基于一致引导的不完全多视图聚类

安 萍[1] 彭军龙[2]

[1](自然资源陕西省卫星应用技术中心 陕西 西安 710119)

[2](长沙理工大学交通工程学院 湖南 长沙 410114)

摘 要 为了解决传统聚类方法存在的效果差、泛化能力弱等问题,提出一种基于一致引导的不完全多视图聚类方法。将图学习和一致性表示学习集成到一个联合框架中,从而充分利用多视图数据信息。引入的自适应学习权值向量可以平衡不同视图的影响,联合正则化表示学习策略则为一致表示学习提供了更大的自由度。提出交替迭代优化算法对聚类进行优化。在七个数据集上的实验结果表明,提出的方法能够有效提升不完全多视图聚类的效果。

关键词 多视图聚类 一致引导 图学习 正则化 自适应

中图分类号 TP391.41 TP18 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.05.039

INCOMPLETE MULTIPLE VIEW CLUSTERING BASED ON CONSISTENT GUIDANCE

An Ping[1] Peng Junlong[2]

[1](*Shaanxi Satellite Application Center for Nature Resources, Xi'an 710119, Shaanxi, China*)

[2](*School of Traffic & Transportation Engineering, Changsha University of Science & Technology, Changsha 410114, Hunan, China*)

Abstract In order to solve the problems of poor effect and weak generalization ability of traditional clustering methods, an incomplete multiple view clustering method based on consistent guidance is proposed. Graph learning and consistent representation learning were integrated into a joint framework to make full use of multiple view data information. The adaptive learning weight vector was introduced to balance the influence of different views, and the joint regularization representation learning strategy provided more freedom for consistent representation learning. An alternative iterative optimization algorithm was proposed to optimize the clustering. Experimental results on seven data sets show that the proposed method can effectively improve the effect of incomplete multiple view clustering.

Keywords Multiple view clustering Consistent guidance Graph learning Regularization Adaptive algorithm

0 引 言

多视图聚类侧重于探索从不同来源收集的多个特征表示或多个模式所提供的互补信息,从而提高聚类性能[1]。传统的多视图聚类方法只关注完整多视图数据,由于实际应用中不完整的多视图数据是很常见的,因此不完全多视图聚类成为数据挖掘领域关注的重点[2-3]。

为了对视图不完整的数据进行分割,近年来提出了许多方法,一般分为两类:基于局部视图和基于缺失信息恢复方法。基于局部视图的方法一般寻求对可用视图的某些信息进行挖掘。部分多视图聚类(PMVC)[4]、不完全多模式分组(IMG)[5]通过图正则化矩阵分解的 IMC(MCGRMF)[6]、在线多视图聚类(OMV)[7]等是代表性的基于部分视图的方法。这些方法设计了部分对齐的多重矩阵分解(MF)或加权 MF 模型,以获得所有视图共享的一致表示,用于聚类。其中,观察到的视图中某些信息被排除,并且丢失的视图的负面影响可以通过部分视图对齐或加权正则化策略

收稿日期:2021-02-03。湖南省自然科学基金重大项目(2015JJ2004)。**安萍**,讲师,主研领域:计算机科学与技术,地理信息技术。**彭军龙**,副教授。

缓解。与基于 MF 的方法不同的是,基于图的方法将特征缺失问题转化为图空间,并探索利用各视图中观测实例间的相似度信息,学习基于谱聚类的正交一致性表示。该方法对噪声具有较强的鲁棒性,适用于非线性可分离数据,代表性方法有面向扰动的 IMC(PIC)[8] 和不完全多视图自适应图学习光谱聚类(MVSCAGL)[9]。基于缺失信息恢复的方法寻求获取缺失视图或缺失视图对应的核元素,从而解决不完全学习问题。此外,还可以利用恢复的丢失信息来提高聚类性能。代表性方法包括基于核典型相关分析(KCCA)[10] 和集体核学习(CKL)[11] 等。

上述方法仍存在一些不足,例如,基于缺失信息恢复的方法无论从理论上还是实验上都不能保证恢复的内核或视图的合理性。另外上述大部分方法都不适用于具有任意缺失视图的不完全情况,虽然基于图的方法可以处理各种不完全情况,对非线性可分数据具有聚类优势,但大多忽略了跨视图的判别差异。另外,由于要从不完整数据中预先构造的固定图学习表示,而将缺失元素作为平均实例填充,因此无法得到最优的表示。

为了解决上述问题,提出一种基于一致引导的不完全多视图聚类方法。通过同时考虑视图内信息和视图间信息,从而充分利用多视图数据的信息,提升聚类精度,另外,引入自适应学习权值向量和联合正则化策略可以提升方法的自由度,进一步利用交替迭代优化提升算法性能。

1　方　法

1.1　总体结构

对于给定的多视图数据 $\{\boldsymbol{Y}^{(v)}\}_{v=1}^{l}$($\boldsymbol{Y}^{(v)}=[y_1^{(v)},y_2^{(v)},\cdots,y_n^{(v)}]\in\mathbf{R}^{m_v\times n}$),缺失实例用"*NaN*"表示,并且可用视图和缺失信息记录在对角矩阵 $\{\boldsymbol{W}^{(v)}\}_{v=1}^{l}$($\boldsymbol{W}^{(v)}\in\mathbf{R}^{n\times n}$)中,不完全多视图聚类主要是将缺少视图的这 n 个样本分割为 c 个合理的组。m_v 是第 v 个视图的特征维数,$W_{i,i}^{(v)}=1$ 表示第 i 个样本具有 v 视图,否则 $W_{i,i}^{(v)}=0$。

现有的基于图的 IMC 方法要么忽略视图之间的信息差异,要么使用固定关联图,这限制了它们的性能,因为这些方法无法获得最优的一致表示。因此提出一种图学习和一致性表示学习联合的学习框架,然后提供一种有效的优化算法。本文模型的流程如图 1 所示。

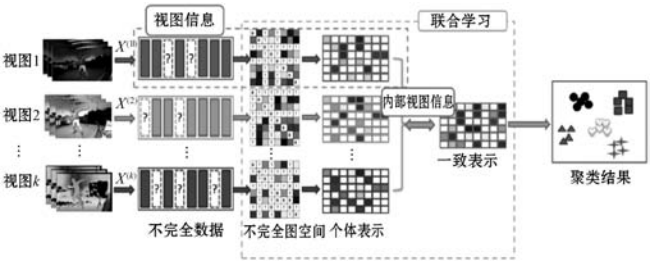


图 1　本文模型 CGIMVSC

1.2　CGIMVSC 的学习模型

多视图聚类模型可以转换为以下目标模型:

$$\min_{\alpha_v,P}\quad\sum_{v=1}^{l}(\alpha_v)^r\sum_{j=1}^{n}\sum_{i=1}^{n}\|\boldsymbol{P}_{:,i}-\boldsymbol{P}_{:,j}\|_2^2\,|Z_{i,j}^{(v)}|\qquad(1)$$

$$\text{s.t.}\quad\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}=\boldsymbol{I},\sum_{v=1}^{l}\alpha_v=1,\alpha_v>0$$

式中:P 表示视图的新表示形式;$\boldsymbol{P}_{:,i}$ 表示 P 的第 i 列向量;Z 表示视图数据。

在式(1)中,较大的 $\sum_{v=1}^{l}(\alpha_v)^r|Z_{i,j}^{(v)}|$ 将把 $\boldsymbol{P}_{:,i}$ 和 $\boldsymbol{P}_{:,j}$ 归于一起。这表明多视图聚类的核心是通过图形控制表示空间分布。但是,当给定的数据不完整且缺少视图时,就不可能在图中获得与这些缺少视图对应的相似性元素。因此,一致表示将受到不平等的约束,从而导致不合理的一致表示。不完全多视图聚类的关键问题是如何从不同大小的不完全图中得到一致表示。本文提出一种一致引导的不完全多视图聚类方法来解决上述问题,该方法寻求从个体表征中获得一个感知表征:

$$\min_{\alpha_v,P^{(v)},U}\quad\sum_{v=1}^{l}(\alpha_v)^r\sum_{j=1}^{n}\sum_{i=1}^{n}\|P_{:,i}^{(v)}-P_{:,j}^{(v)}\|_2^2\,|Z_{i,j}^{(v)}|+\lambda_1\Psi(\boldsymbol{P}^{(v)},\boldsymbol{U})\qquad(2)$$

$$\text{s.t.}\quad\boldsymbol{P}^{(v)}\boldsymbol{P}^{(v)T}=\boldsymbol{I},\sum_{v=1}^{l}\alpha_v=1,\alpha_v>0,\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{I}$$

式中:λ_1 是一个惩罚参数;n 是在一个视图中观察到的实例数量;$\boldsymbol{P}^{(v)}$ 和 $\boldsymbol{U}$ 分别表示第 v 个视图的新表示形式和所有视图共享的一致表示形式;$\boldsymbol{Z}^{(v)}$ 是第 v 个视图,其中与缺失实例相对应的元素填充为 0,这样就消除了缺失实例的不确定信息。$\Psi(\boldsymbol{P}^{(v)},\boldsymbol{U})$ 是一种度量个体表示和一致表示之间的分歧函数,定义如下:

$$\Psi(\boldsymbol{P}^{(v)},\boldsymbol{U})=\left\|\frac{\boldsymbol{K}_{P^{(v)}}}{\|\boldsymbol{K}_{P^{(v)}}\|_F^2}-\frac{\boldsymbol{K}_U}{\|\boldsymbol{K}_U\|_F^2}\right\|_F^2\qquad(3)$$

式中:选择线性核 $\boldsymbol{K}_U=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ 来度量 $\boldsymbol{U}$ 中点的相似性,因为它能很好地捕捉到数据之间的非线性关系。考虑到正交条件 $\boldsymbol{P}^{(v)}\boldsymbol{P}^{(v)\mathrm{T}}=\boldsymbol{I}$ 和 $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{I}$,可以将式(3)转换为:

$$\Psi(\boldsymbol{P}^{(v)},\boldsymbol{U})=\frac{2(c-\mathrm{Tr}(\boldsymbol{P}^{(v)T}\boldsymbol{P}^{(v)}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}))}{c^2}\qquad(4)$$

式中:c 为 $\boldsymbol{P}^{(v)}$ 和 $\boldsymbol{U}$ 的特征维数。

固定图的使用明显地限制了一致性表示学习的自由,而且,这种基于固定图的方法通常会对预构建图的质量比较敏感。为了解决这些问题,将自适应图学习策略集成到式(2)中,生成如下目标模型:

$$\min_{\alpha_v,\boldsymbol{P}^{(v)},\boldsymbol{U},\{\boldsymbol{S}^{(v)}\}_{v=1}^{l}} \sum_{v=1}^{l}(\alpha_v)^r \left(\begin{array}{l} \sum_{i,j}^{n_v}\left(\left\|\boldsymbol{x}_i^{(v)}-\boldsymbol{x}_j^{(v)}\right\|_2^2 S_{i,j}^{(v)}\right)+\lambda_1\left\|\boldsymbol{S}^{(v)}\right\|_F^2+ \\ \frac{\lambda_2}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\left\|\boldsymbol{P}_{:,i}^{(v)}-\boldsymbol{P}_{:,j}^{(v)}\right\|_2^2\left|Z_{i,j}^{(v)}\right|+ \\ \lambda_3\left(c-\mathrm{Tr}\left(\boldsymbol{P}^{(v)\mathrm{T}}\boldsymbol{P}^{(v)}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right)\right) \end{array} \right) \tag{5}$$

$$\text{s.t.}\quad \boldsymbol{P}^{(v)}\boldsymbol{P}^{(v)\mathrm{T}}=\boldsymbol{I},\ \sum_{v=1}^{l}\alpha_v=1,\ \alpha_v>0,\ \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{I},$$
$$\boldsymbol{0}\leqslant\boldsymbol{S}^{(v)}\leqslant\boldsymbol{1},\boldsymbol{S}^{(v)}\boldsymbol{1}=1,\mathrm{diag}(\boldsymbol{S}^{(v)})=\boldsymbol{0}$$

式中:λ_1、λ_2 和 λ_3 是惩罚参数。$\boldsymbol{X}^{(v)}=[x_1^{(v)},x_2^{(v)},\cdots,x_{n_v}^{(v)}]\in\mathbf{R}^{m_v\times n_v}$ 是 $Y^{(v)}$ 的一个子集,$Y^{(v)}$ 表示第 v 个视图中观察到的实例的集合。$\boldsymbol{S}^{(v)}$ 是要学习的相似度图,其中 $S_{i,j}^{(v)}$ 表示观察到的实例 $\boldsymbol{x}_i^{(v)}$ 和 $\boldsymbol{x}_j^{(v)}$ 之间的相似度。边界约束 $\{\boldsymbol{0}\leqslant\boldsymbol{S}^{(v)}\leqslant\boldsymbol{1},\boldsymbol{S}^{(v)}\boldsymbol{1}=1,\mathrm{diag}(\boldsymbol{S}^{(v)})=\boldsymbol{0}\}$ 使学习图直观地反映了样本的相似程度。通过引入约束 $\sum_{i,j}^{n_v}\left(\left\|\boldsymbol{x}_i^{(v)}-\boldsymbol{x}_j^{(v)}\right\|_2^2 S_{i,j}^{(v)}\right)+\lambda_1\left\|\boldsymbol{S}^{(v)}\right\|_F^2$,提出的方法可以自适应地学习所有视图,其中每个图 $\boldsymbol{S}^{(v)}$ 中的元素反映了该视图中实例的内在最接近的关系。例如,如果实例 $\boldsymbol{x}_i^{(v)}$ 和 $\boldsymbol{x}_j^{(v)}$ 在原特征空间中接近,则会得到较大的 $S_{i,j}^{(v)}$,说明这两个实例属于同一簇的概率很高。此外,可以发现第一项等价于 $\left\|\boldsymbol{A}^{(v)}\otimes\boldsymbol{S}^{(v)}\right\|_1$,这是一个典型的加权稀疏算法,其中⊗是一个元素化乘法算子,同时 $A_{i,j}^{(v)}=\left\|\boldsymbol{x}_i^{(v)}-\boldsymbol{x}_j^{(v)}\right\|_2^2$。根据稀疏约束的性质,式(5)将从不完整数据中获得一些稀疏图,并且只有那些最近邻的样本是连通的。

对于图 $\boldsymbol{Z}^{(v)}$ 而言,其对应于可用实例的元素就是 $\boldsymbol{S}^{(v)}$ 中的元素,而对应于缺失实例的其他元素则填入 0。实际上,$\boldsymbol{Z}^{(v)}$ 可以通过矩阵 $\boldsymbol{S}^{(v)}$ 上的行和列变换简单地得到,如下:

$$\boldsymbol{Z}^{(v)}=\boldsymbol{G}^{(v)}\boldsymbol{S}^{(v)}\boldsymbol{G}^{(v)\mathrm{T}} \tag{6}$$

式中:$\boldsymbol{G}^{(v)}\in\mathbf{R}^{n\times n_v}$。$\boldsymbol{G}^{(v)}$ 定义如下:

$$G_{i,j}^{(v)}=\begin{cases}1 & y_i^{(v)}\text{是第 }j\text{ 个观测值}\\ 0 & \text{其他}\end{cases} \tag{7}$$

式中:$y_i^{(v)}$ 表示第 v 个视图中的第 i 个实例。$\boldsymbol{G}^{(v)}$ 可以通过去掉 $\boldsymbol{W}^{(v)}$ 中所有元素都为 0 的列得到。另外,假设 $\boldsymbol{L}_{Z^{(v)}}$ 和 $\boldsymbol{L}_{S^{(v)}}$ 分别为 $\boldsymbol{Z}^{(v)}$ 和 $\boldsymbol{S}^{(v)}$ 的拉普拉斯图,则可以推导出 $\boldsymbol{L}_{Z^{(v)}}=\boldsymbol{G}^{(v)}\boldsymbol{L}_{S^{(v)}}\boldsymbol{G}^{(v)\mathrm{T}}$。根据式(5)和式(6),CGIMVSC 的最终目标学习模型转化为以下问题:

$$\min_{\{\boldsymbol{S}^{(v)},\boldsymbol{P}^{(v)}\}_{v=1}^{l},\boldsymbol{U},\alpha} \sum_{v=1}^{l}(\alpha_v)^r \left(\begin{array}{l} \sum_{i,j}^{n_v}\left(\left\|\boldsymbol{x}_i^{(v)}-\boldsymbol{x}_j^{(v)}\right\|_2^2 S_{i,j}^{(v)}\right)+\lambda_1\left\|\boldsymbol{S}^{(v)}\right\|_F^2+ \\ \lambda_2\mathrm{Tr}\left(\boldsymbol{P}^{(v)}\boldsymbol{G}^{(v)}\boldsymbol{L}_{S^{(v)}}\boldsymbol{G}^{(v)\mathrm{T}}\boldsymbol{P}^{(v)\mathrm{T}}\right)+ \\ \underbrace{\lambda_3\left(c-\mathrm{Tr}\left(\boldsymbol{P}^{(v)}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{P}^{(v)\mathrm{T}}\right)\right)}_{\text{视图信息}} \end{array} \right) \tag{8}$$

$$\text{s.t.}\quad \boldsymbol{0}\leqslant\boldsymbol{S}^{(v)}\leqslant\boldsymbol{1},\boldsymbol{S}^{(v)}\boldsymbol{1}=1,\mathrm{diag}(\boldsymbol{S}^{(v)})=\boldsymbol{0},\boldsymbol{P}^{(v)}\boldsymbol{P}^{(v)\mathrm{T}}=\boldsymbol{I},$$
$$\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{I},\ \sum_{v=1}^{l}\alpha_v=1,\ 0\leqslant\alpha_v\leqslant1$$

式中:前三项研究了关于每个视图的视图内信息,分别寻求获得所有视图的相似图和紧凑表示;第四项可以作为视图间约束,它从所有视图的特定表示中得到了一致表示。通过将视图间约束和一致表示无缝地结合到联合学习模型中,可以得到最优的相似图、紧凑的视图特定表示和一致表示,从而获得更好的性能。

此外,作为基于图的多视图聚类方法的扩展,本文方法与之前所提出的方法有许多异同。例如,本文方法和这些先前的工作都集中在探索数据的几何结构上。不同之处在于:(1) 本文方法尝试从自适应学习的相似图中获得最优的紧凑结果,而不是从数据构造的固定图中获得结果;(2) 先前的方法不适用于不完整的情况,而本文方法可以灵活地处理各种不完整和完整的多视图数据。

1.3 优化过程

本节采用交替迭代优化方法来解决式(8),式(8)的优化过程主要包括以下四个步骤:

步骤 1 计算 $\boldsymbol{U}$。通过固定式(8)中的其他变量,可将变量 $\boldsymbol{U}$ 的子问题转换为:

$$\max_{\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{I}}\sum_{v=1}^{l}(\alpha_v)^r\mathrm{Tr}\left(\boldsymbol{U}\boldsymbol{P}^{(v)\mathrm{T}}\boldsymbol{P}^{(v)}\boldsymbol{U}^{\mathrm{T}}\right)\Leftrightarrow$$
$$\max_{\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{I}}\mathrm{Tr}\ \boldsymbol{U}\left(\left(\sum_{v=1}^{l}(\alpha_v)^r\boldsymbol{P}^{(v)\mathrm{T}}\boldsymbol{P}^{(v)}\right)\boldsymbol{U}^{\mathrm{T}}\right) \tag{9}$$

变量 $\boldsymbol{U}$ 的最优解由 $\left(\sum_{v=1}^{l}(\alpha_v)^r\boldsymbol{P}^{(v)\mathrm{T}}\boldsymbol{P}^{(v)}\right)$ 的 c 个最大特征值所对应的 c 个特征向量组成。

步骤 2 计算 $\boldsymbol{S}^{(v)}$。通过求解以下问题得到变量 $\boldsymbol{S}^{(v)}$:

$$\min_{S^{(v)}}\quad \sum_{i,j}^{n_v}\left(\left\|\boldsymbol{x}_i^{(v)}-\boldsymbol{x}_j^{(v)}\right\|_2^2 S_{i,j}^{(v)}\right)+\lambda_1\left\|\boldsymbol{S}^{(v)}\right\|_F^2+$$
$$\lambda_2\mathrm{Tr}\left(\boldsymbol{P}^{(v)}\boldsymbol{G}^{(v)}\boldsymbol{L}_{S^{(v)}}\boldsymbol{G}^{(v)\mathrm{T}}\boldsymbol{P}^{(v)\mathrm{T}}\right)$$
$$\text{s.t.}\quad \boldsymbol{0}\leqslant\boldsymbol{S}^{(v)}\leqslant\boldsymbol{1},\boldsymbol{S}^{(v)}\boldsymbol{1}=1,\mathrm{diag}(\boldsymbol{S}^{(v)})=\boldsymbol{0} \tag{10}$$

对于式(10),有:$A_{i,j}^{(v)}=\left\|\boldsymbol{x}_i^{(v)}-\boldsymbol{x}_j^{(v)}\right\|_2^2$,$\boldsymbol{H}^{(v)}=\boldsymbol{P}^{(v)}\boldsymbol{G}^{(v)}$ 和 $M_{i,j}^{(v)}=\left\|\boldsymbol{H}_{:,i}^{(v)}-\boldsymbol{H}_{:,j}^{(v)}\right\|_2^2$,则:

$$\sum_{i,j}^{n_v}\left(\left\|\boldsymbol{x}_i^{(v)}-\boldsymbol{x}_j^{(v)}\right\|_2^2 S_{i,j}^{(v)}\right)+\lambda_1\left\|\boldsymbol{S}^{(v)}\right\|_F^2+\lambda_2\mathrm{Tr}\left(\boldsymbol{P}^{(v)}\boldsymbol{G}^{(v)}\boldsymbol{L}_{S^{(v)}}\boldsymbol{G}^{(v)\mathrm{T}}\boldsymbol{P}^{(v)\mathrm{T}}\right)=$$

$$\sum_{i,j}^{n_v}(A_{i,j}^{(v)}S_{i.j}^{(v)})+\lambda_1\|S^{(v)}\|_F^2+\lambda_2\mathrm{Tr}(P^{(v)}G^{(v)}L_{S^{(v)}}G^{(v)\mathrm{T}}P^{(v)\mathrm{T}})=$$
$$\sum_{i,j}^{n_v}(A_{i,j}^{(v)}S_{i.j}^{(v)})+\lambda_1\|S^{(v)}\|_F^2+\frac{\lambda_2}{2}\sum_{i,j}^{n_v}\|H_{:,i}^{(v)}-H_{:,j}^{(v)}\|_2^2S_{i,j}^{(v)}=$$
$$\sum_{i,j}^{n_v}(A_{i,j}^{(v)}S_{i.j}^{(v)})+\lambda_1\|S^{(v)}\|_F^2+\frac{\lambda_2}{2}\sum_{i,j}^{n_v}M_{i,j}^{(v)}S_{i,j}^{(v)} \tag{11}$$

式(10)等价于:

$$\min\sum_{i,j}^{n_v}\left(\left(A_{i,j}^{(v)}+\frac{\lambda_2}{2}M_{i,j}^{(v)}\right)S_{i.j}^{(v)}+\lambda_1S_{i.j}^{(v)2}\right)\Leftrightarrow$$
$$\min\sum_{i,j}^{n_v}\left(S_{i.j}^{(v)}+\frac{A_{i,j}^{(v)}}{2\lambda_1}+\frac{\lambda_2}{4\lambda_1}M_{i,j}^{(v)}\right)^2 \tag{12}$$

式中:$\psi=\{\mathbf{0}\leqslant S^{(v)}\leqslant\mathbf{1},S^{(v)}1=1,\mathrm{diag}(S^{(v)})=0\}$。显然,式(12)可以逐行优化,其最优解为:

$$S_{i.j}^{(v)}=\begin{cases}\left(-\frac{A_{i,j}^{(v)}}{2\lambda_1}-\frac{\lambda_2}{4\lambda_1}M_{i,j}^{(v)}+\eta_i\right)_+ & i\neq j\\ 0 & i=j\end{cases} \tag{13}$$

式中:如果 $a>0$ 函数 $(a)_+=a$,否则 $(a)_+=0$。根据 $S^{(v)}1=1$,η_i 表示如下:

$$\eta_i=\frac{1}{n-1}+\frac{1}{n-1_j}\sum_{j=1,j\neq i}^{n_v}\left(\frac{A_{i,j}^{(v)}}{2\lambda_1}+\frac{\lambda_2}{4\lambda_1}M_{i,j}^{(v)}\right) \tag{14}$$

步骤 3　计算 $P^{(v)}$。固定式(8)中的其他变量,将变量 $P^{(v)}$ 的子问题转化为:

$$\min_{P^{(v)}P^{(v)T}=I}\lambda_2\mathrm{Tr}(P^{(v)}G^{(v)}L_{S^{(v)}}G^{(v)T}P^{(v)\mathrm{T}})+$$
$$\lambda_2(c-\mathrm{Tr}(P^{(v)\mathrm{T}}P^{(v)}U^{\mathrm{T}}U))\Leftrightarrow$$
$$\min_{P^{(v)}P^{(v)\mathrm{T}}=I}\mathrm{Tr}(P^{(v)}(\lambda_3U^{\mathrm{T}}U-\lambda_2G^{(v)}L_{S^{(v)}}G^{(v)\mathrm{T}})P^{(v)\mathrm{T}}) \tag{15}$$

式(15)是与式(9)类似的特征值分解问题。假设前 c 个最大特征值是 $\sigma_1>\sigma_2>\cdots>\sigma_c$,相应的特征向量是 $p_1^{(v)},p_2^{(v)},\cdots,p_c^{(v)}$,则式(15)的最优解为 $P^{(v)}=[p_1^{(v)},p_2^{(v)},\cdots,p_c^{(v)}]^{\mathrm{T}}$。

步骤 4　计算 α_v:定义 $\gamma_v=\sum_{i,j}^{n_v}(\|x_i^{(v)}-x_j^{(v)}\|_2^2S_{i.j}^{(v)})+\lambda_2\mathrm{Tr}(P^{(v)}G^{(v)}L_{S^{(v)}}G^{(v)\mathrm{T}}P^{(v)\mathrm{T}})+\lambda_1\|S^{(v)}\|_F^2+\lambda_3(c-\mathrm{Tr}(P^{(v)}U^{\mathrm{T}}UP^{(v)\mathrm{T}}))$,并且固定变量 $\{S^{(v)},P^{(v)}\}_{v=1}^{l}$ 和 U,向量 α 的子问题被分解为:

$$\min_{\sum_{v=1}^{l}\alpha_v=1,0\leqslant\alpha_v\leqslant1}\sum_{v=1}^{l}(\alpha_v)^r\gamma_v \tag{16}$$

容易推导得到最佳 α_v 为:

$$\alpha_v=\frac{(\gamma_v)^{\frac{1}{(1-r)}}}{\sum_{v=1}^{l}(\gamma_v)^{\frac{1}{(1-r)}}} \tag{17}$$

通过根据上述优化过程迭代更新变量 U,$\{S^{(v)}\}_{v=1}^{l}$、α 和 $\{P^{(v)}\}_{v=1}^{l}$,最终可以获得目标问题的局部最优解。算法 1 总结了上述优化过程。通过对一致表示 U 进行 K-means 作为常规的内模聚集方法,得到最终的聚类结果。

算法 1　CGIMVSC

输入:不完整多视图数据 $\{X^{(v)}\}_{v=1}^{l}$,索引矩阵 $\{G^{(v)}\}_{v=1}^{l}$,参数为 λ_1、λ_2 和 λ_3。

输出:U。

(1) 初始化:将 $\{S^{(v)}\}_{v=1}^{l}$ 初始化为每个视图的 k 近邻图;初始化 $P^{(v)}$ 为拉普拉斯图 $G^{(v)}S^{(v)}G^{(v)\mathrm{T}}$ 的 c 个最小特征值对应的 c 个特征向量的集合;

(2) **while** 不收敛 **do**

(3)　通过式(9)更新变量 U;

(4)　通过求解式(13)更新变量 $\{S^{(v)}\}_{v=1}^{l}$;

(5)　通过式(15)更新变量 $\{P^{(v)}\}_{v=1}^{l}$;

(6)　通过式(17)更新变量 α;

(7) **end while**

2　算法分析

2.1　计算复杂度分析

算法 1 将目标问题分解为 4 个主要子问题,其中图 $\{S^{(v)}\}_{v=1}^{l}$ 的第二个子问题和权向量的第四个子问题只涉及一些矩阵和向量的基本运算。因此,可以忽略步骤 2 和步骤 4 的计算复杂度。在步骤 1 和步骤 3 中,主要的计算代价是特征值分解,对于一个 $n\times n$ 的矩阵,其代价约为 $O(n^3)$。对于一个 $n\times n$ 矩阵,其计算复杂度为 $O(cn^2)$ 左右。因此,步骤 1 和步骤 3 的计算复杂度分别为 $O(cn^2)$ 和 $O(lcn^2)$。通过充分考虑上述四个步骤,并假设总迭代数为 τ,算法 1 的总计算复杂度约为 $O(\tau(lcn^2+cn^2))$。

2.2　收敛性分析

定理 1　利用算法 1 所列的交替优化算法,式(8)的目标函数值单调递减。

证明:从这四个变量的子目标问题:即式(9)、式(10)、式(15)和式(16)来看,它们都属于凸优化问题。假设 $\Gamma(U_t,\{S_t^{(v)}\}_{v=1}^{l},\{P_t^{(v)}\}_{v=1}^{l},\alpha_t)$ 是第 t 次迭代的目标函数值,通过对变量 U、$S^{(v)}$、$P^{(v)}$ 和 α 逐一求解,得出如下不等式:

$$\begin{aligned}&\Gamma(U_t,\{S_t^{(v)}\}_{v=1}^{l},\{P_t^{(v)}\}_{v=1}^{l},\alpha_t)\geqslant\\&\Gamma(U_{t+1},\{S_t^{(v)}\}_{v=1}^{l},\{P_t^{(v)}\}_{v=1}^{l},\alpha_t)\geqslant\\&\Gamma(U_{t+1},\{S_{t+1}^{(v)}\}_{v=1}^{l},\{P_t^{(v)}\}_{v=1}^{l},\alpha_t)\geqslant\\&\Gamma(U_{t+1},\{S_{t+1}^{(v)}\}_{v=1}^{l},\{P_{t+1}^{(v)}\}_{v=1}^{l},\alpha_t)\geqslant\\&\Gamma(U_{t+1},\{S_{t+1}^{(v)}\}_{v=1}^{l},\{P_{t+1}^{(v)}\}_{v=1}^{l},\alpha_{t+1})\end{aligned} \tag{18}$$

式(18)验证了算法 1 中所列的交替优化方法的单调递减性质,从而证明了定理 1。

从式(8)可以看出,它的目标函数值下界为 0。结

合定理 1 所证明的单调递减性质,可以得出结论,式(8)通过使用算法 1 所列的交替优化算法,最终收敛于局部最优点。

3 实 验

3.1 数据集和评估指标

表 1 总结了七个多视图数据集的基本信息。

表 1 多视图数据集的描述

数据集	类别数量	视图数量	样本数量	所有视图的特征维度
BBCSport	5	4	116	1 991 206 321 132 158
Handwritten	10	5	2 000	240 762 164 764
NNIST	10	2	4 000	784 784
Citeseer	6	2	3 321	33 123 703
ORL	40	3	400	409 633 046 750
COIL20	20	4	1 440	10 245 126 801 024
Caltech101	102	4	9 144	2 541 984 512 928

(1) BBCsport:一个体育新闻文章数据集,在这个数据集中,每篇文章被分成 4 个子部分,这些子部分被认为是 4 个视图。在该实验中,选取一个包含 116 个样本的子集,充分观察了 4 个视图,以验证所提出方法的有效性。

(2) Handwritten:每个类包含 10 个数字和 200 幅图像。从荷兰效用图收集的原始公共数据集中,每幅图像中提取了六种特征,即傅里叶系数、轮廓相关性、卡洪恩-洛夫系数、像素平均、泽尼克矩和形态逻辑特征。

(3) MNIST:一个非常流行的大规模手写数字数据集,其中有 60 000 个训练样本和由 10 个数字提供的 10 000 个测试样本。

(4) Citeseer:一个多视图文档数据集,包含 3 312 篇科学出版物。这些文档分为六类,即 Agent、AI、DB、IR、ML 和 HCI。文档之间的引文链接被提取为引文视图,然后从文档中提取一个二进制单词向量作为内容视图。

(5) ORL:在人脸识别领域中流行的人脸数据库,它包含 40 位志愿者提供的 400 幅面部图像,其中每个人都有 10 幅面部图像,通过 LBP、Gabor 和 intensity 作为三个视图来表示每幅人脸图像。

(6) COIL20:包含 20 个对象,每个对象提供 72 幅不同角度的图像。从图像中提取四种类型的特征,即 LBP、PHOG、GIST 和像素灰度值作为 4 个视图。

(7) Caltech101:包含 9 发 144 幅杂乱背景图像和 101 个物体,包括飞机、蚂蚁、鲈鱼和海狸等。从每幅图像中提取四种特征,即 Cenhist、HOG、GIST 和 LBP。

不完全数据构建:在 MNIST 和 Citeseer 数据集上,随机选择了 $p\%$ ($p \in \{10,30,50,70\}$) 个样本作为配对样本,它们的所有视图都会被完全观察到。本文从剩余的样本中随机选择 50% 的样本,并删除其第一个视图。对于其余样本,删除其第二个视图。以这种方式,构建了具有 $p\%$ 配对视图的不完整 MNIST 和 Citeseer 数据集。对于其他数据集,采用文献[12]中相同方法来构建不完整的多视图数据,其丢失视图率为 $p\%$。

评估指标:三个标准评价指标,选取了聚类评价中广泛使用的聚类精度(ACC)、非标准化互信息(NMI)和纯度(Purity)。这些度量的域为[0,1]。

3.2 比较方法

对比方法包括九种 IMC 方法,即 PMVC[4]、IMG[5]、多不完整视图聚类(MIC)[13]、OMVC[7]、DAIMC[14]、OPIMC[15]、PIC[8]、不完整的多核 k-means 和相互核完成(MKKM-IK-MKC)[16] 和通过一致性 GAN(PMVC-GAN)进行的局部多视图聚类[17] 比较了以下四种基准方法:

(1) 最佳单视图(BSV)[18]:BSV 可被视为单视图 k-means 聚类方法,它不利用多视图中的任何补充信息进行数据聚类。即它将对所有视图独立执行 k-means,然后输出性能最佳的一个视图的聚类结果。

(2) Concat: Concat 在一个由所有视图的特性连接起来的单视图上执行 k-means 聚类。

(3) 多视图非负矩阵分解(Multiple-view Non-negative Matrix Factorization, MultiNMF)[8]:多项式是一种著名的基于非负矩阵分解的多视图聚类方法,它设计了一个联合的 NMF 模型,在约束下,从所有视图的聚类解中获得联合一致表示。

(4) 基于质心的联合正则化多视图光谱聚类(CCo-MVSC)[19]:CCo-MVSC 将光谱聚类和协同学习集成到一个框架中,该框架旨在通过最大程度地减少它们之间的分歧,将所有聚类解决方案推向一个一致表示。

对于 BSV、Concat 和多项式,本文将每个视图缺少的实例填充为预先在该视图中观察到的实例的平均值。对于 CCo-MVSC,先分别从所有视图的可用实例中构造图,然后将图转换成相同大小的图,大小为 $n \times n$,其中与缺失实例对应的元素填充为 0,n 表示样本数量。

3.3　实验结果与分析

表 2 给出了 ORL 数据集上聚类结果的平均值和标准差。这些方法在其他 6 个数据集上的实验结果如表 3 – 表 8 所示。实验结果反映了以下几点:

(1) 在几乎所有情况下,所提出的 CGIMVSC 在具有不同视角缺失率和配对视角率的七个数据集上均达到最佳性能。例如,CGIMVSC 在手写数据集的聚类 ACC 方面比第二优方法提高了 4% ~8%。这验证了所提出的 CGIMVSC 对不完整多视图聚类任务的有效性。

(2) 可以发现基于图学习的方法和基于核学习的方法比基于矩阵分解的方法表现更好。这说明,将特征缺失问题转移到图空间或核空间中有可能改善 IMC 性能。另外,基于图学习和基于核学习的内模控制方法可以减少不完整数据中出现的信息不平衡问题的负面影响。

(3) 在 7 个数据集上,本文方法的性能优于 PIC 方法。PIC 是一种基于代表性图的 IMC 方法,由于它从固定图生成聚类结果,因此对从数据中预先构造的图的质量很敏感。与 PIC 不同的是,提出的方法在一个框架中同时进行一致表示学习和图的构建,这两个学习任务可以相互促进。通过这种方式,本文方法可以获得最优的一致表示,并捕获每个视图的可用实例的内在结构,因此提出方法比 PIC 性能更好。

(4) 在大多数情况下,BSV、Concat 和 MultiNMF 的性能比大多数 IMC 方法差。除此之外,可以发现,随着缺失视图数量的增加,这三种非 IMC 方法的聚类性能明显下降。这些方法的共同点是,它们在聚类之前将缺少的视图填充为平均实例。这样的填充实例将严重干扰聚类决策,因为这些填充实例可以自然地视为同一聚类,因此,将缺少的实例填充为平均实例或其他向量对 IMC 任务是不利的,尤其是对于具有较大丢失率的情况。

表 2　ORL 数据集上不同方法的 ACC、NMI 和纯度的平均值和标准偏差

方法	ACC			NMI			Purity		
	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$
BSV	60.98 ±2.76	50.25 ±1.94	36.63 ±1.50	74.75 ±1.06	59.37 ±0.85	43.26 ±0.88	65.67 ±2.41	53.78 ±1.28	39.77 ±1.51
Concat	53.85 ±2.09	44.82 ±1.95	34.52 ±0.53	69.90 ±1.06	56.50 ±1.04	41.51 ±0.65	58.92 ±1.36	48.55 ±1.66	37.58 ±0.70
MultiNMF	63.55 ±2.75	50.70 ±2.06	38.85 ±1.94	77.17 ±1.81	64.09 ±2.16	46.39 ±2.48	67.95 ±2.61	55.05 ±1.80	41.20 ±0.70
CCo-MVSC	74.83 ±1.49	70.55 ±0.92	60.97 ±0.89	85.92 ±0.67	82.55 ±0.95	77.22 ±0.72	77.11 ±1.64	74.28 ±1.32	65.33 ±1.24
MIC	59.92 ±2.32	54.67 ±1.36	46.61 ±2.07	76.91 ±1.75	69.53 ±0.87	60.11 ±1.76	65.64 ±1.97	59.53 ±0.83	50.41 ±2.07
OMVC	45.72 ±2.65	35.25 ±2.89	25.22 ±1.10	63.11 ±1.55	54.39 ±2.65	49.33 ±1.54	49.17 ±2.45	38.01 ±2.70	30.03 ±1.25
DAIMC	68.90 ±2.00	68.52 ±1.53	55.00 ±3.96	84.11 ±0.63	80.88 ±0.68	71.50 ±2.47	73.15 ±1.39	73.08 ±1.21	59.40 ±3.61
OPIMC	43.80 ±1.43	36.65 ±1.34	29.52 ±1.55	62.87 ±1.82	54.89 ±1.54	49.41 ±1.78	46.85 ±2.67	37.65 ±1.09	32.75 ±2.33
MKKM-IK-MKC	72.36 ±1.75	71.13 ±0.73	60.52 ±1.98	85.27 ±0.95	82.03 ±0.36	76.35 ±0.88	75.88 ±1.37	74.44 ±0.69	65.42 ±1.67
PIC	71.52 ±3.34	69.80 ±1.24	56.63 ±0.92	84.98 ±2.23	81.12 ±0.84	73.76 ±0.75	75.17 ±3.39	72.88 ±1.05	60.43 ±1.43
本文方法	**76.21 ±1.11**	**73.50 ±0.66**	**62.54 ±2.09**	**87.39 ±0.33**	**83.67 ±0.18**	**78.38 ±1.36**	**78.40 ±0.84**	**75.25 ±0.66**	**66.52 ±1.97**

表 3　BBCSport 数据集上不同方法的 ACC 和 NMI 的平均值和标准偏差

方法	ACC				NMI			
	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$
BSV	58.62 ±3.94	51.31 ±5.33	44.03 ±3.78	36.43 ±2.95	43.73 ±7.43	31.03 ±2.08	21.40 ±2.61	12.05 ±2.28
Concat	70.62 ±3.76	58.72 ±5.42	33.21 ±2.19	35.95 ±4.11	61.69 ±6.72	38.92 ±7.87	18.61 ±1.44	8.58 ±3.05
MultiNMF	48.58 ±4.25	42.75 ±7.33	40.34 ±8.07	35.69 ±3.74	23.48 ±3.15	18.25 ±7.74	14.79 ±8.36	7.84 ±1.49
CCo-MVSC	72.76 ±4.13	70.06 ±4.50	61.38 ±6.17	35.86 ±4.80	62.79 ±6.52	57.69 ±3.54	39.55 ±8.00	14.11 ±4.49
MIC	51.21 ±4.21	46.21 ±4.71	46.03 ±5.19	36.66 ±3.50	29.90 ±6.25	25.84 ±3.24	24.01 ±5.39	9.29 ±2.44
OMVC	53.33 ±3.21	51.38 ±3.06	48.79 ±3.10	38.13 ±4.67	30.64 ±2.00	41.57 ±2.79	40.63 ±2.45	11.74 ±5.35
DAIMC	68.62 ±4.59	63.45 ±10.97	56.89 ±5.59	39.59 ±6.17	56.62 ±4.60	50.17 ±9.91	37.89 ±6.22	17.16 ±6.11
OPIMC	54.14 ±4.78	52.93 ±4.93	45.69 ±6.00	44.34 ±6.08	35.66 ±4.71	31.56 ±6.10	21.75 ±6.44	14.65 ±3.51
MKKM-IK-MKC	77.55 ±2.01	75.66 ±3.01	67.07 ±3.51	45.07 ±6.54	72.91 ±3.29	64.42 ±4.69	53.52 ±4.74	22.07 ±6.13
PIC	75.52 ±1.57	74.48 ±3.32	69.48 ±6.02	31.89 ±2.20	70.94 ±2.22	64.18 ±2.74	53.91 ±6.22	9.99 ±3.08
本文方法	**79.17 ±1.85**	**76.21 ±4.16**	**71.21 ±1.88**	**46.21 ±4.75**	**74.68 ±2.51**	**69.34 ±3.67**	**57.56 ±5.24**	**23.87 ±4.56**

表 4 手写数据集上不同方法的 ACC 和 NMI 的平均值和标准偏差

方法	ACC				NMI			
	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$
BSV	68.27±5.66	51.49±2.29	38.24±2.25	27.15±1.31	62.82±3.24	47.01±1.71	32.21±1.00	19.48±0.69
Concat	75.06±3.86	55.48±1.57	42.19±0.99	28.31±0.75	73.05±2.11	51.66±0.99	38.24±1.59	23.50±0.95
MultiNMF	82.35±2.30	71.74±4.53	52.03±3.71	31.85±1.79	72.05±2.11	60.11±3.08	41.99±1.94	20.88±1.60
CCo-MVSC	74.61±2.99	73.17±4.64	70.15±4.11	64.62±3.61	70.89±1.68	68.23±2.79	62.86±1.84	53.14±2.22
MIC	77.59±2.41	73.29±3.41	61.27±3.16	41.34±2.69	70.84±2.08	65.39±2.08	52.95±1.33	34.71±2.11
OMVC	65.04±6.50	55.00±5.06	36.40±4.93	29.80±4.63	56.72±5.05	44.99±4.56	35.16±4.62	25.83±8.37
DAIMC	88.86±0.63	86.73±0.79	81.92±0.88	60.44±6.87	79.78±0.71	76.65±1.07	68.77±0.99	47.10±4.79
OPIMC	80.20±5.40	76.45±5.15	69.50±6.54	56.66±10.06	77.26±3.11	73.74±3.42	66.57±4.18	51.86±7.97
MKKM-IK-MKC	71.78±1.74	69.07±0.73	66.08±3.25	55.55±1.39	69.43±1.28	65.42±0.61	59.04±2.69	47.36±1.78
PIC	84.20±0.98	83.90±0.17	83.24±0.28	80.97±1.70	85.41±0.91	84.79±1.79	82.25±1.66	77.56±0.59
本文方法	**96.33±0.24**	**92.37±6.15**	**89.50±6.17**	**84.71±5.33**	**92.31±0.29**	**88.45±2.65**	**83.46±2.78**	**77.81±2.30**

表 5 MNIST 数据集上不同方法的 ACC 和 NMI 的平均值和标准偏差指标

方法	ACC				NMI			
	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$
BSV	33.25±1.79	37.37±0.69	42.76±1.30	47.95±1.36	27.20±0.89	31.39±1.28	37.45±1.42	42.49±1.47
Concat	36.88±1.96	39.24±1.47	43.79±1.71	47.37±1.08	34.48±1.09	33.38±0.54	37.42±1.38	43.17±0.69
MultiNMF	40.95±3.14	43.24±2.17	43.57±1.79	44.25±3.45	33.20±2.67	36.99±1.03	37.35±0.94	37.14±1.08
CCo-MVSC	45.10±1.81	45.89±1.88	46.86±1.43	47.86±1.14	37.72±0.59	39.22±1.05	40.43±0.78	41.36±0.77
PMVC	41.36±2.29	43.42±2.99	44.68±1.23	45.84±1.59	35.46±0.25	38.51±1.63	39.43±1.37	39.83±1.71
IMG	46.34±3.36	47.13±2.24	46.88±1.51	48.31±1.22	39.74±2.42	40.71±2.56	39.87±1.05	44.16±1.09
MIC	43.96±2.38	44.42±2.28	44.17±1.37	45.38±2.82	38.77±1.35	40.81±1.28	40.53±0.67	41.61±1.50
OMVC	40.44±2.95	42.23±2.17	40.36±2.20	41.44±3.39	36.21±1.47	36.68±2.16	35.64±1.89	32.25±2.95
DAIMC	45.33±4.12	48.19±1.38	49.25±1.67	49.36±1.87	37.46±3.04	41.09±1.58	43.47±0.82	44.15±0.75
OPIMC	41.40±2.51	48.02±2.63	47.77±3.39	48.71±2.44	34.29±2.33	43.98±1.98	44.63±1.47	45.65±1.15
MKKM-IK-MKC	47.56±2.18	51.02±0.66	51.72±0.58	52.45±0,41	40.39±1.17	42.76±0.70	43.88±0.54	45.10±0.39
PIC	56.88±1.59	58.66±0.55	58.80±2.21	59.09±1.19	57.19±1.58	60.60±0.76	61.29±0.66	61.59±0.44
PMVC_CGAN	45.17±0.86	48.36±0.71	52.80±0.78	52.02±0.70	39.33±	43.22±	49.61±	48.22±
本文方法	**57.84±2.00**	**59.33±2.91**	**60.92±3.23**	**60.94±2.56**	**57.31±0.87**	**61.48±0.24**	**61.97±2.66**	**62.46±2.48**

表 6 Citeseer 数据集上不同方法的 ACC 和 NMI 的平均值和标准偏差

方法	ACC				NMI			
	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$
BSV	30.56±1.80	32.61±2.29	33.46±1.49	33.75±3.34	8.10±1.21	10.23±2.10	11.00±1.44	11.12±2.87
Concat	30.93±0.72	33.07±4.23	37.25±1.85	41.14±1.84	8.71±0.51	10.23±2.86	13.81±1.57	18.60±1.26
MultiNMF	33.77±1.95	37.02±1.21	36.55±3.28	36.09±5.79	10.16±1.31	13.58±1.57	14.13±2.96	13.02±4.40
CCo-MVSC	33.13±1.21	37.05±0.37	39.65±1.88	41.53±0.78	11.11±1.37	13.46±2.41	15.99±2.09	16.21±0.79
PMVC	34.93±2.77	40.65±3.22	43.94±3.81	44.98±4.33	12.10±1.72	16.91±1.49	18.14±2.88	19.26±2.78
IMG	35.31±1.80	36.55±0.56	39.07±1.29	40.35±4.82	14.36±1.31	14.91±1.19	16.61±1.11	17.65±3.33
MIC	36.22±2.03	38.98±0.85	42.38±1.83	46.09±1.12	13.57±1.75	15.35±0.73	19.25±3.13	21.72±1.98
OMVC	26.20±1.04	27.25±2.01	33.71±2.35	40.35±1.98	11.15±1.01	13.55±2.44	16.64±1.66	18.36±3.21
DAIMC	30.29±2.97	34.66±4.59	35.72±4.40	45.39±5.03	7.83±2.99	11.34±3.59	12.56±3.94	20.39±2.93
OPIMC	29.99±3.01	30.95±3.45	33.53±4.82	34.92±5.09	6.10±1.16	7.62±2.27	9.98±2.92	11.23±3.17
MKKM-IK-MKC	39.36±1.93	39.82±3.14	43.86±2.46	46.39±2.37	16.14±2.62	15.80±2.01	18.24±2.80	19.78±1.38
PIC	32.91±2.19	33.52±2.27	36.08±3.02	37.33±1.83	12.23±1.96	14.57±2.25	15.44±3.97	16.78±2.77
PMVC_CGAN	—	—	—	—	—	—	—	—
本文方法	51.25±2.43	6.92±4.36	60.04±1.95	63.43±0.66	22.64±0.69	28.56±2.92	32.28±1.67	34.92±0.94

表 7　COIL20 数据集上不同方法的 ACC 和 NMI 的平均值和标准偏差

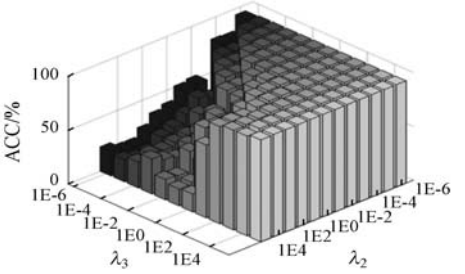
方法	ACC				NMI			
	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$
BSV	53.73±3.00	42.20±1.40	30.88±1.49	21.74±0.78	64.26±1.77	50.15±0.57	34.61±1.26	21.75±0.30
Concat	46.64±2.51	39.06±1.64	29.91±1.59	20.14±1.05	60.07±1.44	46.08±0.77	32.94±1.48	20.12±0.89
MultiNMF	67.65±2.39	62.81±1.93	49.50±1.93	34.86±2.58	74.71±1.88	67.01±0.85	56.20±1.09	43.37±2.59
CCo-MVSC	66.45±1.20	66.41±2.00	65.83±2.32	54.51±1.72	77.92±0.31	76.76±0.83	74.94±1.60	62.05±1.63
MIC	61.33±2.03	60.95±1.96	60.28±1.93	37.28±3.33	73.92±0.93	71.07±1.09	67.45±0.93	45.22±2.31
OMVC	52.34±1.32	49.44±2.51	38.54±2.48	25.32±1.58	59.86±0.63	55.77±2.13	44.02±1.97	30.68±2.15
DAIMC	72.25±2.68	71.61±1.92	71.32±3.23	51.26±4.65	79.71±1.31	78.95±1.70	77.03±1.41	57.19±2.79
OPIMC	55.43±4.22	47.85±5.87	45.06±4.45	27.97±1.64	68.40±1.51	60.54±3.37	53.37±2.72	35.46±1.52
MKKM-IK-MKC	71.41±0.70	71.33±0.91	69.78±0.74	48.76±3.09	80.14±0.64	79.57±0.68	77.07±0.76	57.71±2.51
PIC	74.15±2.64	73.37±1.56	72.59±0.71	62.78±1.96	83.67±1.14	82.21±0.71	80.66±0.45	69.84±1.39
提出方法	**77.05±1.02**	**75.77±1.23**	**74.46±0.83**	**64.77±1.94**	**85.75±1.45**	**84.33±1.56**	**82.17±0.03**	**70.13±1.22**

表 8　Caltech 101 数据集上不同方法的 ACC 和 NMI 的平均值和标准偏差指标

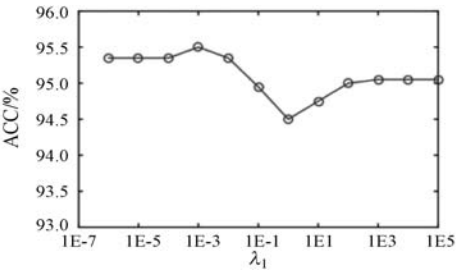
方法	ACC				NMI			
	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$	$p\%=0.1$	$p\%=0.3$	$p\%=0.5$	$p\%=0.7$
BSV	22.57±0.25	19.36±0.37	16.55±0.33	13.66±0.09	42.94±0.12	33.57±0.11	23.57±0.18	13.87±0.17
Concat	22.51±0.66	17.59±0.46	14.12±0.32	12.43±0.54	43.38±0.22	37.13±0.14	31.38±0.31	31.12±0.25
MultiNMF	21.42±0.62	9.28±0.95	14.98±0.86	9.97±0.48	44.71±0.41	39.70±0.47	31.92±0.66	26.67±0.43
CCo-MVSC	26.26±0.46	24.92±0.25	22.53±0.25	16.37±0.51	46.77±0.31	45.86±0.16	40.97±0.29	34.89±0.43
MIC	22.82±0.57	20.12±0.75	18.44±0.36	13.79±0.69	44.55±0.79	38.12±0.94	31.78±0.56	31.33±0.84
OMVC	23.45±0.77	20.59±0.34	19.75±0.89	14.55±0.40	45.05±0.53	38.52±0.51	33.81±1.11	33.16±0.45
DAIMC	24.93±1.41	25.15±1.47	**23.21±1.12**	16.43±1.39	47.81±0.50	46.81±0.25	40.97±1.29	35.53±0.51
OPIMC	26.78±1.24	25.04±1.47	21.63±1.17	13.58±1.32	31.39±0.57	25.63±1.37	33.58±0.18	17.35±1.54
MKKM-IK-MKC	15.11±0.28	14.35±0.21	14.83±0.50	12.39±0.16	32.72±0.14	32.88±0.28	33.58±0.29	33.05±0.16
PIC	25.33±2.38	24.53±1.02	22.24±1.11	14.72±0.38	46.46±3.24	45.40±1.32	41.94±1.10	35.09±0.25
本文方法	**27.11±0.67**	**26.32±0.71**	22.88±0.56	**17.23±0.45**	**48.02±0.35**	**47.33±0.75**	**42.44±1.01**	**36.05±0.07**

3.4　参数分析及消融研究

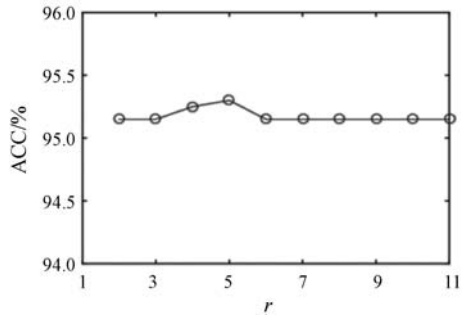
参数分析:式(8)具有四个可调参数,即 λ_1、λ_2、λ_3 和比例参数 r。本节结合在手写和 MNIST 数据集上的实验结果来分析这些参数相对于聚类 ACC 的敏感性。从图 2 和图 3 的实验结果可以看出:(1) 当参数 λ_2 和 λ_3 分别在 $[10^{-6},10^{5}]$ 和 $[10^{2},10^{5}]$ 的范围内选取时,本文方法可以获得较好的性能。(2) 当将 λ_1 设置为 $[10^{-6},10^{-1}]$ 中的较小值时,可以获得良好的性能。(3) 所提出的方法对手写数据集上的参数 r 不敏感,当 r 位于 $[2,11]$ 中时,可以获得一致的良好性能。在 MNIST 数据集上,最好在 $[2,3]$ 的范围内选择较小的 r。以上结果表明,所提出的 CGIMVSC 对这些参数在一定程度上是不敏感的。



(a)

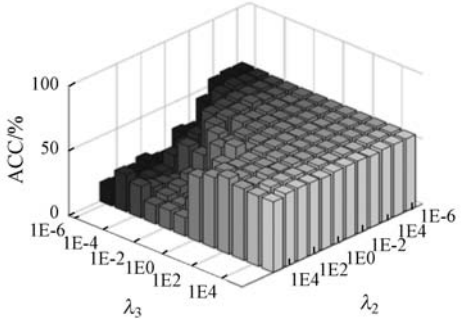


(b)

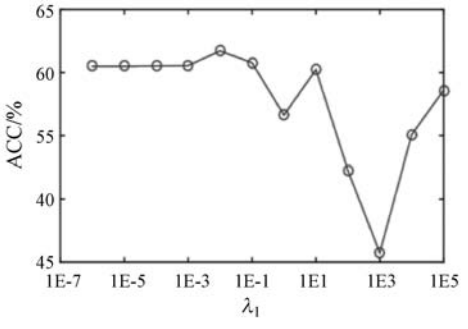


(c)

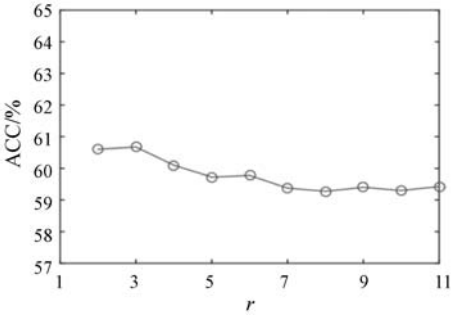
图 2 本文方法在不同参数的手写数据集 ACC



(a)



(b)



(c)

图 3 本文方法在不同参数的 MNIST 数据集 ACC

消融研究:从式(5)中,发现存在三个主要约束,其中第二约束 $\frac{\lambda_2}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\|\boldsymbol{P}_{:,i}^{(v)}-\boldsymbol{P}_{:,j}^{(v)}\|_2^2|Z_{i,j}^{(v)}|$ 和第三约束 $\lambda_3(c-\mathrm{Tr}(\boldsymbol{P}^{(v)\mathrm{T}}\boldsymbol{P}^{(v)}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}))$ 用于学习所有视图的表示和视图共享一致表示。引入第一项 $\lambda_1\|\boldsymbol{S}^{(v)}\|_F^2$ 来避免出现无效解。此外,引入此约束可以使本文的模型获得变量 $\boldsymbol{S}^{(v)}$ 的闭式最优解。为了验证该约束的有效性,在缺失视图或有偿视图率为 30% 的情况下,在表 1 所列的前 5 个数据集上进行了实验,并将本文方法与无此约束的退化模型进行了比较。实验结果如图 4 所示,可以看出,引入第一约束条件 $\lambda_1\|\boldsymbol{S}^{(v)}\|_F^2$ 有利于获得较好的聚类性能。

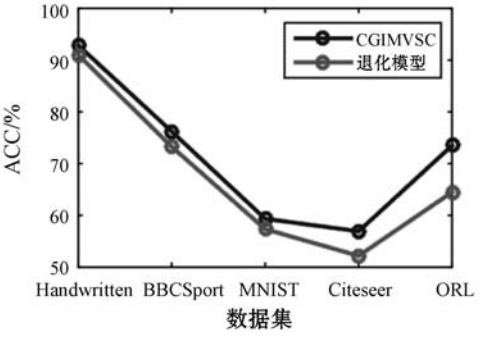
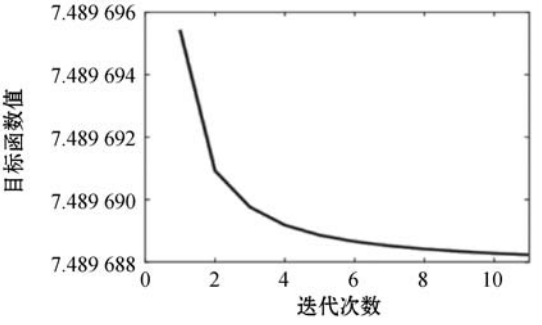


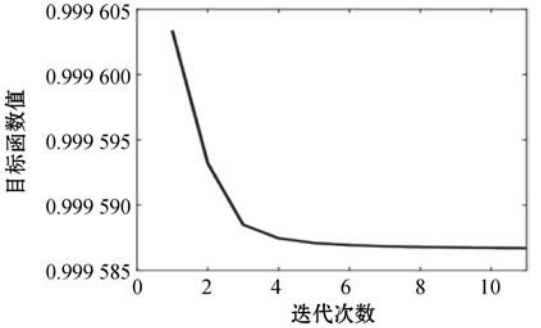
图 4 本文方法和其退化模型的 ACC 比较

3.5 收敛性实验

由图 5 所示的实验结果可以发现,在缺失视图率为 30% 的 BBCSport 和 Handwritten 数据集上,本文模型的目标函数损失呈单调递减趋势。可以看出,该方法在两个数据集上迭代 12 次后收敛速度很快。说明提出的方法具有良好的收敛性能。



(a) BBCSport 数据集



(b) Handwritten 数据集

图 5 目标函数损失与迭代次数的关系

4 结 语

为了解决传统方法存在的聚类效果差和泛化能力弱等问题,本文提出一种基于一致引导的不完全多视图聚类方法。分析和归纳在七个数据集上的实验结果得到如下结论:

(1) 相对于其他方法,本文方法能够完成精度更高的不完整多视图聚类任务。说明本文方法能够有效解决不完整多视图数据的聚类。

(2) 将特征缺失问题转移到图空间或核空间中可以有效改善聚类性能。另外,基于图学习和基于核学习的内模控制方法可以减轻不完整数据中信息不平衡问题造成的负面影响。

(3) 将缺少的实例填充为平均实例或其他向量不利于数据的聚类,尤其对于具有较大丢失率的情况。另外,引入第一约束条件有利于更好的聚类性能。

参 考 文 献

[1] 徐霜,余琍. 利用正则化矩阵分解技术的多视图聚类方法[J]. 计算机工程与应用,2019,55(14):142 - 147,161.

[2] 敬明旻. 基于深度神经网络的多模态特征自适应聚类方法[J]. 计算机应用与软件,2020,37(10):262 - 269.

[3] 朱丹,陈晓红,吴卿源,等. 自适应图学习诱导的子空间聚类[J]. 计算机工程与应用,2020,56(21):30 - 37.

[4] 郭圣,仲兆满,李存华. 基于深度自编码的多视图子空间聚类网络[J]. 计算机工程与应用,2020,56(17):60 - 68.

[5] Hu M, Chen S. Doubly aligned incomplete multi-view clustering[C]//27th International Joint Conference on Artificial Intelligence,2019:2262 - 2268.

[6] Huang Z, Zhou J, Peng X, et al. Multi-view spectral clustering network[C]//28th International Joint Conference on Artificial Intelligence,2019:2563 - 2569.

[7] Kang Z, Pan H, Hoi S, et al. Robust graph learning from noisy data[J]. IEEE Transactions on Cybernetics,2020,50(5):1833 - 1843.

[8] Li J, Li Z, Lu G, et al. Asymmetric gaussian process multi-view learning for visual classification[J]. Information Fusion,2020,65:108 - 118.

[9] Koohi H, Kiani K. User based collaborative filtering using fuzzy C-means[J]. Measurement,2018,91:134 - 139.

[10] Li X, Chen M, Wang Q. Adaptive consistency propagation method for graph clustering[J]. IEEE Transactions on Knowledge and Data Engineering,2020,32(4):797 - 802.

[11] Tian C, Xu Y, Zuo W. Image denoising using deep CNN with batch re-normalization[J]. Neural Networks,2020,121(6):461 - 473.

[12] 邓强,杨燕,王浩. 一种改进的多视图聚类集成算法[J]. 计算机科学,2017,44(1):65 - 70.

[13] Wen J, Zhang B, Xu Y. Adaptive weighted non-negative low-rank representation[J]. Pattern Recognition,2018,81:326 - 340.

[14] Wang Q, Chen M, Nie F. Detecting coherent groups in crowd scenes by multiview clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(1):46 - 58.

[15] 宋悦. 不完整多视觉数据聚类分析[D]. 西安:西安电子科技大学,2019.

[16] 姬名书. 基于稀疏嵌入框架的不完全多视图聚类[D]. 南昌:南昌大学,2019.

[17] Yang Y, Wang H. Multi-view clustering: A survey[J]. Big Data Mining and Analytics,2018,1(2):83 - 107.

[18] Zhan K, Niu C, Chen C. Graph structure fusion for multiview clustering[J]. IEEE Transactions on Knowledge and Data Engineering,2018,31(10):1984 - 1993.

[19] Zhang C, Fu H, Wang J. Tensorized multi-view subspace representation learning[J]. International Journal of Computer Vision,2020,128:2344 - 2361.

(上接第 157 页)

[12] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger[C]//Computer Vision and Pattern Recognition,2017:6517 - 6525.

[13] Redmon J, Farhadi A. YOLOv3: An incremental improvement[EB]. arXiv:1804.02767,2018.

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large scale image recognition[EB]. arXiv:1409.1556,2015.

[15] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence,2017:2999 - 3007.

(上接第 188 页)

[12] 王顶,吴玥瑶,曹旺辉,等. 基于 Dice 匹配的 SWOMP 压缩感知重构算法[J]. 西北工业大学学报,2017,35(5):774 - 779.

[13] Zhang Y Y, Sun G L. Stagewise arithmetic orthogonal matching pursuit[J]. International Journal of Wireless Information Networks,2018, 25(2): 221 - 228.

[14] 江晓林,唐征宇,渠苏苏. 基于 SWOMP 分段回溯的压缩感知改进算法[J]. 黑龙江科技大学学报,2019,29(4):501 - 505.

[15] Zhao S, Zhang Q Y, Yang H. Orthogonal matching pursuit based on tree-structure redundant dictionary[J]. Advances in Computer Science, Environment, Ecoinformatics, and Education, 2011, 215: 310 - 315.

[16] Needell D, Tropp J A. Cosamp: Iterative signal recovery from incomplete and inaccurate samples[J]. Applied and Computational Harmonic Analysis,2009,26(3):301 - 321.

[17] Song L, Yan R. Bearing fault diagnosis based on cluster-contraction stage-wise orthogonal-matching-pursuit[J]. Measurement, 2019, 140: 240 - 253.