软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(4):1354–1372 [doi: 10.13328/j.cnki.jos.006471]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于多阶近邻融合的不完整多视图聚类算法*

刘晓琳[1,2], 白 亮[1,2], 赵兴旺[1,2], 梁吉业[1,2]

[1](山西大学 计算机与信息技术学院, 山西 太原 030006)

[2](计算智能与中文信息处理教育部重点实验室(山西大学), 山西 太原 030006)

通信作者: 梁吉业, E-mail: ljy@sxu.edu.cn

摘 要: 在实际应用中, 聚类多视图数据是一项重要的数据挖掘任务. 样本缺失所导致的多视图不完整给聚类任务带来了巨大的挑战. 大部分已有的不完整多视图聚类方法主要基于浅层图结构信息, 易受到噪声及缺失数据的影响, 且难以准确刻画并兼容所有视图的潜在结构, 从而降低了聚类性能. 为此, 提出了一种更为鲁棒和灵活的基于多阶近邻扩散融合的不完整多视图聚类算法. 该算法在利用多阶相似性学习不完整视图潜在结构的基础上, 通过跨视图交叉扩散的方式, 将不同阶的深层结构信息进行非线性融合, 以此挖掘视图间更全面的结构信息,从而降低了缺失样本所导致的视图结构不确定性. 进一步证明了所提算法的收敛性. 实验结果表明, 相比已有方法, 所提出的算法在处理不完整多视图聚类问题上是更加有效的.

关键词: 不完整多视图聚类; 结构信息; 多阶近邻; 交叉扩散; 非线性融合

中图法分类号: TP311

中文引用格式: 刘晓琳, 白亮, 赵兴旺, 梁吉业. 基于多阶近邻融合的不完整多视图聚类算法. 软件学报, 2022, 33(4): 1354–1372. http://www.jos.org.cn/1000-9825/6471.htm

英文引用格式: Liu XL, Bai L, Zhao XW, Liang JY. Incomplete Multi-view Clustering Algorithm Based on Multi-order Neighborhood Fusion. Ruan Jian Xue Bao/Journal of Software, 2022, 33(4): 1354–1372 (in Chinese). http://www.jos.org.cn/1000-9825/6471.htm

Incomplete Multi-view Clustering Algorithm Based on Multi-order Neighborhood Fusion

LIU Xiao-Lin[1,2], BAI Liang[1,2], ZHAO Xing-Wang[1,2], LIANG Ji-Ye[1,2]

[1](School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

[2](Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education (Shanxi University), Taiyuan 030006, China)

Abstract: In real applications, it is an important field for clustering the multi-view data in data mining. The incompleteness of multi- view caused by missing samples brings great challenge to multi-view clustering task. The shallow graph structure information is easily affected by noise and missing data. Most of the existing multi-view clustering methods are difficult to describe the underlying structure of all views accurately and comprehensively, which reduces the performance of incomplete multi-view clustering. To this end, this study proposes a robust incomplete multi-view clustering algorithm based on the strategies of diffusing and fusing among multi-order neighborhoods. Firstly, the proposed algorithm obtains the potential structural information from incomplete views by utilizing multi-order similarities. Then, the deep structural information of multi-views is nonlinearly fused by the way of cross-view diffusion. Through all above, the much more comprehensive structural information among views can be extracted from the proposed algorithm, thereby reducing the uncertainty of views-structure caused by missing samples. In addition, this paper presents detailed steps to prove the convergence of the proposed algorithm. Experimental results show that the proposed method is more effective in solving the problem of incomplete multi-view clustering than other existing methods.

* 基金项目: 国家重点研发计划(2020AAA0106100); 国家自然科学基金(62022052, 62072293)
本文由“面向开放场景的鲁棒机器学习”专刊特约编辑陈恩红教授、李宇峰副教授、邹权教授推荐.
收稿时间: 2021-05-19; 修改时间: 2021-07-16; 采用时间: 2021-08-27; jos 在线出版时间: 2021-10-26

Key words: incomplete multi-view clustering; structure information; multi-order neighborhood; cross diffusion; nonlinear fusion

聚类是机器学习、模式识别领域最重要且最基础的一种数据分析方法. 聚类旨在根据数据间的相似性准则将无标记的数据集合划分为若干有意义的子集, 使得同一子集中的样本相似性较大,不同子集中的样本相异性较大, 以此挖掘出数据中潜在的类簇结构信息. 聚类分析作为一种典型的无监督学习方法, 由于其具有无需标注训练样本的优点, 在过去的几十年里, 已被广泛应用于众多实际场景并取得了长足的发展. 然而, 随着信息技术的发展, 数据的采集方式趋于多样, 数据逐渐呈现出多元化的特点. 数据可以来源于不同的表示空间, 或通过不同的特征采集器提取得到. 我们称这种不同来源或不同模态的数据为多视图数据[1,2]. 目前, 已经有大量先进的聚类算法被提出, 但是传统的聚类算法[3−5]只能处理单一特征空间下的数据集合, 而无法直接处理多视图数据. 一种简单的操作就是将多视图数据进行特征拼接, 将多视图聚类问题转化为单视图聚类问题. 然而, 这种简单的特征拼接不仅使样本的特征维度急剧增加, 引发更加严峻的“维度灾难”问题, 而且拼接后的特征在物理意义上也缺乏合理的解释, 继而导致样本“距离失效”的问题. 此外, 这种简单的拼接操作忽略了视图之间的内在关联以及每个视图所包含的多样性信息, 导致聚类效果不佳.

多视图聚类[6,7]作为一种新的机器学习范式, 近年来受到了科研学者和工业技术人员的广泛关注. 多视图聚类旨在通过联合学习多个视图的特征信息, 将视图中相似的样本划分到同一个类簇, 将视图中不相似的样本划分至不同类簇, 并且要求多个视图之间具有一致的划分结果. 目前已经有大量的多视图聚类算法被提出, 其中: 基于协同学习的方法[8−11]主要利用不同视图的先验信息或学习得到的知识交互式地指导其他视图的学习, 强调视图之间的协作作用, 算法通过不同视图之间的互补信息得到所有视图一致的划分结果. 基于矩阵分解的方法[12−15]假设多个视图之间共享某些潜在的空间结构, 通过挖掘多视图的相似结构, 学习不同视图之间的共识信息, 以此达到视图融合的目的. 从度量学习的角度来看, 核方法是一种度量方法, 通过度量数据间的相似性, 隐式地考虑了特征之间的关系. 基于多核的方法[16−18]假设每个视图存在一种相似性度量, 不同的相似性度量对应于不同的核空间, 因此多个视图必然存在多个核. 从多个核中学习一个组合最优的核, 是该类算法的核心, 多核聚类算法可以在最优的核上进行后续的聚类操作来完成多视图聚类任务. 基于图学习的方法[19−21]主要从多个视图间构建不同的亲和图, 挖掘多个视图之间的结构关联, 以此来学习一个兼容图, 最后将多视图聚类问题转换成图分割问题.

传统的多视图聚类算法通常要求数据在所有视图下是样本完整的. 由于实际中存在诸多风险因素, 这样的假设对于现实开放环境下的数据采集工作来说, 要求是相对苛刻的. 完整视图的设置很可能并不成立, 因为并不是所有的视图都能被完整地观测到, 即每个样本不一定在所有视图上均被采集并且一一对应[22−24]. 例如: 在同一场景下的多个视频监控中, 某些摄像头可能由于设备故障等原因无法正常工作, 从而采集不到某个角度的视频画面, 造成该故障设备视图下的样本缺失; 在医疗诊断过程中, 通过血液检测和磁共振扫描得到的检查结果可以被认为是诊断疾病的两个重要视图, 通常情况下, 某些患者由于高昂的检查费用或自身的一些原因只参加两种测试中的一种, 从而造成检查结果视图下的患者样本部分非对齐的现象. 以上情况都导致了多视图数据的不完整性. 由于视图中部分样本的缺失或非对齐, 使得传统的多视图聚类算法基本失效, 不能获得满意的聚类性能. 如何充分利用不同视图间的一致性和互补性信息, 减少视图中缺失样本的影响, 是不完整多视图聚类中最具挑战性的问题.

在传统多视图聚类算法的基础上, 目前已经有大量的工作被提出拟解决不完整多视图聚类问题. 然而多数算法仅利用数据的浅层图结构信息, 线性融合后的图难以准确刻画并兼容所有视图的潜在结构, 从而降低不完整多视图数据的聚类性能[25]. 为此, 本文提出了一种更为鲁棒且有效的基于多阶近邻扩散融合的不完整多视图聚类算法(incomplete multi-view clustering algorithm based on multi-order neighborhood diffusion and fusion, MNDF). 具体来说, MNDF 在利用多阶相似性学习不完整视图潜在结构信息的基础上, 通过跨视图交叉扩散的方式将不同阶的深层结构信息进行交互与融合. 算法所采用的交叉融合方式实际上是一种非线性融合方法. 该方法基于信息传递理论动态更新每个图, 并使得迭代后的图与其他参与交叉扩散过程的图相似性

逐渐增大. 另外, MNDF 充分利用每个图自身的局部结构信息为别的图更新提供指导, 最终获得的一致性收敛图充分保留每个图的有用信息, 并兼容其他视图的结构信息. 此外, 算法的整个过程无需为每个图分配特定的权重, 这样也避免了优化权重的问题. 总而言之, MNDF 不仅利用了视图内不同阶的潜在结构信息, 还利用了视图间的互补结构信息, 多角度、多层次、多粒度、多维度的构造近似完整图, 从而挖掘不完整视图间更全面的结构信息, 降低了缺失样本导致的视图结构不确定性. 此外, 本文不仅证明了所提算法的收敛性, 还在合成数据集和真实数据集上进行了大量的实验, 实验结果表明: 与其他经典的不完整多视图聚类算法相比, MNDF 在解决不完整多视图聚类问题上是更加有效的.

本文第 1 节简要介绍不完整多视图聚类相关工作. 第 2 节提出基于多阶近邻扩散融合的不完整多视图聚类算法. 第 3 节对所提的算法进行实验验证和分析. 最后, 第 4 节对本文的工作进行总结与展望.

1 相关工作

从对缺失数据的处理方式上来看, 现有的不完整多视图聚类算法一般可以分为以下 3 类: (1) 基于数据补全的方法; (2) 利用对齐信息的方法; (3) 基于近似完整图学习的方法. 具体方法介绍如下.

- 基于数据补全的方法

对不完整多视图聚类问题而言, 一个很自然的想法是, 通过补全不完整的数据来重用现有的基于完整性假设的多视图聚类算法. 已知多个视图是对同一对象的不同特征描述, 所以视图之间必然存在着一定的相关性. 从如何利用这种相关性进行数据补全的角度来看, 又能将基于数据补全的方法分为以下两类.

1) 直接对缺失的数据矩阵进行填充. Trivedi 等人[26]提出一种利用视图相关性补全核矩阵的算法, 该算法利用 CCA 最小化视图之间的不一致性将缺失的视图进行补全, 填充之后的多个视图就可以用已有的多视图聚类算法进行求解. 该算法最大的局限在于, 假设视图中至少有一个视图是完整的. Shao 等人[27]提出了一种基于协同补全的多视图聚类算法, 该算法假设两个视图都不完整, 首先将两个视图相互协同补全, 并将最小化两个视图的不一致性作为优化目标, 通过交替迭代使得两个视图达到较好的收敛结果, 最后采用现有的多视图聚类算法进行后续的聚类操作. 此外, 作为一个经典的不完整多视图聚类算法(multiple incomplete views clustering, MIC)[28], Shao 等人利用每个视图的数据均值去补全各视图中的缺失样本, 再利用已有的基于 Multi-NMF 的多视图聚类算法进行求解;

2) 在多核学习框架下, 一些不完整多视图聚类算法将缺失的部分作为需要优化的变量进行学习, 通过填充和聚类的交替迭代来达到一个较好的聚类结果. Liu 等利用多核 K 均值聚类作为基础算法, 将聚类损失作为优化目标, 把缺失的部分作为需要优化的变量进行学习, 提出了一系列基于多核框架下的不完整多视图聚类算法[29−32].

然而, 基于信息补全的方法中, 数据补全的质量会直接影响最终的聚类结果, 现有的方法并不能有效地弥补和挖掘缺失视图的潜在结构信息, 从而得不到可靠的聚类结果.

- 利用对齐信息的方法

假设多个视图之间存在样本部分非对齐的情况, 那么视图之间必然存在一些可用的对齐样本信息, 非对齐的样本信息便可利用这些对齐样本的信息来进行刻画. 因此, 该类方法一般将不完整视图的对齐部分和非对齐部分划分开来分而治之: 对于对齐部分, 可以将其视为一个完整视图的聚类问题, 用已有的多视图学习算法学习对齐样本的低维表示; 对于视图的非对齐部分, 可以通过相应的投影算法将视图特有的样本映射到与对齐部分一致的空间中, 最后, 算法得到全部样本在同一空间中的潜在表示, 继而对所有样本进行后续聚类操作. 基于这一思想, Li 等人[33]于 2014 年首次基于 NMF 提出了部分对齐的多视图聚类算法(partial multi-view clustering, PVC). Xu 等人[34]通过自表示子空间聚类方法对 PVC 的结果进行后续聚类操作, 该方法可以看作是对 PVC 联合表示的改进方法. Zhao 等人[35]提出了不完整的多模态视觉数据聚类算法(incomplete multi-modal visual data grouping, IMG), 该方法通过求解投影子空间的同时约束其保留局部流形结构, 取得了较好的聚类结果. 在此之后, Qian 等人[36]也通过添加流形约束对 PVC 进行改进, 从而对不完整的多视图数据进行

聚类. 综上,基于对齐信息的不完整多视图聚类算法最大的局限性在于缺失视图中必须存在对齐部分作为聚类指导, 因此不能处理视图任意缺失的情况. 此外, 这种方法由于非对齐样本的存在, 很难挖掘不完整视图的全局结构信息, 强加结构约束也不一定能获得令人信服的聚类结果.

- 基于近似完整图学习的方法

这类方法旨在寻找一个能够兼容所有不完整视图的近似完整图. 换句话说, 这种方法旨在学习一个能够刻画所有不完整视图的几何结构并描述样本之间相互关系的相似性矩阵, 继而将不完整多视图聚类问题转化为图学习及图分割的问题. 2018 年, Yu 等人[37]利用自适应加权的相似性补全算法, 通过不同视图间相似性传递的方法, 进而学习所有样本的相似性矩阵. 次年, Yang 等人[38]正式提出了基于近似完整图学习的概念, 所提的算法(graph-based incomplete multi-view clustering, GIMC)有效地为每个视图学习一个近似完整图, 并自动对每个视图构造的图进行加权学习一致图. Guo 等人[39]提出的不完整多视图聚类算法(anchor-based partial multi-view clustering, APMC)利用锚点来重构样本-样本之间的相似性关系用于聚类任务. Hu 等人[40]通过一种基于样本级的自适应图融合过程学习到一个潜在的子空间, 该子空间能够整合来自多个视图的一致信息, 可以有效提高不完整多视图数据的聚类性能. Hou 等人[41]借鉴了基于对齐信息方法的思想, 将数据划分为对齐部分和非对齐部分, 通过结合二者的局部结构来挖掘一个理想的结构化图, 然后将所有的样本按照学习到的图结构进行图划分, 完成聚类任务. 除此之外, Wen 等人还提出了一系列基于近似完整图学习的方法, 如基于自适应图学习的不完整多视图聚类算法[42]、基于结构保持和表示学习的图正则化不完整多视图聚类算法[43]、基于一致张量框架下的缺失视图推断的不完整多视图聚类算法[44]、基于一致仿射图学习并结合谱聚类和协同正则化的不完整多视图聚类算法[45]等. 2021 年, Liu 等人[46]提出的算法(consensus learning approach to incomplete multi-view clustering, CLIMC)利用原始数据表示的互补信息和协同信息, 联合学习多个视图的一致表示和近似完整相似性图. 同年, Li 等人[47]提出的算法(joint partition and graph, JPG)旨在迭代构造局部不完整图矩阵, 随后生成不完整基划分矩阵并拉伸生成统一的划分矩阵, 利用其学习一致图矩阵.基于近似完整图学习的方法因其简单、高效, 近年来得到了国内外学者的广泛关注, 是处理不完整多视图聚类问题的主流方法之一. 然而, 现有的基于近似完整图学习的方法大都使用原始数据构图, 或者从多个不完整视图中直接学习共享的完整图用于聚类. 用这种方法学习到的图鲁棒性差且容易受到不同视图的异构信息及噪声的影响, 导致所获得的图难以准确刻画所有视图的潜在结构, 从而降低聚类性能.

本文提出的 MNDF 算法是一种基于近似完整图学习的方法, 以克服上述方法存在的不足. 该方法不仅考虑了数据的多阶结构信息, 还充分考虑了视图之间的互补信息, 并通过交叉扩散过程将视图的深层结构信息和多样性信息进行非线性融合. 因而, 该算法可以挖掘不完整视图间更全面的结构信息, 以此确保算法在应对现实世界复杂多变的多视图数据缺失情况时具有更好的聚类性能, 亦增强算法的鲁棒性和适应性.

2 基于多阶近邻扩散融合的不完整多视图聚类算法

2.1 不完整多视图聚类问题描述

给定一个包含 n 个样本、m 个视图的数据集 $\{\boldsymbol{X}^v\}_{v=1}^m$, 假设每个视图中都存在随机缺失的样本, 那么 $\boldsymbol{X}^v=[\boldsymbol{x}_1^v,\boldsymbol{x}_2^v,...,\boldsymbol{x}_{n_v}^v]\in\mathbb{R}^{n_v\times d_v}(n_v<n)$ 表示第 v 个视图中包含 n_v 个维度为 d_v 的可见样本, 其中, 第 $i(1\leqslant i\leqslant n_v)$个样本表示为 $\boldsymbol{x}_i^v=[x_{i1}^v,x_{i2}^v,...,x_{id_v}^v]$. 为了便于从原始数据集中识别每个视图缺失的样本, 本文引入一个指示矩阵 $\boldsymbol{M}^v\in\mathbb{R}^{n_v\times n}$, 该矩阵的每个元素 m_{ij}^v 表示为

$$m_{ij}^v=\begin{cases}1, & \text{if } \boldsymbol{x}_i^v \text{ corresponds to the } j\text{-th original instance}\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

与大多数现有方法一样, 本文假设每个不完整视图所包含样本的个数 n_v 少于原始数据集样本数量 n, 但是每个样本至少存在于一个视图中. 不完整多视图聚类的目标是: 将原始的 n 个样本划分为 c 个类簇, 并且要求多个视图之间具有一致的划分结果.

本文对所提算法中使用的重要符号进行了总结, 见表 1.

表 1 MNDF 算法中使用符号汇总

符号表示	符号说明
$\boldsymbol{X}^v \in \mathbb{R}^{n_v \times d_v}$	第 v 个视图的数据集合
$\boldsymbol{M}^v \in \mathbb{R}^{n_v \times n}$	第 v 个视图的样本指示矩阵
$\boldsymbol{W}(o)^v \in \mathbb{R}^{n_v \times n_v}$	第 v 个视图的第 o 阶近邻矩阵
$\boldsymbol{A}(o) \in \mathbb{R}^{n \times n}$	第 o 阶多个视图的互补近邻图
$\bar{\boldsymbol{A}}(o) \in \mathbb{R}^{n \times n}$	第 o 阶归一化互补近邻图
$\bar{\boldsymbol{A}}(o)^{(t)} \in \mathbb{R}^{n \times n}$	第 o 阶第 t 次迭代的状态矩阵
$\boldsymbol{A}^* \in \mathbb{R}^{n \times n}$	多图融合后的最优图
n	所有样本个数
n_v	第 v 个视图的可见样本个数
m	视图个数
c	类簇个数
$\|O\|$	图的阶数
t	算法迭代次数
α	正则化参数
k	m-kNN 近邻参数

2.2 基于多阶近邻扩散融合的不完整多视图聚类算法

目前, 大部分不完整多视图聚类算法主要去解决如何利用视图之间的互补性去构建一个完整的数据结构图问题. 然而, 已有方法仅仅利用数据不同视图下的单阶结构信息, 忽略了数据的高阶结构信息. 为解决这一问题, 本文提出了基于多阶近邻扩散融合的不完整多视图聚类算. 该算法将回答一个关键问题: 如何充分提取并融合多阶结构信息代替单阶信息构造近似完整图. 所提算法主要由两个核心步骤构成: (1) 构造初始多阶近邻图; (2) 非线性融合多阶近邻图. 通过这两个步骤, 算法可以从多个不完整视图不同层次的结构中获得互补信息, 并以交叉扩散的方式进行协同学习, 以此达到视图深层结构信息和多样性信息融合的目的, 学习到一个鲁棒性更强、结构信息更丰富的近似完整图.

2.2.1 构造初始多阶近邻图

由于图中所包含的结构信息对基于近似完整图学习的算法来说是至关重要的, 现有的学习方法大都从原始数据构图, 即直接计算数据集中两个样本间的相似性, 学习到的图鲁棒性差, 易受到噪声和缺失样本的影响, 导致所获得的图难以准确刻画数据的结构信息, 从而降低聚类性能. 在相似性计算过程中, 多阶相似性信息可以为数据提供不同层次的结构描述信息, 融合这些不同阶的相似性关系, 能够为聚类算法提供更加明确的聚类指导. 因此, 利用不完整多视图数据的底层结构信息来提高聚类性能是非常有必要的.

如图 1(a)所示, 样本间的一阶结构信息通过数据的原始特征信息计算得出, 捕获的是原始数据的局部连接关系, 即样本 B 分别与样本 A 和样本 C 之间存在连边. 高阶相似性刻画的是一种更深层次的近邻关系, 即共享邻居越多的数据点越有可能相似. 如图 1(b)所示, 样本 A 和样本 C 存在共享的邻居节点 B, 那么就可以利用数据间的高阶结构信息对样本 A 和样本 C 之间的潜在连边进行刻画, 以此挖掘出样本间更加全面的潜在结构信息.

(a) 一阶结构信息 (b) 二阶结构信息

图 1 举例说明样本间的多阶结构信息

在复杂的样本关系中, 多阶相似性对数据潜在结构的挖掘更为重要. 如图 2 所示, 样本 4 和样本 5 在一阶结构中不存在连接关系, 即它们在一阶相似性的定义中具有较低的相似度. 而样本 4 与样本 6 在一阶结构中

具有较高的相似度. 因此, 以一阶相似性进行数据划分时, 样本 4 与样本 6 会大概率同属于一类. 然而, 以二阶相似性计算样本相似度时, 样本 4 与样本 5 都和样本 1、样本 2、样本 3 相连, 即它们共享相同的邻域连接结构. 因此, 以二阶相似性进行数据划分时, 样本 4 与样本 5 应该同属于一类, 这也更加拟合数据真实的潜在结构. 在多视图数据中, 样本间的连接关系愈加复杂. 低阶结构信息侧重描述数据的局部成对关系, 这种局部结构信息对于刻画整体数据的潜在结构来说是片面且不够准确的. 高阶结构信息由于度量的是样本间共享相连节点的深层相似性关系, 获得的是一种更加全面的邻域结构信息, 因此在刻画多视图数据潜在结构方面是更加准确的.

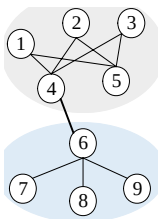


图 2 举例说明高阶结构信息的重要性

本文采用多阶相似性来挖掘多视图数据的潜在结构信息. 以第 v 个视图为例, 定义 $\boldsymbol{W}(o)^v \in \mathbb{R}^{n_v \times n_v}$ 为样本对之间的第 o 阶近邻矩阵, 其中, 矩阵的每个元素计算方法为

$$w(o)_{ij}^v = \begin{cases} \exp\left(-\dfrac{d(o)_{ij}^v}{(\sigma^v)^2}\right), & \text{if } i\text{-th and } j\text{-th data are } m\text{-}k\text{NN} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

其中,

$$d(o)_{ij}^v = \begin{cases} \| \boldsymbol{x}_i^v - \boldsymbol{x}_j^v \|_2^2, & o = 1 \\ \| \boldsymbol{w}(o-1)_i^v - \boldsymbol{w}(o-1)_j^v \|_2^2, & o > 1 \end{cases} \tag{3}$$

由于视图存在样本缺失的情况, 因此多个不完整视图的近邻图也存在缺失顶点和缺失边的情况. 基于视图之间的互补性, 假设每个视图的数据样本信息是缺失的, 但是多个视图的结构信息是互补且完备的. 为此, 利用每个视图的指示矩阵 $\boldsymbol{M}^v$, 对同阶的多个不完整视图进行互补性对齐融合, 旨在获得视图间同阶的互补近邻图. 定义 $\boldsymbol{A}(o) \in \mathbb{R}^{n \times n}$ 为第 o 阶的互补近邻图, 其计算公式为

$$\boldsymbol{A}(o) = \frac{1}{m} \sum_{v=1}^{m} (\boldsymbol{M}^v)^T \boldsymbol{W}(o)^v \boldsymbol{M}^v \tag{4}$$

基于每个样本至少存在于一个视图的基本假设, 所获得的初始互补近邻图将是包含 n 个顶点的近似完整图. 由公式(4)可知: 若多个视图存在一致连边, 则互补近邻图中边的权重越大. 虽然 MNDF 算法在第 1 步获得了多个初始近似完整图, 包含了多个视图不同阶的结构信息, 但是如何有效地融合这些近邻图, 也是本文算法所考虑的关键.

2.2.2 非线性融合多个互补近邻图

线性图融合方法本质上是一种组合优化问题, 图中的强弱连接通过加权的方式进行融合. 这种方式将弱化连边信息的表达能力, 从而降低聚类性能. 为了克服线性融合的缺点, 基于信息传递理论[48], 本文提出一种针对多图的非线性图融合方法, 即 MNDF 算法. 该方法以一种扩散的方式对第 2.2.1 节中获得的初始多阶完整图(即互补近邻图 $\boldsymbol{A}(o)$, 其中, o=1,2,…,$|O|$)的边缘信息进行相互交换, 从而增强图中的强连接并减弱图中的弱连接, 交叉扩散过程将不断地重复迭代, 直到最终的统一图收敛.

为使不同的图保持尺度的一致性, MNDF 首先需要对图数据进行归一化操作. 以第 o 阶互补近邻图 $\boldsymbol{A}(o)$ 可以以一个简单的方法对其进行归一化, 即 $\bar{\boldsymbol{A}}(o) = \boldsymbol{D}(o)^{-1}\boldsymbol{A}(o)$, 其中, $\boldsymbol{D}(o)$ 为度矩阵, 对角线元素为 $d(o)_{ii} = \sum_{j=1}^{n} a(o)_{ij}$. 虽然这种方式保证了 $\sum_{j=1}^{n} \bar{a}(o)_{ij} = 1$, 但是理论证明: 这种归一化方式由于使用了对角项

的自相似性, 在数值上是不稳定的. 因此, MNDF 采用一种新的方式对互补近邻图的矩阵进行归一化得到 $\bar{\boldsymbol{A}}(o)$, 矩阵元素计算公式为

$$\bar{a}(o)_{ij}=\begin{cases}\dfrac{a(o)_{ij}}{2\sum_{j=1}^{n}a(o)_{ij}}, & i\neq j\\ \dfrac{1}{2}, & i=j\end{cases} \tag{5}$$

这种归一化方式将每一行的自相似性设置为 1/2, 其余元素总和为 1/2, 每一行的元素总和仍然为 1. 这种标准化方法能够避免相似性矩阵对角元素的自相似性, 从而使得数值更加稳定. 归一化后的 $\bar{\boldsymbol{A}}(o)$ 矩阵虽然是非对称的, 但是却保留了每个样本和其他样本之间全部的相似性信息. 如何有效地融合这些多阶近邻图, 是 MNDF 算法的核心步骤. 受单视图下正则化图扩散过程(regularized diffusion process, RDP)[49]的启发, 本文将提出一种适用于多图非线性融合的交叉扩散过程.

本质上来说, RDP 算法对交叉扩散过程进行了流形解释, 即利用平行四边形法则进行相似性传播, 所获得的相似性度量(即状态矩阵 A)为以下优化问题的闭式解:

$$\min_{\boldsymbol{A}}\frac{1}{2}\sum_{i,j,p,q=1}^{n}w_{ij}w_{pq}\left(\frac{a_{ip}}{\sqrt{d_{ii}d_{pp}}}-\frac{a_{jq}}{\sqrt{d_{jj}d_{qq}}}\right)^2+\mu\|\boldsymbol{A}-\boldsymbol{W}\|_F^2 \tag{6}$$

其中, W 为原始数据的初始化相似性矩阵, 对角阵元素 $d_{ii}=\sum_{j=1}^{n}w_{ij}$.

该目标函数的闭式解为

$$\hat{\boldsymbol{A}}^*=(1-\alpha)vec^{-1}((\boldsymbol{I}-\alpha\mathbb{L})^{-1}vec(\boldsymbol{W})) \tag{7}$$

其中, $\alpha=\dfrac{1}{1+\mu}$, $vec(\cdot)$是一个通过将输入矩阵的列依次堆叠来对其进行向量化的运算符, 它对应的逆算子为 $vec(\cdot)^{-1}$. 令 $\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$, 则$\mathbb{L}=\boldsymbol{L}\otimes\boldsymbol{L}$, 其中, $\mathbb{L}\in\mathbb{R}^{n^2\times n^2}$ 是 L 的张量积.

公式(7)的求解过程等价于一种相似性的迭代过程. 优化过程可以改写成如下等式:

$$\boldsymbol{A}^{(t)}=\alpha\boldsymbol{L}\boldsymbol{A}^{(t-1)}\boldsymbol{L}^T+(1-\alpha)\boldsymbol{W} \tag{8}$$

公式(8)仅限于单视图上的相似性传播, 而不能应用于多图传播. 因此, MNDF 算法提出了核心公式(9), 使其适用于$|O|$(经实验表明, $|O|$一般取 3 即可)个多阶图的交叉扩散过程. 算法将归一化互补近邻图作为交叉扩散过程中的初始状态矩阵, 即 $\bar{\boldsymbol{A}}(o)^{(1)}=\bar{\boldsymbol{A}}(o)$, 则第 o 阶的第 t 次迭代更新公式为

$$\bar{\boldsymbol{A}}(o)^{(t)}=\alpha\boldsymbol{L}(o)\left\{\frac{1}{|O|-1}\left[\sum_{l=1,\ l\neq o}^{|O|}\bar{\boldsymbol{A}}(l)^{(t-1)}\right]\right\}(\boldsymbol{L}(o))^T+(1-\alpha)\bar{\boldsymbol{A}}(o)^{(1)} \tag{9}$$

从公式(9)可以看出, 每次迭代都并行地进行$|O|$个相互影响却又相互独立的扩散过程. MNDF 设计的好处在于: 一方面, 通过迭代交换不同图中的连接信息, 在每次迭代中, 不同图矩阵的相似值被传播到其他图矩阵, 这样可以利用多图之间的互补性来改善扩散过程; 另一方面, 通过正则项的控制, 原始图中的结构信息也会被部分保留, 从而约束扩散过程, 避免扩散过程造成较大偏差.

公式(9)的迭代过程会重复进行多次直至收敛, 即不同图所包含的结构信息会为了结果趋于一致而连续进行多次信息交换. 在交叉扩散的过程中, 样本间不同阶的相似性关系将会彼此传播. 图中的强连接会添加到其他图中, 即相似的样本之间相似性会被增强, 而图中的弱连接会断开, 即相异的样本之间的相似性会被减弱, 这样便可以降低融合后图的噪声信息, 使算法更加鲁棒. 进行 t 次交叉扩散之后, 求得最终融合了$|O|$个高阶相似性关系的最优图 $\boldsymbol{A}^*$为

$$\boldsymbol{A}^*=\frac{1}{|O|}\left(\sum_{o=1}^{|O|}\bar{\boldsymbol{A}}(o)^{(t)}\right) \tag{10}$$

通过公式(10)得到的最优图 $\boldsymbol{A}^*$融合了多种不同状态的相似性关系, 将多阶图中的边缘信息进行协同交互, 使得同类样本间的连接更加紧密而不同类样本的连接相对减弱, 增强了图中连边信息的表达能力, 从而提高

后续图划分的准确率.

2.2.3 MNDF 算法框架及算法流程

MNDF 算法框架如图 3 所示. 基于三阶结构信息, 该算法可以从多个不完整视图的结构中获得不同层次的互补信息, 并以交叉扩散的方式进行协同学习, 以此达到视图深层结构信息和多样性信息融合的目的.

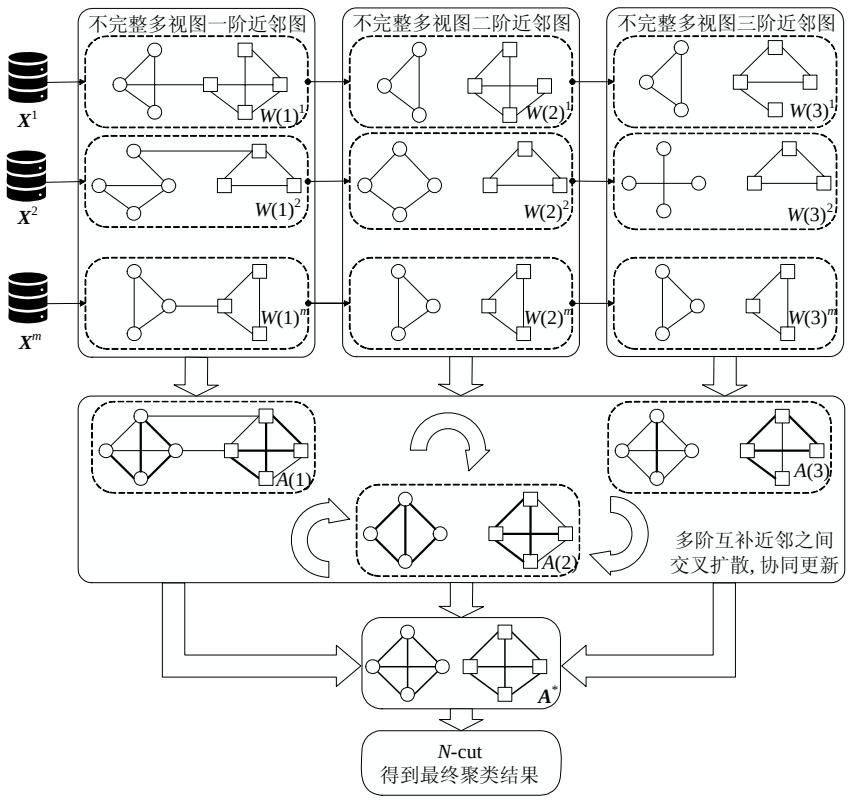


图 3 基于多阶近邻扩散融合的不完整多视图聚类算法框架

MNDF 算法的具体算法过程如下.

Algorithm. 基于多阶近邻扩散融合的不完整多视图聚类算法.

Input: 不完整多视图数据集 $\{\boldsymbol{X}^v\}_{v=1}^m$, 类数 c, 参数 α, 阶数 $|O|$, 迭代次数 t;

1. 根据公式(1)为每个视图构造指示矩阵 $\boldsymbol{M}^v$.
2. 根据公式(2)计算每个不完整视图的多阶近邻矩阵 $\boldsymbol{W}(1)^v,\boldsymbol{W}(2)^v,\ldots,\boldsymbol{W}(|O|)^v$.
3. 根据公式(4)构造视图间同阶的互补近邻图, 如第 o 阶的互补近邻图即 $\boldsymbol{A}(o)$.
4. 根据公式(5)对第 3 步中的互补近邻图进行归一化, 得到 $\bar{\boldsymbol{A}}(o)$.
5. 初始化 $|O|$ 个状态矩阵, 如第 o 阶状态矩阵 $\bar{\boldsymbol{A}}(o)^{(1)}=\bar{\boldsymbol{A}}(o)$.
6. **while** $iter<t$ or not converged **do**
7. **for** o=1 to $|O|$ **do**
8. 根据公式(9)的交叉扩散过程更新每次迭代的状态矩阵 $\bar{\boldsymbol{A}}(o)^{(iter)}$.
9. **end for**
10. **end while**
11. 根据公式(10)得到最终融合后的最优图 $\boldsymbol{A}^*$.

Output: 在最优图 $\boldsymbol{A}^*$ 上进行 N-cut 并得到最终的聚类结果.

2.2.4 MNDF 算法的时间复杂度分析

MNDF 算法的核心在于构造多阶近邻图以及多个图之间的交叉扩散过程. 由于每次迭代都能并行地进行$|O|$个相互影响却又相互独立的过程, 因此构造多阶近邻图的计算代价为$\mathcal{O}(|O|mn_v^2)$, 交叉扩散过程的计算代价为$\mathcal{O}(|O|tn^3)$. 算法的整体时间复杂度为$\mathcal{O}(|O|mn_v^2+|O|tn^3)$, 其中 t 为最终的迭代次数.

2.2.5 MNDF 算法的收敛性理论证明

MNDF 算法的收敛性也可以得到理论保证. 在给出证明之前, 我们先给出两个引理.

引理 1. 给定 $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, $\boldsymbol{B}\in\mathbb{R}^{n\times p}$, $\boldsymbol{C}\in\mathbb{R}^{p\times q}$ 这 3 个矩阵, 则 $vec(\boldsymbol{ABC})=(\boldsymbol{C}^T\otimes\boldsymbol{A})vec(\boldsymbol{B})$.

引理 2. 给定两个方阵 $\boldsymbol{A}\in\mathbb{R}^{n\times n}$, $\boldsymbol{B}\in\mathbb{R}^{n\times n}$, 令 $\lambda_{\boldsymbol{A}}^i(1\leqslant i\leqslant n)$ 为矩阵 A 的特征值, $\lambda_{\boldsymbol{B}}^j(1\leqslant j\leqslant n)$ 为矩阵 B 的特征值, 则 $\boldsymbol{A}\otimes\boldsymbol{B}$ 的特征值为 $\lambda_{\boldsymbol{A}}^i\lambda_{\boldsymbol{B}}^j(1\leqslant i,j\leqslant n)$.

定理 1. 如果 $\bar{\boldsymbol{A}}(o)^{(1)}=\bar{\boldsymbol{A}}(o)$, 则有第 o 阶的第 t 次迭代更新公式:

$$\bar{\boldsymbol{A}}(o)^{(t)}=\alpha\boldsymbol{L}(o)\left\{\frac{1}{|O|-1}\left[\sum_{l=1,l\neq o}^{|O|}\bar{\boldsymbol{A}}(l)^{(t-1)}\right]\right\}(\boldsymbol{L}(o))^T+(1-\alpha)\bar{\boldsymbol{A}}(o)^{(1)} \tag{11}$$

根据引理 1 和引理 2, 可以得到其极限形式. 具体地, 当 $t\to\infty$时:

$$\lim_{t\to\infty}vec(\bar{\boldsymbol{A}}(o)^{(t)})=(1-\alpha)(\boldsymbol{I}-\alpha\mathbb{L}(o))^{-1}vec(\bar{\boldsymbol{A}}(o)) \tag{12}$$

因此, 我们有第 o 阶互补近邻图在交叉扩散过程中的闭式解为

$$\bar{\boldsymbol{A}}(o)^{(*)}=(1-\alpha)vec^{-1}((\boldsymbol{I}-\alpha\mathbb{L}(o))^{-1}vec(\bar{\boldsymbol{A}}(o))) \tag{13}$$

证明: 基于引理 1, 将 $vec(\cdot)$应用于公式(9)的两边. 以第 o 阶归一化互补近邻 $\bar{\boldsymbol{A}}(o)^{(t)}$ 更新公式为例, 可以得出:

$$vec(\bar{\boldsymbol{A}}(o)^{(t)})=\alpha\mathbb{L}(o)vec\left(\frac{1}{|O|-1}\left(\sum_{l=1,l\neq o}^{|O|}\bar{A}(l)^{(t-1)}\right)\right)+(1-\alpha)vec(\bar{\boldsymbol{A}}(o)^{(1)}) \tag{14}$$

假设算法进行了多次迭代, 则:

$$\left.\begin{aligned}&vec(\bar{\boldsymbol{A}}(o)^{(t-1)})=\alpha\mathbb{L}(o)vec\left(\frac{1}{|O|-1}\left(\sum_{l=1,l\neq o}^{|O|}\bar{\boldsymbol{A}}(l)^{(t-2)}\right)\right)+(1-\alpha)vec(\bar{\boldsymbol{A}}(o)^{(1)})\\&\vdots\\&vec(\bar{\boldsymbol{A}}(o)^{(2)})=\alpha\mathbb{L}(o)vec\left(\frac{1}{|O|-1}\left(\sum_{l=1,l\neq o}^{|O|}\bar{\boldsymbol{A}}(l)^{(1)}\right)\right)+(1-\alpha)vec(\bar{\boldsymbol{A}}(o)^{(1)})\\&\bar{\boldsymbol{A}}(o)^{(1)}=\bar{\boldsymbol{A}}(o)\end{aligned}\right\} \tag{15}$$

将公式(15)的迭代公式代入公式(14), 则公式(14)等价于:

$$vec(\bar{\boldsymbol{A}}(o)^{(t)})=(\alpha\mathbb{L}(o))^{t-1}vec\left(\frac{1}{|O|-1}\left(\sum_{l=1,l\neq o}^{|O|}\bar{\boldsymbol{A}}(l)\right)\right)+(1-\alpha)\sum_{k=0}^{t-2}(\alpha\mathbb{L}(o))^k vec(\bar{\boldsymbol{A}}(o)) \tag{16}$$

由于 $\boldsymbol{L}(o)$的谱半径小于等于 1, 根据引理 2, $\mathbb{L}(o)=\boldsymbol{L}(o)\otimes\boldsymbol{L}(o)$的特征值取值范围为[−1,1];

此外, 由于 $0<\alpha<1$, 可以得出:

$$\begin{cases}\lim\limits_{t\to\infty}(\alpha\mathbb{L}(o))^{t-1}vec\left(\dfrac{1}{|O|-1}\left(\sum\limits_{l=1,l\neq o}^{|O|}\bar{\boldsymbol{A}}(l)\right)\right)=0\\\lim\limits_{t\to\infty}\sum\limits_{k=0}^{t-2}(\alpha\mathbb{L}(o))^k vec(\bar{\boldsymbol{A}}(o))=(\boldsymbol{I}-\alpha\mathbb{L}(o))^{-1}vec(\bar{\boldsymbol{A}}(o))\end{cases} \tag{17}$$

因此, 当 $t\to\infty$时:

$$\lim_{t\to\infty}vec(\bar{\boldsymbol{A}}(o)^{(t)})=(1-\alpha)(\boldsymbol{I}-\alpha\mathbb{L}(o))^{-1}vec(\bar{\boldsymbol{A}}(o)) \tag{18}$$

随后, 利用 $vec(\cdot)^{-1}$ 用于公式(18)的两边, 得出第 o 阶互补近邻图在交叉扩散过程中的闭式解为

$$\bar{\boldsymbol{A}}(o)^{*}=(1-\alpha)vec^{-1}((\boldsymbol{I}-\alpha\mathbb{L}(o))^{-1}vec(\bar{\boldsymbol{A}}(o))) \tag{19}$$

定理 1 通过给出互补近邻图的闭式解, 而证明了所提出的 MNDF 算法的收敛性.

3 实验与分析

3.1 人工合成数据集实验结果

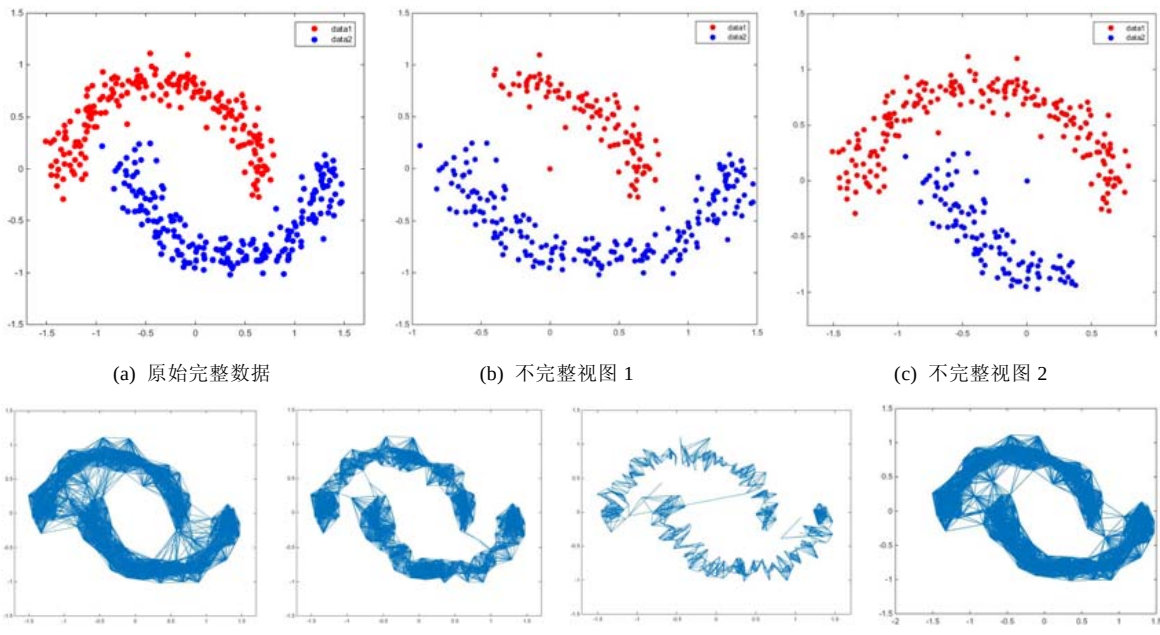
本节以合成数据集 TwoMoon 为例, 设计多种缺失方式, 以更加清晰直观的方式展现 MNDF 算法的实验过程及结果, 以此来验证 MNDF 算法的有效性. 首先, 随机生成一个 TwoMoon 数据集, 该数据集包含 400 个完整样本, 自然划分为 2 类(红色类和蓝色类), 每类包含 200 个可见样本. 本文在所构造的不完整多视图数据上运行 MNDF 算法. 最后, 本文分别展示了不同缺失情况下, 一阶、二阶、三阶互补近邻图和最终扩散融合图的实验结果.

为了模拟现实生活中复杂多变的不完整多视图缺失情况, 本文设置了多种缺失方案, 并在参数为{α=0.2, k=35}的情况下进行实验.

(1) 如图 4(a)–图 4(c)所示, 本文设置从完整视图的两个类中分别非随机性地删除 100 个样本, 以此构造两个不完整的多视图数据;

(2) 如图 5(a)–图 5(c)所示, 本文设置视图的缺失率为 20%, 分别运行两次随机缺失过程, 以此构造两个不完整的多视图数据;

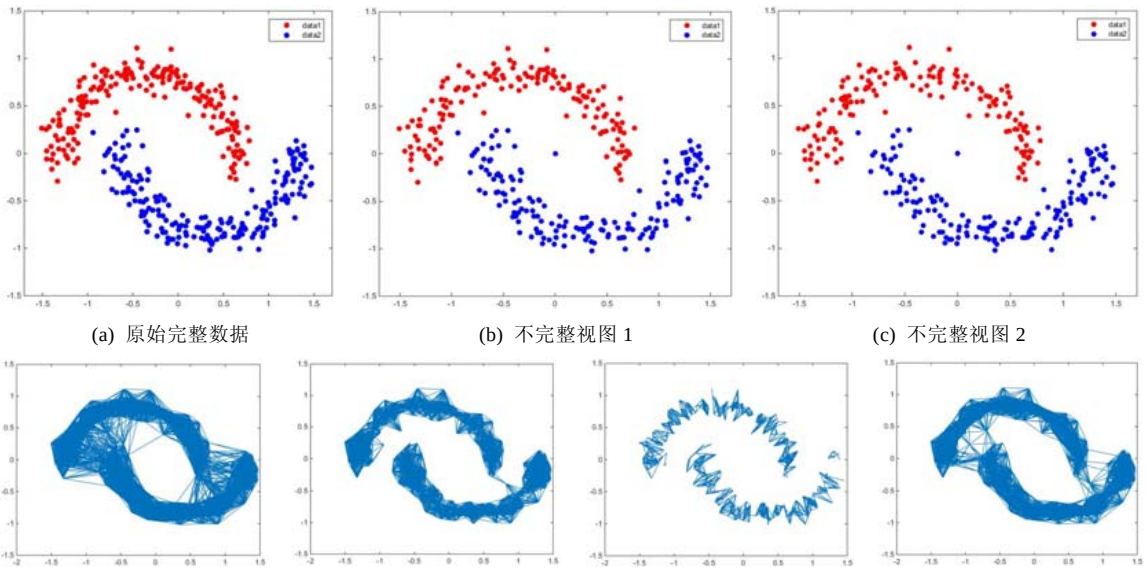
(3) 如图 6(a)–图 6(c)所示, 本文设置数据的缺失率为 40%, 分别运行两次随机缺失过程, 以此构造两个不完整的多视图数据.



(a) 原始完整数据 (b) 不完整视图 1 (c) 不完整视图 2

(d) 一阶互补近邻图 (e) 二阶互补近邻图 (f) 三阶互补近邻图 (g) 最终扩散融合图

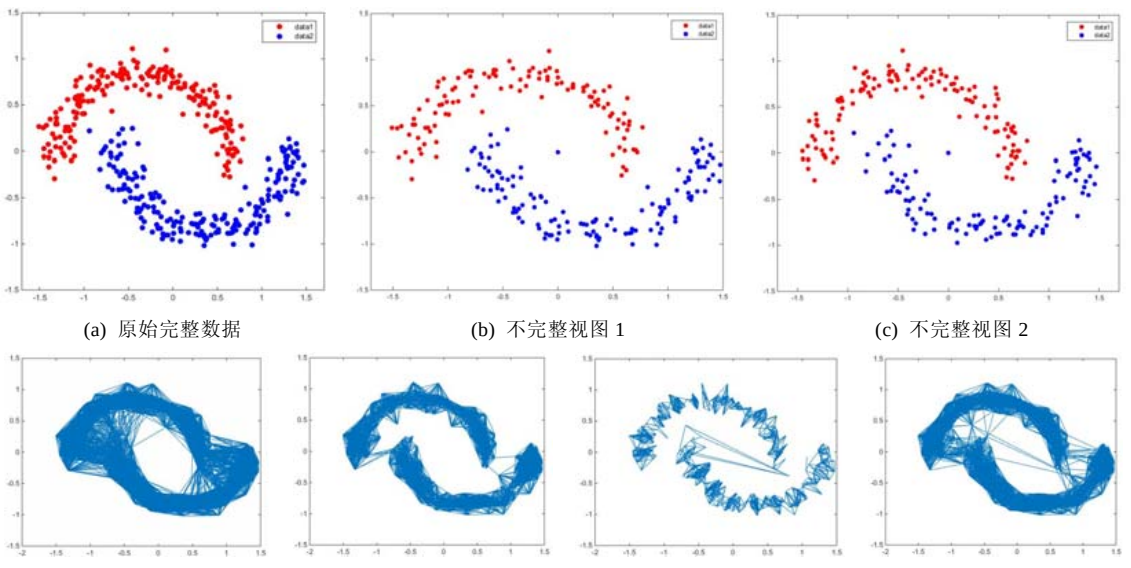
图 4 非随机删除 100 个样本的不完整多视图数据实验结果



(a) 原始完整数据 (b) 不完整视图 1 (c) 不完整视图 2

(d) 一阶互补近邻图 (e) 二阶互补近邻图 (f) 三阶互补近邻图 (g) 最终扩散融合图

图 5 缺失率为 20%的实验结果



(a) 原始完整数据 (b) 不完整视图 1 (c) 不完整视图 2

(d) 一阶互补近邻图 (e) 二阶互补近邻图 (f) 三阶互补近邻图 (g) 最终扩散融合图

图 6 缺失率为 40%的实验结果

由于图中所包含的结构信息对基于图学习的算法来说是至关重要的，而现有的算法大都从原始数据构图即直接计算数据集中两个样本间的相似性，由图 4–图 6 中的一阶互补近邻图所示：浅层图结构信息容易受到噪声和缺失数据的影响，导致所获得的图难以准确刻画数据的结构信息，从而降低聚类性能. 以图 4–图 6 中的二阶、三阶互补近邻图为例，高阶相似性所构的图可以挖掘数据更加潜在可靠的结构信息，并且阶数越高，所挖掘的结构信息越细致. 因此在相似性计算过程中，不同的多阶相似性信息可以给数据提供不同的结构描述信息，融合这些不同阶的相似性关系，能够为聚类算法提供更加明确的聚类指导.

MNDF 算法中，样本对间的相似性在每次交叉扩散过程时都会被传播，图中的强连接会添加到其他图中，

而图中的弱连接会断开, 这样便可以降低融合后图的噪声信息. 此外, 通过这种扩散融合过程, 相似的样本之间相似性被增强, 而相异的样本之间的相似性被减弱, 最终融合后的图也更能准确刻画数据的真实结构信息, 得到的图也更加鲁棒, 从而提高图划分的聚类性能.

3.2 真实多视图数据集

本实验选取 5 个真实多视图数据集作为实验数据, 数据集的详细介绍如下.

(1) HW 数据集: 手写数字数据集, 包含 6 个视图, 每个视图包含 2 000 个来自数字“0–9”的数字图像;

(2) MSRCV1 数据集: 该数据集由来自 210 个图像的 7 个类组成, 分别用 6 种不同的特征提取方法将图像构造成不同的视图;

(3) BBC 数据集: 该数据集由 BBC 网站收集的 685 份新闻文档构成, 每个文档被分为 4 个视图, 且最多包含来自单个文档的一个片段, 所有文档被分为 5 个主题;

(4) NGs 数据集: 该数据集由 500 个新闻文档组成, 每个原始文档都用 3 种不同的特征提取方法进行预处理, 并划分为 5 个主题;

(5) 3sources 数据集: 该数据集包含 BBC, Reuters 和 The Guardianz 这三大新闻资源的 419 篇文章, 这 419 篇文章被分为商业、娱乐、健康、政治、体育和科技 6 个类.

前 4 个数据集是完整的多视图数据. 而 3sources 数据集是一种天然缺失的不完整多视图数据集, 3 个视图的缺失率分别为 15.38%, 29.33%, 27.40%. 表 2 总结了 5 个数据集的详细情况.

表 2 数据集的统计信息

数据集	HW	MSRCV1	BBC	NGs	3sources
$\#d_1$	216	1 302	4 659	2 000	3 560
$\#d_2$	76	48	4 633	2 000	3 631
$\#d_3$	64	521	4 665	2 000	3 069
$\#d_4$	6	100	4 684	–	–
$\#d_5$	240	256	–	–	–
$\#d_6$	47	210	–	–	–
#viwe	6	6	4	3	3
#instances	2 000	210	685	500	416
#class	10	7	5	5	6

在多视图数据中, 不同视图之间存在尺度差异. 为了消除这些差异, 提高聚类性能, 有必要对数据集进行归一化处理. 如果对比实验没有明确声明归一化的方法, 本文将按照 L_2 范数对数据进行统一归一化.

3.3 对比算法

实验过程中,我们将和以下算法进行聚类性能比较.

(1) 单视图谱聚类算法: BSV 算法[1]在每个视图上单独进行标准化谱聚类算法, 并选取这些视图中最好的聚类结果进行对比.

(2) 基于数据补全的方法: MIC 算法[25]首先利用视图的均值对不完整多视图进行数据填充, 然后利用现有的基于 Multi-NMF 的多视图聚类算法进行后续的聚类操作; LF-IMVC 算法[31]是一个基于核矩阵补全的算法, 用集成的方法将多个核矩阵进行融合.

(3) 利用对齐信息的方法: PVC[32]是不完整多视图聚类的一项先驱工作, PVC 旨在寻找对齐样本和非对齐样本共享的潜在空间; IMG[34]在 PVC 算法的基础上, 通过使用一个图拉普拉斯正则项来约束不完整多视图数据的全局结构. 上述两种方法仅适用于两视图数据, 因此我们评估所有两视图组合, 并报告最佳聚类结果.

(4) 基于近似完整图学习的方法: GIMC 算法[37]首次提出了基于近似完整图学习的概念, 旨在其他视图的帮助下, 有效地为每个视图构造一个完整的图, 并自动对多个图进行加权学习一致图; IMSC-AGL[44]将不完整视图的图构造和数据的一致性表示集成于一个联合优化框架中, IMSC-AGL 算法旨在利用仿射图学习并结合谱聚类和协同正则化方法来解决视角缺失问题, 是一种极具代表性的近似完整图学习方法.

3.4 实验设置及评价指标

为了模拟不完整的多视图设置，本文从完整的多视图数据的每个视图中随机删除一些实例. 具体来说，从每个视图中随机删除 *per*%(*per*∈{10,20,30,40,50})实例，构造缺失率为 *per*%的不完整数据. 对于每个样本，确保它的实例至少存在于一个视图中. 本文的实验环境在 Intel i7-7700 CPU 48G 内存的个人计算机上进行. 此外，在 *N*-cut 图划分过程中，算法需要使用 *k*-means 算法来实现最终的聚类结果. 由于 *k*-means 对初始化很敏感，为了减少由 *k*-means 引起的随机性影响，本文用随机初始化重复聚类过程 20 次，并报告平均值. 本文采用聚类精度(ACC)、归一化互信息(NMI)和纯度(purity)这 3 个外部指标评价聚类性能. 一般来说，我们期望这些评价标准的值尽可能大.

3.5 真实数据集实验结果

表 3–表 7 分别显示了不同缺失率下，各对比算法在多个真实数据集上的 ACC 和 NMI 以及 Purity 的对比.

表 3　各聚类方法在 3sources 数据集上的聚类结果

	ACC	NMI	Purity
BSV	0.540 9	0.383 7	0.564 9
PVC	0.572 5	0.508 9	0.664 3
IMG	0.558 2	0.501 8	0.610 4
MIC	0.651 4	0.619 2	0.742 8
LF-IMVC	0.634 6	0.541 8	0.706 7
IMSC-AGL	0.813 9	0.690 1	0.823 9
GIMC	0.733 2	0.502 4	0.628 3
MNDF	**0.887 0**	**0.751 1**	**0.880 2**

表 4　不同缺失率下各聚类方法在 HW 数据集上的聚类结果

		BSV	PVC	IMG	MIC	LF-IMVC	IMSC-AGL	GIMC	MNDF
ACC	10%	0.550 8	0.595 4	0.663 7	0.739 0	0.895 5	0.957 5	0.891 5	**0.961 2**
	20%	0.504 6	0.574 5	0.603	0.665 3	0.885 5	0.956 0	0.883 0	**0.957 7**
	30%	0.457 5	0.552 5	0.568 1	0.601 4	0.880 5	0.939 4	0.875 5	**0.943 2**
	40%	0.436 3	0.532 0	0.542 8	0.524 6	0.806 5	0.911 6	0.779 0	**0.920 0**
	50%	0.383 5	0.522 0	0.523 9	0.512 3	0.722 0	0.897 7	0.585 0	**0.901 1**
NMI	10%	0.515 7	0.541 0	0.572 9	0.648 1	0.804 8	0.917 0	0.897 8	**0.921 7**
	20%	0.489 0	0.512 1	0.533 2	0.582 7	0.787 7	0.910 8	0.887 5	**0.921 1**
	30%	0.428 9	0.480 5	0.5118	0.536 0	0.779 4	**0.899 4**	0.888 1	0.899 1
	40%	0.384 4	0.461 3	0.478 1	0.483 5	0.677 9	0.860 1	**0.874 7**	0.871 4
	50%	0.324 3	0.448 5	0.461 2	0.448 1	0.625 6	0.813 4	0.778 5	**0.824 4**
Purity	10%	0.512 5	0.548 5	0.587 4	0.648 5	0.800 2	0.914 5	0.891 2	**0.919 8**
	20%	0.492 2	0.517 8	0.551 2	0.594 1	0.791 6	0.909 4	0.884 3	**0.910 2**
	30%	0.451 5	0.498 7	0.524 9	0.537 4	0.785 6	0.887 0	0.879 8	**0.900 0**
	40%	0.406 2	0.481 2	0.509 5	0.510 0	0.713 3	0.870 6	**0.875 4**	0.875 3
	50%	0.357 5	0.469 7	0.498 8	0.470 1	0.637 4	0.851 4	0.850 1	**0.852 1**

表 5　不同缺失率下各聚类方法在 MSRCV1 数据集上的聚类结果

		BSV	PVC	IMG	MIC	LF-IMVC	IMSC-AGL	GIMC	MNDF
ACC	10%	0.501 1	0.561 7	0.628 0	0.761 9	0.600 0	0.887 6	0.580 1	**0.895 2**
	20%	0.492 1	0.539 2	0.547 3	0.720 9	0.594 7	0.842 9	0.561 9	**0.885 7**
	30%	0.412 1	0.478 3	0.531 2	0.647 6	0.538 1	0.773 3	0.538 1	**0.814 3**
	40%	0.267 7	0.417 9	0.512 4	0.569 5	0.514 3	0.728 6	0.452 4	**0.754 8**
	50%	0.214 0	0.373 4	0.493 6	0.434 2	0.433 3	0.624 7	0.385 7	**0.675 2**
NMI	10%	0.415 2	0.447 4	0.474 6	0.614 5	0.447 1	0.797 7	0.636 8	**0.800 5**
	20%	0.392 5	0.431 8	0.425 6	0.556 8	0.443 5	0.749 2	0.614 2	**0.784 0**
	30%	0.274 5	0.386 0	0.418 6	0.509 9	0.394 1	0.628 1	0.548 4	**0.662 6**
	40%	0.112 5	0.308 5	0.378 5	0.399 4	0.330 8	0.614 9	0.427 4	**0.631 0**
	50%	0.105 4	0.286 4	0.360 2	0.354 9	0.256 5	0.495 3	0.375 5	**0.560 0**
Purity	10%	0.417 4	0.455 4	0.487 5	0.611 7	0.427 4	0.800 1	0.617 4	**0.801 2**
	20%	0.398 9	0.434 9	0.426 7	0.580 1	0.435 5	0.763 4	0.601 2	**0.792 0**
	30%	0.282 4	0.403 4	0.418 1	0.515 4	0.408 1	0.622 1	0.582 3	**0.654 2**
	40%	0.157 4	0.315 8	0.387 5	0.417 4	0.324 5	0.618 7	0.454 7	**0.627 9**
	50%	0.143 2	0.297 5	0.371 0	0.370 1	0.261 4	0.515 2	0.384 0	**0.587 3**

表 6 不同缺失率下各聚类方法在 BBC 数据集上的聚类结果

		BSV	PVC	IMG	MIC	LF-IMVC	IMSC-AGL	GIMC	MNDF
ACC	10%	0.605 8	0.645 3	0.655 3	0.781 0	0.699 3	0.908 0	0.563 5	**0.914 3**
	20%	0.505 1	0.634 5	0.639 4	0.674 5	0.632 1	0.868 6	0.557 7	**0.872 6**
	30%	0.430 7	0.632 2	0.608 3	0.636 5	0.557 7	0.835 0	0.519 7	**0.838 0**
	40%	0.395 6	0.570 7	0.579 9	0.620 4	0.535 8	0.816 1	0.513 9	**0.821 1**
	50%	0.351 8	0.562 5	0.557 5	0.505 1	0.529 9	0.804 4	0.509 5	**0.812 2**
NMI	10%	0.379 9	0.510 4	0.502 1	0.603 0	0.513 8	0.762 8	0.613 0	**0.797 5**
	20%	0.269 3	0.480 4	0.485 2	0.545 2	0.471 0	0.692 1	0.659 0	**0.701 2**
	30%	0.223 0	0.489 0	0.419 4	0.434 7	0.404 2	0.643 2	0.637 5	**0.664 6**
	40%	0.203 0	0.448 3	0.409 9	0.408 0	0.307 8	0.596 8	0.616 2	**0.631 2**
	50%	0.166 2	0.422 6	0.377 5	0.237 4	0.292 3	0.571 3	0.569 3	**0.602 2**
Purity	10%	0.381 0	0.507 5	0.500 1	0.601 0	0.501 4	0.761 2	0.601 9	**0.800 1**
	20%	0.261 4	0.483 3	0.486 7	0.527 4	0.479 6	0.693 0	0.674 5	**0.701 4**
	30%	0.223 5	0.493 2	0.452 1	0.457 0	0.449 8	0.651 4	0.642 2	**0.681 0**
	40%	0.201 2	0.460 1	0.451 1	0.450 2	0.321 7	0.601 4	0.617 2	**0.625 4**
	50%	0.172 6	0.456 6	0.401 2	0.227 9	0.291 5	0.597 5	0.591 2	**0.612 0**

表 7 不同缺失率下各聚类方法在 NGs 数据集上的聚类结果

		BSV	PVC	IMG	MIC	LF-IMVC	IMSC-AGL	GIMC	MNDF
ACC	10%	0.571 2	0.713 4	0.690 4	0.898 0	0.848 0	0.950 0	0.784 0	**0.970 0**
	20%	0.542 5	0.662 4	0.647 5	0.828 0	0.784 0	0.924 0	0.772 0	**0.924 0**
	30%	0.493 0	0.635 4	0.602 2	0.574 0	0.644 0	0.812 0	0.748 0	**0.896 0**
	40%	0.412 0	0.623 0	0.582 8	0.520 0	0.610 0	0.804 0	0.592 0	**0.876 0**
	50%	0.370 0	0.568 6	0.579 3	0.424 0	0.620 0	0.770 0	0.586 0	**0.791 2**
NMI	10%	0.342 8	0.525 2	0.489 8	0.752 8	0.652 2	0.856 4	0.822 2	**0.910 3**
	20%	0.322 7	0.444 0	0.458 8	0.600 9	0.540 4	0.785 7	0.787 7	**0.796 6**
	30%	0.312 6	0.394 5	0.412 6	0.384 9	0.420 5	0.574 1	0.712 9	**0.754 6**
	40%	0.211 7	0.381 2	0.384 6	0.282 4	0.366 0	0.557 3	0.622 9	**0.713 5**
	50%	0.172 2	0.363 4	0.344 0	0.142 4	0.382 3	0.482 3	0.611 6	**0.682 1**
Purity	10%	0.381 29	0.512 2	0.498 2	0.748 8	0.651 2	0.873 1	0.809 2	**0.909 8**
	20%	0.312 2	0.443 2	0.451 0	0.600 1	0.547 3	0.791 2	0.792 0	**0.800 1**
	30%	0.300 1	0.396 5	0.407 8	0.395 1	0.409 9	0.568 9	0.701 2	**0.720 1**
	40%	0.202 1	0.390 1	0.393 0	0.297 4	0.387 0	0.559 1	0.632 8	**0.703 2**
	50%	0.187 5	0.378 8	0.367 4	0.159 9	0.390 4	0.501 2	0.627 9	**0.683 4**

从这些实验结果中能够得出以下结论.

MIC 算法用视图内的平均值来填充缺失的实例, MIC 的性能退化速度较其他方法要快. 从这个结果可以看出: 基于视图内统计信息的填充方法可能不是一个有效的方法, 因为不同的缺失样本并不一定存在于同一类中. 用视图内的均值填充缺失样本将弱化它们之间的差异, 特别是当缺失率较大时, 聚类的性能会急速下降. 本文可以得出结论: 基于视图填充的方法只有在视图缺失率相对较低的情况下才有效, 当缺失率较大时, 基于填充的不完整多视图聚类算法将失效.

与多视图聚类算法相比, 单视图聚类算法结果往往是最差的. 这表明单视图聚类算法由于缺乏其他视图的互补信息, 并不能获得准确的聚类结果. 另外, 当缺失比例固定时, 基于对齐信息的不完整两视图聚类算法(PVC 和 IMG)较其他基于多个视图的不完整多视图聚类算法结果要略低 0.1–0.2. 究其原因, PVC 和 IMG 仅适用于两个不完整视图, 视图之间的互补信息没有进行充分利用. 此外, 基于对齐信息的不完整多视图聚类算法对缺失条件有一定的限制, 视图内的样本不能任意缺失, 多个视图中必须存在一定的公共样本作为对齐信息. 综上可以得出如下结论: 利用多视图的互补信息是解决不完全问题的有效方法, 使用的视图越多, 聚类效果越好.

在大多数情况下, 基于 NMF 的不完整多视图聚类方法比基于多核和图的方法聚类效果差, 因为基于矩阵分解的方法并不能有效地处理非线性数据. 然而, 基于图和基于多核的聚类算法仅在原始数据上构造图和核, 因此, 算法对初始图和核是敏感的. 本文所提的算法在大多数数据集上都取得了相对较好的结果: 首先, MNDF 算法实则为一种非线性融合的方法, 可以有效地处理非线性数据, 整个过程无需为每个图分配特定的权重, 这样也避免了优化权重的问题; 其次, MNDF 算法首先利用多阶相似性学习不完整视图潜在结构信息,

从浅到深地对不完整多视图数据构图, 从而避免了算法对初始图敏感的问题. 另外, 交叉扩散融合过程充分利用每个图的局部结构信息, 为别的图更新提供指导, 因此这种方法不仅可以最终获得一致性完整图, 还充分保留了每个图的有用信息.

随着多视图数据样本缺失率的增加, 所有不完整多视图聚类算法的聚类性能在大多数情况下都急剧下降. 但是相比于其他对比算法, MNDF 算法的精度保持相对较高, 且性能下降率相对较低. 如表 4–表 6 所示: 缺失率从 10%升至 50%时, HW 数据集的 ACC 和 Purity 降低约 5%, NMI 降低约 10%; MSRCV1 数据集的 ACC 和 Purity 降低约 20%, NMI 降低约 25%; BBC 数据集的 ACC 下降约 10%, NMI 和 Purity 下降约 20%; NGs 数据集的 ACC, NMI 和 Purity 均下降约 20%. 另外, 如表 3 所示: 天然不完整多视图数据集 3sources 在每个视图缺失率不同的情况下, 仍然获得最优的聚类性能. 综上, 可充分体现 MNDF 算法在不同情况的多视图数据缺失情况下均取得较好的聚类性能, 算法亦具有较强的鲁棒性.

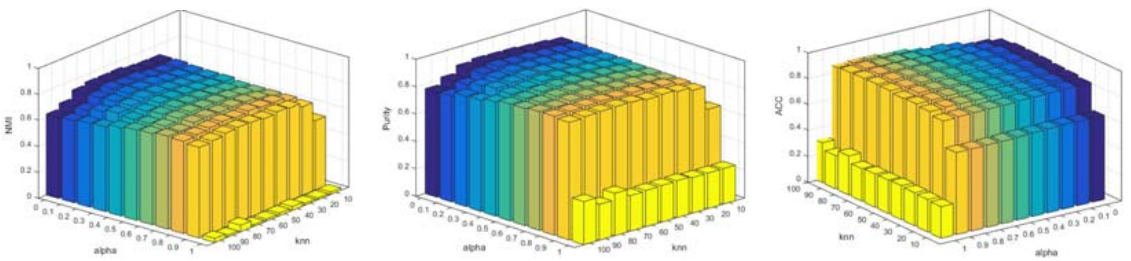
3.6 参数敏感性及算法性能分析

根据前文的理论分析已知, 不同阶的互补近邻可以为不完整多视图的图学习提供不同深度的结构信息. 根据我们的一般认知: 阶数越高, 挖掘的图结构信息越深. 然而事实并非如此, 随着阶数的增加, 图的稀疏性也逐渐增大. 倘若不加区分地构造足够深的图结构, 在后续的交叉扩散过程中, 有效的连边会受高阶稀疏图的影响而被切断. 因此, 以真实不完整多视图数据集 3sources 为例, 在实验中设置保留不同的阶数对比实验, 本文分别比较了保留二阶–五阶互补近邻图时算法对聚类结果的影响. 从表 8 可以看出, 保留三阶互补近邻图获得了最好的聚类性能. 因此, 在保证聚类性能的前提下, 为了提高计算效率, 在上述的所有实验中, 本文将算法的阶数都固定为三阶.

表 8 保留不同阶近邻关系时的性能比较

聚类评价	二阶	三阶	四阶	五阶
ACC	0.834 1	**0.887 0**	0.875 0	0.870 2
NMI	0.682 7	**0.751 1**	0.732 4	0.712 8
Purity	0.812 4	**0.880 2**	0.852 1	0.834 7

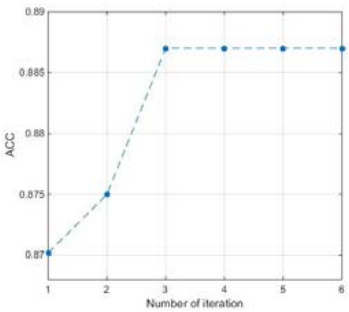
以真实不完整多视图数据集 3sources 为例, 本文还分析了 MNDF 算法的聚类性能与算法中 k 近邻参数和 α的关系, 以及迭代次数对算法收敛性的影响. 如图 7(a)–图 7(c)所示, 本文首先分析了两个参数在算法中的敏感性. k 近邻个数在候选参数集{10,20,30,40,50,60,70,80,90,100}中选取, α在候选参数集{0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1}中选取. 本文使用网格搜索方法来寻找最优的参数组合. 根据实验结果显示, 3sources 中 k 近邻个数在 50–70 之间, 参数α在 0.2–0.5 之间, MNDF 算法即可取得较好的聚类性能.



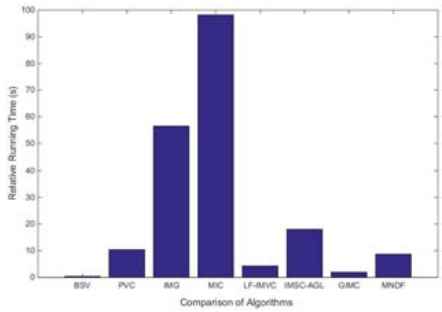
(a) k 近邻和α的参数分析实验 (b) k 近邻和α的参数分析实验 (c) k 近邻和α的参数分析实验

图 7 参数敏感性分析实验

在算法性能分析方面, 本节进行了算法收敛性分析和算法运行时间对比. 如图 8(a)所示: 随着迭代次数的增多, 聚类的准确率总体呈先上升后平稳的趋势. 如图可见: MNDF 算法收敛速度很快, 一般在迭代运算 3 次左右即可达到收敛. 如图 8(b)所示: MNDF 算法的运行时间相对较低, 耗时排名居中. 但是考虑到其在聚类精度方面的优越性能, MNDF 算法在实际应用中具有更广阔的前景.



(a) 所提算法的收敛性实验



(b) 算法的相对运行时间对比

图 8 算法性能分析实验

3.7 最优图融合方案实验分析

MNDF 算法的最优图是通过融合多种不同状态的相似性关系得到的, 目的是将多阶图中的边缘信息进行协同交互, 使得同类样本间的连接更加紧密而不同类样本的连接相对减弱, 从而增强图中连边信息的表达能力, 提高后续图划分的准确率. 本节就最优图的融合计算过程进行分析与讨论, 分别设置了权重有所偏重的加权平均方案以及本文 MNDF 算法设置的均值计算方案.

如表 9 所示, 通过对各方案的聚类性能进行比较可以看出, 不同的多阶图融合方案对最终的算法结果有一定的影响. 当一阶图的权值较大时, 算法的聚类性能相对最优; 当三阶图的权值较大时, 算法的性能相对最差. 虽然合适的超参会一定程度地提升算法性能, 但是在可行域范围内寻找最优的参数也会给算法带来额外的计算成本. 通过实验表明: 在聚类性能和计算成本的双重权衡考虑之下, 本文所提方案的聚类性能尚佳, 且避免了权重选择, 在算法的实际应用中表现出了一定的实用性和优越性.

表 9 最优图融合方案的实验分析

权重设置	ACC	NMI	Purity
[0.2,0.3,0.5]	0.882 2	0.741 5	0.851 5
[0.2,0.5,0.3]	0.884 6	0.746 0	0.863 7
[0.3,0.2,0.5]	0.883 1	0.745 5	0.872 4
[0.3,0.5,0.2]	0.885 1	0.748 5	0.884 0
[0.5,0.2,0.3]	0.899 0	0.773 3	0.891 2
[0.5,0.3,0.2]	0.894 2	0.762 5	0.890 4
Ours	0.887 0	0.751 1	0.880 2

4 总 结

本文提出了一种基于多阶近邻扩散融合的不完整多视图聚类算法, 该算法具有鲁棒性强、准确度高且高效的优点. MNDF 算法在利用多阶相似性学习不完整视图潜在结构的基础上, 通过跨视图图扩散的方式将不同阶的深层结构信息进行非线性融合. 换言之, MNDF 算法能够挖掘多个视图的潜在结构信息, 利用视图间的多样性推断出更加合理的多阶互补近邻图, 进而能够充分地利用这些视图的高阶信息和互补信息来学习更加鲁棒的近似完整图. 通过大量的实验, 验证了本文所提方法在不完整多视图聚类问题上的有效性. 在未来的发展方面, 使用多阶相似性进行的非线性融合的方法可以应用于其他存在数据缺失、数据稀少的机器学习子领域, 同时, 对多模态信息交互研究也具有启发意义. 此外, 如何利用不完整多视图数据的其他高阶结构信息来解决样本缺失的多视图聚类问题, 是我们未来关注的重点.

References:

[1] Xu XM, Li KK, He SF. GDFace: Gated deformation for multi-view face image synthesis. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 12532−12540.

[2] Fei HL, Li P. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 5759−5771.

[3] Ulrike VL. A tutorial on spectral clustering. Statistics and Computing, 2007, 17(4): 395−416.

[4] Alex R, Alessandro L. Clustering by fast search and find of density peaks. Science, 2014, 344(6191): 1492−1496.

[5] Bai L, Liang JY. Sparse subspace clustering with entropy-norm. In: Proc. of the 37th Int'l Conf. on Machine Learning. Vienna, 2020. 561−568.

[6] Zhao J, Xie XJ, Xu X, *et al*. Multi-view learning overview: Recent progress and new challenges. Information Fusion, 2017, 38: 43−54.

[7] Fu LL, Lin PF, Athanasios VV, *et al*. An overview of recent multi-view clustering. Neurocomputing, 2020, 402: 148−161.

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

[9] Zhang YP, Zhou J, Deng ZH, *et al*. Multi-view fuzzy clustering approach based on medoid invariant constraint. Ruan Jian Xue Bao/Journal of Software, 2019, 30(2): 282−301 (in Chinese with English abstract). http://www. jos.org.cn/1000-9825/5625.htm [doi: 10.13328/j.cnki.jos.005625]

[10] Kumar A, Daume III H. A co-training approach for multi-view spectral clustering. In: Proc. of the 28th Int'l Conf. on Machine Learning. Washington: Omnipress, 2011. 393−400.

[11] Kumar A, Rai P, Daume III H. Co-regularized multi-view spectral clustering. In: Proc. of the 25th Annual Conf. on Neural Information Processing Systems. 2011. 1413−1421.

[12] Jiang YZ, Deng ZH, Wang J, *et al*. Collaborative partition multi-view fuzzy clustering algorithm using entropy weighting. Ruan Jian Xue Bao/Journal of Software, 2014, 25(10): 2293−2311 (in Chinese with English abstract). http://www.jos. org.cn/1000-9825/4510.htm [doi: 10.13328/j.cnki.jos.004510]

[13] Zhang Y, Kong XW, Wang ZF, *et al*. Matrix factorization for multi-view clustering. Acta Automatica Sinica, 2018, 44(12): 2160−2169 (in Chinese with English abstract).

[14] Zhang CQ, Fu HZ, Hu QH, *et al*. Generalized latent multi-view subspace clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020, 42(1): 86−99.

[15] Chen MS, Huang L, Wang CD, *et al*. Multi-view clustering in latent embedding space. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 3513−3520.

[16] Liu J, Cao FY, Gao XZ, *et al*. A cluster-weighted kernel *K*-means method for multi-view clustering. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 4860−4867.

[17] Zhou SH, Liu XW, Liu JY, *et al*. Multi-view spectral clustering with optimal neighborhood laplacian matrix. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 6965−6972.

[18] Xia DX, Yang Y, Wang H, *et al*. Late fusion multi-view clustering based on local multi-kernel learning. Journal of Computer Research and Development, 2020, 57(8): 1627−1638 (in Chinese with English abstract).

[19] Zhou SH, Liu XW, Li MM, *et al*. Multiple kernel clustering with neighbor-kernel subspace segmentation. IEEE Trans. on Neural Networks and Learning Systems, 2019, 31(4): 1351−1362.

[20] Tang C, Liu XW, Zhu XZ, *et al*. CGD: Multi-view clustering via cross-view graph diffusion. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 5924−5931.

[21] Zhan K, Niu CX, Chen CL, *et al*. Graph structure fusion for multiview clustering. IEEE Trans. on Knowledge and Data Engineering, 2018, 31(10): 1984−1993.

[22] Wang H, Yang Y, Liu B. GMC: Graph-based multi-view clustering. IEEE Trans. on Knowledge and Data Engineering, 2020, 32(6): 1116−1129.

[23] Yang X, Zhu ZF, Xu MX, *et al*. Missing view completion for multi-view data. Ruan Jian Xue Bao/Journal of Software, 2018, 29(4): 945−956 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5416.htm [doi: 10.13328/j.cnki.jos.005416]

[24] Lin YJ, Gou YB, Liu ZT, *et al*. COMPLETER: Incomplete multi-view clustering via contrastive prediction. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. IEEE, 2021. 11174−11183.

[25] Zhao BY, Zhang CQ, Chen L, *et al*. Generative model for partial multi-view clustering. Acta Automatica Sinica, 2020, 47(8): 1867−1875 (in Chinese with English abstract). [doi: 10.16383/j.aas.c200121]

[26] Trivedi A, Rai P, Daumé III H, *et al*. Multiview clustering with incomplete views. In: Proc. of the 24th Annual Conf. on Neural Information Processing Systems. 2010. 1−7.

[27] Shao WX, Shi XX, Philip SY. Clustering on multiple incomplete datasets via collective kernel learning. In: Proc. of the 13th Int'l Conf. on Data Mining. Dallas: IEEE, 2013. 1181−1186.

[28] Shao WX, He LF, Philip SY. Multiple incomplete views clustering via weighted nonnegative matrix factorization with L21 regularization. In: Proc. of the European Conf. on Machine Learning and Knowledge Discovery in Databases. Porto: Springer, 2015. 318−334.

[29] Zhu XZ, Liu XW, Li MM, *et al*. Localized incomplete multiple kernel *k*-means. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: IJCAI, 2018. 3271−3277.

[30] Liu XW, Zhu XZ, Li MM, *et al*. Efficient and effective incomplete multi-view clustering. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 43292−4399.

[31] Liu XW, Zhu XZ, Li MM, *et al*. Multiple kernel *k*-means with incomplete kernels. IEEE Tans. on Pattern Analysis and Machine Intelligence, 2019, 42(5): 1191−1204.

[32] Liu XW, Zhu XZ, Li MM, *et al*. Late fusion incomplete multi-view clustering. IEEE Tans. on Pattern Analysis and Machine Intelligence, 2019, 41(10): 2410−2423.

[33] Li SY, Jiang Y, Zhou ZH. Partial multi-view clustering. In: Proc. of the 28th AAAI Conf. on Artificial Intelligence. Quebec: AAAI, 2014. 1968−1974.

[34] Xu N, Guo YQ, Zheng X, *et al*. Partial multi-view subspace clustering. In: Proc. of the ACM Multimedia Conf. on Multimedia Conf. Seoul: ACM, 2018. 1794−1801.

[35] Zhao HD, Liu HF, Fu Y. Incomplete multi-modal visual data grouping. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence. Stockholm: IJCAI, 2016. 2392−2398.

[36] Qian B, Shen XB, Gu YY, *et al*. Double constrained NMF for partial multi-view clustering. In: Proc. of the 6th Int'l Conf. on Digital Image Computing: Techniques and Applications. Gold Coast: IEEE, 2016. 1−7.

[37] Min C, Cheng MM, Yu J, *et al*. Partial multi-view clustering via auto-weighting similarity completion. In: Proc. of the 13th Chinese Conf. on Biometric Recognition. Urumqi: Springer, 2018. 214−222.

[38] Zhou W, Wang H, Yang Y. Consensus graph learning for incomplete multi-view clustering. In: Proc. of the 23rd Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Macau: Springer, 2019. 529−540.

[39] Guo J, Ye JH. Anchors bring ease: An embarrassingly simple approach to partial multi-view clustering. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. Hawaii: AAAI, 2019. 118−125.

[40] Liu Y, Shen CY, Hu QH, *et al*. Adaptive sample-level graph combination for partial multiview clustering. IEEE Trans. on Image Processing, 2019, 29: 2780−2794.

[41] Wu J, Zhuge WZ, Tao H, *et al*. Incomplete multi-view clustering via structured graph learning. In: Proc. of the 15th Pacific Rim Int'l Conf. on Artificial Intelligence. Nanjing: Springer, 2018. 98−112.

[42] Wen J, Yan K, Zhang Z, *et al*. Adaptive graph completion based incomplete multi-view clustering. IEEE Trans. on Multimedia, 2020, 23: 2493−2504.

[43] Wen J, Zhang Z, Zhang Z, *et al*. Generalized incomplete multiview clustering with flexible locality structure diffusion. IEEE Trans. on Cybernetics, 2020, 51(1): 101−114.

[44] Wen J, Zhang Z, Zhang Z, *et al*. Unified tensor framework for incomplete multi-view clustering and missing-view inferring. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 10273−10281.

[45] Wen J, Xu Y, Liu H. Incomplete multi-view spectral clustering with adaptive graph learning. IEEE Trans. on Cybernetics, 2020, 50: 1418−1429.

[46] Liu JL, Teng SH, Fei LK, *et al*. Consensus learning approach to incomplete multi-view clustering. Pattern Recognition, 2021, 115: 107890.

[47] Li LS, Wan ZQ, He HB. Incomplete multi-view clustering with joint partition and graph learning. IEEE Trans. on Knowledge and Data Engineering, 2021. [doi: 10.1109/TKDE.2021.3082470]

[48] Wang B, Jiang JY, Wang W, *et al*. Unsupervised metric fusion by cross diffusion. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 2997−3004.

[49] Bai S, Bai X, Latecki LJ. Regularized diffusion process for visual retrieval. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. California: AAAI, 2017. 3967−3973.

附中文参考文献:

[9] 张远鹏, 周洁, 邓赵红, 等. 代表点一致性约束的多视角模糊聚类算法. 软件学报, 2019, 30(2): 282−301. http://www.jos.org.cn/1000-9825/5625.htm [doi: 10.13328/j.cnki.jos.005625]

[12] 蒋亦樟, 邓赵红, 王骏, 等. 熵加权多视角协同划分模糊聚类算法. 软件学报, 2014, 25(10): 2293−2311. http:// www.jos.org.cn/1000-9825/4510.htm [doi: 10.13328/j.cnki.jos.004510]

[13] 张祎, 孔祥维, 王振帆, 等. 基于多视图矩阵分解的聚类分析. 自动化学报, 2018, 44(12): 2160−2169.

[18] 夏冬雪, 杨燕, 王浩, 等. 基于邻域多核学习的后融合多视图聚类算法. 计算机研究与发展, 2020, 57(8): 1627−1638.

[23] 杨旭, 朱振峰, 徐美香, 等. 多视角数据缺失补全. 软件学报, 2018, 29(4): 945−956. http://www.jos.org.cn/1000-9825/5416.htm [doi: 10.13328/j.cnki.jos.005416]

[25] 赵博宇, 张长青, 陈蕾, 等. 生成式不完整多视图数据聚类. 自动化学报, 2020, 47(8): 1867−1875. [doi: 10.16383/j.aas.c200121]

刘晓琳(1990－), 女, 博士生, CCF 学生会员, 主要研究领域为机器学习, 数据挖掘.

白亮(1982－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 数据挖掘.

赵兴旺(1984－), 男, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 数据挖掘.

梁吉业(1962－), 男, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为数据挖掘, 机器学习, 人工智能.