计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190600155

一种基于块对角表示和近邻约束的子空间聚类方法

高方远[1]　王秀美[2]
1 北京航空航天大学数学与系统科学学院　北京 102206
2 西安电子科技大学电子工程学院　西安 710071
(fygao.huaa@gmail.com)

摘　要　聚类分析是机器学习与数据挖掘中的重要工具,而子空间聚类是高维数据分析中常用的聚类方法。基于谱图的子空间聚类方法首先学习数据在子空间中的自表示系数矩阵,然后基于此进行谱聚类分析。通过研究子空间聚类的过程和模型设计,发现基于子空间的聚类方法存在难以保持数据非线性和局部几何结构的问题。为此,文中提出了一种可以提取非线性结构的子空间聚类方法。首先,使用非线性映射函数将原始数据空间映射到高维的线性空间,利用块对角表示保持子空间的独立性。此外,为了对聚类过程中数据的局部结构进行约束,文中引入了基于拉普拉斯矩阵的流形正则项。然后,采用3种计算拉普拉斯矩阵的方法设计不同的基于流形正则和块对角约束的非线性子空间聚类模型。最后,在不同数据集上的实验结果验证了所提算法的有效性。

关键词:子空间聚类;块对角约束;非线性映射;流形正则
中图法分类号　TP391

Subspace Clustering Method Based on Block Diagonal Representation and Neighbor Constraint

GAO Fang-yuan[1] and WANG Xiu-mei[2]
1 School of Mathematics and System Science,Beihang University,Beijing 102206,China
2 School of Electronic Engineering,Xidian University,Xi'an 710071,China

Abstract　Clustering is an important tool for machine learning and data mining,and subspace clustering is a popular method in high-dimensional data analysis. Spectral clustering based subspace clustering method learns the self-representation coefficient matrix of data in subspace,and then the spectral clustering is carried out on the coefficient matrix. It is found that the subspace-based clustering cannot deal with nonlinear problem and neglect the local geometric structure of the data. To this end,this paper proposes a new subspace clustering method which first projects the data to a high-dimensional linear space by a nonlinear mapping function and applies a Laplacian-based manifold regularization constraint on the subspace clustering model to preserve the local structure of the data at the same time. Three kinds of Laplacian matrix are used to establish the different nonlinear subspace clustering models based on manifold regularization and block diagonal constraints. Experimental results on different data sets show the effectiveness of the proposed methods.

Keywords　Subspace clustering,Block diagonal constraint,Nonlinear mapping,Manifold regularization

1　引言

随着信息时代的到来,各种传感器收集的数据越来越丰富,样本的维度也随之升高。对高维数据进行分析,提取其内在结构是一项非常有意义但又非常有挑战性的任务。聚类是数据挖掘与机器学习的重要手段之一,在计算机视觉、模式识别等领域有着广泛的应用。常见的聚类方法有k-means方法[1]、谱聚类方法[2]和子空间聚类方法[3]等。相比k-means和谱聚类,子空间聚类方法因在处理高维数据时有较好的鲁棒性而受到广泛关注。基于谱聚类的子空间学习是一种以谱图理论为基础的聚类方法,根据子空间上线性自表示的性质,先学习高维数据在低维子空间上的表示,再进行聚类分析。该方法处理高维复杂数据的能力更强,聚类结果也更好。

现有的子空间聚类方法大致可分为4类:基于统计学习的方法[4-5]、基于矩阵分解的方法[6]、基于代数的方法[7]和基于谱聚类的方法[8-10]。其中,基于谱聚类的方法因为形式简单,便于计算,能够处理噪声和奇异点等而受到广泛关注,成为子空间算法中最常用的方法。基于谱聚类的子空间聚类方

到稿日期:2019-06-26　返修日期:2019-09-13　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61972305,61871308,61772402);国家大学生创新创业训练计划项目经费;陕西省自然科学基金(2019JM-090)
This work was supported by the National Natural Science Foundation of China(61972305,61871308,61772402),College Students' Innovative Entrepreneurial Training Plan Program and Natural Science Basic Research Plan in Shaanxi Province of China(2019JM-090).
通信作者:王秀美(wangxm@xidian.edu.cn)

法使用谱聚类算法作为聚类框架，需要学习数据在子空间的自表示系数矩阵，然后根据该矩阵进行谱聚类分析。在高维空间中，来自某子空间的数据点可以由其他数据点的线性或仿射组合表示。基于此，Elhamifar 等提出了稀疏子空间聚类(Sparse Subspace Clustering，SSC)[8]方法，在数据无噪声且子空间互相独立的前提下，通过对系数矩阵施加 ℓ_1 范数来学习每一个数据点的最稀疏表示，求得一个结构简洁、特征明确的稀疏系数矩阵，从而提高谱聚类的分析精度。虽然 SSC 方法中稀疏约束的方法可以更好地划分子集，但所得系数矩阵可能太稀疏，因此对相关度高的数据的聚类效果不好。基于此，Liu 等提出了基于低秩表示(Low Rank Representation，LRR)的子空间聚类方法[9]，该方法利用矩阵的核范数寻求所有数据的最低秩表示，得到的低秩矩阵能够更好地保持自表示系数矩阵的内在结构，从而优化并提升了聚类性能。LRR 方法能更好地分析相关数据，但是由于该方法获得的表示系数矩阵很稠密，故缺乏对不同类样本的剔除能力，不能简洁有效地反映原始数据的空间结构，影响了聚类的效果。上述两种方法的目的都是学习一个满足块对角条件的子空间自表示系数矩阵，以保证原始数据空间的固有结构，从而提高子空间聚类的效果。但这两种方法对噪声数据都比较敏感，且只有在子空间独立的情况下才能获得较好的结果。为此，Lu 等提出一种基于块对角约束的子空间聚类方法(Subspace Clustering by Block Diagonal Representation，BDR)[11]，在聚类目标函数中直接加入对系数矩阵的块对角约束。该方法更好地保证了原始数据空间的固有结构，提高了聚类效果。但上述几种子空间聚类模型仍存在两方面不足：1)原始数据空间的固有结构可能是非线性的，不能直接在子空间上进行准确的自表示；2)缺少对数据流形结构的保护，得到的数据自表示系数矩阵的判别性较差。

基于上述分析，本文首先使用一个非线性映射函数将原始数据空间映射到一个更高维的线性可分空间上进行聚类分析，建立一个非线性子空间聚类模型；然后在基于块对角约束的子空间聚类模型上施加一个流形正则约束项，来保持子空间聚类过程中数据的局部几何结构。进一步，为了研究基于不同度量方法的流形正则条件对聚类结果的影响，本文通过研究数据相似度计算的方法，选取高斯核方法[12]、基于局部尺度(Local-Scale，LS)[13]的相似度度量方法和基于共享最近邻(Shared Nearest Neighbors，SNN)[14]的相似度度量方法来计算拉普拉斯矩阵，从而建立了 3 种基于不同流形正则和块对角约束的非线性子空间聚类模型。本文提出的方法能得到判别性更强的自表示系数矩阵，同时能够保持数据的固有空间结构和局部相似性关系，提高了聚类的效果。模型优化采用了交替极小化方法，实验结果表明了所提方法的有效性。

本文的主要贡献如下：

1)利用非线性映射，将原始数据空间映射到更高维度的线性可分的空间上，以方便处理高维非线性数据，保证子空间的可分性；

2)提出了一种基于块对角约束的子空间聚类方法，尽可能地保持子空间的独立性；

3)利用拉普拉斯正则保持数据的局部结构，同时分析了 3 种相似性度量方法在构造拉普拉斯矩阵时对聚类结果的影响。

2 相关工作

子空间聚类的目标是对高维数据在子空间上的低维表示进行聚类。基于谱聚类的子空间聚类算法主要包含两个步骤：首先，学习原始数据空间在子空间上的自表示系数矩阵；然后，根据所得的自表示系数矩阵构建相似度矩阵进行谱聚类分析，获得聚类结果。

给定数据矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in R^{d\times n}$，其中 $\boldsymbol{x}_i$ 为 $\boldsymbol{X}$ 的列向量，每个 $\boldsymbol{x}_i$ 代表 $\boldsymbol{x}$ 数据集中的一个样本；d 表示每个样本的维度；n 表示数据集 $\boldsymbol{X}$ 中样本的个数。则子空间聚类的模型可以表示为如下形式：

$$\begin{aligned}&\min\ \Omega(\boldsymbol{Z})+\Phi(E)\\&\text{s.t. }\boldsymbol{X}=\boldsymbol{X}\boldsymbol{Z}+\mathrm{E}\end{aligned}\tag{1}$$

其中，$\boldsymbol{Z}\in R^{n\times n}$ 为子空间自表示系数矩阵，$E\in R^{d\times n}$ 为噪声项，$\Omega(\cdot)$ 表示对自表示系数矩阵构造正则项，$\Phi(\cdot)$ 表示对噪声矩阵构造正则项。优化上述目标函数得到最优系数矩阵 $\boldsymbol{Z}$，根据 $\boldsymbol{Z}$ 构建相似度矩阵进行谱聚类分析，即可得到最终的聚类结果。

Elhamifar 等提出的 SSC 算法采用 ℓ_1 范数构造 $\Omega(\cdot)$ 约束项，目标函数为：

$$\begin{aligned}&\min_{\boldsymbol{Z},E}\|\boldsymbol{Z}\|_1+\lambda\|E\|_1\\&\text{s.t. }\boldsymbol{X}=\boldsymbol{X}\boldsymbol{Z}+E\end{aligned}\tag{2}$$

其中，$\|\cdot\|_1$ 表示矩阵 ℓ_1 范数，第一项最小化该范数可以求得一个稀疏矩阵，第二项为噪声正则项。该方法能学习一个符合数据内在结构的稀疏系数矩阵，以提升聚类效果。但是该方法对噪声比较敏感，不适用于相关性强的数据。

针对上述问题，Liu 等提出了 LRR 方法，采用低秩约束方法构造目标函数，并将 NP 难的低秩约束优化松弛为一个核范数的目标函数，该目标函数为：

$$\begin{aligned}&\min_{\boldsymbol{Z},E}\|\boldsymbol{Z}\|_*+\|E\|_{2,1}\\&\text{s.t. }\boldsymbol{X}=\boldsymbol{X}\boldsymbol{Z}+E\end{aligned}\tag{3}$$

其中，$\|\cdot\|_*$ 为矩阵核范数，$\|\cdot\|_{2,1}$ 表示矩阵 2.1 范数。LRR 在具有相关性的数据聚类方面优于 SSC 方法，但两种方法均要求子空间尽可能独立，且对噪声数据的处理不够好，算法鲁棒性不高。

为此，Lu 等提出块对角约束理论，建立了基于块对角约束(BDR)的子空间聚类模型，其目标函数如下：

$$\begin{aligned}&\min_{\boldsymbol{Z}}\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_F^2+\gamma\|\boldsymbol{Z}\|_k\\&\text{s.t. }diag(\boldsymbol{Z})=0,\boldsymbol{Z}\geqslant 0,\boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}}\end{aligned}\tag{4}$$

其中，k 表示样本集数据的类数；$\|\cdot\|_k$ 表示 k 块对角约束范数；$\|\boldsymbol{Z}\|_k=\sum_{i=n-k+1}^{n}\lambda_i(\boldsymbol{L}_{\boldsymbol{Z}})$，且 $\boldsymbol{L}_{\boldsymbol{Z}}=Diag(\boldsymbol{Z}\cdot\mathbf{1})-\boldsymbol{Z}$，$Diag(\cdot)$ 表示将向量对角化为矩阵，$\mathbf{1}$ 表示元素全为 1 的列向量，$\lambda_i(\boldsymbol{L}_{\boldsymbol{Z}})$ 表示 $\boldsymbol{L}_{\boldsymbol{Z}}$ 的特征值从大到小排序的第 i 个特征值。$diag(\cdot)$ 表示将矩阵对角化。式(4)的第二项即表

示对 $\boldsymbol{Z}$ 的块对角约束。

BDR 模型的优化过程采用交替极小化的方法。Lu 等不仅详细介绍了该方法的理论推导过程，还通过实验验证了该方法的性能。该模型学习的块对角的系数矩阵能更好地保持自表示系数矩阵的固有结构，聚类效果更好，算法鲁棒性高。但是该方法仍有不足，如数据空间原始结构的非线性不利于子空间表示，数据局部几何结构难以保证等。

基于上述不同目标函数求得相应的自表示系数矩阵 $\boldsymbol{Z}$ 后，采用下式建立相似度矩阵：

$$\boldsymbol{S}=\frac{|\boldsymbol{Z}|+|\boldsymbol{Z}|^{\mathrm{T}}}{2} \tag{5}$$

根据所得相似度矩阵，使用谱聚类方法，即可得到相应的子空间聚类结果。

3 基于块对角约束的子空间聚类方法

本节介绍提出的子空间聚类方法的构造过程，同时将 3 种相似性度量函数用于构造拉普拉斯正则，最后给出目标函数优化步骤。

3.1 目标函数

3.1.1 流形正则化

现有的谱聚类方法常采用流形正则约束来保持数据点的相似性关系，即数据的局部几何结构，流形正则约束条件可以概括为一个迹形式的约束：

$$\arg\min_{\boldsymbol{z}} \mathrm{tr}(\boldsymbol{ZLZ}^{\mathrm{T}}) \tag{6}$$

其中，$\boldsymbol{Z}$ 为子空间自表示系数矩阵；$\boldsymbol{L}$ 为拉普拉斯矩阵，计算方式为 $\boldsymbol{L}=Diag(\boldsymbol{S}\cdot\boldsymbol{1})-\boldsymbol{S}$，$\boldsymbol{S}$ 表示相似度矩阵。

相似矩阵是保持流形正则条件有效性的重要因素。对于数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，其中任意两个数据点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的相似度 $\boldsymbol{S}_{i,j}$ 的计算方式有以下几种。

(1)在不具有任何关于数据集的先验知识的情况下，构建数据集对应的相似度矩阵最常用的方法是 Gaussian 核函数法[12]：

$$\boldsymbol{S}_{i,j}=\exp\left(\frac{-d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)}{2\sigma^2}\right) \tag{7}$$

其中，$d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2$，$d(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 表示第 i 个数据点和第 j 个数据点之间的欧氏距离；σ 为尺度参数。

(2)由于 Gaussian 核函数中，参数 σ 根据数据集的不同而取不同值，并且数据集中数据点的分布是多重尺度的，因此，Zelnik-Manor 等提出了基于局部尺度的相似度度量[13]：

$$\boldsymbol{S}_{i,j}=\begin{cases}\exp(-d(\boldsymbol{x}_i,\boldsymbol{x}_j)^2/(\sigma_i\sigma_j)), & i\neq j\\ 1, & i=j\end{cases} \tag{8}$$

其中，$\sigma_i=d(\boldsymbol{x}_i,\boldsymbol{x}_m)$ 表示第 i 个数据点和第 m 个数据点的欧氏距离，$\sigma_j=d(\boldsymbol{x}_j,\boldsymbol{x}_m)$ 表示第 j 个数据点和第 m 个数据点的欧氏距离。

(3)由于基于局部相似度度量未充分利用数据点邻域信息，为了解决这个问题，Liu 等提出了基于共享最近邻的相似度度量[14]。

设数据点 $\boldsymbol{x}_i$ 的最近邻是 $N(\boldsymbol{x}_i)$，数据点 $\boldsymbol{x}_j$ 的最近邻是 $N(\boldsymbol{x}_j)$，则数据点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的共享最近邻数目为：

$$SNN(\boldsymbol{x}_i,\boldsymbol{x}_j)=|N(\boldsymbol{x}_i)\cap N(\boldsymbol{x}_j)| \tag{9}$$

相似矩阵计算公式为：

$$\boldsymbol{S}_{i,j}=\begin{cases}\exp(-d(\boldsymbol{x}_i,\boldsymbol{x}_j)^2/(\sigma_i\sigma_j(SNN(\boldsymbol{x}_i,\boldsymbol{x}_j)+1))), & i\neq j\\ 1, & i=j\end{cases} \tag{10}$$

其中，$\sigma_i\sigma_j(SNN(\boldsymbol{x}_i,\boldsymbol{x}_j)+1)$ 表示 σ_i，σ_j 和 $(SNN(x_i,x_j)+1)$ 的乘积。则两个样本点的共享近邻样本点越多，两个样本点的相似度越大。

本文分别使用上述 3 种不同的相似度矩阵计算方法，构造 3 种基于流形假设和块对角约束的子空间聚类算法。总体目标函数描述如下：

$$\begin{aligned}&\min_{\boldsymbol{z}}\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{XZ}\|_F^2+\lambda_1\|\boldsymbol{Z}\|_k+\frac{\lambda_2}{2}\mathrm{tr}(\boldsymbol{ZLZ}^{\mathrm{T}})\\ &\text{s.t. } diag(\boldsymbol{Z})=\boldsymbol{0},\boldsymbol{Z}\geqslant 0,\boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}}\end{aligned} \tag{11}$$

3.1.2 非线性映射

子空间聚类最基本的假设是原始数据可以在其子空间上进行线性自表示。因此，子空间聚类过程除了对自表示系数矩阵和噪声矩阵构造的目标函数进行优化，更重要的是要保证数据结构的线性化。

根据 Lee 等的研究[18]，当原始空间线性不可分时，通过非线性映射的方法可以将原始空间映射到一个高维的空间，进而实现数据的线性划分。基于此，本文使用非线性函数映射将原始低维数据空间映射到一个高维线性空间中，在该高维线性空间中进行子空间自表示系数的学习，非线性映射后的子空间自表示的形式如下所示：

$$\phi(\boldsymbol{x})=\phi(\boldsymbol{X})\boldsymbol{Z} \tag{12}$$

从数据原始空间的线性可分性、系数矩阵内在结构的保持和数据局部流形结构的保持等方面考虑，本文建立了一种基于流形假设和块对角约束的非线性子空间聚类模型。总体目标函数如下所示：

$$\begin{aligned}&\min_{\boldsymbol{z}}\frac{1}{2}\|\phi(\boldsymbol{X})-\phi(\boldsymbol{X})\boldsymbol{Z}\|_F^2+\lambda_1\|\boldsymbol{Z}\|_k+\frac{\lambda_2}{2}\mathrm{tr}(\boldsymbol{ZLZ}^{\mathrm{T}})\\ &\text{s.t. } \mathrm{diag}(\boldsymbol{Z})=\boldsymbol{0},\boldsymbol{Z}\geqslant 0,\boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}}\end{aligned} \tag{13}$$

3.2 目标函数优化

上述目标函数中，需要优化的变量是自表示系数矩阵 $\boldsymbol{Z}$，但是该变量对应的子问题无法直接求解，因此引入 $\boldsymbol{Z}$ 的辅助变量 $\boldsymbol{B}$，则原目标函数转化为：

$$\begin{aligned}&\min_{\boldsymbol{z}}\left\{\frac{1}{2}\|\phi(\boldsymbol{X})-\phi(\boldsymbol{X})\boldsymbol{Z}\|\alpha^2\sigma_F+\lambda_1\|\boldsymbol{B}\|_k+\right.\\ &\left.\frac{\lambda_2}{2}\mathrm{tr}(\boldsymbol{ZLZ}^{\mathrm{T}})+\frac{\lambda_3}{2}\|\boldsymbol{Z}-\boldsymbol{B}\|_F^2\right\}\\ &\text{s.t. } diag(\boldsymbol{B})=\boldsymbol{0},\boldsymbol{B}\geqslant 0,\boldsymbol{B}=\boldsymbol{B}^{\mathrm{T}}\end{aligned} \tag{14}$$

采用交替迭代的方法寻找变量 $\boldsymbol{B}$ 和变量 $\boldsymbol{B}$ 的局部最优解，过程如下。

当 $\boldsymbol{B}$ 固定时，目标函数转化为 $\boldsymbol{Z}$ 的子问题：

$$\begin{aligned}&\min_{\boldsymbol{z}}\frac{1}{2}\|\phi(\boldsymbol{X})-\phi(\boldsymbol{X})\boldsymbol{Z}\|_F^2+\frac{\lambda_2}{2}\mathrm{tr}(\boldsymbol{ZLZ}^{\mathrm{T}})+\\ &\frac{\lambda_3}{2}\|\boldsymbol{Z}-\boldsymbol{B}\|_F^2\\ &\text{s.t. } diag(\boldsymbol{B})=\boldsymbol{0},\boldsymbol{B}\geqslant 0,\boldsymbol{B}=\boldsymbol{B}^{\mathrm{T}}\end{aligned} \tag{15}$$

对变量 $\boldsymbol{Z}$ 求导并令其为 0 得到：

$$-\phi(\boldsymbol{X})^{\mathrm{T}}\phi(\boldsymbol{X})+\phi(\boldsymbol{X})^{\mathrm{T}}\phi(\boldsymbol{X})\boldsymbol{Z}+\lambda_2(\boldsymbol{ZL})+\lambda_3(\boldsymbol{Z}-\boldsymbol{B})=0 \tag{16}$$

记 $\phi(\boldsymbol{X})^{\mathrm{T}}\phi(\boldsymbol{X})$ 为核 $\boldsymbol{K}$，则上式为：

$$(\boldsymbol{K}+\lambda_3\boldsymbol{I})\boldsymbol{Z}+\lambda_2\boldsymbol{ZL}-\boldsymbol{K}-\lambda_3\boldsymbol{B}=0 \tag{17}$$

根据 Sorensen 等的研究[16]，得出该方程是标准的 Sylvester 方程，具有唯一解。

当 $\boldsymbol{Z}$ 固定时，由文献[11]中的定理 5 可知：

$$\begin{aligned}&\sum_{i=n-k+1}^{n}\lambda_i(\boldsymbol{L})=\min_{\boldsymbol{W}}\langle\boldsymbol{L},\boldsymbol{W}\rangle\\&\text{s. t. } 0\leq\boldsymbol{W}\leq I,\mathrm{tr}(\boldsymbol{W})=k\end{aligned} \tag{18}$$

则关于 $\boldsymbol{B}$ 的子问题可变化为：

$$\begin{aligned}&\min_{\boldsymbol{B}}\lambda_1\langle Diag(\boldsymbol{B}\mathbf{1})-\boldsymbol{B},\boldsymbol{W}\rangle+\frac{\lambda_3}{2}\|\boldsymbol{Z}-\boldsymbol{B}\|_F^2\\&\text{s. t. } diag(\boldsymbol{B})=\mathbf{0},\boldsymbol{B}\geqslant 0,\boldsymbol{B}=\boldsymbol{B}^{\mathrm{T}},0\leq\boldsymbol{W}\leq I,\mathrm{tr}(\boldsymbol{W})=k\end{aligned} \tag{19}$$

其中，$\boldsymbol{W}$ 的近似解为 $\boldsymbol{UU}^{\mathrm{T}}$，$\boldsymbol{U}$ 是由 $Diag(\boldsymbol{B}\mathbf{1})-\boldsymbol{B}$ 的 k 个最小的特征值对应的特征向量组成的 $n\times k$ 的矩阵。

根据 Lu 等的分析，可求得 $\boldsymbol{B}$ 的局部最优解为：

$$\boldsymbol{B}=\left[\frac{(\hat{\boldsymbol{A}}+\hat{\boldsymbol{A}}^{\mathrm{T}})}{2}\right]_+ \tag{20}$$

其中，$\boldsymbol{A}=\boldsymbol{Z}-\frac{\lambda_1}{\lambda_3}(diag(\boldsymbol{W})\mathbf{1}^{\mathrm{T}}-\boldsymbol{W})$，$\hat{\boldsymbol{A}}=A-Diag(diag(\boldsymbol{A}))$，$[\boldsymbol{C}]_+=\begin{cases}\boldsymbol{c}_{ij}, & \text{if } \boldsymbol{c}_{ij}>0\\0, & \text{if } \boldsymbol{c}_{ij}\leqslant 0\end{cases}$。

两个子问题交替优化求解，直到达到收敛条件。其中，收敛条件为[11]：

$$\max\{\|\boldsymbol{Z}^{t+1}-\boldsymbol{Z}^t\|_\infty,\|\boldsymbol{B}^{t+1}-\boldsymbol{B}^t\|_\infty\}\leqslant\varepsilon \tag{21}$$

其中，t 表示迭代次数；ε 为预先设定的阈值误差，一般取为 1×10^{-3}。

根据求得的系数矩阵 $\boldsymbol{Z}$ 构造相似矩阵进行谱聚类，即可得到最终聚类结果。

4 实验结果

本节使用常用的数据库验证算法的有效性。

4.1 聚类结果

为了验证算法的性能和效果，在不同数据集上进行测试实验。在实验中，核 K 采用线性核进行计算。选取 UCI 数据库[17]中的 Iris，Glass 和 WDBC，以及 YALE-B[18]，mnist-fou[19] 5 个数据集。其中，mnist-fou 表示 mnist 数据集中傅里叶系数特征视图上的数据集。几个数据集的具体信息如表 1 所列。

表 1 数据集特征

Table 1 Features of datasets

编号	数据集	样本数目	类别数	特征维度
1	Iris	150	3	4
2	Glass	214	6	9
3	WDBC	569	2	30
4	YALE-B	2414	38	1024
5	mnist-fou	2000	10	76

分别将本文 3 个基于不同流形正则的算法与谱聚类方法(SC)、稀疏子空间聚类方法(SSC)、低秩子空间聚类方法(LRR)和基于块对角约束的聚类方法(DBR)在上述数据集上进行对比实验，以比较几种算法的聚类性能。使用聚类准确度指标(Accuracy，ACC)[20]来衡量聚类效果。为了保证结果的可靠性，每个方法均进行 10 次循环，取均值进行比较。实验结果如表 2 所列。

表 2 聚类结果

Table 2 Clustering results

方法	Iris	Glass	WDBC	YALE-B	mnist-fou
SC	0.6800	0.4589	0.8910	0.3236	0.6118
SSC	0.8127	0.4907	0.8752	0.7105	0.4902
LRR	0.9297	0.5294	0.8981	0.5650	0.6025
BDR	0.9600	0.5308	0.8805	0.4312	0.6787
BDR+Gaussian	**0.9800**	**0.5374**	0.8998	0.7249	**0.7205**
BDR+LS	0.9733	**0.5374**	**0.9367**	0.7154	0.6935
BDR+SNN	0.9733	**0.5374**	**0.9367**	**0.7353**	0.6920

从表 2 可以看出，相比其他算法中最优的聚类结果，本文提出的算法在 5 个数据集上的聚类结果仍有至少 0.8% 的提高，由此可见本文算法的聚类性能更好。与 BDR 方法相比，本文算法的聚类结果始终较高，说明流形正则约束条件对聚类效果的提高具有重要作用，能很好地保持数据局部几何结构；另外，BDR 方法在 YALE-B 数据集上的聚类效果比传统 SSC 方法和 LRR 方法略差。对于本文的 3 种不同的流形正则构造方法，其聚类结果各有优劣，3 种方法在结构简单的数据集上性能没有太大的差别，在稍大数据集上则需从更多方面进行分析。最后，实验采用多次循环的方法处理数据，得到的聚类结果方差近似为 0，说明在参数不变的情况下，本文算法具有稳定性。

4.2 参数分析

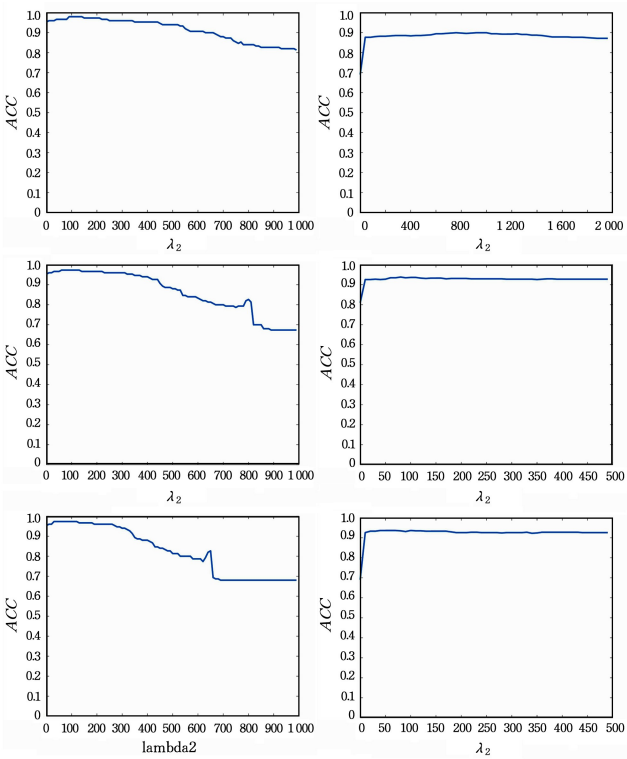
本节给出本文方法在各个数据集上的实验参数选取，如表 3 所列，其中，m_1 表示采用基于局部尺度的相似度度量计算数据集相似度矩阵的数据点的近邻个数，m_2 表示采用基于共享最近邻的相似度度量计算数据集相似度矩阵的数据点的近邻个数。

表 3 各数据集上的参数选取

Table 3 Parameter choice of each dataset

方法	参数	Iris	Glass	WDBC	YALE-B	Mnist-fou
BDR+Gaussian	λ_1	0.01	0.001	0.001	0.001	0.001
	λ_2	100	100	1000	5000	100
	λ_3	10000	10000	10	1000	12500
	σ	0.25	0.25	10000	1000	0.1
BDR+LS	λ_1	0.01	0.001	0.001	0.001	0.001
	λ_2	100	100	100	12500	100
	λ_3	1000	10000	1	1000	11000
	m_1	6	8	8	8	8
BDR+SNN	λ_1	0.01	0.001	0.001	0.001	0.001
	λ_2	100	100	100	12500	100
	λ_3	10000	10000	10	1000	12600
	m_2	6	8	8	8	8

进一步，分析参数设置对算法性能的影响，本文算法主要使用流形正则的方法来提高聚类效果，因此对流形正则条件的参数 λ_2 进行实验，在其他参数固定的情况下，设置不同的 λ_2，观察聚类结果。图 1 给出了 Iris 数据集和 WDBC 数据集上不同参数设置下的聚类结果。



(a)Iris 数据集 (b)WDBC 数据集

图 1 不同参数设置下的聚类准确度

Fig. 1 Clustering accuracy under different parameter settings

结束语 本文主要针对子空间聚类中的数据空间非线性、数据局部几何结构难保持等问题，提出一种基于流形正则和块对角约束的非线性子空间聚类算法。在流形条件的设计方面，采用 3 种不同的相似矩阵计算方法，研究不同的流形正则约束项对算法的影响。在算法优化过程，使用核方法处理目标函数的非线性映射部分，采用交替极小化方法优化得到子空间自表示系数矩阵。最后，通过在 5 个不同维数数据集上的实验结果验证了该算法的有效性。

参 考 文 献

[1] MAC Q J. Some methods for classification and analysis of multivariate observations[C]// Proceedings of Berkeley Symposium on Mathematical Statistics and Probability. 1967:281-297.

[2] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm[C]// Proceedings of the International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2001:849-856.

[3] VIDAL R. Subspace clustering [J]. IEEE Signal Processing Magazine, 2011, 28(2):52-68.

[4] FISCHLER M, BOLLES R R. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Journal of ACM, 1981, 24(6):381-395.

[5] MA Y, DERKSEN H, HONG W, et al. Segmentation of multivariate mixed data via lossy coding and compression[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(9):1546-1562.

[6] LU L, VIDAL R. Combined central and subspace clustering for computer vision applications[C]// Proceedings of International Conference on Machine Learning. New York: ACM Press, 2006: 593-600.

[7] VIDAL R, MA Y, SASTRY S. Generalized principal component analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(12):1-15.

[8] ELHAMIFAR E, VIDAL R. Sparse subspace clustering [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2009:2790-2797.

[9] LIU G, LIN Z, YAN S, et al. Robust recovery of subspace structures by low-rank representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1):171-184.

[10] GAO X, ZHANG K, TAO D, et al. Image super-resolution with sparse neighbor embedding[J]. IEEE Transactions on Image Processing, 2012, 21(7):3194-3205.

[11] LU C, FENG J, LIN Z, et al. Subspace clustering by block diagonal representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2):487-501.

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

[13] ZELNIK-MANOR L, PERONA P. Self-tuning spectral clustering[C]// Proceedings of the International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2004: 1601-1608.

[14] LIU X Y, LI J W, YU H, et al. Adaptive spectral clustering based on shared nearest neighbors[J]. Journal of Chinese Computer Systems, 2011, 32:1876-1880.

[15] TOLIĆD, ANTULOV-FANTULIN N, KOPRIVA I. A nonlinear orthogonal non-negative matrix factorization approach to subspace clustering[J]. Pattern Recognition, 2018, 82:40-55.

[16] SORENSEN D C, ANTOULAS A C. The sylvester equation and approximate balanced reduction[J]. Linear Algebra and its Applications, 2002, 351/352:671-700.

[17] https://archive.ics.uci.edu/ml/datasets.html.

[18] LEE K C, HO J, KRIEGMAN D J. Acquiring linear subspaces for face recognition under variable lighting[J]. IEEE Transactions on Pattern Recognition and Machine Intelligence, 2005, 27(5):684-698.

[19] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceeding of IEEE, 1998, 86(11):2278-2324.

[20] WANG F, ZHAO B, ZHANG C. Linear time maximum margin clustering[J]. IEEE Transactions on Neural Networks, 2010, 21(2):319-332.

GAO Fang-yuan, born in 2000, underduate student. His current research interests include clustering analysis and image processing.

WANG Xiu-mei, born in 1978, professor. Her main research interests include statistical machine learning and image processing.