

中图分类号: TP181

单位代号: 10280

密 级: 公开

学 号: 22721948

上海大学



硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题 目	基于稀疏优化的无监督 特征选择方法研究
-----	------------------------

作 者 杨安宁

学科专业 控制科学与工程

导 师 修贤超 副教授

完成日期 二〇二五年四月

姓 名：杨安宁

学号：22721948

论文题目：基于稀疏优化的无监督特征选择方法研究

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主席：

委员：

导师：

答辩日期： 年 月 日

姓 名：杨安宁

学号：22721948

论文题目：基于稀疏优化的无监督特征选择方法研究

上海大学学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师指导下，独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他研究者对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：

日期： 年 月 日

上海大学学位论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

(保密的论文在解密后应遵守此规定)

学位论文作者签名：

导师签名：

日期： 年 月 日

日期： 年 月 日

上海大学工学硕士学位论文

基于稀疏优化的无监督 特征选择方法研究

作 者: 杨安宁

导 师: 修贤超 副教授

学科专业: 控制科学与工程

机电工程与自动化学院

上海大学

2025 年 4 月

A Dissertation Submitted to Shanghai University for the
Degree of Master in Engineering

Research on Unsupervised Feature Selection Methods via Sparse Optimization

Candidate: Anning Yang

Supervisor: Associate Prof. Xianchao Xiu

Major: Control Science and Engineering

**School of Mechatronic Engineering and Automation
Shanghai University
April, 2025**

摘要

在大数据时代，如何挖掘隐藏在高维复杂数据中的有效信息是数据分析的关键。无监督特征选择作为一种无需标签指导的降维技术，因其能有效剔除冗余特征并保留数据本质结构而备受关注。尽管现有基于稀疏优化的方法已取得显著进展，但在处理高维复杂数据结构时仍面临诸多挑战，如稀疏结构表示不充分、局部特征辨别不准确和重构误差度量不合理。针对上述挑战，本文从单稀疏约束、双稀疏约束和双稀疏约束对比学习等三个方面进行深入研究。主要工作如下：

- (1) 针对稀疏结构表示不充分的问题，探讨了现有 $\ell_{2,0}$ 范数约束和正则的单稀疏约束模型，设计了基于信赖域算法和硬阈值的交替最小化策略。实验表明了 $\ell_{2,0}$ 范数约束的有效性和灵活性。具体地，在所有对比方法中， $\ell_{2,0}$ 范数约束模型的平均聚类准确率和归一化互信息分别提高了至少 2.83% 和 5.09%，验证了 $\ell_{2,0}$ 范数约束对稀疏结构表示的优越性。
- (2) 针对局部特征辨别不准确的问题，提出了 $\ell_{2,0}$ 范数和 ℓ_0 范数的双稀疏约束模型，称为 DSCOFS。通过引入 ℓ_0 范数约束过滤掉不规则的噪声特征，该模型弥补了 $\ell_{2,0}$ 范数约束单一性的限制，从而提高了特征辨别能力。算法方面，设计了基于一阶精确罚函数和硬阈值的近端交替最小化策略，并在理论上证明了算法的收敛性。实验表明，DSCOFS 的平均聚类准确率和归一化互信息分别提高了至少 3.34% 和 3.02%，验证了双稀疏约束对局部特征辨别有效性。
- (3) 针对重构误差度量不合理的问题，构建了融合对比学习和 DSCOFS 的增强型模型，称为 DSCOFS-CL。通过自表示框架重构样本数据，引入对比学习损失建模重构误差，结合低秩约束保持全局结构，从而在原始空间与投影空间自适应地学习样本之间的关系。算法方面，设计了基于梯度下降和硬阈值的近端交替最小化策略，并在理论上证明了算法的收敛性。实验表明，DSCOFS-CL 相较于 DSCOFS，平均聚类准确率和归一化互信息分别提高了 1.85% 和 1.06%，验证了对比学习损失对重构误差度量的实用性。

关键词：无监督特征选择；稀疏优化；双稀疏；对比学习；近端交替最小化

ABSTRACT

In the era of big data, how to mine the effective information hidden in high-dimensional and complex data is the key to data analysis. Unsupervised feature selection, as a dimensionality reduction technique that does not require label guidance, has attracted much attention due to its ability to effectively remove redundant features while preserving the essential structure of the data. Although existing methods based on sparse optimization have made significant progress, there are still many challenges when dealing with high-dimensional complex data structures, such as insufficient sparse structure representation, inaccurate local feature discrimination, and unreasonable reconstruction error measurement. To address these challenges, this paper conducts in-depth research from three aspects: single sparsity constraint, double sparsity constraint, and double sparsity constraint with contrastive learning. The main contributions are as follows:

(1) To address the issue of insufficient sparse structure representation, existing $\ell_{2,0}$ -norm constrained and regularized single sparsity constraint models are investigated, and an alternating minimization strategy based on the trust region algorithm and hard thresholding is designed. The experiments demonstrate the effectiveness and flexibility of the $\ell_{2,0}$ -norm constraint. Specifically, among all the comparison methods, the average clustering accuracy and normalized mutual information of the $\ell_{2,0}$ -norm constrained model are improved by at least 2.83% and 5.09%, respectively, validating the superiority of the $\ell_{2,0}$ -norm constraint for sparse structure representation.

(2) To address the issue of inaccurate local feature discrimination, a double sparsity constrained model combining $\ell_{2,0}$ -norm and ℓ_0 -norm, called DSCOFS, is proposed. By introducing the ℓ_0 -norm constraint to filter out irregular noise features, the model compensates for the limitation of the single sparsity of the $\ell_{2,0}$ -norm constraint, thereby enhancing the ability to discriminate features. In algorithms, a proximal alternating minimization strategy based on the first-order exact penalty function and hard thresholding is designed, and the convergence of the algorithm is theoretically proven. Experimental results show that the

average clustering accuracy and normalized mutual information of DSCOFS are improved by at least 3.34% and 3.02%, respectively, validating the effectiveness of the double sparsity constraint in local feature discrimination.

(3) To address the issue of unreasonable reconstruction error measurement, an enhanced model integrating contrastive learning and DSCOFS, called DSCOFS-CL, is constructed. By reconstructing sample data within the self-representation framework, contrastive learning loss is introduced to model the reconstruction error, while a low-rank constraint is applied to maintain the global structure. This enables the adaptive learning of relationships between samples in both the original and projected spaces. In algorithms, a proximal alternating minimization strategy based on the gradient descent method and hard thresholding is designed, and the convergence of the algorithm is theoretically proven. Experimental results show that compared to DSCOFS, the average clustering accuracy and normalized mutual information of DSCOFS-CL are improved by 1.85% and 1.06%, respectively, validating the practicality of contrastive learning in reconstruction error measurement.

Keywords: Unsupervised feature selection; sparse optimization; double sparsity; contrastive learning; proximal alternating minimization

主要符号对照表

$X \in \mathbb{R}^{d \times m}$	d 行 m 列的实数矩阵 X
$\mathbf{x}^i \in \mathbb{R}^m$	矩阵 X 的第 i 行的 m 维向量
$\mathbf{x}_i \in \mathbb{R}^d$	矩阵 X 的第 i 列的 d 维向量
X_{ij}	矩阵 X 的第 i 行 j 列的元素
$\ X\ _1$	矩阵 X 的 ℓ_1 范数, 即 $\ X\ _1 = \sum_{i=1}^d \sum_{j=1}^m X_{ij}$
$\ X\ _{\text{F}}$	矩阵 X 的 Frobenius 范数, 即 $\ X\ _{\text{F}} = \sqrt{\sum_{i=1}^d \sum_{j=1}^m X_{ij}^2}$
$\ X\ _0$	矩阵 X 的 ℓ_0 范数, 即 X 中非零元素的个数
$\ X\ _{2,0}$	矩阵 X 的 $\ell_{2,0}$ 范数, 即 X 中非零行的个数
$\ X\ _{2,1}$	矩阵 X 的 $\ell_{2,1}$ 范数, 即 $\ X\ _{2,1} = \sum_{i=1}^d \sqrt{\sum_{j=1}^m X_{ij}^2}$
$\ X\ _{2,p}$	矩阵 X 的 $\ell_{2,p}$ 范数, 即 $\ X\ _{2,p} = \sqrt[p]{\sum_{i=1}^d (\sqrt{\sum_{j=1}^m X_{ij}^2})^p}$
$\ X\ _{1,2}$	矩阵 X 的 $\ell_{1,2}$ 范数, 即 $\ X\ _{1,2} = \sqrt{\sum_{j=1}^m (\sum_{i=1}^d X_{ij})^2}$
$\ \mathbf{x}^i\ _0$	向量 \mathbf{x}^i 的 ℓ_0 范数, 即 \mathbf{x}^i 中非零元素的个数
$\ \mathbf{x}^i\ _1$	向量 \mathbf{x}^i 的 ℓ_1 范数, 即 $\ \mathbf{x}^i\ _1 = \sum_{j=1}^m X_{ij} $
$\ \mathbf{x}^i\ _2$	向量 \mathbf{x}^i 的 ℓ_2 范数, 即 $\ \mathbf{x}^i\ _2 = \sqrt{\sum_{j=1}^m X_{ij}^2}$
$\ \mathbf{x}^i\ _p$	向量 \mathbf{x}^i 的 ℓ_p 范数, 即 $\ \mathbf{x}^i\ _p = \sqrt[p]{\sum_{j=1}^m X_{ij} ^p}$

目 录

摘 要	I
ABSTRACT	II
主要符号对照表	IV
第一章 绪论	1
1.1 研究背景及意义.....	1
1.2 国内外研究现状.....	2
1.2.1 无监督特征选择.....	2
1.2.2 稀疏优化	5
1.3 本文主要研究内容和结构组织	7
第二章 基于单稀疏约束的无监督特征选择方法	9
2.1 相关工作	9
2.2 数学模型与算法.....	10
2.2.1 数学模型	10
2.2.2 优化算法	12
2.3 数值实验与分析.....	16
2.3.1 实验设置	17
2.3.2 实验结果	19
2.3.3 稀疏度分析	24
2.3.4 讨论	25
2.4 本章小结	28
第三章 基于双稀疏约束的无监督特征选择方法	29
3.1 相关工作	29
3.2 数学模型与算法.....	30
3.2.1 数学模型	30
3.2.2 优化算法	31

3.2.3 理论分析	36
3.3 数值实验与分析	40
3.3.1 实验设置	40
3.3.2 实验结果	41
3.3.3 消融实验	48
3.3.4 讨论	50
3.4 本章小结	55
第四章 基于对比学习和双稀疏约束的无监督特征选择方法	56
4.1 相关工作	56
4.2 数学模型与算法	59
4.2.1 数学模型	59
4.2.2 优化算法	60
4.2.3 理论分析	63
4.3 数值实验与分析	67
4.3.1 实验设置	67
4.3.2 实验结果	68
4.3.3 消融实验	73
4.3.4 讨论	75
4.4 本章小结	80
第五章 总结与展望	82
5.1 总结	82
5.2 展望	83
插图索引	84
表格索引	86
参考文献	87
攻读硕士学位期间取得的研究成果	97
致 谢	98
附录 A 本文英文缩写对照表	99

第一章 绪论

1.1 研究背景及意义

随着信息化技术的高速发展，人们所面临的数据出现了井喷式增长，并且这些数据大多呈现高维特性。根据国际数据公司的统计和预测，2025年数据总量将突破175ZB，并在2028年将增长至393.8ZB，相比于2025年增长了1.25倍。然而，海量高维数据中普遍存在稀疏性^[1]。以典型文本数据为例，大多数词在每篇文档中并没有出现或仅出现一两次，因此统计所有词出现频率的矩阵大部分位置为零。合理利用数据的稀疏结构不仅能对高维数据进行充分地压缩，从而节约储存空间、减少传输量，还可以从海量的数据中挖掘出有价值的信息，让复杂问题得以简化^[2]。因此，对于需要大量数据分析的模式识别^[3]、机器学习^[4]、数据挖掘^[5]、生物工程^[6]以及物联网^[7]等研究领域，如何基于稀疏性特征挖掘隐藏在数据中的有效信息是大数据分析的关键。

数据降维是一个基本而重要的预处理方法，根据对特征处理方式的不同可以分为特征提取^[8]和特征选择^[9]。特征提取通过投影矩阵将原始特征映射到低维特征空间，生成的新特征是原始特征经过线性或非线性组合而成，这意味着特征是从原始特征中“提取”出来的。与特征提取不同，特征选择不改变原始的特征空间，而是通过对原始的数据按照一定的评价指标选择包含更多信息的特征子空间，这意味着特征是从原始特征中“选择”出来的。特征选择能够通过选择出包含最多有效信息的特征子空间来消除冗余信息和噪声特征，从而提高模型的效率和性能^[10]。此外，特征选择的数据来源于原始特征，这使得可解释性更强^[11]，更适用于大数据分析。毫无疑问，特征选择已经成为人工智能中的核心技术之一。

然而，现实世界的数据往往呈现部分未标注或标注不完整的情形^[12]。特别是在医学成像分析领域，标签的获取尤为困难。针对这种情形，无监督特征选择应运而生，并因其能够在不依赖标签的情况下选择出最具代表性的特征而备受关注。目前，无监督特征选择已经成功应用于诸多工程领域，包括图像分类^[13]、遥感成像^[14]、无线通信^[15]以及基因表达^[16]。如图1.1所示，Web of Science中关于无监督特征选择主题的文章近十年（2014-2024）增长将300%。

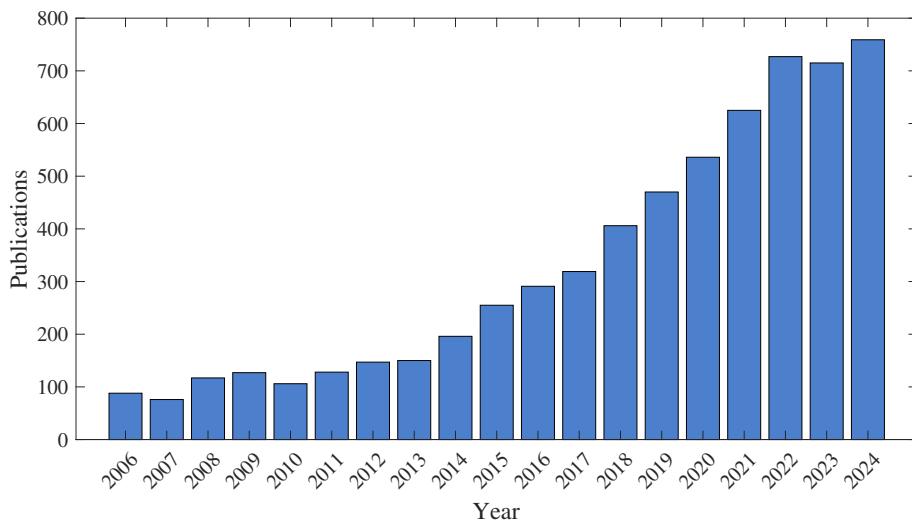


图 1.1 关于无监督特征选择主题的文章统计

Figure 1.1 The publication statistics on the topic of unsupervised feature selection

综上所述，无监督特征选择是模式识别领域的研究热点，也是人工智能领域的前沿课题。因此，本文以此为出发点，旨在探讨稀疏优化框架下无监督特征选择的新方法，具有重要的科学意义。

1.2 国内外研究现状

1.2.1 无监督特征选择

一般来说，无监督特征选择方法可以根据模型训练和特征选择的结合方式分为三种类型：过滤式^[17]、包裹式^[18]和嵌入式^[19]。过滤式方法将特征选择和模型训练分开，首先根据统计指标或规定的评价指标从数据集中选择出最相关的特征，然后使用选择出的特征来训练模型。He 等人^[20]提出的拉普拉斯评分 (Laplace Score, LapScore) 是典型的过滤式无监督特征选择方法。LapScore 首先通过维持数据局部结构的能力来评估特征的重要性，随后对特征按照重要性排序，最后根据需求选取前列的特征用于后续的模型训练。然而，过滤式无监督特征选择方法由于脱离模型性能进行选择，可能会选出对特定模型而言不理想的特征。相反地，包裹式方法通过使用特定的模型来评估特征子集的优劣。它把特征选择“包裹”在一个机器学习模型中，选择出最契合模型性能的特征。拉斯维加斯包裹式方法 (Las Vegas Wrapper, LVW^[21]) 是结合了拉斯维加斯算法的典型包裹式无监督特征选择方法。LVW 首先随机产生多个特征子集，随后将产生的随机子集放入模型中进行训练并评估其误差，最后保留

误差最小的特征子集，循环此过程直至误差足够小。显然，包裹式无监督特征选择方法每次特征选择都需要训练一个新的模型，因此计算成本较高。嵌入式方法结合了过滤式和包裹式方法的特点，将特征选择和模型训练过程构建成一个整体，允许特征选择在模型训练的过程中自动进行。与包裹式方法的“包裹”不同，嵌入方法在训练模型的过程中对特征进行一定的处理（如 ℓ_1 正则化），从而实现特征的自动选择。图 1.2 展示了三种类型的无监督特征选择方法框架。事实上，这些方法各有优缺点，实际选择时应根据数据集的特性、计算资源的限制以及任务的需求来决定。

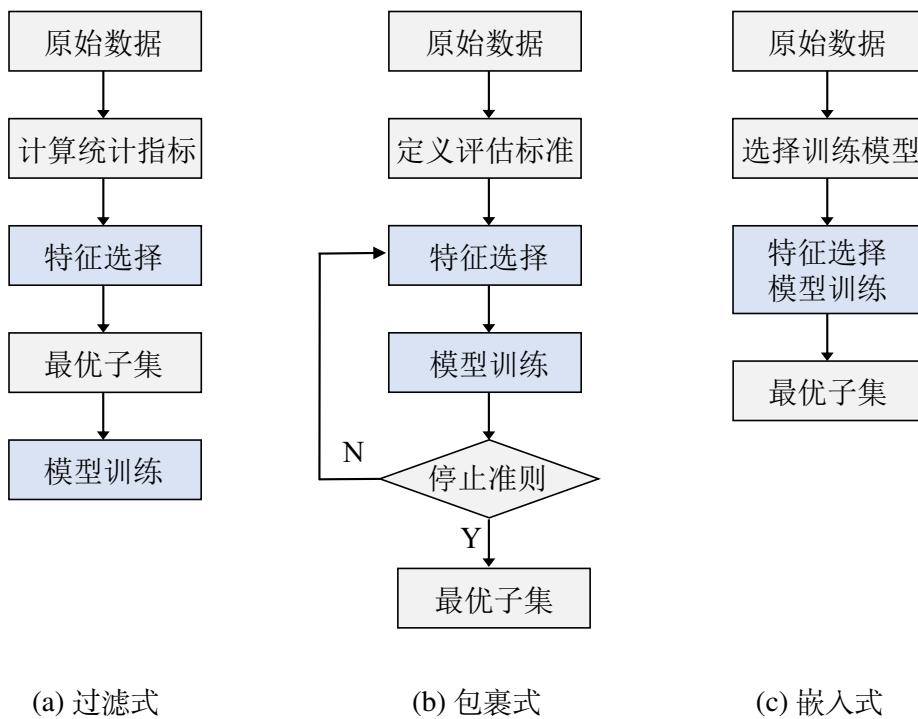


图 1.2 无监督特征选择方法框架

Figure 1.2 The framework of unsupervised feature selection methods

近年来，嵌入式无监督特征选择方法在机器学习领域取得了显著突破。而大多数嵌入式方法，利用谱分析和流形学习来选择具有辨别力的特征^[22]。基于谱分析的方法通常构造邻接矩阵或者图亲和矩阵描述数据的局部几何结构，然后利用稀疏正则化来探索数据聚类结构，最后通过稀疏优化模型实现特征选择。Cai 等人^[23] 将谱分析与 ℓ_1 范数结合，提出了用于多聚类结构的特征选择框架。为了提高鲁棒性，Yang 等人^[24] 考虑到局部判别信息比全局判别信息更重要，将构造的局部判别分析得分矩阵和 $\ell_{2,1}$ 范数相结合，提出了无监督判别特征选择（Unsupervised Discriminative Feature Selection, UDFS）。随后，Li 等人^[25] 将 UDFS 扩展到具有非负谱分析的情况，

取得了良好的数值结果。Liu 等人^[26]通过局部线性嵌入算法获得特征权重矩阵，并使用 ℓ_1 范数描述损失函数，提出了鲁棒邻域嵌入（Robust Neighborhood Embedding, RNE）。然而，上述基于谱分析的方法将图的构造和特征选择的过程分离，这使得它们对冗余特征和噪声特征十分敏感。为了解决此问题，Nie 等人^[27]在低维空间中学习自适应的图并嵌入 $\ell_{2,p}$ ($0 < p \leq 1$) 范数来表征稀疏性，进而将局部结构的学习融合到特征选择的过程中，最终提出了结构化最优图特征选择（Structured Optimal Graph Feature Selection, SOGFS）。Li 等人^[28]基于最大熵原理构建了自适应图，同时引入了广义无相关约束和 $\ell_{2,1}$ 范数，提出了一种改进的稀疏回归模型。为了保留原始空间数据的局部几何结构，Shi 等人^[29]通过自适应构建动态相似性图来建模数据的内在结构，将二进制哈希码学习为弱监督的多标签，同时使用学习到的标签来指导特征选择。Tang 等人^[30]采用自表示技术自动学习样本的相似性图，进而保留了数据的局部几何结构。Chen 等人^[31]利用 $\ell_{2,1}$ 范数学习到灵活的最优图，并将其与特征选择结合到一起。Zhou 等人^[32]通过在原始空间和低维空间中联合学习自表示矩阵，提出了基于联合自表示图学习的无监督特征选择方法。

最近，基于主成分分析（Principal Component Analysis, PCA）^[33]和稀疏主成分分析（Sparse PCA, SPCA）^[34]的特征选择方法呈现新的发展趋势。例如，Chang 等人^[35]通过引入 $\ell_{2,1}$ 范数来衡量损失和正则化，构建了有效的 SPCA 模型。Yi 等人^[36]提出了自适应加权 SPCA，能够从噪声数据中稳健地选择出重要的特征。Zheng 等人^[37]使用重构矩阵将 SPCA 描述为一个凸优化问题，并加入了 $\ell_{2,1}$ 范数和核范数使其稀疏且低秩，同时证明了该重构矩阵的最优解落在半正定锥上，最终通过半正定投影实现稀疏主成分分析（SPCA via Positive Semidefinite projection, SPCA-PSD）。Li 等人^[22]引入了非凸的 $\ell_{2,p}$ ($0 < p \leq 1$) 范数，提出了用于特征选择的稀疏主成分分析（SPCA for Feature Selection, SPCAFS），同时也验证了 $\ell_{2,p}$ ($0 < p \leq 1$) 范数相比 $\ell_{2,1}$ 范数的优势。Zhou 等人^[38]引入了对比学习^[39]来度量 PCA 的重构误差，同时对投影矩阵和自表示矩阵稀施加了 $\ell_{1,2}$ 范数约束，最终实现了更好的无监督特征选择性能。

值得注意的是，上述无监督特征选择方法仅考虑了 $\ell_{2,0}$ 范数的凸松弛 ($\ell_{2,1}$ 范数) 或非凸松弛 ($\ell_{2,p}$ 范数形式)。正如文献^[40-41] 中所讨论的，这可能会导致稀疏表示不足，我们将会在下一小节中详细介绍。最近，Chen 等人^[42]在子空间中嵌入了自适应二分图，并对投影矩阵应用 $\ell_{2,0}$ 范数约束进行特征选择。Nie 等人^[43]基于数据协方差矩阵秩探究了 $\ell_{2,0}$ 范数约束的特征稀疏约束主成分分析（Feature-Sparsity constrained

PCA, FSPCA), 该方法把特征选择和 PCA 相结合, 进而通过行稀疏选择出相应的特征。Zhang 等人^[44]在样本和锚点之间构建了二分图, 利用 $\ell_{2,0}$ 范数引导的相似关系更新图结构, 从而提高了所选特征的质量, 同时降低了求解过程中的计算成本。

1.2.2 稀疏优化

2006 年, Donoho^[45]提出了压缩感知 (Compressed Sensing, CS) 的概念, 即如果一个信号是稀疏的或可压缩的, 那么它可以通过少量的测量值来重构原始信号, 而这些测量值远少于以前的理论 (如香农采样定理) 所建议的数量。压缩感知理论为高维数据的处理带来了新的思路, 极大推动了信息技术工作的发展, 也为后续稀疏优化打下了基础。一般地, 稀疏优化是指带有 ℓ_0 范数正则或约束的一类非凸非连续优化问题, 也是非确定性多项式 (Nondeterministic Polynomial, NP) 难的。在理论方面, 人们借助变分分析工具研究了稀疏优化的变分性质, 如次微分、邻近点算子、切锥与法锥等^[46]。同时, 在算法方面, 已经有一些成熟的求解算法框架, 例如交替最小化算法 (Alternating Minimization Algorithm, AMA), 交替方向乘子法 (Alternating Direction Method of Multipliers, ADMM)^[47] 和近端交替最小化 (Proximal Alternating Minimization, PAM)^[48] 等。此外, 也发展了一些速度快且精度高的二阶算法^[49], 例如牛顿硬阈值追踪法^[50]、拉格朗日对偶法^[51]、半光滑牛顿法^[52]等, 并且在许多工程问题中验证了其高效性, 包括信号处理^[53]、图像处理^[54]、数据降维^[55]、故障诊断^[56]等。图 1.3 展示了二维向量下 ℓ_p 范数值不大于 1 的约束区域。从图中可以直观地看出, 随着 p 的减小, ℓ_p 范数约束区域越小, 这也意味着向量值越小。因此, 只有当 $p = 0$ 时, 二维向量才能得到 0 解, 而其他都是接近 0 的松弛解, 这也解释了为什

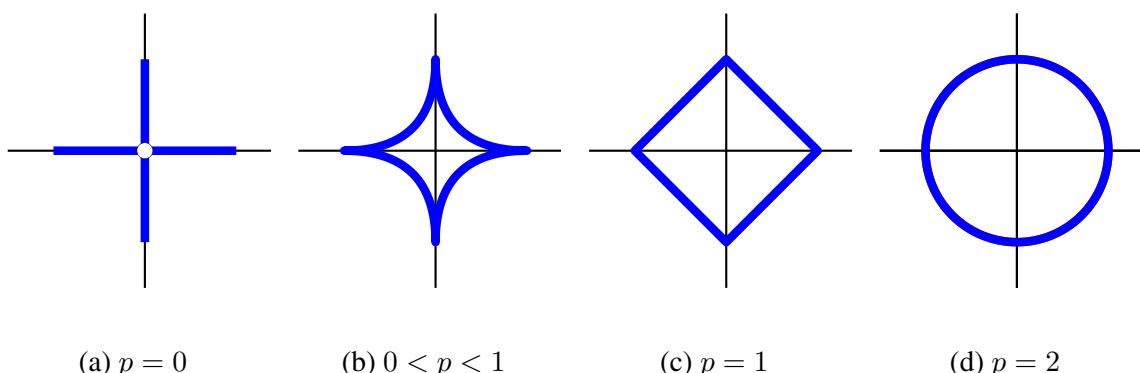


图 1.3 不同 p 取值下的 ℓ_p 范数约束区域

Figure 1.3 The constraint regions of the ℓ_p norm under different values of p

么 ℓ_0 范数可以刻画更加充分的稀疏结构。此外，当 $p < 1$ 时， ℓ_p 范数约束区域呈现非凸性质，而 ℓ_1 范数是 ℓ_0 范数最优的凸松弛。

本质上说，基于 SPCA 的无监督特征选择问题是一个带有稀疏性特征的 Stiefel 流形（又称正交流形）优化问题。对于这类特殊的非凸优化问题，人们通常采用松弛方法近似求解，即寻找不同形式的凸松弛^[57-60]或非凸松弛^[22,61-62]，然后通过松弛问题进行近似求解。例如 Chen 等人^[63]探讨了带 ℓ_1 范数的 Stiefel 流形优化问题，通过引入新的变量将非光滑项分开，然后基于增广拉格朗日函数开发了 PAM 优化策略。Chen 等人^[58]针对偏最小二乘回归问题，构造了带 $\ell_{2,1}$ 范数的 Stiefel 流形和 Grassmann 流形优化形式，进而利用求解器进行计算。为了提升在 Stiefel 流形上计算效率，Chen 等人^[64]分析了带 $\ell_{2,1}$ 范数的 Stiefel 流形优化问题，并在 Stiefel 流形切线空间上采用近端梯度技术进行计算。而当子问题没有显示表达式时，还设计了半光滑牛顿算法近似计算。同样对于带 $\ell_{2,1}$ 范数的 Stiefel 流形优化问题，Xiao 等人^[60]利用拉格朗日乘子的显示表达式建立了精确罚理论，设计了非精确的近端梯度方法。由于子问题都可以显示求解，算法的效率得到了保障。进一步地，Zhou 等人^[65]充分考虑了 $\ell_{2,1}$ 范数和 Stiefel 流形的二阶信息，基于半光滑牛顿技巧提出了更快速的增广拉格朗日算法，每一步迭代都自动满足流形约束，同时严格证明了全局收敛性和局部超线性收敛率。虽然 ℓ_1 范数和 $\ell_{2,1}$ 范数等凸松弛可以快速求解，但他们无法充分地表示稀疏结构。于是，人们尝试使用非凸松弛 ℓ_p ($0 < p < 1$) 范数和 $\ell_{2,p}$ ($0 < p < 1$) 范数。Fung 等人^[66]证明了当 p 趋近于 0 时， ℓ_p 范数松弛问题与原问题（即 ℓ_0 范数）等价。Breloy 等人^[61]细致分析了不同类型的 SPCA 模型，考虑了 ℓ_0 范数的非凸松弛近似，设计了有效的 Majorization-Minimization 算法。Li 等人^[22]针对无监督特征提选择问题，研究了带 $\ell_{2,p}$ ($0 < p \leq 1$) 范数正则的优化模型，最后利用光滑化技巧求解。

经过多年的发展，稀疏优化的研究热点已由最初的凸函数或非凸函数松弛发展到直接处理原问题。相比松弛方法来说， ℓ_0 和 $\ell_{2,0}$ 范数可以对稀疏度进行更精准的刻画，因而更值得研究。最近，Bertsimas 等人^[67]研究了带 ℓ_0 范数约束的 SPCA 问题，将问题转化为混合整数半正定规划形式，并最终通过切平面法求解。Xiu 等人^[68]利用硬阈值算子开发了一种基于 ADMM 的优化算法，并成功应用到故障诊断。Nie 等人^[43]研究了带 $\ell_{2,0}$ 范数约束的 SPCA 问题，根据数据协方差矩阵秩的情况，设计了全局优化算法和迭代代理更新算法，并证明了算法的收敛性。进一步，Nie 等人^[69]将 SPCA 问题转化为一个新的等效问题，提出了一种不需要调整任何参数的高

效坐标下降算法。

综上所述，尽管基于稀疏优化松弛方法的无监督特征选择已形成较为完整的理论体系，但面向原始稀疏约束（即 ℓ_0 范数和 $\ell_{2,0}$ 范数）的研究仍处于持续探索阶段。特别是在处理高维复杂数据结构时，现有方法仍面临诸多挑战，如稀疏结构表示不充分、局部特征辨别不准确和重构误差度量不合理等问题。更值得关注的是，针对这类特殊的稀疏 Stiefel 流形优化问题，设计有效且收敛的算法也是亟待突破的重要科学问题。

1.3 本文主要研究内容和结构组织

本文从单稀疏约束、双稀疏约束和双稀疏约束对比学习三个方面有序展开，逐层深入，对无监督特征选择方法进行了系统的研究，主要内容和章节安排总体框架如图 1.4 所示。具体地，

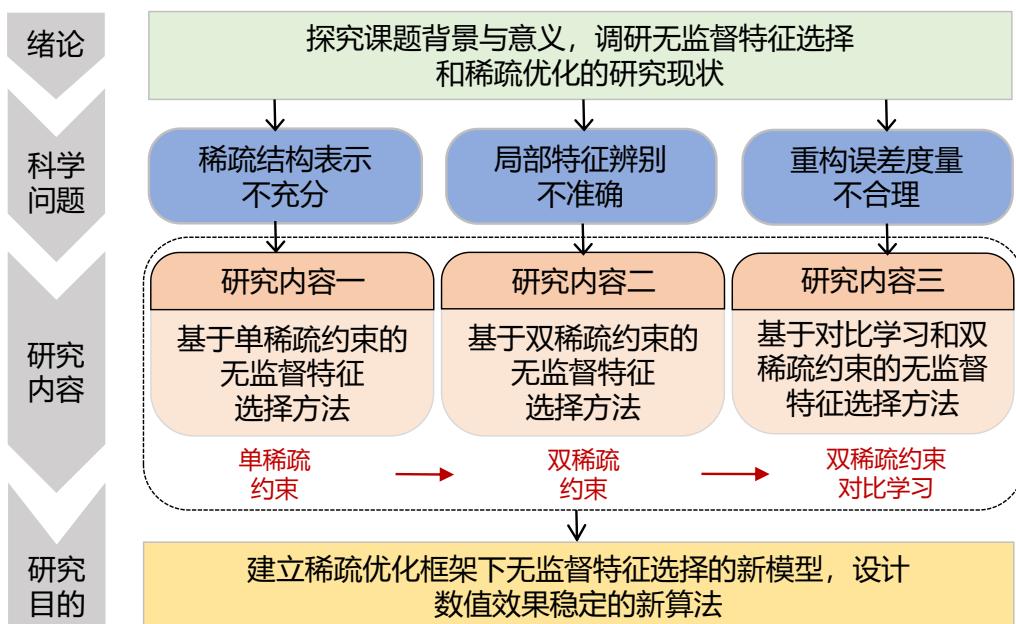


图 1.4 文章主要内容与章节安排

Figure 1.4 The main content of the article and chapter organization

第一章为绪论。阐述了本文的研究背景与意义以及无监督特征选择和稀疏优化的研究现状。同时，给出了文章的研究内容和后续章节的结构组织。

第二章针对稀疏结构表示不充分的问题进行研究。首先探讨了 $\ell_{2,0}$ 范数正则的稀疏主成分分析 (SPCA with $\ell_{2,0}$ -norm Regularization, SPCA-R) 模型和 $\ell_{2,0}$ 范数约束

的稀疏主成分分析 (SPCA with $\ell_{2,0}$ -norm Constraint, SPCA-C) 模型。其次，设计了基于信赖域算法和硬阈值的优化策略。数值实验验证了 $\ell_{2,0}$ 范数可以更加充分地表示稀疏结构，从而提升无监督特征选择的性能。同时，通过稀疏度分析展现了 SPCA-C 比 SPCA-R 在实际特征选择过程中的优势。最后，与 SPCA-C 同模型算法进行了对比，反映了算法设计对无监督特征选择性能的影响。

第三章针对局部特征辨别不准确的问题进行研究。首先在 SPCA-C 的基础上，引入了 ℓ_0 范数用于辨识局部稀疏性，建立了 $\ell_{2,0}$ 范数和 ℓ_0 范数的双稀疏约束优化特征选择 (Double Sparsity Constrained Optimization Feature Selection, DSCOFS) 模型。其次，设计了基于一阶精确罚函数方法和硬阈值的优化策略，严格证明了算法所产生的序列能够收敛到驻点（又称稳定点）。特别地，为了评估双稀疏约束的性能，引入了一种新的特征相似度评价指标。最后，实验验证了 DSCOFS 的有效性、稳健型和收敛性，表明了双稀疏约束相较于单稀疏约束的优势。

第四章针对重构误差度量不合理的问题进行研究。首先在 DSCOFS 的基础上，利用自表示矩阵学习样本间的数据结构，引入了对比学习损失作为重构误差的度量，建立了融合对比学习的双稀疏约束优化特征选择 (DSCOFS with Contrastive Learning, DSCOFS-CL) 模型。其次，设计了基于近端梯度下降算法和硬阈值的优化策略，严格证明了算法所产生的序列能够收敛到驻点（又称稳定点）。最后，通过数值实验表明了 DSCOFS-CL 性能相较于 DSCOFS 有了进一步的提升，验证了对比学习作为重构误差度量的有效性。

第五章为总结与展望。总结了本文的主要研究内容和成果，同时针对本文工作中的不足之处进行了分析，最后在本文工作的基础上对未来的工作进行了展望。

第二章 基于单稀疏约束的无监督特征选择方法

绪论部分梳理了当前主流的无监督特征选择方法，然而这些方法多采用 $\ell_{2,0}$ 范数的松弛形式去表示数据的稀疏结构，这可能会导致对稀疏结构的表示不够充分。因此，本章基于稀疏主成分分析，聚焦于直接使用 $\ell_{2,0}$ 范数的无监督特征选择方法，重点探讨了 $\ell_{2,0}$ 范数正则和约束的两类数学模型。在算法方面，设计了基于信赖域算法和硬阈值的交替优化策略。实验结果表明，与松弛形式范数相比， $\ell_{2,0}$ 范数对数据稀疏结构的表示更加充分，能够有效提升无监督特征选择的性能。

2.1 相关工作

作为经典的统计分析方法，PCA 的核心思想是通过正交的线性变换矩阵将原始高维数据投影到新的低维空间，使投影后的数据方差最大，从而在较低的维度下尽可能地保留数据的主要信息。若将数据降维至 m 维，可以记投影矩阵为 $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m] \in \mathbb{R}^{d \times m}$ ，其中 $\mathbf{x}_i \in \mathbb{R}^d$ 是投影矩阵 X 的第 i 个列向量，也表示第 i 个投影方向。同时，每个投影方向还应满足 $\mathbf{x}_i^\top \mathbf{x}_i = 1$ 和 $\mathbf{x}_i^\top \mathbf{x}_j = 0$ ($i \neq j$)，其中 \top 表示转置。对于中心化处理的数据 $A = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n] \in \mathbb{R}^{d \times n}$ ，应满足 $\sum_{i=1}^n \mathbf{a}_i = 0$ ，即每个特征的均值为零，其中 $\mathbf{a}_i \in \mathbb{R}^d$ 是数据 A 的第 i 个样本。PCA 在方向 \mathbf{x}_i 上的投影结果为 $A^\top \mathbf{x}_i$ ，投影的方差为

$$\text{Var}(A^\top \mathbf{x}_i) = \frac{1}{n} (A^\top \mathbf{x}_i)^\top (A^\top \mathbf{x}_i) = \frac{1}{n} \mathbf{x}_i^\top A A^\top \mathbf{x}_i, \quad (2.1)$$

在所有方向上的方差和为

$$\text{Var}(A^\top X) = \sum_{i=1}^n \frac{1}{n} \mathbf{x}_i^\top A A^\top \mathbf{x}_i = \text{Tr}(X^\top A A^\top X), \quad (2.2)$$

其中 $\text{Tr}(\cdot)$ 表示矩阵的迹。因此，PCA 可以看作是求解投影矩阵 X 最大化方差的问题，其数学模型为

$$\begin{aligned} \min_X \quad & -\text{Tr}(X^\top A A^\top X) \\ \text{s.t.} \quad & X^\top X = I_m. \end{aligned} \quad (2.3)$$

从优化的角度来看，模型(2.3)的求解可转化为对协方差矩阵 $\frac{1}{n}AA^\top$ 进行特征分解，并计算其最大 d 个特征值对应的特征向量。然而，这种方法在特征可解释性方面较差。为突破此限制，一种常见的方法是引入稀疏正则化^[70]。通过调整正则化参数，可以较容易地选择出具有判别性的特征。Zou 等人^[34]提出了下面的 SPCA 模型

$$\begin{aligned} \min_X \quad & -\text{Tr}(X^\top AA^\top X) + \lambda \|X\|_{2,1} \\ \text{s.t.} \quad & X^\top X = I_m, \end{aligned} \tag{2.4}$$

其中 $\lambda > 0$ 是影响稀疏度的正则化参数， $\|X\|_{2,1}$ 是 X 的 $\ell_{2,1}$ 范数，即 $\|X\|_{2,1} = \sum_{i=1}^d \sqrt{\sum_{j=1}^m X_{ij}^2}$ ，这里 X_{ij} 表示矩阵 X 第 i 行 j 列的元素。通过将 $\ell_{2,1}$ 范数加入到模型(2.3)的目标函数中，模型(2.4)可以得到一个稀疏的投影矩阵 X ，从而增强了模型的可解释性，并减轻了噪声的干扰。

相比于凸优化，非凸优化虽然不易得到全局最优解，但却能够提供更多的可能解^[71]。例如，对于非凸函数 ℓ_p ($0 < p < 1$) 范数^[72]，当 $p = 1/2, 2/3, 3/4$ 时，存在显示解且效果优于 ℓ_1 范数。受此启发，Li 等人^[22]引入了 $\ell_{2,0}$ 范数的合理松弛 $\ell_{2,p}$ ($0 < p \leq 1$) 范数，并构建如下SPCAFS模型

$$\begin{aligned} \min_X \quad & -\text{Tr}(X^\top AA^\top X) + \lambda \|X\|_{2,p}^p \\ \text{s.t.} \quad & X^\top X = I_m, \end{aligned} \tag{2.5}$$

其中 $\|X\|_{2,p}^p = \sum_{i=1}^d (\sqrt{\sum_{j=1}^m X_{ij}^2})^p$ 。实验表明，相较于模型(2.4)中的 $\ell_{2,1}$ 范数，非凸松弛的 $\ell_{2,p}$ 范数得到了更稀疏的投影矩阵，进而提高了无监督特征选择的性能。

2.2 数学模型与算法

本节首先给出 $\ell_{2,0}$ 范数正则和约束的无监督特征选择模型，其次设计基于信赖域算法和硬阈值的优化算法。

2.2.1 数学模型

根据前文相关模型的介绍，基于 $\ell_{2,0}$ 范数正则的 SPCA 模型可以描述为

$$\begin{aligned} \min_X \quad & -\text{Tr}(X^\top AA^\top X) + \lambda \|X\|_{2,0} \\ \text{s.t.} \quad & X^\top X = I_m, \end{aligned} \tag{2.6}$$

其中 $\|X\|_{2,0} = \sum_{i=1}^d \mathbb{I}(\|\mathbf{x}^i\|_2)$ 表示 X 的 $\ell_{2,0}$ 范数。这里， \mathbf{x}^i 是投影矩阵 X 的第 i 个行向量。 $\mathbb{I}(\|\mathbf{x}^i\|_2)$ 用于判断向量 \mathbf{x}^i 的 ℓ_2 范数是否为零，即

$$\mathbb{I}(\|\mathbf{x}^i\|_2) = \begin{cases} 1, & \|\mathbf{x}^i\|_2 \neq 0, \\ 0, & \|\mathbf{x}^i\|_2 = 0. \end{cases} \quad (2.7)$$

因此，与松弛形式的 $\ell_{2,p}$ ($0 < p < 1$) 范数相比， $\ell_{2,0}$ 范数会得到全零行的稀疏解，进而实现稀疏结构的充分表示。

在无监督特征选择中，矩阵 X^\top 列向量 (X 行向量) 的稀疏结构对应数据矩阵 A 相应行的稀疏结构，而 A 的列表示相应的特征向量，这意味着投影矩阵 X 行向量的稀疏结构可以反映原始数据相应特征向量的稀疏结构。图 2.1 展示了原始数据的投影过程。从图中可以看出， X^\top 的列向量影响原始数据相应特征投影后的数据分量。具体地，当 X^\top 第 i 列向量为全零元素时，第 i 个特征参与投影的数据分量为零，这意味着第 i 个特征在投影过程中是可稀疏的无效信息。

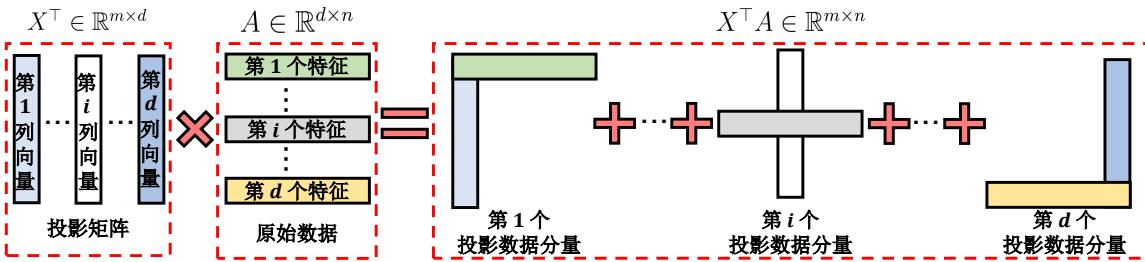


图 2.1 原始数据的投影过程

Figure 2.1 The projection process of the original data

基于上述分析，若要选择出更具辨识性的特征，只需评估投影矩阵 X 相应行的向量。对 X 施加 $\ell_{2,0}$ 范数约束，可构建一个约束形式的无监督特征选择模型

$$\begin{aligned} \min_X \quad & -\text{Tr}(X^\top A A^\top X) \\ \text{s.t.} \quad & X^\top X = I_m, \|X\|_{2,0} \leq s, \end{aligned} \quad (2.8)$$

其中 $s > 0$ 用于控制投影矩阵 X 的稀疏度，即保留非零行的个数，其对应需要选择的特征数量。

在本文中，记 $\ell_{2,0}$ 范数正则的 SPCA 模型 (2.6) 为 SPCA-R， $\ell_{2,0}$ 范数约束的 SPCA 模型 (2.8) 为 SPCA-C。

2.2.2 优化算法

由于模型 (2.6) 和 (2.8) 中包含 $\ell_{2,0}$ 范数，因此它们是非凸、非光滑的优化模型。此外，模型中还存在正交约束 $X^\top X = I_m$ ，这也给模型的求解带来了较大的挑战。

2.2.2.1 求解 SPCA-R

首先引入中间变量 $X = Y$ ，将模型 (2.6) 改写为

$$\begin{aligned} \min_{X,Y} \quad & -\text{Tr}(X^\top A A^\top X) + \lambda \|Y\|_{2,0} \\ \text{s.t.} \quad & X^\top X = I_m, \quad X = Y. \end{aligned} \quad (2.9)$$

利用惩罚函数方法，可以把模型 (2.9) 转化为

$$\begin{aligned} \min_{X,Y} \quad & -\text{Tr}(X^\top A A^\top X) + \lambda \|Y\|_{2,0} + \mu \|X - Y\|_F^2 \\ \text{s.t.} \quad & X^\top X = I_m, \end{aligned} \quad (2.10)$$

其中 $\mu > 0$ 是用于控制约束违反程度的惩罚参数， $\|\cdot\|_F$ 是矩阵的 Frobenius 范数。设 X^k 和 Y^k 是第 k 次迭代的变量，下面将介绍如何通过 AMA 策略进行求解。

(1) 固定 Y ，更新 X ：

$$\begin{aligned} \min_X \quad & -\text{Tr}(X^\top A A^\top X) + \mu \|X - Y^k\|_F^2 \\ \text{s.t.} \quad & X^\top X = I_m. \end{aligned} \quad (2.11)$$

记模型 (2.11) 的目标函数为 $g(X)$ ，则 $g(X)$ 的欧式梯度为

$$\nabla g(X) = -2A A^\top X + 2\mu(X - Y^k), \quad (2.12)$$

欧式 Hessian 矩阵为

$$\nabla^2 g(X) = -2I_m \otimes A A^\top + 2\mu I_{dm}, \quad (2.13)$$

其中 \otimes 表示克罗内克积。记 $\text{St}(d, m) = \{X \in \mathbb{R}^{d \times m} \mid X^\top X = I_m\}$ 是黎曼空间中的一个 Stiefel 流形，则模型 (2.11) 可以改写为 Stiefel 流形优化模型

$$\min_{X \in \text{St}(d,m)} -\text{Tr}(X^\top A A^\top X) + \mu \|X - Y^k\|_F^2. \quad (2.14)$$

众所周知，信赖域算法^[73]是处理带有复杂约束非线性优化模型的强大工具，不仅简单易用，而且收敛速度较快。下面将使用信赖域算法求解优化模型(2.14)。

首先，Stiefel流形在点 X 处的切空间为

$$T_X \text{St}(d, m) = \{\eta \in \mathbb{R}^{d \times m} \mid X^\top \eta + \eta^\top X = 0\}. \quad (2.15)$$

随后，黎曼梯度可以通过将其欧式梯度投影到黎曼流形的切空间上来获得，即

$$\begin{aligned} \text{grad } g(X) &= \mathcal{P}_X(\nabla g(X)) \\ &= \nabla g(X) - X \text{sym}(X^\top \nabla g(X)), \end{aligned} \quad (2.16)$$

其中 $\mathcal{P}_X(\nabla g(X))$ 是将欧式梯度 $\nabla g(X)$ 投影到黎曼流形切空间上的结果，而 $\text{sym}(X^\top \nabla g(X))$ 表示提取 $X^\top \nabla g(X)$ 的对称部分，定义为

$$\text{sym}(X^\top \nabla g(X)) = (X^\top \nabla g(X) + (X^\top \nabla g(X))^\top)/2. \quad (2.17)$$

同样地，黎曼Hessian矩阵也可以通过将欧式Hessian矩阵投影到黎曼流形的切空间上来获得，即

$$\begin{aligned} \text{Hess } g(X) &= \mathcal{P}_X(\nabla^2 g(X)) \\ &= \nabla^2 g(X) - X \text{sym}(X^\top \nabla^2 g(X)). \end{aligned} \quad (2.18)$$

然而，黎曼Hessian矩阵的计算十分困难，这里采用以下近似

$$\text{Hess } g(X) \approx \frac{\text{grad } g(X + \varepsilon I) - \text{grad } g(X)}{\varepsilon}, \quad (2.19)$$

其中 $\varepsilon > 0$ 是一个极小的常数，目的是防止 $X + \varepsilon I$ 过于远离流形。信赖域算法的搜索方向通过求解以下模型得出

$$\begin{aligned} \min_{\eta \in T_W \text{St}(d, m)} \quad & m_X(\eta) = g(X) + \text{Tr}(\eta^\top \text{grad } g(X)) \\ & + \frac{1}{2} \text{vec}(\eta)^\top \text{Hess } g(X) \text{vec}(\eta) \\ \text{s.t.} \quad & \text{Tr}(\eta^\top \eta) \leq \Delta^2, \end{aligned} \quad (2.20)$$

其中 $\text{vec}(\eta)$ 表示通过将 η 列堆叠成向量， Δ 是信赖域半径。此外，信赖域比率由以下方式确定

$$\rho = \frac{g(X) - g(R_X(\eta))}{m_X(0) - m_X(\eta)}, \quad (2.21)$$

算法 1 信赖域算法求解模型 (2.14)

输入: 数据 A , 变量 Y^k , 参数 $\mu, \varepsilon, 0 < \rho_{\min} < \rho_{\max} < 1, 0 < \gamma_1 < 1 < \gamma_2, \Delta' > 0$,
 $\rho' \in [0, \frac{1}{4})$

初始化: $i = 0, X_i^{k+1} \in \text{St}(d, m), \Delta_i \in (0, \Delta')$

当 未收敛时 执行

1: 将 $X = X_i^{k+1}, \Delta = \Delta_i$ 代入模型 (2.20) 并求解得到 η_i

2: 将 $X = X_i^{k+1}, \eta = \eta_i$ 代入式 (2.21) 并计算得到 ρ_i

3: **如果** $\rho_i < \rho_{\min}$ **执行**

4: $\Delta_{i+1} = \gamma_1 \Delta_i$

5: **否则如果** $\rho_i > \rho_{\max}$ **执行**

6: $\Delta_{i+1} = \min(\gamma_2 \Delta_i, \Delta')$

7: **否则**

8: $\Delta_{i+1} = \Delta_i$

9: **结束判断**

10: **如果** $\rho_i > \rho'$ **执行**

11: $X_{i+1}^k = R_X(\eta_i)$

12: **否则**

13: $X_{i+1}^k = X_i^k$

14: **结束判断**

15: 检查收敛性

结束循环

输出: X^{k+1}

其中 $R_X(\eta)$ 是将 η 约束到黎曼流形上的收缩算子。综上, 求解模型 (2.14) 的信赖域算法见算法 1。

(2) 固定 X , 更新 Y :

$$\min_Y \quad \lambda \|Y\|_{2,0} + \mu \|X^{k+1} - Y\|_{\text{F}}^2. \quad (2.22)$$

根据范数的性质, 这里可以将模型 (2.22) 的目标函数按照行向量拆分

$$\begin{cases} \|Y\|_{2,0} = \sum_{i=1}^d \mathbb{I}(\|\mathbf{y}^i\|_2), \\ \|X^{k+1} - Y\|_F^2 = \sum_{i=1}^d \|\mathbf{x}^{i,k+1} - \mathbf{y}^i\|_2^2. \end{cases} \quad (2.23)$$

根据拆分的结果, Y 的每一个行向量可以通过下式求解

$$\min_{\mathbf{y}^i} \lambda \mathbb{I}(\mathbf{y}^i) + \mu \|\mathbf{x}^{i,k+1} - \mathbf{y}^i\|_2^2. \quad (2.24)$$

记模型(2.24)的目标函数为 $f(\mathbf{y}^i)$ 。当 $\mathbf{y}^i = 0$ 时, $f(\mathbf{y}^i) = \mu \|\mathbf{x}^{i,k+1}\|_F^2$; 当 $\mathbf{y}^i \neq 0$ 时, 取 $\mathbf{y}^i = \mathbf{x}^{i,k+1}$ 即可得到最小值 $f(\mathbf{y}^i) = \lambda$ 。最终可以得到 Y^{k+1} 的更新公式

$$\mathbf{y}^{i,k+1} = \begin{cases} 0, & \|\mathbf{x}^{i,k+1}\|_F < \sqrt{\lambda/\mu}, \\ \{0, \mathbf{x}^{i,k+1}\}, & \|\mathbf{x}^{i,k+1}\|_F = \sqrt{\lambda/\mu}, \\ \mathbf{x}^{i,k+1}, & \|\mathbf{x}^{i,k+1}\|_F > \sqrt{\lambda/\mu}. \end{cases} \quad (2.25)$$

综上所述, 求解 SPCA-R 的完整过程见算法 2。注意, \mathbf{y}^i 的求解和正则化参数 λ 、惩罚参数 μ 和变量 $\mathbf{x}^{i,k+1}$ 有关, 后续实验部分将会进行分析。

算法 2 求解 SPCA-R 的优化算法

输入: 数据 A , 参数 λ, μ

初始化: $k = 0$, 根据初始化策略得到 (X^0, Y^0)

当 未收敛时 执行

- 1: 通过算法 1 得到 X^{k+1}
- 2: 通过式 (2.25) 得到 Y^{k+1}
- 3: 检查收敛性

结束循环

输出: (X^{k+1}, Y^{k+1})

2.2.2.2 求解 SPCA-C

仿照求解 SPCA-R 的思路, 引入中间变量 $X = Y$ 并将模型(2.8)改写为

$$\begin{aligned} \min_X \quad & -\text{Tr}(X^\top A A^\top X) + \mu \|X - Y^k\|_F^2 \\ \text{s.t.} \quad & X^\top X = I_m, \|Y\|_{2,0} \leq s. \end{aligned} \quad (2.26)$$

(1) 固定 Y , 更新 X :

$$\begin{aligned} \min_X \quad & -\text{Tr}(X^\top A A^\top X) + \mu \|X - Y^k\|_F^2 \\ \text{s.t.} \quad & X^\top X = I_m. \end{aligned} \tag{2.27}$$

由于模型 (2.27) 与模型 (2.11) 相同, 因此可以直接调用算法 1, 此处省略。

(2) 固定 X , 更新 Y :

$$\begin{aligned} \min_Y \quad & \|X^{k+1} - Y\|_F^2 \\ \text{s.t.} \quad & \|Y\|_{2,0} \leq s. \end{aligned} \tag{2.28}$$

当行向量 $\mathbf{y}^i = 0$ 时, 相应行的目标函数最小值为 $\|\mathbf{x}^{i,k+1}\|_2$; 当 $\mathbf{y}^i \neq 0$ 时, 相应行的目标函数在 $\mathbf{y}^i = \mathbf{x}^{i,k+1}$ 处取到最小值 0。根据上述分析, 计算 X^{k+1} 每一行向量的 ℓ_2 范数 $\|\mathbf{x}^{i,k+1}\|_2$, 并将其中第 s 大的值记为 t_s^{k+1} , 从而得到 Y^{k+1} 的显示解

$$\mathbf{y}^{i,k+1} = \begin{cases} \mathbf{x}^{i,k+1}, & \|\mathbf{x}^{i,k+1}\|_2 \geq t_s^{k+1}, \\ 0, & \|\mathbf{x}^{i,k+1}\|_2 < t_s^{k+1}. \end{cases} \tag{2.29}$$

综上所述, 求解 SPCA-C 的完整过程见算法 3。

算法 3 求解 SPCA-C 的优化算法

输入: 数据 A , 参数 s, μ

初始化: $k = 0$, 根据初始化策略得到 (X^0, Y^0)

当 未收敛时 执行

- 1: 通过算法 1 得到 X^{k+1}
- 2: 通过式 (2.29) 得到 Y^{k+1}
- 3: 检查收敛性

结束循环

输出: (X^{k+1}, Y^{k+1})

2.3 数值实验与分析

本节通过与代表性无监督特征选择方法的比较验证 SPCA-R 和 SPCA-C 的有效性和优越性, 这些方法包括 LapScore^[20]、UDFS^[24]、SOGFS^[27]、RNE^[26] 和 SPCAFS^[22]。

其中 LapScore、UDFS、SOGFS 和 RNE 通过 AutoUFSTool^①实现，SPCAFS^②则是通过下载作者提供的代码实现。

2.3.1 实验设置

2.3.1.1 实验数据

在实验中，使用六个真实数据集来比较无监督特征选择方法的性能表现，包括一个物体图像数据集 COIL20^③，两个面部图像数据集 warpPIE10P^③ 和 UMIST^④，一个手写图像数据集 USPS^③，两个生物数据集 GLIOMA^③ 和 lung_discrete^③。在这六个真实数据集中，lung_discrete 数据集是离散的，而其余的数据集是连续的。为了书写方便，后续把 lung_discrete 简写为 lungd，有关这些数据集的详细信息如表 2.1 所示。此外，COIL20 和 UMIST 数据集的可视化如图 2.2 所示。

表 2.1 数据集信息

Table 2.1 The dataset information

数据集	特征数	样本数	类别数
COIL20	1024	1440	20
USPS	256	1000	10
lungd	325	73	7
GLIOMA	4434	50	4
UMIST	644	575	20
warpPIE10P	2420	210	10

2.3.1.2 参数设置

为了实验结果的公平性，下面首先规定所有比较方法的参数设置和选取规则。对于 LapScore，采用热核模式构造相似度矩阵，并设置温度参数 $t = 1$ 。对于 LapScore、UDFS、SOGFS 和 RNE，构造加权矩阵时“邻居数”设置为 5。对于 UDFS 和 SOGFS，聚类数设置为数据的类别数。对于 SPCAFS， $\ell_{2,p}$ 范数选择 $p = 1/2$ 。对于 SOGFS、SPCAF-S、SPCA-R 和 SPCA-C，投影维度固定为数据的类别数。

① <https://github.com/farhadabedinzadeh/AutoUFSTool>

② <https://github.com/quiter2005/algorithm>

③ <https://jundongl.github.io/scikit-feature/datasets.html>

④ https://github.com/yewang-libra/LpCNMF_github/tree/main/datasets



图 2.2 数据集可视化结果

Figure 2.2 The dataset visualization results

根据文献^[22]中的参数设定，正则化参数和惩罚参数通过网格搜索策略从给定的候选集合中调优，候选集设定为 $\{10^{-6}, 10^{-4}, 10^{-2}, 10^0, 10^2, 10^4, 10^6\}$ 。对于所有数据集，用于比较的特征数量从集合 $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ 中选取。

2.3.1.3 初始化和停止准则

对于变量 X ，采用随机正交矩阵作为初始值 X^0 ，并设 $Y^0 = X^0$ 。然而，随机正交矩阵有很大的随机性。为了减小随机性带来的影响和增加可复现性，实验中获取 10 个随机正交矩阵，并选择使 PCA 方差 $\text{Tr}(X^\top A A^\top X)$ 最大的正交矩阵作为最终的初始解^[74]。记模型 (2.10) 和 (2.26) 的目标函数分别为 $f_1(X, Y)$ 和 $f_2(X, Y)$ ，下面介绍停止准则。

(1) 算法 1 检查收敛性时满足

$$\text{grad } g(W_{i+1}^k) < 10^{-6} \quad (2.30)$$

或最大迭代次数达到 100 时停止。

(2) 算法 2 检查收敛性时满足

$$\frac{|f_1(X^{k+1}, Y^{k+1}) - f_1(X^k, Y^k)|}{1 + |f_1(X^k, Y^k)|} \leq 10^{-3} \quad (2.31)$$

或最大迭代次数达到 100 时停止。

(3) 算法 3 检查收敛性时满足

$$\frac{|f_2(X^{k+1}, Y^{k+1}) - f_2(X^k, Y^k)|}{1 + |f_2(X^k, Y^k)|} \leq 10^{-3} \quad (2.32)$$

或者最大迭代次数达到 100 时停止。

2.3.1.4 评价指标

为了评估无监督特征选择方法的性能，首先需要通过相应的方法获取所需的特征子集，随后基于选定的特征子集使用 K 均值聚类算法获得伪标签，最后利用库恩-蒙克雷斯（Kuhn-Munkres，KM）算法匹配伪标签与真实标签之间的最佳对应关系。在实验中，选取准确率（Accuracy，ACC）和归一化互信息（Normalized Mutual Information，NMI）评估聚类的性能。需要注意的是， K 均值聚类算法受初始点的影响较大，所以选择执行 50 次 K 均值聚类并计算出平均值和标准差，同时记录下最优参数下的聚类结果。

ACC 可以直观地看出聚类的准确度，定义为

$$ACC = \frac{1}{n} \sum_{i=1}^n \delta(\phi(i), \varphi(i)), \quad (2.33)$$

其中 $\phi(i)$ 和 $\varphi(i)$ 分别表示通过 KM 算法最佳匹配之后的第 i 个伪标签和真实标签，而 $\delta(\cdot)$ 是一个比较伪标签和真实标签的算子。如果 $\phi(i) = \varphi(i)$ ，则 $\delta(\phi(i), \varphi(i)) = 1$ ，否则 $\delta(\phi(i), \varphi(i)) = 0$ 。

NMI 反映了聚类结果与真实结果之间的相似度，定义为

$$NMI = \frac{I(P, Q)}{\sqrt{H(P)H(Q)}}, \quad (2.34)$$

其中 P 和 Q 分别表示伪标签和真实标签集合， $I(P, Q)$ 表示两者的互信息，而 $H(P)$ 和 $H(Q)$ 则表示各自的熵。

ACC 和 NMI 的值越大说明聚类结果越好，也反映出相应的无监督特征选择方法性能越好。后续章节如果没有特殊说明，评价指标都采用 ACC 和 NMI。

2.3.2 实验结果

图 2.3 和图 2.4 展示了不同特征数量下 ACC 和 NMI 的均值曲线，其中作为参考基准的 ALLfea 表示使用所有特征（即原始数据集）进行聚类。此外，表 2.2 和表 2.3

给出了在 100 个特征范围内最佳 ACC 和 NMI 的平均值、标准差和相应的特征数量。同时，最好和第二好的结果（除 ALLfea 外）分别使用红色和蓝色标记。

从图 2.3 中可以直观地看到，使用 $\ell_{2,0}$ 范数的 SPCA-R 和 SPCA-C 在六个数据集上都表现出了优越的性能，ACC 曲线都处于所有曲线的上方，且在除 GLIOMA 数据集外二者的表现在较为接近。使用 $\ell_{2,p}$ 范数的 SPCAFS 在 USPS 和 warpPIE10P 数据集上与 SPCA-R 和 SPCA-C 较为接近，但是在其他数据集上有一定的差距。从表 2.2

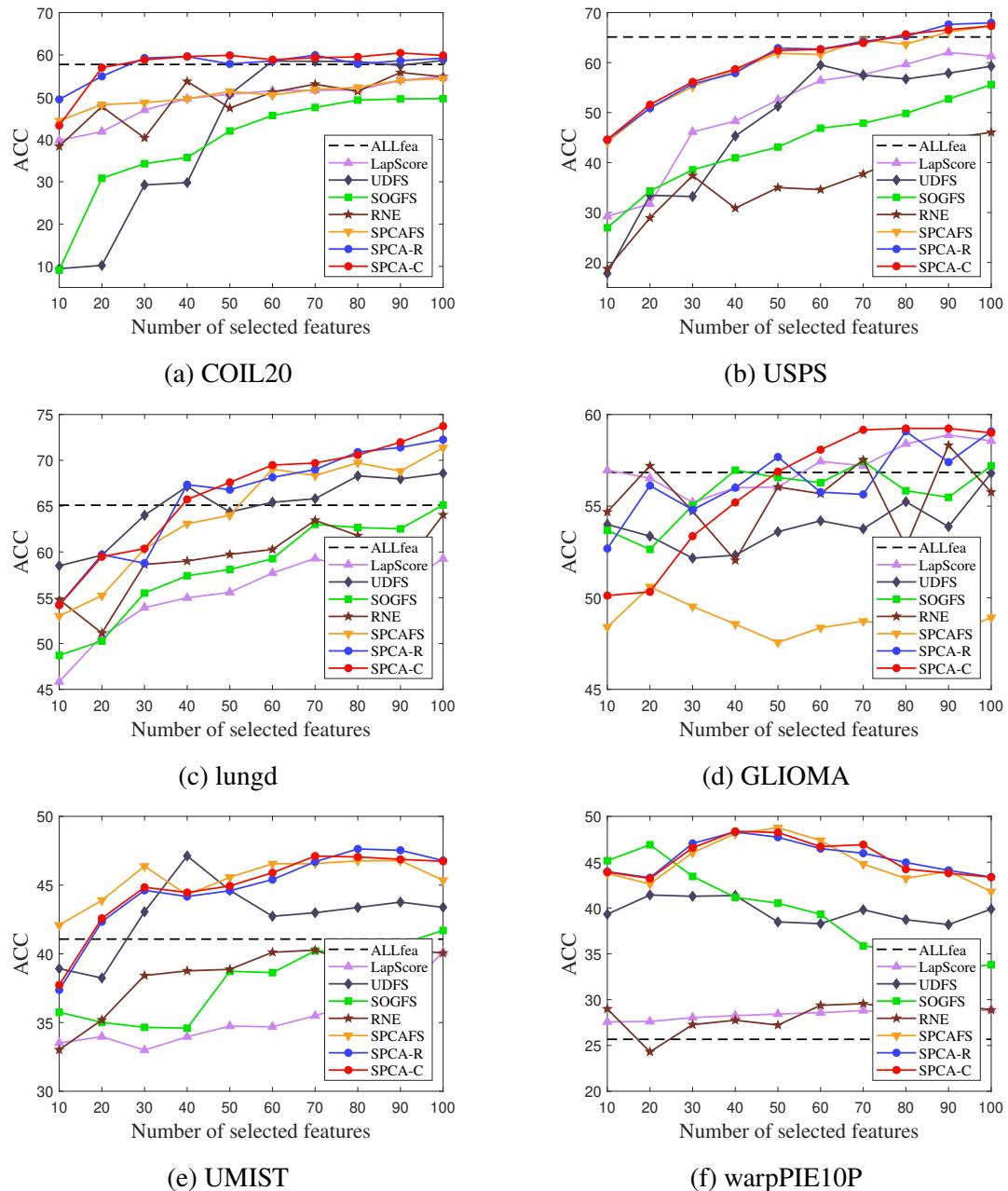


图 2.3 对比方法在六个真实数据集上的 ACC (%) 曲线

Figure 2.3 The ACC (%) curves of compared methods on six real-world datasets

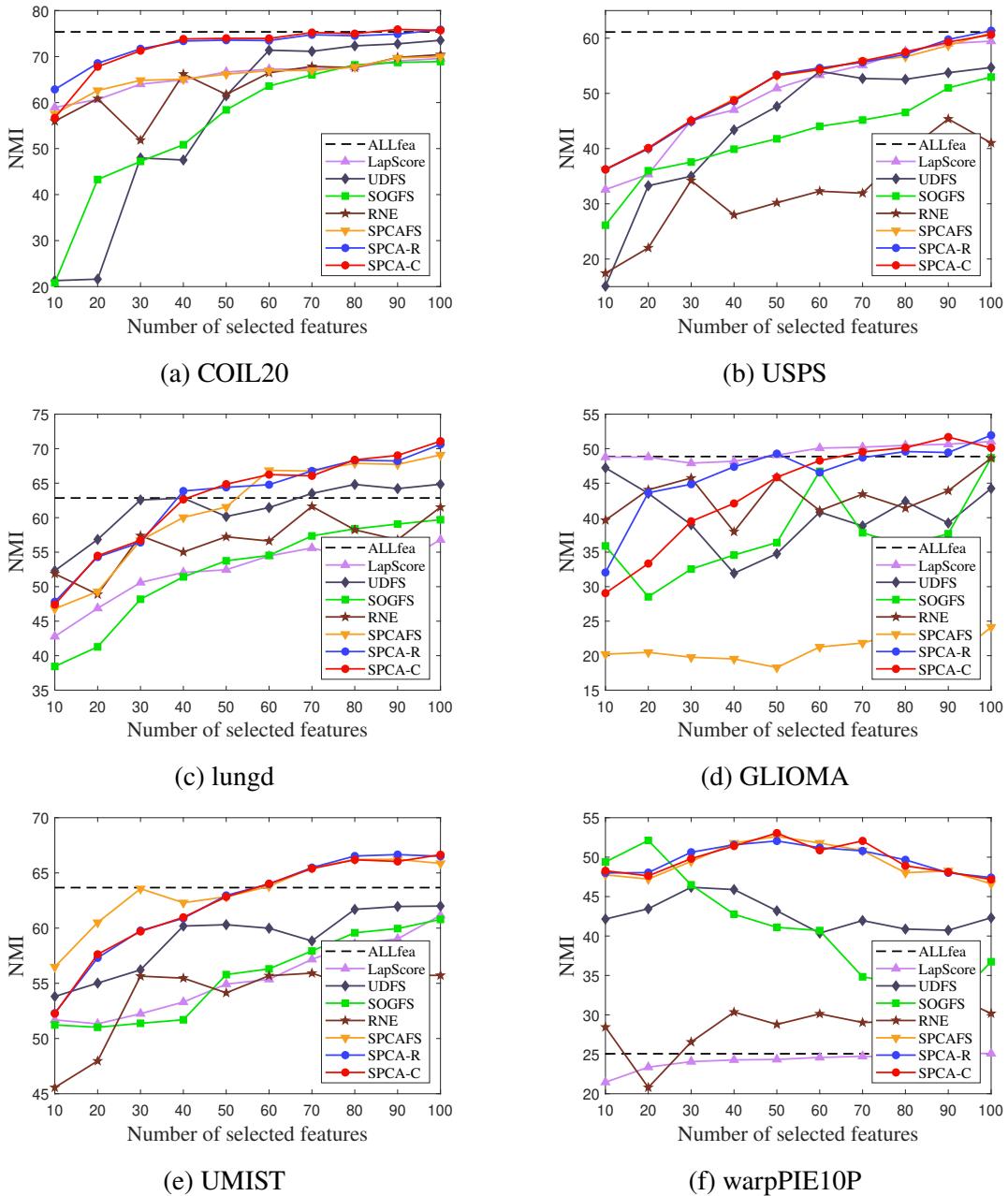


图 2.4 对比方法在六个真实数据集上的 NMI (%) 曲线

Figure 2.4 The NMI (%) curves of compared methods on six real-world datasets

中可以看到，除在 UMIST 数据集外，ACC 最好以及第二好的结果均来自 SPCAFS、SCPA-R 和 SPCA-C。特别地，利用 $\ell_{2,0}$ 范数的 SPCA-R 和 SPCA-C 性能相较于使用 $\ell_{2,p}$ 范数的 SPCAFS 表现更加优异，在 COIL20 和 GLIOMA 数据集上分别有 6.06% 和 8.64% 的提升，同时在所有方法的对比上也分别有 1.74% 和 0.36% 的提升。综合六个数据集上的 ACC 平均结果，SCPA-R 和 SPCA-C 有 2.83% 的明显提升。

从图 2.4 中也可以直观地看到与图 2.3 类似的结果，使用 $\ell_{2,0}$ 范数的 SPCA-R 和

表 2.2 对比方法在六个真实数据集上的 ACC (平均值% ± 标准差%) 结果

Table 2.2 The ACC (mean% ± std%) results of compared methods on six real-world datasets

数据集	ALLfea	LapScore	UDFS	SOGFS	RNE	SPCAFS	SPCA-R	SPCA-C
COIL20	57.74±4.93	54.82±3.91 (100)	58.71±3.47 (100)	49.66±4.81 (100)	55.84±4.41 (90)	54.39±3.67 (100)	59.94±4.70 (70)	60.45±5.43 (90)
USPS	65.12±4.95	62.02±4.09 (90)	59.52±2.97 (60)	55.58±3.07 (100)	46.04±2.69 (100)	67.34±4.49 (100)	67.96±4.46 (100)	67.31±4.55 (100)
lungd	65.10±6.44	59.29±6.33 (70)	68.58±6.99 (100)	65.12±6.89 (100)	64.05±6.65 (100)	71.37±7.68 (100)	72.25±8.62 (100)	73.73±6.35 (100)
GLIOMA	56.84±5.24	58.88±3.96 (90)	56.80±4.85 (100)	57.44±6.16 (70)	58.32±7.31 (90)	50.60±5.02 (20)	59.08±5.65 (80)	59.24±5.00 (80)
UMIST	41.07±2.38	40.13±2.79 (100)	47.12±2.49 (40)	41.70±3.17 (100)	40.35±2.26 (90)	46.78±2.51 (90)	47.63±2.68 (80)	47.11±2.99 (70)
warpPIE10P	25.67±1.90	28.94±1.66 (100)	41.42±3.18 (20)	46.90±3.89 (20)	29.57±2.96 (90)	48.76±3.86 (50)	48.30±3.95 (40)	48.37±4.02 (40)
Average	51.92±4.31	50.68±3.79	55.36±4.08	52.73±4.67	49.03±4.38	56.54±4.54	59.19±5.01	59.37±4.72

表 2.3 对比方法在六个真实数据集上的 NMI (平均值% \pm 标准差%) 结果Table 2.3 The NMI (mean% \pm std%) results of compared methods on six real-world datasets

数据集	ALLfea	LapScore	UDFS	SOGFS	RNE	SPCAFS	SPCA-R	SPCA-C
COIL20	75.37 \pm 1.96	69.59 \pm 1.48	73.54 \pm 1.76	68.92 \pm 1.84	70.43 \pm 1.92	69.98 \pm 1.45	75.86\pm1.86	75.91\pm2.03
USPS	61.12 \pm 2.01	59.46 \pm 1.80	54.69 \pm 2.11	52.96 \pm 1.54	45.36 \pm 1.93	60.98\pm2.37	61.36\pm2.45	60.64 \pm 2.55
lungd	62.85 \pm 5.13	56.79 \pm 3.99	64.84 \pm 5.09	59.70 \pm 5.24	61.63 \pm 5.83	69.09 \pm 5.61	70.62\pm6.18	71.08\pm4.39
GLIOMA	48.86 \pm 5.72	51.03 \pm 2.48	47.22 \pm 3.53	48.67 \pm 10.98	48.62 \pm 6.32	24.14 \pm 6.97	51.94\pm6.38	51.68\pm6.02
UMIST	63.67 \pm 1.85	61.16 \pm 1.71	62.00 \pm 1.58	60.79 \pm 1.54	55.92 \pm 1.57	66.23 \pm 1.60	66.66\pm1.82	66.65\pm1.75
warpPIE10P	25.07 \pm 2.88	25.13 \pm 1.73	46.18 \pm 3.30	52.12 \pm 3.25	32.67 \pm 3.31	52.63\pm3.33	52.05 \pm 2.96	53.05\pm3.16
Average	56.16 \pm 3.26	53.86 \pm 2.20	58.08 \pm 2.90	57.19 \pm 4.07	52.44 \pm 3.48	57.18 \pm 3.56	63.08\pm3.61	63.17\pm3.32

SPCA-C 的 NMI 曲线也都处于所有曲线的上方。虽然 SPCAFS 有着不错的表现，但是在 GLIOMA 数据集远低于其他方法。从表 2.3 中可以看到，NMI 最好以及第二好的结果均来自 SPCAFS、SCPA-R 和 SPCA-C。特别地，使用 $\ell_{2,0}$ 范数的 SCPA-R 和 SPCA-C 只在 USPS 和 warpPIE10P 数据集上没有同时取得前二，并且在这两个数据集上也只比取得第二好结果的 SPCAFS 分别低 0.34% 和 0.58%。然而，在 COIL20 和 GLIOMA 数据集上，使用 $\ell_{2,0}$ 范数的 SCPA-R 和 SPCA-C 相较于使用 $\ell_{2,p}$ 范数的 SPCAFS 分别有 5.93% 和 27.80% 的提升，同时在所有方法的对比上也分别有 2.37% 和 0.91% 的提升。需要注意的是，SPCAF 的 NMI 在 GLIOMA 数据集上非常低，但是在 ACC 上较为正常，这也意味着 ACC 和 NMI 并不是完全正相关的，因此使用两个指标来评估特征选择的性能更加全面。综合六个数据集上的 NMI 平均结果，SCPA-R 和 SPCA-C 有 5.09% 的明显提升。

经过上述实验结果可知，使用 $\ell_{2,0}$ 范数的 SCPA-R 和 SPCA-C 可以更充分地表示数据的稀疏结构，选择出更具有判别性并包含更多信息的特征。此外，在 PCA 的基础上无论是将 $\ell_{2,0}$ 范数加入到正则项还是约束项，其性能十分接近，且两者的平均 ACC 和 NMI 仅分别相差 0.18% 和 0.05%。综上所述，与松弛类方法相比，使用 $\ell_{2,0}$ 范数来表示稀疏结构可以提升无监督特征选择的性能。

2.3.3 稀疏度分析

从求解 SPCA-R 的更新公式 (2.25) 可以看出，求解 SPCA-R 得到变量 Y 的稀疏度与正则化参数 λ 以及惩罚参数 μ 的比值 λ/μ 有关。当 λ/μ 比值过大时容易得到全零解，而过小时可能会学习不到稀疏结构。相反地，从求解 SPCA-C 的更新公式(2.29)可以发现，求解 SPCA-C 得到变量 Y 的稀疏度和 s 相关，并且 s 可以根据实际的需求直接手动调整。由于变量 Y 的更新与变量 X 相关，并且变量 X 的更新满足 $X^\top X = I_m$ ，因此阈值 $\lambda/\mu < 1$ 更为合理。从实际的实验结果也会发现， $\lambda/\mu \geq 100$ 时可能会出现全零解，而 $\lambda/\mu \leq 10^{-4}$ 时学习不到稀疏结构。对此，本实验选择 $\lambda/\mu = 0.01$ 分析求解 SPCA-R 得到变量 Y 的稀疏度，同时设定 $s = 100$ 分析求解 SPCA-C 得到变量 Y 的稀疏度。选取 USPS 和 lungd 数据集进行试验，设定 $s = 100$ 、 $\lambda = 100$ 和 $\mu = 10000$ 。为了更完整地观察稀疏度的变化，把停止准则的相对误差设为 0，防止算法因为满足收敛条件而提前停止。变量 Y 的稀疏度在更新过程中的变化如图 2.5 所示。

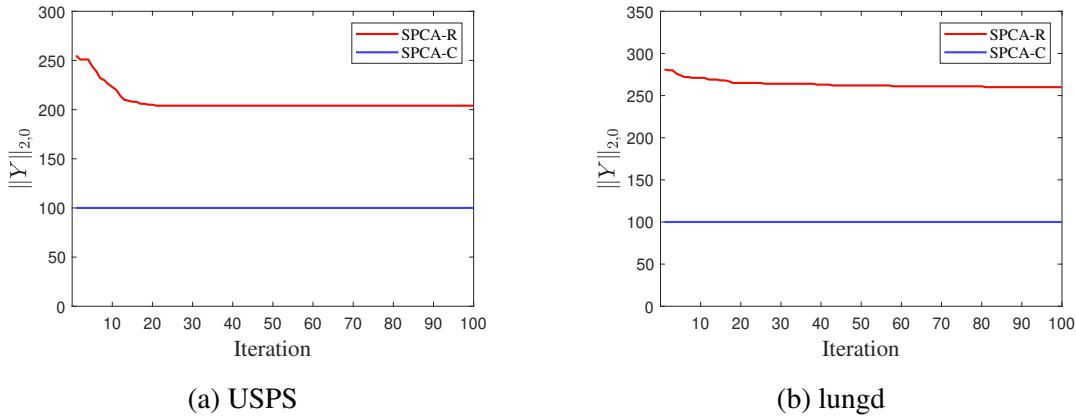


图 2.5 SPCA-R 和 SPCA-C 在两个真实数据集上的稀疏度变化曲线

Figure 2.5 The sparsity variation curves of SPCA-R and SPCA-C on two real-world datasets

根据图 2.5 可以观察到，通过 SPCA-R 得到变量 Y 的稀疏度会随着迭代次数逐渐下降，但是仍高于 SPCA-C。相反地，通过 SPCA-C 得到变量 Y 的稀疏度并不改变，仅与初始设置相关。此外，实际实验中 SPCA-R 在正则化参数 λ 与惩罚参数 μ 的不同取值下，存在全零解和学习不到稀疏结构的可能。因此，SPCA-C 相较于 SPCA-R 可以避免无效解，并且拥有更低的稀疏度和更灵活的特征选择。

2.3.4 讨论

2.3.4.1 参数敏感度分析

正如前文所述，参数的取值会影响稀疏度，进而影响无监督特征选择的性能。本小节对 SPCA-R 的两个参数 λ 和 μ 以及 SPCA-C 的两个参数 μ 和 s 进行了参数敏感度分析，不同参数下的聚类结果分别如图 2.6 和 2.7 所示。

从图 2.6 可以看出，参数 λ 和 μ 同时影响 SPCA-R 特征选择的性能。特别在 lungd 数据集上，当 $\mu = 10^6$ 且 $\lambda \neq 10^6$ 时，ACC 和 NMI 明显降低。在稀疏度分析中提到阈值 λ/μ 过小可能学习不到稀疏结构，而 μ 过大恰恰会导致阈值过小，因此性能的影响可能是阈值变化的结果。此外， μ 控制着约束 $X = Y$ 的违反程度，也可能会影响求解变量的结果。从图 2.7 可以看出，参数 μ 和 s 同时影响 SPCA-C 特征选择的性能。同时，可以看到如图 2.6 类似的结果，当 μ 过大时会影响 SPCA-C 的性能。此外， s 就是所选取的特征，这也是为什么 SPCA-C 对 s 的变化敏感，并且结果的变化趋势和实验结果保持一致。

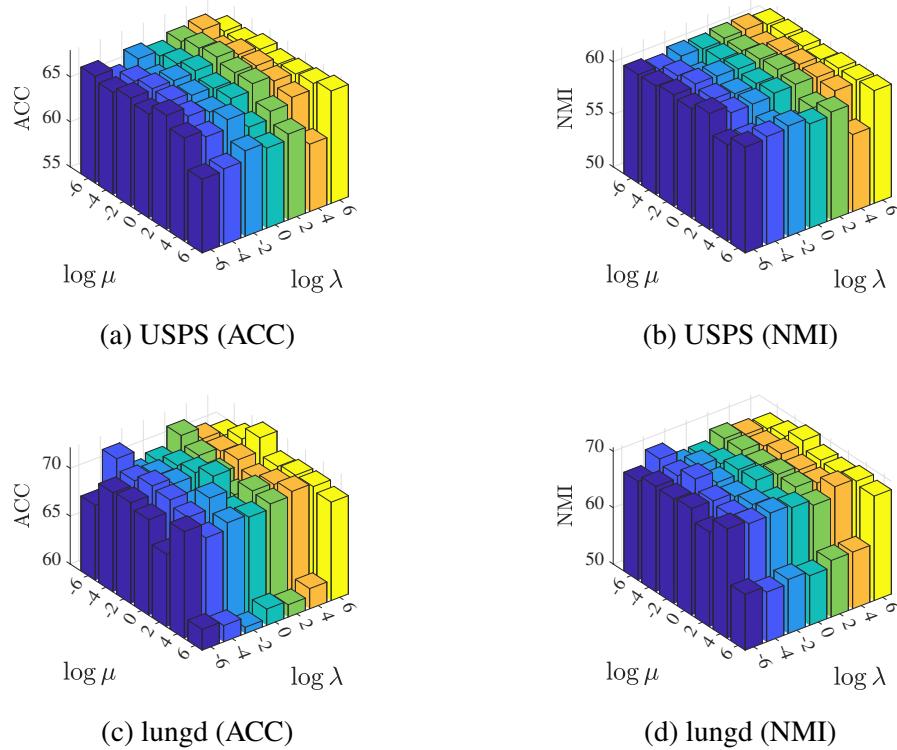


图 2.6 SPCA-R 在两个真实数据集上的参数敏感度分析结果

Figure 2.6 The parameter sensitivity analysis results of SPCA-R on two real-world datasets

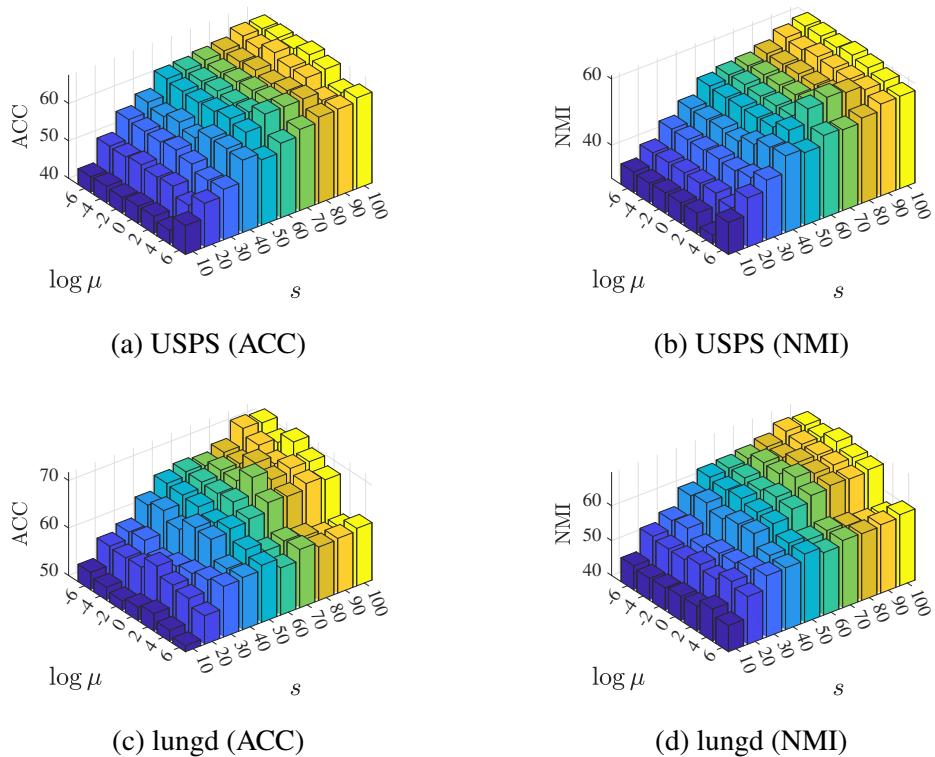


图 2.7 SPCA-C 在两个真实数据集上的参数敏感度分析结果

Figure 2.7 The parameter sensitivity analysis results of SPCA-C on two real-world datasets

2.3.4.2 模型收敛性分析

图 2.8 显示了进行特征选择时，SPCA-R 和 SPCA-C 的目标函数值 $f_1(X, Y)$ 和 $f_2(X, Y)$ 随迭次数变化的曲线。结果表明，SPCA-R 和 SPCA-C 在大多数情况下能够持续下降，并在 100 次迭代内达到稳定状态，这保证了它们在实际应用中的有效性。

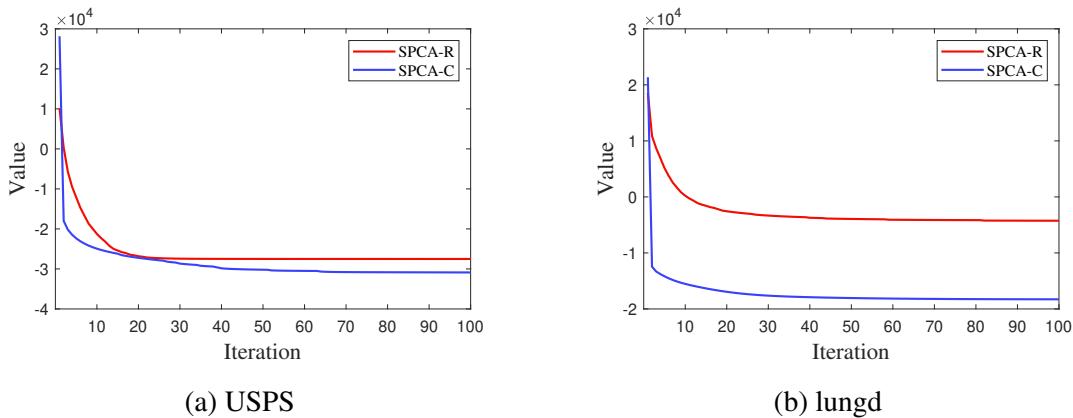


图 2.8 SPCA-R 和 SPCA-C 在两个真实数据集上的收敛曲线

Figure 2.8 The convergence curves of SPCA-R and SPCA-C on two real-world datasets

2.3.4.3 同模型对比分析

通过前面的实验已经验证了相比松弛模型 (2.5)，原始模型 (2.6) 在无监督特征选择上有更优越的性能。而对 SPCA-C 模型，Nie 等人^[43]研究了基于数据协方差矩阵秩的全局最优算法和迭代代理更新算法，在文中把 SPCA-C 模型描述为 FSPCA 问题。使用作者提供的程序^①与算法 3 分别求解模型 (2.8)，执行 K 均值聚类并记录 ACC 和 NMI，表 2.4 给出了对比实验的结果。

从表 2.4 中的结果可以观察到，SPCA-C 和 FSPCA 在 USPS 和 UMIST 数据集上的性能相差不大，而在其他数据集上 SPCA-C 地表现优于 FSPCA。特别地，在 GLIOMA 和 warpPIE10P 数据集上，FSPCA 的性能表现较差。这可能是因为 FSPCA 根据原始数据协方差矩阵的秩来设计算法，容易受到噪声和冗余影响。综合结果来看，与 FSPCA 相比，SPCA-C 拥有更好的性能表现。同时也说明，对于非凸、非光滑的 SPCA 问题，即便模型相同，不同的优化算法也会导致差距较大的结果。

^① <https://github.com/tianlai09/FSPCA>

表 2.4 SPCA-R 和 FSPCA 在六个真实数据集上的 ACC (%) 和 NMI (%)

Table 2.4 The ACC (%) and NMI (%) of SPCA-R and FSPCA on six real-world datasets

数据集	方法	ACC	NMI
COIL20	SPCA-C	60.45±5.43	75.91±2.03
	FSPCA	50.15±4.70	68.50±1.56
USPS	SPCA-C	67.31±4.55	60.64±2.55
	FSPCA	67.38±4.36	62.00±1.87
lungd	SPCA-C	73.73±6.35	71.08±4.39
	FSPCA	60.19±6.55	58.26±6.39
GLIOMA	SPCA-C	59.24±5.00	51.68±6.02
	FSPCA	47.92±4.61	21.94±5.28
UMIST	SPCA-C	47.11±2.99	66.65±1.75
	FSPCA	46.70±2.29	65.27±1.58
warpPIE10P	SPCA-C	48.37±4.02	53.05±3.16
	FSPCA	28.01±2.27	23.90±2.01

2.4 本章小结

本章从 PCA 出发, 探讨了 $\ell_{2,0}$ 范数正则的 SPCA-R 模型和 $\ell_{2,0}$ 范数约束的 SPCA-C 模型。针对两个模型都存在的正交约束, 首先将其描述为一个黎曼空间中的 Stiefel 流形模型, 然后利用信赖域算法求解。针对 $\ell_{2,0}$ 范数约束模型, 利用硬阈值算子求得一个灵活的显示解。大量的数值实验表明, 相较于传统松弛方法而言, $\ell_{2,0}$ 范数更加充分地表示稀疏结构, 并且展现出其在无监督特征选择的性能优势。具体地, SPCA-C 在六个真实数据集上的平均 ACC 和 NMI 相较 SPCAFS 分别有 2.83% 和 5.99% 的提升。虽然 SPCA-R 和 SPCA-C 的平均 ACC 和 NMI 仅分别相差 0.18% 和 0.05%, 但是 SPCA-R 的阈值会随正则化参数变化, 可能会导致完全稀疏或者完全不稀疏的情况, 而 SPCA-C 通过 $\ell_{2,0}$ 范数约束避免了上述情况的出现。值得注意的是, 通过与 FSPCA 的对比分析说明了设计有效的算法也是提升无监督特征选择性能的关键。

第三章 基于双稀疏约束的无监督特征选择方法

上一章论述了 $\ell_{2,0}$ 范数约束在无监督特征选择任务中的性能优势，但它仅能描述单一的稀疏结构。当实际数据呈现出更加复杂的结构稀疏时， $\ell_{2,0}$ 范数容易忽略局部结构信息，从而无法很好地完成特征选择。因此，本章将 $\ell_{2,0}$ 范数和 ℓ_0 范数双稀疏约束嵌入到稀疏主成分分析框架中，建立了新的无监督特征选择模型，即 DSCOFS。其核心思想在于， $\ell_{2,0}$ 范数能够去除冗余特征，而 ℓ_0 范数能够过滤掉不规则的噪声特征，从而在一定程度上实现了与 $\ell_{2,0}$ 范数的互补，提升了特征判别的能力。在算法方面，设计了基于一阶精确罚函数和硬阈值的优化策略，并且从理论上严格证明了该算法所产生序列的收敛性。最后，大量的数值实验表明了双稀疏约束的有效性和所提出 DSCOFS 的优越性。

3.1 相关工作

文献^[75-76]表明，相比其他形式的范数， $\ell_{2,0}$ 范数更适合用于特征选择，并且能够更有效地表示稀疏结构。回顾上一章用于同模型对比分析的 FSPCA 和 SPCA-C，其本质就是在 PCA 中加入 $\ell_{2,0}$ 范数约束选择判别特征。具体的数学模型为

$$\begin{aligned} \min_X \quad & -\text{Tr}(X^\top A A^\top X) \\ \text{s.t.} \quad & X^\top X = I_m, \|X\|_{2,0} \leq s, \end{aligned} \tag{3.1}$$

其中 $s > 0$ 表示选择特征的数量。具体分析见模型 (2.8)，此处不再赘述。

当数据呈现出元素稀疏时，模型 (3.1) 可能会导致较差的特征选择，从而影响聚类结果^[77-78]。因此，一个自然的想法是考虑多个稀疏的混合以应对不同的稀疏结构。经过查阅文献，本文发现一些研究已经将双稀疏应用于信号恢复^[79]、雷达成像^[80]、压缩感知^[81] 和特征选择^[82] 等领域，但它们并未对同一变量进行约束。值得注意的是，在脑成像预测因子识别^[83] 和基因表达研究^[84] 中，成功地将 $\ell_{2,1}$ 范数与 ℓ_1 范数正则化引入目标函数，且应用于同一变量。

然而， $\ell_{2,0}$ 范数和 ℓ_0 范数约束在确定稀疏结构时，比传统的正则化方法和松弛技术更具灵活性和精确性^[85]。对于不同范数约束下所获得的稀疏情况如图 3.1 所示。

$\ell_{2,0}$ 范数和 ℓ_0 范数分别实现相应的行稀疏和元素稀疏，而 $\ell_{2,0}$ 和 ℓ_0 范数的双稀疏约束去除了更多冗余特征，并实现了更加复杂的稀疏结构表示。此外， $\ell_{2,0}$ 范数和 ℓ_0 范数双稀疏选择出了特征 1、4 和 5，这与 $\ell_{2,0}$ 稀疏选择出的特征 1、3 和 5 是不同的。实际上，双稀疏可以看作连接单稀疏与实际稀疏的桥梁，更符合实际的稀疏分布。

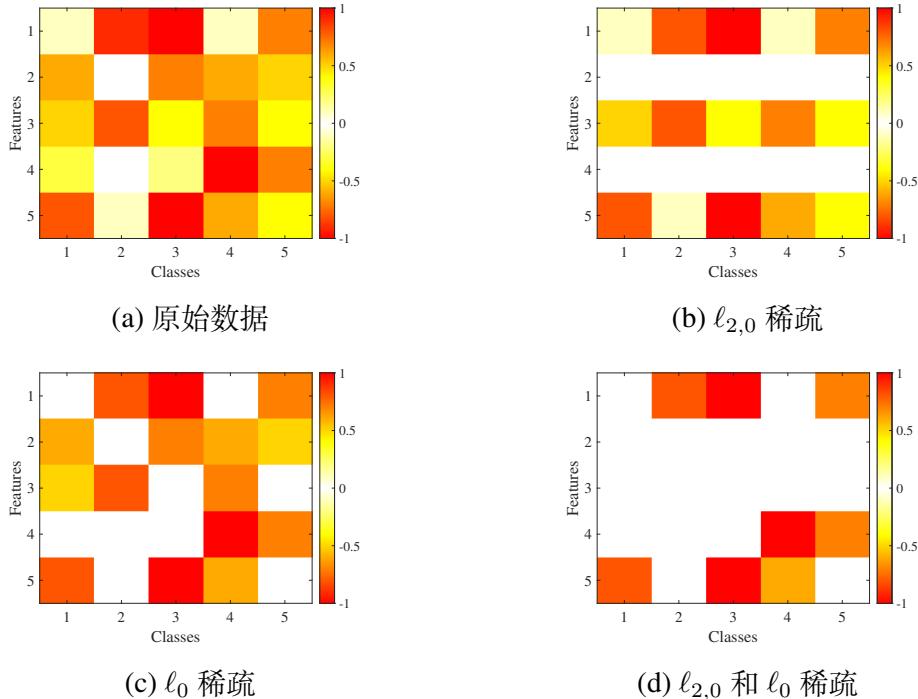


图 3.1 不同稀疏约束下获得的结果示例

Figure 3.1 The examples of results obtained by different sparsity constraints

3.2 数学模型与算法

本节首先提出双稀疏约束的模型，其次设计基于一阶精确罚函数和硬阈值的优化算法，最后详细证明算法的收敛性。

3.2.1 数学模型

基于上述观察，本章借助非松弛的 $\ell_{2,0}$ 范数和 ℓ_0 范数构建了一种新的双稀疏约束优化特征选择 (Double Sparsity Constrained Optimization Feature Selection, DSCOFS)，

其数学模型为

$$\begin{aligned} \min_X \quad & -\text{Tr}(X^\top A A^\top X) \\ \text{s.t.} \quad & X^\top X = I_m, \|X\|_{2,0} \leq s_1, \|X\|_0 \leq s_2, \end{aligned} \quad (3.2)$$

其中 $s_1 > 0$ 表示非零行的数量, $s_2 > 0$ 表示非零元素的数量。与单一 $\ell_{2,0}$ 范数模型(3.1)相比, DSCOFS 的优势为

- $\|X\|_0 \leq s_2$ 可以过滤掉不规则的噪声和局部的元素稀疏项, 弥补了 $\|X\|_{2,0} \leq s_1$ 的局限性, 最终更容易识别出具有差异性的特征。
- 双稀疏在特征选择中提供了更多的灵活性, 参数 s_1 和 s_2 可以根据实际任务直接设定和调整。

最终, DSCOFS 特征选择和聚类的流程如图 3.2 所示。

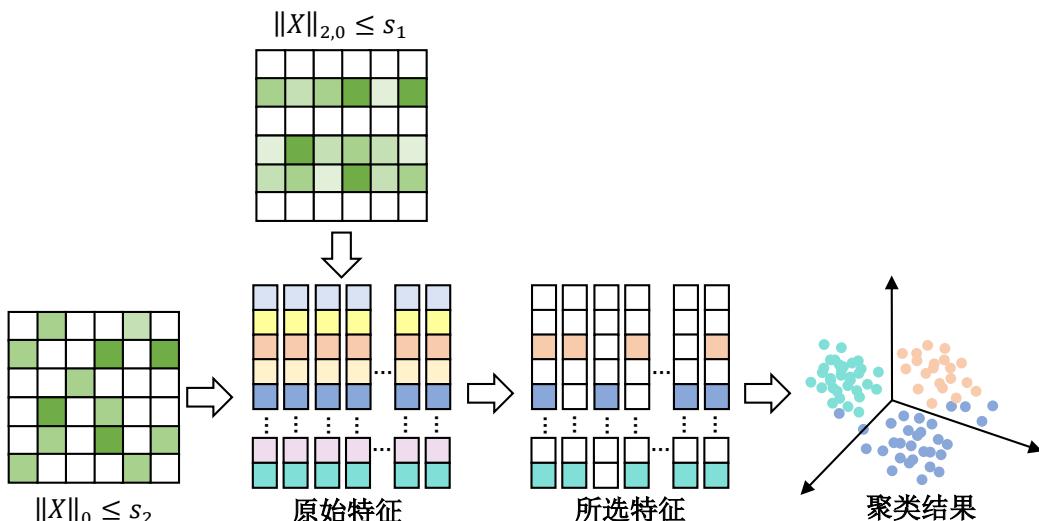


图 3.2 DSCOFS 特征选择和聚类的流程图

Figure 3.2 The flowchart of feature selection and clustering of DSCOFS

3.2.2 优化算法

对于模型 (3.2), 观察到它有三个非凸约束, 即 $X^\top X = I_m$ 、 $\|X\|_{2,0} \leq s_1$ 和 $\|X\|_0 \leq s_2$ 。这些约束同时求解十分困难, 受文献^[59]的启发, 这里采用一种更为有效的方法——近端交替最小化 (Proximal Alternating Minimization, PAM), 它可以在保证一定收敛性的同时交替更新变量。通过引入辅助变量 $X = Y$ 和 $X = Z$, 模型

(3.2) 可以改写为

$$\begin{aligned} \min_{X,Y,Z} \quad & -\text{Tr}(X^\top A A^\top X) \\ \text{s.t.} \quad & X^\top X = I_m, \|Z\|_{2,0} \leq s_1, \|Y\|_0 \leq s_2, \\ & X = Y, X = Z. \end{aligned} \tag{3.3}$$

记

$$\begin{aligned} \mathcal{M} &= \{X \in \mathbb{R}^{d \times m} \mid X^\top X = I_m\}, \\ \mathcal{S}_1 &= \{Z \in \mathbb{R}^{d \times m} \mid \|Z\|_{2,0} \leq s_1\}, \\ \mathcal{S}_2 &= \{Y \in \mathbb{R}^{d \times m} \mid \|Y\|_0 \leq s_2\}. \end{aligned} \tag{3.4}$$

于是，利用惩罚函数方法，可以将模型 (3.3) 转化为

$$\begin{aligned} \min_{X,Y,Z} \quad & -\text{Tr}(X^\top A A^\top X) + \mu_1 \|X - Y\|_{\text{F}}^2 + \mu_2 \|X - Z\|_{\text{F}}^2 \\ \text{s.t.} \quad & X \in \mathcal{M}, Z \in \mathcal{S}_1, Y \in \mathcal{S}_2, \end{aligned} \tag{3.5}$$

其中 $\mu_1, \mu_2 > 0$ 是惩罚参数。设 X^k, Y^k 和 Z^k 是第 k 次更新的变量，同时在迭代过程中引入近端参数 $\tau_1, \tau_2, \tau_3 > 0$ ，用于保证算法的收敛性。

(1) 固定 Y 和 Z ，更新 X :

$$\begin{aligned} \min_X \quad & -\text{Tr}(X^\top A A^\top X) + \mu_1 \|X - Y^k\|_{\text{F}}^2 + \mu_2 \|X - Z^k\|_{\text{F}}^2 + \tau_1 \|X - X^k\|_{\text{F}}^2 \\ \text{s.t.} \quad & X \in \mathcal{M}. \end{aligned} \tag{3.6}$$

这是一个 Stiefel 流形优化模型。上一章中采用了信赖域算法求解，不过 Hessian 矩阵的计算十分复杂且效率较低。尽管存在一些针对 Stiefel 流形的优化方法，但其中多是依赖于沿着 \mathcal{M} 的测地线生成点，这使得计算效率依旧不高。为了保持 Stiefel 流形的结构并避免计算测地线，下面采用了一种称为精确罚函数^[86]的有效不可行方法。这里“不可行”指的是变量在更新过程中不会严格处于流形约束中，而是控制在一定的小范围内。根据文献^[87]，模型 (3.6) 可以转化

$$\min_{X \in \mathcal{C}} h(X) = l(X) + g(X), \tag{3.7}$$

其中 \mathcal{C} 为紧致凸约束，这里采用 Frobenius 范数下半径为 ρ 的球体，即 $\mathcal{B}_\rho = \{X \in$

$\mathbb{R}^{d \times m} \mid \|X\|_{\text{F}} \leq \rho\}$ 。模型 (3.7) 的第一项为

$$l(X) = -\text{Tr}(X^T A A^T X) + \mu_1 \|X - Y^k\|_{\text{F}}^2 + \mu_2 \|X - Z^k\|_{\text{F}}^2 + \tau_1 \|X - X^k\|_{\text{F}}^2, \quad (3.8)$$

第二项为

$$g(X) = -\frac{1}{2} \langle \Lambda(X), X^T X - I_m \rangle + \frac{\beta}{4} \|X^T X - I_m\|_{\text{F}}^2. \quad (3.9)$$

在式 (3.9) 中, 惩罚参数 $\beta > 0$, 同时有

$$\Lambda(X) = \frac{1}{2} (X^T \nabla l(X) + \nabla l(X)^T X). \quad (3.10)$$

模型 (3.6) 和模型 (3.7) 在全局最小值的意义下的等价性已在文献^[86]中定理 3.2 证明。通过观察式 (3.10) 可以发现, $h(X)$ 中包含了 $\nabla l(X)$, 这使得对 $h(X)$ 求导需要计算 $l(X)$ 的 Hessian 矩阵。为此, 本章考虑了 $h(X)$ 的近似梯度, 即

$$\begin{aligned} D(X) = & -2AA^T X + 2\mu_1(X - Y^k) + 2\mu_2(X - Z^k) \\ & + 2\tau_1(X - X^k) - X\Lambda(X) + \beta X(X^T X - I_m). \end{aligned} \quad (3.11)$$

根据式 (3.11), 模型 (3.7) 可以通过近似梯度下降法来求解。这里, 梯度下降法的迭代公式为

$$\hat{X}^{k+1} = X^k - \eta_k D(X^k), \quad (3.12)$$

其中, $D(X^k)$ 是 $h(X)$ 在点 X^k 处的近似梯度, 而 $\eta_k > 0$ 是 Barzilai-Borwein 步长^[88]。

如果任由迭代更新而不加以限制, 更新得到的 \hat{X}^{k+1} 不能总满足 Stiefel 流形约束 \mathcal{M} 。因此, 这里选择一个半径为 $\rho > \sqrt{m}$ 的球域。记 $\mathcal{P}_{\mathcal{B}_\rho}$ 是数据在球域 \mathcal{B}_ρ 上的投影, 则 X^{k+1} 的更新可以表述为

$$X^{k+1} = \mathcal{P}_{\mathcal{B}_\rho}(\hat{X}^{k+1}). \quad (3.13)$$

通过收缩 X^{k+1} , 可以保证更新的 X^{k+1} 始终处在球域 \mathcal{B}_ρ 内。而当更新的 X^{k+1} 足够接近 Stiefel 流形约束 \mathcal{M} 时, 则可以直接使用式 (3.12) 来更新 X^{k+1} 。综上, 更新 X^{k+1} 的详细迭代见算法 4。

算法 4 精确罚函数方法求解模型 (3.6)

输入: 数据 A , 参数 β, ρ, η **初始化:** $k = 0, X^0 \leftarrow X^k$ **当 未收敛时 执行**1: 通过式 (3.11) 计算 $D(X^k)$ 2: 通过式 (3.12) 计算 \hat{X}^{k+1} 3: 如果 $\|\hat{X}^{k+1}\|_F > \rho$ 执行

4:
$$X^{k+1} = \frac{\rho}{\|\hat{X}^{k+1}\|_F} \hat{X}^{k+1}$$

5: 否则

6:
$$X^{k+1} = \hat{X}^{k+1}$$

7: 结束判断

8: 检查收敛性

结束循环**输出:** X^{k+1} (2) 固定 X 和 Z , 更新 Y :

$$\begin{aligned} \min_Y \quad & \|X^{k+1} - Y\|_F^2 + \tau_2 \|Y - Y^k\|_F^2 \\ \text{s.t.} \quad & Y \in \mathcal{S}_2. \end{aligned} \tag{3.14}$$

合并模型 (3.14) 目标函数中的 Frobenius 范数, 可以改写为

$$\begin{aligned} \min_Y \quad & \left\| \frac{X^{k+1} + \tau_2 Y^k}{1 + \tau_2} - Y \right\|_F^2 \\ \text{s.t.} \quad & Y \in \mathcal{S}_2. \end{aligned} \tag{3.15}$$

记 $W^{k+1} = \frac{X^{k+1} + \tau_2 Y^k}{1 + \tau_2}$ 。对 W^{k+1} 取绝对值, 将其中第 s_2 大的值记为 $t_{s_2}^{k+1}$ 。通过硬阈值算子^[89]可以直接得到 Y^{k+1} 的显示解

$$Y_{ij}^{k+1} = \begin{cases} W_{ij}^{k+1}, & |W_{ij}^{k+1}| \geq t_{s_2}^{k+1}, \\ 0, & |W_{ij}^{k+1}| < t_{s_2}^{k+1}, \end{cases} \tag{3.16}$$

其中 $|\cdot|$ 表示绝对值。在实际应用中, 可以根据需求设定不同稀疏度 s_2 , 显示解 Y^{k+1} 都可以很容易地按照式 (3.16) 直接确定。

(3) 固定 X 和 Y , 更新 Z :

$$\begin{aligned} \min_Z \quad & \|X^{k+1} - Z\|_F^2 + \tau_3 \|Z - Z^k\|_2^2 \\ \text{s.t.} \quad & Z \in \mathcal{S}_1. \end{aligned} \tag{3.17}$$

类似于更新 Y^{k+1} , 模型 (3.17) 可以改写为

$$\begin{aligned} \min_Z \quad & \left\| \frac{X^{k+1} + \tau_3 Z^k}{1 + \tau_3} - Z \right\|_2^2 \\ \text{s.t.} \quad & Z \in \mathcal{S}_1. \end{aligned} \tag{3.18}$$

同样地, 记 $V^{k+1} = \frac{X^{k+1} + \tau_3 Z^k}{1 + \tau_3}$ 。计算 V^{k+1} 每一行的 ℓ_2 范数 $\|\mathbf{v}^{i,k+1}\|_2$, 将其中第 s_1 大的值记为 $t_{s_1}^{k+1}$ 。考虑到 $\ell_{2,0}$ 范数的行稀疏性质, Z^{k+1} 有显示解

$$\mathbf{z}^{i,k+1} = \begin{cases} \mathbf{v}^{i,k+1}, & \|\mathbf{v}^{i,k+1}\|_F \geq t_{s_1}^{k+1}, \\ 0, & \|\mathbf{v}^{i,k+1}\|_F < t_{s_1}^{k+1}. \end{cases} \tag{3.19}$$

通过式 (3.19) 可以看出, Z^{k+1} 仅保留 ℓ_2 范数意义下最大的 s_1 行。同时, 稀疏度 s_1 可以如同 s_2 一样自由调整, 这表明双稀疏约束在特征选择中能够提供更多的灵活性。

综上所述, 求解 DSCOFS 的完整过程见算法 5。从优化的角度来看, 更新 Y^{k+1} 和 Z^{k+1} 的顺序对结果没有影响。然而, 在特征选择的应用中, 应该首先在 \mathcal{S}_2 上更新 Y^{k+1} 达到去除冗余的目的, 然后在 \mathcal{S}_1 上更新 Z^{k+1} 达到选择特征的目的。否则, 很难有效地影响特征选择的结果。

算法 5 求解 DSCOFS 的优化算法

输入: 数据 A , 参数 $s_1, s_2, \mu_1, \mu_2, \tau_1, \tau_2, \tau_3$

初始化: $k = 0$, 根据初始化策略得到 (X^0, Y^0, Z^0)

当 未收敛时 执行

- 1: 通过算法 4 得到 X^{k+1}
- 2: 通过式 (3.16) 得到 Y^{k+1}
- 3: 通过式 (3.19) 得到 Z^{k+1}
- 4: 检查收敛性

结束循环

输出: $(X^{k+1}, Y^{k+1}, Z^{k+1})$

3.2.3 理论分析

为了简化符号，将模型 (3.5) 的目标函数记作

$$f(X, Y, Z) = -\text{Tr}(X^\top A A^\top X) + \mu_1 \|X - Y\|_{\text{F}}^2 + \mu_2 \|X - Z\|_{\text{F}}^2. \quad (3.20)$$

显然， $f(X, Y, Z)$ 是连续可微的，其梯度为

$$\begin{aligned} \nabla f(X, Y, Z) &= \begin{bmatrix} \frac{\partial}{\partial X} f(X, Y, Z) \\ \frac{\partial}{\partial Y} f(X, Y, Z) \\ \frac{\partial}{\partial Z} f(X, Y, Z) \end{bmatrix} \\ &= \begin{bmatrix} -2(AA^\top X - \mu_1(X - Y) - \mu_2(X - Z)) \\ 2\mu_1(Y - X) \\ 2\mu_2(Z - X) \end{bmatrix}. \end{aligned} \quad (3.21)$$

设 $N_{\mathcal{M} \times \mathcal{S}_2 \times \mathcal{S}_1}(X, Y, Z)$ 表示为 $\mathcal{M} \times \mathcal{S}_2 \times \mathcal{S}_1$ 在点 (X, Y, Z) 处的法锥，并记

$$\begin{aligned} \lambda_0 &= \sup_{X \in \mathcal{B}_\rho} \max\{1, \|\nabla l(X)\|_{\text{F}}\}, \\ \lambda_1 &= \sup_{X \in \mathcal{B}_\rho} \max\{1, \|\Lambda(X)\|_{\text{F}}\}, \\ \lambda_2 &= \sup_{X_1, X_2 \in \mathcal{B}_\rho} \max\left\{1, \frac{\|\Lambda(X_1) - \Lambda(X_2)\|}{\|X_1 - X_2\|}\right\}, \end{aligned} \quad (3.22)$$

其中， $\|\cdot\|$ 表示矩阵的谱范数，即矩阵的最大奇异值， \sup 表示上确界。

引理 3.1 设 $\{(X^k, Y^k, Z^k)\}$ 是算法 5 产生的序列，则 $\{f(X^k, Y^k, Z^k)\}$ 严格递减。

证明：设 X^{k+1} 、 Y^{k+1} 和 Z^{k+1} 分别为模型 (3.6)、(3.14) 和 (3.17) 的解。对于任意的 $X^k \in \mathcal{M}$ 、 $Y^k \in \mathcal{S}_2$ 和 $Z^k \in \mathcal{S}_1$ ，有

$$\begin{aligned} f(X^{k+1}, Y^k, Z^k) &\leq f(X^k, Y^k, Z^k) - \tau_1 \|X^{k+1} - X^k\|_{\text{F}}^2, \\ f(X^{k+1}, Y^{k+1}, Z^k) &\leq f(X^{k+1}, Y^k, Z^k) - \tau_2 \|Y^{k+1} - Y^k\|_{\text{F}}^2, \\ f(X^{k+1}, Y^{k+1}, Z^{k+1}) &\leq f(X^{k+1}, Y^{k+1}, Z^k) - \tau_3 \|Z^{k+1} - Z^k\|_{\text{F}}^2. \end{aligned} \quad (3.23)$$

由此可得

$$\begin{aligned} f(X^{k+1}, Y^{k+1}, Z^{k+1}) + \tau_1 \|X^{k+1} - X^k\|_{\text{F}}^2 + \tau_2 \|Y^{k+1} - Y^k\|_{\text{F}}^2 + \tau_3 \|Z^{k+1} - Z^k\|_{\text{F}}^2 \\ \leq f(X^k, Y^k, Z^k). \end{aligned} \quad (3.24)$$

因此，迭代更新规则使得目标函数序列是严格递减的。证毕。 \square

引理 3.2 设 $\{(X^k, Y^k, Z^k)\}$ 是算法 5 产生的序列，则 $\{(X^k, Y^k, Z^k)\}$ 有界。

证明：序列 $\{(X^k, Y^k, Z^k)\}$ 的有界性通过反证法证明。假设序列 $\{(X^k, Y^k, Z^k)\}$ 是无界的，因此有

$$\lim_{k \rightarrow \infty} \|(X^k, Y^k, Z^k)\|_F = \infty. \quad (3.25)$$

根据式 (3.25) 和 $f(X, Y, Z)$ 的强制性，序列 $\{f(X^k, Y^k, Z^k)\}$ 应发散到无穷大。记

$$\|E^{k+1} - E^k\|_F^2 = \tau_1 \|X^{k+1} - X^k\|_F^2 + \tau_2 \|Y^{k+1} - Y^k\|_F^2 + \tau_3 \|Z^{k+1} - Z^k\|_F^2. \quad (3.26)$$

从式 (3.24) 得到

$$\begin{aligned} & f(X^{k+1}, Y^{k+1}, Z^{k+1}) \\ & \leq f(X^{k+1}, Y^{k+1}, Z^{k+1}) + \|E^{k+1} - E^k\|_F^2 \leq f(X^k, Y^k, Z^k) \\ & \leq f(X^k, Y^k, Z^k) + \|E^k - E^{k-1}\|_F^2 \\ & \leq \dots \\ & \leq f(X^0, Y^0, Z^0), \end{aligned} \quad (3.27)$$

这意味着 $f(X^k, Y^k, Z^k)$ 对于任何 k 都有界，从而与假设矛盾。因此，序列 $\{(X^k, Y^k, Z^k)\}$ 有界。证毕。 \square

推论 3.1 设 $\{(X^k, Y^k, Z^k)\}$ 是算法 5 产生的有界序列，则 $\{(X^k, Y^k, Z^k)\}$ 满足

$$\lim_{k \rightarrow \infty} \|(X^{k+1}, Y^{k+1}, Z^{k+1}) - (X^k, Y^k, Z^k)\|_F = 0. \quad (3.28)$$

证明：设 K 是一个正整数且 $K > 1$ 。对式 (3.24) 在 $k = 0, 1, \dots, K-1$ 范围内求和，得到

$$\begin{aligned} & \sum_{k=0}^{K-1} \tau_1 \|X^{k+1} - X^k\|_F^2 + \tau_2 \|Y^{k+1} - Y^k\|_F^2 + \tau_3 \|Z^{k+1} - Z^k\|_F^2 \\ & \leq \sum_{k=0}^{K-1} (f(X^k, Y^k, Z^k) - f(X^{k+1}, Y^{k+1}, Z^{k+1})) \\ & \leq f(X^0, Y^0, Z^0) - f(X^K, Y^K, Z^K) \\ & < +\infty, \end{aligned} \quad (3.29)$$

其中最后一个不等式成立是因为 $f(X, Y, Z)$ 有下界。因此

$$\lim_{k \rightarrow \infty} \tau_1 \|X^{k+1} - X^k\|_{\text{F}}^2 + \tau_2 \|Y^{k+1} - Y^k\|_{\text{F}}^2 + \tau_3 \|Z^{k+1} - Z^k\|_{\text{F}}^2 = 0, \quad (3.30)$$

于是

$$\lim_{k \rightarrow \infty} \|(X^{k+1}, Y^{k+1}, Z^{k+1}) - (X^k, Y^k, Z^k)\|_{\text{F}} = 0. \quad (3.31)$$

证毕。 \square

定理 3.1 设 $\{(X^k, Y^k, Z^k)\}$ 是算法 5 产生的序列，则 $\{(X^k, Y^k, Z^k)\}$ 的任何聚点 (X^*, Y^*, Z^*) 都是模型 (3.5) 的驻点（又称稳定点），即

$$0 \in \nabla f(X^*, Y^*, Z^*) + N_{\mathcal{M} \times \mathcal{S}_2 \times \mathcal{S}_1}(X^*, Y^*, Z^*).$$

如果满足 $\beta \geq \max\{2(\lambda_0 + \lambda_1), 2m\lambda_2\}$ ，可以进一步得到， $\{(X^k, Y^k, Z^k)\}$ 收敛到模型 (3.5) 的驻点。

证明：根据模型 (3.6) 的一阶最优化条件，得到

$$0 \in \nabla_X f(X^{k+1}, Y^{k+1}, Z^{k+1}) + N_{\mathcal{M}}(X^{k+1}) + 2\tau_1(X^{k+1} - X^k). \quad (3.32)$$

同样地，对于模型 (3.14) 和 (3.17)，有

$$\begin{aligned} 0 &\in \nabla_Y f(X^{k+1}, Y^{k+1}, Z^{k+1}) + N_{\mathcal{S}_2}(X^{k+1}) + 2\tau_2(Y^{k+1} - Y^k), \\ 0 &\in \nabla_Z f(X^{k+1}, Y^{k+1}, Z^{k+1}) + N_{\mathcal{S}_1}(X^{k+1}) + 2\tau_3(Z^{k+1} - Z^k). \end{aligned} \quad (3.33)$$

它们满足

$$\begin{aligned} A^{k+1} &= (A_X^{k+1}, A_Y^{k+1}, A_Z^{k+1}) \\ &\in \nabla f(X^{k+1}, Y^{k+1}, Z^{k+1}) + N_{\mathcal{M} \times \mathcal{S}_2 \times \mathcal{S}_1}(X^{k+1}, Y^{k+1}, Z^{k+1}). \end{aligned} \quad (3.34)$$

进而得到

$$\begin{aligned} A_X^{k+1} &\in \nabla_X f(X^{k+1}, Y^{k+1}, Z^{k+1}) + N_{\mathcal{M}}(X^{k+1}), \\ A_Y^{k+1} &\in \nabla_Y f(X^{k+1}, Y^{k+1}, Z^{k+1}) + N_{\mathcal{S}_2}(Y^{k+1}), \\ A_Z^{k+1} &\in \nabla_Z f(X^{k+1}, Y^{k+1}, Z^{k+1}) + N_{\mathcal{S}_1}(Z^{k+1}). \end{aligned} \quad (3.35)$$

使得

$$\begin{aligned} 0 &= A_X^{k+1} + 2\tau_1(X^{k+1} - X^k), \\ 0 &= A_Y^{k+1} + 2\tau_2(Y^{k+1} - Y^k), \\ 0 &= A_Z^{k+1} + 2\tau_3(Z^{k+1} - Z^k). \end{aligned} \quad (3.36)$$

因此，当取 $\tau = 2 \max\{\tau_1, \tau_2, \tau_3\}$ 时，有

$$\|(A_X^{k+1}, A_Y^{k+1}, A_Z^{k+1})\|_F \leq \tau \|(X^{k+1}, Y^{k+1}, Z^{k+1}) - (X^k, Y^k, Z^k)\|_F. \quad (3.37)$$

其次，约束集合 \mathcal{M} 、 \mathcal{S}_2 和 \mathcal{S}_1 以及它们的指示函数是半代数的，因此它们的二次函数 $f(X, Y, Z)$ 也是半代数的。利用半代数函数的复合函数仍然是半代数函数的性质，可以推导出

$$f(X, Y, Z) + \delta_{\mathcal{M}}(X) + \delta_{\mathcal{S}_2}(Y) + \delta_{\mathcal{S}_1}(Z) \quad (3.38)$$

也是半代数函数。这里， $\delta_{\mathcal{M}}(X)$ 表示集合 \mathcal{M} 上的指示函数，即当 $X \in \mathcal{M}$ 时， $\delta_{\mathcal{M}}(X) = 0$ ，否则 $\delta_{\mathcal{M}}(X) = \infty$ 。因此，它在每个点上都满足 Kurdyka-Łojasiewicz 性质。

最后，设 X^* 是模型 (3.7) 的驻点。根据文献^[86] 中定理 3.1，对于 $\beta \geq \max\{2(\lambda_0 + \lambda_1), 2m\lambda_2\}$ ，可以很容易验证 X^* 也是模型 (3.6) 的驻点。根据文献^[86] 中定义 2.1，对于 $U \in \mathbf{T}_{\mathcal{M}}(X^*)$ 可以推导出

$$\text{Tr}(U^\top \nabla_X f(X^*, Y^*, Z^*)) \geq 0, \quad X^{*\top} X^* = I_m. \quad (3.39)$$

这意味着

$$0 \in \nabla_X f(X^*, Y^*, Z^*) + \mathbf{N}_{\mathcal{M}}(X^*). \quad (3.40)$$

此外， $Y^* \in \mathcal{S}_2$ 是模型 (3.14) 的显示解，因此有

$$0 \in \nabla_Y f(X^*, Y^*, Z^*) + \mathbf{N}_{\mathcal{S}_2}(Y^*). \quad (3.41)$$

同样地，也可以得到

$$0 \in \nabla_Z f(X^*, Y^*, Z^*) + \mathbf{N}_{\mathcal{S}_1}(Z^*). \quad (3.42)$$

结合式(3.40)、(3.41)和(3.42)，有

$$0 \in \nabla f(X^*, Y^*, Z^*) + N_{\mathcal{M} \times \mathcal{S}_2 \times \mathcal{S}_1}(X^*, Y^*, Z^*). \quad (3.43)$$

根据上述分析，序列 $\{(X^k, Y^k, Z^k)\}$ 的任何聚点 (X^*, Y^*, Z^*) 都是模型(3.5)的驻点。根据文献^[90]，结合引理3.1、引理3.2、式(3.37)以及Kurdyka-Łojasiewicz性质得到最终结论，即算法5产生的序列 $\{(X^k, Y^k, Z^k)\}$ 收敛到模型(3.5)的驻点。证毕。□

定理3.1表明，由算法5产成的序列严格递减，这与文献^[22]中关于SPCAFS的结果类似。此外，在Kurdyka-Łojasiewicz性质的帮助下，本章还建立了全局收敛性。

3.3 数值实验与分析

为了验证DSCOFS的有效性，本节将与LapScore^[20]、UDFS^[24]、SOGFS^[27]、RNE^[26]、FSPCA^[43]、SPCAFS^[22]和SPCA-PSD^[37]进行比较。其中PCA-PSD^①则通过下载作者提供的代码实现，其余方法的实现与章节2.3保持一致。

3.3.1 实验设置

3.3.1.1 实验数据

在实验中，使用三个合成数据集和八个真实数据集。其中三个合成数据包括2Spiral^②，Banana^③和Dartboard^④数据集，八个真实数据集中除了章节2.3.1.1的六个，还增加了一个语音字母识别数据集，即Isolet^⑤，以及一个深度学习数据集，即MSTAR_SOC_CNN^⑥。为了方便书写，后续把数据集MSTAR_SOC_CNN简写为MSTARSC，有关这些数据集的详细信息如表3.1所示。

3.3.1.2 参数设置

关于参数，沿用了章节2.3.1.2的设置。对于DSCOFS，投影维度固定为数据的类别数，元素稀疏度参数设置为 $s_2 = cdm$ ，其中 c 是稀疏度百分比，表示保留元素个数的百分比，而 dm 是变换矩阵 X 中的元素总数。此外，稀疏度百分比 c 从 $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ 中选择。

^① <https://github.com/zjj20212035/SPCA-PSD>

^② <https://github.com/milaan9/Clustering-Datasets/>

^③ <https://jundongl.github.io/scikit-feature/datasets.html>

^④ <https://github.com/zjj20212035/SPCA-PSD>

表 3.1 数据集信息

Table 3.1 The dataset information

类型	数据集	特征数	样本数	类别数
合成数据	2Spiral	9	1000	2
	Banana	9	1000	2
	Dartboard	9	1000	4
真实数据	COIL20	1024	1440	20
	USPS	256	1000	10
	lungd	325	73	7
	GLIOMA	4434	50	4
	UMIST	644	575	20
	warpPIE10P	2420	210	10
	Isolet	617	1560	26
	MSTARSC	1024	2425	10

3.3.1.3 初始化和停止准则

对于 DSCOFS，初始变量 X^0 采用与章节 2.3.1.3 中相同的初始化策略，并设 $Z^0 = Y^0 = X^0$ 。此外，算法 4 的停止准则与精确罚函数方法^[86]中一致。算法 5 检查收敛性时满足

$$\frac{|f(X^{k+1}, Y^{k+1}, Z^{k+1}) - f(X^k, Y^k, Z^k)|}{1 + |f(X^k, Y^k, Z^k)|} \leq 10^{-3} \quad (3.44)$$

或最大迭代次数达到 100 时停止。

3.3.2 实验结果

3.3.2.1 合成数据集上的实验

最初的三个合成数据集只包含两个真实特征和许多数据点，所有数据点被分为若干类别并构成直观的图案。为了体现特征选择的能力，利用原始特征的均值和方差生成七个高斯噪声特征，并将原始特征放置在第 4 和第 5 个特征的位置。Banana 数据集从原始数据集的两个类别中分别随机选择 500 个样本组合而成。本次实验主要选择了基于 SPCA 的无监督特征选择方法，包括 FSPCA、SPCAF、SPCA-PSD 和 DSCOFS。值得注意的是，FSPCA 采用 $\ell_{2,0}$ 范数，SPCAF 采用 $\ell_{2,p}$ ($0 < p \leq 1$) 范数，SPCA-PSD 采用 $\ell_{2,1}$ 范数，而 DSCOFS 采用 $\ell_{2,0}$ 范数和 ℓ_0 范数构成的双稀疏约束。

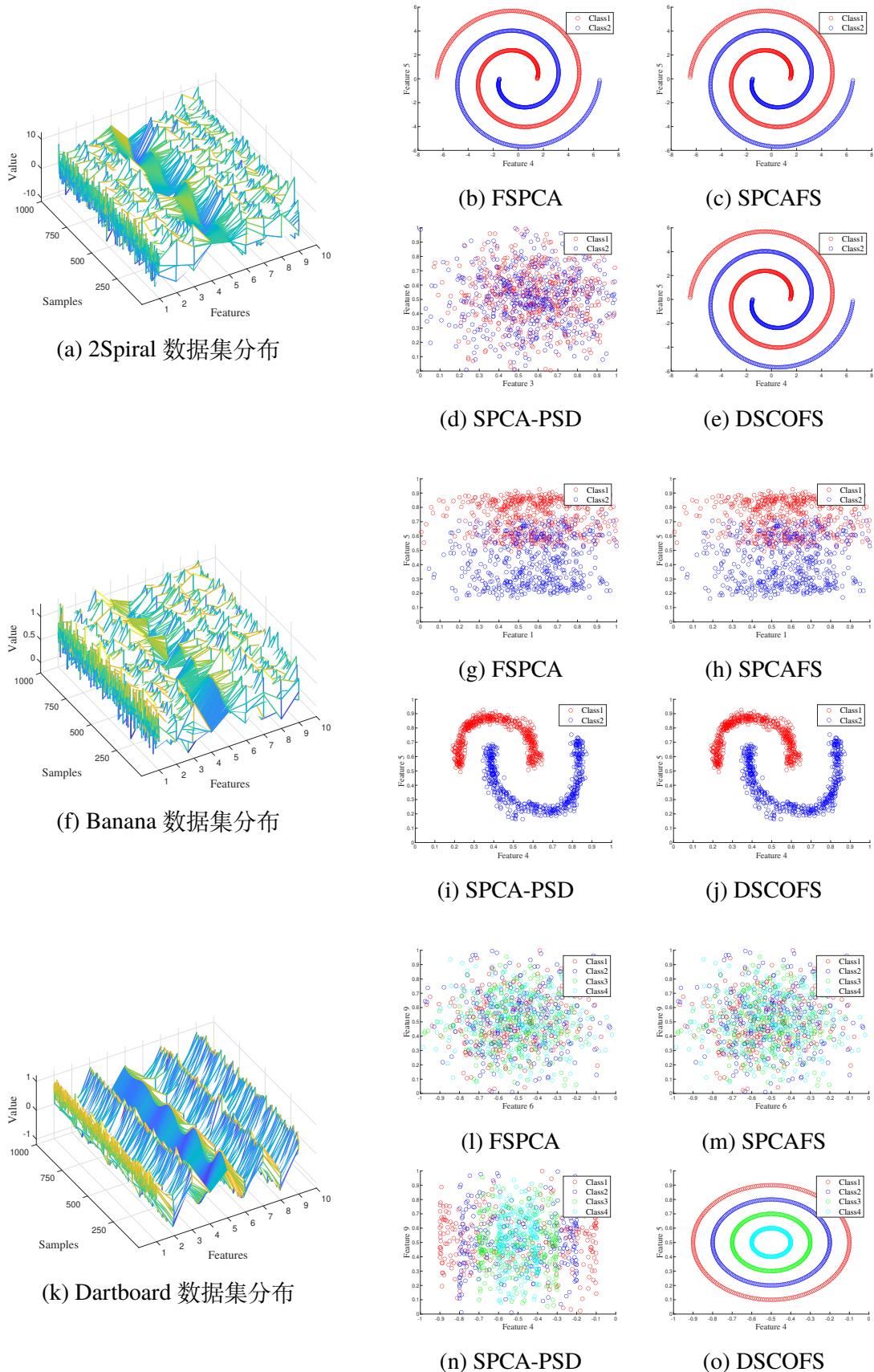


图 3.3 合成数据集的原始分布和特征选择结果

Figure 3.3 The original distribution and feature selection results on the synthetic datasets

执行所有对比的无监督特征选择方法，选取相应方法下排名前二的特征，利用两个特征作为坐标并绘制散点图。图 3.3 显示了三个合成数据集的特征选择可视化结果。图 3.3 中的 (a)、(f) 和 (k) 是所选的三个合成数据集经过加噪之后的数据分布，(b) - (e) 是对比方法在 2Spiral 数据集上的特征选择结果，(g) - (j) 是在 Banana 数据集上的特征选择结果，(l) - (o) 是在 Dartboard 数据集上的特征选择结果。与前面规定的一致，正确的特征是特征 4 和特征 5。从图中可以看出，DSCOFS 在所有三个数据集上都选择了正确的特征。对于 FSPCA 和 SPCAFS，只有在 2Spiral 数据集上选择了正确的特征，而 SPCA-PSD 仅在 Banana 数据集上选择了正确的特征。此外，除了 DSCOFS 方法外，其他方法在 Dartboard 数据集上都无法选择正确的特征。显然，不同的稀疏结构会导致特征选择结果的不同。相比之下，本章提出的 DSCOFS 同时考虑了全局结构稀疏和局部元素稀疏，在面对不同数据结构时依然有强大的特征辨识能力，这为无监督特征选择提供了更多可能性。

3.3.2.2 真实数据集上的实验

图 3.4 和图 3.5 展示了不同特征数量下 ACC 和 NMI 的平均曲线，其中作为参考基准的 ALLfea 表示使用所有特征进行聚类。表 3.2 和表 3.3 给出了在 100 个特征范围内最佳 ACC 和 NMI 的平均值、标准差和相应的所选特征数量。同时，最好和第二好的结果（除 ALLfea 外）分别用红色和蓝色标记。

从图 3.4 可以看出，本章提出的 DSCOFS 在所有数据集上都表现出卓越的性能，并且是唯一在所有数据集上都超过基线的方法。在 COIL20 和 Isolet 数据集上，DSCOFS 几乎在所有特征数下都位于所有方法的第一位且明显高于第二位。从表 3.2 可以观察到，对于 COIL20 数据集，UDFS 的表现超过了 FSPCA、SPCAF 和 SPCA-PSD，但仍低于 DSCOFS。对于 Isolet 数据集，DSCOFS 比 FSPCA、SPCAF 和 SPCA-PSD 分别提高了 6.05%、6.63% 和 7.76%。在八个数据集上的平均 ACC 结果，DSCOFS 相比其他无监督特征选择方法提升了至少 3.34%。

从图 3.5 可以观察到与图 3.4 类似的结果，DSCOFS 在所有数据集上都有不错的效果。与其他对比方法相比，DSCOFS 在除了 GLIOMA 的所有数据集上都取得了最佳结果。从表 3.3 也可以直观地看到，在 COIL20 和 Isolet 数据集上，NMI 结果有较明显的提升，这和前面 ACC 的结果一致。具体地，DSCOFS 在 COIL20 和 Isolet 数据集上分别提高了 2.71% 和 4.22%。值得注意的是，FSPCA、SPCAF、SPCA-PSD

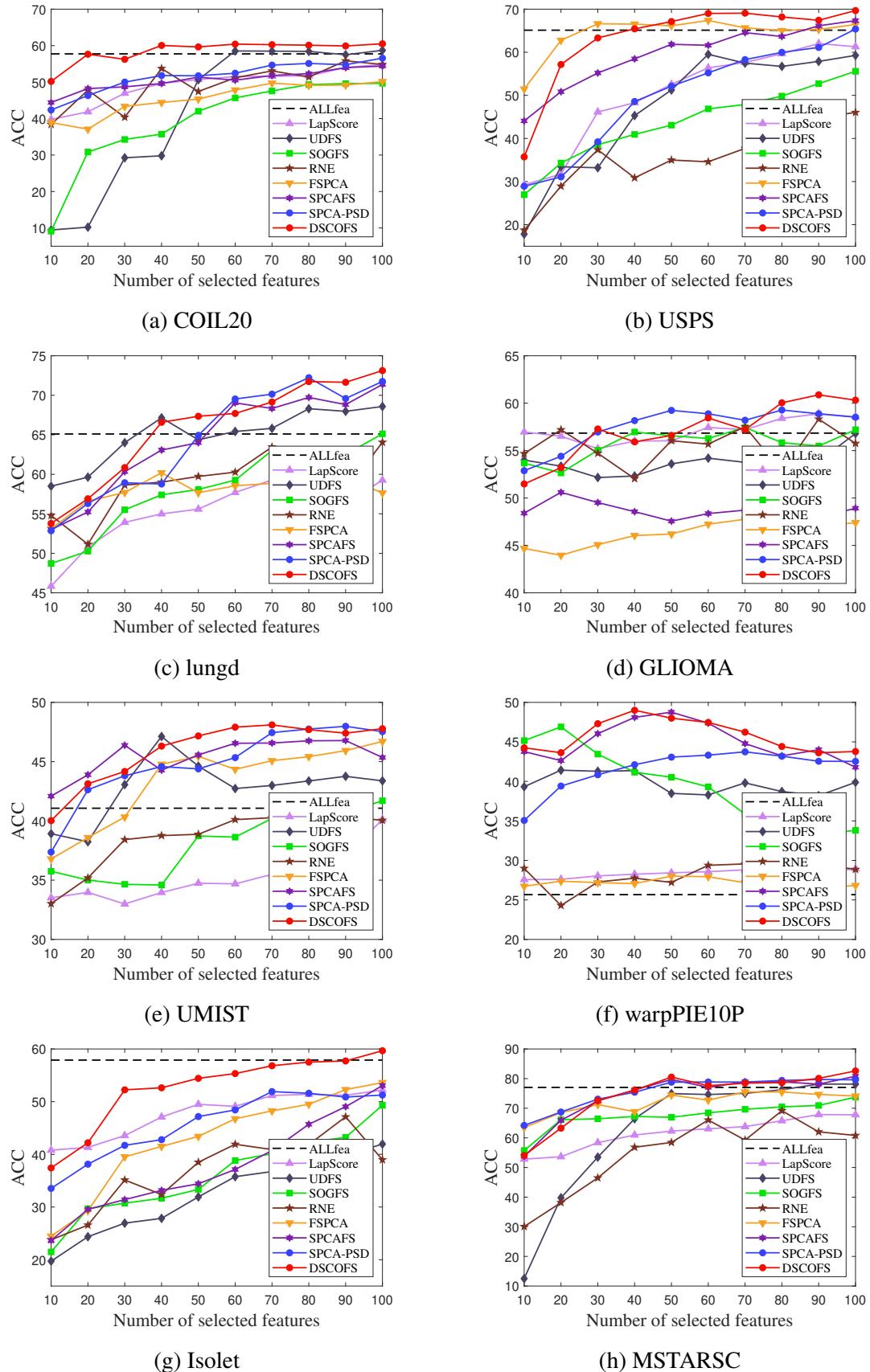


图 3.4 对比方法在八个真实数据集上的 ACC (%) 曲线

Figure 3.4 The ACC (%) curves of compared methods on eight real-world datasets

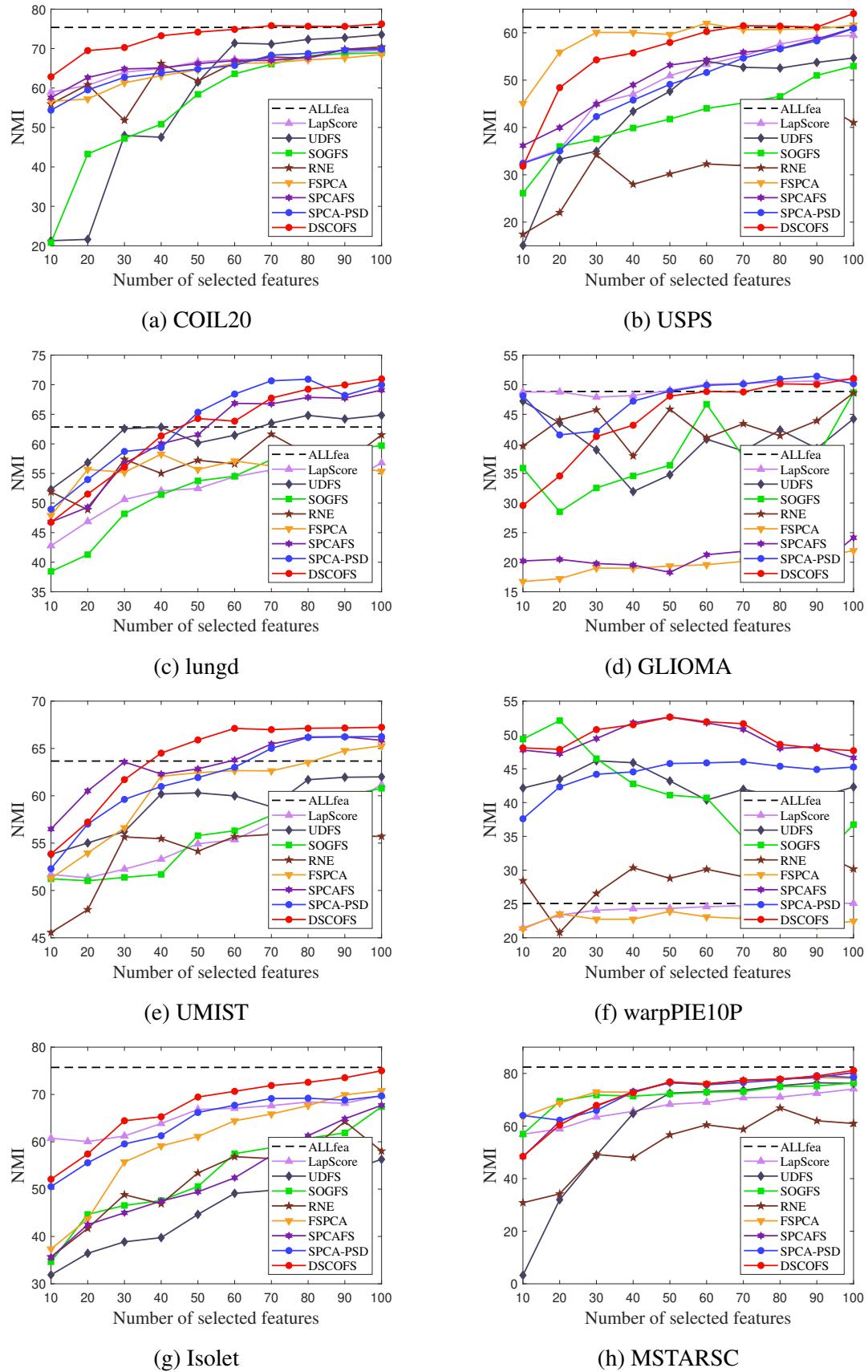


图 3.5 对比方法在八个真实数据集上的 NMI (%) 曲线

Figure 3.5 The NMI (%) curves of compared methods on eight real-world datasets

表 3.2 对比方法在八个真实数据集上的 ACC (平均值% ± std%) 结果

Table 3.2 The ACC (mean% ± std%) results of compared methods on eight real-world datasets

数据集	ALLfea	LapScore	UDFS	SOGFS	RNE	FSPCA	SPCAFS	SPCA-PSD	DSCOFS
COIL20	57.74±4.93	54.82±3.91 (100)	58.71±3.47 (100)	49.66±4.81 (100)	55.84±4.41 (90)	50.15±4.70 (100)	54.39±3.67 (100)	56.57±3.05 (100)	60.51±4.63 (100)
USPS	65.12±4.95	62.02±4.09 (90)	59.52±2.97 (60)	55.58±3.07 (100)	46.04±2.69 (100)	67.38±4.36 (60)	67.34±4.49 (100)	65.38±4.26 (100)	69.67±4.97 (100)
lungd	65.10±6.44	59.29±6.33 (70)	68.58±6.99 (100)	65.12±6.89 (100)	64.05±6.65 (100)	60.19±6.55 (40)	71.37±7.68 (100)	72.22±8.02 (80)	73.12±8.48 (100)
GLIOMA	56.84±5.24	58.88±3.96 (90)	56.80±4.85 (100)	57.44±6.16 (70)	58.32±7.31 (90)	47.92±4.61 (80)	50.60±5.02 (20)	59.28±5.01 (90)	60.88±6.31 (80)
UMIST	41.07±2.38	40.13±2.79 (100)	47.12±2.49 (40)	41.70±3.17 (100)	40.35±2.26 (90)	46.70±2.29 (100)	46.78±2.51 (90)	47.98±2.91 (90)	48.10±3.01 (70)
warpPIE10P	25.67±1.90	28.94±1.66 (100)	41.42±3.18 (20)	46.90±3.89 (20)	29.57±2.96 (90)	28.01±2.27 (50)	48.76±3.86 (50)	43.74±3.91 (70)	49.00±3.88 (40)
Isolet	57.89±3.82	52.21±2.76 (100)	41.95±2.07 (100)	49.31±2.32 (100)	47.12±2.06 (90)	53.62±2.36 (100)	53.04±2.33 (100)	51.91±2.15 (70)	59.67±3.46 (100)
MSTARSC	77.04±7.98	67.87±3.49 (90)	78.15±5.80 (90)	73.74±5.89 (100)	69.16±6.03 (100)	75.52±6.22 (70)	80.80±5.95 (100)	79.70±6.43 (90)	82.59±7.41 (100)
Average	55.81±4.71	53.02±3.62	56.53±4.04	54.93±4.53	51.31±4.30	53.69±4.47	59.14±4.44	59.60±4.47	62.94±5.27

表 3.3 对比方法在八个真实数据集上的 NMI (平均值% ± 标准差%) 结果

Table 3.3 The NMI (mean% ± std%) results of compared methods on eight real-world datasets

数据集	ALLfea	LapScore	UDFS	SOGFS	RNE	FSPCA	SPCAFS	SPCA-PSD	DSCOFS
COIL20	75.37±1.96	69.59±1.48 (100)	73.54±1.76 (100)	68.92±1.84 (100)	70.43±1.92 (100)	68.50±1.56 (100)	69.98±1.45 (100)	69.85±1.41 (100)	76.25±1.71 (100)
USPS	61.12±2.01	59.46±1.80 (100)	54.69±2.11 (100)	52.96±1.54 (100)	45.36±1.93 (90)	62.00±1.87 (60)	60.98±2.37 (100)	60.90±2.02 (100)	64.06±2.58 (100)
lungd	62.85±5.13	56.79±3.99 (100)	64.84±5.09 (100)	59.70±5.24 (100)	61.63±5.83 (70)	58.26±6.39 (40)	69.09±5.61 (100)	70.93±5.46 (80)	70.98±7.00 (100)
GLIOMA	48.86±5.72	51.03±2.48 (100)	47.22±3.53 (10)	48.67±10.98 (100)	48.62±6.32 (100)	21.94±5.28 (100)	24.14±6.97 (100)	51.44±5.62 (90)	51.06±6.19 (80)
UMIST	63.67±1.85	61.16±1.71 (100)	62.00±1.58 (100)	60.79±1.54 (100)	55.92±1.57 (70)	65.27±1.58 (100)	66.23±1.60 (90)	66.25±1.72 (100)	67.24±1.85 (100)
warpPIE10P	25.07±2.88	25.13±1.73 (90)	46.18±3.30 (20)	52.12±3.25 (20)	32.67±3.31 (90)	23.90±2.01 (50)	52.63±3.33 (50)	46.02±3.70 (70)	52.65±3.29 (50)
Isolet	75.72±1.70	69.77±1.20 (100)	56.29±1.11 (100)	67.40±1.44 (100)	64.27±0.95 (90)	70.79±1.12 (100)	67.71±1.33 (100)	69.69±0.80 (100)	75.01±1.35 (100)
MSTARSC	82.42±3.31	74.10±1.76 (100)	76.45±2.47 (90)	76.39±1.70 (100)	66.87±1.99 (80)	78.39±2.17 (90)	80.33±2.50 (100)	79.17±2.77 (90)	81.14±3.13 (100)
Average	61.89±3.07	58.38±2.02	60.15±2.62	60.87±3.44	55.72±2.98	56.13±2.75	61.39±3.15	64.28±2.94	67.30±3.39

和 DSCOFS 在除了 COIL20 的所有数据集上获得了最好和第二好的结果，这表明了 SPCA 在特征选择中具有较好的前景。在八个数据集上的平均 NMI 结果，DSCOFS 相比其他无监督特征选择方法提升了至少 3.02%。

由此可以得出结论，DSCOFS 中的双稀疏约束能够处理比单稀疏约束更复杂的稀疏结构，增强了局部特征的辨别能力，从而提高了特征选择的性能。

3.3.3 消融实验

虽然实验结果验证了 DSCOFS 的有效性，但无法直观地观察到加入元素稀疏的作用。下面将通过消融实验分析双稀疏的效果。

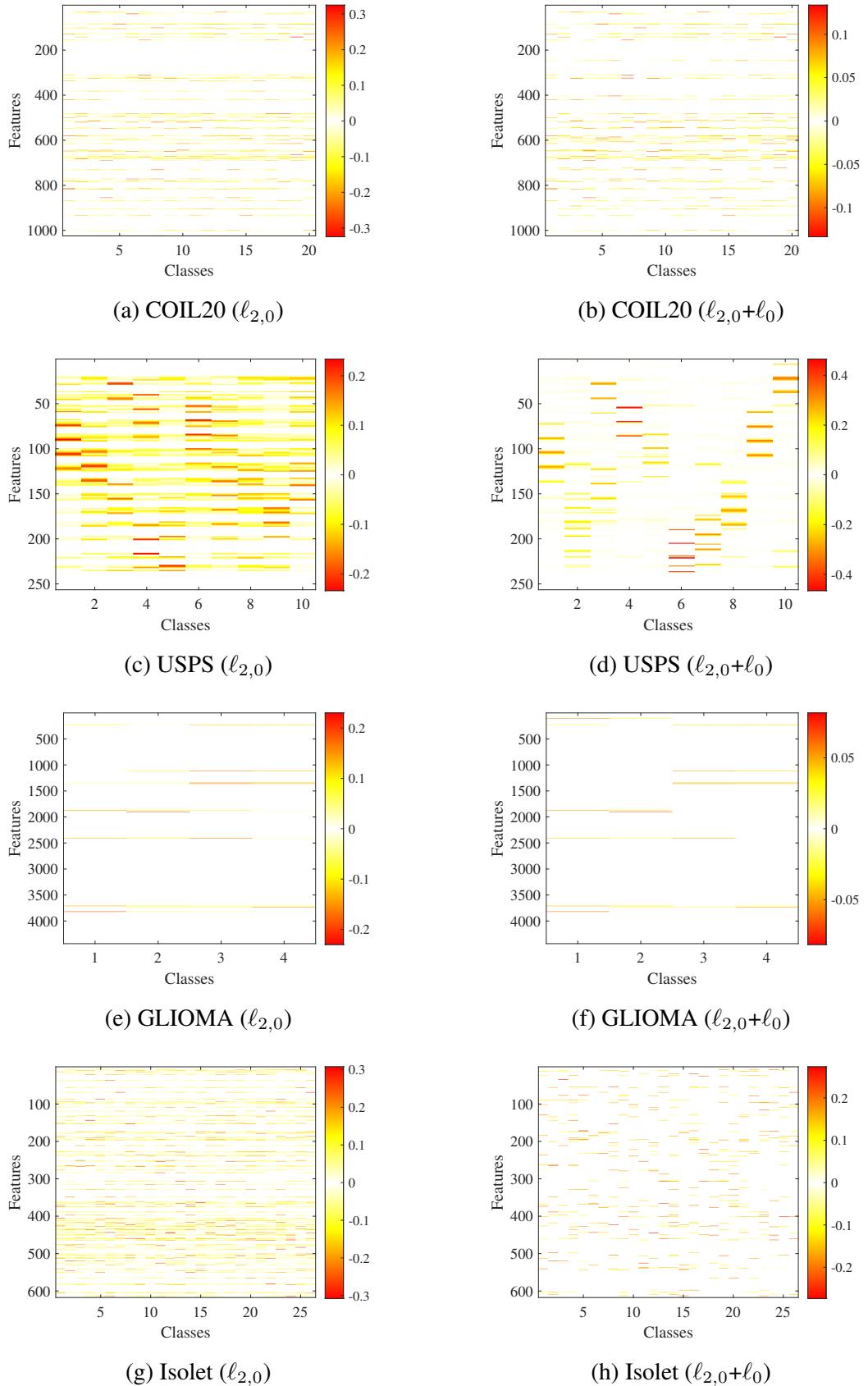
表 3.4 在八个真实数据集上消融实验的 ACC(%)、NMI(%) 和 FSR(%)

Table 3.4 The ACC (%), NMI (%) and FSR (%) of ablation study on eight real-world datasets

类型	$\ X\ _0 \leq s$	ACC	NMI	FSR
COIL20	✗	60.25 ± 4.52	75.89 ± 1.58	84
	✓	60.51 ± 4.42	76.25 ± 1.71	
USPS	✗	67.84 ± 3.71	60.90 ± 1.95	68
	✓	69.67 ± 4.97	64.06 ± 2.58	
lungd	✗	71.42 ± 7.95	69.74 ± 6.11	92
	✓	73.12 ± 8.48	70.98 ± 7.00	
GLIOMA	✗	58.24 ± 5.04	49.76 ± 6.12	85
	✓	60.88 ± 6.31	51.06 ± 6.19	
UMIST	✗	47.33 ± 3.05	67.44 ± 1.88	95
	✓	48.10 ± 3.01	67.24 ± 1.85	
warpPIE10P	✗	47.91 ± 4.99	51.19 ± 3.79	89
	✓	49.00 ± 3.88	52.65 ± 3.29	
Isolet	✗	57.29 ± 3.44	72.82 ± 1.87	52
	✓	59.67 ± 3.46	75.01 ± 1.35	
MSTARSC	✗	82.06 ± 6.87	81.01 ± 2.41	99
	✓	82.59 ± 7.41	81.14 ± 3.13	

首先，本章引入了一种新的评估指标，称为特征相似率（Feature Similarity Rate, FSR）。设 $\mathbb{T}_{\text{DSCOFS}}$ 和 $\mathbb{T}_{2,0}$ 分别表示 DSCOFS 模型和仅使用 $\ell_{2,0}$ 范数模型选择出的特征集合。FSR 定义为

$$\text{FSR} = \frac{1}{n} \text{card}(\mathbb{T}_{\text{DSCOFS}} \cap \mathbb{T}_{2,0}), \quad (3.45)$$

图 3.6 稀疏投影矩阵 X 在四个真实数据集上的可视化Figure 3.6 The sparse visualization of the projection matrix X on four real-world datasets

其中, n 是计算 FSR 时选择的特征数量, $\text{card}(\cdot)$ 表示集合中的个数。从式 (3.45) 可以得知, FSR 本质上表示两个特征选择结果集合之间重叠特征的百分比。因此 FSR 越小, 两个特征集差异越大。表 3.4 记录了在 100 个特征内最佳 ACC 和相应的 NMI, 其中 \times 表示仅考虑 $\ell_{2,0}$ 范数约束, 而 \checkmark 表示本章提出的双稀疏约束。

从表 3.4 可以看出, 本章提出的 DSCOFS 通过引入元素稀疏后在 ACC 和 NMI 都有一定程度的提升。此外, 可以观察到 FSR 在不同的数据集上变化较大。特别是在 USPS 和 Isolet 数据集上, FSR 值分别为 68% 和 52%, 这表明添加元素稀疏选择的特征与仅使用 $\ell_{2,0}$ 范数选择的特征存在一定差异。进一步, 图 3.6 可视化了 COIL20、USPS、GLIOMA 和 Isolet 四个 FSR 较小的真实数据集的投影矩阵。从图中可以清晰地观察到, 相比于仅使用 $\ell_{2,0}$ 范数, 使用双稀疏得到了更加稀疏的变换矩阵, 并且保留了不同的特征。尤其在 USPS 和 Isolet 数据集上差距明显, 这也直观地解释了为什么 DSCOFS 在这两个数据集上有较明显的提升。

3.3.4 讨论

3.3.4.1 统计检验

Friedman 检验是一种基于排名的统计方法, 常用于比较多种方法的平均性能是否存在显著差异。以 ACC 指标为例, 对每种方法在八个数据集上的 ACC 进行排名, 从最好到最差赋予排名值 1 到 8, 并最终取所有数据集上的平均排名。当多种方法在同一数据集上的准确度相同时, 采用平均排名值, 如排名 4 和 5 的方法一样就都赋予 4.5。在本实验中, Friedman 检验的原假设 H_0 表示所有对比方法的性能没有显著差异。在显著性水平设定为 $\alpha = 0.05$ 的情况下, 从表 3.5 中可以看到 $p = 0.0001$, 这意味着结果拒绝原假设 H_0 。因此, Friedman 检验表明所有对比方法之间确实存在显著差异。

遗憾的是, Friedman 检验是对整体差异性的检验, 无法确定其中两种方法之间是否存在显著差异。为了进一步探究不同方法的差异性, 下面进行后验 Nemenyi 检验。它可以通过临界差异 (Critical Difference, CD) 值来衡量差异, 当两种方法在同一个 CD 值内意味着这两种方法没有显著的差异性。

图 3.7 显示了后验 Nemenyi 检验的相应结果。本章提出的 DSCOFS 与 LapScore、UDFS、SOGFS、RNE 和 FSPCA 有显著差异, 但与 SPCA-PSD 和 SPCAFS 没有显著差异。而 SPCA-PSD 和 SPCAFS 与 LapScore、UDFS、SOGFS、RNE 和 FSPCA 没有

表 3.5 ACC 指标下 DSCOFS 的 Friedman 检验结果

Table 3.5 The Friedman test results of DSCOFS in terms of ACC

方法	平均排名	<i>p</i> 值	假设
LapScore	6.000		
UDFS	4.750		
SOGFS	5.750		
RNE	6.125		
FSPCA	5.500	0.0001	拒绝
SPCAFS	3.750		
SPCA-PSD	3.125		
DSCOFS	1.000		

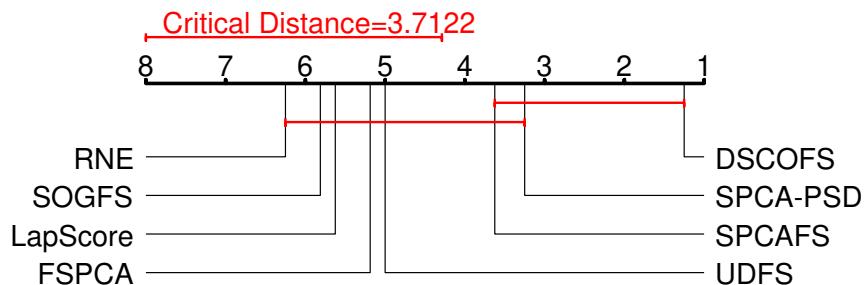


图 3.7 ACC 指标下 DSCOFS 的后验 Nemenyi 检验结果

Figure 3.7 The post-hoc Nemenyi test results of DSCOFS in terms of ACC

显著差异。值得注意的是，DSCOFS 与具有 $\ell_{2,0}$ 范数约束的 FSPCA 之间存在显著差异，这也表明加入了 ℓ_0 范数双稀疏约束的优势。

3.3.4.2 参数敏感度分析

对于 DSCOFS 而言, 参数 s_1 和 $s_2 = cdm$ 分别控制着两个稀疏度, 是影响无监督特征选择性能的关键。下面选取 s_1 和 c 作为敏感度分析的参数, 结果如图 3.8 所示, 其中 c 表示稀疏度的百分比。可以看出, ACC 和 NMI 会受到稀疏度百分比 c 的影响而波动。但是, 大部分变化集中在高稀疏性, 即低稀疏度百分比 $c = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ 范围内。对于 USPS 数据集来说, 当 $c = 0.1$ 时, 性能有明显地提高, 这也表明了元素稀疏可以增强模型的特征选择能力。

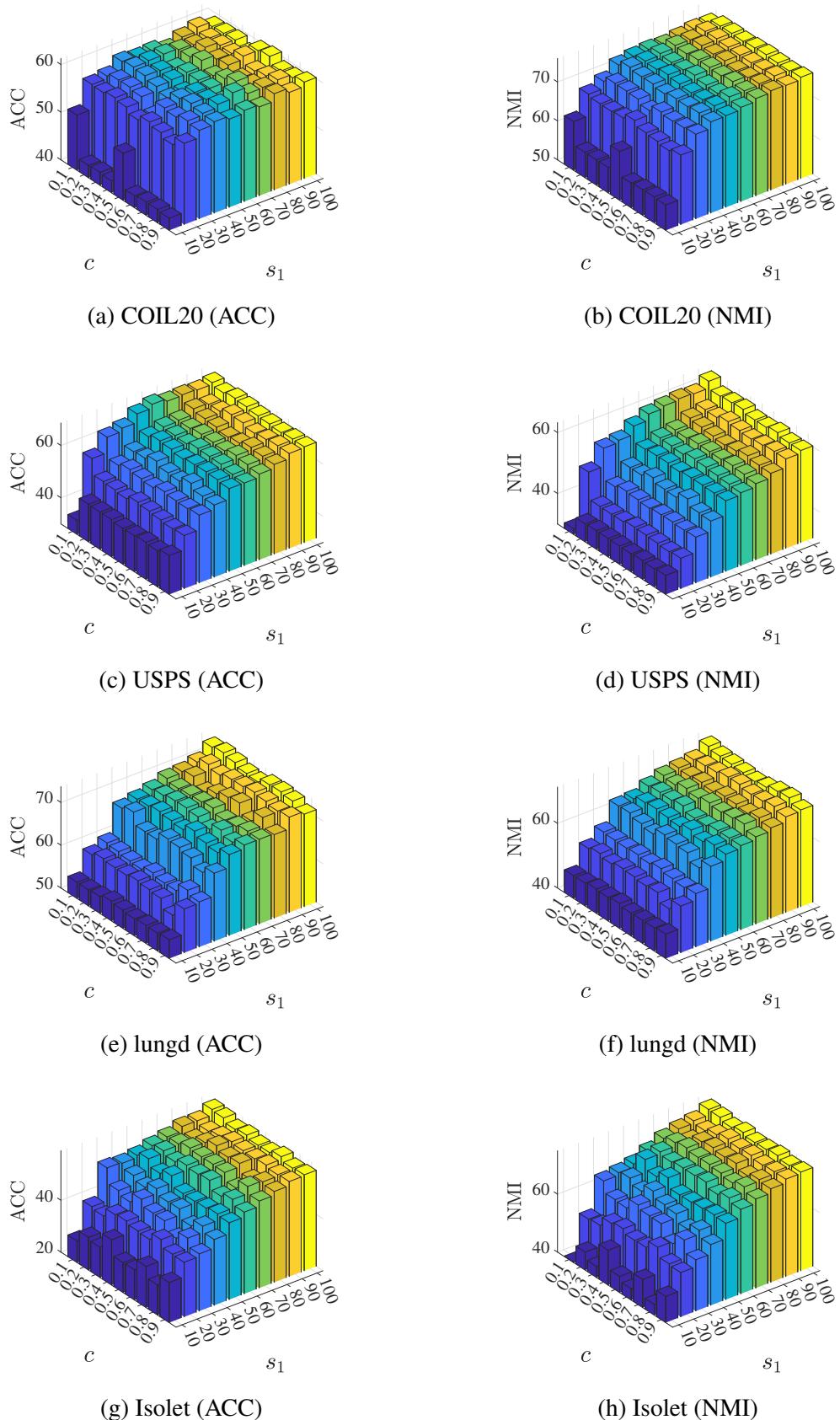


图 3.8 DSCOFS 在四个真实数据集上的参数敏感度分析结果

Figure 3.8 The parameter sensitivity analysis results of DSCOFS on four real-world datasets

3.3.4.3 模型收敛性分析

图 3.9 显示了在执行特征选择时，DSCOFS 的目标函数值随迭代次数的变化曲线。结果表明，DSCOFS 在大多数情况下能够持续下降，并在 100 次迭代内达到稳定状态。这与定理 3.1 中给出的结论一致，验证了算法的有效性。

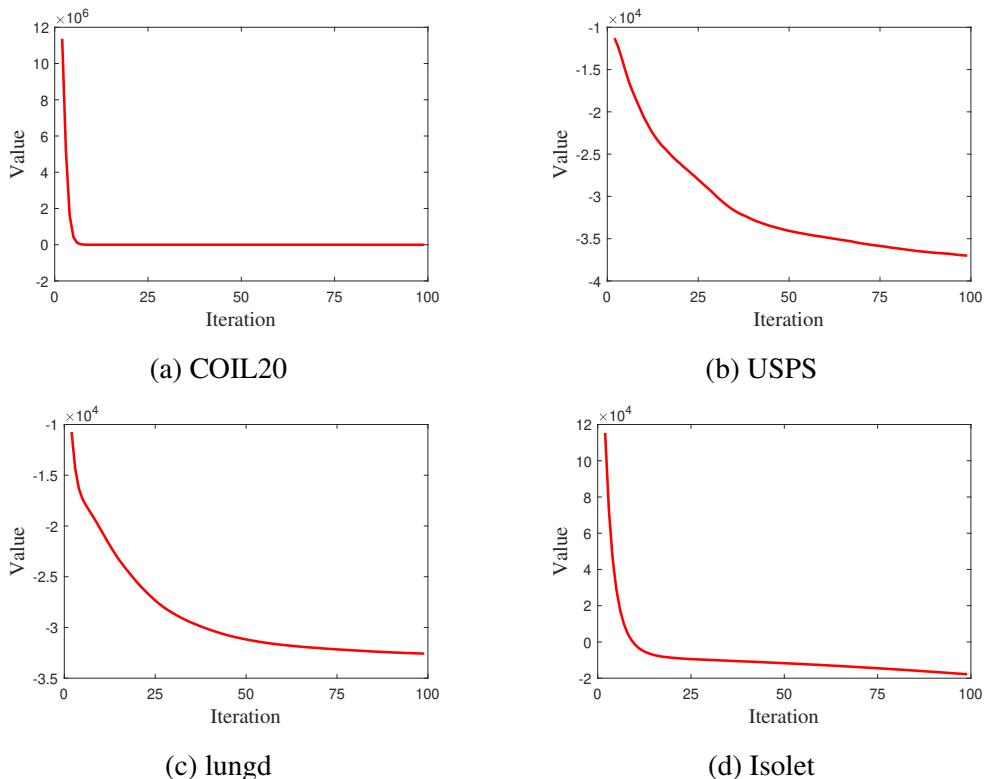


图 3.9 DSCOFS 在四个真实数据集上的收敛曲线

Figure 3.9 The convergence curves of DSCOFS on four real-world datasets

3.3.4.4 模型稳定性分析

为了探究模型的稳定性，本实验记录了最好的聚类均值结果中 50 次的值，如图 3.10 所示。图中将 50 次聚类结果排序，去除离群值（用红色加号表示）后，分别在排名百分之 0%，25%，50%，75% 和 100% 处绘制横线。将排名 25% 至 75% 的数据绘制为“箱体”，排名 50% 的数据作为中位线。可以观察到，DSCOFS 聚类结果有一定的波动，但整体结果优于其他对比方法。特别是在 Isolet 数据集上，DSCOFS 最低值已经超过或接近其他方法的中位数值。尽管 DSCOFS 在 lungd 数据集上波动较大，但平均结果仍略优于其他方法。因此，这些结果说明了 DSCOFS 具有不错的稳定性。

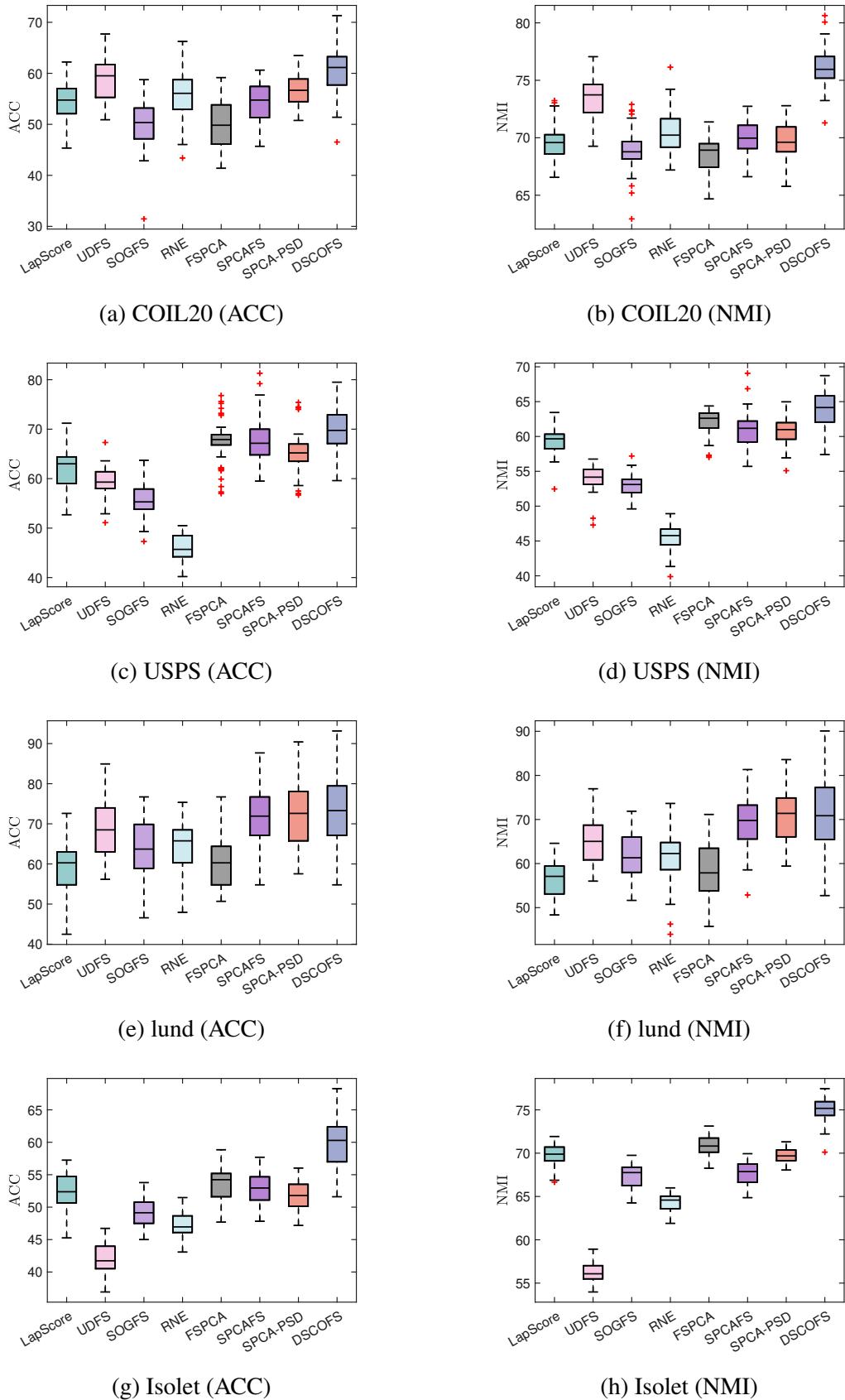


图 3.10 对比方法在四个真实数据集上的模型稳定性

Figure 3.10 The model stability of compared methods on four real-world datasets

3.3.4.5 与深度学习的对比分析

本实验将选取了一个基于深度学习的无监督特征选择方法（即 TSFS+TSNE^[91]）进行简单的讨论，如表 3.6 所示。可以看到，与 TSFS+TSNE 相比，本章提出的 DSCOFS 表现优越。在 USPS 和 lungd 数据集上，DSCOFS 的表现优于 TSFS+TSNE，其中 ACC 指标分别提高了 7.71% 和 8.76%。这可能是因为像 TSFS+TSNE 这类基于深度学习的无监督特征选择方法需要大量的样本数据集进行训练，而 DSCOFS 由于其增强的可解释性和泛化能力，在较小的数据集上也可以有不错的表现。

表 3.6 DSCOFS 和 TSFS+TSNE 在四个真实数据集上的 ACC (%) 和 NMI (%)

Table 3.6 The ACC (%) and NMI (%) of DSCOFS and TSFS+TSNE on four real-world datasets

方法	方法	ACC	NMI
COIL20	TSFS+TSNE	60.80±3.83	71.59±1.46
	DSCOFS	60.51±4.42	76.25±1.71
USPS	TSFS+TSNE	61.96±3.96	56.20±1.20
	DSCOFS	69.67±4.97	64.06±2.58
lung_discrete	TSFS+TSNE	64.36±7.24	61.61±5.70
	DSCOFS	73.12±8.48	70.98±7.00
Isolet	TSFS+TSNE	60.40±4.34	76.13±1.54
	DSCOFS	59.67±3.46	75.01±1.35

3.4 本章小结

本章通过巧妙地结合 $\ell_{2,0}$ 范数和 ℓ_0 范数约束，提出了基于双稀疏约束的 DSCOFS 模型。利用双稀疏特性，DSCOFS 能够学习更多的稀疏结构，从而过滤掉冗余和不相关的特征，达到更准确和有效的特征辨别。在算法方面，将精确罚函数方法和硬阈值引入到 PAM 框架中，设计了一种有效的优化算法。在理论方面，严格证明了算法的收敛性，即序列全局收敛到驻点。最后，大量的数值实验证证了 DSCOFS 的优越性和局部特征辨别的准确性。具体地，DSCOFS 在八个真实数据集上的平均 ACC 和 NMI 相较其他无监督特征选择方法分别至少提升了 3.34% 和 3.02%。此外，本章还引入了一种新的特征差异性度量，直观地分析了元素稀疏如何增强特征选择的能力并成功弥补单稀疏的局限性。

第四章 基于对比学习和双稀疏约束的无监督特征选择方法

上一章阐述了双稀疏约束优化 DSCOFS 模型的优势，然而该模型采用了欧式距离来度量样本的邻接关系，缺乏对特征间相关性的建模能力，无法有效捕捉潜在的结构信息。因此，本章在 DSCOFS 的基础上，提出了一种融合自表示学习和对比学习的无监督特征选择模型，即 DSCOFS-CL。具体而言，通过在原始空间和投影空间联合学习自表示矩阵，同时引入对比学习损失充分考虑样本之间的关系，进而使构建的最优图可以很好地刻画空间结构。此外，对自表示矩阵施加低秩约束来代替传统的稀疏约束，可以保留图的全局结构。在算法方面，设计了基于梯度下降和硬阈值的优化策略，并在理论上证明了该算法产生序列的收敛性。最后，大量的数值实验验证了对比学习的有效性和所提出 DSCOFS-CL 的优越性。

4.1 相关工作

除了第二章提到了最大相关性描述，PCA 还有另外一种描述。具体地，寻找一个低维的变换矩阵 X ，使得数据 A 可以通过重构数据 $XX^\top A$ 近似并使得误差最小。因此，最小化重构误差的 PCA 模型为

$$\begin{aligned} \min_X \quad & \|A - XX^\top A\|_F^2 \\ \text{s.t.} \quad & X^\top X = I_m. \end{aligned} \tag{4.1}$$

最小化重构误差模型通常采用 Frobenius 范数作为度量标准，而该范数隐含地假设各特征之间具有相同的尺度和重要性。但是在实际数据中，特征之间的尺度和分布往往存在差异性。这意味着，即使仅有少数噪声和离群值，其平方差也可能显著增加，甚至可能导致误差评估失真。虽然 SPCA 一定程度缓解了数据噪声和冗余的情况，但是使用何种范数度量重构误差仍是一个值得深入研究的问题。

为了提高稳定性，Ke 等人^[92]采用 ℓ_1 范数度量重构误差，其数学模型为

$$\begin{aligned} \min_X \quad & \|A - XX^\top A\|_1 \\ \text{s.t.} \quad & X^\top X = I_m. \end{aligned} \tag{4.2}$$

与 Frobenius 范数不同， ℓ_1 范数刻画的是所有元素的绝对值之和，因此对少数噪声和离群值的影响相对较小。当数据中存在较多不相关或冗余特征时， ℓ_1 范数倾向于找到较少的非零元素。

值得注意的是， ℓ_1 范数不具有旋转不变性^[93]，且依赖数据本身结构的方向性。此外， ℓ_1 范数不能正确计算重构数据和原始数据之间的欧氏距离。为了解决上述问题，Wang 等人^[94]使用 $\ell_{2,p}$ ($0 < p < 2$) 范数度量重构误差，其数学模型为

$$\begin{aligned} \min_X \quad & \|A - XX^\top A\|_{2,p}^p \\ \text{s.t.} \quad & X^\top X = I_m. \end{aligned} \tag{4.3}$$

这里， $\ell_{2,p}$ 范数与 Frobenius 范数相比，降低了少数噪声和离群值的影响。因此， $\ell_{2,p}$ 范数在保留原本旋转不变性的同时又具有一定的鲁棒性。在实际的应用中， p 常取值为 $(0, 1)$ ，而当 p 取一个非常小的值时， $\ell_{2,p}$ 范数将会失去区分正确样本的能力^[38]。

上述模型没有很好地考虑样本之间的关系，这使得判别性能较差，难以满足下游任务的需要。随着机器学习和深度学习技术的发展，基于对比学习^[95]的判别方法得到了广泛的关注。对比学习通过最大化正对之间的相似性和负对之间的距离提高判别性^[96]，一方面有效地缩小相似样本表示之间的距离，另一方面又将不同类的样本分开，进而学习到数据结构中存在的潜在类别信息。

给定数据 $A = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n] \in \mathbb{R}^{d \times n}$ 和重构数据 $\hat{A} = [\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \dots, \hat{\mathbf{a}}_n] \in \mathbb{R}^{d \times n}$ 。受到负样本共享的采样策略^[97]的启发，给定一个样本 \mathbf{a}_i ，其相应的正样本为 $\hat{\mathbf{a}}_i$ ，而剩余的 $2(n - 1)$ 个均视为负样本。为了度量对比学习损失误差，可以使用归一化温度缩放交叉熵损失 (NT-Xent)^[98]来实现最大化正对之间的相似性和最小化负对之间的相似性。具体地，对于 \mathbf{a}_i ，其交叉熵损失为

$$L_c(\mathbf{a}_i) = -\log \frac{\exp(s(\mathbf{a}_i, \hat{\mathbf{a}}_i)/\tau)}{\sum_{j=1, j \neq i}^n \exp(s(\mathbf{a}_i, \mathbf{a}_j)/\tau) + \sum_{j=1}^n \exp(s(\mathbf{a}_i, \hat{\mathbf{a}}_j)/\tau)}, \tag{4.4}$$

其中 τ 是用来控制相似度计算的温度参数， $s(\mathbf{a}_i, \hat{\mathbf{a}}_j)$ 为相似度度量，这里取 $s(\mathbf{a}_i, \hat{\mathbf{a}}_j) = \mathbf{a}_i^\top \hat{\mathbf{a}}_j$ 。同样地， $\hat{\mathbf{a}}_i$ 的交叉熵损失为

$$L_c(\hat{\mathbf{a}}_i) = -\log \frac{\exp(s(\hat{\mathbf{a}}_i, \mathbf{a}_i)/\tau)}{\sum_{j=1, j \neq i}^n \exp(s(\hat{\mathbf{a}}_i, \hat{\mathbf{a}}_j)/\tau) + \sum_{j=1}^n \exp(s(\hat{\mathbf{a}}_i, \mathbf{a}_j)/\tau)}. \tag{4.5}$$

对于交叉熵损失 (4.4) 和 (4.5) 而言，分母越大表示样本的相似度越高，相反地，分母越小表示样本的相似度越低，即负样本对的相似度越低。最终可以得到对比学习的

损失函数

$$L_c(A, \hat{A}) = \frac{1}{2n} \sum_{i=1}^n (L_c(\mathbf{a}_i) + L_c(\hat{\mathbf{a}}_i)). \quad (4.6)$$

图 4.1 展示了分别利用欧式距离度量和对比学习损失得到的投影空间结果。从图中可以看出，利用欧式距离评估相似度可能会导致不同类别样本之间的相似度较高，而同类别的相似度较低，这使得投影空间中不同类别样本之间的划分不清晰。相比之下，采用对比学习损失设定正负样本对后，投影空间中的同类样本因为包含了大量相似信息而更加紧凑，且不同类别样本之间的边界更加明显，从而提升了判别性能。

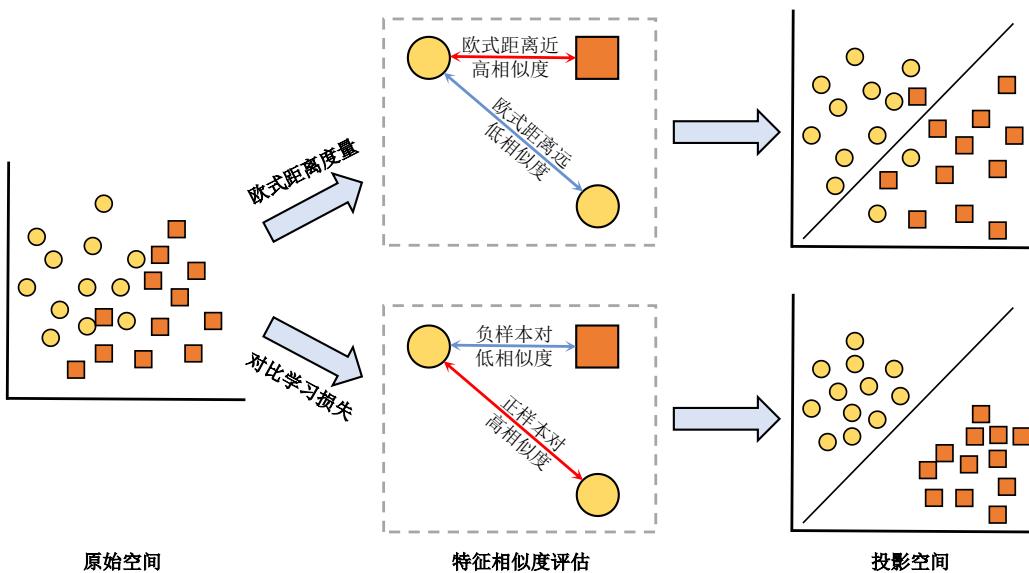


图 4.1 不同度量得到的投影空间

Figure 4.1 The projection spaces obtained from different metrics

基于上述分析，Qian 等人^[38]结合 PCA 和对比学习，构建了如下无监督特征选择模型

$$\begin{aligned} \min_X \quad & L_c(A, XX^\top A) + \alpha \|X\|_{1,2}^2 + \beta \|X^\top A\|_{1,2}^2 \\ \text{s.t.} \quad & X^\top X = I_m, \end{aligned} \quad (4.7)$$

其中 $\alpha, \beta > 0$ 是正则化参数。这里， $\|X\|_{1,2} = \sqrt{\sum_{j=1}^m (\sum_{i=1}^d |X_{ij}|)^2}$ ，用于实现自表示矩阵 Z 的列稀疏。实验表明，使用对比学习损失度量重构误差可以挖掘数据的有效信息，实现更好的无监督特征选择性能。

考虑到削弱原始空间噪声以及在投影空间保持相似结构的重要性, Qian 等人^[32]引入了自表示框架, 进一步建立了如下模型

$$\begin{aligned} \min_{X,Z} \quad & L_c(A, AZ) + \lambda L_c(X^\top A, X^\top AZ) + \alpha \|X\|_{1,2}^2 + \beta \|Z\|_{1,2}^2 \\ \text{s.t.} \quad & X^\top X = I_m, \text{Diag}(Z) = 0, \end{aligned} \quad (4.8)$$

其中 Z 是自表示矩阵, $\lambda > 0$ 用于控制投影空间数据对特征选择的影响程度。 AZ 和 $X^\top AZ$ 分别表示原始空间数据 A 和投影空间数据 $X^\top A$ 通过自表示的重构数据, $\text{Diag}(Z) = 0$ 是为了防止简单解 $Z = E$ 。通过联合学习自表示矩阵, 可以很好地刻画原始空间和投影空间的几何结构。后续把模型 (4.8) 记为融合对比学习的稀疏主成分分析 (SPCA with Contrastive Learning, SPCA-CL)。

4.2 数学模型与算法

本节首先提出基于对比学习和双稀疏约束的数学模型, 其次设计基于梯度下降和硬阈值的优化算法, 最后证明算法的收敛性。

4.2.1 数学模型

利用对比学习考虑数据结构信息的特性, 并借助上一章构建的 DSCOFS, 提出融合对比学习的双稀疏约束优化特征选择(DSCOFS with Contrastive Learning, DSCOFS-CL), 其数学模型为

$$\begin{aligned} \min_{X,Z} \quad & \lambda L_c(A, AZ) + (1 - \lambda) L_c(X^\top A, X^\top AZ) \\ \text{s.t.} \quad & X^\top X = I_m, \|X\|_{2,0} \leq s_1, \|X\|_0 \leq s_2, \\ & \text{rank}(Z) \leq r, \text{Diag}(Z) = 0, \end{aligned} \quad (4.9)$$

其中 $0 < \lambda < 1$ 用于调节原始空间和投影空间参与自表示学习的比重, $r > 0$ 用于控制矩阵 Z 的秩, 可根据实际需求进行设置。与模型 (4.8) 相比, 本章提出的 DSCOFS-CL 方法具有以下优势:

- 引入 $\text{rank}(Z) \leq r$ 可以保持自表示矩阵的全局结构。
- 嵌入 $\|X\|_{2,0} \leq s_1$ 和 $\|X\|_0 \leq s_2$ 的双稀疏约束, 能够提升特征选择的性能 (已在第三章验证)。
- 参数 λ 可平衡原始空间与投影空间中对比学习损失项的权重。

最终，DSCOFS-CL 特征选择的流程如图 4.2 所示。

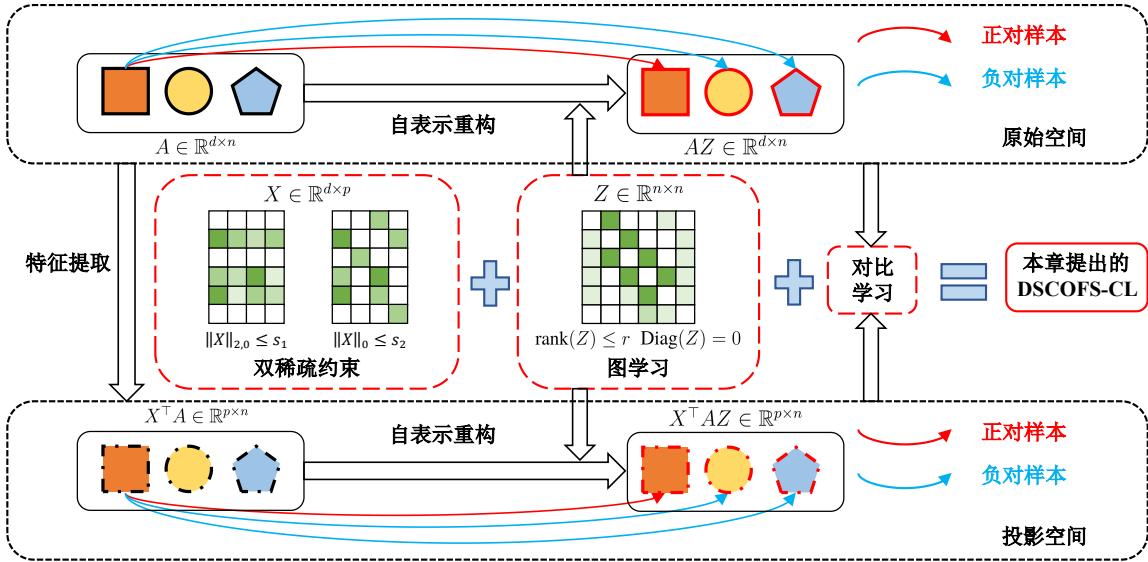


图 4.2 DSCOFS-CL 特征选择的流程图

Figure 4.2 The flowchart of feature selection of DSCOFS-CL

4.2.2 优化算法

显然，模型(4.9)是非凸、非光滑的优化模型，且对比学习损失的计算较为繁琐。

首先引入辅助变量 $X = P$ 、 $X = Q$ 和 $Z = Y$ ，将模型(4.9)等价转化为

$$\begin{aligned} & \min_{X, Z, Y, P, Q} \quad \lambda L_c(A, AZ) + (1 - \lambda) L_c(X^T A, X^T AZ) \\ \text{s.t.} \quad & X^T X = I_m, \quad P \in \mathcal{S}_1, \quad Q \in \mathcal{S}_2, \\ & Y \in \mathcal{R}, \quad Z \in \mathcal{D}, \\ & Z = Y, \quad X = P, \quad X = Q, \end{aligned} \tag{4.10}$$

其中

$$\begin{aligned} \mathcal{S}_1 &= \{P \in \mathbb{R}^{d \times m} \mid \|P\|_{2,0} \leq s_1\}, \\ \mathcal{S}_2 &= \{Q \in \mathbb{R}^{d \times m} \mid \|Q\|_0 \leq s_2\}, \\ \mathcal{R} &= \{Y \in \mathbb{R}^{n \times n} \mid \text{rank}(Y) \leq r\}, \\ \mathcal{D} &= \{Z \in \mathbb{R}^{n \times n} \mid \text{Diag}(Z) = 0\}. \end{aligned} \tag{4.11}$$

利用惩罚函数方法，将模型 (4.10) 转化为以下形式

$$\begin{aligned} \min_{X, Z, Y, P, Q} \quad & \lambda L_c(A, AZ) + (1 - \lambda)L_c(X^\top A, X^\top AZ) + \mu \|X^\top X - I_m\|_F^2 \\ & + \alpha \|Z - Y\|_F^2 + \beta \|X - P\|_F^2 + \gamma \|X - Q\|_F^2 \\ \text{s.t.} \quad & Z \in \mathcal{D}, Y \in \mathcal{R}, P \in \mathcal{S}_1, Q \in \mathcal{S}_2, \end{aligned} \quad (4.12)$$

其中 $\mu, \alpha, \beta, \gamma > 0$ 是惩罚参数。设 X^k, Z^k, Y^k, P^k 和 Q^k 是第 k 次更新的变量，同时在迭代过程中，引入近端参数 $0 < \tau_i < \infty$ ($i = 1, 2, 3, 4, 5$)，从而保证算法的收敛性。

(1) 固定 Y, Z, P 和 Q ，更新 X ：

$$\begin{aligned} \min_X \quad & (1 - \lambda)L_c(X^\top A, X^\top AZ^k) + \mu \|X^\top X - I_m\|_F^2 \\ & + \beta \|X - P^k\|_F^2 + \gamma \|X - Q^k\|_F^2 + \tau_1 \|X - X^k\|_F^2. \end{aligned} \quad (4.13)$$

由于模型 (4.13) 的目标函数涉及对比学习计算，无法直接写出显示解。然而，模型 (4.13) 是一个无约束的优化模型，可以通过梯度下降法求解。设目标函数为 $f(X)$ ，则 X^{k+1} 的更新可以以下式计算

$$X^{k+1} = X^k - \eta_1 \frac{\partial f(X)}{\partial X}, \quad (4.14)$$

其中 $\eta_1 > 0$ 表示更新的步长。

(2) 固定 X, Y, P 和 Q ，更新 Z ：

$$\begin{aligned} \min_Z \quad & \lambda L_c(A, AZ) + (1 - \lambda)L_c(X^{\top, k+1} A, X^{\top, k+1} AZ) \\ & + \alpha \|Z - Y^k\|_F^2 + \tau_2 \|Z - Z^k\|_F^2 \\ \text{s.t.} \quad & \text{Diag}(Z) = 0 \end{aligned} \quad (4.15)$$

为了使 Z 的对角元素为零，可以将 Z 设为 $M - \text{Diag}(M)$ ，其中 $\text{Diag}(M)$ 是由 M 对角元素组成的对角矩阵。于是，模型 (4.15) 可以改写为

$$\begin{aligned} \min_M \quad & \lambda L_c(A, A(M - \text{Diag}(M))) \\ & + (1 - \lambda)L_c(X^{\top, k+1} A, X^{\top, k+1} A(M - \text{Diag}(M))) \\ & + \alpha \|M - \text{Diag}(M) - Y^k\|_F^2 \\ & + \tau_2 \|M - \text{Diag}(M) - M^k + \text{Diag}(M^k)\|_F^2. \end{aligned} \quad (4.16)$$

设模型 (4.16) 的目标函数为 $h(M)$, 则 M^{k+1} 可以通过梯度下降计算

$$M^{k+1} = M^k - \eta_2 \frac{\partial h(M)}{\partial M}, \quad (4.17)$$

其中 $\eta_2 > 0$ 表示更新的步长。从而可以得到

$$Z^{k+1} = M^{k+1} - \text{Diag}(M^{k+1}). \quad (4.18)$$

(3) 固定 X 、 Z 、 P 和 Q , 更新 Y :

$$\begin{aligned} \min_Y \quad & \|Z^{k+1} - Y\|_{\text{F}}^2 + \tau_3 \|Y - Y^k\|_{\text{F}}^2 \\ \text{s.t.} \quad & \text{rank}(Y) \leq r. \end{aligned} \quad (4.19)$$

将模型 (4.19) 的两项 Frobenius 范数合并, 得到

$$\begin{aligned} \min_Y \quad & \left\| \frac{Z^{k+1} + \tau_3 Y^k}{1 + \tau_3} - Y \right\|_{\text{F}}^2 \\ \text{s.t.} \quad & \text{rank}(Y) \leq r. \end{aligned} \quad (4.20)$$

低秩模型 (4.20) 可以通过奇异值分解求解。设 $B^{k+1} = \frac{Z^{k+1} + \tau_3 Y^k}{1 + \tau_3}$, 对 B^{k+1} 进行奇异值分解 $B^{k+1} = U\Sigma V^\top$, 其中 $U, V, \Sigma \in \mathbb{R}^{n \times n}$, Σ 的对角线是矩阵 B^{k+1} 按降序排列的奇异值。分别取 U 和 V^\top 的前 r 列和前 r 行得到 $U_r \in \mathbb{R}^{n \times r}$ 和 $V_r^\top \in \mathbb{R}^{r \times n}$, 取 Σ 的前 r 个奇异值得到 $\Sigma_r \in \mathbb{R}^{r \times r}$ 。因此, Y^{k+1} 有显示解

$$Y^{k+1} = U_r \Sigma_r V_r^\top. \quad (4.21)$$

(4) 固定 X 、 Y 、 Z 和 Q , 更新 P :

$$\begin{aligned} \min_P \quad & \|X^{k+1} - P\|_{\text{F}}^2 + \tau_4 \|P - P^k\|_{\text{F}}^2 \\ \text{s.t.} \quad & \|P\|_{2,0} \leq s_1, \end{aligned} \quad (4.22)$$

等价于

$$\begin{aligned} \min_P \quad & \left\| \frac{X^{k+1} + \tau_4 P^k}{1 + \tau_4} - P \right\|_{\text{F}}^2 \\ \text{s.t.} \quad & \|P\|_{2,0} \leq s_1. \end{aligned} \quad (4.23)$$

设 $C^{k+1} = \frac{X^{k+1} + \tau_4 P^k}{1 + \tau_4}$, 计算 C^{k+1} 每一行的 ℓ_2 范数 $\|\mathbf{c}^{i,k+1}\|_2$, 并将其中第 s_1 大的值

记为 $t_{s_1}^{k+1}$ 。考虑到 $\ell_{2,0}$ 范数的行稀疏性, P^{k+1} 有显示解

$$\mathbf{p}^{i,k+1} = \begin{cases} \mathbf{c}^{i,k+1}, & \|\mathbf{c}^{i,k+1}\|_2 \geq t_{s_1}^{k+1}, \\ 0, & \|\mathbf{c}^{i,k+1}\|_2 < t_{s_1}^{k+1}. \end{cases} \quad (4.24)$$

(5) 固定 X 、 Y 、 Z 和 P , 更新 Q :

$$\begin{aligned} \min_Q \quad & \|X^{k+1} - Q\|_F^2 + \tau_5 \|Q - Q^k\|_F^2 \\ \text{s.t.} \quad & \|Q\|_0 \leq s_2, \end{aligned} \quad (4.25)$$

等价于

$$\begin{aligned} \min_Q \quad & \left\| \frac{X^{k+1} + \tau_5 Q^k}{1 + \tau_5} - Q \right\|_F^2 \\ \text{s.t.} \quad & \|Q\|_0 \leq s_2. \end{aligned} \quad (4.26)$$

记 $D^{k+1} = \frac{X^{k+1} + \tau_5 Q^k}{1 + \tau_5}$, 对 D^{k+1} 取绝对值, 并将其中第 s_2 大的绝对值记为 $t_{s_2}^{k+1}$ 。通过硬阈值算子可以直接得到 Q^{k+1} 的显示解, 即

$$Q_{ij}^{k+1} = \begin{cases} D_{ij}^{k+1}, & |D_{ij}^{k+1}| \geq t_{s_2}^{k+1}, \\ 0, & |D_{ij}^{k+1}| < t_{s_2}^{k+1}. \end{cases} \quad (4.27)$$

综上所述, 求解 DSCOFS-CL 的完整过程见算法 6。事实上, 更新 Y 、 P 和 Q 时, r 、 s_1 和 s_2 可以按照实际的需求设定, 并都有相应的显示解。而 X 和 Z 由于设计了交叉熵的计算, 这里直接调用 Pytorch 下的梯度求解工具。

4.2.3 理论分析

定义 $W = (X, Z, Y, P, Q)$, 并记

$$\begin{aligned} f(W) = & \lambda L_c(A, AZ) + (1 - \lambda) L_c(X^\top A, X^\top AZ) \\ & + \mu \|X^\top X - I_m\|_F^2 + \alpha \|Z - Y\|_F^2 + \beta \|X - P\|_F^2 \\ & + \gamma \|X - Q\|_F^2 + \delta_D(Z) + \delta_R(Y) + \delta_{S_1}(P) + \delta_{S_2}(Q). \end{aligned} \quad (4.28)$$

显然, $f(W)$ 是适当下半连续函数。如果 $0 \in \partial f(W)$, 则称 W 是模型(4.12)的驻点, 其中 $\partial f(W)$ 为 f 在点 W 的极限次微分。

算法 6 求解 DSCOFS-CL 的优化算法**输入:** 数据 A , 参数 $\lambda, \mu, \alpha, \beta, \gamma, s_1, s_2, r, \tau_i (i = 1, 2, 3, 4, 5)$ **初始化:** $k = 0$, 根据初始化策略得到 $(X^0, Z^0, Y^0, P^0, Q^0)$ **当 $k < 500$ 执行**

- 1: 通过式 (4.14) 得到 X^{k+1}
- 2: 通过式 (4.18) 得到 Z^{k+1}
- 3: 通过式 (4.21) 得到 Y^{k+1}
- 4: 通过式 (4.24) 得到 P^{k+1}
- 5: 通过式 (4.27) 得到 Q^{k+1}
- 6: $k = k + 1$

结束循环**输出:** $(X^{k+1}, Z^{k+1}, Y^{k+1}, P^{k+1}, Q^{k+1})$

引理 4.1 若 $\{W^k\} = \{(X^k, Z^k, Y^k, P^k, Q^k)\}$ 是算法 6 产生的迭代序列, 则在迭代点处的函数值序列 $\{f(W^k)\}$ 单调下降, 且存在 $\rho_1 > 0$ 使得

$$\rho_1 \|W^{k+1} - W^k\|_{\text{F}}^2 \leq f(W^k) - f(W^{k+1}). \quad (4.29)$$

证明: 令 $X^{k+1}, Z^{k+1}, Y^{k+1}, P^{k+1}$ 和 Q^{k+1} 是模型 (4.13)、(4.15)、(4.19)、(4.22) 和 (4.25) 的最优解, 则下列不等式系统成立

$$\begin{cases} f(X^{k+1}, Z^k, Y^k, P^k, Q^k) + \tau_1 \|X^{k+1} - X^k\|_{\text{F}}^2 \leq f(X^k, Z^k, Y^k, P^k, Q^k), \\ f(X^{k+1}, Z^{k+1}, Y^k, P^k, Q^k) + \tau_2 \|Z^{k+1} - Z^k\|_{\text{F}}^2 \leq f(X^{k+1}, Z^k, Y^k, P^k, Q^k), \\ f(X^{k+1}, Z^{k+1}, Y^{k+1}, P^k, Q^k) + \tau_3 \|Y^{k+1} - Y^k\|_{\text{F}}^2 \leq f(X^{k+1}, Z^{k+1}, Y^k, P^k, Q^k), \\ f(X^{k+1}, Z^{k+1}, Y^{k+1}, P^{k+1}, Q^k) + \tau_4 \|P^{k+1} - P^k\|_{\text{F}}^2 \leq f(X^{k+1}, Z^{k+1}, Y^{k+1}, P^k, Q^k), \\ f(X^{k+1}, Z^{k+1}, Y^{k+1}, P^{k+1}, Q^{k+1}) + \tau_5 \|Q^{k+1} - Q^k\|_{\text{F}}^2 \leq f(X^{k+1}, Z^{k+1}, Y^{k+1}, P^{k+1}, Q^k). \end{cases} \quad (4.30)$$

这意味着

$$\begin{aligned} & f(X^{k+1}, Z^{k+1}, Y^{k+1}, P^{k+1}, Q^{k+1}) + \tau_1 \|X^{k+1} - X^k\|_{\text{F}}^2 + \tau_2 \|Z^{k+1} - Z^k\|_{\text{F}}^2 \\ & + \tau_3 \|Y^{k+1} - Y^k\|_{\text{F}}^2 + \tau_4 \|P^{k+1} - P^k\|_{\text{F}}^2 + \tau_5 \|Q^{k+1} - Q^k\|_{\text{F}}^2 \\ & \leq f(X^k, Z^k, Y^k, P^k, Q^k). \end{aligned} \quad (4.31)$$

于是, 存在 $\rho_1 = \min\{\tau_1, \tau_2, \tau_3, \tau_4, \tau_5\} > 0$ 使得 $f(W^{k+1}) + \rho_1 \|W^{k+1} - W^k\|_{\text{F}}^2 \leq f(W^k)$ 。因此, 式 (4.29) 成立, 可知 $\{f(W^k)\}$ 是单调下降的。证毕。 \square

上述引理 4.1 说明了迭代序列对应的目标函数 $\{f(W^k)\}$ 下降量的下界可被相邻迭代点之间的距离控制。进一步, 设 N 为任意整数, 在式 (4.29) 中对 k 求和, 得

$$\sum_{k=0}^{N-1} \|W^{k+1} - W^k\|_{\text{F}}^2 \leq \frac{1}{\rho_1} (f(W^0) - f(W^N)) \leq \frac{1}{\rho_1} (f(W^0) - f(W^*)), \quad (4.32)$$

这里 W^* 为最优解。令 $N \rightarrow \infty$, 可得 $\sum_{k=0}^{\infty} \|W^{k+1} - W^k\|_{\text{F}}^2 < +\infty$, 因此

$$\lim_{k \rightarrow \infty} \|W^{k+1} - W^k\|_{\text{F}} = 0. \quad (4.33)$$

下面引理将表明随着迭代的进行, $\partial f(W^k)$ 的模长的上界将被一个模长不断趋于 0 的矩阵控制。

引理 4.2 设 $\{W^k\} = \{(X^k, Z^k, Y^k, P^k, Q^k)\}$ 是算法 6 产生的序列, 则存在 $\rho_2 > 0$ 和 $\Xi^k \in \partial f(W^k)$, 使得

$$\|\Xi^k\|_{\text{F}} \leq \rho_2 \|W^k - W^{k-1}\|_{\text{F}}. \quad (4.34)$$

证明: 考虑到 $f(W)$ 的结构, 可以拆分为

$$f(W) = f_1(X) + f_2(Z) + f_3(Y) + f_4(P) + f_5(Q) + g(W), \quad (4.35)$$

其中拆分函数为

$$\begin{cases} f_1(X) = \mu \|X^\top X - I_m\|_{\text{F}}^2, \\ f_2(Z) = \lambda L_c(A, AZ) + \delta_{\mathcal{D}}(Z), \\ f_3(Y) = \delta_{\mathcal{R}}(Y), \\ f_4(P) = \delta_{\mathcal{S}_1}(P), \\ f_5(Q) = \delta_{\mathcal{S}_2}(Q), \end{cases} \quad (4.36)$$

与

$$\begin{aligned} g(W) = & (1 - \lambda)L_c(X^\top A, X^\top AZ) \\ & + \alpha \|Z - Y\|_{\text{F}}^2 + \beta \|X - P\|_{\text{F}}^2 + \gamma \|X - Q\|_{\text{F}}^2. \end{aligned} \quad (4.37)$$

根据文献^[99] 中 8.8 (c), 模型 (4.13)、(4.15)、(4.19)、(4.22) 和 (4.25) 的一阶最优性条件为

$$\left\{ \begin{array}{l} \nabla f_1(X^{k+1}) + \nabla_X g(X^{k+1}, Z^k, Y^k, P^k, Q^k) + \tau_1(X^{k+1} - X^k) = 0, \\ \partial f_2(Z^{k+1}) + \nabla_Z g(X^{k+1}, Z^{k+1}, Y^k, P^k, Q^k) + \tau_2(Z^{k+1} - Z^k) = 0, \\ \partial f_3(Y^{k+1}) + \nabla_Y g(X^{k+1}, Z^{k+1}, Y^{k+1}, P^k, Q^k) + \tau_3(Y^{k+1} - Y^k) = 0, \\ \partial f_4(P^{k+1}) + \nabla_P g(X^{k+1}, Z^{k+1}, Y^{k+1}, P^{k+1}, Q^k) + \tau_4(P^{k+1} - P^k) = 0, \\ \partial f_5(Q^{k+1}) + \nabla_Q g(X^{k+1}, Z^{k+1}, Y^{k+1}, P^{k+1}, Q^{k+1}) + \tau_5(Q^{k+1} - Q^k) = 0. \end{array} \right. \quad (4.38)$$

因此, 存在 $\Xi^{k+1} = (\Xi_X^{k+1}, \Xi_Z^{k+1}, \Xi_Y^{k+1}, \Xi_P^{k+1}, \Xi_Q^{k+1}) \in \partial f(W^{k+1})$, 其中

$$\left\{ \begin{array}{l} \Xi_X^{k+1} = \nabla f_1(X^{k+1}) + \nabla_X g(X^{k+1}, Z^k, Y^k, P^k, Q^k), \\ \Xi_Z^{k+1} \in \partial f_2(Z^{k+1}) + \nabla_Z g(X^{k+1}, Z^{k+1}, Y^k, P^k, Q^k), \\ \Xi_Y^{k+1} \in \partial f_3(Y^{k+1}) + \nabla_Y g(X^{k+1}, Z^{k+1}, Y^{k+1}, P^k, Q^k), \\ \Xi_P^{k+1} \in \partial f_4(P^{k+1}) + \nabla_P g(X^{k+1}, Z^{k+1}, Y^{k+1}, P^{k+1}, Q^k), \\ \Xi_Q^{k+1} \in \partial f_5(Q^{k+1}) + \nabla_Q g(X^{k+1}, Z^{k+1}, Y^{k+1}, P^{k+1}, Q^{k+1}), \end{array} \right. \quad (4.39)$$

满足

$$\left\{ \begin{array}{l} \Xi_X^{k+1} + \tau_1(X^{k+1} - X^k) = 0, \\ \Xi_Z^{k+1} + \tau_2(Z^{k+1} - Z^k) = 0, \\ \Xi_Y^{k+1} + \tau_3(Y^{k+1} - Y^k) = 0, \\ \Xi_P^{k+1} + \tau_4(P^{k+1} - P^k) = 0, \\ \Xi_Q^{k+1} + \tau_5(Q^{k+1} - Q^k) = 0. \end{array} \right. \quad (4.40)$$

注意到

$$\|\Xi^{k+1}\|_{\text{F}} \leq \|\Xi_X^{k+1}\|_{\text{F}} + \|\Xi_Z^{k+1}\|_{\text{F}} + \|\Xi_Y^{k+1}\|_{\text{F}} + \|\Xi_P^{k+1}\|_{\text{F}} + \|\Xi_Q^{k+1}\|_{\text{F}}. \quad (4.41)$$

结合式(4.40), 存在 $\rho_2 = \max\{\tau_1, \tau_2, \tau_3, \tau_4, \tau_5\}$ 使得 $\|\Xi^{k+1}\|_{\text{F}} \leq \rho_2 \|W^{k+1} - W^k\|_{\text{F}}$ 。因此, 式 (4.34) 成立。证毕。 \square

上述引理 4.2 说明 $\lim_{k \rightarrow \infty} \partial f(W^k) = 0$, 这蕴含着算法 6 的收敛性。进一步, 下面定理说明, 在迭代序列 $\{W^k\}$ 有界的假设下, 从初始点 W^0 开始迭代, 产生的点列 $\{W^k\}$ 的极限点都是模型 (4.12) 的驻点。

定理 4.1 设 $\{W^k\} = \{(X^k, Z^k, Y^k, P^k, Q^k)\}$ 是算法 6 产生的序列, 若 $\{W^k\}$ 有界, 设 W^* 为 $\{W^k\}$ 子序列的极限点, 则有 W^* 是模型 (4.12) 的驻点, 即 $0 \in \partial f(W^*)$ 。

证明: 由于 $\{W^k\}$ 有界, 令 $W^* = (X^*, Z^*, Y^*, P^*, Q^*)$ 是点列 $\{W^k\}$ 的一个极限点, 这意味着存在子序列 $(X^{k_i}, Z^{k_i}, Y^{k_i}, P^{k_i}, Q^{k_i})$, 当 $i \rightarrow \infty$ 时使得

$$(X^{k_i}, Z^{k_i}, Y^{k_i}, P^{k_i}, Q^{k_i}) \rightarrow (X^*, Z^*, Y^*, P^*, Q^*). \quad (4.42)$$

由于式 (4.28) 中 $f(W)$ 具有下半连续性。根据利用引理 4.1、引理 4.2 和文献^[100] 中引理 5 可得, $0 \in \partial f(W^*)$ 。因此 W^* 是模型 (4.12) 的驻点。证毕。 \square

4.3 数值实验与分析

本节将通过实验验证 DSCOFS-CL 的有效性和优越性, 对比的方法包括 LapScore^[20]、UDFS^[24]、SOGFS^[27]、RNE^[26]、FSPCA^[43]、SPCAF^[22]、DSCOFS 和 SPCA-CL^[32]。其中 SPCA-CL 的程序由作者提供, 其它与章节 3.3 保持一致。

4.3.1 实验设置

4.3.1.1 实验数据

在实验中, 使用从章节 3.3.1.1 选择的六个真实数据集, 即一个物体图像数据集 COIL20, 一个面部图像数据集 UMIST, 一个手写图像数据集 USPS, 一个生物数据集 GLIOMA, 一个语音字母识别数据集 Isolet, 以及一个深度学习数据集 MSTARSC, 有关这些数据集的详细信息如表 4.1 所示。

4.3.1.2 参数设置

对于 LapScore、UDFS、SOGFS、RNE、FSPCA、SPCAF^[22] 和 DSCOFS, 参数设置和基本的实验候选集与章节 3.3.1.2 保持一致, 此处不再赘述。对于 SPCA-CL 和 DSOFS-CL, 投影维度固定为数据的类别数。对于 DSOFS-CL, 稀疏度参数 s_1 和 s_2 的设置与 DSCOFS 中的 s_1 和 s_2 类似。由于在 DSCOFS 中已经验证了低稀疏度的有效

表 4.1 数据集信息

Table 4.1 The dataset information

数据集	特征数	样本数	类别数
COIL20	1024	1440	20
USPS	256	1000	10
GLIOMA	4434	50	4
UMIST	644	575	20
Isolet	617	1560	26
MSTARSC	1024	2425	10

性，因此本节的稀疏度百分比 c 从 $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ 中选择。此外，对于 DSOFS-CL，设定平衡参数 $\lambda = 0.5$ ，保证原始空间和投影空间的同等重要性。设定低秩约束 $r = 0.1n$ ，保证自表示矩阵的低秩性。

4.3.1.3 初始化和停止准则

上一章采用了随机正交解作为初始值，经过分析本章将采用更为一般的非正交解作为初始值。Xavier 均匀初始化可以从一个均匀分布中采样，常用于神经网络的参数初始化，并能够保持深度神经网络中梯度的稳定性。Xavier 均匀初始化的范围由输入和输出的维度决定，对于变量 $X \in \mathbb{R}^{d \times m}$ 和 $M \in \mathbb{R}^{n \times n}$ ，初始化范围为

$$X^0 \sim U\left(-\sqrt{\frac{6}{d+m}}, \sqrt{\frac{6}{d+m}}\right), \quad M^0 \sim U\left(-\sqrt{\frac{3}{n}}, \sqrt{\frac{3}{n}}\right). \quad (4.43)$$

设 $B^0 = Z^0/(1 + \tau_3)$ 、 $C^0 = X^0/(1 + \tau_4)$ 和 $D^0 = X^0/(1 + \tau_5)$ ，则 Y^0 、 P^0 和 Q^0 可以根据式 (4.21)、(4.24) 和 (4.27) 得到。此外，算法 6 的迭代次数满足 500 时停止。

4.3.2 实验结果

图 4.3 和图 4.4 展示了不同特征数量 ACC 和 NMI 的均值曲线，其中 ALLfea 作为参考基准。表 4.2 和表 4.3 给出了在 100 个特征范围内最佳 ACC 和 NMI 的平均值、标准差以及相应的特征数量，并且最好和第二好的结果（除 ALLfea 外）分别用红色和蓝色标记。

从图 4.3 可以看出，本章提出的 DSCOFS-CL 在所有数据集上都表现出卓越的性能。在 DSCOFS 有着优越性能的前提下，DSCOFS-CL 融合了对比学习后相比 DSCOFS 有了进一步的提升，同时相比 SPCA-CL 也有一定的提升，这在 UMIST、

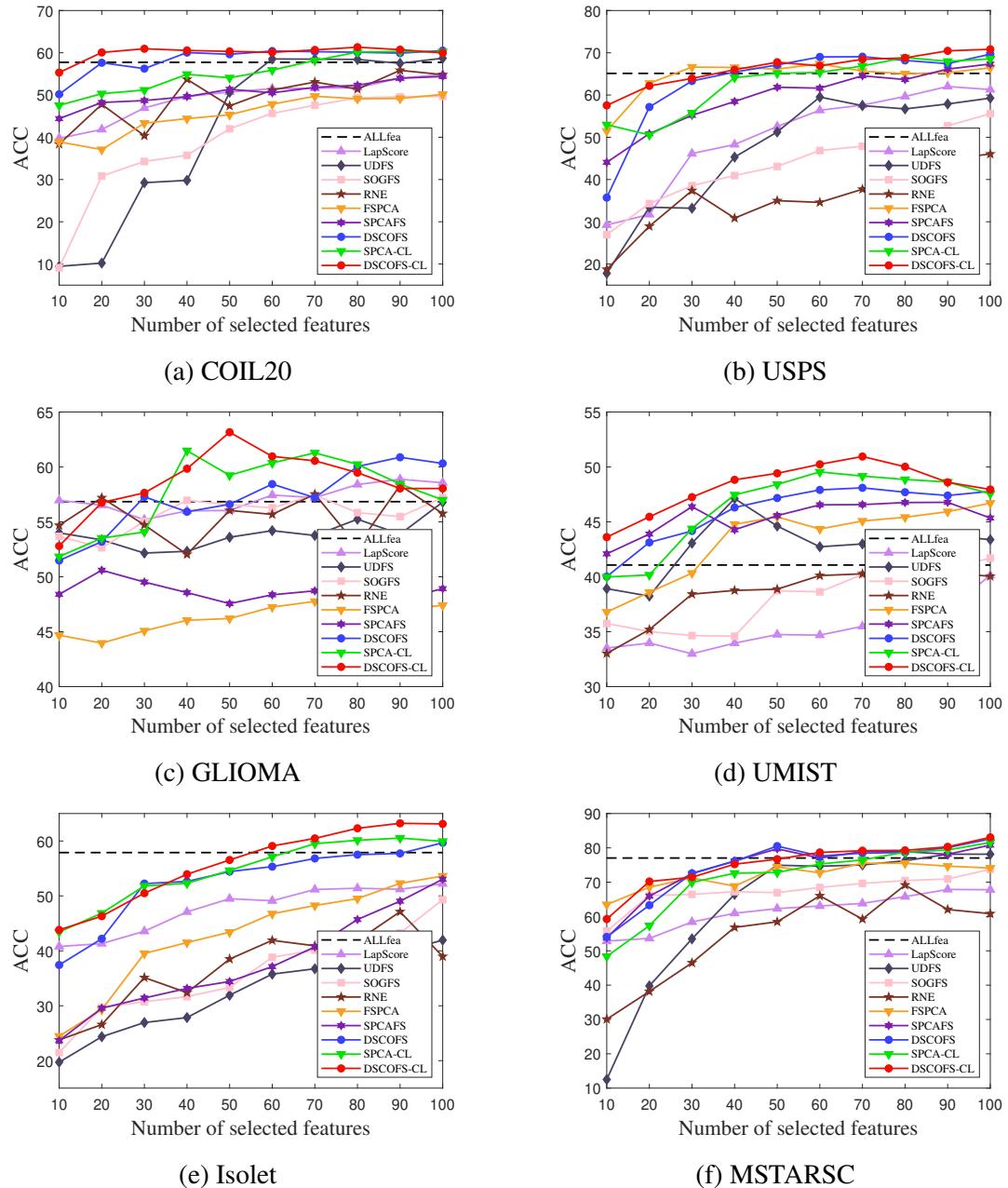


图 4.3 对比方法在六个真实数据集上的 ACC (%) 曲线

Figure 4.3 The ACC (%) curves of compared methods on six real-world datasets

GLIOMA 和 Isolet 数据集上的表现较为明显。从表 4.2 可以观察到，DSCOFS-CL 在所有数据集上均表现最好，同时第二好的结果均由 DSCOFS 和 SPCA-CL 得到。特别地，DSCOFS-CL 在 GLIOMA、UMIST 和 Isolet 数据集上有不错的提升，相较于第二好的结果分别有 1.68%、1.40% 和 2.69% 的增长。在六个数据集上的平均 ACC 结果，DSCOFS 和 SPCA-CL 仅相差 0.15%，而 DSCOFS-CL 相较于 DSCOFS 和 SPCA-CL 分别提升了 1.85% 和 1.7%。

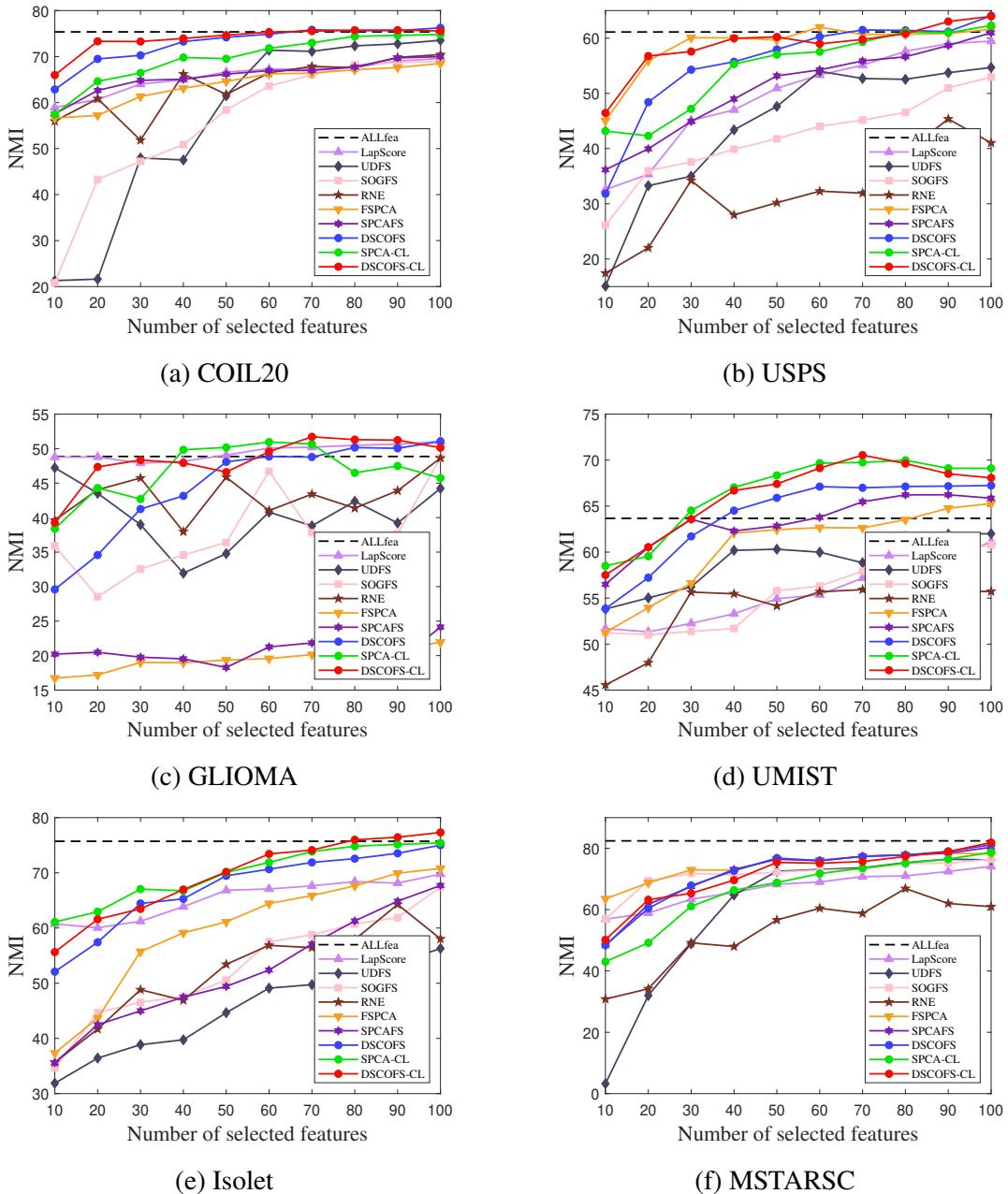


图 4.4 对比方法在六个真实数据集上的 NMI (%) 曲线

Figure 4.4 The NMI (%) curves of compared methods on six real-world datasets

从图 4.4 也可以观察到，本章提出的 DSCOFS-CL 展现出卓越的性能。DSCOFS-CL 在 Isolet 数据集上有着一定的提升，且是唯一超过基线的方法。值得注意的是，在 GLIOMA 数据集上，当 DSCOFS-CL 所选特征为 50 时，NMI 结果较低。这与 ACC 曲线上的表现相反，进一步说明了 ACC 和 NMI 在同组参数下不一定是完全正相关的。从表 4.3 可以发现，DSCOFS-CL 均取到了最好或者第二好的结果。与 ACC 结果类似，DSCOFS-CL 在 Isolet 数据集上依然有不错的提升，相比于第二好的 SPCA-CL 有

表 4.2 对比方法在六个真实数据集上的 ACC (平均值% ± 标准差%) 结果

Table 4.2 The ACC (mean% ± std%) results of compared methods on six real-world datasets

数据集	ALLfea	LapScore	UDFS	SOGFS	RNE	FSPCA	SPCAFS	DSCOFS	SPCA-CL	DSCOFS-CL
COIL20	57.74±4.93	54.82±3.91	58.71±3.47	49.66±4.81	55.84±4.41	50.15±4.70	54.39±3.67	60.51±4.63	60.31±3.49	61.32±5.18
USPS	65.12±4.95	62.02±4.09	59.52±2.97	55.58±3.07	46.04±2.69	67.38±4.36	67.34±4.49	69.67±4.97	68.88±4.05	70.82±4.77
GLIOMA	56.84±5.24	58.88±3.96	56.80±4.85	57.44±6.16	58.32±7.31	47.92±4.61	50.60±5.02	60.88±6.31	61.48±6.20	63.16±7.46
UMIST	41.07±2.38	40.13±2.79	47.12±2.49	41.70±3.17	40.35±2.26	46.70±2.29	46.78±2.51	48.10±3.01	49.55±3.00	50.95±3.15
Isolet	57.89±3.82	52.21±2.76	41.95±2.07	49.31±2.32	47.12±2.06	53.62±2.36	53.04±2.33	59.67±3.46	60.53±3.75	63.22±3.50
MSTARSC	77.04±7.98	67.87±3.49	78.15±5.80	73.74±5.89	69.16±6.03	75.52±6.22	80.80±5.95	82.59±7.41	81.57±6.28	83.06±6.31
Average	59.28±4.88	55.99±3.50	57.04±3.69	54.57±4.24	52.81±4.13	56.88±4.09	58.83±4.00	63.57±4.96	63.72±4.46	65.42±5.06

表 4.3 对比方法在六个真实数据集上的 NMI (平均值% \pm 标准差%) 结果Table 4.3 The NMI (mean% \pm std%) results of compared methods on six real-world datasets

数据集	ALIfea	LapScore	UDFS	SOGFS	RNE	FSPCA	SPCAFS	DSCOFS	SPCA-CL	DSCOFS-CL
COIL20	75.37 \pm 1.96	69.59 \pm 1.48	73.54 \pm 1.76	68.92 \pm 1.84	70.43 \pm 1.92	68.50 \pm 1.56	69.98 \pm 1.45	76.25\pm1.71	74.79 \pm 1.48	75.76\pm1.76
USPS	61.12 \pm 2.01	59.46 \pm 1.80	54.69 \pm 2.11	52.96 \pm 1.54	45.36 \pm 1.93	62.00 \pm 1.87	60.98 \pm 2.37	64.06\pm2.58	62.29 \pm 2.40	63.95\pm2.67
GLIOMA	48.86 \pm 5.72	51.03 \pm 2.48	47.22 \pm 3.53	48.67 \pm 10.98	48.62 \pm 6.32	21.94 \pm 5.28	24.14 \pm 6.97	51.06\pm6.19	50.95 \pm 4.10	51.71\pm5.03
UMIST	63.67 \pm 1.85	61.16 \pm 1.71	62.00 \pm 1.58	60.79 \pm 1.54	55.92 \pm 1.57	65.27 \pm 1.58	66.23 \pm 1.60	67.24 \pm 1.85	69.98\pm1.84	70.54\pm1.70
Isolet	75.72 \pm 1.70	69.77 \pm 1.20	56.29 \pm 1.11	67.40 \pm 1.44	64.27 \pm 0.95	70.79 \pm 1.12	67.71 \pm 1.33	75.01 \pm 1.35	75.41\pm1.51	77.32\pm1.37
MSTARSC	82.42 \pm 3.31	74.10 \pm 1.76	76.45 \pm 2.47	76.39 \pm 1.70	66.87 \pm 1.99	78.39 \pm 2.17	80.33 \pm 2.50	81.14\pm3.13	78.63 \pm 2.50	81.88\pm2.03
Average	67.86 \pm 2.76	64.19 \pm 1.74	61.70 \pm 2.09	62.52 \pm 3.17	58.58 \pm 2.45	61.15 \pm 2.26	61.56 \pm 2.70	69.13\pm2.80	68.68 \pm 2.31	70.19\pm2.43

1.91% 的增加。在六个数据集上的平均 NMI 结果，DSCOFS 取得第二好的表现，并且相较于 SPCA-CL 提升了 0.45%，而 DSCOFS-CL 相较于 DSCOFS 和 SPCA-CL 分别提升了 1.06% 和 1.51%。

综合 ACC 和 NMI 的结果可以看到，本章提出的 DSCOFS-CL 在所有数据集上都表现良好。实际上，上一章所提出的 DSCOFS 已经拥有不错的性能，并且通过双稀疏约束提升了局部结构的辨别能力。而 SPCA-CL 使用对比学习损失度量重构误差，通过样本间的关系来寻找更具判别性的特征。从结果上来看，SPCA-CL 和 DSCOFS 都有着不错的表现且性能接近，而 DSCOFS-CL 结合了两者的特点，采用双稀疏约束充分表示稀疏结构并辅助自表示矩阵学习到更好的数据结构，最终实现出色的特征选择性能。由此可以得出结论，对比学习损失作为重构误差的度量能够进一步挖掘数据的有效信息，这也展现了 DSCOFS-CL 的优越性和未来潜力。

4.3.3 消融实验

关于对比学习损失项，在实验中选择了平衡参数 $\lambda = 0.5$ ，从而保证原始空间和投影空间的同等重要性。下面考虑当 $\lambda = 0$ 的情形，即仅通过投影空间学习自表示矩阵。需要注意的是，DSCOFS-CL 的特征选择是根据投影矩阵 X 确定的，因此当 $\lambda = 1$ 时无法获得投影矩阵 X ，于是在本实验中不考虑 $\lambda = 1$ 的情形。为了便于描述，记模型 (4.9) 为 Case1，而当 $\lambda = 0$ 时模型为 Case2，具体为

$$\begin{aligned} \min_{X,Z} \quad & L_c(X^\top A, X^\top AZ) \\ \text{s.t.} \quad & X^\top X = I_m, \|X\|_{2,0} \leq s_1, \|X\|_0 \leq s_2, \\ & \text{rank}(Z) \leq r, \text{Diag}(Z) = 0. \end{aligned} \tag{4.44}$$

在 COIL20 和 UMIST 数据集上进行特征选择实验，记录自表示矩阵 Z 以及 ACC 和 NMI 结果。利用 Z 计算数据之间的相似度矩阵，即 $S = \frac{|Z|+|Z|^\top}{2}$ 。相似度矩阵 S 的可视化结果如图 4.5 所示，ACC 和 NMI 的结果如图 4.6 所示。

从图 4.5 可见，Case1 的相似度矩阵 S 显示出明显的簇结构，反映了数据在自表示过程中的内在关联性。同时，Case1 还呈现低秩结构，表明低秩约束在模型中得到了有效体现。与之相比，Case2 虽然揭示了低秩结构，但未能有效学习到簇结构。因此，仅通过投影空间学习最优图的效果并不理想，其原因在于数据经过投影降维后，可能会破坏原有的结构特征。从图 4.6 可以看到，Case2 的性能与 Case1 的相比存在

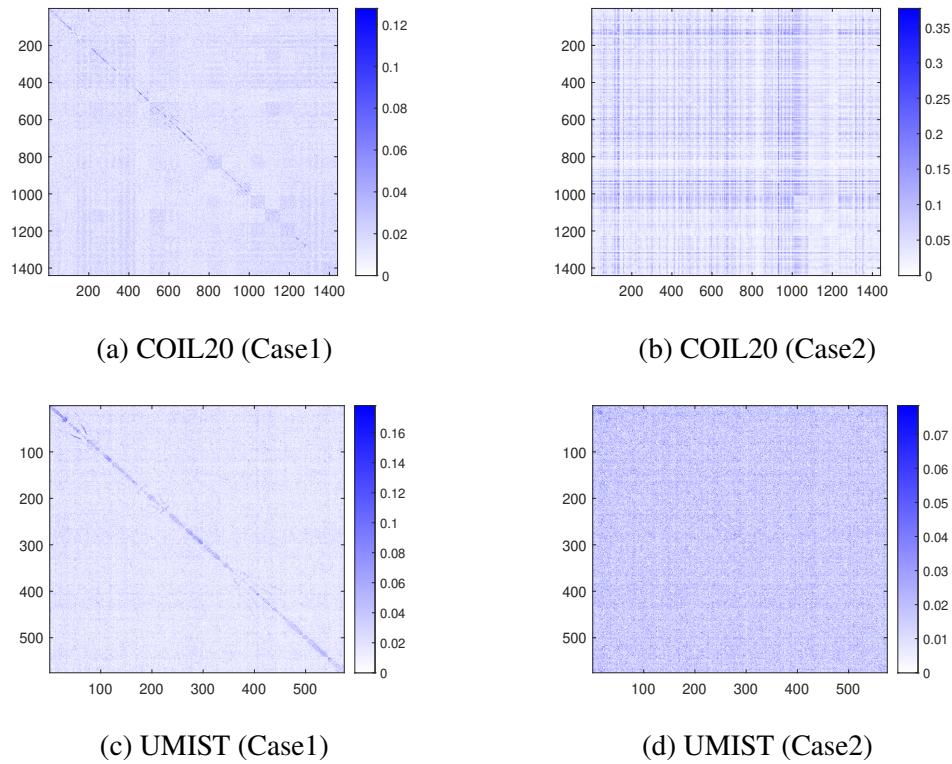
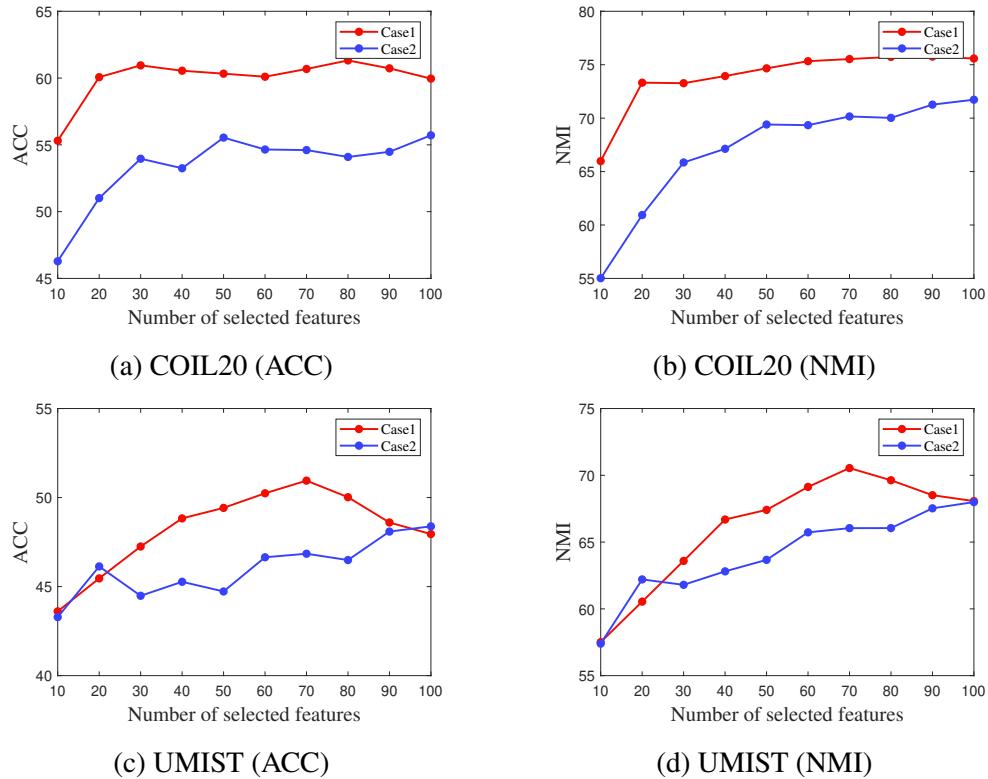
图 4.5 两个真实数据集上相似度矩阵 S 的可视化Figure 4.5 The visualization of the similarity matrix S on two real-world datasets

图 4.6 DSCOFS-CL 在两个真实数据集上的消融实验结果

Figure 4.6 The ablation experiment results of DSCOFS-CL on two real-world datasets

一定的差距。结合图 4.5 中的相似度矩阵 S , 可以得出结论, 学习原始空间中的数据结构对提升无监督特征选择性能是有效的。综合实验结果表明, DSCOFS-CL 通过在原始空间和投影空间中联合学习自表示矩阵, 能够获得更优的特征选择效果。

4.3.4 讨论

4.3.4.1 统计检验

仿照章节 3.3.4.1 的统计检验, 假设 \mathcal{H}_0 表示所有对比方法的性能没有显著差异, 在显著性水平设定为 $\alpha = 0.05$ 的情况下对 DSCOFS-CL 进行 Friedman 检验和后验 Nemenyi 检验, 其结果分别如表 4.4 和图 4.7 所示。

表 4.4 ACC 指标下 DSCOFS-CL 的 Friedman 检验结果

Table 4.4 The Friedman test result of DSCOFS-CL in terms of ACC

方法	平均排名	p 值	假设
LapScore	6.667		
UDFS	6.000		
SOGFS	7.333		
RNE	7.167		
FSPCA	6.167	2.280×10^{-5}	拒绝
SPCAF	5.667		
DSCOFS	2.500		
PCA-CL	2.500		
DSCOFS-CL	1.000		

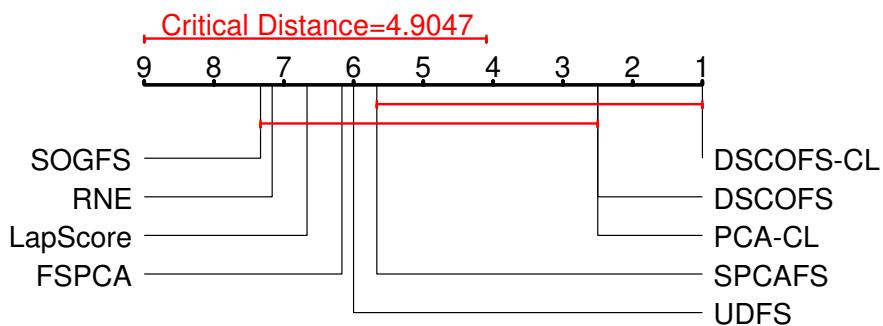


图 4.7 ACC 指标下 DSCOFS-CL 的后验 Nemenyi 检验结果

Figure 4.7 The post-hoc Nemenyi test results of DSCOFS-CL in terms of ACC

从表 4.4 可以看到 $p = 2.280 \times 10^{-5}$, 这意味着结果拒绝原假设 \mathcal{H}_0 , 即所有对

比方法之间确实存在显著差异。从图 4.7 则可以看到，本章提出的 DSCOFS-CL 与 LapScore、UDFS、SOGFS、RNE 和 FSPCA 有显著差异，但与 SPCAFS、DSCOFS 和 SPCA-CL 没有显著差异。然而，DSCOFS、SPCA-CL、和 SPCAFS 与其他的方法都在同一个 CD 值内，这说明除 DSCOFS-CL 外的方法之间没有显著的差异。需要注意的是，上一章中 DSCOFS 与 LapScore、UDFS、SOGFS、RNE 和 FSPCA 存在显著差异。然而，由于本章中使用的数据集以及对比方法的平均排名不同，导致 DSCOFS 在后验 Nemenyi 检验结果上与 LapScore、UDFS、SOGFS、RNE 和 FSPCA 之间没有显著差异。通过融入对比学习，DSCOFS-CL 与 LapScore、UDFS、SOGFS、RNE 和 FSPCA 之间产生了明显的差异，这表明对比学习可以进一步挖掘数据中的有效信息，从而实现更好的无监督特征选择性能。

4.3.4.2 参数敏感度分析

对于 DSCOFS-CL，双稀疏约束参数 s_1 和 $s_2 = cdm$ 是控制模型学习稀疏结构的关键，而 r 是控制自表示矩阵 Z 学习数据结构的关键。此外，惩罚参数 α 、 β 和 γ 在模型 (4.12) 也会影响自表示矩阵和双稀疏约束。在实验中，将低秩结构固定为 $r = 0.1n$ ，选择 s_1 、 c 、 α 、 β 和 γ 分析这些参数对特征选择性能的影响。在 USPS 数据集上的参数敏感度结果如图 4.8 所示。

从图 4.8 中的 (a) 和 (b) 可以看出，元素稀疏度 s_2 对性能有较明显地影响。尤其在稀疏度百分比 $c = 0.3$ 时，性能有一定的提升，而在 $c = 0.1$ 时，性能明显下降，这说明稀疏度并不是越低越好，过于低的稀疏度可能造成有效数据的丢失而影响性能。从图 4.8 中的 (c) - (h) 可以观察到，惩罚参数 α 、 β 和 γ 对性能都有一定的影响，其中 β 和 γ 的影响相较于 α 更大。这些参数都是影响模型求解的关键，因此在实验中要谨慎选择。未来，我们可以尝试使用自适应方法进行参数调节，例如深度展开网络，以进一步优化模型的性能。

4.3.4.3 模型收敛性分析

为了验证算法 6 的收敛性，记模型 (4.12) 的目标函数为 Loss1，模型 (4.13) 和 (4.16) 的目标函数为 Loss2 和 Loss3。损失函数在最优参数下迭代过程的变化如图 4.9 所示。从图中可以看出，Loss1、Loss2 和 Loss3 在迭代过程中呈现一致的下降趋势，同时在 100 次迭代内完成了快速的收敛，并最终缓慢平稳。值得注意的是，Loss2 在

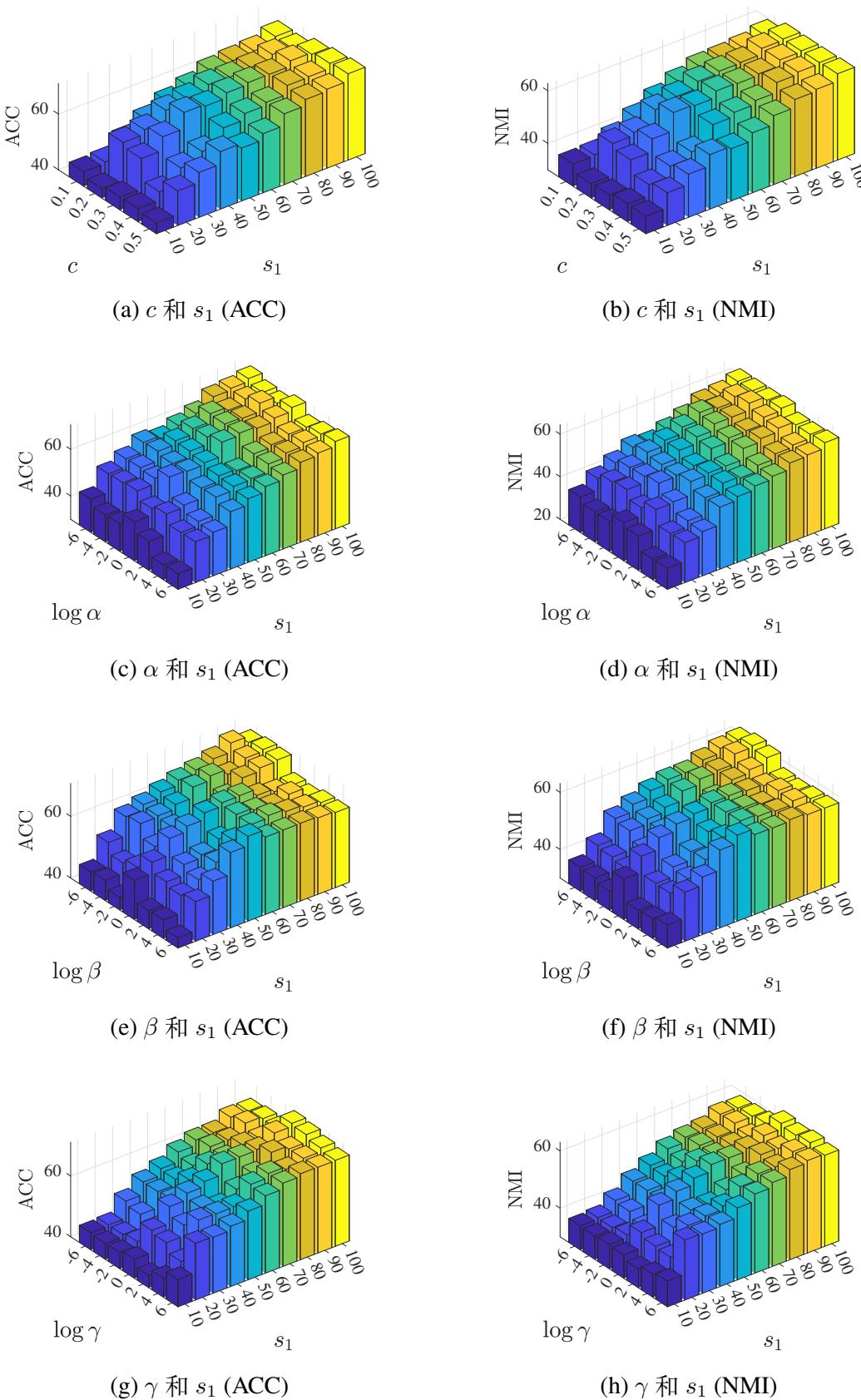


图 4.8 DSCOFS-CL 在 USPS 数据集上的参数敏感度分析结果

Figure 4.8 The parameter sensitivity analysis results of DSCOFS-CL on the USPS dataset

COIL20 和 Isolet 数据集上接近 Loss1，而在另外两个数据集上 Loss1，这与 Loss3 相反。这表明，原始空间和投影空间在不同数据集上对自表示矩阵学习的影响程度存在差异。最终，收敛曲线的变化验证了定理 4.1 所证明的收敛性。

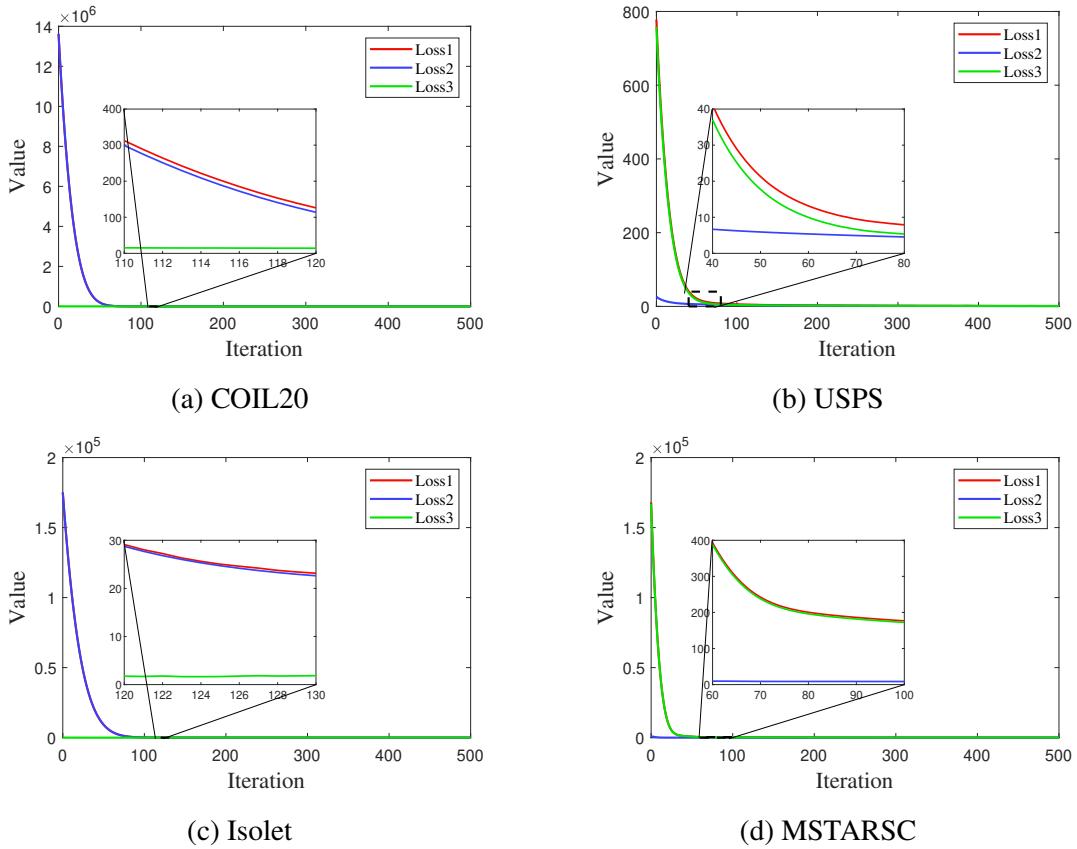


图 4.9 DSCOFS-CL 在四个真实数据集上的收敛曲线

Figure 4.9 The convergence curves of DSCOFS-CL on four real-world datasets

4.3.4.4 模型稳定性分析

图 4.10 显示了最佳聚类结果的 50 次聚类分布。可以看到 DSCOFS-CL 聚类结果有一定的波动，但 DSCOFS-CL 的整体结果优于其他对比方法。特别是在 Isolet 数据集上，相比于 DSCOFS 以及对比的 SPCA-CL，本章提出的 DSCOFS-CL 有着较为稳定的聚类结果和更高的性能表现。同时也注意到，DSCOFS-CL 在 USPS 数据集上的最大值和最小值波动较大，这可能是由于模型得到的特征区分度仍然不够所带来的影响。综合考虑所有数据集的性能表现，DSCOFS-CL 展现出了良好的稳定性。

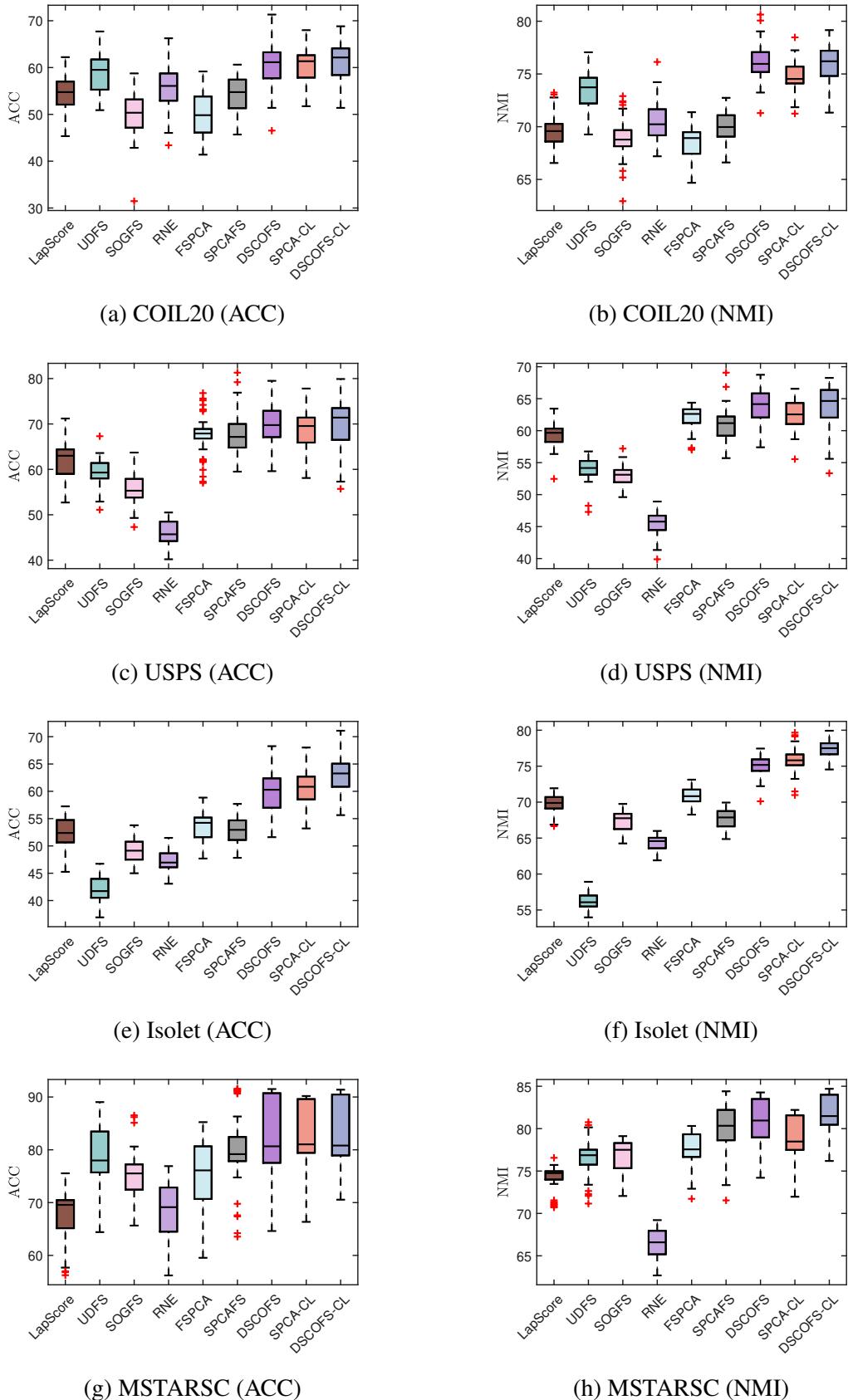


图 4.10 对比方法在四个真实数据集上的模型稳定性

Figure 4.10 The model stability of comparison methods on four real-world datasets

4.3.4.5 数据可视化分析

为了直观地观察数据的分布以及理解聚类的结果，本实验采用 t-随机邻近嵌入 (t-distributed Stochastic Neighbor Embedding, t-SNE)^[101] 技术展示特征子空间的数据分布。t-SNE 是一种高维数据降维的非线性技术，常用于在二维或三维的低维空间中表示高维数据集，从而实现数据可视化。首先通过特征选择得到数据子集，然后利用 t-SNE 技术将数据降维至二维，最后通过散点图呈现数据的低维分布。从图 4.11 可以观察到，在低维空间中数据呈现明显的聚类特征，因此使用聚类来检验特征选择的性能是合理的。

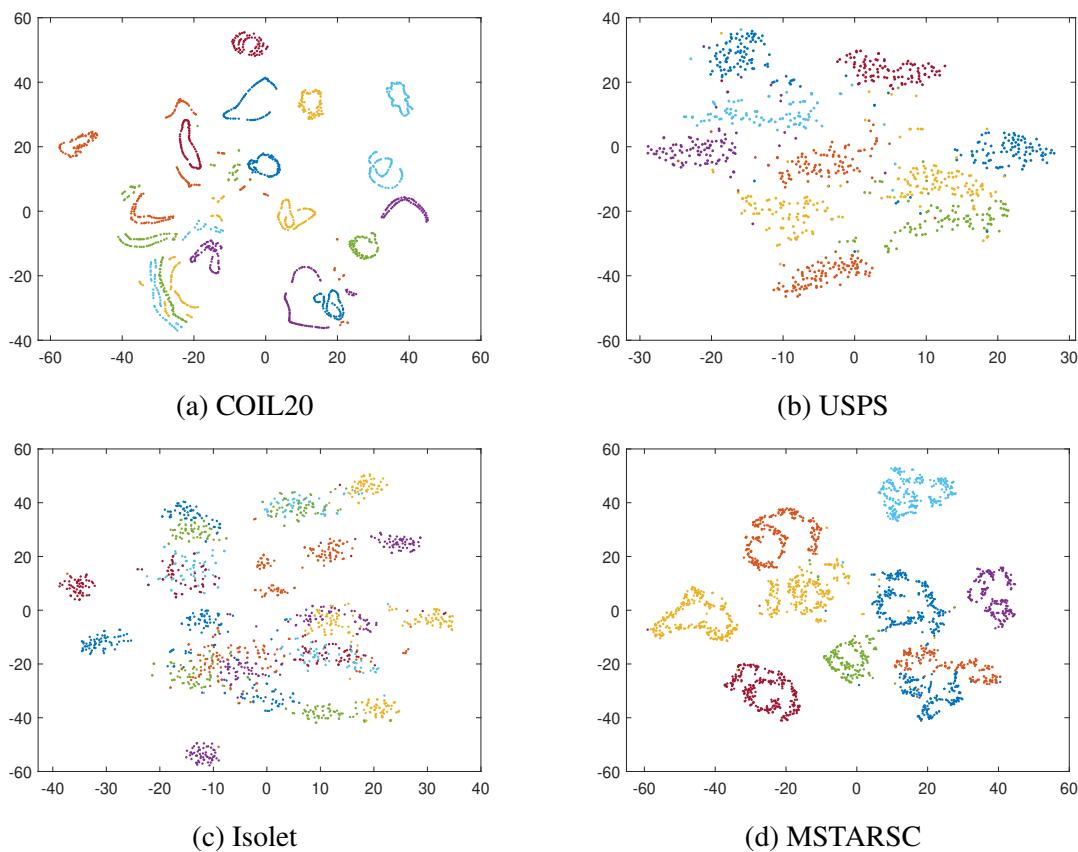


图 4.11 t-SNE 在四个真实数据集上的可视化结果

Figure 4.11 The visualization results of t-SNE on four real-world datasets

4.4 本章小结

本章将双稀疏约束和对比学习融合，提出了一种新的无监督特征选择模型，称为 DSCOFS-CL。相较于传统基于欧氏距离度量的重构误差方法，DSCOFS-CL 融合

对比学习策略，能够更充分地挖掘样本之间的关系。同时，利用双稀疏约束表示稀疏结构，使得图学习可以在低维空间中更有效捕捉数据特性。此外，通过对自表示矩阵施加低秩约束，模型能够保留图的全局结构。在算法方面，结合梯度下降法和硬阈值设计了 PAM 优化算法，并证明了算法的收敛性。最后，通过在真实数据集上的大量数值实验，验证了对比学习损失作为重构误差度量的有效性。具体地，相较于上一章验证过优良性能的 DSCOFS，本章提出的 DSCOFS-CL 在六个真实数据集上的平均 ACC 和 NMI 进一步分别提升了 1.85% 和 1.06%。而对于同样融合了对比学习的 SPCA-CL，平均 ACC 和 NMI 分别提升了 1.7% 和 1.51%。此外，消融实验结果还表明，自表达框架下原始空间数据结构也是指导无监督特征选择的关键。

第五章 总结与展望

5.1 总结

无监督特征选择作为一种不依赖标签的降维技术，近年来在模式识别、机器学习、统计学与自动化等领域引发了广泛的关注。稀疏优化的蓬勃发展为无监督特征选择注入了新的活力，合理利用稀疏性可以提高变量的可解释性，更有效地进行无监督特征选择。针对稀疏结构表示不充分、局部特征辨别不准确和重构误差度量不合理的问题，本文从单稀疏约束、双稀疏约束和双稀疏约束对比学习三个方面有序展开，逐层深入，建立了无监督特征选择的新方法，取得了重要的研究成果。

首先，探讨了 $\ell_{2,0}$ 范数约束和正则的无监督特征选择模型。通过实验发现，相较于传统松弛方法，SPCA-R 和 SPCA-C 展现出较为明显的性能优势，这验证了 $\ell_{2,0}$ 范数的有效性。进一步，SPCA-R 的阈值会随正则化参数改变，可能会导致完全稀疏或者完全不稀疏的情况，而 SPCA-C 通过 $\ell_{2,0}$ 范数约束避免了上述情况的出现。值得注意的是，通过与 FSPCA 的对比分析，揭示了该模型对优化算法设计的敏感性。

其次，提出了 $\ell_{2,0}$ 范数和 ℓ_0 范数双稀疏约束的 DSCOFS 模型。其中 $\ell_{2,0}$ 范数确保了全局结构的稀疏性并具有更好的可解释性，而 ℓ_0 范数则考虑了数据元素的局部个体稀疏性。设计了有效的 PAM 优化算法，并且严格证明了算法的收敛性。大量的数值实验表明了 DSCOFS 的优越性。更重要的是，通过设计新的特征选择评价指标 FSR，验证了 ℓ_0 范数对无监督特征选择的有效性。

最后，构建了融合 DSCOFS 和对比学习的 DSCOFS-CL 模型。这里对比学习考虑到样本之间的关系，通过在原始空间与投影空间联合学习最优图结构，实现了数据全局与局部分布的自适应表达。设计了有效的 PAM 优化算法，并证明了收敛性。大量的数值实验证明了 DSCOFS 和对比学习结合的有效性，进一步提升了无监督特征选择的性能。特别地，消融实验表明了 DSCOFS-CL 自表达框架下原始空间数据结构的重要性。

5.2 展望

尽管本文基于稀疏优化在无监督特征选择方面取得了一定的进展，但仍有若干问题亟待深入探索。

首先，如何建立更一般的稀疏结构。双稀疏可以看作连接单稀疏与实际稀疏的桥梁，然而实际场景中可能存在更为复杂的稀疏结构，双稀疏仍然不足以学习到完整的稀疏结构。因此，未来研究需探讨基于可学习范数的自适应稀疏表征框架，建立更一般的稀疏先验模型。

其次，如何有效地进行参数选择。对于所提出的 DSCOFS 和 DSCOFS-CL 模型，本文采用了手动网格搜索策略。然而随着参数量的增大，手动调优变得异常困难。受深度展开网络的启发，可以把传统的优化的迭代展开为神经网络的网络层，然后通过神经网络的反向传播和梯度下降自动地调整参数。当然，也可以考虑使用强化学习和大语言模型等策略。因此，模型参数自适应选择也是一个很重要的研究方向。

最后，如何设计高效的优化策略。虽然本文基于一阶方法和硬阈值设计了有效的数值算法，并且严格证明了收敛性，但面对样本过大的数据集时难以高效进行科学计算。因此，未来可以考虑充分挖掘稀疏约束的结构特点，开发适合大规模计算的二阶牛顿法、随机优化策略和量子算法等。

插图索引

图 1.1	关于无监督特征选择主题的文章统计	2
图 1.2	无监督特征选择方法框架	3
图 1.3	不同 p 取值下的 ℓ_p 范数约束区域	5
图 1.4	文章主要内容与章节安排	7
图 2.1	原始数据的投影过程	11
图 2.2	数据集可视化结果	18
图 2.3	对比方法在六个真实数据集上的 ACC (%) 曲线	20
图 2.4	对比方法在六个真实数据集上的 NMI (%) 曲线	21
图 2.5	SPCA-R 和 SPCA-C 在两个真实数据集上的稀疏度变化曲线	25
图 2.6	SPCA-R 在两个真实数据集上的参数敏感度分析结果	26
图 2.7	SPCA-C 在两个真实数据集上的参数敏感度分析结果	26
图 2.8	SPCA-R 和 SPCA-C 在两个真实数据集上的收敛曲线	27
图 3.1	不同稀疏约束下获得的结果示例	30
图 3.2	DSCOFS 特征选择和聚类的流程图	31
图 3.3	合成数据集的原始分布和特征选择结果	42
图 3.4	对比方法在八个真实数据集上的 ACC (%) 曲线	44
图 3.5	对比方法在八个真实数据集上的 NMI (%) 曲线	45
图 3.6	稀疏投影矩阵 X 在四个真实数据集上的可视化	49
图 3.7	ACC 指标下 DSCOFS 的后验 Nemenyi 检验结果	51
图 3.8	DSCOFS 在四个真实数据集上的参数敏感度分析结果	52

图 3.9 DSCOFS 在四个真实数据集上的收敛曲线.....	53
图 3.10 对比方法在四个真实数据集上的模型稳定性.....	54
图 4.1 不同度量得到的投影空间	58
图 4.2 DSCOFS-CL 特征选择的流程图.....	60
图 4.3 对比方法在六个真实数据集上的 ACC (%) 曲线	69
图 4.4 对比方法在六个真实数据集上的 NMI (%) 曲线.....	70
图 4.5 两个真实数据集上相似度矩阵 S 的可视化	74
图 4.6 DSCOFS-CL 在两个真实数据集上的消融实验结果.....	74
图 4.7 ACC 指标下 DSCOFS-CL 的后验 Nemenyi 检验结果.....	75
图 4.8 DSCOFS-CL 在 USPS 数据集上的参数敏感度分析结果	77
图 4.9 DSCOFS-CL 在四个真实数据集上的收敛曲线	78
图 4.10 对比方法在四个真实数据集上的模型稳定性.....	79
图 4.11 t-SNE 在四个真实数据集上的可视化结果	80

表格索引

表 2.1 数据集信息.....	17
表 2.2 对比方法在六个真实数据集上的 ACC (平均值% ± 标准差%) 结果	22
表 2.3 对比方法在六个真实数据集上的 NMI (平均值% ± 标准差%) 结果	23
表 2.4 SPCA-R 和 FSPCA 在六个真实数据集上的 ACC (%) 和 NMI (%)	28
表 3.1 数据集信息.....	41
表 3.2 对比方法在八个真实数据集上的 ACC (平均值% ± 标准差%) 结果	46
表 3.3 对比方法在八个真实数据集上的 NMI (平均值% ± 标准差%) 结果	47
表 3.4 在八个真实数据集上消融实验的 ACC(%)、NMI(%) 和 FSR(%).....	48
表 3.5 ACC 指标下 DSCOFS 的 Friedman 检验结果	51
表 3.6 DSCOFS 和 TSFS+TSNE 在四个真实数据集上的 ACC (%) 和 NMI (%)	55
表 4.1 数据集信息.....	68
表 4.2 对比方法在六个真实数据集上的 ACC (平均值% ± 标准差%) 结果	71
表 4.3 对比方法在六个真实数据集上的 NMI (平均值% ± 标准差%) 结果	72
表 4.4 ACC 指标下 DSCOFS-CL 的 Friedman 检验结果.....	75

参考文献

- [1] 胡耀华, 李昱帆, 刘艳艳, 等. 结构稀疏优化模型的理论与算法[J]. 中国科学: 数学, 2024, 54(7): 1045-1070.
- [2] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
- [3] Zhuang H, Lin Z, Yang Y, et al. An analytic formulation of convolutional neural network learning for pattern recognition[J]. Information Sciences, 2025, 686: 121317.
- [4] Bracco A, Brajard J, Dijkstra H A, et al. Machine learning for the physics of climate [J]. Nature Reviews Physics, 2025, 7(1): 6-20.
- [5] Marshoodulla S Z, Saha G. A survey of data mining methodologies in the environment of IoT and it's variants[J]. Journal of Network and Computer Applications, 2024: 103907.
- [6] Maghini D G, Dvorak M, Dahlen A, et al. Quantifying bias introduced by sample collection in relative and absolute microbiome measurements[J]. Nature Biotechnology, 2024, 42(2): 328-338.
- [7] Ge W, Cui Z, Wang J, et al. Metacluster: A universal interpretable classification framework for cybersecurity[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 3829-3843.
- [8] Bai Y, Xu Y, Chen S, et al. Tops-speed complex-valued convolutional accelerator for feature extraction and inference[J]. Nature Communications, 2025, 16(1): 292.
- [9] Liao H, Chen H, Yin T, et al. A general adaptive unsupervised feature selection with auto-weighting[J]. Neural Networks, 2025, 181: 106840.
- [10] Li G, Yu Z, Yang K, et al. Exploring feature selection with limited labels: A comprehensive survey of semi-supervised and unsupervised approaches[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 6124-6144.
- [11] Li Y, Hu L, Gao W. Multi-label feature selection with high-sparse personalized and low-redundancy shared common features[J]. Information Processing & Management, 2024, 61(3): 103633.

- [12] Solorio-Fernández S, Carrasco-Ochoa J A, Martínez-Trinidad J F. A review of unsupervised feature selection methods[J]. Artificial Intelligence Review, 2020, 53(2): 907-948.
- [13] Huang P, Kong Z, Xie M, et al. Robust unsupervised feature selection via data relationship learning[J]. Pattern Recognition, 2023, 142: 109676.
- [14] Uddin M P, Mamun M A, Afjal M I, et al. Information-theoretic feature selection with segmentation-based folded principal component analysis (PCA) for hyperspectral image classification[J]. International Journal of Remote Sensing, 2021, 42(1): 286-321.
- [15] Liu Z, Yang Y, Gao F, et al. Deep unsupervised learning for joint antenna selection and hybrid beamforming[J]. IEEE Transactions on Communications, 2022, 70(3): 1697-1710.
- [16] Saberi-Movahed F, Rostami M, Berahmand K, et al. Dual regularized unsupervised feature selection based on matrix factorization and minimum redundancy with application in gene selection[J]. Knowledge-Based Systems, 2022, 256: 109884.
- [17] Capó M, Pérez A, Lozano J A. A cheap feature selection approach for the k-means algorithm[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(5): 2195-2208.
- [18] El Aboudi N, Benhlima L. Review on wrapper feature selection approaches[C]//2016 International Conference on Engineering MIS (ICEMIS). IEEE, 2016: 1-5.
- [19] Wu J S, Song M X, Min W, et al. Joint adaptive manifold and embedding learning for unsupervised feature selection[J]. Pattern Recognition, 2021, 112: 107742.
- [20] He X, Cai D, Niyogi P. Laplacian score for feature selection[J]. Advances in Neural Information Processing Systems, 2005, 18.
- [21] Liu H, Setiono R. Feature selection and classification—a probabilistic wrapper approach[M]//Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. CRC Press, 2022: 419-424.

- [22] Li Z, Nie F, Bian J, et al. Sparse PCA via $\ell_{2,p}$ -norm regularization for unsupervised feature selection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 5322-5328.
- [23] Cai D, Zhang C, He X. Unsupervised feature selection for multi-cluster data[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010: 333-342.
- [24] Yang Y, Shen H T, Ma Z, et al. $\ell_{2,1}$ -norm regularized discriminative feature selection for unsupervised learning[C]//IJCAI International Joint Conference on Artificial Intelligence. 2011: 1589-1594.
- [25] Li Z, Yang Y, Liu J, et al. Unsupervised feature selection using nonnegative spectral analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 26. 2012: 1026-1032.
- [26] Liu Y, Ye D, Li W, et al. Robust neighborhood embedding for unsupervised feature selection[J]. Knowledge-Based Systems, 2020, 193: 105462.
- [27] Nie F, Zhu W, Li X. Unsupervised feature selection with structured graph optimization [C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 30. 2016.
- [28] Li X, Zhang H, Zhang R, et al. Generalized uncorrelated regression with adaptive graph for unsupervised feature selection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 30(5): 1587-1595.
- [29] Shi D, Zhu L, Li J, et al. Unsupervised adaptive feature selection with binary hashing [J]. IEEE Transactions on Image Processing, 2023, 32: 838-853.
- [30] Tang C, Liu X, Li M, et al. Robust unsupervised feature selection via dual self-representation and manifold regularization[J]. Knowledge-Based Systems, 2018, 145: 109-120.
- [31] Chen H, Nie F, Wang R, et al. Unsupervised feature selection with flexible optimal graph[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(2): 2014-2027.

- [32] Zhou Q, Wang Q, Gao Q, et al. Unsupervised discriminative feature selection via contrastive graph learning[J]. *IEEE Transactions on Image Processing*, 2024, 33: 972-986.
- [33] Abdi H, Williams L J. Principal component analysis[J]. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2010, 2(4): 433-459.
- [34] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis[J]. *Journal of Computational and Graphical Statistics*, 2006, 15(2): 265-286.
- [35] Chang X, Nie F, Yang Y, et al. Convex sparse PCA for unsupervised feature learning [J]. *ACM Transactions on Knowledge Discovery from Data*, 2016, 11(1): 1-16.
- [36] Yi S, He Z, Jing X, et al. Adaptive weighted sparse principal component analysis for robust unsupervised feature selection[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 31(6): 2153-2163.
- [37] Zheng J, Zhang X, Liu Y, et al. Fast sparse pca via positive semidefinite projection for unsupervised feature selection[J]. 2023, arXiv:2309.06202.
- [38] Zhou Q, Gao Q, Wang Q, et al. Sparse discriminant PCA based on contrastive learning and class-specificity distribution[J]. *Neural Networks*, 2023, 167: 775-786.
- [39] Boileau P, Hejazi N S, Dudoit S. Exploring high-dimensional biological data with sparse contrastive principal component analysis[J]. *Bioinformatics*, 2020, 36(11): 3422-3430.
- [40] Pang T, Nie F, Han J, et al. Efficient feature selection via $\ell_{2,0}$ -norm constrained sparse regression[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2019, 31(5): 880-893.
- [41] Wang J, Wang H, Nie F, et al. Sparse feature selection via fast embedding spectral analysis[J]. *Pattern Recognition*, 2023, 139: 109472.
- [42] Chen H, Nie F, Wang R, et al. Fast unsupervised feature selection with bipartite graph and $\ell_{2,0}$ -norm constraint[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 35(5): 4781-4793.

- [43] Nie F, Tian L, Wang R, et al. Learning feature-sparse principal subspace[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 4858-4869.
- [44] Zhang C, Fang Y, Liang X, et al. Efficient multi-view unsupervised feature selection with adaptive structure learning and inference[C]//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24). 2024: 5443-5452.
- [45] Donoho D. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306.
- [46] 赵晨, 罗自炎, 修乃华. 稀疏优化理论与算法若干新进展[J]. 运筹学学报, 2020, 24(4): 1-24.
- [47] 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.
- [48] Attouch H, Bolte J, Redont P, et al. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality[J]. Mathematics of Operations Research, 2010, 35(2): 438-457.
- [49] 王锐, 修乃华. 稀疏优化二阶算法研究进展[J]. 数值计算与计算机应用, 2022, 43 (3): 314-328.
- [50] Zhou S, Xiu N, Qi H D. Global and quadratic convergence of Newton hard-thresholding pursuit[J]. Journal of Machine Learning Research, 2021, 22(12): 1-45.
- [51] Zhao C, Xiu N, Qi H, et al. A Lagrange–Newton algorithm for sparse nonlinear programming[J]. Mathematical Programming, 2022, 195(1): 903-928.
- [52] Milzarek A, Xiao X, Cen S, et al. A stochastic semismooth Newton method for nonsmooth nonconvex optimization[J]. SIAM Journal on Optimization, 2019, 29(4): 2916-2948.
- [53] Xiang M, Zhang Z. Fast recursive greedy methods for sparse signal recovery[J]. IEEE Transactions on Signal Processing, 2024, 72: 4381-4394.
- [54] Zhang C, Li H, Qian Y, et al. Locality-constrained discriminative matrix regression for robust face identification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(3): 1254-1268.

- [55] Zhou S. Sparse SVM for sufficient data reduction[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(9): 5560-5571.
- [56] Xiu X, Miao Z, Yang Y, et al. Deep canonical correlation analysis using sparsity-constrained optimization for nonlinear process monitoring[J]. *IEEE Transactions on Industrial Informatics*, 2022, 18(10): 6690-6699.
- [57] Nikolova M. Relationship between the optimal solutions of least squares regularized with ℓ_0 -norm and constrained by k -sparsity[J]. *Applied and Computational Harmonic Analysis*, 2016, 41(1): 237-265.
- [58] Chen H, Sun Y, Gao J, et al. Solving partial least squares regression via manifold optimization approaches[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 30(2): 588-600.
- [59] Chen S, Ma S, Xue L, et al. An alternating manifold proximal gradient method for sparse principal component analysis and sparse canonical correlation analysis[J]. *INFORMS Journal on Optimization*, 2020, 2(3): 192-208.
- [60] Xiao N, Liu X, Yuan Y x. Exact penalty function for $\ell_{2,1}$ -norm minimization over the Stiefel manifold[J]. *SIAM Journal on Optimization*, 2021, 31(4): 3097-3126.
- [61] Breloy A, Kumar S, Sun Y, et al. Majorization-minimization on the Stiefel manifold with application to robust sparse PCA[J]. *IEEE Transactions on Signal Processing*, 2021, 69: 1507-1520.
- [62] Zeng J, Lin S, Wang Y, et al. $\ell_{1/2}$ -regularization: Convergence of iterative half thresholding algorithm[J]. *IEEE Transactions on Signal Processing*, 2014, 62(9): 2317-2329.
- [63] Chen W, Ji H, You Y. An augmented lagrangian method for ℓ_1 -regularized optimization problems with orthogonality constraints[J]. *SIAM Journal on Scientific Computing*, 2016, 38(4): B570-B592.
- [64] Chen S, Ma S, Man-Cho So A, et al. Proximal gradient method for nonsmooth optimization over the Stiefel manifold[J]. *SIAM Journal on Optimization*, 2020, 30(1): 210-239.

- [65] Zhou Y, Bao C, Ding C, et al. A semismooth Newton based augmented Lagrangian method for nonsmooth optimization on matrix manifolds[J]. Mathematical Programming, 2023, 201(1): 1-61.
- [66] Fung G, Mangasarian O. Equivalence of minimal ℓ_0 -norm and ℓ_p -norm solutions of linear equalities, inequalities and linear programs for sufficiently small p [J]. Journal of Optimization Theory and Applications, 2011, 151: 1-10.
- [67] Bertsimas D, Cory-Wright R, Pauphilet J. Solving large-scale sparse PCA to certifiable (near) optimality[J]. Journal of Machine Learning Research, 2022, 23(13): 1-35.
- [68] Xiu X, Miao Z, Liu W. A sparsity-aware fault diagnosis framework focusing on accurate isolation[J]. IEEE Transactions on Industrial Informatics, 2023, 19(2): 1356-1365.
- [69] Nie F, Chen Q, Yu W, et al. Row-sparse principal component analysis via coordinate descent metho[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3460-3471.
- [70] Zou H, Xue L. A selective overview of sparse principal component analysis[J]. Proceedings of the IEEE, 2018, 106(8): 1311-1320.
- [71] Zhang X, Zheng J, Wang D, et al. Structured sparsity optimization with non-convex surrogates of $\ell_{2,0}$ -norm: A unified algorithmic framework[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(5): 6386-6402.
- [72] Xiu X, Kong L, Li Y, et al. Iterative reweighted methods for ℓ_1 - ℓ_p minimization[J]. Computational Optimization and Applications, 2018, 70(1): 201-219.
- [73] Absil P A, Mahony R, Sepulchre R. Optimization algorithms on matrix manifolds [M]. Princeton University Press, 2008.
- [74] Bai H, Huang M, Zhong P. Precise feature selection via non-convex regularized graph embedding and self-representation for unsupervised learning[J]. Knowledge-Based Systems, 2024, 296: 111900.
- [75] Wang Z, Li Q, Zhao H, et al. Simultaneous local clustering and unsupervised feature selection via strong space constraint[J]. Pattern Recognition, 2023, 142: 109718.

- [76] Xiu X, Pan L, Yang Y, et al. Efficient and fast joint sparse constrained canonical correlation analysis for fault detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(3): 4153-4163.
- [77] Cui L, Bai L, Wang Y, et al. Fused lasso for feature selection using structural information[J]. Pattern Recognition, 2021, 119: 108058.
- [78] Liu J, Feng M, Xiu X, et al. Towards robust and sparse linear discriminant analysis for image classification[J]. Pattern Recognition, 2024, 153: 110512.
- [79] Bian X, Xu W, Wang Y, et al. Joint compressed signal recovery and ris diagnosis via double-sparsity optimization[J]. IEEE Internet of Things Journal, 2024, 11(8): 13327-13339.
- [80] Zhang S, Liu Y, Li X. Micro-doppler effects removed sparse aperture isar imaging via low-rank and double sparsity constrained ADMM and linearized admm[J]. IEEE Transactions on Image Processing, 2021, 30: 4678-4690.
- [81] Zhou S, Luo Z, Xiu N, et al. Computing one-bit compressive sensing via double-sparsity constrained optimization[J]. IEEE Transactions on Signal Processing, 2022, 70: 1593-1608.
- [82] Guo Y, Sun Y, Wang Z, et al. Double-structured sparsity guided flexible embedding learning for unsupervised feature selection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(10): 13354-13367.
- [83] Wang H, Nie F, Huang H, et al. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance[C]//2011 International Conference on Computer Vision. IEEE, 2011: 557-562.
- [84] Hu Y, Liu J, Gao Y, et al. DSTPCA: Double-sparse constrained tensor principal component analysis method for feature selection[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(4): 1481-1491.
- [85] Sun J, Kong L, Zhou S. Gradient projection Newton algorithm for sparse collaborative learning using synthetic and real datasets of applications[J]. Journal of Computational and Applied Mathematics, 2023, 422: 114872.

- [86] Xiao N, Liu X, Yuan Y. A class of smooth exact penalty function methods for optimization problems with orthogonality constraints[J]. Optimization Methods and Software, 2022, 37(4): 1205-1241.
- [87] Gao B, Liu X, Chen X, et al. A new first-order algorithmic framework for optimization problems with orthogonality constraints[J]. SIAM Journal on Optimization, 2018, 28 (1): 302-332.
- [88] Huang Y, Dai Y, Liu X. Equipping the Barzilai–Borwein method with the two dimensional quadratic termination property[J]. SIAM Journal on Optimization, 2021, 31(4): 3068-3096.
- [89] Blumensath T, Davies M E. Iterative hard thresholding for compressed sensing[J]. Applied and Computational Harmonic Analysis, 2009, 27(3): 265-274.
- [90] Bolte J, Sabach S, Teboulle M. Proximal alternating linearized minimization for non-convex and nonsmooth problems[J]. Mathematical Programming, 2014, 146(1): 459-494.
- [91] Mirzaei A, Pourahmadi V, Soltani M, et al. Deep feature selection using a teacher-student network[J]. Neurocomputing, 2020, 383: 396-408.
- [92] Ke Q, Kanade T. Robust ℓ_1 -norm factorization in the presence of outliers and missing data by alternative convex programming[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition: Vol. 1. IEEE, 2005: 739-746.
- [93] Ng A Y. Feature selection, ℓ_1 vs. ℓ_2 regularization, and rotational invariance[C]// Proceedings of the Twenty-first International Conference on Machine Learning. 2004: 78.
- [94] Wang Q, Gao Q, Gao X, et al. $\ell_{2,p}$ -norm based PCA for image recognition[J]. IEEE Transactions on Image Processing, 2018, 27(3): 1336-1346.
- [95] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International Conference on Machine Learning. PMLR, 2020: 1597-1607.

- [96] Hu H, Wang X, Zhang Y, et al. A comprehensive survey on contrastive learning[J]. *Neurocomputing*, 2024: 128645.
- [97] Chen T, Sun Y, Shi Y, et al. On sampling strategies for neural network-based collaborative filtering[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 767-776.
- [98] Wu Z, Xiong Y, Yu S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3733-3742.
- [99] Rockafellar R T, Wets R J B. Variational analysis: Vol. 317[M]. Springer Science & Business Media, 2009.
- [100] Attouch H, Bolte J, Svaiter B F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods[J]. *Mathematical Programming*, 2013, 137(1): 91-129.
- [101] Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. *Journal of Machine Learning Research*, 2008, 9: 2579-2605.

攻读硕士学位期间取得的研究成果

一、学术论文

- [1] **Yang A**, Li X, Xiu X. Sparse PCA via $\ell_{2,0}$ -norm constrained optimization for unsupervised feature selection[C]//2024 43rd Chinese Control Conference (CCC). IEEE, 2024: 7375-7379. (已发表, EI 会议, 对应论文第二章)
- [2] Xiu X, **Yang A**, Huang C, et al. Enhancing unsupervised feature selection via double sparsity constrained optimization[J]. Applied Mathematical Modelling. (审稿中, SCI 一区, 对应论文第三章)
- [3] 修贤超, **杨安宁**, 李鑫荣. 基于对比学习的稀疏低秩无监督特征选择方法 [J]. 运筹学学报. (审稿中, 中文核心, 对应论文第四章)

二、知识产权

- [1] 修贤超, **杨安宁**, 柳春. 基于 RNN 的互补双残差生成器的故障监测方法和系统: CN117194880B[P]. 2025-02-11. (已授权)
- [2] 修贤超, **杨安宁**, 孙腾. 基于典型相关分析快速算法的多视角数据分类方法和系统: CN117312953B[P]. 2025-04-11. (已授权)

三、参与项目

- [1] 国家自然科学基金面上项目, 大规模黎曼流形稀疏优化算法及应用, 2024.01-2027.12.
- [2] 上海大学青年英才启航计划, 数据驱动故障诊断的优化模型与算法, 2022.01-2024.12.

致 谢

现在是 2024 年，4 月 15 日，随着最终论文敲定的最后一个字落下，三年的求学也即将告一段落。回头观望记录自己成果的论文，却似乎写不出一丝自己的三年的经历。三年过的很快，仿佛不久前还在考场上为自己的考研梦而奋斗；三年时间也很慢，脑海里播不完自己一幕幕喜忧交杂的片段。求学之路漫漫，从最初的小乡村走到如今的大都市，曾经的自己可能没想到如今的自己会在样一个阶段，这样一个让我收获颇多、对未来充满期待的阶段，这样一个给予我在人生道路上前进的养料的阶段，这样一个何其有幸的阶段。论文写不尽我的三年，就像此刻的我也说不清道不明心中的情绪。

首先我要感谢我的导师修贤超老师，在循循善诱、不厌其烦的教诲下，让我从懵懂走向小有所成，让我在我的科研白纸上写下了属于自己的诗篇。您严谨务实的学术态度令我受益良多，让我逐渐成为一个更加认真严谨的人，相信您的言传身教会在我未来的道路上持续指引我前行。

然后我要感谢我的父母，不仅是此刻更是时时刻刻，从小你们就支持我为自己的梦想去努力，告诫我人要以诚为本，鼓励我面对失败虽然有遗憾但是不要被轻易打败。现在的我也可以告诉我的父母，我做到了，我走到了大都市闯荡了一番并且我也会带着全新的自己走回那个小乡村。

我还想感谢陪我走过三年的所有人，无论是室友、同学、师弟、师兄还是那些一瞬间的路人，都在我三年的画卷里留下了独特的痕迹。

三年复三年，人生能有几个三年？既然这个三年即将画上句号，那下一个三年也将是新的起点，再次回看这记录自己成果的论文，似乎也是一个人生的笔触，在我这个三年上画上圆满的句点。最后我想祝福我自己，希望未来的旅途永远坦荡，一生永远纯粹善良。

杨安宁

机自大楼 345

2025 年 04 月 27 日

附录 A 本文英文缩写对照表^①

拉普拉斯评分	Laplace Score, LapScore
拉斯维加斯包裹式	Las Vegas Wrapper, LVW
无监督判别特征选择	Unsupervised Discriminative Feature Selection, UDFS
鲁棒的邻域嵌入法	Robust Neighborhood Embedding, RNE
结构化最优图特征选择	Structured Optimal Graph Feature Selection, SOGFS
主成分分析	Principal Component Analysis, PCA
稀疏主成分分析	Sparse PCA, SPCA
半正定投影实现稀疏主成分分析	SPCA via Positive Semidefinite projection, SPCA-PSD
用于特征选择的稀疏主成分分析	SPCA for Feature Selection, SPCAFS
特征稀疏性约束主成分分析	Feature-Sparsity constrained PCA, FSPCA
压缩感知	Compressed Sensing, CS
非确定性多项式	Non-deterministic Polynomial, NP
交替最小化算法	Alternating Minimization Algorithm, AMA
交替方向乘子法	Alternating Direction Method of Multipliers, ADMM
近端交替最小化	Proximal Alternating Minimization, PAM
$\ell_{2,0}$ 范数正则的稀疏主成分分析	SPCA with $\ell_{2,0}$ -norm Regularization, SPCA-R
$\ell_{2,0}$ 范数约束的稀疏主成分分析	SPCA with $\ell_{2,0}$ -norm Constraint, SPCA-C
双稀疏约束优化特征选择	Double Sparsity Constrained Optimization Feature Selection, DSCOFS
融合对比学习的双稀疏约束优化特征选择	DSCOFS with Contrastive Learning, DSCOFS-CL
库恩-蒙克雷斯	Kuhn-Munkres, KM
准确率	Accuracy, ACC
归一化互信息	Normalized Mutual Information, NMI

^① 按文中第一次出现的先后顺序排列

特征相似率

Feature Similarity Rate, FSR

临界差异

Critical Difference, CD

t-随机邻近嵌入

t-distributed Stochastic Neighbor Embedding, t-SNE

融合对比学习的稀疏主成分分析 SPCA with Contrastive Learning, SPCA-CL