控制与决策
Control and Decision

基于多级自表示约束的不完备多视图聚类

陈梅, 马学艳, 钱罗雄, 张锦宏, 张弛

引用本文:
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.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2024.0055

基于多级自表示约束的不完备多视图聚类

陈 梅†, 马学艳, 钱罗雄, 张锦宏, 张 弛

(兰州交通大学 电子与信息工程学院，兰州 730000)

摘 要: 针对现有的不完备多视图聚类方法存在无法准确利用缺失视图的潜在信息和未能充分利用视图间的互补信息以及高阶相关性等问题,提出一种新的基于多级自表示约束的不完备多视图聚类(CMLC)方法. CMLC利用公共潜在表示来恢复缺失值,从而有效获取缺失部分的潜在信息. 为了获得多视图数据的统一低秩表示,CMLC首先通过多级自表示约束捕获多视图数据内部的一致信息和视图间的互补信息,同时利用多级误差表示提高模型对噪声的鲁棒性,接着通过张量对数行列式捕获视图间的高阶相似信息,最后引入距离正则项捕获数据的局部信息. 与9个对比方法在多种缺失率下的6个仿真不完备多视图数据集上进行实验对比,结果表明CMLC均获得了最好的聚类性能.

关键词: 不完备多视图聚类；一致表示；张量对数行列式；低秩张量；不完备数据；张量分析

中图分类号: TP311 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.0055

引用格式: 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.

Incomplete multi-view clustering based on multi-level self-representation constraints

CHEN Mei†, MA Xue-yan, QIAN Luo-xiong, ZHANG Jin-hong, ZHANG Chi

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

Abstract: The existing incomplete multi-view clustering methods have some problems, such as failing to make full use of the potential information of missing views, and failing to make full use of the complementary information and high order correlation among views. In this paper, a new incomplete multi-view clustering method based on multi-level self representation constraints (CMLC) is proposed. The CMLC uses the common latent representation to recover the missing value and thus effectively obtain the latent information of the missing part. In order to obtain a unified low-rank representation of multi-view data, the CMLC first captures the consistent information within the multi-view data and the complementary information between views through multi-level self-representation constraints. At the same time, it uses multi-level error representation to improve the robustness of the model against noise, and then captures the higher-order similar information between views through the logarithmic tensor. Finally, the distance regular term is introduced to capture the local information of the data. The results show that the CMLC has the best clustering performance compared with nine methods on six imperfectly simulated multi-view datasets with different miss rates.

Keywords: incomplete multi-view clustering；consensus representation；tensor log-determinant；low rank tensor；incomplete data；tensor analysis

0 引 言

信息技术的飞跃发展带来了多视图数据的普及,即从多个角度捕捉单一对象信息. 随之,挖掘这些数据背后的复杂结构信息成为紧迫任务. 因此,多视图聚类技术显得至关重要,它能整合多源信息,从而挖掘多视图数据内部的结构信息.

为了捕获多视图数据中存在的结构信息,子空间聚类方法利用数据的自表示性构建各视图的低维表征矩阵,从而在多个视图中捕获更准确的数据结构[1-4]. Xue等[2]采用深度矩阵分解捕获数据中的多层结构; Zhang等[3]通过引入各视图共享的潜在表示探索数据集内部的低秩结构. 文献[4]在文献[3]的基础上将各视图的仿射图分为潜在表示和正交矩阵,保留每个视图中的潜在低秩结构,从而强化不同视图间

收稿日期: 2024-01-11；**录用日期:** 2024-06-02.

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

†通讯作者. E-mail: chenmeilz@mail.lzjtu.cn.

的一致性. 上述方法都仅用矩阵表示各视图的表征图,通过捕捉它们中的二维结构信息进行聚类,因此会忽略多视图间的高阶相似信息,从而出现聚类结果不准确以及对噪声敏感等问题.

由于张量可以表示高维数据结构,研究人员通过张量低秩约束捕获多视图数据的高阶相似信息[5-9]. Zhang等[5]通过对所有视图的子空间表示矩阵叠加的张量施加低秩约束来捕捉多视图数据的高阶相似信息. 然而,该方法不能获得Tucher秩的紧凸松弛,且缺乏明确的物理意义. Xie等[9]提出了一种基于t-SVD (tensor-singular value decomposition)的多视图子空间聚类方法,它有效地将离散傅里叶变换与张量优化问题联系起来,从而得到张量的最紧凸松弛. 但是这种方法忽略了不同大小奇异值的贡献差异性,有可能保留不重要特征的奇异值,而放弃具有高判别特性的奇异值. 而且在真实的应用中,通常数据会有缺失,现有的多视图聚类方法无法处理含有缺失数据的数据集.

为了充分利用缺失数据的潜在信息,学者们提出了很多不完备多视图聚类方法[10-15],其缺失值处理策略大致可分为两类. 一类是在数据预处理阶段填充缺失值[10-12]. Shao等[10]、Gao等[11]均使用各视图存在样本的均值填充缺失值; Wang等[12]通过学习完整的相似度矩阵弥补失去的信息. 这类方法填充的数据可能不够准确. 另一类是在学习共识表示的过程中填充缺失值[13-15]. Hu等[13]强制对齐基矩阵以减少缺失数据的负面影响; Liu等[14]采用矩阵分解获得潜在因子,并将视图特定的相似性矩阵集成为一个一致性矩阵; Li等[15]使用低秩张量正则化恢复完整的kNN图. 这类方法受噪声影响严重.

综上所述,现有的方法仍然具有以下局限性: 1) 不能准确发现和利用缺失视图的信息; 2) 忽略了视图间的高阶相关性; 3) 没有充分利用视图间互补信息. 为解决这些问题,本文提出一种新的聚类方法: 基于多级自表示约束的不完备多视图聚类(CMLC). 该方法利用公共潜在表示恢复缺失值有效获取缺失部分的潜在信息. 为了得到统一低秩表示,该方法通过多级自表示约束捕获多视图数据内的一致信息和视图间的互补信息. 同时,考虑到视图间的高阶相似信息,本文引入张量对数行列式、低秩张量正则化增强恢复数据的对齐性,保证潜在表示的一致性,而且对于对数行列式可避免过度惩罚大的奇异值可能导致期望的秩近似不准确的问题. 为了捕获数据的局部信息,CMLC引入距离正则项,使同簇内的样本间相似度小而异簇内的样本间相似度大.

1 相关理论

1.1 潜在多视图子空间聚类

LMSC(latent multi-view subspace clustering)[3]是一种典型的潜在多视图子空间聚类算法. 给定多视图数据集$\{X^{(v)}\in\mathbb{R}^{d_v\times n}\}(v=1,2,\dots,m)$,LMSC的目标函数如下:

$$\min_{W,H,Z,E_h,E_r}\|E_h\|_{2,1}+\lambda_1\|E_r\|_{2,1}+\lambda_2\|Z\|_*;$$
$$\text{s.t. } X=PH+E_h,\ H=HZ+E_r,\ PP^{\mathrm{T}}=I. \quad (1)$$

其中: H为潜在表示, P为线性重构模型, Z为低秩表示矩阵, E_h和E_r为误差矩阵, λ_1和λ_2为平衡参数.

LMSC通过低秩表示捕获数据的全局结构,但忽略了数据点之间的邻域关系,不能充分利用视图间的互补信息,且无法处理不完备数据.

1.2 基于t-SVD张量分解的多视图谱聚类

基于t-SVD的图表示方法将聚类问题转化为张量核范数最小化问题,从而有效捕捉各视图间的高阶相似信息. Xie等[9]提出了一种基于t-SVD的多视图子空间聚类算法,目标函数如下:

$$\begin{aligned}&\lim_{\mathcal{L},\mathcal{E}}\|\mathcal{A}\|_{\circledast}+\lambda\|E\|_{2,1}.\\&\text{s.t. } X^{(v)}=X^{(v)}\mathcal{A}^{(v)}+E^{(v)},\ v=1,2,\dots,m;\\&\quad\mathcal{A}=\Phi(A^{(1)},A^{(2)},\dots,A^{(v)});\\&\quad E=[E^{(1)},E^{(2)},\dots,E^{(v)}].\end{aligned}\quad (2)$$

其中: $\Phi(\ )$将所有视图的表示矩阵A合并为一个三阶张量结构$\mathcal{A}$,张量$\mathcal{A}$通过旋转操作探索视图之间的高阶低秩相关性; $E^{(v)}$为误差表示矩阵; 张量核范数$\|\mathcal{A}\|_{\circledast}=\sum_{k=1}^{n_3}\|\mathcal{A}_f^{(k)}\|_*=\sum_{i=1}^{\min(n_1,n_2)}\sum_{k=1}^{n_3}|\mathcal{S}_f^{(k)}(i,i)|$.

这种方法不能利用矩阵中的先验知识,同时忽略了不同大小奇异值的贡献差异性.

2 方 法

2.1 视图恢复和多级图学习

CMLC通过学习公共潜在表示来捕获不同视图间的互补信息,从而提高数据表示的准确性和鲁棒性. 为了有效弥补缺失样本包含的结构信息,CMLC首先通过各视图的存在样本初始化公共潜在表示,再通过初始化值重构样本集. 具体地, n_v为第v个视图存在的样本数,设$\widetilde{X}^{(v)}\in\mathbb{R}^{d_v\times n_v}$为第$v$个视图未缺失样本集, n为总样本数,期望重构相应的完整样本$X^{(v)}\in\mathbb{R}^{d_v\times n}$,即

$$X^{(v)}=P^{(v)}H+E_1^{(v)},\widetilde{X}^{(v)}=X^{(v)}W^{(v)}. \quad (3)$$

其中: $H\in\mathbb{R}^{k\times n}$为公共潜在表示, k为特征维数; $P^{(v)}\in\mathbb{R}^{d_v\times n}$为第$v$个视图的重构模型; $E_1^{(v)}$为第v个视图的重构误差; $W^{(v)}\in\mathbb{R}^{d_v\times n}$为索引矩阵. 特别需

要指出的是, $P^{(v)}$ 在初始化过程中维度变为 $(d_v \times k)$. $W^{(v)}$ 具体如下:

$$W^{(v)} = \begin{cases} 1,\ X_{(i)}^{(v)} = \tilde{X}_{(j)}^{(v)}; \\ 0,\ \text{Otherwise.} \end{cases} \tag{4}$$

约束 $\widetilde{X}^{(v)} = X^{(v)}W^{(v)}$ 从现有样本中的潜在表示重构样本集, 保证了潜在表示中的区分信息并有效恢复了缺失视图的部分结构信息.

多视图数据通常具有复杂的变异性和多样性, 如果模型只学习各视图子空间的相似度图或只学习各视图的公共潜在表示, 则难以充分利用多视图数据的复杂结构信息. 因此, CLMC把潜在表示学习与各视图的子空间表示学习有效结合, 将潜在表示 Z 称为第 $m+1$ 个亲和矩阵, 这时 Z 与 m 个亲和矩阵统称为多级图表示, 具体为

$$H = HZ + E_2,\ X^{(v)} = X^{(v)}Q^{(v)} + E_3^{(v)}. \tag{5}$$

其中: Z 为基于潜在表示学习的亲和矩阵, $Q^{(v)}$ 为基于子空间表示学习的亲和矩阵.

2.2　视图间的高阶相似性

现有的基于张量的不完备多视图聚类方法常用张量核范数约束, 它通过张量旋转操作探索各视图间的高阶相关信息[9]. 由于张量核范数与奇异值的和成正比, 当一些奇异值很大而另一些奇异值很小时, 采用张量核范数可能会过度抵消大的奇异值, 导致期望的张量秩近似不准确. 因此, 为了更好地最小化张量秩, 本文引入张量对数行列式约束项对多级自表示进行约束, 使用对数行列式代替核范数进行秩近似, 对奇异值进行对数约束, 减少奇异值之间的差距, 避免过度惩罚大奇异值, 从而克服张量核范数的缺点. 具体地, 本文将 m 个亲和矩阵与潜在表示 Z 堆叠得到三维张量 $\mathcal{Q} \in \mathbb{R}^{n\times n\times(m+1)}$, $\mathcal{Q} = \Phi(Q^{(1)}, \ldots, Q^{(v)}, Q^{(m+1)})$, 并将其旋转为 $\mathcal{Q} \in \mathbb{R}^{n\times(m+1)\times n}$. 对数张量核范数定义为

$$\|\mathcal{Q}\|_{T\text{Log}} = \sum_{i=1}^{m+1} \log(I + Q_f^{(i)\mathrm{T}} Q_f^{(i)}) = \sum_{i=1}^{m+1}\sum_{k=1}^{n} \log|1 + S_f^2(i,i,k)|. \tag{6}$$

其中: $Q_f^{(i)}$ 为张量 $\mathcal{Q}$ 的第 i 个正切片; S_f 为张量 $\mathcal{Q}$ 沿着第3维的快速傅里叶变换, 即 $S_f = \text{fft}(Q, [\,], 3)$.

2.3　数据集的局部流形结构

实际获取的数据集具有非线性流形结构, 因此, 为了捕获各视图的局部几何结构, CMLC引入局部流形结构约束项, 使同一簇内的样本间相似度大, 不同簇内的样本间相似度小, 具体如下:

$$\min \sum_{v=1}^{m}\sum_{i,j=1}^{N} \|x_i - x_j\|_2^2 q_{ij}^{(v)}, \tag{7}$$

其中 $q_{ij}^{(v)}$ 为 x_i 与 x_j 之间的相似关系.

模型目标是求解出式(7)的最小值, 局部流形结构约束项通过捕捉样本间的邻域关系, 使同簇内的样本间相似度大, 异簇内的样本间相似度小.

2.4　总体目标函数

为了处理数据集中包含的噪声, CMLC对多级图表示的每一级均施加了 $l_{2,1}$ 噪声约束, 称为多级误差表示. 它可以在多个角度约束噪声, 从而增强模型的鲁棒性. 根据式(3)、(5)～(7), CMLC的总体目标函数表示如下:

$$\begin{aligned} &\lim_{\Omega} \|\mathcal{Q}\|_{T\text{Log}} + \beta\sum_{v=1}^{m}\sum_{i,j=1}^{N} \|x_i - x_j\|_2^2 q_{ij}^{(v)} + \\ &\sum_{v=1}^{m} \lambda_1 \|E_1^{(v)}\|_{2,1} + \lambda_2 \|E_2\|_{2,1} + \sum_{v=1}^{m} \lambda_3 \|E_3^{(v)}\|_{2,1}; \\ &\text{s.t. } X^{(v)} = P^{(v)}H + E_1^{(v)},\ \widetilde{X}^{(v)} = X^{(v)}W^{(v)}, \\ &\quad H = HZ + E_2,\ X^{(v)} = X^{(v)}Q^{(v)} + E_3^{(v)}, \\ &\quad P^{(v)}P^{(v)\mathrm{T}} = I,\ \mathcal{Q} = \Phi(Q^{(1)}, \ldots, Q^{(v)}, Q^{(m+1)}). \end{aligned} \tag{8}$$

其中: $\Omega = \{X^{(v)}, P^{(v)}, H, E_1^{(v)}, E_2, E_3^{(v)}, Q^{(v)}, Z, \mathcal{Q}\}$, λ_1、λ_2、λ_3 和 β 为惩罚项参数, $P^{(v)}P^{(v)\mathrm{T}} = I$ 可避免过小表示出现, $\Phi(\cdot)$ 表示堆叠所有亲和矩阵并旋转.

得到式(8)的解之后, CMLC通过多级图融合得到最终的相似度矩阵, 即

$$S = \frac{1}{m+1}\Big(\frac{|Z| + |Z^{\mathrm{T}}|}{2} + \sum_{v=1}^{m} \frac{|Q^{(v)}| + |Q^{(v)\mathrm{T}}|}{2}\Big). \tag{9}$$

3　模型优化

3.1　目标模型优化

本节提出一种迭代交替优化方法求解目标函数(8). 引入辅助变量 G, 并将目标函数(8)的求解问题转化为其增广拉格朗日函数的最小化问题, 即

$$\begin{aligned} &\mathcal{L}(\Omega) = \\ &\|\mathcal{Q}\|_{T\text{Log}} + \beta\sum_{v=1}^{m}\sum_{i,j=1}^{N} \|x_i - x_j\|_2^2 q_{ij}^{(v)} + \lambda_2\|E_2\|_{2,1} + \\ &\sum_{v=1}^{m} \lambda_1 \|E_1^{(v)}\|_{2,1} + \frac{\mu}{2}\sum_{v=1}^{m} \|G^{(v)} - Q^{(v)} + F_5^{(v)}/\mu\|_F^2 + \\ &\frac{\mu}{2}\sum_{v=1}^{m} \|X^{(v)} - P^{(v)}H - E_1^{(v)} + F_1^{(v)}/\mu\|_F^2 + \\ &\sum_{v=1}^{m} \lambda_3 \|E_3^{(v)}\|_{2,1} + \frac{\mu}{2}\|H - HZ - E_2 + F_2/\mu\|_F^2 + \\ &\frac{\mu}{2}\|Z - Q^{(m+1)} + F_4^{(v)}/\mu\|_F^2 + \end{aligned}$$

$$\frac{\mu}{2}\sum_{v=1}^{m}\|X^{(v)}-X^{(v)}G^{(v)}-E_3^{(v)}+F_3^{(v)}/\mu\|_F^2;$$
$$\text{s.t. } \widetilde{X}^{(v)}=X^{(v)}W^{(v)},\ P^{(v)}P^{(v)\mathrm{T}}=I. \tag{10}$$

其中: $F_1^{(v)}$、F_2、$F_3^{(v)}$、$F_4^{(v)}$ 和 $F_5^{(v)}$ 为拉格朗日乘子, $\mu>0$ 为惩罚因子.

本文将相关变量进行固定,对问题(10)中的剩余变量逐个优化得到最优解,详细优化过程如下:

1) 更新 P,固定其他变量,令 $M^{(v)}=X^{(v)}-E_1^{(v)}+F_1^{(v)}/\mu$,利用下式求解 P:

$$\min_{P^{(v)}}\sum_{v=1}^{m}\|M^{(v)}-P^{(v)}H\|_F^2;\ \text{s.t. } P^{(v)}P^{(v)\mathrm{T}}=I. \tag{11}$$

最优解为 $P^{(v)}=VU^{\mathrm{T}}$. 其中: U 和 V 由SVD操作计算,即 $HM^{(v)\mathrm{T}}=U\mathfrak{T}V^{\mathrm{T}}$.

2) 更新 H,固定其他变量, H 的求解变为

$$\min_{H}\sum_{v=1}^{m}\|M^{(v)}-P^{(v)}H\|_F^2+\|H-HZ-E_2+F_2^{(v)}/\mu\|_F^2. \tag{12}$$

对式(12)求 H 的偏导数,并令其为零,可得

$$\Big(\sum_{v}P^{(v)\mathrm{T}}P^{(v)}\Big)H+H(I-Z)(I-Z)^{\mathrm{T}}=N. \tag{13}$$

问题(13)为典型的西尔维斯特方程(Sylvester equation),可以使用Bartels−Stewart算法求解[3].

3) 更新 X,固定其他变量,变量 X 的子问题为

$$\sum_{v=1}^{m}\|X^{(v)}-P^{(v)}H-E_1^{(v)}+F_1^{(v)}/\mu\|_F^2+\sum_{v=1}^{m}\|X^{(v)}-X^{(v)}G^{(v)}-E_3^{(v)}+F_3^{(v)}/\mu\|_F^2;$$
$$\text{s.t. } \widetilde{X}^{(v)}=X^{(v)}W^{(v)}. \tag{14}$$

求 X 的偏导数,并令其为零,可得如下解:

$$X^{(v)}=U(I+(I-G^{(v)})(I-G^{(v)})^{\mathrm{T}}). \tag{15}$$

其中: $U=P^{(v)}H+E_1^{(v)}-F_1^{(v)}/\mu+(E_3^{(v)}-F_3^{(v)}/\mu)\cdot(I-G^{(v)})^{\mathrm{T}}$,由约束 $\widetilde{X}^{(v)}=X^{(v)}W^{(v)}$ 更新缺失数据.

4) 更新 Q,固定其他变量,变量 Q 的子问题为

$$\beta\sum_{v=1}^{m}\sum_{i,j=1}^{N}\|x_i-x_j\|_2^2q_{ij}^{(v)}+\|Z-Q^{(m+1)}+F_4^{(v)}/\mu\|_F^2+\sum_{v=1}^{m}\|G^{(v)}-Q^{(v)}+F_5^{(v)}/\mu\|_F^2. \tag{16}$$

令 $d_{ij}^{(v)}=\|x_i-x_j\|_2^2$,其中 $d_{ij}^{(v)}$ 为第 v 个视图的数据点 i 与 j 之间的欧氏距离. 则式(16)变为

$$\operatorname*{argmin}_{Q^{(v)}}\beta\mathrm{Tr}\,(D^{(v)}Q^{(v)})+\frac{\mu}{2}\|Z-Q^{(m+1)}+F_4^{(v)}/\mu\|_F^2+\frac{\mu}{2}\sum_{v=1}^{m}\|G^{(v)}-Q^{(v)}+F_5^{(v)}/\mu\|_F^2. \tag{17}$$

令 $M_1=Z+F_4^{(v)}/\mu$, $M_2=G^{(v)}+F_5^{(v)}/\mu$,则 $\widehat{Q}^{(v)}$ 为

$$\hat{Q}^{(v)}=\beta\mathrm{Tr}(D^{(v)}Q^{(v)})+\frac{\mu}{2}\|M_1-Q^{(m+1)}\|_F^2+\frac{\mu}{2}\sum_{v=1}^{m}\|M_2-Q^{(v)}\|_F^2=(\mu X^{(v)\mathrm{T}}X^{(v)}+\mu I)^{-1}[\mu X^{(v)\mathrm{T}}M_1+\mu M_2-\beta D^{(v)}]. \tag{18}$$

求解如下最小化问题得到 $\widehat{Q}^{(v)}$ 的最优解

$$\min\|Q^{(v)}-\widehat{Q}^{(v)}\|_F^2. \tag{19}$$

采用逐行优化法求解 $Q^{(v)}$ 各行的最优值,即

$$q_i^{(v)}=\max(\xi_i\overline{1}^{\mathrm{T}}+\overline{q}_i^{(v)},0). \tag{20}$$

其中: $\overline{Q}_i^{(v)}=[\overline{Q}_{i1}^{(v)},\ldots,\overline{Q}_{ii}^{(v)},\ldots,\overline{Q}_{in}^{(v)}]$, $q_i^{(v)}$ 为潜在解 $\widehat{Q}^{(v)}$ 的第 i 行,拉格朗日乘子 ξ_i 的解为

$$\xi_i=\frac{1-\overline{q}_i^{(v)}1}{n-1}. \tag{21}$$

5) 更新 $\mathcal{Q}$,变量 $\mathcal{Q}$ 的子问题为

$$\min_{\mathcal{Q}}\|\mathcal{Q}\|_{T\mathrm{Log}}+\frac{\mu}{2}\|Z-Q^{(m+1)}+F_4^{(v)}/\mu\|_F^2+\frac{\mu}{2}\sum_{v=1}^{m}\|G^{(v)}-Q^{(v)}+F_5^{(v)}/\mu\|_F^2;$$
$$\text{s.t. } \mathcal{Q}=\Phi(Q^{(1)},\ldots,Q^{(m)},Q^{(m+1)}). \tag{22}$$

令 $\mathcal{T}=\Phi(G^{(1)}-F_5^{(v)}/\mu,\ldots,G^{(m)}-F_5^{(v)}/\mu,Z-F_4^{(v)}/\mu)$,则式(22)变为

$$\min_{\mathcal{Q}}\|\mathcal{Q}\|_{T\mathrm{Log}}+\frac{\mu}{2}\|\mathcal{Q}-\mathcal{T}\|_F^2, \tag{23}$$

则有

$$Q_f^{(k)*}=\arg\min_{Q_f^{(k)}}\|Q_f^{(k)}\|_{T\mathrm{Log}}+\frac{\mu}{2}\|Q_f^{(k)}-T_f^{(k)}\|_F^2. \tag{24}$$

其中: $Q_f^{(k)}$ 为张量 $\mathcal{Q}$ 第 k 个正切片, $T_f^{(k)}$ 为张量 $\mathcal{T}$ 第 k 个正切片. $\sigma_i(T_f^{(k)})$ 为张量 $\mathcal{T}$ 第 k 个切片的第 i 个最大奇异值, $\sigma_i^{(k)}$ 为张量 $\mathcal{Q}$ 第 k 个切片第 i 个最大奇异值.

令 $T_f^{(k)}=\bar{\mathcal{U}}_{T_f}^{(k)}\cdot\bar{S}_{T_f}^{(k)}\cdot\bar{\mathcal{V}}_{T_f}^{(k)\mathrm{T}}$,则有 $\Delta=\bar{\mathcal{U}}_{T_f}^{(k)\mathrm{T}}\cdot\overline{Q}_f^{(k)}\cdot\bar{\mathcal{V}}_{T_f^{(k)}}$. 可知 $\overline{Q}_f^{(k)\mathrm{T}}$ 与 $\overline{Q}_f^{(k)}$ 具有相同的特征值, Δ 与 $\overline{Q}_f^{(k)}$ 具有相同的特征值. S_Δ 和 $S_{Q_f}^{(k)}$ 分别为 Δ 和 $\overline{Q}_f^{(k)}$ 的奇异值矩阵,则 $S_\Delta=S_{Q_f}^{(k)}$. 根据Von-Neumann的迹不等式和 F 范数的酉不变性可得 $\|T_f^{(k)}-Q_f^{(k)}\|_F^2=\|S_{T_f}^{(k)}-\Delta\|_F^2\geqslant\|S_{T_f}^{(k)}-S_\Delta\|_F^2=\|S_{T_f}^{(k)}-S_{Q_f}^{(k)}\|_F^2$,式(24)的相应求解可变为

$$\|Q_f^{(k)}\|_{T\mathrm{Log}}+\frac{\mu}{2}\|Q_f^{(k)}-T_f^{(k)}\|_F^2\geqslant\sum_{i=1}^{N}\log(1+\sigma_i^{(k)})+\frac{\mu}{2}[\sigma_i^{(k)}-\sigma_i(T_f^{(k)})]^2, \tag{25}$$

则式(24)的求解可变为

$$\arg\min\sum_{i=1}^{N}\log(1+\sigma_i^{(k)2})+\frac{\mu}{2}[\sigma_i^{(k)}-\sigma_i(T_f^{(k)})]^2. \tag{26}$$

对式(26)求σ_i的偏导数,并令其为0,可得

$$\frac{2\sigma_i^{(k)}}{1+{\sigma_i^{(k)}}^2}+\frac{1}{\rho}(\sigma_i^{(k)}-\sigma_i(T_f^{(k)}))=0;$$
$$\text{s.t. } \sigma_i^{(k)}\geqslant 0,\ i=1,2,\ldots,V. \tag{27}$$

6) 更新Z和G,变量Z和G的子问题为

$$\min_Z \|K-HZ\|_F^2+\|Z-Q^{(m+1)}+F_4^{(v)}/\mu\|_F^2; \tag{28}$$

$$\min_G \sum_{v=1}^{m}\|J^{(v)}-X^{(v)}G^{(v)}\|_F^2+\sum_{v=1}^{m}\|G^{(v)}-Q^{(v)}+F_5^{(v)}/\mu\|_F^2. \tag{29}$$

其中: $K=H-E_2+F_2^{(v)}/\mu, J^{(v)}=X^{(v)}-E_3^{(v)}+F_3^{(v)}/\mu$.

对式(28)、(29)分别求Z和G的偏导,并令其为0,可得如下最优解:

$$Z=(H^{\mathrm{T}}H+I)^{\mathrm{T}}(H^{\mathrm{T}}K+Q^{(m+1)}-F_4^{(v)}/\mu), \tag{30}$$

$$G^{(v)}=({X^{(v)}}^{\mathrm{T}}X^{(v)}+I)^{\mathrm{T}}({X^{(v)}}^{\mathrm{T}}J+Q^{(v)}-F_5^{(v)}/\mu). \tag{31}$$

7) 更新多级误差表示矩阵$E_1^{(v)}$、E_2和$E_3^{(v)}$,有

$$\min_{E_1^{(v)}}\frac{\lambda_1}{\mu}\|E_1^{(v)}\|_{2,1}+\frac{1}{2}\sum_{v=1}^{m}\|E_1^{(v)}-(X^{(v)}-P^{(v)}H+F_1^{(v)}/\mu)\|_F^2; \tag{32}$$

$$\min_{E_2^{(v)}}\frac{\lambda_2}{\mu}\|E_2\|_{2,1}+\frac{1}{2}\|E_2-(H-HZ+F_2^{(v)}/\mu)\|_F^2, \tag{33}$$

$$\min_{E_2^{(v)}}\frac{\lambda_3}{\mu}\|E_3^{(v)}\|_{2,1}+\frac{1}{2}\sum_{v=1}^{m}\|E_3^{(v)}-(X^{(v)}-X^{(v)}G^{(v)}+F_3^{(v)}/\mu)\|_F^2. \tag{34}$$

这3项采用相同形式更新,即

$$\min_W \tau\|W\|_{2,1}+\frac{1}{2}\|W-A\|_F^2. \tag{35}$$

假设最优解为$\widehat{W}$,则$\widehat{W}$的第i列为

$$\widehat{W}_{:,i}=\begin{cases}\dfrac{\|A_{:,i}\|_2-\tau}{\|A_{:,i}\|_2}A_{:,i}, & \text{if } \|A_{:,i}\|_2>\tau;\\ 0, & \text{otherwise.}\end{cases} \tag{36}$$

通过上述方式更新$E_1^{(v)}$、E_2和$E_3^{(v)}$.

8) 更新乘子和惩罚因子,即

$$\begin{cases}F_1^{(v)}=F_1^{(v)}+\mu R_1^{(v)},\ R_1^{(v)}=X^{(v)}-P^{(v)}H-E_1^{(v)};\\ F_2=F_2+\mu R_2,\ R_2=H-HZ-E_2;\\ F_3^{(v)}=F_3^{(v)}+\mu R_3^{(v)},\\ R_3^{(v)}=X^{(v)}-X^{(v)}G^{(v)}-E_3^{(v)};\\ F_4=F_4+\mu R_4,\ R_4=Z-Q^{(m+1)};\\ F_5^{(v)}=F_5^{(v)}+\mu R_5^{(v)},\ R_5^{(v)}=G^{(v)}-Q^{(v)};\\ \mu=\min(\mu_{\max},\rho\mu).\end{cases} \tag{37}$$

其中: $\rho>1$为正标量, $\mu_{\max}$为预定义μ的最大值.

令$R_*=[R_*^{(1)},R_*^{(2)},\ldots,R_*^{(v)}]$, $*=1,2,\ldots,5$,收敛条件为

$$\|R_i\|_\infty\leqslant\varepsilon,\ i=1,2,\ldots,5, \tag{38}$$

其中ε为最小误差. 通过迭代优化所有步骤,增广拉格朗日函数可以逐渐最小化直到收敛.

3.2 变量初始化

本节初始化潜在表示H,并使用现有样本和恢复的缺失样本初始化重构矩阵$P^{(v)}$和数据集$X^{(v)}$, $\widetilde{X}^{(v)}$为第v个视图的存在样本, $W^{(v)}$为索引矩阵

$$\min_{P^{(v)},H}\sum_{v=1}^{m}\|\widetilde{X}^{(v)}-P^{(v)}HW^{(v)}\|_F^2;$$
$$\text{s.t. } P^{(v)}{P^{(v)}}^{\mathrm{T}}=I. \tag{39}$$

问题(39)通过以下方式更新:

1) 更新$P^{(v)}$. $P^{(v)}$的求解形式与式(11)相同,即

$$P^{(v)}=VU^{\mathrm{T}}. \tag{40}$$

其中U和V由SVD精简分解得出,即$HW^{(v)}(\widetilde{X}^{(v)})^{\mathrm{T}}=U\mathfrak{T}V^{\mathrm{T}}$,此时$U$的维数为$(d_v\times k)$, V的维数为$(k\times k)$,则$P^{(v)}$的维度被初始化为$(d_v\times k)$.

2) 更新H. 使用梯度下降法

$$H=H-\eta_H\nabla_H. \tag{41}$$

其中:梯度$\nabla_H=\sum_{v=1}^{m}{P^{(v)}}^{\mathrm{T}}P^{(v)}HW^{(v)}{W^{(v)}}^{\mathrm{T}}-{P^{(v)}}^{\mathrm{T}}\widetilde{X}^{(v)}W^{(v)}$, η_H为armijo规定的步长.

上述为初始化模型和CMLC总模型的全部优化过程. 算法过程总结如下.

算法1

step 1: 输入数据集$\{X^{(v)}\}_{v=1}^m$,参数λ_1、λ_2、λ_3、$\beta=0.0001$、$\mu=1$、$\rho=1.3$、$\mu_{\max}=10^6$、$\varepsilon=10^{-6}$,初始化H,其他矩阵初始化为全0矩阵.

step 2: 根据式(4)初始化$\{W^{(v)}\}_{v=1}^m$.

step 3: while未收敛do

step 3.1: 根据式(40)更新$P^{(v)}$;

step 3.2: 通过梯度下降法(41)更新H;

step 3.3: end while

step 4: 初始化$X^{(v)}$,从$P^{(v)}H$中恢复缺失实例.

step 5: while未收敛do

step 5.1: 根据式(11)更新$P^{(v)}$;

step 5.2: 根据式(13)更新H;

step 5.3: 根据式(15)更新$X^{(v)}$;

step 5.4: 根据式(18)更新$Q^{(v)}$;

step 5.5: 根据式(27)更新$\mathcal{Q}$;

step 5.6: 根据式(30)和(31)更新Z和$G^{(v)}$;

step 5.7: 根据式(36)更新$E_1^{(v)}$、$E_2^{(v)}$和$E_3^{(v)}$;

step 5.8: 根据式(37)求解乘数和惩罚因子;

step 5.9: end while

step 6: 输出多级图$Z, \{G^{(v)}\}_{v=1}^{m}$.

3.3 时间复杂度分析

初始化过程主要包括更新重构模型$P^{(v)}$和公共潜在表示H. 其中: $P^{(v)}$主要时间花费在SVD运算, 矩阵大小为$d_v \times k$时, 时间复杂度为$O(k^2 d_v)$; 更新H主要时间花销在矩阵乘法, 时间复杂度为$O(mk\tau_0 \cdot (kd_v + kn + d_v n))$, τ_0为梯度下降的迭代次数. 由于$k, m \ll n$, 初始化过程的时间复杂度为$O(\tau_1\tau_0(n))$, 其中τ_1为外循环的迭代次数. 忽略矩阵基本运算时间, 主算法更新$P^{(v)}$所需的SVD运算, 其时间复杂度为$O(k^2 n)$, 更新H时间复杂度为$O(k^3)$, 更新$X^{(v)}, Z$和$G^{(v)}$的时间复杂度为$O(n^3)$, 更新$Q^{(v)}$主要花费在于矩阵逆运算, 这一步在迭代循环外, 所以更新时间忽略不计, 更新$\mathcal{Q}$的快速傅里叶变换和快速傅里叶逆变换所需的时间复杂度为$O((m+1)n^2\log(n) + (m+1)^2 n^2)$. 综上所述, 由于$k, m \ll n$, 总时间复杂度为$O(\tau_2 n^3)$, 其中$\tau_2$为迭代次数.

4 实 验

4.1 实验设置

1) 数据集: 如表1所示, 本文采用常用的5个图像数据集、一个文档数据集以及一个随机生成的二维多视图噪声数据集评估算法性能. 本文对前6个真实完备数据集按缺失率进行随机删除, 得到仿真不完备多视图数据集, 缺失比率(PER)分别为0.0、0.1、0.3、0.5、0.7、0.9.

表 1 实验数据集描述

数据集	视图数	簇个数	样本数	各视图维数	类型
ORL[1]	4	40	400	256/256/256/256	人脸图像
NGs[2]	3	5	500	2 000/2 000/2 000	新闻文档
100leaves[3]	3	100	1 600	64/64/64	叶片图像
Scene-15	3	15	4 485	20/59/40	场景图像
LandUse-21	3	21	2 100	20/59/40	卫星图像
Uci_digits[4]	3	20	2 000	64/76/216	手写字
Noise_Data	3	3	300	2/2/2	—

注: 1 https://cam-orl.co.uk/facedatabase.html;
2 http://qwone.com/ jason/20Newsgroups/;
3 https://gitee.com/zhangfk/multi-view-dataset;
4 http://archive.ics.uci.edu/dataset/72/multiple+features.

2) 对比方法: 本文选取9个聚类方法作为对比方法, 参数选择范围详见其原论文. 本文使用准确性(accuracy)、归一化互信息(NMI)和纯度(purity)来衡量聚类结果, 其中指标数值越大, 聚类效果越好. 本文提出的CLMC方法的参数β固定为0.000 1, 其余参数选择范围为$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$. 对于所有算法, 由于$k$-Means对初始值有一定依赖性, 将实验运行30次k-Means, 并记录评价指标的平均值.

4.2 实验结果与分析

表2给出了CMLC与对比方法在真实多视图数据集上的聚类结果. 表2中对比算法分别为: PIC (perturbation-oriented incomplete multi-view clustering)[12], DAIMC (doubly aligned incomplete multi-view clustering)[13], BSV (best single view)[16], MIC (multi-incomplete view clustering)[17], TCIMC (tensor completionbased incomplete multi view clustering)[18], TTGL (tensorized topological graph learning for generalized incomplete multi-view clustering)[19], HCPIMSC (high order correlation preserved incomplete multi-view subspace clustering)[20], HCLS_CGL (highly confident local structure based consensus graph learning for incomplete multi-view clustering)[21], LSGMC (multi view subspace clustering via low-rank symmetric affinity graph)[22].

从表2和图1可以得出如下结论:

1) BSV在所有数据集上的表现均低于其他方法. 这表明简单地选择最佳视图会因为缺失其他视图的特有信息以及视图间的互补信息而取得较差的结果. 其余对比方法和CMLC均通过不同方式捕获视图间的互补信息和视图内的特有信息寻求所有视图的低维一致表征. 其中, DAIMC利用给定的样本对齐信息来学习所有视图的共同潜在特征矩阵. MIC学习多个视图的潜在特征矩阵, 通过协同正则化方法生成共识矩阵. HCLS_CGL基于相似近邻的假设设计新的置信图, 然后通过置信结构驱动的共识图学习模型直接学习跨视图的共识图. 相比这3种方法, CMLC取得了最好的结果, 这是因为CMLC通过多级自表示约束充分利用了视图内的一致信息和视图间的互补性信息, 并融合得到一个鲁棒聚类的内在相似矩阵. 此外, CMLC还引入了距离正则项, 能够有效保持数据的局部流形结构.

2) Uci_digits数据集中MIC行的结果远低于MIC# 行, 这是由于Uci_digits数据集包含负视图, MIC方法无法处理负数据, 而本文提出的CMLC方法可以处理负视图并取得优异的结果.

3) 数据集特征缺失的数量能够直接影响聚类效果. 实验结果显示, 本文提出的CMLC方法均优于其他方法. 同时在ORL、NGs和Uci_digits数据集上, 缺失率对CMLC方法的实验结果影响不大. 这是因为CMLC方法利用全部视图共有的潜在表示和已存在样本在多视图的特征空间之间的关系恢复了缺失数

表2　9种对比方法与CMLC在不同多视图数据集的accuracy和NMI　　100%

dataset	metrics	accuracy						NMI					
	PER	0.0	0.1	0.3	0.5	0.7	0.9	0	0.1	0.3	0.5	0.7	0.9
ORL	BSV	5.85	25.55	23.00	24.83	23.55	24.83	13.46	43.07	39.46	44.55	46.43	47.03
	DAIMC	61.00	60.48	56.08	53.38	47.27	43.50	77.93	76.91	73.44	70.96	65.65	61.62
	MIC	62.20	59.15	50.30	44.83	35.40	28.80	78.69	75.15	65.31	57.61	49.10	42.33
	PIC	74.08	72.43	71.12	70.12	66.50	60.38	85.61	84.44	83.25	81.54	79.60	75.96
	TCIMC	77.25	74.00	70.00	66.50	62.25	62.25	87.04	85.95	82.36	82.52	79.08	77.05
	HCLS_CGL	72.25	72.00	67.75	65.00	59.50	51.50	83.82	82.44	79.32	76.10	73.84	70.03
	TTGL	53.75	51.50	49.25	44.25	39.75	32.25	71.51	69.02	65.08	57.87	52.20	50.54
	HCPIMSC	74.99	85.85	85.90	82.85	81.14	79.38	86.50	74.27	74.30	71.40	68.06	66.13
	LSGMC*	79.00	—	—	—	—	—	88.80	—	—	—	—	—
	CMLC	**96.50**	**96.00**	**98.50**	**96.75**	**96.00**	**97.50**	**97.00**	**96.00**	**98.50**	**96.75**	**96.25**	**97.50**
NGs	BSV	27.12	43.08	46.00	39.66	37.94	41.30	11.62	23.90	26.37	17.74	12.09	13.64
	DAIMC	91.02	91.08	77.78	63.12	60.54	55.98	78.73	78.37	60.29	43.87	42.43	33.63
	MIC	25.88	24.40	26.26	26.74	27.68	25.20	5.35	4.13	5.54	5.21	4.42	3.79
	PIC	98.00	97.00	76.54	92.40	90.80	63.48	93.56	90.65	62.36	78.06	74.05	44.15
	TCIMC	97.40	95.80	93.40	89.60	87.20	83.40	91.62	87.55	80.94	72.58	67.34	61.04
	HCLS_CGL	96.40	93.40	90.20	87.60	83.40	81.40	89.67	82.08	74.24	68.10	65.75	60.08
	TTGL	67.20	65.20	65.80	60.20	54.80	54.00	53.60	51.36	48.89	43.49	35.21	34.43
	HCPIMSC	98.40	92.88	85.13	80.27	74.04	72.22	94.60	97.80	95.00	92.80	89.40	89.60
	LSGMC*	79.50	—	—	—	—	—	89.80	—	—	—	—	—
	CMLC	**100.00**	**100.00**	**99.80**	**100.00**	**99.60**	**99.20**	**100.00**	**100.00**	**99.30**	**100.00**	**98.61**	**97.39**
100leaves	BSV	23.71	23.46	21.64	20.21	17.60	16.32	54.17	53.63	47.47	43.59	39.61	35.52
	DAIMC	72.49	63.77	56.14	43.41	34.99	28.19	87.54	84.06	76.28	68.33	63.02	58.13
	MIC	66.46	65.50	56.39	46.77	39.64	32.84	84.88	81.70	74.17	67.35	63.17	59.44
	PIC	86.38	85.07	76.26	69.05	61.04	55.62	93.59	92.81	86.77	81.92	77.00	74.79
	TCIMC	81.31	79.56	42.94	45.56	49.50	65.44	89.37	88.60	69.01	69.82	70.84	78.28
	HCLS_CGL	79.44	78.56	71.31	67.06	63.25	59.69	88.30	87.33	81.04	79.86	77.60	75.40
	TTGL	59.81	57.31	53.06	44.88	40.56	33.88	78.56	75.03	73.20	67.68	64.80	60.12
	HCPIMSC	77.51	88.03	84.58	80.38	77.14	**72.90**	98.40	76.27	72.52	66.52	61.47	54.40
	LSGMC*	85.56	—	—	—	—	—	92.81	—	—	—	—	—
	CMLC	**94.31**	**91.13**	**91.50**	**86.19**	**84.06**	68.94	**98.56**	**96.02**	**96.16**	**93.03**	**91.13**	**83.16**
Scene-15	BSV	9.48	9.39	9.5	9.49	9.53	9.59	0.74	0.58	0.76	0.75	0.79	0.88
	DAIMC	33.03	33.22	31.14	28.37	25.74	23.55	30.48	30.53	27.91	24.68	20.92	18.36
	MIC	33.68	32.33	29.07	26.79	22.83	20.87	34.40	32.41	28.74	25.07	21.16	19.21
	PIC	44.67	44.94	41.53	43.56	41.57	41.01	41.94	41.39	39.63	39.32	35.40	34.81
	TCIMC	25.12	41.05	42.05	43.10	40.89	38.64	21.58	40.45	41.02	39.94	36.60	34.71
	HCLS_CGL	41.56	40.69	39.58	31.42	34.36	35.81	39.99	38.91	36.26	29.59	32.10	30.74
	TTGL	37.64	40.74	33.67	34.54	29.28	28.38	39.14	37.63	33.36	31.10	26.47	25.00
	HCPIMSC	38.45	33.12	33.09	31.20	29.06	28.62	35.04	36.63	37.48	35.03	33.56	32.88
	LSGMC*	46.85	—	—	—	—	—	48.45	—	—	—	—	—
	CMLC	**53.00**	**53.18**	**51.57**	**50.06**	**50.48**	**45.53**	**50.52**	**49.76**	**48.92**	**46.72**	**44.88**	**39.63**
LandUse-21	BSV	5.38	5.29	5.33	5.43	5.62	5.57	2.16	1.87	1.88	1.66	2.38	2.21
	DAIMC	24.43	24.62	22.78	20.68	17.01	16.74	28.74	27.79	24.18	20.96	16.51	15.35
	MIC	23.56	22.40	21.84	19.29	16.00	15.41	28.34	26.53	24.11	20.83	16.37	14.46
	PIC	24.24	23.20	23.93	24.36	25.10	25.23	30.96	28.68	28.77	29.95	29.43	29.04
	TCIMC	19.71	17.28	15.95	14.28	17.57	13.47	25.17	22.11	20.69	18.89	21.54	14.99
	HCLS_CGL	24.61	24.71	24.28	25.76	23.66	22.28	30.44	30.07	29.55	30.43	28.79	24.62
	TTGL	23.14	22.10	20.71	19.33	18.10	16.19	26.17	24.71	22.44	20.31	19.02	16.90
	HCPIMSC	32.68	32.41	31.57	30.21	27.92	26.73	36.78	36.30	34.70	31.54	30.20	25.78
	LSGMC*	33.25	—	—	—	—	—	38.88	—	—	—	—	—
	CMLC	**33.33**	**33.24**	**32.90**	**33.10**	**31.38**	**30.76**	**40.26**	**38.58**	**38.55**	**38.18**	**37.27**	**35.04**
Uci_digits	BSV	45.62	33.12	28.25	30.41	29.37	34.23	52.52	33.95	29.96	25.26	26.81	33.04
	DAIMC	83.92	84.55	81.34	80.54	76.79	69.06	76.09	75.64	72.83	69.10	65.48	58.36
	MIC	10.05	10.05	10.05	10.05	10.05	10.05	0.45	0.45	0.45	0.45	0.45	0.45
	MIC#	65.78	61.92	57.97	49.63	33.23	28.98	63.59	58.57	53.18	44.39	28.56	23.93
	PIC	83.65	83.65	84.10	82.98	81.05	82.79	86.38	86.02	86.04	84.38	80.11	82.25
	TCIMC	76.05	62.35	69.20	59.75	63.20	43.20	77.54	67.23	69.79	60.83	57.97	39.73
	HCLS_CGL	83.70	83.85	84.75	83.55	83.65	83.35	85.70	86.46	87.12	84.10	84.36	84.84
	TTGL	93.65	90.75	82.55	68.15	67.25	61.00	87.62	82.77	72.35	63.69	56.39	50.52
	HCPIMSC	89.80	91.26	91.99	92.54	89.72	87.69	82.25	83.88	84.53	85.07	81.67	78.10
	LSGMC*	96.00	—	—	—	—	—	92.59	—	—	—	-	-
	CMLC	**96.40**	**95.75**	**96.40**	**96.00**	**96.45**	**95.95**	**93.10**	**91.71**	**92.38**	**91.60**	**92.13**	**91.37**

注: ‘*’表示该算法为完备多视图聚类方法, ‘#’表示使用该算法在去除负视图数据情况下的聚类结果.

据的部分潜在信息,从而取得了优异的效果.

4) 对于完备数据集,CMLC在所有数据集上均优于完备多视图聚类方法LSGMC和其他方法,因此CMLC在处理完备数据集方面也具有显著优势.

5) TCIMC、HCPIMSC和CMLC方法取得了较好的结果,这是由于TCIMC和LSGMC引入SP张量,HCPIMSC引入张量分解项,利用张量直接挖掘了多视图数据中视图间的高阶相似信息,同时可以更有效

地挖掘不同视图之间的互补信息,但由于其忽略了奇异值之间大小的差距,可能出现过度惩罚大奇异值的问题,从而影响聚类性能. TTGL 也引入低秩张量结构,但它利用张量因式分解逼近张量的低秩,容易受到局部极小值的干扰,所以得到的结果较差. 而本文方法引入张量对数行列式,对多级自表示堆叠旋转进行低秩约束,从而捕获视图间的高阶信息,并取得了最好的聚类结果.

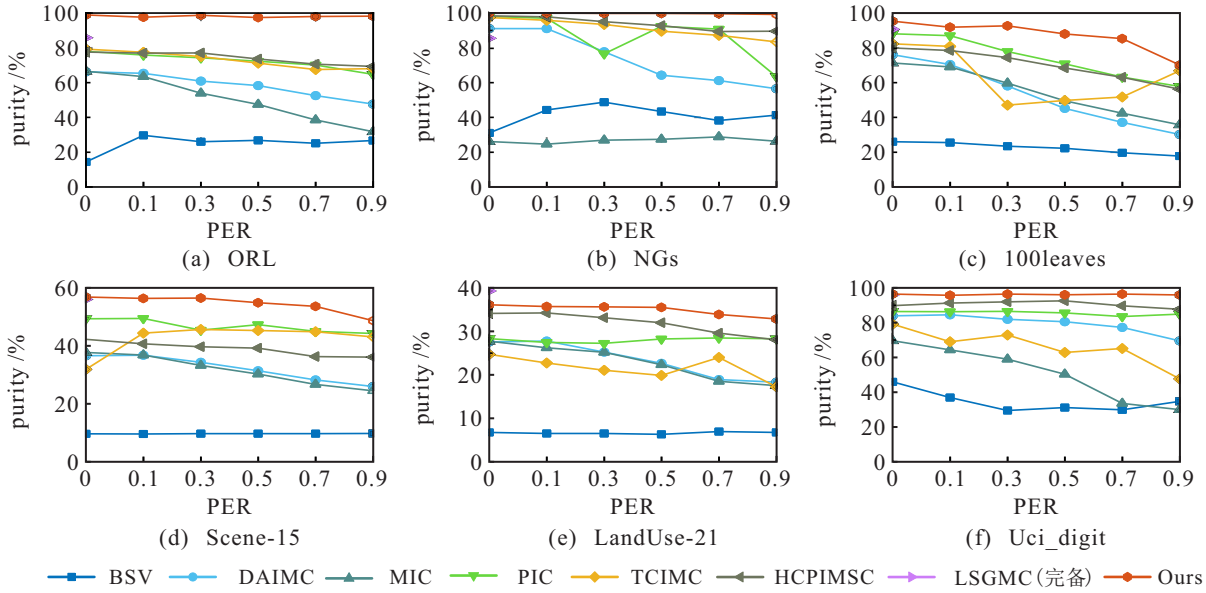


图 1 不同方法在不同数据集上的purity

4.3 噪声测试

噪声和离群点能够直接影响算法的聚类性能. 本节为了测试CMLC方法对噪声的鲁棒性,随机生成了一个二维多视图数据集NoiseData,并按照不同噪声比例随机生成噪声向量,最后将不同比例的噪声向量乘在NoiseData的每个视图上生成含有噪声的多视图数据集. 表1为具体参数,为了更直观地表达噪声数据,本节将含有不同比例噪声的NoiseData数据集的首个视图可视化,如图2所示.

表3展示了CMLC与对比方法在3个含有不同比例噪声的NoiseData数据集上的聚类结果.

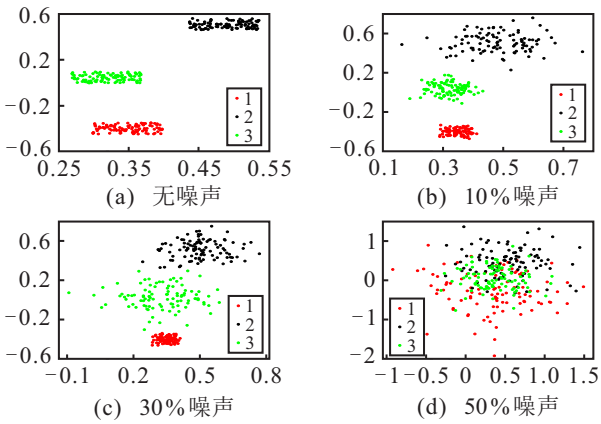


图 2 不同比例噪声数据集首个视图的可视化

表3 不同方法在不同比例噪声数据集上的聚类结果 100%

dataset	NoiseData								
	accuracy			NMI			purity		
	10%	30%	50%	10%	30%	50%	10%	30%	50%
BSV	45.00	65.67	41.33	20.92	32.84	3.29	52.00	65.67	42.00
DAIMC	99.00	78.00	60.33	95.88	49.62	23.40	99.00	78.00	60.33
MIC	33.67	33.67	33.67	0.67	0.67	0.67	34.00	34.00	34.00
PIC	53.13	47.33	39.07	20.8	13.53	3.65	55.4	49.33	40.8
TCIMC	**100.00**	96.33	58.33	**100.00**	86.39	21.99	**100.00**	96.33	59.00
HCLS_CGL	99.66	97.66	49.00	98.29	89.37	9.12	99.66	97.66	50.00
TTGL	98.67	92.67	54.67	94.86	78.88	11.13	98.67	92.67	54.67
HCPIMSC	**100.00**	97.67	60.67	**100.00**	91.34	25.95	**100.00**	97.67	60.67
LSGMC*	99.33	98.00	64.67	96.59	90.80	26.95	99.33	98.00	64.67
CMLC	**100.00**	**99.33**	**69.67**	**100.00**	**97.02**	**28.72**	**100.00**	**99.33**	**69.67**

从图2可以看出,当噪声比例为10%时,3个簇结构清晰,簇中的点很集中,簇间界限明显. 因此,TCIMC、HCPIMSC和CMLC方法均可达到100%的精度值. 当噪声比例为30%时,黑色和绿色的两个簇开始交接,各簇间的界限开始模糊,CMLC方法达到了最高的99.33%的精度值. 而噪声比例为50%时,

从图2(d)可以看出,3个簇大部分重叠在一起,此时,3个簇不易区分,但CMLC方法仍达到最好的聚类性能. 这是因为,CMLC对多级图表示的每一级均施加了噪声约束,使得模型对噪声和离群点具有比对比方法更好的鲁棒性.

4.4　消融实验

本文为进一步探索CMLC中包含的公共潜在表示、张量对数行列式约束项和距离正则项对聚类结果的影响,将CMLC方法、仅移除距离正则项后的CMLC方法、仅移除张量对数行列式约束项后的CMLC方法以及同时移除距离正则项和张量对数行列式约束项后的CMLC方法在PER = 0.5的6个不完备多视图数据上进行消融实验,如图3所示. 可以看出,移除上述3个约束项后精度均有明显下降,因此这些约束项为提高聚类性能做出了显著贡献.

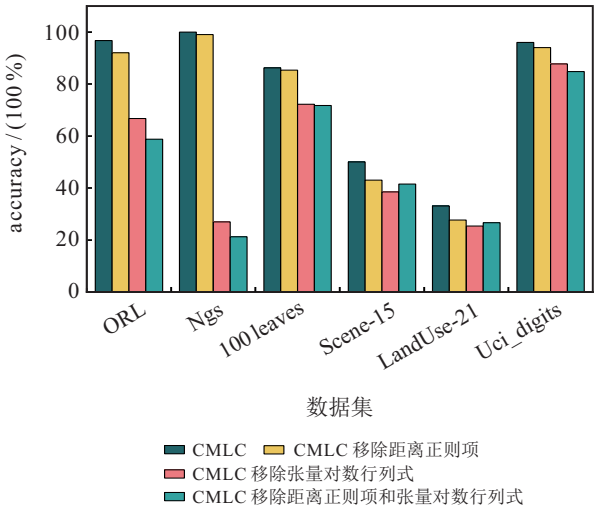
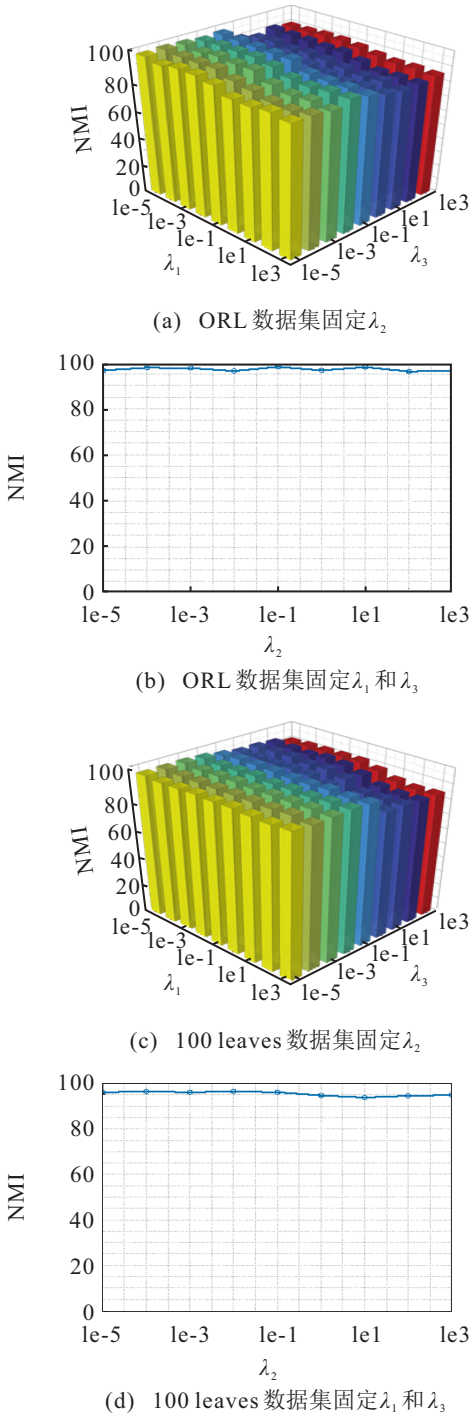


图 3　该方法在6种不完备多视图数据集上的消融实验

4.5　参数敏感性

本节在PER = 0.5下的ORL数据集和PER = 0.3下的100 leaves数据集上测试CMLC方法约束多级误差表示的3个惩罚项参数λ_1、λ_2和λ_3对聚类评价指标NMI的敏感性,如图4所示. 参数选择范围均为$\{10^{-5},10^{-4},10^{-3},10^{-2},10^{1},10^{2},10^{3},10^{4}\}$. 图4(a)和图4(c)展示了当$\lambda_2$固定时,CMLC在$\lambda_1$和$\lambda_3$取不同组合值时的NMI值变化曲线. 由图4(a)和图4(c)可以看出,对参数λ_1和λ_3取不同值时,NMI值都稳定在80%以上. 图4(b)和图4(d)为当λ_1和λ_3固定时,CMLC在λ_2取不同值时得到的NMI值变化曲线. 由图4(b)和图4(d)可以明显看出,NMI值均在95%左右,λ_2的取值对聚类结果的影响不大. 以上结果表明,CMLC对参数λ_1、λ_2和λ_3不敏感.



(a)　ORL数据集固定λ_2

(b)　ORL数据集固定λ_1和λ_3

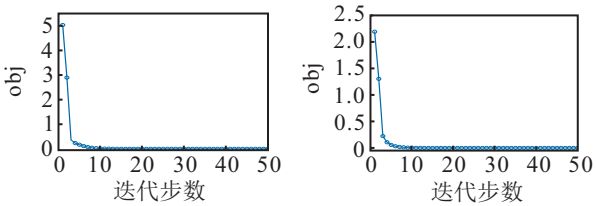
(c)　100 leaves数据集固定λ_2

(d)　100 leaves数据集固定λ_1和λ_3

图 4　不同参数关于NMI指标的变化分布

4.6　收敛性分析

本节通过实验验证CMLC模型的收敛性,令$\text{obj} = \sum_i \|R_i\|_\infty$,根据式(38)计算每次迭代的原始残差,当它们均小于设定值时,算法收敛. 图5为PER = 0.5的ORL数据集和PER = 0.3的100 leaves数据集上obj值的变化曲线. 由图5可以清晰地观察到,在两个数据集上obj值均随着迭代步数的增加先急剧下降后马上趋于平稳,表明CMLC方法能够很快达到收敛,具有强收敛性.



(a) ORL数据集 (b) 100 leaves数据集

图 5 obj值变化

5 结 论

针对现有算法无法充分挖掘不完备多视图数据中的结构信息等问题,本文提出了一种基于多级自表示约束的不完备多视图聚类方法. 该方法利用所有视图的公共潜在表示重构样本集,获取缺失数据中的潜在信息;引入张量对数行列式约束捕获视图间的高阶相似信息,引入局部流形结构约束项捕获各视图的局部几何结构;对多级图表示的每一级均施加了$l_{2,1}$噪声约束,增强模型的鲁棒性. 在7种不同的数据集上的实验结果表明,本文方法表现出了更好的性能和鲁棒性. 未来将在本文方法的基础上继续探究挖掘数据完整结构信息的方法.

参考文献(References)

[1] Cao X C, Zhang C Q, Fu H Z, et al. Diversity-induced multi-view subspace clustering[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015: 586-594.

[2] Xue Z, Du J P, Du D W, et al. Deep low-rank subspace ensemble for multi-view clustering[J]. Information Sciences, 2019, 482: 210-227.

[3] Zhang C Q, Hu Q H, Fu H Z, et al. Latent multi-view subspace clustering[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 4279-4287.

[4] Zhang G Y, Huang D, Wang C D. Facilitated low-rank multi-view subspace clustering[J]. Knowledge-Based Systems, 2023, 260: 110141.

[5] Zhang C Q, Fu H Z, Liu S, et al. Low-rank tensor constrained multiview subspace clustering[C]. 2015 IEEE International Conference on Computer Vision. Santiago, 2015: 1582-1590.

[6] Wen J, Yan K, Zhang Z, et al. Low-rank tensor graph learning based incomplete multi-view clustering[J]. Acta Automatica Sinica, 2023, 49(7): 1433-1445.

[7] Wu J L, Lin Z C, Zha H B. Essential tensor learning for multi-view spectral clustering[J]. IEEE Transactions on Image Processing, 2019, 28(12): 5910-5922.

[8] Li A, Chen J J, Yu X Y, et al. Robust multiview graph learning with applications to clustering for incomplete data[J]. Control and Decision, 2022, 37(12): 3251-3258.

[9] 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.

[10] Shao W X, He L F, Yu P S. Multiple incomplete views clustering via weighted nonnegative matrix factorization with $L_{2,1}$ regularization[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham, 2015: 318-334.

[11] Gao H, Peng Y, Jian S. Incomplete multi-view clustering[C]. International Conference of Intelligent Information Processing. Melbourne, 2016: 245-255.

[12] Wang H, Zong L L, Liu B, et al. Spectral perturbation meets incomplete multi-view data[J/OL]. 2019, arXiv: 1906. 00098.

[13] Hu M L, Chen S C. Doubly aligned incomplete multi-view clustering[J/OL]. 2019, arXiv: 1903. 02785.

[14] Liu J L, Teng S H, Fei L K, et al. A novel consensus learning approach to incomplete multi-view clustering[J]. Pattern Recognition, 2021, 115: 107890.

[15] Li X L, Chen M S, Wang C D, et al. Refining graph structure for incomplete multi-view clustering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(2): 2300-2313.

[16] Zhao H, Liu H, Fu Y . Incomplete multi-modal visual data grouping[C]. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16). New York, 2016: 2392-2398.

[17] Xia W, Gao Q, Wang Q, et al. Tensor completion-based incomplete multiview clustering[J]. IEEE Transactions on Cybernetics, 2022: 13635-13644.

[18] 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.

[19] Zhang Z, He W J. Tensorized topological graph learning for generalized incomplete multi-view clustering[J]. Information Fusion, 2023, 100: 101914.

[20] 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.

[21] Wen J, Liu C L, Xu G H, et al. Highly confident local structure based consensus graph learning for incomplete multi-view clustering[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 15712-15721.

[22] Lan W, Yang T C, Chen Q F, et al. Multiview subspace clustering via low-rank symmetric affinity graph[J]. IEEE Transactions on Neural Networks and Learning Systems. DOI: 10.1109/TNNLS.2023.3260258.

作者简介

陈梅(1973−), 女, 教授, 博士生导师, 主要研究方向为数据挖掘、机器学习, E-mail: chenmeilz@mail.lzjtu.cn;

马学艳(1995−), 女, 硕士生, 主要研究方向为数据挖掘、图聚类, E-mail: 1563344797@qq.com;

钱罗雄(1998−), 男, 硕士生, 主要研究方向为复杂数据挖掘、图表示聚类, E-mail:1309564426@qq.com;

张锦宏(1998−), 男, 硕士生, 主要研究方向为复杂数据挖掘、模式识别, E-mail: 1336478235@qq.com;

张弛(1999−), 男, 硕士生, 主要研究方向为复杂数据挖掘、时间序列聚类, E-mail: 904048197@qq.com.