

控制与决策

Control and Decision

基于缺失值学习和结构保留的不平衡不完备多视图聚类

陈梅, 郭爱霞, 王洁, 杨嘉怡, 詹苏宝

引用本文:

陈梅, 郭爱霞, 王洁, 等. 基于缺失值学习和结构保留的不平衡不完备多视图聚类[J]. 控制与决策, 2025, 40(9): 2901–2912.

在线阅读 View online: <https://doi.org/10.13195/j.kzyjc.2025.0191>

您可能感兴趣的其他文章

Articles you may be interested in

[基于Hessian正则的自适应损失半监督特征选择](#)

Adaptive loss semi-supervised feature selection based on Hessian regularization

控制与决策. 2021, 36(8): 1862–1870 <https://doi.org/10.13195/j.kzyjc.2019.1510>

[小样本下多稀疏表示分类器的决策融合方法](#)

Decision fusion of multiple sparse representation-based classifiers in case of small samples

控制与决策. 2021, 36(8): 1984–1990 <https://doi.org/10.13195/j.kzyjc.2019.1839>

[嵌入重采样技术的C4.5决策树集成分类算法的临床医学预测](#)

Clinical prediction of C4.5 decision tree classification algorithm with embedded resampling technique

控制与决策. 2021, 36(6): 1342–1350 <https://doi.org/10.13195/j.kzyjc.2019.1247>

[乘型一致性毕达哥拉斯模糊偏好关系](#)

Multiplicative consistent Pythagorean fuzzy preference relation

控制与决策. 2021, 36(4): 1010–1016 <https://doi.org/10.13195/j.kzyjc.2019.0967>

[基于边缘峰度度量的特征缩减模糊聚类算法](#)

Feature-reduction fuzzy clustering algorithm based on marginal kurtosis measure

控制与决策. 2021, 36(11): 2665–2673 <https://doi.org/10.13195/j.kzyjc.2020.0220>

基于缺失值学习和结构保留的不平衡不完备多视图聚类

陈梅[†], 郭爱霞, 王洁, 杨嘉怡, 詹苏宝

(兰州交通大学 电子与信息工程学院, 兰州 730000)

摘要: 现有的不完备多视图聚类算法虽然取得了一定的进展, 但是仍然存在局限性: 1) 无法准确挖掘缺失数据的潜在信息, 特别是在各视图缺失率不一致的情况下; 2) 难以同时保留数据的全局结构和局部结构; 3) 无法有效挖掘不同视图的高阶相关性和互补信息. 为解决这些问题, 提出基于缺失值学习和结构保留的不平衡不完备多视图聚类算法. 首先, 算法通过线性投影将原始高维数据映射至低维空间; 然后, 结合基于近邻关系学习的补全矩阵和完整性约束机制对缺失值进行填充, 从而确保填充的缺失值尽可能保持数据的原始结构; 接着, 算法采用子空间聚类技术、超拉普拉斯正则化和加权张量 Schatten- p 范数, 有效捕获数据的全局结构、局部结构以及高阶相关性; 最后, 与 10 个对比算法在多种缺失率的 8 个仿真不完备多视图数据集上进行对比实验, 实验结果表明所提出方法的性能优于其他对比算法.

关键词: 不平衡不完备多视图聚类; 缺失值学习; 结构保留; 加权张量 Schatten- p 范数; 完整性约束

中图分类号: TP311 文献标志码: A

DOI: 10.13195/j.kzyjc.2025.0191

引用格式: 陈梅, 郭爱霞, 王洁, 等. 基于缺失值学习和结构保留的不平衡不完备多视图聚类 [J]. 控制与决策, 2025, 40(9): 2901-2912.

Unbalanced incomplete multi-view clustering based on missing-value learning and structure preservation

CHEN Mei[†], GUO Ai-xia, WANG Jie, YANG Jia-yi, ZHAN Su-bao

(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730000, China)

Abstract: Existing incomplete multi-view clustering algorithms have made some progress, however, they have certain limitations: 1) Most existing methods fail to accurately uncover the latent information of missing data, especially when handling varying missing rate across different views; 2) They cannot preserve both global and local structures of data; 3) They fail to effectively extract high-order correlations and complementary information between ones. To address these issues, the paper proposes an unbalanced incomplete multi-view clustering algorithm based on missing-value learning and structure preservation. The algorithm maps the original high-dimensional data to a low-dimensional space using linear projection, and combines a completion matrix based on neighborhood relationships and an integrity constraint to fill in the missing values, ensuring that the filled missing values closely match the original structural features of the data. Additionally, the algorithm employs subspace clustering techniques, hyper-Laplacian regularization, and the weighted tensor Schatten- p norm, to simultaneously capture the global and local structure and high-order correlation. Experimental results with ten benchmark algorithms on eight simulated incomplete multi-view datasets with various missing values show that the performance of the proposed algorithm is superior to the comparison algorithms.

Keywords: unbalanced incomplete multi-view clustering; missing-value learning; structure preserved; the weighted tensor Schatten- p norm; integrity constraint

0 引言

聚类技术作为一种无监督学习方法, 旨在将数据集划分为由相似数据点构成的簇, 以挖掘数据中的潜在结构和信息^[1]. 随着信息采集技术的进步, 研究者能够从多个角度获取同一对象的多样化数据表

示, 形成多视图数据. 相较于传统的单视图聚类, 多视图聚类通过挖掘视图间的一致性和互补性信息, 提升了聚类性能^[2-3]. 然而, 在实际应用中, 由于机器故障、人为因素或隐私限制等, 收集到的多视图数据常存在信息缺失, 产生了不完备多视图聚类问题

收稿日期: 2025-02-25; 录用日期: 2025-05-03.

基金项目: 国家自然科学基金项目 (62266029); 甘肃省重点研发计划项目 (24YFGA036); 甘肃省高等学校产业支撑计划项目 (2022CYZC-36).

[†]通信作者. E-mail: mei.chen.lzjtu@hotmail.com.

(IMVC)^[4].

近年来, IMVC 领域取得了显著进展. 文献 [5] 通过引入低秩约束利用局部视图内数据相关性进行数据重建, 实现了缺失数据的补全; 文献 [6] 通过重构项推断缺失数据, 并引入跨视图图匹配方法增强了视图间的关联性. 然而, 这些方法主要针对不同视图缺失率相同的情形, 对于视图间缺失率不同的数据关注较少^[7-8]. 视图间缺失率的差异导致样本数量和可用信息的不平衡, 破坏了数据的整体结构, 影响聚类准确性.

目前, IMVC 算法主要采用两种策略处理缺失数据^[9]: 1) 将所有视图的数据映射至公共子空间进行聚类, 但是, 这一策略通常仅适用于双视图数据聚类场景; 2) 补全缺失值后再聚类, 这类方法通常使用特征平均值填充缺失值, 但是, 由于填充值与真实数据存在偏差, 影响聚类效果^[10]. 近年来, 为了更好地处理缺失数据问题, 文献 [11] 提出了自适应加权策略, 为缺失样本分配较小的权重来减轻其对于聚类结果的影响; 文献 [12] 使用所有视图的共识图和每个视图特有的不完整图, 重建了相应视图的完整图. 然而, 这些方法仍然难以有效获取缺失数据的准确信息, 特别是对于缺失率较高、各视图缺失率不同的多视图数据, 易导致较大偏差.

为了更好地学习数据的结构特征, 一些新颖的算法逐渐涌现. 文献 [13] 通过多级自表示约束捕获了数据内部的一致信息和视图间的互补信息; 文献 [14] 使用矩阵分解恢复缺失数据, 并利用线性表示捕获了数据的内在关系和高阶信息; 文献 [15] 通过超拉普拉斯正则化, 提出了一种噪声鲁棒多视图子空间聚类算法, 有效捕获了数据的局部几何结构. 然而, 这些方法难以同时捕获数据的全局结构和局部结构, 影响聚类结果.

张量技术在多视图聚类算法中已得到广泛应用. 文献 [16] 首次将各视图的子空间表示矩阵构建为张量, 并通过低秩张量约束捕获了视图间的高阶相关性, 但是, 其解并非 Tucker 秩的凸松弛形式, 导致无法保证全局最优解, 且缺乏明确的物理意义; 文献 [17] 提出了一种基于张量奇异值分解的多视图子空间聚类算法, 该算法对所有视图的子空间表示矩阵构造的张量施加低秩张量约束, 但是, 重构多视图数据是在矩阵形式而非张量形式下, 导致聚类性能不佳. 此外, 这些方法在处理缺失数据时仍然存在不足.

针对上述问题, 本文提出基于缺失值学习和结构保留的不平衡不完备多视图聚类 (UIMVC/MLSP). 该算法包括缺失值学习和结构保留两部分: 1) 通过

线性投影将原始高维数据映射至低维空间, 并结合基于近邻关系学习的补全矩阵和完整性约束填充缺失值, 以最大程度保留数据的原始特征; 2) 结合子空间聚类技术和超拉普拉斯正则化, 兼顾全局结构和局部结构的捕获, 同时, 通过加权张量 Schatten- p 范数挖掘数据的高阶相关性.

1 相关理论

为便于理解所提出算法, 本节首先介绍张量的相关概念, 文中所有重要符号如表 1 所示.

表1 重要符号及其含义

符号	含义
z, Z, \mathcal{Z}	向量, 矩阵, 张量
d_v	第 v 个视图的原始特征维度
$X^v \in \mathbb{R}^{d_v \times n}$	第 v 个视图的原始数据矩阵
$E_1^v \in \mathbb{R}^{p_v \times n}$	第 v 个视图的投影误差矩阵
$E_H^v \in \mathbb{R}^{p_v \times n}$	第 v 个视图的补全误差矩阵
$M^v \in \mathbb{R}^{n \times n}$	第 v 个视图的补全矩阵
$\mathcal{Z}(i, :, :)$	\mathcal{Z} 的第 i 个水平切片
$\mathcal{Z}(:, :, k)$	\mathcal{Z} 的第 k 个正切片
n	数据点的个数
p_v	第 v 个视图的投影特征维度
$P^v \in \mathbb{R}^{p_v \times d_v}$	第 v 个视图的投影矩阵
$E_2^v \in \mathbb{R}^{p_v \times n}$	第 v 个视图子空间表示误差矩阵
$H^v \in \mathbb{R}^{p_v \times n}$	第 v 个视图的低维特征矩阵
$T^v \in \mathbb{R}^{p_v \times n}$	第 v 个视图补全后的投影矩阵
$\mathcal{Z}(:, j, :)$	\mathcal{Z} 的第 j 个侧切片
$\ \mathcal{Z}\ _*$	张量核范数

定义 1^[18] 已知三阶张量 $\mathcal{Z} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, 其对应的 Frobenius 范数定义为

$$\|\mathcal{Z}\|_F = \sqrt{\sum_{ijk} |z_{ijk}|^2}. \quad (1)$$

定义 2^[18] 已知三阶张量 $\mathcal{Z} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, 其对应的张量奇异值分解 (t -SVD) 为

$$\mathcal{Z} = \mathcal{U} * \mathcal{S} * \mathcal{V}^T. \quad (2)$$

其中: $\mathcal{U} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ 和 $\mathcal{V} \in \mathbb{R}^{n_2 \times n_2 \times n_3}$ 为正交张量, $\mathcal{S} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ 为 f -对角张量.

定义 3^[18] 已知三阶张量 $\mathcal{Z} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, 其基于张量奇异值分解的张量核范数 (t -SVD-TNN) 定义为

$$\|\mathcal{Z}\|_{\otimes} = \sum_{j=1}^{n_3} \|\bar{\mathcal{Z}}(:, :, j)\|_* = \sum_{i=1}^{\min(n_1, n_2)} \sum_{j=1}^{n_3} |\bar{\mathcal{S}}(i, i, j)|. \quad (3)$$

其中: $\bar{\mathcal{Z}}$ 为 \mathcal{Z} 第三维度的傅里叶变换, $\bar{\mathcal{Z}} = \bar{\mathcal{U}} * \bar{\mathcal{S}} * \bar{\mathcal{V}}$, 即 $\bar{\mathcal{S}}$ 是通过 $\bar{\mathcal{Z}}$ 的正切片进行奇异值分解得到的.

2 基于缺失值学习和结构保留的不平衡不完备多视图聚类

所提出基于缺失值学习和结构保留的不平衡不完备多视图聚类算法包含两部分: 1) 缺失值学习: 首先通过线性投影将原始高维数据映射至低维潜在空间; 然后结合基于近邻关系学习的补全矩阵和完整性约束机制填充缺失值, 确保填充值不破坏数据的原始特征结构. 2) 结构保留: 算法结合子空间聚类技术和超拉普拉斯正则化, 同时捕获全局结构和局部结构, 并通过加权张量 Schatten- p 范数挖掘数据的高阶相关性和互补信息.

2.1 缺失值学习

在不完备多视图数据集中, 由于原始数据本身存在一定的噪声和冗余信息, 直接填充缺失值可能会降低聚类性能. 文献 [19] 利用线性投影, 将原始数据投影至低维潜在空间, 并将低维特征矩阵与基于随机数学学习的补全矩阵相乘, 在补全缺失数据的同时不仅有效减少了冗余信息, 还更好地恢复了数据的真实结构. 基于这一思想, 本文首先采用线性投影将原始高维数据映射至低维空间, 具体模型可表示为

$$H^v = P^v X^v + E_1^v. \quad (4)$$

然后, 本文基于数据的邻接关系学习补全矩阵来填充缺失值, 从而更加准确地反映数据的内在结构. 具体补全策略如下所示:

$$T^v = H^v M^v. \quad (5)$$

其中: T^v 为补全缺失值后的完备低维特征表示矩阵; 补全矩阵 M^v 的定义如下所示:

$$\begin{cases} M_{i,i}^v = 1, M_{i,j}^v = 0, & \text{if } i\text{th instance exists;} \\ M_{i,i}^v = 0, M_{i,j}^v = R, & \text{if } i\text{th instance is missing.} \end{cases} \quad (6)$$

这里 R 为 H^v 中第 i 个样本的 k 近邻.

通过这种方式, 补全矩阵将根据邻接关系为缺失数据分配合理的填充值, 捕获了数据的结构特征. 然而, 在填充缺失值的过程中, 填充值可能会破坏原始数据的结构特征, 尤其是在缺失值较多的情况下. 为避免这一问题, 本文引入了填充误差矩阵 E_H^v , 以确保填充的缺失值尽可能接近原始数据, 从而有效避免破坏数据的内在特征结构. 通过引入填充误差矩阵, 本文能够在填充过程中对填充值与原始数据间的差异进行控制, 从而提高数据补全的准确性. 具体如下:

$$E_H^v = T^v - H^v. \quad (7)$$

2.2 结构保留

2.2.1 全局结构

多视图子空间聚类技术是指每个数据可以用同一子空间内其他数据线性表示. 对于含有 V 个视图的数据集 $X^v \in \mathbb{R}^{d_v \times n}$, 多视图子空间聚类模型可表示为

$$\begin{aligned} & \min_{Z^v, E^v} L(X^v, X^v Z^v) + \lambda \mathfrak{R}(Z^v). \\ & \text{s.t. } X^v = X^v Z^v + E^v, \quad v = 1, 2, \dots, V; \\ & \quad 0 \leq Z^v \leq 1, (Z^v)^T \mathbf{1} = 1, \text{diag}(Z^v) = 0. \end{aligned} \quad (8)$$

其中: Z^v 为亲和矩阵; E^v 为子空间误差矩阵; $X^v = X^v Z^v + E^v$ 将数据样本表示为自身的线性组合; $L(X^v, X^v Z^v)$ 为子空间聚类重构误差的损失函数, 最初采用 Frobenius 范数和 l_1 范数进行优化^[20], 然而, 由于 $l_{2,1}$ 范数通过增加列稀疏性, 在一定程度上改善了算法的鲁棒性, 在多视图聚类算法中得到了广泛应用^[20]. 此外, 文献 [15] 提出了一种鲁棒性更强的柯西伪范数, 用于处理大规模、高强度的噪声数据, 可以显著提升聚类算法的鲁棒性. $\mathfrak{R}(Z^v)$ 为与亲和矩阵 Z^v 相关的正则化项, 通常表示视图间先验信息的约束; λ 为大于 0 的权衡参数. 各视图的亲和矩阵 $\{Z^v\}_{v=1}^V$ 被融合为一个统一的公共亲和矩阵 Z , 表示所有视图间的总体关系, 从而可以捕获多视图数据的全局结构^[9].

2.2.2 局部结构

在超图模型中: 顶点集 V 和边集 E 通过权重矩阵 W 进行关联, 指示矩阵 $H \in \mathbb{R}^{|V| \times |E|}$ 用于描述顶点与边的关系, 其元素被定义为

$$h(v_i, e_j) = \begin{cases} 1, & v_i \in e_j; \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

在超图中: $d(v_i) = \sum_{e_j \in E} w(e_j) h(v_i, e_j)$ 为顶点的度, $d(e_j) = \sum_{v_i \in V} h(v_i, e_j)$ 为边的度, $d(v_i)$ 和 $d(e_j)$ 分别为对角矩阵 D_v 和 D_e 的对角元素. 超图对应的超拉普拉斯矩阵^[9] 被定义为

$$L_h = D_v - H W D_e^{-1} H^T. \quad (10)$$

在普通图中, 每条边只连接两个顶点, 主要用于描述数据点间的二元关系; 而在超图中, 每条边可以连接多个顶点, 能够更全面地表示数据点间的多元关系. 正是由于这种多元连接的特性, 超图不仅可以保留数据的局部结构, 还能有效捕获数据间的多元关系. 因此, 本文引入了基于超图的拉普拉斯正则化 $\text{tr}(Z^v L_h^v (Z^v)^T)$ ($v = 1, 2, \dots, V$) 来保留数据的局部

结构, 捕获多元关系.

2.2.3 高阶相关性

张量结构能够有效揭示不同视图间的高阶相关性和互补信息, 在多视图聚类算法中得到了广泛应用. 现有的基于张量的多视图聚类方法通常采用张量核范数进行约束, 但是, 该方法通过线性求取所有奇异值来近似秩函数, 平等对待每个奇异值, 这可能会过度惩罚较大奇异值, 导致部分重要信息丢失, 使得张量秩的近似不准确^[21]. 相比之下, 加权张量 Schatten- p 范数作为更精确的秩估计方法, 通过 ω 和 p 可以更好地平衡不同奇异值的贡献, 保留数据中的关键信息和重要特征, 避免对较大奇异值的过度收缩^[21], 从而更精确地进行秩估计. 其数学定义如下所示:

$$\|\mathcal{Z}\|_{\omega, S_p} = \left(\sum_{i=1}^{n_3} \|\bar{\mathcal{Z}}^{(i)}\|_{\omega, S_p}^p \right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n_3} \sum_{j=1}^{\min(n_1, n_2)} \omega_j \cdot \sigma_j(\bar{\mathcal{Z}}^{(i)})^p \right)^{\frac{1}{p}}. \quad (11)$$

其中: $\bar{\mathcal{Z}}^{(i)}$ 为 \mathcal{Z} 的第 i 个正切片, $\sigma_j(\bar{\mathcal{Z}}^{(i)})$ 为 $\bar{\mathcal{Z}}^{(i)}$ 的第 j 个奇异值.

为了更精确地近似张量秩函数, 本文引入了加权张量 Schatten- p 范数对亲和张量进行约束. 具体而言, 首先将各视图的亲和矩阵构建为三阶张量, 并通过旋转维度来保留原结构, 同时降低时间复杂度^[22]. 在此基础上, 采用加权张量 Schatten- p 范数处理旋转后的张量, 精准捕获数据的高阶相关性和重要特征. 此外, 本文引入了鲁棒性更强的柯西伪范数处理各阶段的误差矩阵, 提高算法对噪声数据的鲁棒性.

综上所述, 所提出总模型如下所示:

$$\begin{aligned} \min_{\substack{P^v, H^v, M^v, \\ Z^v, E^v, E_H^v, T^v}} &= \sum_{v=1}^V \{ \lambda \|E^v\|_c + \beta \|E_H^v\|_c + \\ &\alpha \text{tr}(Z^v L_h^v (Z^v)^T) + \|\mathcal{Z}\|_{\omega, S_p}^p \}. \\ \text{s.t. } &H^v = P^v X^v + E_1^v, T^v = H^v M^v; \\ &E_H^v = T^v - H^v; \\ &T^v = T^v Z^v + E_2^v, P^v (P^v)^T = I; \\ &E^v = [E_1^v; E_2^v]; \\ &\mathcal{Z} = \Phi(Z^1, Z^2, \dots, Z^V), v = 1, 2, \dots, V. \end{aligned} \quad (12)$$

其中 λ 、 β 和 α 为3个平衡参数.

2.3 模型求解

为获得式(12)的最优解, 本文采用交替方向乘子法进行变量优化. 为方便求解, 在式(12)的基础上引入辅助变量 Q^v 和 \mathcal{J} , 将式(12)改写为

$$\begin{aligned} \min_{\substack{P^v, H^v, M^v, \\ Z^v, E^v, E_H^v, \\ T^v, Q^v, \mathcal{J}}} &= \sum_{v=1}^V \{ \lambda \|E^v\|_c + \beta \|E_H^v\|_c + \\ &\alpha \text{tr}(Q^v L_h^v (Q^v)^T) + \|\mathcal{J}\|_{\omega, S_p}^p \}. \\ \text{s.t. } &H^v = P^v X^v + E_1^v, T^v = H^v M^v; \\ &E_H^v = T^v - H^v; \\ &T^v = T^v Z^v + E_2^v, P^v (P^v)^T = I; \\ &E^v = [E_1^v; E_2^v]; \\ &\mathcal{Z} = \Phi(Z^1, Z^2, \dots, Z^V), Z^v = Q^v, \mathcal{Z} = \mathcal{J}; \\ &v = 1, 2, \dots, V. \end{aligned} \quad (13)$$

模型(13)可通过增广拉格朗日函数转换为无约束优化问题, 分离各变量依次求解. 模型(13)的增广拉格朗日函数为

$$\begin{aligned} L(P^v, H^v, M^v, Z^v, E^v, E_H^v, T^v, Q^v, \mathcal{J}, \mathcal{W}, Y_1^v, \\ Y_2^v, Y_Q^v, Y_H^v, Y_M^v) = \\ \left\{ \sum_{v=1}^V \lambda \|E^v\|_c + \beta \|E_H^v\|_c + \alpha \text{tr}(Q^v L_h^v (Q^v)^T) + \right. \\ \left. \|\mathcal{J}\|_{\omega, S_p}^p + \frac{\mu}{2} \left\| \mathcal{Z} - \mathcal{J} + \frac{\mathcal{W}}{\mu} \right\|_F^2 + \right. \\ \left. \frac{\mu}{2} \left\| P^v X^v - H^v + E_1^v + \frac{Y_1^v}{\mu} \right\|_F^2 + \right. \\ \left. \frac{\mu}{2} \left\| Z^v - Q^v + \frac{Y_Q^v}{\mu} \right\|_F^2 + \right. \\ \left. \frac{\mu}{2} \left\| T^v - H^v - E_H^v + \frac{Y_H^v}{\mu} \right\|_F^2 + \right. \\ \left. \frac{\mu}{2} \left\| T^v - T^v Z^v - E_2^v + \frac{Y_2^v}{\mu} \right\|_F^2 + \right. \\ \left. \frac{\mu}{2} \left\| H^v M^v - T^v + \frac{Y_M^v}{\mu} \right\|_F^2 \right\}. \end{aligned} \quad (14)$$

其中: μ 为惩罚项参数, Y_1^v 、 Y_2^v 、 Y_Q^v 、 Y_H^v 、 Y_M^v 、 \mathcal{W} 为拉格朗日乘子. 通过固定其他变量, 对相关变量逐个求解优化, 得到模型(14)的最优解, 下面将详细介绍所有变量的更新过程.

step 1: 更新变量 P^v . 固定其他变量后, 对变量 P^v 的求解可简化为如下问题:

$$\begin{aligned} L(P^v) = \min_{P^v} \frac{\mu}{2} \left\| P^v X^v - H^v + E_1^v + \frac{Y_1^v}{\mu} \right\|_F^2; \\ \text{s.t. } P^v (P^v)^T = I. \end{aligned} \quad (15)$$

式(15)参见文献[19]的方法获得 P^v 的最优解.

step 2: 更新变量 H^v . 固定其他变量后, 对变量 H^v 求导, 置导数为0, 得到 H^v 的解为

$$\begin{aligned} H^v = (2I + M^v (M^v)^T)^{-1} \times \\ \left(P^v X^v + E_1^v + T^v - E_H^v + T^v (M^v)^T + \right. \\ \left. \frac{Y_1^v + Y_H^v - Y_M^v (M^v)^T}{\mu} \right). \end{aligned} \quad (16)$$

step 3: 更新变量 Z^v . 固定其他变量后, 对变量

Z^v 求导, 置导数为 0, 得到 Z^v 的解为

$$Z^v = (2I + T^v(T^v)^T)^{-1} \times \left((T^v)^T(T^v - E_2^v) + \frac{(T^v)^T Y_2^v - Y_Q^v - W^v}{\mu} + Q^v + J^v \right). \quad (17)$$

step 4: 更新变量 M^v . 固定其他变量后, 对 M^v 求导, 置导数为 0, 得到 M^v 的解为

$$M^v = ((H^v)^T H^v)^{-1} \times \left((H^v)^T T^v - \frac{(H^v)^T Y_M^v}{\mu} \right). \quad (18)$$

step 5: 更新变量 T^v . 固定其他变量后, 对 T^v 求导, 置导数为 0, 得到 T^v 的解为

$$T^v = (2I + (I - Z^v)(Z^v)^T)^{-1} \times \left(H^v + E_H^v + H^v M^v + E_2^v (Z^v)^T + \frac{Y_M^v - Y_H^v - Y_2^v (Z^v)^T}{\mu} \right). \quad (19)$$

step 6: 更新变量 E^v . 固定其他变量后, 对变量 E^v 的求解将变为如下优化问题:

$$(E^v)^* = \min_{E^v} \left\{ \lambda \|E^v\|_c + \frac{\mu}{2} \left\| H^v - P^v X^v - E_1^v + \frac{Y_1^v}{\mu} \right\|_F^2 + \frac{\mu}{2} \left\| T^v - T^v Z^v - E_2^v + \frac{Y_2^v}{\mu} \right\|_F^2 \right\} = \min_{E^v} \frac{\lambda}{2} \|E^v\|_c + \frac{1}{2} \|E^v - D^v\|_F^2. \quad (20)$$

其中: $D^v = \left[H^v - P^v X^v + \frac{Y_1^v}{\mu}; T^v - T^v Z^v + \frac{Y_2^v}{\mu} \right]$; $E^v = [E_1^v; E_2^v]$, 由于矩阵 E^v 的各列元素间不相关, 可将变量 E^v 分解为 n 个子问题进行求解, 具体求解过程参见文献 [15]. 最终得到 E^v 的解为

$$(E_{:,i}^v)^* = \begin{cases} \left(1 - \frac{\lambda/\mu}{\lambda/\mu + \tau(\lambda + \tau)} \right) d_j^v, & \frac{\lambda}{\mu} < \|d_j^v\|_2; \\ 0, & \text{otherwise.} \end{cases} \quad (21)$$

这里: $\tau = \frac{\|d_j^v\|_2 - \gamma + \sqrt{(\|d_j^v\|_2 + \gamma)^2 + 4\lambda/\mu}}{2}$, d_j^v

为矩阵 D^v 的第 j 列.

step 7: 更新变量 E_H^v . 类似于更新 E^v , E_H^v 的更新如下所示:

$$(E_H^v)^* = \min_{E_H^v} \left\{ \beta \|E_H^v\|_c + \frac{\mu}{2} \left\| T^v - H^v - E_H^v + \frac{Y_H^v}{\mu} \right\|_F^2 \right\} = \min_{E_H^v} \frac{\beta}{2} \|E_H^v\|_c + \frac{1}{2} \|E_H^v - D_H^v\|_F^2. \quad (22)$$

其中: $D_H^v = \left[T^v - H^v + \frac{Y_H^v}{\mu} \right]$; E_H^v 的最终解为

$$((E_H^v)_{:,i})^* = \begin{cases} \left(1 - \frac{\beta/\mu}{\beta/\mu + \tau_2(\beta + \tau_2)} \right) (d_H^v)_j^v, \\ \frac{\beta}{\mu} < \|(d_H^v)_j\|_2; \\ 0, & \text{otherwise.} \end{cases} \quad (23)$$

这里: $\tau_2 = \frac{\|(d_H^v)_j\|_2 - \gamma + \sqrt{(\|(d_H^v)_j\|_2 + \gamma)^2 + 4\beta/\mu}}{2}$,

$(d_H^v)_j^v$ 为矩阵 D_H^v 的第 j 列.

step 8: 更新变量 Q^v . 固定其他变量后, 对 Q^v 求导, 置导数为 0, 得到 Q^v 的解为

$$(Q^v)^* = (\alpha(L_h^v + (L_h^v)^T) + \mu I)^{-1} \times (\mu Z^v + Y_Q^v). \quad (24)$$

step 9: 更新变量 \mathcal{J} . 固定其他变量, 通过如下模型对变量 \mathcal{J} 进行求解:

$$\mathcal{J}^* = \arg \min_{\mathcal{J}} \left\| \mathcal{J} \right\|_{\omega, S_p}^p + \frac{\mu}{2} \left\| \mathcal{Z} - \mathcal{J} + \frac{\mathcal{W}}{\mu} \right\|_F^2 = \arg \min_{\mathcal{J}} \left\| \mathcal{J} \right\|_{\omega, S_p}^p + \frac{\mu}{2} \left\| \mathcal{J} - \left(\mathcal{Z} + \frac{\mathcal{W}}{\mu} \right) \right\|_F^2. \quad (25)$$

根据定理 1, 式 (25) 的解为

$$\mathcal{J}^* = \Gamma_{\frac{1}{\mu}, n_3, \omega} \left(\mathcal{Z} + \frac{\mathcal{W}}{\mu} \right). \quad (26)$$

定理 1^[18] 假设张量 $\mathcal{G} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, $l = \min(n_1, n_2)$, $0 \leq \omega_1 \leq \dots \leq \omega_l$, $\mathcal{G} = \mathcal{U} * \mathcal{S} * \mathcal{V}^T$, 则对于

$$\arg \min_{\mathcal{J}} \frac{1}{2} \left\| \mathcal{J} - \mathcal{G} \right\|_F^2 + \mu \left\| \mathcal{J} \right\|_{\omega, S_p}^p, \quad (27)$$

式 (27) 的解为

$$\mathcal{J}^* = \Gamma_{\mu, n_3, \omega}(\mathcal{G}) = \mathcal{U} * \text{iff}(\mathcal{P}_{\mu, n_3, \omega}(\bar{\mathcal{G}})) * \mathcal{V}^T, \quad (28)$$

其中 $\mathcal{P}_{\mu, n_3, \omega}(\bar{\mathcal{G}}^{(i)})$ 为张量 $\mathcal{P}_{\mu, n_3, \omega}(\bar{\mathcal{G}})$ 的第 i 个正切片.

step 10: 更新 Y_1^v , Y_2^v , Y_Q^v , Y_M^v , Y_H^v , \mathcal{W} 和 μ , 即

$$Y_1^v = Y_1^v + \mu R_1^v, \quad R_1^v = H^v - P^v X^v - E_1^v; \quad (29)$$

$$Y_2^v = Y_2^v + \mu R_2^v, \quad R_2^v = T^v - T^v Z^v - E_2^v; \quad (30)$$

$$Y_Q^v = Y_Q^v + \mu R_3^v, \quad R_3^v = Z^v - Q^v; \quad (31)$$

$$Y_H^v = Y_H^v + \mu R_4^v, \quad R_4^v = T^v - H^v - E_H^v; \quad (32)$$

$$Y_M^v = Y_M^v + \mu R_5^v, \quad R_5^v = H^v M^v - T^v; \quad (33)$$

$$\mathcal{W} = \mathcal{W} + \mu R_6, \quad R_6 = \mathcal{Z} - \mathcal{J}; \quad (34)$$

$$\mu = \min(\mu\rho, \mu_{\max}). \quad (35)$$

其中 ρ 和 μ_{\max} 为常数.

令 $R_* = [R_*^1, R_*^2, \dots, R_*^6]$ ($*$ = 1, 2, ..., 6), 算法的收敛条件为

$$\|R_i\|_\infty \leq \varepsilon, \quad i = 1, 2, \dots, 6, \quad (36)$$

其中 ε 为最小误差. 当算法收敛后, 得到统一亲和矩阵 $S = \frac{1}{V} \sum_{v=1}^V (|Z^v| + |(Z^v)^T|)$, 在 S 上执行谱聚类得到聚类结果. 过程如算法1所示.

算法1 UIMVC/MLSP.

step 1: 输入多视图数据 X^v , 参数 λ 、 β 和 α , 投影特征维度 p_v , 簇个数 c ;

step 2: 初始化 $Y_1^v = Y_2^v = Y_Q^v = Y_H^v = Y_M^v = M^v = T^v = Z^v = Q^v = E^v = E_H^v = 0, v = 1, 2, \dots, V; \mu = 10^{-6}, \varepsilon = 10^{-6}, \rho = 1.2, \mu_{\max} = 10^6$;

step 3: while 未收敛 do

step 3.1: 由式(15), 更新 P^v ;

step 3.2: 由式(16), 更新 H^v ;

step 3.3: 由式(17), 更新 Z^v ;

step 3.4: 由式(18), 更新 M^v ;

step 3.5: 由式(19), 更新 T^v ;

step 3.6: 由式(21), 更新 E^v ;

step 3.7: 由式(23), 更新 E_H^v ;

step 3.8: 由式(24), 更新 Q^v ;

step 3.9: 由式(26), 更新 \mathcal{J} ;

step 3.10: 由式(29) ~ (35), 更新拉格朗日乘子和惩罚项参数 $Y_1^v, Y_2^v, Y_Q^v, Y_M^v, Y_H^v, \mathcal{W}, \mu$;

step 3.11: 由式(36), 判断算法收敛;

step 3.12: end while

step 4: 得到亲和矩阵 $S, S = \sum_{v=1}^V \frac{|Z^v| + |(Z^v)^T|}{V}$;

step 5: 将 S 作为谱聚类的输入得到聚类结果.

2.4 时间复杂度分析

本节将对算法的时间复杂度进行分析. 主要是更新 $L_h^v, P^v, H^v, M^v, Z^v, E^v, E_H^v, T^v, Q^v, \mathcal{J}$. 1) 构建超拉普拉斯矩阵 L_h^v 的时间复杂度为 $O(n^2V \log(n))$; 2) 更新投影矩阵 P^v 时需要奇异值分解, 时间复杂度为 $O((d_v)^2Vn)$; 3) 更新 H^v, M^v, Z^v, T^v, Q^v 的时间复杂度为 $O(Vn^3)$; 4) 更新 E^v, E_H^v 的时间复杂度为 $O(Vn^2)$; 5) 更新 \mathcal{J} 的过程中, 涉及到张量沿第三维度的傅里叶变换以及逆变换, 其时间复杂度为 $O(n^2V \log(n))$, 且需要在傅里叶域内对 n 个正切片进行奇异值分解, 时间复杂度为 $O(n^2V^2)$, 因此, 更新 \mathcal{J} 的时间复杂度为 $O(n^2V^2 + n^2V \log(n))$. 本文忽略简单操作的时间复杂度, 如矩阵的加减、点乘和点除等操作. 综上所述, 由于 $1 < V \ll \log(n) < n$, 所提出算法总的时间复杂度为

$O(t((d_v)^2Vn + Vn^3))$, 其中 t 为迭代次数.

3 实验分析

本文所有实验的编程环境为 Matlab 2019a, 操作系统为 Windows 10, 内存为 16 GB, 处理器为 AMD Ryzen 5 5600H.

3.1 数据集介绍

为了验证所提出方法的性能, 本节选取 8 个常见的公开多视图数据集, 基本信息如表 2 所示. 其中: V 为视图个数, c 为簇个数, n 为数据点的个数, d 为原始特征维度. 为全面评估算法在不同缺失程度下的聚类性能, 本文随机删除一定比例的数据来构建不完备多视图数据集, 缺失比率 (PER) 分别设置为 0.1、0.3、0.5、0.7 和 0.9, 构建过程如表 3 所示.

表2 数据集的描述

datasets	V	c	n	d
BBC	4	5	685	4659/4633/4665/4684
BBCSport	2	5	544	3183/3203
ORL	3	40	400	4096/3304/6750
NGs	3	5	500	2000/2000/2000
Digit	5	10	2000	216/76/64/240/47
MSRCV1	6	7	210	1302/48/512/1020/256/210
YaleA	3	15	165	9/512/50
3Sources	3	6	416	3560/3631/3068

表3 不完备数据集的描述

datasets	BBCSport	BBC/ORL/NGs/ YaleA/3Sources	Digit	MSRCV1
v_1	0.25 × PER	0.25 × PER	0.25 × PER	0.25 × PER
v_2	0.75 × PER	0.75 × PER	0.75 × PER	0.75 × PER
v_3	—	PER	PER	PER
v_4	—	—	1.25 × PER	1.25 × PER
v_5	—	—	1.5 × PER	1.5 × PER
v_6	—	—	—	1.75 × PER

3.2 对比算法介绍

为全面评估所提出算法的有效性, 本文将 UIMVC/MLSP 与以下 10 种聚类算法进行比较: 不完备单视图聚类算法 BSV^[23], 基于谱扰动的不完备多视图聚类 (PIC)^[10], 双对齐不完备多视图聚类 (DAIMC)^[24], 基于潜在空间中缺失数据学习的不完备多视图聚类 (MISS)^[19], 基于高阶相关性保留的不完备多视图子空间聚类 (HCPIMSC)^[25], 投影不完备多视图聚类 (PIMVC)^[26], 平衡引导不完备多视图谱聚类 (BGIMVSC)^[27], 基于张量化图学习的广义不完备多视图聚类 (TTGL)^[28], 基于生物进化论的不平衡不完备多视图聚类 (TETCI_UIMC)^[8] 以及基于跨视图投影不平衡不完备多视图聚类 (PCL_UIMVC)^[29].

上述所有算法的代码均由作者提供, 且根据提供的参数范围进行了参数寻优, 将最佳结果作为相应算法最终的聚类结果. 所有结果均由本文作者复现.

3.3 实验结果与分析

为全面评估所提出算法的性能, 本文采用 Accuracy、NMI 和 Purity 共 3 种指标衡量聚类效果. 此外, 通过网格搜索为每个数据集寻找最佳参数组合, 并在最优参数组合下重复实验 10 次, 最终取平均值作为各数据集的最终指标值. 表 4 ~ 表 6 为所有算法在 8 个数据集所有 PER 下聚类指标的平均值, 图 1 为 NGs 数据集在每个 PER 下各指标的变化

趋势. 由实验结果可得出如下结论.

1) 表 4 ~ 表 6 的数据表明, PIC、MISS 以及 HCPIMSC 等不完备多视图聚类算法相较于不完备单视图聚类算法 BSV 表现更优异, 因为不完备多视图聚类算法可有效捕获数据各视图间的互补信息, 获得更全面、更丰富的信息, 有助于提升聚类性能.

2) TETCI_UIMC、PCL_UIMVC 与 UIMVC/MLSP 相较于其他对比算法表现更出色, 因为这 3 个算法是一种不平衡不完备多视图聚类算法, 能够有效处理视图缺失率不同的问题, 更符合现实数据集的特点, 有助于提升聚类性能.

表4 8个数据集上所有 PER 的 Accuracy 均值

	BBC	BBCSport	NGs	ORL	YaleA	MSRCV1	Digit	3Sources
BSV	0.4387	0.4892	0.4161	0.3929	0.3448	0.4138	0.3142	0.4212
PIC	0.7357	0.7407	0.8404	0.6811	0.4462	0.7624	0.8384	0.4549
DAIMC	0.5321	0.5712	0.6970	0.5214	0.6178	0.6507	0.8467	0.4982
MISS	0.7091	0.6335	0.7652	0.4465	0.4897	0.6305	0.8385	0.6010
HCPIMSC	0.8149	0.8779	0.8616	0.8302	0.7201	0.7982	0.7674	0.5423
PIMVC	0.7616	0.7657	0.8864	0.7285	0.6798	0.8295	0.8123	0.6532
BGIMVSC	0.6403	0.6992	0.6644	0.6220	0.5757	0.5228	0.5165	0.4935
TTGL	0.6619	0.7665	0.6000	0.4340	0.5265	0.6209	0.7802	0.5088
TETCI_UIMC	0.8654	0.7796	0.8252	0.4340	0.5261	0.6209	0.7802	0.5088
PCL_UIMVC	0.7887	0.7566	0.8240	0.7940	0.7912	0.8038	0.8054	0.7289
ours	0.9743	0.9849	0.9996	0.9745	0.9031	0.9762	0.9749	0.7799

表5 8个数据集上所有 PER 的 NMI 均值

	BBC	BBCSport	NGs	ORL	YaleA	MSRCV1	Digit	3Sources
BSV	0.1978	0.3231	0.1913	0.5417	0.4188	0.4679	0.3521	0.2755
PIC	0.6642	0.6284	0.6984	0.8096	0.5041	0.6948	0.8603	0.2872
DAIMC	0.3557	0.4263	0.5178	0.6972	0.6819	0.5515	0.7652	0.4451
MISS	0.5449	0.4388	0.5791	0.6101	0.5379	0.5021	0.7645	0.4706
HCPIMSC	0.6422	0.7552	0.6972	0.7083	0.7368	0.6434	0.7114	0.4461
PIMVC	0.7616	0.7657	0.8864	0.7285	0.6798	0.8295	0.8123	0.6532
BGIMVSC	0.6403	0.6992	0.6644	0.6220	0.5757	0.5228	0.5165	0.4935
TTGL	0.6619	0.7665	0.6000	0.4340	0.5265	0.6209	0.7802	0.5088
TETCI_UIMC	0.8654	0.7796	0.8252	0.4340	0.5261	0.6209	0.7802	0.5088
PCL_UIMVC	0.7887	0.7566	0.8240	0.7940	0.7912	0.8038	0.8054	0.7289
ours	0.9743	0.9849	0.9996	0.9745	0.9031	0.9762	0.9749	0.7799

表6 8个数据集上所有 PER 的 Purity 均值

	BBC	BBCSport	NGs	ORL	YaleA	MSRCV1	Digit	3Sources
BSV	0.4387	0.5479	0.4161	0.4152	0.3934	0.4584	0.3401	0.4827
PIC	0.7357	0.7545	0.8404	0.6811	0.4603	0.7624	0.8631	0.5636
DAIMC	0.5932	0.6693	0.7017	0.5695	0.6388	0.6668	0.8239	0.5902
MISS	0.7542	0.6452	0.7736	0.4705	0.5043	0.6314	0.8416	0.6828
HCPIMSC	0.8236	0.8779	0.8616	0.8303	0.7237	0.8029	0.7689	0.6519
PIMVC	0.7868	0.7794	0.8864	0.7760	0.6798	0.8372	0.8454	0.7424
BGIMVSC	0.6596	0.7374	0.6708	0.6345	0.5903	0.5391	0.5260	0.5887
TTGL	0.6029	0.7627	0.6000	0.4345	0.5364	0.6276	0.7836	0.5179
TETCI_UIMC	0.8654	0.7961	0.8272	0.4340	0.5366	0.6278	0.7836	0.5779
PCL_UIMVC	0.7941	0.7943	0.8279	0.7955	0.8148	0.8076	0.8189	0.7917
ours	0.9743	0.9849	0.9996	0.9785	0.9067	0.9762	0.9749	0.8189

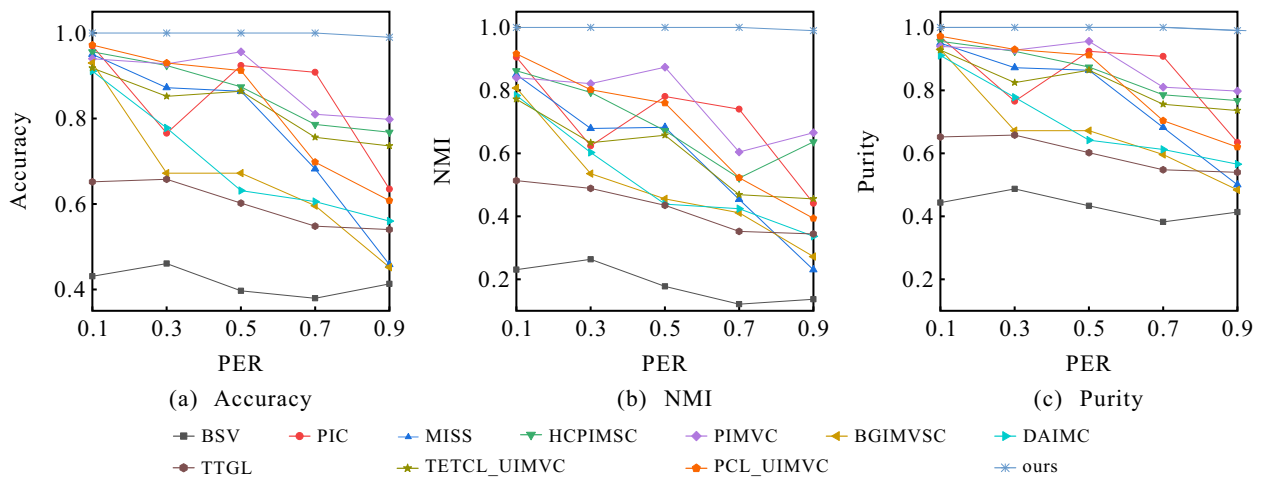


图1 NGs数据集上不同缺失率与聚类指标关系

3) UIMVC/MLSP 算法的聚类性能优于其他对比算法. 因为该算法不仅通过基于近邻学习的补全矩阵和完整性约束机制填充缺失值, 确保了填充后的缺失数据尽可能保留原始特征结构, 且结合子空间聚类技术和超拉普拉斯正则化, 捕获了数据的全局结构和局部结构, 通过结合缺失值学习和结构保留, 增强了算法的聚类效果.

4) 相较于 BSV、DAIMC 与 PIC, UIMVC/MLSP 算法的聚类性能更好, 因为 UIMVC/MLSP 利用加权张量 Schatten- p 范数, 能够更精确地捕获数据的高阶相关性和各视图的互补信息, 从而有效保留了数据的关键特征, 提升了聚类性能.

5) 由图 1 可以发现, 随着缺失率的增加, 所有方法的 Accuracy、NMI 和 Purity 均呈现下降趋势, 表明数据缺失对于聚类性能会造成负面影响. 然而, UIMVC/MLSP 的各项聚类指标始终优于其他算法, 表明该方法在处理缺失数据时具有较强的鲁棒性.

3.4 模型可视化分析

本节在 PER = 0.5 的 NGs 数据集上采用 t 分布随机邻居嵌入方法 (t -SNE)^[30] 对每个视图的原始数据和最终学习到的亲和矩阵进行可视化, 如图 2 所示. 限于篇幅, 本节仅展示当 PER = 0.5 时, NGs 数据集上 Accuracy 值排名前 5 的对比算法的亲和矩阵可视化结果. 图 2(a) ~ 图 2(c) 分别为 NGs 数据集原始视图 1、原始视图 2 和原始视图 3 的数据分布. 由图 2(a) ~ 图 2(c) 可见, 各视图中数据的分布较为混乱, 簇间的区分度较低, 簇结构不明显. 图 2(d) ~ 图 2(i) 为各方法获得的最终亲和矩阵. 与其他算法相比, 所提出方法使得数据由原始的混乱状态变为清晰的簇结构, 不同簇间的区分度显著增强, 可以准确识别出 5 个簇. 这是因为本文采用了补全矩阵和完整性约束机制填充缺失数据, 确保了填充后的数据

尽可能保持原始结构特征. 此外, 本文还同时捕获了数据的全局结构、局部结构和高阶结构, 提升了聚类效果.

3.5 参数敏感性实验

所提出模型包含 3 个惩罚项参数: λ 、 β 和 α , 分别调节投影误差、补全误差和超拉普拉斯正则化对于模型的贡献. 本节分析了 PER = 0.5 时 BBC 和 NGs 数据集上 3 个参数对于算法性能的影响. 参数的取值范围为 $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. 由于 3 个参数与 Accuracy 的关系难以在一张图中展示, 本文采用控制变量法, 分别展示不同参数与 Accuracy 的关系, 如图 3 所示. 图 3(a) 和图 3(c) 为固定参数 α 后, λ 和 β 的取值与 Accuracy 的关系. 由图 3(a) 和图 3(c) 可见, 无论是 BBC 还是 NGs 数据集, 参数 λ 和 β 的变化对于 Accuracy 的影响较小, 表明算法对于这两个参数不敏感. 图 3(b) 和图 3(d) 为固定 λ 和 β 后, α 与 Accuracy 的关系. 由图 3(b) 和图 3(d) 可见, 随着 α 的变化, Accuracy 始终保持在 0.85 以上, 表明 α 对于聚类结果的影响较小. 综上所述, 3 个参数 λ 、 β 和 α 对于 UIMVC/MLSP 的影响较小, 算法对于 λ 、 β 和 α 不敏感.

3.6 有效性实验

为了进一步探究 UIMVC/MLSP 算法中各模块对于最终聚类性能的影响, 本文在不同缺失率下的 3Sources 和 BBCSport 数据集上进行了有效性实验. 为验证完整性约束、超拉普拉斯正则化和加权张量 Schatten- p 范数的作用, 分别移除这些约束, 得到模型 EH、模型 HL 和模型 SP. 图 4 为不同缺失率下的 Accuracy 值. 实验结果表明: 1) UIMVC/MLSP 在各缺失率下表现优于其他 3 种变体, 尤其是在高缺失率下, 表明其对于不平衡不完备多视图数据具有较强的适应性; 2) 与变体 EH 相比, UIMVC/MLSP 的

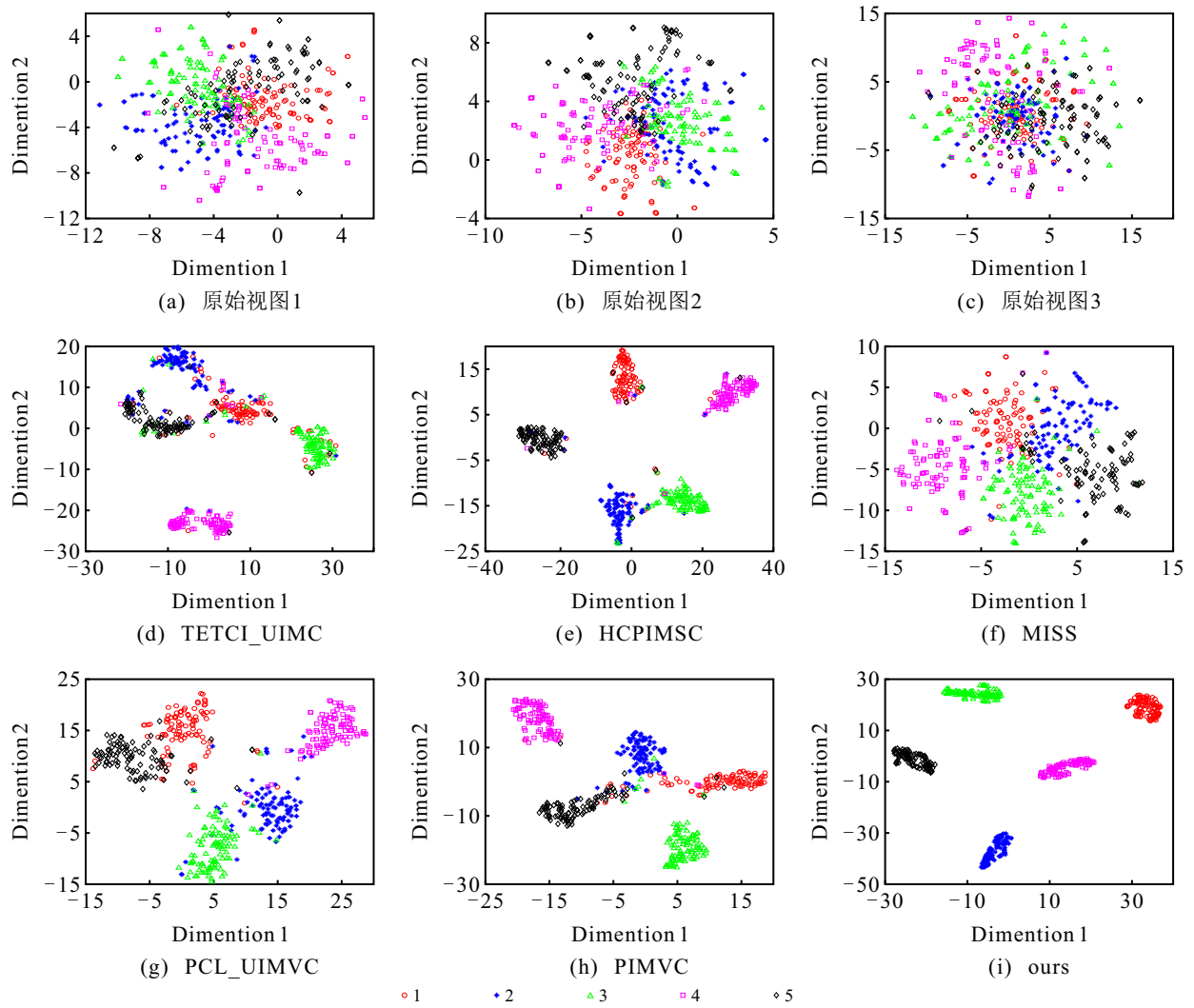


图2 NGs数据集t-SNE可视化

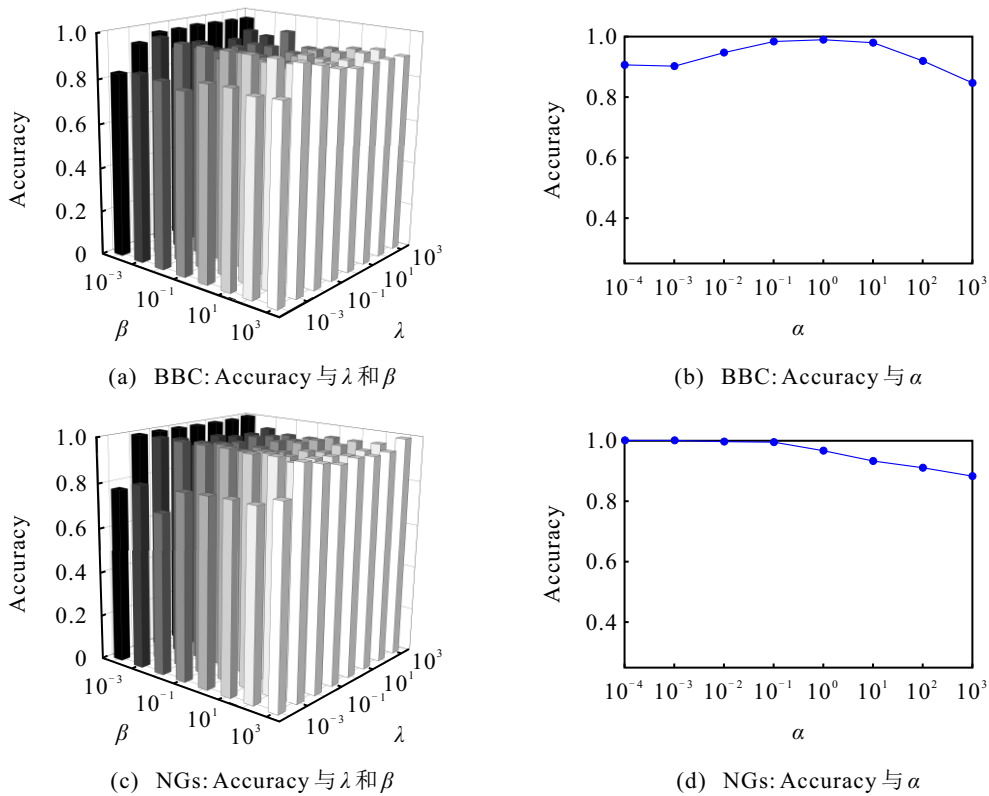
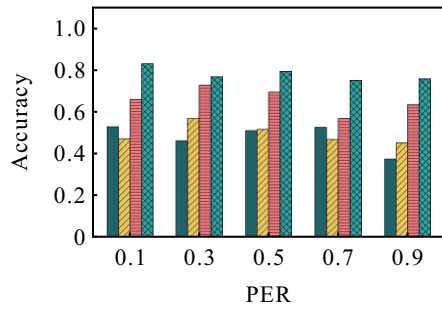
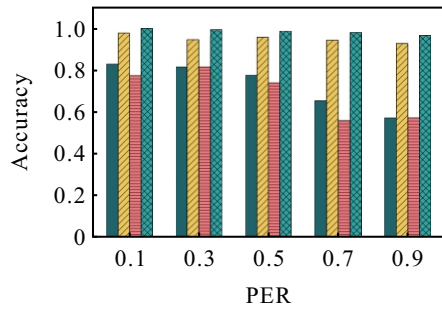


图3 不同参数分布与 Accuracy 的关系



(a) 3Sources



(b) BBCSport

■ 模型EH ■ 模型HL
■ 模型SP ■ 原模型MLSP/UMVC

图4 有效性分析

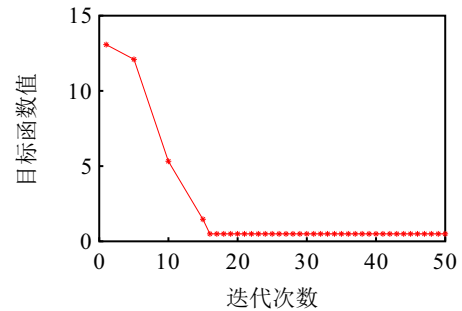
Accuracy 更高,表明完整性约束使得填充的缺失值有效恢复了数据的原始特征; 3) UIMVC/MLSP 比模型 HL 的聚类性能更好,因其通过子空间聚类技术和超拉普拉斯正则化同时捕获了数据的全局结构和局部结构,提升了聚类准确性; 4) 相比于模型 SP, UIMVC/MLSP 通过加权张量 Schatten- p 范数更好地捕获了数据的高阶相关性和互补信息,进一步提升了聚类性能. 综上所述, UIMVC/MLSP 通过各模块组合提升了聚类性能,验证了模型的有效性.

3.7 收敛性实验

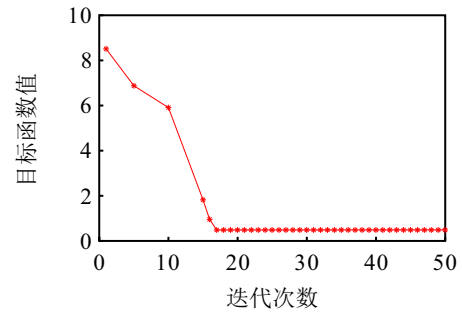
本节通过实验验证所提出算法的收敛性. 目标函数值的定义如式 (36) 所示,由式 (36) 计算每次迭代的原始残差,当原始残差小于设定值时,算法判定收敛. 图 5(a) 和图 5(b) 分别为 PER = 0.5 时 BBC 和 ORL 数据集的收敛曲线. 由图 5 可见,目标函数值在两个数据集上呈现先下降后平稳的趋势,表明了所提出算法具有良好的收敛性.

3.8 时间对比实验

为了更好地验证所提出方法的可行性,本节进行了时间对比实验,验证算法的实际运行时间. 在第 2.4 节中对所提出算法的时间复杂度进行了理论分析,同时本节整理了对比算法的时间复杂度,主要参考其原论文中给出的分析结果,具体如表 7 所示. 鉴于 BSV 算法的时间复杂度在现有文献中尚未明确给出,本文在表 7 中暂未整理. 为保证实验的可比性,所有算法统一迭代 15 次,并去除了参数寻优过程.



(a) BBC



(b) ORL

图5 收敛性分析

表7 各算法时间复杂度

算法	时间复杂度
BSV	—
PIC	$O(V^3n^2k + V^3)$
HCPIMSC	$O(tVn^3)$
BGIMVSC	$O(tnc^2)$
TETCI_UIMC	$O\left(t\sum_{v=1}^V(n_v^3)\right)$
DAIMC	$O(VLd_{\max}^3 + LMd_{\max}KN)$
MISS	$O(t(Vd_v^2n + n^3))$
PIMVC	$O\left(n^3 + t\sum_{v=1}^V d_v c^2\right)$
TTGL	$O(t(n^3 + n^2 \log(n)))$
PCL_UIMVC	$O\left(n^3 + t\sum_{v=1}^V (d_v^3 + d_v c^2)\right)$

本文选取 BBCSport 和 3Sources 数据集,展示各方法的运行时间,如图 6 所示. 由图 6 可见: 1) 所提出算法的运行时间低于 HCPIMSC、TETCI_UIMC 与 PCL_UIMVC 等算法. 因为 UIMVC/MLSP 利用投影矩阵将原始高维数据映射至低维潜在空间,有效去除了冗余信息,在一定程度上降低了运行时间. 此外,本文在进行张量奇异值分解时,通过旋转维度有效提升了计算效率. 2) 与 BSV 算法相比, UIMVC/MLSP 的运行时间较长. 因为 BSV 主要针对单一视图进行聚类,计算负担小,但是 BSV 的聚类性能欠佳,结果偏低. 3) 相较于 MISS、PIMVC 与 BGIMVSC 等算法, UIMVC/MLSP 的运行时间略高.

因为本文在模型中引入了加权张量 Schatten- p 范数, 有效捕获了数据的高阶信息与各视图的互补关系, 虽然增加了一定的时间开销, 但是显著提升了聚类性能. 综上所述, 虽然所提出方法的运行时间并非最低, 但是通过结合缺失值学习和结构保留, 可获得卓越的聚类性能.

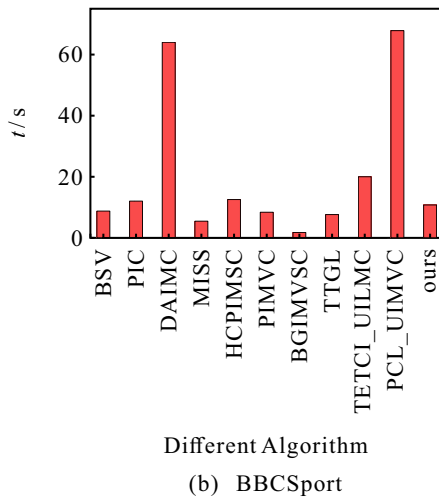
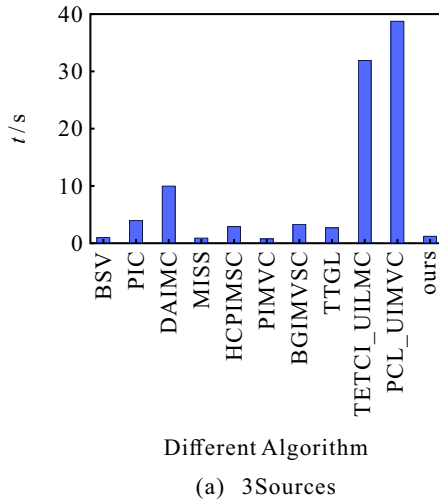


图6 时间对比实验

4 结论

针对现有不完备多视图聚类算法在挖掘数据结构信息等问题时表现出的局限性, 本文提出了基于缺失值学习和结构保留的不平衡不完备多视图聚类算法. 该算法通过补全矩阵填充缺失值, 并引入了完整性约束机制, 最大程度保留了数据的原始特征. 同时, 结合子空间聚类技术、超拉普拉斯正则化和加权张量 Schatten- p 范数, 充分捕获了数据的全局结构、局部结构以及高阶相关性. 在 8 个数据集上与 10 个对比算法进行了实验对比, 实验结果表明了所提出方法的聚类性能优异.

参考文献 (References)

[1] Chen M, Zhang J H, Zhang C, et al. Density change

consistency clustering from density extreme[J]. *Pattern Recognition*, 2025, 158: 110932.

[2] Chen M, Yao Y Y, You Y Y X, et al. Consensus multi-view subspace clustering based on graph filtering[J]. *Neurocomputing*, 2024, 591: 127742.

[3] Pan B C, Li C D, Che H J, et al. Low-rank tensor regularized graph fuzzy learning for multi-view data processing[J]. *IEEE Transactions on Consumer Electronics*, 2024, 70(1): 2925-2938.

[4] Wen J, Zhang Z, Fei L K, et al. A survey on incomplete multiview clustering[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, 53(2): 1136-1149.

[5] 李骛, 陈嘉佳, 于晓洋, 等. 不完备数据的鲁棒多视图学习及其聚类应用[J]. *控制与决策*, 2022, 37(12): 3251-3258.
(Li A, Chen J J, Yu X Y, et al. Robust multiview graph learning with application to clustering for incomplete data[J]. *Control and Decision*, 2022, 37(12): 3251-3258.)

[6] Yang J H, Fu L L, Chen C, et al. Cross-view graph matching for incomplete multi-view clustering[J]. *Neurocomputing*, 2023, 515: 79-88.

[7] Liu X W, Zhu X Z, Li M M, et al. Efficient and effective incomplete multi-view clustering[C]. *Proceedings of the AAAI Conference on Artificial Intelligence*. Hawaii, 2019: 4392-4399.

[8] Fang X, Hu Y C, Zhou P, et al. Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat[J]. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022, 6(4): 913-927.

[9] Wang Z Y, Li L S, Ning X, et al. Incomplete multi-view clustering via structure exploration and missing-view inference[J]. *Information Fusion*, 2024, 103: 102123.

[10] Wang H, Zong L L, Liu B, et al. Spectral perturbation meets incomplete multi-view data[J/OL]. 2019, arXiv: 1906.00098.

[11] Ji G Y, Lu G F, Cai B, et al. Unbalanced incomplete multi-view clustering based on low-rank tensor graph learning[J]. *Expert Systems with Applications*, 2023, 225: 120055.

[12] Sun S L, Zhang N. Incomplete multiview nonnegative representation learning with graph completion and adaptive neighbors[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, 35(3): 4017-4031.

[13] 陈梅, 马学艳, 钱罗雄, 等. 基于多级自表示约束的不完备多视图聚类[J]. *控制与决策*, 2025, 40(2): 645-654.
(Chen M, Ma X Y, Qian L X, et al. Incomplete multi-view clustering based on multi-level self-representation constraints[J]. *Control and Decision*, 2025, 40(2): 645-654.)

[14] Liang C, Wang D Y, Zhang H X, et al. Robust tensor subspace learning for incomplete multi-view clustering[J]. *IEEE Transactions on Knowledge and*

- [Data Engineering](#), 2024, 36(11): 6934-6948.
- [15] Pan B C, Li C D, Che H J. Error-robust multi-view subspace clustering with nonconvex low-rank tensor approximation and hyper-Laplacian graph embedding[J]. [Engineering Applications of Artificial Intelligence](#), 2024, 133: 108274.
- [16] Zhang C Q, Fu H Z, Liu S, et al. Low-rank tensor constrained multiview subspace clustering[C]. IEEE International Conference on Computer Vision. Santiago, 2015: 1582-1590.
- [17] Xie Y, Tao D C, Zhang W S, et al. On unifying multi-view self-representations for clustering by tensor multi-rank minimization[J]. [International Journal of Computer Vision](#), 2018, 126(11): 1157-1179.
- [18] Xia W, Zhang X D, Gao Q X, et al. Multiview subspace clustering by an enhanced tensor nuclear norm[J]. [IEEE Transactions on Cybernetics](#), 2022, 52(9): 8962-8975.
- [19] Yu Z Q, Ye M, Xiao S Y, et al. Learning missing instances in latent space for incomplete multi-view clustering[J]. [Knowledge-Based Systems](#), 2022, 250: 109122.
- [20] Wang Q, He X, Jiang X, et al. Robust bi-stochastic graph regularized matrix factorization for data clustering[J]. [IEEE Transactions on Pattern Analysis and Machine Intelligence](#), 2022, 44(1): 390-403.
- [21] Xia W, Gao Q X, Wang Q Q, et al. Tensor completion-based incomplete multiview clustering[J]. [IEEE Transactions on Cybernetics](#), 2022, 52(12): 13635-13644.
- [22] Luo Q L, Yang M, Li W, et al. Hyper-Laplacian regularized multi-view clustering with exclusive L21 regularization and tensor log-determinant minimization approach[J]. [ACM Transactions on Intelligent Systems and Technology](#), 2023, 14(3): 1-29.
- [23] Zhao H, Liu H, Fu Y. Incomplete multi-modal visual data grouping[C]. Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, 2016: 2392-2398.
- [24] Hu M L, Chen S C. Doubly aligned incomplete multi-view clustering[J/OL]. 2019, arXiv: 1903.02785.
- [25] Li Z L, Tang C, Zheng X, et al. High-order correlation preserved incomplete multi-view subspace clustering[J]. [IEEE Transactions on Image Processing](#), 2022, 31: 2067-2080.
- [26] Deng S J, Wen J, Liu C L, et al. Projective incomplete multi-view clustering[J]. [IEEE Transactions on Neural Networks and Learning Systems](#), 2024, 35(8): 10539-10551.
- [27] Sun L L, Wen J, Liu C L, et al. Balance guided incomplete multi-view spectral clustering[J]. [Neural Networks](#), 2023, 166: 260-272.
- [28] Zhang Z, He W J. Tensorized topological graph learning for generalized incomplete multi-view clustering[J]. [Information Fusion](#), 2023, 100: 101914.
- [29] Cai Y R, Che H J, Pan B C, et al. Projected cross-view learning for unbalanced incomplete multi-view clustering[J]. [Information Fusion](#), 2024, 105: 102245.
- [30] van der Maaten L, Hinton G. Visualizing data using t -SNE[J]. [Journal of Machine Learning Research](#), 2008, 9(11): 2579-2605.

作者简介

陈梅 (1973-), 女, 教授, 博士, 主要研究方向为人工智能、数据挖掘、复杂网络, E-mail: mei.chen.lzjtu@hotmail.com;

郭爱霞 (1999-), 女, 硕士生, 主要研究方向为复杂数据挖掘和图学习聚类, E-mail: 3042996124@qq.com;

王洁 (1999-), 女, 硕士生, 主要研究方向为复杂数据挖掘和图学习聚类, E-mail: 3276044359@qq.com;

杨嘉怡 (2001-), 女, 硕士生, 主要研究方向为复杂数据挖掘和图学习聚类, E-mail: 3114993637@qq.com;

詹苏宝 (2000-), 女, 硕士生, 主要研究方向为复杂数据挖掘和图学习聚类, E-mail: 452024856@qq.com.