doi: 10.13413/j.cnki.jdxblxb.2023071

基于自表示和投影映射的不完整多视图聚类

赵翠娜, 杨有龙
(西安电子科技大学 数学与统计学院, 西安 710126)

摘要: 针对不完整多视图聚类存在的缺陷, 提出一种融合自表示和投影映射的统一框架. 首先, 利用自表示和样本存在指示矩阵学习一致相似图, 它反映了样本间的公共相似关系; 其次, 利用投影映射将样本矩阵投影到超球面上, 得到公共低维表示; 最后, 将两者通过谱表示嵌入在一起, 解决了因多视图数据缺失引起的不完整多视图聚类问题. 该算法在真实数据集上的实验结果优于其他算法, 证明了算法的有效性.

关键词: 多视图聚类; 不完整视图; 自表示学习; 投影映射
中图分类号: TP391 **文献标志码**: A **文章编号**: 1671-5489(2024)02-0331-08

Incomplete Multi-view Clustering Based on Self-representation and Projection Mapping

ZHAO Cuina, YANG Youlong
(*School of Mathematics and Statistics, Xidian University, Xi'an* 710126, *China*)

Abstract: Aiming at the shortcomings of incomplete multi-view clustering, we proposed a unified framework that integrated self-representation and projection mapping. Firstly, self-representation and sample presence indication matrices were used to learn a uniform similarity graph, which reflected the common similarity relationship between samples. Secondly, the sample matrices were projected onto the hypersphere by using projection mapping to obtain a common low-dimensional representation. Finally, the two were embedded together through spectral representation to solve the incomplete multi-view clustering problem caused by missing multi-view data. The experimental results of this algorithm on real datasets are better than other algorithms, which proves the effectiveness of the proposed algorithm.

Keywords: multi-view clustering; incomplete view; self-representation learning; projection mapping

0 引 言

由于不同的采集源或对同一物体的不同处理方法, 数据可被分成多个特征集, 这些特征集有不同的物理意义和鉴别能力, 因此可将这些不同的特征集视为多视图数据. 随着多视图数据的大量涌现, 多视图聚类(multi-view clustering, MVC)[1-7]考虑了来自多个视图的互补信息和一致信息, 因此在数

收稿日期: 2023-03-01.

第一作者简介: 赵翠娜(1997—), 女, 汉族, 硕士研究生, 从事多视图聚类的研究, E-mail: 20071212564@stu.xidian.edu.cn.
通信作者简介: 杨有龙(1967—), 男, 汉族, 博士, 教授, 从事数据分类、数据聚类融合以及数据预测的研究, E-mail: ylyang@mail.xidian.edu.cn.

基金项目: 国家自然科学基金(批准号: 61573266)和陕西省自然科学基础研究计划项目(批准号: 2021JM-133).

据挖掘和分析领域已引起广泛关注.

但由于一些现实因素的影响,多视图数据经常是不完整的. 不完整性可能由于随机缺少特征或缺少样本的某个视图特征导致. 例如,在图像聚类时,获取到的图像可能由于遭到遮挡受到破坏,这属于前者;在多语言文档聚类中,一种语言可视为文档的一个视图,一个文档可能没有所有的语言版本,这属于后者. 本文主要关注后者,即一个样本可以在所有或部分视图上找到. 在不完整多视图数据上进行的聚类问题称为不完整多视图聚类(incomplete multi-view clustering, IMC).

近年来,已有许多方法处理不完整多视图聚类. 不完整多视图聚类与多视图聚类相同,可分为基于核的不完整多视图聚类、基于矩阵分解的不完整多视图聚类和基于图的不完整多视图聚类,但现有的很多方法已将上述种类融合在同一个框架中,进一步提高学习性能. 从视图缺失处理角度,不完整多视图聚类可分为基于部分视图的方法和基于缺失视图推断的方法.

基于缺失视图补充的方法对缺失视图或缺失视图转化部分进行填充,占据了不完整多视图聚类很大一部分. 例如,Shao 等[8]提出了一种集体核学习的多视图补全算法,该算法对于两个缺失的视图,利用共享实例的对齐信息,去优化不完全核矩阵的信息,最后使用核典型相关分析完成聚类;Wen 等[9]提出了缺失视图推断的统一嵌入对齐不完整多视图聚类,通过将位置重建项引入到矩阵分解和反向图学习中,使所有视图都能自然对齐;Zhang 等[10]提出了基于自适应缺失视图归算和合作学习的端到端不完全多视图模糊聚类,将缺失视图植入、隐藏视图学习和聚类整合到同一个框架中,实现了隐藏视图和可见视图之间的合作学习;Wen 等[11]提出了基于不完全多视图聚类的自适应图补全,通过借用其他视图的相似性信息恢复不完整视图信息,并且最大程度上保留了自身的相似性信息;Liu 等[12]提出了具有不完整核的多核 k-均值算法,并不直接填充缺失部分,而将缺失部分作为多核聚类过程中需要优化的辅助变量,通过填充和聚类的交替迭代达到一个较好的聚类结果;Gao 等[13]提出了不完全多视图聚类,基于谱理论,将原始数据投影到新的空间中,并将各视图的投影进行整合更新每个视图中的缺失部分,进而趋向共识.

基于部分视图的方法通常将视图信息分为对齐部分和非对齐部分学习视图的潜在表示去聚类. 例如,Li 等[14]提出了部分多视图聚类,利用矩阵分解在两个视图的共有部分学习相同的隐藏表示,在局部视图学习各自的潜在表示去实现聚类;Shao 等[15]提出了基于 $L_{2,1}$ 正则化加权非负矩阵分解的多重不完全视图聚类,通过引入一个值位于 0～1 间的权重矩阵在非负矩阵分解中学习共识矩阵,使缺失实例的影响低于存在实例,并加入了 $L_{2,1}$ 正则化使模型对噪声和异常值具有鲁棒性;Zhuge 等[16]提出了基于联合表示和聚类(joint representation learning and clustering, JRLC)框架的不完全多视图聚类,提出了带谱嵌入的 JRLC(JRLC-SE)及通过集成非负嵌入和谱嵌入的 JRLC(JRL-NS)两种方法;Wang 等[17]提出了谱扰动满足不完整多视图数据,先将样本缺失转化为样本间相似性缺失问题,然后基于谱扰动理论,使用每个视图的 Laplace 矩阵进行一致性学习,最后使用谱聚类在一致 Laplace 矩阵上进行聚类.

虽然现有的方法在不完整多视图聚类方面都取得了较好的效果,但仍存在一些缺点:

1) 虽然在聚类过程中完成了缺失数据的填充,但不准确的填补可能会使聚类结果和真实结果产生较大误差;

2) 利用各视图上存在的样本获取的共识表示并未充分探索视图间的互补信息,仅学习到存在样本的一致信息;

3) 仅探索低维表示的信息,原始样本矩阵中的公共结构信息被忽略.

基于以上限制,本文提出一种基于自表示和投影映射的不完整多视图聚类(IMSP). 首先,在每个视图上引入样本存在指示矩阵学习每个视图上的自表示矩阵,既使缺失视图可以对齐,又充分探索了样本点之间的公共相似度关系;其次,在减少缺失样本权重影响的情况下,将不同视图的样本矩阵投影在一个超球面上获取共识低维表示;最后,将上述两部分通过谱表示嵌入在同一个框架中相互交替迭代和更新,使样本间的公共表示和视图矩阵的局部表示更具有一致性. 本文使用 IMSP-P 和 IMSP-S 两种算法获得聚类结果,进一步提高聚类结果的质量.

本文用 $\boldsymbol{X}^{(v)} \in \mathbb{R}^{d_v \times n_v}$ 表示第 v 个视图的样本矩阵，$\boldsymbol{Y}^{(v)} \in \mathbb{R}^{d_v \times n}$ 表示第 v 个视图的缺失部分用 0 填充的样本矩阵，$\boldsymbol{G}^{(v)} \in \mathbb{R}^{n_v \times n}$ 表示第 v 个视图的样本存在指示矩阵，$\boldsymbol{Z}^{(v)} \in \mathbb{R}^{n \times n}$ 表示第 v 个视图的自表示矩阵，$\boldsymbol{S} \in \mathbb{R}^{n \times n}$ 表示一致相似图矩阵，$\boldsymbol{E}^{(v)} \in \mathbb{R}^{n \times n}$ 表示样本权重矩阵，$\boldsymbol{P} \in \mathbb{R}^{c \times n}$ 为共识低维表示，$\boldsymbol{F}^{(v)} \in \mathbb{R}^{c \times d_v}$ 表示第 v 个视图的投影矩阵，d_v 表示第 v 个视图的特征维度，n_v 表示第 v 个视图存在的样本数量，n 为总样本数量，c 为类的数量.

1　相关工作和模型

1.1　投影映射

文献[18]通过学习一组优秀的同构特征捕获不同视图之间的语义互补和相同分布. 该方法将多个异质特征空间线性映射到特征同构子空间. 同时，来自不同视图的相关表示被耦合在一起，以捕捉不同视图之间的语义互补和相同分布. 该方法的目标函数为

$$\| \boldsymbol{A}\boldsymbol{X}^{(1)} - \boldsymbol{B}\boldsymbol{X}^{(2)} \|_{\mathrm{F}}^{2}, \tag{1}$$

其中：$\boldsymbol{X}^{(1)} = (\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \cdots, \boldsymbol{x}_{n_1}^{(1)}) \in \mathbb{R}^{d_1 \times n_1}$，$\boldsymbol{X}^{(2)} = (\boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \cdots, \boldsymbol{x}_{n_1}^{(2)}) \in \mathbb{R}^{d_2 \times n_1}$，$d_1$ 和 d_2 分别为 $\boldsymbol{X}^{(1)}$ 视图和 $\boldsymbol{X}^{(2)}$ 视图的特征维度，n_1 为两个视图共同存在的样本数量；$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 分别为两个视图上的投影矩阵. 将该方法推广到不完整多视图聚类中面临两个困难：1）成对匹配循环过多，且不好优化；2）成对视图之间的共同可用样本数量不同. 因此，考虑将不同视图的特征映射到同一个共识低维表示，缺失样本用 0 填补，并且引入样本权重赋给缺失样本较低的权重. 目标函数表示为

$$\min_{\boldsymbol{P}} \| (\boldsymbol{P} - \boldsymbol{F}^{(v)}\boldsymbol{Y}^{(v)})\boldsymbol{E}^{(v)} \|_{\mathrm{F}}^{2}, \tag{2}$$

其中：$\boldsymbol{P}$ 是投影映射后的低维矩阵；$\boldsymbol{Y}^{(v)}$ 是填充后的样本矩阵；$\boldsymbol{F}^{(v)}$ 是第 v 个视图的映射矩阵；$\boldsymbol{E}^{(v)}$ 是样本权重矩阵，可表示为

$$\boldsymbol{E}_{ii}^{(v)} = \begin{cases} 1, & \text{第 } i \text{ 个样本在第 } v \text{ 个视图中存在}, \\ 0, & \text{第 } i \text{ 个样本在第 } v \text{ 个视图中缺失}. \end{cases} \tag{3}$$

1.2　自表示和一致相似图

自表示是非常有效的学习任意两个样本之间相似度的子空间聚类方法，其思想为每个样本是其余样本的线性组合，系数是两个样本之间相似度的表示. 多视图自表示的目标函数通常表示为

$$\min_{\boldsymbol{Z}^{(v)}} \{ R(\boldsymbol{X}^{(v)}, \boldsymbol{X}^{(v)}\boldsymbol{Z}^{(v)}) + \lambda\varphi(\boldsymbol{Z}^{(v)}) \}, \tag{4}$$

其中：$R(\boldsymbol{X}^{(v)}, \boldsymbol{X}^{(v)}\boldsymbol{Z}^{(v)})$ 表示 $\boldsymbol{X}^{(v)}$ 与 $\boldsymbol{X}^{(v)}\boldsymbol{Z}^{(v)}$ 之间的重构误差，通常根据不同需求，选择不同的范数；$\varphi(\boldsymbol{Z}^{(v)})$ 为正则项；λ 为平衡系数. 不能直接将多视图自表示推广到缺失样本的不完整多视图中，因为缺失样本的存在使样本信息不能对齐. 文献[19]中引入了样本存在指示矩阵，多个视图之间的信息可以对齐后互相融合. 目标函数为

$$\min_{\boldsymbol{Z}^{(v)}} \| \boldsymbol{X}^{(v)} - \boldsymbol{X}^{(v)}\boldsymbol{G}(v)^{\mathrm{T}}\boldsymbol{Z}^{(v)}\boldsymbol{G}^{(v)} \|_{\mathrm{F}}^{2}. \tag{5}$$

通过自表示学习到的不同视图的 $\boldsymbol{Z}^{(v)}$ 应该反映相同的聚类结构，即无论在哪个视图中样本相似度关系应该一致. 一致相似图 $\boldsymbol{S}$ 最直接的融合应该是所有视图的自表示矩阵的质心，但由自表示得到的 $\boldsymbol{Z}^{(v)}$ 中的系数含有负数，并不能代表样本之间的相似性关系. 所以添加了一些限制条件，$\mathrm{diag}(\boldsymbol{S}) = \boldsymbol{0}$ 避免平凡解，$\boldsymbol{S}^{\mathrm{T}}\boldsymbol{1} = 1$ 可以给出更好的概率解释.

1.3　低维共识矩阵与一致相似图的嵌入

低维共识矩阵已成为样本矩阵的新表示，因此它的几何结构应符合原始样本之间的相似度关系，即样本之间距离大的相似度较小. 目标函数为

$$\min_{\boldsymbol{P},\boldsymbol{S}} \| \boldsymbol{P}_{:,i} - \boldsymbol{P}_{:,j} \|_2^2 \boldsymbol{S}_{ij}. \tag{6}$$

上述目标的等效表述可转化为以下形式：

$$\min_{\boldsymbol{P},\boldsymbol{S}} \mathrm{tr}(\boldsymbol{P}\boldsymbol{L}_{\boldsymbol{S}}\boldsymbol{P}^{\mathrm{T}}), \tag{7}$$

其中：$\boldsymbol{L}_{\boldsymbol{S}} = \boldsymbol{D} - \boldsymbol{S}$，$\boldsymbol{D}$ 为一个元素为 $\boldsymbol{S}$ 的行和或列和的对角矩阵；$\mathrm{tr}(\cdot)$ 为矩阵的迹. 当 $\boldsymbol{S}$ 固定时，上述目标函数与谱聚类的目标函数相同. 同理，对 $\boldsymbol{P}$ 施加一个正交约束 $\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}} = \boldsymbol{I}$，同时在式(2)中施加此约

束，P 不会简单地成为各视图的样本矩阵投影后的质心.

1.4 总目标函数

本文将自表示学习、图融合、超球面上的投影映射融合到一个框架中. 分别在一致相似图 $\boldsymbol{S}$ 上进行谱聚类(IMSP-S)和超球面低维矩阵 $\boldsymbol{P}$ 进行 K-means 聚类(IMSP-P)得到聚类结果，目标函数为

$$\begin{aligned}\min_{\boldsymbol{P},\boldsymbol{S},\boldsymbol{Z}^{(v)}}\{&\|\boldsymbol{X}^{(v)}-\boldsymbol{X}^{(v)}\boldsymbol{G}^{(v)\mathrm{T}}\boldsymbol{Z}^{(v)}\boldsymbol{G}^{(v)}\|_{\mathrm{F}}^2+\lambda_1\|\boldsymbol{S}-\boldsymbol{Z}^{(v)}\|_{\mathrm{F}}^2+\\&\lambda_2\|(\boldsymbol{P}-\boldsymbol{F}^{(v)}\boldsymbol{Y}^{(v)})\boldsymbol{E}^{(v)}\|_{\mathrm{F}}^2+\lambda_3\mathrm{tr}(\boldsymbol{P}\boldsymbol{L}_{\boldsymbol{S}}\boldsymbol{P}^{\mathrm{T}})\},\\\text{s. t. }&\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}=\boldsymbol{I},\quad\boldsymbol{S}^{\mathrm{T}}\boldsymbol{1}=1,\quad\mathrm{diag}(\boldsymbol{S})=\boldsymbol{0},\quad\boldsymbol{S}\geqslant\boldsymbol{0}.\end{aligned}\tag{8}$$

2 优化方法

1) 更新 $\boldsymbol{P}$ 和 $\boldsymbol{F}^{(v)}$. 当其他变量固定后，关于 $\boldsymbol{P}$ 的子问题变为

$$\begin{aligned}&\min_{\boldsymbol{P}}\{\lambda_2\|(\boldsymbol{P}-\boldsymbol{F}^{(v)}\boldsymbol{Y}^{(v)})\boldsymbol{E}^{(v)}\|_{\mathrm{F}}^2+\lambda_3\mathrm{tr}(\boldsymbol{P}\boldsymbol{L}_{\boldsymbol{S}}\boldsymbol{P}^{\mathrm{T}})\},\\&\text{s. t. }\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}=\boldsymbol{I}.\end{aligned}\tag{9}$$

经过变换后，问题(9)变为以下形式：

$$\begin{aligned}&\min_{\boldsymbol{P}}\{\lambda_2\mathrm{tr}\Big(\sum_{v=1}^{V}\boldsymbol{P}\boldsymbol{E}^{(v)}\boldsymbol{E}^{(v)\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}-2\boldsymbol{P}\boldsymbol{E}^{(v)}\boldsymbol{Y}^{(v)\mathrm{T}}\boldsymbol{F}^{(v)\mathrm{T}}\Big)+\lambda_3\mathrm{tr}(\boldsymbol{P}\boldsymbol{L}_{s}\boldsymbol{P}^{\mathrm{T}})\},\\&\text{s. t. }\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}=\boldsymbol{I}.\end{aligned}\tag{10}$$

这是一个经典的广义典型相关分析[20]问题，其中

$$\boldsymbol{F}^{(v)\mathrm{T}}=(\boldsymbol{E}^{(v)}\boldsymbol{Y}^{(v)\mathrm{T}})^{\dagger}(\boldsymbol{E}^{(v)}\boldsymbol{Y}^{(v)\mathrm{T}})^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}},$$

$\boldsymbol{A}^{\dagger}$ 为 $\boldsymbol{A}$ 的广义逆. 所以式(10)转换为

$$\begin{aligned}&\min_{\boldsymbol{P}}\mathrm{tr}\Big(\boldsymbol{P}\Big(\lambda_2\sum_{v=1}^{V}(\boldsymbol{E}^{(v)}\boldsymbol{E}^{(v)\mathrm{T}}-2\boldsymbol{E}^{(v)}\boldsymbol{Y}^{(v)\mathrm{T}}\boldsymbol{F}^{(v)\mathrm{T}})+\lambda_3\boldsymbol{L}_{\boldsymbol{S}}\Big)\boldsymbol{P}^{\mathrm{T}}\Big),\\&\text{s. t. }\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}=\boldsymbol{I},\end{aligned}\tag{11}$$

$\boldsymbol{P}$ 的求解为 $\boldsymbol{A}=\lambda_2\sum_{v}^{V}(\boldsymbol{E}^{(v)}\boldsymbol{E}^{(v)\mathrm{T}}-2\boldsymbol{E}^{(v)}\boldsymbol{Y}^{(v)\mathrm{T}}\boldsymbol{F}^{(v)\mathrm{T}})+\lambda_3\boldsymbol{L}_{\boldsymbol{S}}$ 的特征值分解问题. $\boldsymbol{P}^{\mathrm{T}}$ 对应 $\boldsymbol{A}$ 的前 c 个最小特征值对应的特征向量.

2) 更新 $\boldsymbol{Z}^{(v)}$. 当与 $\boldsymbol{Z}^{(v)}$ 相关的其他变量固定时，优化问题转换为

$$\min_{\boldsymbol{Z}^{(v)}}\|\boldsymbol{X}^{(v)}-\boldsymbol{X}^{(v)}\boldsymbol{G}^{(v)\mathrm{T}}\boldsymbol{Z}^{(v)}\boldsymbol{G}^{(v)}\|_{\mathrm{F}}^2+\lambda_1\|\boldsymbol{S}-\boldsymbol{Z}^{(v)}\|_{\mathrm{F}}^2.\tag{12}$$

通过对 $\boldsymbol{Z}^{(v)}$ 取导为 0，得

$$\boldsymbol{G}^{(v)}\boldsymbol{X}^{(v)\mathrm{T}}\boldsymbol{X}^{(v)}\boldsymbol{G}^{(v)\mathrm{T}}\boldsymbol{Z}^{(v)}+\lambda_1\boldsymbol{Z}^{(v)}(\boldsymbol{G}^{(v)}\boldsymbol{G}^{(v)\mathrm{T}})=(\lambda_1\boldsymbol{S}+\boldsymbol{G}^{(v)}\boldsymbol{X}^{(v)\mathrm{T}}\boldsymbol{X}^{(v)}\boldsymbol{G}^{(v)\mathrm{T}})(\boldsymbol{G}^{(v)}\boldsymbol{G}^{(v)\mathrm{T}})^{-1}.\tag{13}$$

式(13)是经典的 Sylvester 方程，本文使用 Bartels·Stewart 算法[21]解决式(13)的优化.

3) 更新 $\boldsymbol{S}$. 当其他变量不变时，优化 $\boldsymbol{S}$ 可以解决以下问题：

$$\begin{aligned}&\min_{\boldsymbol{S}}\{\lambda_1\|\boldsymbol{S}-\boldsymbol{Z}^{(v)}\|_{\mathrm{F}}^2+\lambda_3\mathrm{tr}(\boldsymbol{P}\boldsymbol{L}_{\boldsymbol{S}}\boldsymbol{P}^{\mathrm{T}})\},\\&\text{s. t. }\boldsymbol{S}^{\mathrm{T}}\boldsymbol{1}=1,\quad\mathrm{diag}(\boldsymbol{S})=\boldsymbol{0},\quad\boldsymbol{S}\geqslant\boldsymbol{0},\end{aligned}\tag{14}$$

其中 $\lambda_3\mathrm{tr}(\boldsymbol{P}\boldsymbol{L}_{\boldsymbol{S}}\boldsymbol{P}^{\mathrm{T}})=\frac{\lambda_3}{2}\mathrm{tr}(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{S})$，$\boldsymbol{M}_{ij}=\|\boldsymbol{P}_{:,i}-\boldsymbol{P}_{:,j}\|_2^2$. 因此优化问题转换为

$$\begin{aligned}&\min_{\boldsymbol{S}}\boldsymbol{V}\left\|\boldsymbol{S}-\left(\sum_{v=1}^{V}\frac{\boldsymbol{Z}^{(v)}}{V}-\frac{\lambda_3}{4\lambda_1 V}\boldsymbol{M}\right)\right\|_{\mathrm{F}}^2,\\&\text{s. t. }\boldsymbol{S}^{\mathrm{T}}\boldsymbol{1}=1,\quad\mathrm{diag}(\boldsymbol{S})=\boldsymbol{0},\quad\boldsymbol{S}\geqslant\boldsymbol{0}.\end{aligned}\tag{15}$$

上述问题的优化问题使用文献[22]中的算法 1 优化.

3 实 验

3.1 数据集

1) BBCSport：该数据集包括来自 BBC 新闻网站的 116 篇关于 5 个主题的新闻文章，每个文档被

划分为 4 个文本，对应于特征维度为 1 991,2 063,2 113 和 2 158 的 4 个视图.

2) ORL：是一个从英国剑桥实验室收集的人脸数据集，40 个目录中有 400 张图像，分别提取对应于 3 个视图的 3 个特征集，3 个特征集的尺寸分别为 4 096,3 304 和 6 750.

3) Wikipedia：该数据集用于检索，包含 693 个样本，其中 128 维图像特征和 10 维文本特征被分为 10 个类别.

4) BUAA：该数据集共有 1 350 张图像，分为 150 个类别，每个类别包含视觉光谱和近红外数据，两个视图的尺寸均为 100.

5) Caltech-7：该数据集是 Caltech-101 的一个子集，包含 7 个类，其中包含 1 474 张图像，6 个视图分别为 48 维 Gabor 特征、40 维 WM 特征、254 维 CENTRIST 特征、1 984 维 HOG 特征、512 维 GIST 特征和 928 维 LBP 特征.

3.2　对比方法

1) 缺失视图推断的统一嵌入对齐的不完整多视图聚类(UEAF)[9]：首先，利用误差矩阵和索引矩阵对缺失数据进行填充，得到低维一致表示；其次，通过反向图正则化得到一致相似度图矩阵；最后，利用谱聚类和 k 均值得到最终聚类结果.

2) 基于不完全多视图聚类的自适应图补全(AGC-IMC)[11]：该方法采用视图内保持和视图间推断的方法恢复完整的相似图矩阵，先用自适应权重表示每个视图的重要性，然后利用恢复后的完整图进行一致表示学习.

3) 基于 $L_{2,1}$ 正则化加权非负矩阵分解的多重不完全视图聚类(GPMVC)[23]：在非负矩阵分解中引入加权指标矩阵，减小缺失样本相对于已有样本的影响，并利用协正则化和 $L_{2,1}$ 范数得到一致的低维表示矩阵.

4) 基于图正则化非负矩阵分解的局部多视图聚类(MIC)[15]：该方法在部分多视图聚类(PVC)的基础上，将两视图的情况扩展为多视图，并增加了图正则化惩罚项，提高了聚类的质量.

5) 不完整视图的在线多视图聚类(OMVC)[24]：针对大规模多视图数据的处理问题，提出一种逐块处理多视图数据的方法. 采用加权非负矩阵分解和 L_1 范数学习一致性特征矩阵.

3.3　评估指标

本文选择聚类精度(ACC)、归一化互信息(NMI)和纯度(Purity)这 3 个外部指标作为评价聚类性能的指标. 这 3 个指标值均位于 0～1 间，其值越大聚类效果越好.

3.4　实验结果与分析

在实验中，确保每个样本至少在一个视图上出现的情况下，在每个视图上随机移除 10%,30%和 50%的样本. 为进行公平有效的比较，使用上述方法在每个缺失率下随机生成 10 组缺失样本，并将这 10 次聚类指标的平均值作为最终聚类结果. 不同算法在真实数据集上的聚类性能比较列于表 1. 由表 1 可见，在大部分实验中，本文算法都优于其他算法，尤其是在数据集 BUAA 上，本文算法显著优于次优算法. 因为本文算法在减少缺失样本的影响下将多个视图的公共相似度关系和超球面低维表示很好地集成在一个框架中，且互相促进彼此的迭代和更新. 在大多数情况下，UEAF 和AGC-IMC算法的效果显著优于 MIC,OMVC 和 GPMVC 算法，这是因为 UEAF 算法中利用图正则化学习到的误差矩阵去填补缺失样本，误差矩阵即选取 $n-n_v$ 条最能代表存在样本特征关系的 d_v 维向量；AGC-IMC 算法是利用其余视图上的相似性关系的线性组合填补缺失样本的相似性关系，在缺失率增加的情况下，甚至在某些数据集上效果反而更好；而 MIC,OMVC 和 GPMVC 算法只是简单均值填充，说明合理地填补缺失样本或相似性关系也可以达到较好的效果.

虽然在一些数据集上，IMSP 算法没有取得最优聚类指标，但与最优算法相差较小. 在数据集 BUAA 上，IMSP 比次优方法的聚类精度超出了 50%，证明了该算法的优越性.

表1 不同算法在真实数据集上的聚类性能比较

Table 1 Comparison of clustering performance of different algorithms on real datasets

数据集	算法	ACC			NMI			纯度		
		0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
BBCSport	UEAF	80.60	77.41	62.93	72.13	67.30	47.57	89.22	86.64	72.24
	AGC-IMC	78.10	79.48	52.84	74.09	67.74	33.20	89.83	85.95	62.41
	GPMVC	58.51	46.81	52.66	29.26	17.55	14.63	87.77	52.66	52.66
	MIC	48.19	46.21	40.78	23.39	22.11	16.32	52.07	49.57	43.79
	OMVC	49.86	47.77	48.06	40.70	32.78	39.53	57.22	52.45	54.87
	IMSP-S	79.31	77.93	68.62	73.89	69.94	52.11	89.22	85.95	76.47
ORL	UEAF	66.35	58.68	52.53	79.97	74.24	67.77	70.23	63.43	56.13
	AGC-IMC	76.38	64.63	44.05	85.88	77.81	65.52	77.93	67.63	46.90
	GPMVC	68.82	63.45	52.25	84.81	79.72	69.48	73.15	67.57	55.70
	MIC	58.35	49.25	35.75	71.94	57.44	46.44	62.00	51.50	37.00
	OMVC	20.52	30.06	28.87	47.86	59.61	57.81	24.87	32.19	31.18
	IMSP-P	78.43	73.38	62.13	88.59	85.20	76.59	80.38	76.55	65.15
	IMSP-S	81.88	78.78	67.08	90.76	89.31	80.84	85.00	82.28	70.27
Wikipedia	UEAF	54.50	45.12	37.72	50.14	39.31	27.18	58.50	48.80	40.65
	AGC-IMC	53.00	44.96	36.52	43.79	34.41	30.10	56.83	48.31	40.78
	GPMVC	44.37	38.25	34.78	30.70	24.43	21.04	47.13	41.89	37.72
	MIC	48.33	41.52	34.65	38.33	28.83	22.75	52.91	44.81	38.86
	OMVC	27.22	16.22	16.56	19.76	5.62	5.08	30.42	16.84	17.41
	IMSP-P	54.79	46.00	37.92	47.91	36.24	30.29	57.78	49.38	41.69
	IMSP-S	55.14	45.86	37.71	46.80	37.96	29.91	57.27	49.34	41.52
BUAA	UEAF	32.98	29.24	26.88	65.30	62.46	61.89	35.07	31.22	28.95
	AGC-IMC	39.77	39.79	28.04	70.04	69.89	64.31	41.33	41.54	29.62
	GPMVC	7.93	7.93	7.93	26.35	26.35	26.35	11.70	11.70	11.70
	MIC	1.93	1.93	1.93	13.07	13.07	13.07	11.70	11.70	11.70
	OMVC	27.44	27.92	27.85	68.40	68.28	68.28	29.55	30.20	30.16
	IMSP-P	74.64	62.23	41.04	87.40	81.60	72.06	75.50	64.09	43.99
	IMSP-S	75.01	64.18	41.33	87.45	82.18	70.46	75.93	65.86	45.36
Caltech-7	UEAF	46.21	51.23	56.95	43.66	42.18	32.66	85.39	79.37	77.25
	AGC-IMC	65.18	65.99	65.90	53.54	52.50	52.59	85.52	85.52	85.41
	GPMVC	38.41	41.42	38.79	37.18	36.00	29.76	79.74	78.70	75.98
	MIC	57.05	56.15	56.87	23.68	45.11	40.01	73.26	64.79	66.28
	OMVC	38.72	44.59	52.60	29.00	22.46	5.96	73.38	63.32	55.02
	IMSP-P	68.28	67.74	63.14	60.12	59.16	51.65	89.01	88.86	85.77
	IMSP-S	68.88	68.57	64.78	60.59	59.53	52.93	89.41	89.18	86.06

3.5 参数确定及敏感度分析

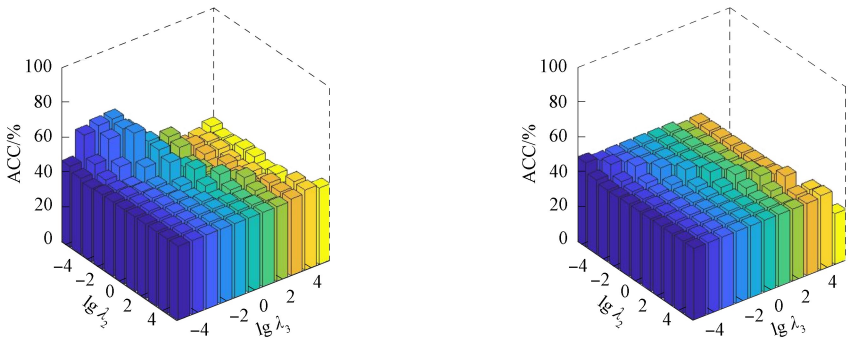
在IMSP算法实验中，3个超参数 $\lambda_1,\lambda_2,\lambda_3$ 需要通过网格搜索最优参数，其中 $\lambda_1,\lambda_2,\lambda_3$ 的取值范围均为 $\{10^{-5},10^{5}\}$. 比较算法的参数会根据相应的参数范围选择最优结果. 图1和图2分别是参数 λ_2 和 λ_3 与 λ_1 在数据集上的敏感性分析. 由图1和图2可见，参数在很大范围内都是相对稳定的.

3.6 算法复杂度分析与收敛性

根据迭代优化算法的更新步骤计算算法的时间复杂度，变量 $\boldsymbol{P},\boldsymbol{F}^{(v)}$ 和 $\boldsymbol{Z}^{(v)}$ 是时间代价最大的，为 $O(n^3)$，计算 $\boldsymbol{S}$ 的时间可以忽略不计，因此总的时间复杂度为 $O(n^3)$.

图3是在数据集Caltech-7缺失率为0.3时的收敛曲线. 由图3可见，在5次迭代以内目标函数快速下降，有效支持了算法的有效性.

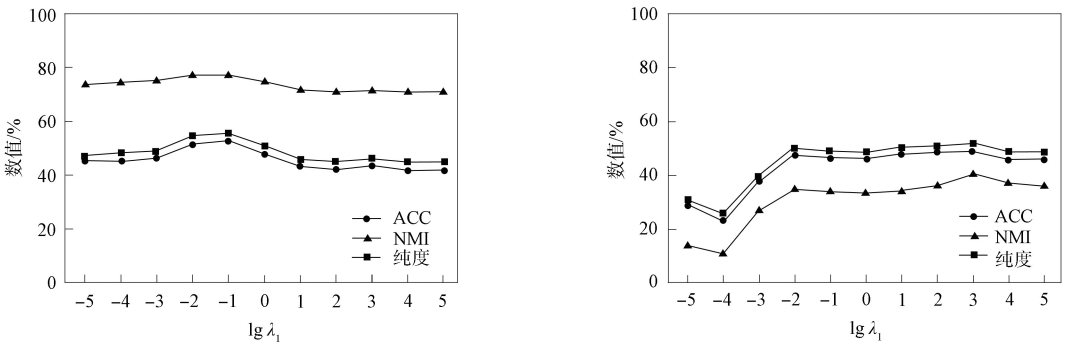
综上所述，针对不完整多视图聚类存在的缺陷，本文提出了一种将自表示学习和视图投影相结合的统一框架去处理不完整多视图聚类的方法. 首先，在自表示中引入了样本存在矩阵，使学习的自表



(A) 数据集BUAA, $\lambda_1=10^{-4}$　　(B) 数据集Wikipedia, $\lambda_1=10^{3}$

图 1　数据集上参数 λ_2 和 λ_3 的敏感性，每个视图的缺失率为 0.3

Fig. 1　Sensitivity of parameters λ_2 and λ_3 on datasets, where missing rate of each view is 0.3



(A) 数据集BUAA, $\lambda_2=\lambda_3=10^{-1}$　　(B) 数据集Wikipedia, $\lambda_2=\lambda_3=10^{-1}$

图 2　数据集上参数 λ_1 的敏感性，每个视图的缺失率为 0.3

Fig. 2　Sensitivity of parameter λ_1 on datasets, where missing rate of each view is 0.3

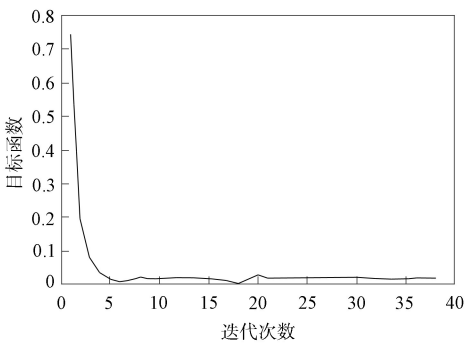


图 3　数据集 Caltech-7 上缺失率为 0.3 的收敛曲线

Fig. 3　Convergence curve with missing rate of 0.3 on Caltech-7 dataset

示系数自然对齐，在一些限制条件下学习相似图矩阵；其次，每个视图上的样本矩阵经过投影映射到一个超球体的表面上；最后，前两者经过 Laplace 正则化组合成一个统一的框架. 该算法以较快的收敛速度迭代和更新，实验结果证明了算法的有效性.

参　考　文　献

[1] CAO X C, ZHANG C Q, FU H Z, et al. Diversity-Induced Multi-view Subspace Clustering [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 586-594.

[2] KUMAR A, DAUMÉ H. A Co-training Approach for Multi-view Spectral Clustering [C]//Proceedings of the 28th International Conference on Machine Learning (ICML-11). Bellevue, WA: Omnipress, 2011: 393-400.

[3] XIA T, TAO D C, MEI T, et al. Multiview Spectral Embedding [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2010, 40(6): 1438-1446.

[4] NIE F P, LI J, LI X L. Parameter-Free Auto-weighted Multiple Graph Learning: A Framework for Multiview

Clustering and Semi-supervised Classification [C]//IJCAI. Palo Alto: AAAI Press, 2016: 1881-1887.

[5] BUCAK S S, GUNSEL B. Incremental Subspace Learning via Non-negative Matrix Factorization [J]. Pattern Recognition, 2009, 42(5): 788-797.

[6] CHAUDHURI K, KAKADE S M, LIVESCU K, et al. Multi-view Clustering via Canonical Correlation Analysis [C]//Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM, 2009: 129-136.

[7] CHEN Y Y, XIAO X L, ZHOU Y C. Multi-view Subspace Clustering via Simultaneously Learning the Representation Tensor and Affinity Matrix [J]. Pattern Recognition, 2020, 106(5): 1348-1353.

[8] SHAO W X, SHI X X, PHILIP S Y. Clustering on Multiple Incomplete Datasets via Collective Kernel Learning [C]//2013 IEEE 13th International Conference on Data Mining. Piscataway, NJ: IEEE, 2013: 1181-1186.

[9] WEN J, ZHANG Z, XU Y, et al. Unified Embedding Alignment with Missing Views Inferring for Incomplete Multi-view Clustering [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 5393-5400.

[10] ZHANG W, DENG Z H, CHOI K S, et al. End-to-End Incomplete Multiview Fuzzy Clustering with Adaptive Missing View Imputation and Cooperative Learning [J]. IEEE Transactions on Fuzzy Systems, 2022, 31(5): 1-14.

[11] WEN J, YAN K, ZHANG Z, et al. Adaptive Graph Completion Based Incomplete Multi-view Clustering [J]. IEEE Transactions on Multimedia, 2020, 23(2): 2493-2504.

[12] LIU X W, LI M M, WANG L, et al. Multiple Kernel k-Means with Incomplete Kernels [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(5): 1191-1204.

[13] GAO H, PENG Y, JIAN S. Incomplete Multi-view Clustering [C]//International Conference on Intelligent Information Processing. Berlin: Springer, 2016: 245-255.

[14] LI S Y, JIANG Y, ZHOU Z H. Partial Multi-view Clustering [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2014: 1968-1974.

[15] SHAO W, HE L, 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. Berlin: Springer, 2015: 318-334.

[16] ZHUGE W Z, TAO H, LUO T J, et al. Joint Representation Learning and Clustering: A Framework for Grouping Partial Multiview Data [J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(8): 3826-3840.

[17] WANG H, ZONG L L, LIU B, et al. Spectral Perturbation Meets Incomplete Multi-view Data [EB/OL]. (2019-03-31)[2023-02-10]. https://arxiv.org/abs/1906.00098.

[18] ZHANG L, ZHAO Y, ZHU Z F, et al. Multi-view Missing Data Completion [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(7): 1296-1309.

[19] 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(7): 1-13.

[20] FU X, HUANG K J, HONG M Y, et al. Scalable and Flexible Multiview MAX-VAR Canonical Correlation Analysis [J]. IEEE Transactions on Signal Processing, 2017, 65(16): 4150-4165.

[21] BARTELS R H, STEWART G W. Solution of the Matrix Equation $AX+XB=C$ [J]. Communications of the ACM, 1972, 15(9): 820-826.

[22] WANG H, YANG Y, LIU B, et al. A Study of Graph-Based System for Multi-view Clustering [J]. Knowledge-Based Systems, 2019, 163(5): 1009-1019.

[23] RAI N, NEGI S, CHAUDHURY S, et al. Partial Multi-view Clustering Using Graph Regularized NMF [C]//2016 23rd International Conference on Pattern Recognition (ICPR). Piscataway, NJ: IEEE, 2016: 2192-2197.

[24] SHAO W X, HE L F, LU C T, et al. Online Multi-view Clustering with Incomplete Views [C]//2016 IEEE International Conference on Big Data. Piscataway, NJ: IEEE, 2016: 1012-1017.

(责任编辑：韩 啸)