73.08	80.29	94.17	83.63	90.42
	NMI	73.05	56.63	84.59	66.00	70.07	84.18	95.21	91.76	90.89
	ARI	83.11	60.70	89.96	69.57	28.80	88.10	97.93	86.63	94.60
Iris	ACC	66.35	78.64	78.10	88.67	73.33	86.34	78.67	90.67	88.67
	NMI	31.18	58.98	58.98	76.18	64.05	73.77	59.45	80.57	76.18
	ARI	66.52	78.45	78.68	88.67	73.33	86.05	78.67	90.67	88.67
Yeast	ACC	34.48	30.77	33.69	34.25	34.76	25.24	34.80	31.20	35.40
	NMI	24.80	14.16	13.03	10.36	24.73	12.18	25.20	17.13	26.82
	ARI	51.96	38.26	35.23	37.73	52.00	40.57	52.57	31.20	54.44
Isolet5	ACC	41.74	24.23	40.13	23.30	29.65	23.91	47.23	23.85	56.81
	NMI	67.56	24.89	45.84	24.93	67.81	63.30	62.73	68.98	72.29
	ARI	53.32	25.51	41.84	35.92	40.14	45.45	49.31	43.85	59.87
COIL20	ACC	63.70	53.74	49.81	59.13	55.12	65.07	68.26	49.38	68.56
	NMI	58.58	57.26	48.52	49.51	60.65	62.66	70.19	67.36	77.63
	ARI	60.92	46.03	42.89	51.14	60.00	66.32	73.90	54.90	69.71

3.3　运行时间

图 3 为 9 种算法在 10 个数据集上运行 20 次的平均运行时间对比结果. 为确保公平，设置所有算法的最大迭代次数为 100. 由图 3 可见，本文算法在大部分数据集上的运行时间都位于中游水平，在所有数据集上运行时间都优于部分算法. 除经典算法 LSR 及核方法 NSC 外，ABDR 与其他算法相比，在运行时间方面有一定的竞争力. 由于 LSR 算法不涉及计算矩阵特征值，核方法 NSC 在迭代前进行

了降维，因此这两种算法与ABDR算法相比计算复杂度较低. ABDR算法的运行时间主要受变量$\boldsymbol{D}$，$\boldsymbol{Z}$,$\boldsymbol{B}$更新的影响，其中$\boldsymbol{D}$的复杂度由变量$\boldsymbol{w}$的迭代决定，$\boldsymbol{Z}$的复杂度由变量$\boldsymbol{Y}$的迭代决定，因此主要分析$\boldsymbol{w}$,$\boldsymbol{Y}$,$\boldsymbol{B}$的复杂度. 由矩阵乘法的计算复杂度可知，更新$\boldsymbol{w}$和$\boldsymbol{B}$的计算复杂度均为$O(mn^2)$，而对于$\boldsymbol{Y}$，由于需要计算矩阵的特征值，故其计算复杂度为$O(n^3)$.

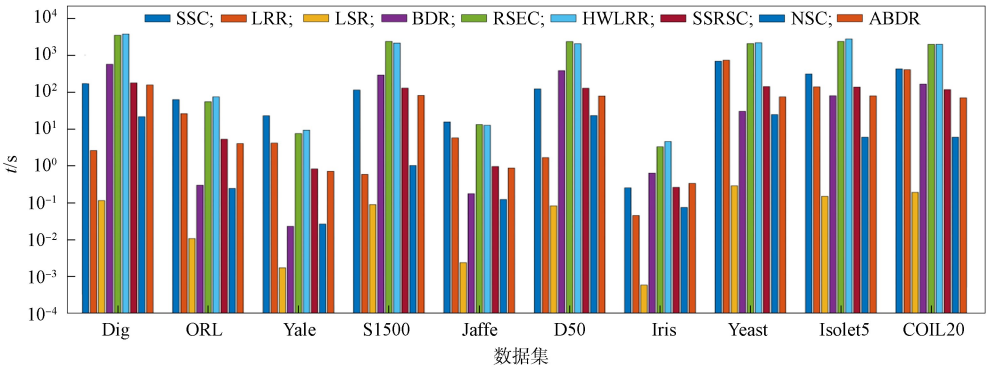


图3 9种算法在10个数据集上的平均运行时间对比结果

Fig. 3 Comparison results of average running time of nine algorithms on ten datasets

3.4 收敛性分析

以STOP为代价函数，图4为本文算法在两个不同数据集Isolet5和ORL上的代价函数值随迭代次数的变化情况. 由图4可见，在数据集Isolet5上，初始代价函数值约为0.14，而在数据集ORL上，初始代价函数值约为0.7，随着迭代次数的增加，两个数据集上的代价函数值都迅速下降并趋于稳定，最终接近于0. 此外，本文算法在两个数据集上都在200次迭代前收敛，表明ADBR算法在数据集Isolet5和ORL上都能快速收敛到局部最小值或全局最小值. 实验结果表明，本文算法在不同数据集上均可快速收敛，能有效找到最优解或近似最优解.

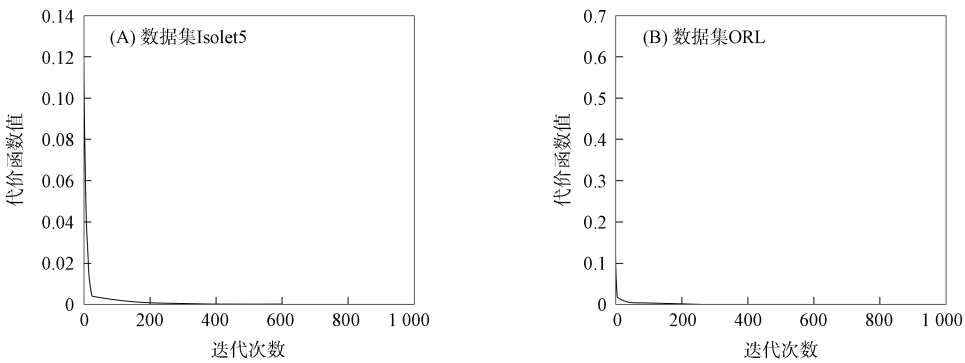


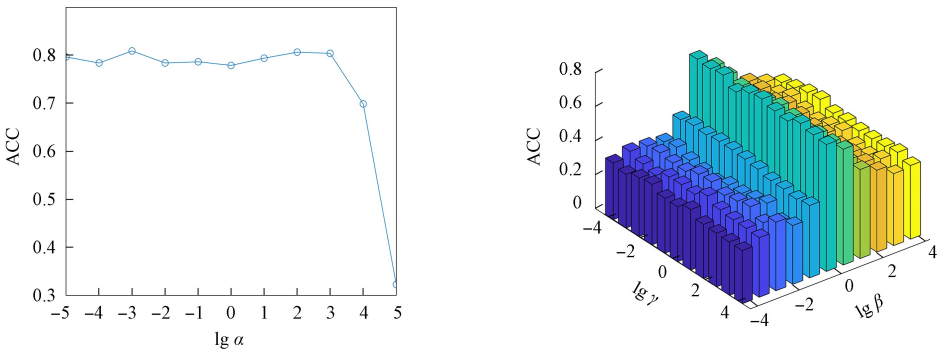
图4 ADBR算法随迭代次数变化在两个数据集上的迭代函数值

Fig. 4 Iteration function values of ADBR algorithm changing with number of iterations on two datasets

3.5 参数分析

本文ABDR模型包含3个关键参数α,β和γ. 为深入理解这些参数对聚类结果的影响，在较大范围内对这些参数进行取值实验，并观察ABDR模型的聚类性能，结果如图5所示. 首先，保持参数β和γ的值不变，将参数α在$[10^{-5},10^{5}]$内进行调节. 由图5可见，α的最佳取值区间为$[10^{-5},10^{3}]$，在此区间内ABDR模型能达到最优的聚类效果. 其次，固定α值，并在一个较大范围内对参数β和γ取值. 通过分析数据集ORL上的ACC值，发现ABDR模型的聚类性能对β的取值非常敏感，而γ的影响则相对较小. 其中，β的最佳取值确定为1，而γ的最佳取值范围为$[10^{-5},10^{-3}]$.

综上所述，针对传统基于谱聚类的子空间聚类方法在高维数据存在离群点时，易受离群点干扰而导致聚类性能下降的问题，本文提出了一种新的基于权重的主动块对角子空间聚类算法(ABDR). ABDR算法利用数据的自表示性质，赋予离群点一个较小的权重，并主动降低了离群点在表示矩阵中



(A) ACC随α的变化情况　　(B) ACC随β和γ的变化情况

图 5　ADBR 模型随不同参数变化在数据集 ORL 上的聚类性能

Fig. 5　Clustering performance of ADBR model changing with different parameters on ORL dataset

的贡献程度，在识别离群点的同时，有效降低了离群点对聚类结果的不良影响. 通过与 8 种对比算法在 10 个数据集上的对比实验，证明了 ABDR 算法的权重分配策略在含离群点聚类任务中的高效性和实用性，以及其在一般聚类任务中的价值. 未来可进一步拓展该方法的应用范围，将其应用于多视图聚类任务或用于处理含离群点的张量数据，以应对更复杂的数据场景.

参　考　文　献

[1] ELHAMIFAR E, VIDAL R. Sparse Subspace Clustering: Algorithm, Theory, and Applications [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(11): 2765-2781.

[2] ABDI H, WILLIAMS L J. Principal Component Analysis [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4): 433-459.

[3] BALAKRISHNAMA S, GANAPATHIRAJU A. Linear Discriminant Analysis—A Brief Tutorial [J]. Institute for Signal and Information Processing, 1998, 18: 1-8.

[4] ROWEISS T, SAUL L K. Nonlinear Dimensionality Reduction by Locally Linear Embedding [J]. Science, 2000, 290(5500): 2323-2326.

[5] 张宪超. 数据聚类 [M]. 北京：科学出版社，2018：233-243.（ZHANG X C. Data Clustering [M]. Beijing: Science Press, 2018: 233-243.）

[6] QU W T, XIU X C, CHEN H Y, et al. A Survey on High-Dimensional Subspace Clustering [J]. Mathematics, 2023, 11(2): 436-1-436-39.

[7] ELHAMIFAR E, VIDAL R. Sparse Subspace Clustering [C]//2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2009: 2790-2797.

[8] LIU G C, LIN Z C, YU Y. Robust Subspace Segmentation by Low-Rank Representation [C]//Proceedings of the 27th International Conference on Machine Learning (ICML-10). New York: ACM, 2010: 663-670.

[9] LUO D J, NIE F P, DING C, et al. Multi-subspace Representation and Discovery [C]//Machine Learning and Knowledge Discovery in Databases, Part Ⅱ. Berlin: Springer, 2011: 405-420.

[10] LU X Q, WANG Y L, YUAN Y. Graph-Regularized Low-Rank Representation for Destriping of Hyperspectral Images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2013, 51(7): 4009-4018.

[11] FENG J S, LIN Z C, XU H, et al. Robust Subspace Segmentation with Block-Diagonal Prior [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 3818-3825.

[12] WANG L J, HUANG J W, YIN M, et al. Block Diagonal Representation Learning for Robust Subspace Clustering [J]. Information Sciences, 2020, 526: 54-67.

[13] LAW J. Robust Statistics: The Approach Based on Influence Functions [J]. Journal of the Royal Statistical Society Series D, 1986, 35(5): 565-566.

[14] ROUSSEEUW P J, LEROY A M. Robust Regression and Outlier Detection [M]. New York: John Wiley & Sons, 2005: 216-247.

[15] SHE Y Y, OWEN A B. Outlier Detection Using Nonconvex Penalized Regression [J]. Journal of the American Statistical Association, 2011, 106: 626-639.

[16] NG A, JORDAN M, WEISS Y. On Spectral Clustering: Analysis and an Algorithm [C]//Proceedings of the 15th International Conference on Neural Information Processing Systems: Natural and Synthetic. New York: ACM, 2001: 849-856.

[17] DONG W H, WU X J, KITTLER J. Subspace Clustering via Joint $l_{1,2}$ and $l_{2,1}$ Norms [J]. Information Sciences, 2022, 612: 675-686.

[18] LU C Y, FENG J S, LIN Z C, et al. Subspace Clustering by Block Diagonal Representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 487-501.

[19] BOYD S, PARIKH N, CHU E, et al. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers [J]. Foundations and Trends in Machine Learning, 2011, 3(1): 1-122.

[20] FU Z Q, ZHAO Y, CHANG D X, et al. Auto-weighted Low-Rank Representation for Clustering [J]. Knowledge-Based Systems, 2022, 251: 109063-1-109063-11.

[21] LU C Y, MIN H, ZHAO Z Q, et al. Robust and Efficient Subspace Segmentation via Least Squares Regression [C]//Computer Vision-ECCV 2012. Berlin: Springer, 2012: 347-360.

[22] TAO Z Q, LIU H F, LI S, et al. Robust Spectral Ensemble Clustering via Rank Minimization [J]. ACM Transactions on Knowledge Discovery from Data, 2019, 13(1): 1-25.

[23] FU Z Q, ZHAO Y, CHANG D X, et al. A Hierarchical Weighted Low-Rank Representation for Image Clustering and Classification [J]. Pattern Recognition, 2021, 112: 107736-1-107736-12.

[24] XU J, YU M Y, SHAO L, et al. Scaled Simplex Representation for Subspace Clustering [J]. IEEE Transactions on Cybernetics, 2019, 51(3): 1493-1505.

[25] JIA H J, ZHU D X, HUANG L X, et al. Global and Local Structure Preserving Nonnegative Subspace Clustering [J]. Pattern Recognition, 2023, 138: 109388-1-109388-12.

[26] SCHÜTZE H, MANNING C D, RAGHAVAN P. Introduction to Information Retrieval [M]. Cambridge: Cambridge University Press, 2008: 39.

[27] ZHANG S H, WONG H S, SHEN Y. Generalized Adjusted Rand Indices for Cluster Ensembles [J]. Pattern Recognition, 2012, 45(6): 2214-2226.

[28] SAMARIA F S, HARTER A C. Parameterisation of a Stochastic Model for Human Face Identification [C]//Proceedings of 1994 IEEE Workshop on Applications of Computer Vision. Piscataway: IEEE, 1994: 138-142.

[29] ZHAO H D, DING Z M, FU Y. Multi-view Clustering via Deep Matrix Factorization [C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. New York: ACM, 2017: 2921-2927.

[30] MÜLLER E, GÜNNEMANN S, ASSENT I, et al. Evaluating Clustering in Subspace Projections of High Dimensional Data [J]. Proceedings of the VLDB Endowment, 2009, 2(1): 1270-1281.

[31] NENE S A, NAYAR S K, MURASE H. Columbia Object Image Library (coil-20) [DB/OL]. (1996-01-01) [2024-12-30]. https://git-disl.github.io/GTDLBench/datasets/coil20.

[32] LI Z C, YANG Y, LIU J, et al. Unsupervised Feature Selection Using Nonnegative Spectral Analysis [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2012: 1026-1032.

[33] JIA Y, LU G, LIU H, et al. Semi-supervised Subspace Clustering via Tensor Low-Rank Representation [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(7): 3455-3461.

(责任编辑：韩 啸)