计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800180

# 基于二阶近邻的核子空间聚类

王中元　刘惊雷
烟台大学计算机与控制工程学院　山东 烟台 264005
(79186799@qq.com)

**摘　要**　高维数据集的处理是计算机视觉领域的核心,子空间聚类是实现高维数据聚类使用最广泛的方法之一。传统的子空间聚类假定数据来自不同的线性子空间,且不同子空间的区域不重叠。然而,现实中的数据往往不满足这两个约束条件,使得子空间聚类的效果受到影响。为了解决这两个问题,引入核化子空间来解决子空间数据的非线性问题,引入子空间系数矩阵的二阶近邻来处理重叠的子空间问题。随后,设计了基于二阶近邻的核化子空间三步聚类算法,首先求取核化子空间数据的自相似系数,然后消除子空间的重叠区域,最后对系数矩阵进行谱聚类。将所设计的子空间聚类算法首先在人工数据集上进行了测试,随后在人脸、场景字符和生物医学3类数据集中共12个真实数据集上进行了实验。实验结果表明,所提算法相比最新的几种算法具有一定的优势。

**关键词**:交替方向乘子法;图像识别;核方法;二阶近邻;子空间聚类
**中图法分类号**　TP18

## Kernel Subspace Clustering Based on Second-order Neighbors

WANG Zhong-yuan and LIU Jing-lei
School of Computer and Control Engineering, Yantai University, Yantai, Shandong 264005, China

**Abstract**　The processing of high-dimensional data sets is the focus of computer vision. Subspace clustering is one of the most widely used methods to achieve high-dimensional data clustering. The traditional subspace clustering assumes that the data comes from different linear subspaces, and different subspace regions do not overlap. However, real data often do not meet these two constraints, which affects the effect of subspace clustering. In order to deal with these two problems, this paper introduces a kernelized subspace to solve the nonlinear problem of subspace data, and introduces the second-order neighbors of the subspace coefficient matrix to deal with the overlapping subspace problem. Then a three-step clustering algorithm based on second-order neighbors of the kernelized subspace is designed. Firstly, the self-similarity coefficients of the kernelized subspace data are obtained. Secondly, the overlapping regions of the subspaces are eliminated. Finally, the coefficient matrix is spectrally clustered. In this paper, the designed subspace clustering algorithm is first tested on three artificial data sets, and then the experiment is performed on 12 real data sets, including face, scene characters and biomedical data sets. Experimental results show that the proposed algorithm has certain advantages over the latest algorithms.

**Keywords**　Alternating direction multiplier method, Image identification, Kernel method, Second-order neighbors, Subspace clustering

## 1　引言

在过去的几十年中,人们提出了许多基于子空间聚类的方法,并成功地将其应用于计算机视觉以及图像处理中。这些方法大致可以分为以下4类:代数方法[1]、统计方法[2]、迭代方法[3]和基于谱聚类的方法[4]。这些方法的局限性比较明显,如代数方法和迭代方法都需要预先设置子空间的维度。当子空间独立时,这些方法表现良好。但在实际应用中,子空间往往是重叠的,上述方法无法同时优化子空间的稀疏性和连通性,因而受到限制。统计方法取决于准确的子空间模型,然而真实数据集往往来源广泛,具有多种数据结构,直接对原始数据集进行处理很难达到很好的效果。因此,目前的主要问题是如何对非线性原始数据集进行处理,同时还要优化子空间的稀疏性和连通性,解决重叠的子空间问题,从而满足实际应用的需要。

随着计算机技术的飞速发展,生成和存储的数据迅速增

到稿日期:2020-08-27　返修日期:2020-09-18　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(61572419,61773331,61703360,61801414)
This work was supported by the National Natural Science Foundation of China (61572419,61773331,61703360,61801414).
通信作者:刘惊雷(jinglei_liu@sina.com)

加，如何有效处理多种多样的数据已成为当下的热点。如支持向量机(Support Vector Machine，SVM)[5]之类的核方法得到了广泛关注，因为它可以将非线性的数据映射到适合的高维特征空间中变成线性数据来处理，能够有效学习原始数据特征。对于高维空间中核矩阵的处理，本文使用类内聚敛、类间分离的方法来处理核矩阵，这样可以约束核矩阵的类内关系和类外关系，提高子空间聚类的准确度。

基于谱聚类的算法是子空间聚类中最常用的算法，因为它容易实现并且受数据初始化和数据损坏的影响较小。大多数传统算法通过解决以下自表达问题来计算系数矩阵 $Z$：

$$\min_{Z} L(XZ,X)+\lambda \| Z \|_{\xi}$$

$$\text{s.t. } \operatorname{diag}(Z)=0$$

其中，$X$ 表示数据集，$\xi$ 代表不同范数的约束。上式中，第一项用于从其他数据中重构每个数据，第二项用于系数矩阵的正则化。不同的 $\xi$ 可以对系数矩阵进行不同的约束，例如 $l_2$ 范数或核范数可以保证子空间内的连通性，$l_0$ 范数和 $l_1$ 范数则可以保证子空间的稀疏性，但是由于子空间是非独立性的，难以在一个模型中同时实现这两种约束。本文受鲁棒的启发式算法中处理自表达弱连接方法[6]的影响，使用具有关键连接的二阶近邻，可以有效约束子空间的非独立性，在保持连通性的同时极大地提高了稀疏性。

总之，我们提出了一种基于二阶近邻的核子空间聚类算法(Kernel Subspace Clustering Based on Second-Order Neighbors，KSCSN)，其主要过程如图 1 所示。KSCSN 将核方法与二阶近邻相结合，共同对传统的基于谱聚类的算法进行优化。首先，它通过核方法将非线性原始数据映射到高维空间中变成线性数据来处理，可以有效学习原始数据特征。其次，通过二阶近邻对系数矩阵进行优化，可以有效解决重叠的子空间问题，同时优化子空间的稀疏性和连通性，大大提高了子空间聚类的准确性。最后，对其进行子空间聚类，并在人脸、场景字符和生物医学数据集上进行测试。

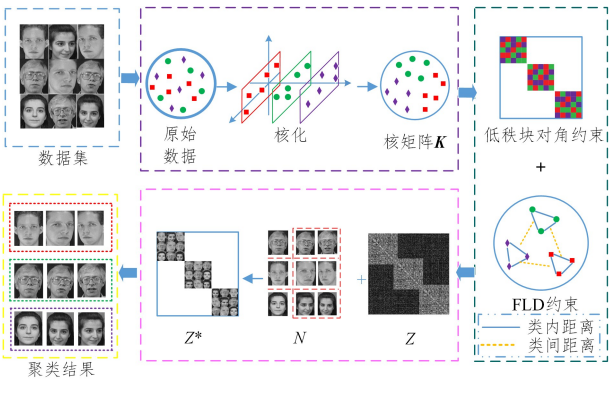



图 1 KSCSN 的整体框架

Fig. 1 Framework of KSCSN

相比传统的子空间聚类算法，本文设计了一种基于二阶近邻的核子空间聚类算法。本文的特点和贡献如下。

(1)引入了一种子空间聚类的框架，能够有效地解决子空间数据的非线性问题。该框架通过核方法将非线性的原始数据映射到合适的高维特征空间中并转化为线性数据来处理，同时能够将转换后的核矩阵划分为不同的特征子集，有效提高了子空间聚类的准确度。

(2)提出了一种基于基于二阶近邻的核子空间聚类算法(KSCSN)，能够准确地对各个子空间重叠部分的数据进行聚类。KSCSN 通过将二阶近邻与子空间聚类相结合，来优化子空间的稀疏性和连通性，以提高聚类准确度。同时，本文设计了一种基于交替方向乘子法(Alternating Direction Method of Multipliers，ADMM)的迭代算法，可以快速有效地解决最终的优化问题。

(3)在人脸、场景字符和生物医学这 3 类数据集上进行了实验验证。实验结果表明，相比传统的子空间聚类算法，本文设计的 KSCSN 算法在聚类精度上有明显提高，同时通过核方法和二阶近邻的优化，为子空间聚类问题提供了新的思路。

本文第 2 节介绍了相关工作，阐述了目前常用的一些核方法和子空间聚类算法；第 3 节阐述了本文所提算法所涉及的相关定义；第 4 节设计了 KSCSN；第 5 节对所提出的 KSCSN 进行了理论分析；第 6 节测试了 KSCSN 在人脸、场景字符和生物医学数据集上的聚类效果；最后总结全文并展望未来。

## 2 相关工作

本节首先简要介绍子空间聚类和核方法的相关知识，然后探讨目前国际上的一些先进方法。

高维数据聚类作为数据挖掘的重点和难点，已经得到了广泛研究。子空间聚类是实现高维数据聚类的有效方法，是传统聚类算法[7-8]在高维数据空间中的应用，主要是对相关维进行局部搜索。传统的聚类算法很难在高维数据空间中进行聚类，主要原因是：1)高维数据集中存在大量噪声，想要在原始空间中进行聚类几乎是不可能的；2)相比低维空间，高维空间中的数据分布要稀疏很多，很多情况下数据间的距离几乎是相等的，而绝大多数传统聚类方法都是基于距离进行聚类，这在高维空间中很难实现。为了解决这个问题，Agrawal 等首先提出了子空间聚类的概念，用于解决高维数据的聚类问题[9]。

子空间聚类算法的核心是利用了子空间的自表达属性[10]，即子空间中的每一个点都可以由同一子空间中其他点的线性组合来表示。例如，给定 $N$ 个数据点 $X=[x_1,\cdots,x_N]\in\mathbb{R}^{D\times N}$，其中 $X=XZ$ 即为子空间的自表达属性，$Z$ 是系数矩阵。由于各个子空间是相对独立的，因此当数据点 $i$ 和 $j$ 来自同一子空间时，可以得到具有块对角结构的 $Z$ 的最优解，$z_{ij}=0$。即通过求解以下优化问题来求解 $Z$：

$$\min_{Z} \| Z \|_{q} \quad \text{s.t. } X=XZ \tag{1}$$

一般来说，子空间聚类通过使用各类范数来对系数矩阵 $Z$ 进行正则化，如核范数[11]、$l_1$ 范数[12]、$l_2$ 范数[13]以及 $l_{21}$ 范数[14]。Lu 等将 Frobenius 范数应用于最小二乘回归(Least Squares Regression，LSR)方法，用于构建高度相关的数据结构[15]。You 等使用弹性网正则化来约束子空间的连通性[16]。类似地，Xu 等提出了一种加权的稀疏子空间聚类算法(Re-

weighted Sparse Subspace Clustering, RSSC), 它通过使用 $l_1$ 范数来逼近 $l_0$ 范数最小化问题, 优化了各个子空间之间的潜在连通性[17]。

上述方法只是基于谱聚类的子空间聚类的一般步骤, 并不能解决数据集来源广泛且具有多种特征结构的问题。本文使用核方法[18]对非线性原始数据矩阵进行处理。核方法是处理复杂的非线性数据的一种有效途径, 其主要思想是通过合适的核函数(通常采用高斯核函数)将原始数据映射到合适的高维特征空间中进行处理。核方法具有明显的优势: 1)其中使用的核函数是针对实际应用设计的, 有利于集成数据的先验知识; 2)核方法在将原始数据映射到高维空间时用到了核技巧, 使得计算复杂度与高维特征空间的维数无关。Zhang 等对传统的低秩表示方法进行核化, 并采用 Frobenius 范数约束误差项, 但此方法对隐式特征空间异常值十分敏感[19]。为了实现鲁棒性, Patel 等将数据映射到潜在空间中进行去噪, 然后进行聚类, 但这限制了数据的连通性[20]。Xiao 等提出了鲁棒非线性的低秩表示方法, 该方法虽然可以有效解决特征空间中的鲁棒聚类问题, 但是计算复杂度很高[21]。

为了解决上述子空间稀疏性及连通性的优化问题, 有效处理子空间边缘重叠区域的数据, 本文引入二阶近邻, 并将其与核方法相结合进行优化, 具体过程如图 1 所示。首先, 使用核方法, 通过高斯核函数将原始数据映射到合适的高维特征空间中并转化为线性核矩阵 $K$, 有效解决了子空间数据的非线性问题。其次, 求取核矩阵 $K$ 的系数矩阵 $Z$, 在此过程中本文使用类内聚敛、类间分离的原则对其进行约束。然后, 通过二阶近邻对有重叠的系数矩阵 $Z$ 进行优化, 其中 $N$ 为寻找左侧人脸数据的二阶近邻的过程, 最终得到新的无重叠的系数矩阵 $Z^*$。二阶近邻可以有效解决子空间边缘重叠的问题, 从而提高聚类准确度。最后, 对得到的系数矩阵 $Z^*$ 进行谱聚类即可。

## 3 相关定义及方法

本节首先介绍文中使用的一些基本符号(见表 1), 然后对本文算法所涉及的相关定义和概念进行简单的介绍。

表 1 基本符号

Table 1 Basic symbols

| 符号 | 含义 |
|---|---|
| $A_{ij}$ | 矩阵 $A$ 的第$(i,j)$项 |
| $A_i$ | 矩阵 $A$ 的第 $i$ 行 |
| $A_j$ | 矩阵 $A$ 的第 $j$ 列 |
| $\|A\|$ | 矩阵 $A$ 的行列式 |
| $I$ | 单位矩阵 |
| $A\geqslant 0$ | 矩阵 $A$ 所有项均为非负数 |
| $\mathrm{Tr}(A)$ | 矩阵 $A$ 的迹 |
| $A=U\Sigma V^{\mathrm{T}}$ | 矩阵 $A$ 的 SVD 分解 |
| $D=\mathrm{diag}(\sum_j A_{ij})\ \forall i$ | 块对角矩阵 |
| $\|A\|_0$ | 矩阵 $A$ 的 $l_0$ 范数 |
| $\|A\|_F=\sqrt{\sum_{ij}A_{ij}^2}$ | 矩阵 $A$ 的 Frobenius 范数 |
| $\|A\|_1=\sum_{ij}\|A_{ij}\|$ | 矩阵 $A$ 的 $l_1$ 范数 |
| $\|A\|_{21}=\sum_j\|[A]_{:j}\|$ | 矩阵 $A$ 的 $l_{21}$ 范数 |

**定义 1**(自表达属性[10]) 来自多个子空间的并集的每个数据实例可以通过其他数据实例的线性组合来表示, 其被称为自表达属性。即:

$$X=XZ$$

(1)本文的目标函数中应包含子空间聚类的框架。最简单的子空间聚类的目标函数如式(1)所示, 通过子空间的自表达属性为原始数据 $X$ 构造一个自表示 $Z$, 同时使用各类范数对其进行正则化。然而, 真实数据集中往往有各种各样的噪声, 这种方法的效果往往很不理想。为了对数据中的噪声进行约束, 本文增加了对误差矩阵 $E$ 的约束。因此, 式(1)可以重新表述为:

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

其中, $\lambda$ 为平衡参数。本文使用 $l_1$ 范数对误差矩阵 $E$ 进行约束, 可以更好地处理随机噪声。同时, 用核范数对系数矩阵 $Z$ 进行约束, 可以有效保证子空间内的连通性。

(2)为了提高聚类准确度, 本文进一步使用类内聚敛、类间分离的方法(FLD 约束)对系数矩阵 $Z$ 进行约束。

假设给定样本 $X=[X_1,\cdots,X_C]\in\mathbb{R}^{d\times n}$, 其中每个列为样本向量, 每个类别的样本中包含子矩阵 $X_i\in\mathbb{R}^{d\times n_i}$, 其中 $n_i$ 为第 $i$ 类($i=1,\cdots,C$)的样本个数。

本文将相同的类尽可能地聚集在一起, 而将不同的类尽可能地分离。这可以通过式(3)进行约束:

$$\begin{aligned}&\min_Z\lambda_1\|A\odot Z\|_F^2+\lambda_2\|D\odot Z\|_0\\&\text{s.t. } X=XZ\end{aligned}\tag{3}$$

其中, $\lambda_1$ 和 $\lambda_2$ 是用于约束类内、类外表示的权衡参数, $\odot$ 表示 Hadamard 乘积, 且 $X\in\mathbb{R}^{d\times n}$。具体来说, 第一项是为了得到最大的类间距离, 其中:

$$A=1_n1_n^{\mathrm{T}}-Y$$

$$Y=\begin{bmatrix}1_{n_1}1_{n_1}^{\mathrm{T}} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1_{n_C}1_{n_C}^{\mathrm{T}}\end{bmatrix}$$

第二项是为了得到最小的类内距离, $d_{ij}$ 用于计算 $x_i$ 和 $x_j$ 之间的距离。目前有很多计算距离的方法, 然而这些方法往往复杂度较高。本文使用欧氏距离来计算样本之间的距离, 即 $\|x_i-x_j\|_2^2$, 因为这种方法计算简单且准确度较高。由于想要最小化 $l_0$ 范数是 NP 难问题, 本文用 $l_1$ 范数代替 $l_0$ 范数进行计算。因此, 问题(3)可以重新表述为:

$$\begin{aligned}&\min_Z\lambda_1\|A\odot Z\|_F^2+\lambda_2\|D\odot Z\|_1\\&\text{s.t. } X=XZ\end{aligned}\tag{4}$$

通过整合式(2)和式(4), 可以得到:

$$\begin{aligned}&\min_Z\|Z\|_*+\lambda_1\|A\odot Z\|_F^2+\lambda_2\|D\odot Z\|_1+\lambda_3\|E\|_{21}\\&\text{s.t. } X=XZ+E\end{aligned}\tag{5}$$

(3)在式(5)的基础上结合核方法。核方法可以通过核函数将原始数据映射到合适的高维特征空间, 从而有效地解决子空间数据的非线性问题。设 $\Phi:\mathbb{R}\to\mathbb{H}$ 是从输入空间到 $Hilbert$ 空间的映射, $K\in\mathbb{R}^{n\times n}$ 为半正定核矩阵。

$K_{ij}=(\Phi(x)^{\mathrm{T}}\Phi(x))_{ij}=\Phi(x_i)^{\mathrm{T}}\Phi(x_{,j})=ker(x_i,x_{,j})$

其中，$ker$ 是核函数 $\Phi(x_1),\cdots\Phi(x_n)$。

首先，式(5)中的误差矩阵 $E$ 可以表示为 $E=X-XZ$，由此可以得到：

$$\min_Z\|Z\|_*+\lambda_1\|A\odot Z\|_F^2+\lambda_2\|D\odot Z\|_1+\lambda_3\|X-XZ\|_{21} \tag{6}$$

其次，定义一个变量 $S=I-Z\in\mathbb{R}^{n\times n}$，可得到 $\|X-XZ\|_{21}=\|XS\|_{21}=\sum_{i=1}^{n}\|Xs_i\|=\sum_{i=1}^{n}(s_i'X'Xs_i)^{1/2}$，其中 $S=[s_1,\cdots,s_n]$。因此，式(6)可以改写为：

$$\min_Z\|Z\|_*+\lambda_1\|A\odot Z\|_F^2+\lambda_2\|D\odot Z\|_1+\lambda_3\sum_{i=1}^{n}\sqrt{s_i'X'Xs_i}\quad \text{s.t. } S=I-Z \tag{7}$$

然后，将式(7)中的 $X$ 替换为 $\Phi(X)$，可以得到：

$$\min_Z\|Z\|_*+\lambda_1\|A\odot Z\|_F^2+\lambda_2\|D\odot Z\|_1+\lambda_3\sum_{i=1}^{n}\sqrt{s'_i\Phi(X)'\Phi(X)s_i}\quad \text{s.t. } S=I-Z \tag{8}$$

最后，引入函数 $g(S)\triangleq\sum_{i-1}^{n}\sqrt{s_i'Ks_i}$，基于以上 3 点，通过整合式(2)、式(4)和式(8)，可以得到 KSCSN 步骤 1 的优化目标。

$$\min_Z\|Z\|_*+\lambda_1\|A\odot Z\|_F^2+\lambda_2\|D\odot Z\|_1+\lambda_3 g(S)\quad \text{s.t. } S=I-Z \tag{9}$$

其中，核矩阵 $K$ 包含在 $g(S)$ 中。

步骤 1　通过交替方向乘子法(ADMM)对优化目标进行迭代求解，其中还涉及求解矩阵秩的最小化。计算矩阵秩的最小化是一个 NP 难题，目前通常使用核范数来求近似解，即核范数最小化可以表示为：

$$X^*=\arg\min_X f(X)+\Gamma\|X\|_*$$

其中，$X=\mathbb{R}^{m\times n}$，$\Gamma>0$ 是正则化参数。

**定义 2**(奇异值阈值(SVT)[22])　求解形如下式的核范数最小化(NNM)问题：

$$X^*=\arg\min_X\Gamma\|X\|_*+\frac{1}{2}\|X-A\|_F^2$$

需要计算奇异值阈值算子 $S_\Gamma(\cdot)$ 给出的闭式解：

$$X=S_\Gamma(A)=U_AS_\Gamma=(\Sigma_A)V_A^{\mathrm{T}}$$

其中，$S_\Gamma(x)=\mathrm{sgn}(x)\cdot\max(|x|-\Gamma,0)$ 是软收缩算子，$U_A\Sigma_AV_A^{\mathrm{T}}$ 是矩阵 $A$ 的 SVD 分解。

步骤 2　通过二阶近邻计算系数矩阵 $Z$ 的邻居，从而获得更准确的系数矩阵 $Z^*$。通过二阶近邻可以同时优化子空间的稀疏性和连通性，解决重叠的子空间问题，提高子空间聚类的准确度。

首先，通过图 2 所示的例子来简单阐述二阶近邻的基本思想。图 2 中，假设红色和紫色的两根线条表示两个有交集的子空间 $N_1$ 和 $N_2$，数据点 $A_1$ 位于子空间 $N_1$ 中，数据点 $A_2$，$A_3$ 为 $A_1$ 的两个邻居，分别位于两个子空间中。可以明显看出，位于同一子空间的两个数据点 $A_1$，$A_3$ 在该子空间上的切线构成的夹角 $\alpha_1$ 明显小于不同子空间的两个数据点 $A_1$，$A_2$ 在该子空间上的切线构成的夹角 $\alpha_2$。本文提出的二阶近邻采用了这种思想，对数据点的邻居进行进一步求导，得到了更深层次的数据特征，从而将原本难以划分的子空间重叠区域的数据准确聚类。

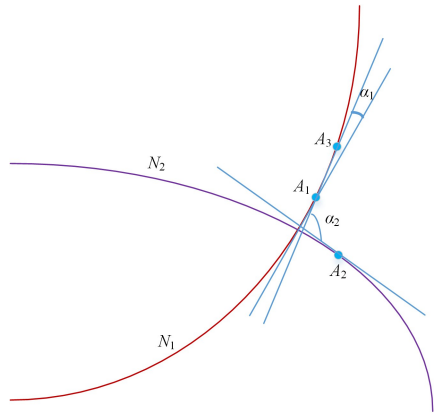



图 2　二阶近邻的局部示意图(电子版为彩色)

Fig. 2　Local schematic diagram of second-order neighbors

其次，本文对二阶近邻的具体概念进行了阐述。前文提到，$x_i$ 是由包含 $x_j$ 在内的点的线性组合表示的，但是通过步骤 1 求解得到的系数矩阵 $Z$ 不一定是对称的。因此，引入非负矩阵 $W$，表达式如下：

$$W=\frac{1}{2}(|Z|+|Z|^{\mathrm{T}}) \tag{10}$$

通过 $W$ 可以保证对称性。随后，在 $W$ 中寻找 $x_i$ 的 $k$ 个邻居。

**定义 3**(邻居[23])　对于每个样本 $x_i$，它的 $k$ 个邻居组成的集合 $N_k(x_i)\in\mathbb{R}^{1\times k}$ 可以通过如下公式计算得到：

$$N_k(x_i)=\{x_j\}_{j=1}^{k}=\arg\max_{x_j}\sum_{j=1}^{k}|w_{ij}|$$

其中，$w_{ij}$ 表示矩阵 $W$ 的第 $(i,j)$ 项，$\mu\leqslant k$ 且 $i\in\mathbb{R}^{\mu}$ 是指数集。

本文中的 $k$ 个邻居保留了 $w_{ij}$ 最大的 $\gamma$ 项，其中 $\gamma<k<n$，$n$ 为子空间中的样本数。$w_{ij}$ 表示 $x_i$ 与 $x_j$ 之间的相似度，$w_{ij}$ 越大，相似度就越高。虽然使用邻居对 $Z$ 进行约束可以保证子空间的稀疏性和其他特性，但是并没有考虑连通性。这样各个子空间的边缘容易被噪声影响，从而导致聚类结果不准确。为了同时约束子空间的稀疏性和连通性，解决子空间边缘重叠问题，本文使用二阶近邻对 $Z$ 进行约束。

**定义 4**(二阶近邻[23])　给定 $x_i$ 的 $k$ 个邻居 $N_k(x_i)$，若 $x_j$ 为 $x_i$ 的二阶近邻，则存在满足以下条件的 $\mu$ 个样本 $\{x_{j_l}\}_{l=1}^{\mu}\subset N_k(x_j)$。

$$\prod_{l=1}^{\mu}1_{x_i\in N_k(x_{jl})}=1$$

其中，$\mu\leqslant k$ 且 $i\in\mathbb{R}^{\mu}$ 是指数集。

从定义 4 可以看出，若 $x_j$ 为 $x_i$ 的二阶近邻，那么在它们所有的邻居中至少有 $\mu$ 个公共邻居。这些二阶近邻约束了子空间之间的潜在联系。对于每个样本，本文从所有 $k$ 个邻居中保留了少数具有潜在联系的二阶近邻，而不是 $w_{ij}$ 的最大数量，从而增强了子空间的连通性。

根据定义 4，若想得到 $x_i$ 的二阶近邻 $x_j$，则需要计算 $x_i$ 的 $k$ 个邻居。通过式(11)可得到 $x_i$ 的 $k$ 个邻居：

$$N_k(x_i)=\{x_j\}_{j=1}^{k}=\arg\max_{x_j}\sum_{j=1}^{k}|w_{ij}| \tag{11}$$

接着，为了判定 $x_j\in N_k(x_i)$ 是否为 $x_i$ 的二阶近邻，根据二阶近邻的定义，需要遍历所有的 $N_k(x_i)$，这是 NP 难的。本文设 $\mu=1$，即 $x_j$ 和 $x_i$ 至少有一个公共邻居。更具体地，$x_i$ 与它的一个二阶近邻之间的路径必须包含 $x_i$ 的一个邻居。为此，本文引入 $s_{ij}$ 作为 $x_i$ 和 $x_j$ 的评判标准。

$$s_{ij}=\sum_{i=1}^{k}1_{x_i\in N_k(x_{ji})}$$

若 $s_{ij}>1$，则 $x_i$ 是 $x_j$ 的一个二阶近邻。

然后，根据式(12)计算新的系数矩阵 $Z^*$：

$$z_{ij}^{*}=\begin{cases}(w_{ij})/(\sum_j w_{ij}), & \text{if } x_j\in N_k(x_i)\\ 0, & \text{if } x_j\notin N_k(x_i)\end{cases} \tag{12}$$

步骤 3　通过二阶近邻对系数矩阵 $Z$ 进行优化得到 $Z^*$，可以同时提升子空间的稀疏性和连通性，有效解决重叠的子空间问题，提高子空间聚类的准确性。

步骤 4　对系数矩阵 $Z^*$ 进行谱聚类。首先根据系数矩阵 $Z^*$ 构建拉普拉斯矩阵；然后计算拉普拉斯矩阵的前 $c$ 个特征值和特征向量，构建特征向量空间；最后通过 $k$-means 算法对特征向量空间中的特征向量进行聚类。

## 4　基于二阶近邻的核子空间聚类算法

基于二阶近邻的核子空间聚类算法如算法 1 所示。

**算法 1**　基于二阶近邻的核子空间聚类算法

输入：特征矩阵 X；参数 $\lambda_1,\lambda_2,\lambda_3$；距离测量矩阵 D

输出：集合 $V_1,\cdots,V_k$

初始化：J=0，Z=0，Q=0，S=0，$\lambda_1,\lambda_2,\lambda_3>0$，$C_1=0$，$C_2=0$，$C_3=0$，$\mu_{\max}=10^8$，$\mathrm{tol}=10^{-6}$，$\rho=1.15$；

1. 计算系数矩阵 Z

while

$\max(\|I-Z^{t+1}-S^{t+1}\|_\infty,\|J^{t+1}-Z^{t+1}\|_\infty,\|Q^{t+1}-Z^{t+1}\|_\infty)>\mathrm{tol}$

do

1.1. 通过式(16)更新系数矩阵 Z；

1.2. 通过式(17)更新辅助变量 J；

1.3. 通过式(19)更新辅助变量 Q；

1.4. 通过式(20)更新辅助变量 S；

1.5. 更新拉格朗日乘子 $C_1$，$C_2$ 和 $C_3$；

$$\begin{cases}C_1^{t+1}=C_1^t+\mu^t(I-Z^{t+1}-S^{t+1})\\ C_2^{t+1}=C_2^t+\mu^t(J^{t+1}-Z^{t+1})\\ C_3^{t+1}=C_3^t+\mu^t(Q^{t+1}-Z^{t+1})\end{cases}$$

1.6. 更新 $\mu$

$$\mu^{t+1}=\min(\mu_{\max},\rho\mu^t)$$

end

2. 通过二阶近邻得到 $Z^*$

2.1. 通过式(10)计算非负矩阵 W；

2.2. 通过式(11)得到 k 个邻居 $N_k$；

2.3. 通过式(12)得到系数矩阵 $Z^*$；

3. 对 $Z^*$ 进行谱聚类

3.1. 计算 $Z^*$ 的拉普拉斯矩阵；

3.2. 得到最小的 c 个特征值和特征向量；

3.3. 用 k-means 得到聚类结果 $V_1,\cdots,V_k$.

为了优化问题(9)，本文首先引入两个辅助变量 $J$ 和 $Q$ 来使问题可分离，然后将问题(9)改写为：

$$\min_{J,Z,Q,S}\|J\|_*+\frac{\lambda_1}{2}\|A\odot Z\|_F^2+\lambda_2\|D\odot Q\|_1+\lambda_3 g(S)$$
$$\text{s.t. } S=I-Z,J=Z,Q=Z,Z\geqslant 0 \tag{13}$$

然后，可以通过增广拉格朗日乘子法得到问题(13)的增广拉格朗日函数。

$$\begin{aligned}&L(J,Z,Q,S,C_1,C_2,C_3)\\&=\|J\|_*+\frac{\lambda_1}{2}\|A\odot Z\|_F^2+\lambda_2\|D\odot Q\|_1+\lambda_3 g(S)+\\&\quad\langle C_1,I-Z-S\rangle+\langle C_2,J-Z\rangle+\langle C_3,Q-Z\rangle+\\&\quad\frac{\mu}{2}(\|J-Z\|_F^2+\|I-Z-S\|_F^2+\|Q-Z\|_F^2)\end{aligned} \tag{14}$$

其中，$\langle J,Q\rangle=\mathrm{tr}(J^{\mathrm{T}}Q)$。$C_1$，$C_2$ 和 $C_3$ 是拉格朗日乘子，$\mu>0$ 是惩罚参数。在每次迭代时最小化一个增广拉格朗日乘子，即在固定其余变量的情况下最小化一个变量。详细的迭代步骤如下：

(1)更新 $Z$。固定其他变量并通过解决以下问题来更新 $Z$。

$$\begin{aligned}L&=\min_Z\frac{\lambda_1}{2}\|A\odot Z\|_F^2+\langle C_1^t,I-Z-S^t\rangle+\langle C_2^t,J^{t+1}-Z\rangle+\\&\quad\langle C_3^t,Q^t-Z\rangle+\frac{\mu^t}{2}(\|J^{t+1}-Z\|_F^2+\|I-Z-S^t\|_F^2+\\&\quad\|Q^t-Z\|_F^2)\\&=\min_Z\frac{\lambda_1}{2}\|A\odot Z\|_F^2+\frac{\mu^t}{2}\left(\left\|I-Z-S^t+\frac{C_1^t}{\mu^t}\right\|_F^2+\right.\\&\quad\left.\left\|J^{t+1}-Z+\frac{C_2^t}{\mu^t}\right\|_F^2+\left\|Q^t-Z+\frac{C_3^t}{\mu^t}\right\|_F^2\right)\end{aligned}$$

这相当于：

$$\begin{aligned}L=\min_Z\frac{\lambda_1}{2}\|Z-R\|_F^2+\frac{\mu^t}{2}&\left(\left\|I-Z-S^t+\frac{C_1^t}{\mu^t}\right\|_F^2+\right.\\&\left.\left\|J^{t+1}-Z+\frac{C_2^t}{\mu^t}\right\|_F^2+\left\|Q^t-Z+\frac{C_3^t}{\mu^t}\right\|_F^2\right)\end{aligned} \tag{15}$$

其中，$R=[Y,0_{n(N-n)}]\odot Z^t$。通过推导 $\partial L/\partial Z=0$，可以很容易得到 $Z$ 的最优解，而问题(15)的闭式解通过以下形式给出：

$$\begin{aligned}Z^{t+1}&=\left[\left(2+\frac{\lambda_1}{\mu_t}\right)I+\Phi(X)^{\mathrm{T}}\Phi(X)\right]^{(-1)}\\&\quad\left(\frac{\lambda_1}{\mu_t}R+U_1+U_2+U_3\right)\\&=\left[\left(2+\frac{\lambda_1}{\mu_t}\right)I+K\right]^{(-1)}\left(\frac{\lambda_1}{\mu_t}R+U_1+U_2+U_3\right)\end{aligned} \tag{16}$$

其中，$U_1=I-S^t+\frac{C_1^t}{\mu^t}$，$U_2=J^{t+1}+\frac{C_2^t}{\mu^t}$，$U_3=Q^t+\frac{C_3^t}{\mu^t}$。

(2)更新 $J$。当固定其他变量时，式(14)的目标函数可以表示为 $J$ 的函数，即：

$$\begin{aligned}J^{t+1}&=\arg\min_J\|J\|_*+\langle C_2^t,J-Z^t\rangle+\frac{\mu^t}{2}\|J-Z^t\|_F^2\\&=\|J\|_*+\frac{\mu^t}{2}\left\|J-(Z^t-\frac{C_2^t}{\mu^t})\right\|_F^2\end{aligned}$$

通过奇异值阈值算子(见定义 2)，可以得到一个封闭形式的解决方案，即：

$$J^{t+1}=\Gamma_{\frac{2}{\mu^t}}\left(Z^t-\frac{C_2^t}{\mu^t}\right)=US_{\frac{1}{\mu^t}}(\Sigma)V^{\mathrm{T}} \tag{17}$$

其中,$U\Sigma V^{\mathrm{T}}$ 是 $(Z^t-C_2^t/\mu^t)$ 的奇异值分解,$S_{\frac{1}{\mu^t}}(\cdot)$ 是软阈值算子[18]。

(3)更新 $Q$。当固定其他变量时,式(14)的目标函数可以表示为 $Q$ 的函数,即:

$$\begin{aligned}L&=\min_{Q}\lambda_2\|D\odot Q\|_1+\langle C_3^t,Q-Z^{t+1}\rangle+\frac{\mu^t}{2}\|Q-Z^{t+1}\|_F^2\\&=\lambda_2\|D\odot Q\|_1+\frac{\mu^t}{2}\left\|Q-\left(Z^{t+1}-\frac{C_3^t}{\mu^t}\right)\right\|_F^2\end{aligned}\tag{18}$$

显然,问题(18)可以等效为解决 $n\times N$ 个子问题。对于第 $i$ 行第 $j$ 列元素 $Q_{ij}$,问题(18)的最优解是:

$$\begin{aligned}Q_{ij}^{t+1}&=\arg\min_{Q_{ij}}\lambda_2 D_{ij}|Q_{ij}|+\frac{\mu^t}{2}(Q_{ij}-M_{ij})^2\\&=S_{\frac{\lambda_2 D_{ij}}{\mu^t}}(M_{ij})\end{aligned}\tag{19}$$

其中,$M_{ij}=Z_{ij}^{t+1}-(C_3^t)_{ij}/\mu^t$。

(4)更新 $S$。当固定其他变量时,式(14)的目标函数可以表示为 $S$ 的函数,即:

$$\begin{aligned}L&=\min_{S}\lambda_3 g(S)+\langle C_1^t,I-Z^{t+1}-S\rangle+\\&\quad\frac{\mu^t}{2}\|I-Z^{t+1}-S\|_F^2\\&=\min_{S}\lambda_3 g(S)+\frac{\mu^t}{2}\left\|I-Z^{t+1}-\frac{C_1^t}{\mu^t}\right\|_F^2\end{aligned}$$

接下来,令 $\Gamma=I-Z^{t+1}-C_1^t/\mu^t$,那么 $S^{t+1}$ 的第 $i$ 行为:

$$S^{t+1}(i,:)=\begin{cases}\dfrac{\|\Gamma^i\|_2-\frac{\lambda_3}{\mu^t}}{\|\Gamma^i\|_2}\Gamma^i, & \text{if } \|\Gamma^i\|_2\geqslant\dfrac{\lambda_3}{\mu^t}\\0, & \text{if } \|\Gamma^i\|_2\leqslant\dfrac{\lambda_3}{\mu^t}\end{cases}\tag{20}$$

其中,$\Gamma^i$ 是矩阵 $\Gamma$ 的第 $i$ 行。

在优化变量 $J,Z,Q$ 和 $S$ 之后,ADMM 算法还需要更新拉格朗日乘子 $C_1,C_2,C_3$ 以及参数 $\mu$,以更快地收敛。算法 1 描述了解决所提出的优化问题(9)的详细过程。

根据算法 1 中描述的步骤 2 和步骤 3,即通过二阶近邻对系数矩阵 $Z$ 进行优化得到 $Z^*$,可以同时优化子空间的稀疏性和连通性,解决了重叠的子空间问题。最后通过谱聚类算法进行聚类,即通过构造亲和矩阵找到数据的低维嵌入,然后通过 $k$-means 聚类得到聚类结果。

## 5 算法的性质分析

### 5.1 算法的收敛性分析

为了解决算法 1 中步骤 1 提出的优化目标,本文使用了 ADMM,详见第 4.1 节。下面证明算法 1 的收敛性。

**定理 1**(算法的收敛性) KSCSN 是收敛的。

证明:传统的 ADMM 算法是为了解决以下问题:

$$\begin{aligned}&\min_{z\in\mathbb{R}^n,w\in\mathbb{R}^m}f(z)+h(w)\\&\text{s.t. } Rz+Tw=\mu\end{aligned}\tag{21}$$

其中,$R\in\mathbb{R}^{p\times n}$,$T=\mathbb{R}^{p\times m}$,$\mu\in\mathbb{R}^p$,$f$ 和 $h$ 是凸函数。显然,问题(21)的 ADMM 可以用于解决矩阵优化问题:

$$\begin{aligned}&\min_{Z\in\mathbb{R}^{n\times N},W\in\mathbb{R}^{m\times M}}f(Z)+h(W)\\&\text{s.t. } RZ+TW=U\end{aligned}\tag{22}$$

其中,$U\in\mathbb{R}^{p\times N}$。问题(22)的增广拉格朗日函数为:

$$\begin{aligned}L_\mu(Z,W,C)=f(Z)+h(W)+\frac{\mu}{2}\|RZ+TW-U\|_F^2+\\\langle C,RZ+TW-U\rangle\end{aligned}\tag{23}$$

其中,$C\in\mathbb{R}^{p\times N}$ 是拉格朗日乘子,$\mu$ 是惩罚系数。

可以看出,问题(13)是问题(22)的一种特殊情况。具体来说,式(13)中的约束条件可以转换为 $RZ+TW=U$ 的形式,其中,$R=(-I_n,-I_n,X_{tr})^{\mathrm{T}}$,$T=\begin{bmatrix}I_n&&\\&I_n&\\&&I_n\end{bmatrix}$,$W=(J,Q,S)^{\mathrm{T}}$,$U=\_(0,0,X)^{\mathrm{T}}$ 并且 $I_n$ 是大小为 $n\times n$ 的单位矩阵。通过这种变换,问题(13)可以被转换为问题(22)。然后,就可以使用 ADMM 算法以交替的方式更新两个原始变量,并迭代地解决问题(23)。

$$\begin{aligned}Z^{t+1}&=\arg\min_{Z\in\mathbb{R}^{n\times N}}L_\mu(Z,W^t,C^t)\\W^{t+1}&=\arg\min_{W\in\mathbb{R}^{m\times N}}L_\mu(Z^{t+1},W,C^t)\\C^{t+1}&=C^t+\mu(RZ^{t+1}+TW^{t+1}-U)\end{aligned}\tag{24}$$

两个原始变量的更新步骤与算法 1 相同。可以看出,式(24)中 $Z$ 的优化等价于式(15)中 $Z$ 的优化。另外,在固定 $Z$ 不变时,式(17)中的 $J$ 和式(19)中的 $Q$ 以及式(20)中的 $S$ 的解是彼此独立的,例如计算 $S^{t+1}$ 取决于 $Z^{k+1}$ 和 $C^{k+1}$ 而不是 $J^{k+1}$ 或 $Q^{k+1}$。因此,可以使用(24)在 $W$ 中进行 $J,Q$ 和 $S$ 的优化。问题(13)是经典 ADMM 问题的一个特殊情况,并且算法 1 具有与经典 ADMM 相同的优化方式,因此,算法 1 相当于使用了两次 ADMM 方法,它的收敛性得以证明。

### 5.2 算法的复杂度分析

**定理 2**(算法的复杂度) KSCSN 的时间复杂度为 $O_k(2n^2N+n^2d+2nN)$,其中 $n$ 为训练样本数,$N$ 为样本总数,$d$ 为样本维数,$k$ 为迭代次数。

证明:算法 1 的主要计算过程是步骤 1 中的迭代过程,因为需要同时进行奇异值分解(SVD)和矩阵运算。因此,当训练样本数 $n$ 和样本总数 $N$ 较大时,KSCSN 的时间复杂度会很高,特别是计算矩阵 $J\in\mathbb{R}^{n\times N}$ 的 SVD 分解需要 $O(n^2N)$ $(N>n)$的复杂度。需要注意的是,由于要计算矩阵的逆,迭代更新 $Z$ 时的时间复杂度为 $O(n^2d+n^2N)$,其中 $d$ 是样本维数。步骤 2 的时间复杂度是 $O(2nN)$。步骤 3 需要对 $Z^*$ 进行 SVD 分解,并且通过 $k$-means 进行聚类,因此时间复杂度为 $O(n^2N+n^2)$。综上,KSCSN 的总时间复杂度为 $O_k(2n^2N+n^2d+2nN)$,其中 $k$ 是迭代次数。

## 6 实验

本节分两个部分对 KSCSN 进行测试。首先在 3 个不同的人工数据集中对 KSCSN 进行测试,然后分别在人脸、场景字符和生物医学数据集上对 KSCSN 进行测试。为了证明 KSCSN 比现有的聚类算法具有更高的准确度,本文针对不同类型的数据集进行了多次实验,并与目前几种先进的聚类算法进行了对比。

### 6.1 实验环境

在实验过程中，本文首先制作了3个不同的人工数据集进行测试，然后使用了3种类型的真实数据集进行测试，包括人脸数据集（YaleB，ORL，AR，JAFFE）、场景字符数据集（SVHN，ICPR，MNIST，MSRA-TD500）和生物医学数据集（LIDC-IDRI，DDSM，Cardiac MRI，MIAS）。图3为3种数据集中的示例图。

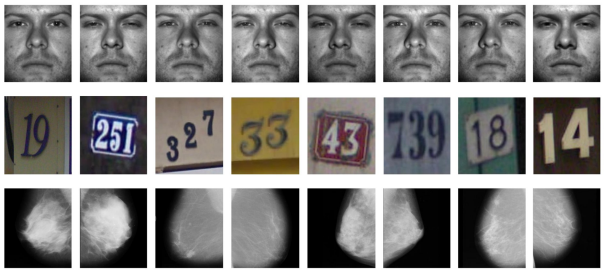


图3 实验中使用的人脸、场景字符和生物医学数据库的示例图片

Fig. 3 Example images of face, handwritten and biomedical datasets used in the experiment

（1）人脸数据集

1）YaleB数据集主要用于身份鉴定，它包含28个人的16128张面部图像，其中每个人的图像都是在9种姿势、64种不同照明方式下获得的，这些人脸图像处于不同的时间和光照，具有不同的表情。

2）ORL数据集共包含10个类别和40个不同的主题，这些图像是在不同的场景、时间、光线和表情下拍摄的。

3）AR数据集共包含120张人的面部图像，每个人都有7张不同的面部图像。人们在图像中的面部表情不同，具有良好的识别能力。

4）JAFFE数据集共包含40个不同的主题，每个主题包含10张不同的面部图像。这些图像是在不同的场景、时间、光线和表情下拍摄的。

（2）场景字符识别数据集

1）SVHN是对阿拉伯数字进行识别的数据集，该数据集中的图像来自真实世界的门牌号，包含训练集和测试集。训练集中包含73257个数字，测试集中包含26032个数字。

2）ICPR数据集为网络图像，主要包括淘宝商家上传到淘宝的一些介绍商品的图像，其中的文本较为清晰，容易识别。

3）MNIST数据集用于手写数字识别。数据集中有60000个训练集和10000个测试集。

4）MSRA-TD500是微软公司使用的一个文本识别数据集，主要包含一些街景图像，环境内容比较复杂，但文本位置比较明显，图像十分清晰且易于识别。

（3）生物医学数据集

1）LIDC-IDRI是由胸部医学图像文件和相应的诊断结果组成的，用于研究一些患者的早期癌症检测。

2）DDSM数据集是由一系列数字乳腺照片组成的。该数据集包含约2500个主题，每个主题包括每个乳房的两张图像，以及一些相关的患者信息和图像信息。

3）Cardiac MRI是心脏病患者心房医疗影像数据集，主要为左心室的外膜和心内膜的图像，共包括33位患者的7980张图像。

4）MIAS数据集包含322张医学图像，该数据库中的图像已经被裁剪为1024×1024的大小。

### 6.2 评估标准

为了有效评估实验结果，本文主要通过3个常用的评估指标将实验结果与其他比较算法进行对比。

（1）准确度（ACC）

$$ACC=\max_{\pi}\frac{1}{n}\sum_{ij}X_{\pi(i)j}^{t}X_{ij}$$

其中，$\pi$是$n$个样本的排列，$X^t$和$X$分别为聚类准确的样本和所有测试样本，如果点$j$属于簇$i$，则它们的第$i$个条目等于1，否则为0。ACC计算的是所有测试样本中分类准确的样本所占的比例。

（2）标准化互信息（NMI）

它通常用于测量两种聚类结果在聚类中的相似度，并计算正确的权重。假设$C$为真实的聚类集，而$C^*$为利用聚类算法计算得到的聚类。它们的互信息$MI(C,C^*)$的定义如下：

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

其中，$p(c_i)$和$p(c_j)$是从数据集中任意选择的数据点，它们分别属于群集$c_i$和$c_j$。$p(c_i,c_j)$是任意选择的联合概率，数据点同时属于群集$c_i$和$c_j$。NMI的定义如下：

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

其中，$H(C)$和$H(C^*)$分别是$C$和$C^*$的熵。NMI越大，聚类性能就越好。

（3）调整兰德系数（ARI）

本文使用的最后一个评估指标是ARI，它用于衡量两个数据分布之间的一致性。ARI的定义如下：

$$ARI=a_{ij}-\frac{\frac{(a_{11}+a_{01})(a_{11}+a_{10})}{a_{00}}}{\frac{(a_{11}+a_{01})(a_{11}+a_{10})}{2}-\frac{(a_{11}+a_{01})(a_{11}+a_{10})}{a_{00}}}$$

其中，$a_{11}$表示将相同类型的样本分配给同一个集合，$a_{00}$表示将不同类型的样本分配给不同的集合，$a_{10}$表示将相同类型的样本分配给不同的集合，$a_{01}$表示将不同类型的样本分配给同一个集合。

### 6.3 实验设计

本文在3个不同的人工数据集上对提出的KSCSN算法进行了测试，目的是证明KSCSN能够准确有效地对子空间边缘重叠区域的数据进行聚类。接下来，本文将KSCSN与目前许多先进的聚类算法进行比较。

首先，将KSCSN与3种不使用核方法的聚类算法进行比较，包括鲁棒的主成分分析（Robust Principal Component Analysis，RPCA）[24]、潜在的低秩表示（Latent Low-Rank Representation，LatLRR）[25]和判别块对角表示学习（Block-Diagonal Low-Rank Representation，BDLRR）[19]。RPCA可以将损坏的数据恢复为具有低秩结构的数据，从而将一个子

空间中的数据分解为低秩和稀疏噪声两部分。LatLRR 主要将低秩表示与子空间聚类相结合，以解决多个子空间的问题。BDLRR 由 Zhang 等[19]提出，通过减少不相关的块外对角元素并添加相关的块对角元素来提高聚类精度。

然后，将 KSCSN 与 3 种使用核方法的聚类算法进行比较，分别是 RKLRR[21]，RKLRS[21] 和 IBDLR[26]。Xiao 等[21]提出了一种基于 LRR 的鲁棒核低秩表示(Robust Kernel Low-Rank Representation，RKLRR)，该方法可以有效地处理非线性数据。Xiao 等还提出了基于 LRS 的鲁棒核低秩表示(Robust Kernel Low-Rank Sparse，RKLRS)，该方法可以有效处理损坏的数据。Xie 等[26]提出了隐式块对角线低秩表示(Implicit Block Diagonal Low-Rank Representation，IBDLR)，该方法将块对角线表示与隐式特征表示相结合，并成功地应用于聚类任务。

与其他核方法类似，本文使用高斯核函数 $K(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|_2^2)$ 对原始数据进行核化，核矩阵已通过 $K(x_i,x_j)=K(x_i,x_j)/\sqrt{K(x_i,x_i)K(x_j,x_j)}$ 进行归一化处理。对于数据集的设置，本文根据表 2 将聚类的数量设置为所有数据集中的实际类别数，并在合理范围内对实验参数进行手动调整以达到较好的实验结果。对于实验中用到的对比算法，本文使用原文献中的初始化方法对数据集进行初始化，从而进行对比实验。

表 2 数据集

Table 2 Data sets

| 数据集类别 | 数据集 | # instances | # features | # classes |
|---|---|---|---|---|
| 人脸数据集 | YaleB | 165 | 1 024 | 15 |
| | ORL | 400 | 1 024 | 40 |
| | AR | 840 | 768 | 120 |
| | JAFFE | 213 | 676 | 10 |
| 场景字符数据集 | SVHN | 258 | 976 | 13 |
| | ICPR | 982 | 2 513 | 44 |
| | MNIST | 1 884 | 2 414 | 58 |
| | MSRA-TD500 | 2 049 | 918 | 33 |
| 生物医学数据集 | LIDC-IDRI | 1 500 | 241 | 22 |
| | DDSM | 1 404 | 320 | 36 |
| | Cardiac MRI | 414 | 6 429 | 130 |
| | MIAS | 878 | 7 454 | 251 |

### 6.4 实验结果

(1)在 3 个人工数据集上，KSCSN 能很好地将 3 个不同的人工数据集进行聚类，可以有效处理子空间边缘重叠区域的数据。在人脸、场景字符和生物医学数据集上，KSCSN 相比其他几种先进的对比算法具有更高的聚类准确度。

(2)在人脸、场景字符和生物医学数据集上，KSCSN 相比其他几种先进的对比算法具有更高的聚类准确度。并且，核方法对非线性数据的处理为聚类提供了良好的条件，可以证明核方法是必要且有效的。

(3)将核方法与二阶近邻相结合，本文提出的子空间聚类算法在图像识别中表现更好。

(4)使用二阶近邻约束子空间的稀疏性和连通性，可以有效解决重叠的子空间问题，大大提高了聚类精度。

综上，本文算法在所有测试数据集上均表现出了良好的性能。KSCSN 主要通过核方法将非线性原始数据映射到合适的高维特征空间中以进行处理，有效解决了子空间数据的非线性问题，然后通过二阶近邻进行子空间聚类，约束了子空间的稀疏性和连通性，解决了重叠的子空间问题，从而提高聚类准确度。

### 6.5 实验分析

首先，从图 4 可以看出，在 3 个人工数据集中，KSCSN 都有不错的聚类结果，这证明 KSCSN 能够有效聚类子空间边缘重叠区域的数据，证明了二阶近邻的必要性。

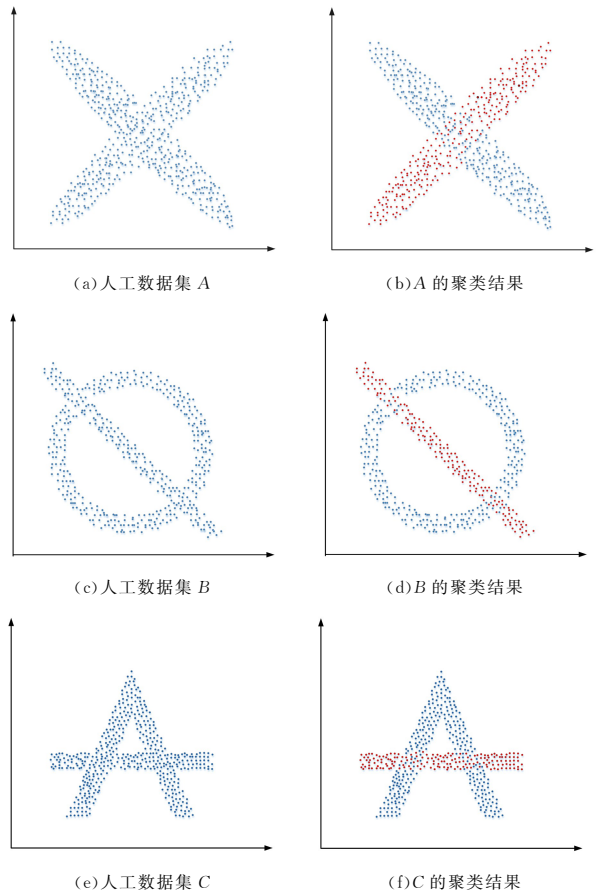



图 4 KSCSN 在 3 个人工数据集上的聚类结果

Fig. 4 KSCSN clustering results on three artificial data sets

其次，在人脸、场景字符和生物医学数据集上，与其他先进对比算法相比，本文提出的 KSCSN 算法可以获得更好的聚类性能。这证明了 KSCSN 可以有效地聚类不同种类的数据。此外，还可以得出结论：当使用核方法和二阶近邻相结合进行聚类时，更有利于图像识别任务的执行。

然后，从表 3－表 5 中可以看出，在这 12 个数据集上所有对比算法中都不是绝对最优的算法，这是因为不同数据集的特征是不同的，并且对于不同的识别任务适合的算法也不一样。但是从表中可以看出，在这些有限训练样本的实验中，KSCSN 优于所有对比算法，主要原因是 KSCSN 具有稀疏性、连通性和核方法的优点。具体而言，一方面，使用核方法将非线性原始数据映射到合适的高维特征空间中进行处理，可以有效解决子空间数据的非线性问题；另一方面，通过二阶近邻对系数矩阵进行优化，可以同时优化子空间的稀疏性和连通性，解决重叠的子空间问题，大大提高了子空间聚类的准确性。

表 3 各方法在不同数据集中的准确度

Table 3 ACC of different methods on different data sets

(单位:%)

| Data | RPCA | LatLRR | BDLRR | RKLRR | RKLRS | IBDLR | KSCSN |
|---|---|---|---|---|---|---|---|
| YaleB | 92.73 | 93.42 | 95.26 | 94.08 | 95.27 | 95.62 | **96.04** |
| ORL | 90.00 | 91.75 | 94.50 | 94.50 | 95.25 | 96.25 | **97.63** |
| AR | 86.76 | 88.83 | 92.63 | 90.88 | 91.90 | 93.11 | **95.44** |
| JAFFE | 92.83 | 93.89 | 95.76 | 95.89 | 96.65 | 96.59 | **97.38** |
| SVHN | 82.04 | 85.67 | 84.24 | 85.35 | 90.23 | 91.14 | **93.24** |
| ICPR | 80.02 | 85.24 | 89.36 | 90.65 | 90.22 | 92.06 | **92.87** |
| MNIST | 87.00 | 86.07 | 90.21 | 93.62 | 93.21 | 94.98 | **96.71** |
| MSRA-TD500 | 78.21 | 79.14 | 80.23 | 82.88 | 83.20 | 84.73 | **88.42** |
| LIDC-IDRI | 62.78 | 65.78 | 64.76 | 67.75 | 69.24 | 70.58 | **73.66** |
| DDSM | 64.56 | 68.34 | 67.17 | 68.27 | 73.25 | 74.24 | **76.17** |
| Caridac MRI | 59.44 | 60.29 | 65.58 | 67.57 | 67.51 | 69.53 | **70.88** |
| MIAS | 63.76 | 61.83 | 58.47 | 64.88 | 67.90 | 70.86 | **71.46** |

表 4 各方法在不同数据集中的标准化互信息

Table 4 NMI of different methods on different data sets

(单位:%)

| Data | RPCA | LatLRR | BDLRR | RKLRR | RKLRS | IBDLR | KSCSN |
|---|---|---|---|---|---|---|---|
| YaleB | 88.94 | 91.33 | 91.86 | 94.50 | 95.25 | 94.50 | **96.03** |
| ORL | 90.28 | 88.73 | 89.24 | 92.63 | 92.76 | 90.88 | **96.72** |
| AR | 91.44 | 91.08 | 94.36 | 95.76 | 94.83 | 95.89 | **95.11** |
| JAFFE | 90.22 | 89.63 | 91.42 | 93.87 | 95.36 | 95.96 | **97.03** |
| SVHN | 81.24 | 82.47 | 82.64 | 84.24 | 86.21 | 85.35 | **87.28** |
| ICPR | 79.68 | 83.33 | 83.24 | 89.36 | 88.44 | 90.65 | **93.91** |
| MNIST | 87.00 | 83.27 | 86.07 | 90.21 | 91.04 | 93.62 | **94.37** |
| MSRA-TD500 | 80.27 | 78.43 | 81.76 | 80.23 | 79.31 | 82.88 | **89.38** |
| LIDC-IDRI | 61.74 | 63.58 | 64.51 | 64.76 | 68.76 | 67.75 | **72.44** |
| DDSM | 65.57 | 67.86 | 68.22 | 67.17 | 68.35 | 68.27 | **75.63** |
| Caridac MRI | 58.33 | 61.84 | 60.61 | 65.58 | 64.26 | 67.57 | **71.35** |
| MIAS | 61.87 | 63.03 | 62.59 | 58.47 | 68.21 | 64.88 | **72.26** |

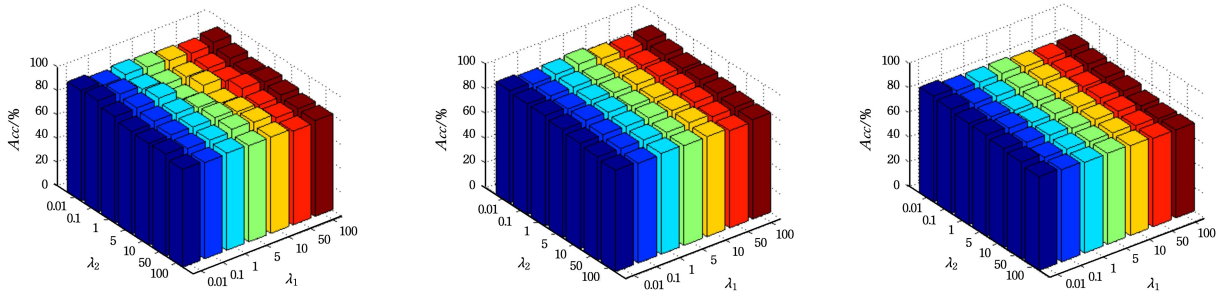
表 5 各方法在不同数据集中的调整兰德系数

Table 5 ARI of different methods on different data sets

(单位:%)

| Data | RPCA | LatLRR | BDLRR | RKLRR | RKLRS | IBDLR | KSCSN |
|---|---|---|---|---|---|---|---|
| YaleB | 71.82 | 73.34 | 72.73 | 75.48 | 74.24 | 75.85 | **84.27** |
| ORL | 68.82 | 67.27 | 73.69 | 75.79 | 77.02 | 82.50 | **85.08** |
| AR | 61.52 | 67.81 | 65.25 | 69.25 | 74.50 | 74.00 | **77.51** |
| JAFFE | 70.91 | 75.41 | 72.29 | 74.87 | 75.59 | 75.12 | **86.64** |
| SVHN | 62.27 | 65.02 | 66.96 | 71.17 | 73.39 | 73.86 | **82.76** |
| ICPR | 71.33 | 75.11 | 73.35 | 76.61 | 81.95 | 80.57 | **84.57** |
| MNIST | 66.41 | 70.28 | 73.71 | 72.94 | 77.24 | 76.14 | **85.33** |
| MSRA-TD500 | 69.71 | 67.69 | 72.89 | 69.62 | 75.82 | 73.79 | **86.29** |
| LIDC-IDRI | 58.89 | 54.36 | 58.67 | 57.55 | 63.84 | 66.41 | **74.22** |
| DDSM | 58.24 | 56.85 | 59.62 | 63.75 | 65.26 | 67.15 | **73.67** |
| Caridac MRI | 55.14 | 58.17 | 60.87 | 59.34 | 62.28 | 69.94 | **71.58** |
| MIAS | 62.71 | 63.37 | 61.27 | 67.78 | 66.43 | 68.82 | **71.17** |

最后,从图 5 中可以看出,随着参数 $\lambda_1$ 和 $\lambda_2$ 取值的变化,KSCSN 算法在各个数据集上的准确度并未发生明显变化。由此可见,KSCSN 的性能受参数 $\lambda_1$ 和 $\lambda_2$ 的影响较小。因此,本文提出的 KSCSN 对参数具有鲁棒性。




(a) YaleB (b) SVHN (c) LIDC-IDRI

图 5 在 YaleB,SVHN 和 LIDC-IDRI 数据集上参数 $\lambda_1$ 和 $\lambda_2$ 对 KSCSN 性能的影响

Fig. 5 Performance evaluation of KSCSN versus parameters $\lambda_1$ and $\lambda_2$ on YaleB, SVHN and LIDC

**结束语** 本文提出了一种基于二阶近邻的核子空间聚类算法,即用于聚类的 KSCSN。KSCSN 的重点在于将核方法与二阶近邻相结合进行子空间聚类,可以有效处理子空间的非线性数据,同时增强了类间不相关表示和类内相关表示。KSCSN 与传统的子空间聚类方法不同,其可以同时优化子空间的稀疏性和连通性,从而解决重叠的子空间问题,有效提高

子空间聚类的准确度。最后针对3种不同的识别任务，分别在人脸、场景字符和生物医学数据集上测试了KSCSN。从实验中可以看出，KSCSN比不使用核方法或二阶近邻的子空间聚类算法更加准确。

未来工作包括：

(1)使用模型选择方法进一步调整本文中的参数，以获得更好的聚类结果。

(2)将核参数的自适应选择与本文相结合，即核函数的带宽自动学习。

(3)学习Grassmannian流形，并将其与子空间聚类相结合以提高聚类性能。

## 参 考 文 献


[1] LIU Q S, LU H Q, MA S D. A review of subspace methods in face recognition Method[J]. Acta Automatica Sinica, 2003(6): 900-911.

[2] DERKSEN H, YI M, WEI H. Segmentation of multivariate mixed data via lossy coding and compression[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(9): 1546-1562.

[3] ZHANG T, SZLAM A, LERMAN G. Median K-flats for hybrid linear modeling with many outliers [C]// IEEE 12th International Conference on Computer Vision Workshops. 2009: 234-241.

[4] HECKEL R, BOLCSKEI H. Robust Subspace Clustering via Thresholding [J]. IEEE Transactions on Information Theory, 2015, 61(11): 6320-6342.

[5] CAO W, ZHAO Y K, GAO S W. Multi-class support vector machine based on fuzzy kernel clustering[C]// Proceedings of 2009 China Process Systems Engineering Annual Conference and China Mes Annual Conference. 2009: 207-210.

[6] PENG X, YU Z, TANG H. Constructing the L2-Graph for Robust Subspace Learning and Subspace Clustering [J]. IEEE Transactions on Cybernetics, 2012, 47(4): 1053.

[7] DING Z C, GE H W, ZHOU J. Density peak clustering algorithm based on KL divergence[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2019, 31(3): 367-374.

[8] ZHOU H, REN H P. Intuitionistic fuzzy similarity clustering algorithm based on chi-square distance[J]. Journal of Chongqing University of Technology(Natural Science), 2020, 34(8): 238-246.

[9] AGRAWAL R, GEHRKE J, GUNOPULOS D. Automatic subspace clustering of high dimensional data for data mining applications [J]. ACM SIGMOD Record, 1998, 27(2): 94-105.

[10] LU C, FENG J, YAN S. A Unified Alternating Direction Method of Multipliers by Majorization Minimization [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(3): 527-541.

[11] VIDAL R, FAVARO P. Low rank subspace clustering(LRSC) [J]. Pattern Recognition Letters, 2014, 43(1): 47-61.

[12] 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.

[13] ZHANG X, CHEN X, SUN X. Schatten-q regularizer constrained low rank subspace clustering model [J]. Neurocomputing, 2016, 182(C): 36-47.

[14] YOU C, LI C G, ROBINSON D P. Oracle Based Active Set Algorithm for Scalable Elastic Net Subspace Clustering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. 3928-3937.

[15] LU C Y, MIN H, ZHAO Z Q. Robust and Efficient Subspace Segmentation via Least Squares Regression [C]// Proceedings of the 12th European Conference on Computer Vision-Volume Part VII. 2012: 347-360.

[16] CHONG Y, LI C G, ROBINSON D P. Oracle Based Active Set Algorithm for Scalable Elastic Net Subspace Clustering [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2016: 3928-3937.

[17] XU K J, RUAN J. Reweighted sparse subspace clustering [J]. Computer Vision & Image Understanding, 2015, 138: 25-37.

[18] ZHANG P T, CHEN X Y. Elastic nuclear subspace clustering [J]. Pattern Recognition and Artificial Intelligence, 2017, 30(9): 779-790.

[19] ZHANG Z, XU Y, SHAO L. Discriminative Block-Diagonal Representation Learning for Image Recognition [J]. IEEE Transactions on Neural Networks & Learning Systems, 2017, 29(7): 3111-3125.

[20] ABAVISANI M, PATEL V M. Multimodal sparse and low-rank subspace clustering [J]. Information Fusion, 2015, 39: 168-177.

[21] XIAO S, TAN M, XU D. Robust Kernel Low-Rank Representation [J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 27(11): 2268-2281.

[22] CHEN Y, JALALI A, SANGHAVI S. Clustering partially observed graphs via convex optimization [J]. Journal of Machine Learning Research, 2014, 15(1): 2213-2238.

[23] YANG J, LIANG J, WANG K. Subspace clustering via good neighbors [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(6): 1537-1544.

[24] CANDES E J, LI X, MA Y. Robust principal component analysis [J]. Journal of the ACM, 2011, 58(3): 11-48.

[25] LIU G, YAN S. Latent Low-Rank Representation for subspace segmentation and feature extraction [C]// International Conference on Computer Vision. 2011: 1615-1622.

[26] XIE X, GUO X, LIU G. Implicit Block Diagonal Low-Rank Representation [J]. IEEE Transactions on Image Processing, 2017, 27(1): 477-489.


**WANG Zhong-yuan**, born in 1996, postgraduate. His main research interests include kernel approximation of low rank block matrix and so on.

**LIU Jing-lei**, born in 1970, Ph.D, professor, master supervisor. His main research interests include artificial intelligent and theoretical computer science.