

融合KL信息的多视图模糊聚类算法

贺娜, 马盈仓

(西安工程大学 理学院, 西安 710600)

摘要: 现有多视图模糊C均值聚类(FCM)算法通常将一个多视图分解为多个单视图进行数据处理, 导致视图数据聚类精度降低, 从而影响全局数据划分结果。为实现高维数据和多视图数据的高效聚类, 提出一种基于KL信息的多视图自加权模糊聚类算法。将多个视图信息及其权重进行拟合融入标准FCM算法, 求解多个隶属度矩阵和质心矩阵。在此基础上, 通过附加KL信息作为模糊正则项进一步修正共识隶属度矩阵并保持权重分布的平滑性, 其中KL信息是视图隶属度与其共识隶属度的比值, 最小化KL信息会使每个视图的隶属度偏向于共识隶属度以得到更好的聚类结果。实验结果表明, 该算法相比于传统聚类算法具有更好的聚类效果和更快的收敛速度, 尤其在3-Sources数据集上相比于MVASM算法的聚类精度、标准化互信息和纯度分别提升了7.46、15.34和5.48个百分点。

关键词: 多视图聚类; 模糊C均值; 权重; KL信息; 共识隶属度矩阵

开放科学(资源服务)标志码(OSID):



中文引用格式: 贺娜, 马盈仓. 融合KL信息的多视图模糊聚类算法[J]. 计算机工程, 2022, 48(7): 114-121, 150.

英文引用格式: HE N, MA Y C. Multi-view fuzzy clustering algorithm fused with KL information[J]. Computer Engineering, 2022, 48(7): 114-121, 150.

Multi-View Fuzzy Clustering Algorithm Fused with KL Information

HE Na, MA Yingcang

(School of Science, Xi'an Polytechnic University, Xi'an 710600, China)

[Abstract] Existing multi-view Fuzzy C-Means(FCM) clustering algorithms usually artificially decompose multi-view data into multiple single-view data for processing, reducing the clustering accuracy of view data and affecting the results of global data division. To achieve efficient clustering of high-dimensional and multi-view data, a multi-view self-weighted fuzzy clustering algorithm based on Kullback-Leibler(KL) information is proposed, fitting multiple view information and their weights into the standard FCM algorithm to solve multiple membership matrices and centroid matrices. On this basis, additional KL information is used as a fuzzy regular term to further correct the consensus membership matrix and maintain the smoothness of the weight distribution, where the KL information is the ratio of a view's membership to its consensus membership, and minimizing the KL information biases each view's membership towards consensus membership, resulting in improved clustering results. The results show that the proposed algorithm has an improved clustering effect and faster convergence speed than traditional clustering algorithms. In particular, the clustering Accuracy (ACC), Normalized Mutual Information (NMI), and Purity of the MVASM algorithm on the 3-Sources dataset increased by 7.46, 15.34, and 5.48 percentage points respectively.

[Key words] multi-view clustering; Fuzzy C-Means (FCM); weight; Kullback-Leibler (KL) information; consensus membership matrix

DOI: 10.19678/j.issn.1000-3428.0061852

0 概述

聚类是一种无监督的机器学习任务, 根据数据自身的距离或相似度将它们划分为若干组, 划分原则是组内距离最小化而组间距离最大化^[1-3]。常见的聚类算法主要包括基于层次^[4-5]、基于模糊C均值

(Fuzzy C-Means, FCM)^[6-7]、基于迭代^[8-9]、基于协作^[10-11]、基于分解^[12-14]、基于谱聚类^[15-17]等, 其中基于FCM的聚类算法由于具有较高的聚类准确率且易于处理、空间复杂度低, 受到学者们的广泛关注。但是基于FCM的聚类算法主要是针对单一视图数据的聚类算法, 当其面对多视图数据时只能对

基金项目: 国家自然科学基金(61976130); 陕西省重点研发计划(2018KW-021); 陕西省自然科学基金(2020JQ-923)。

作者简介: 贺娜(1995—), 女, 硕士研究生, 主研方向为机器学习; 马盈仓(通信作者), 教授、博士。

收稿日期: 2021-06-04 修回日期: 2021-08-01 E-mail: 1428471765@qq.com

各视图样本进行独立的聚类分析以得到每个视图下的聚类结果,然后使用集成学习机制^[18-19]将每个视图下的聚类结果进行统一,最终获取全局意义下的聚类结果。然而人为地将多视图数据分解为多个单视图数据进行处理会因不同视图聚类结果存在明显差异而影响最终获取的全局划分结果。

近些年来,学者们在多视图聚类算法研究方面取得了较大进展。文献[20]通过典型相关分析使用数据的多个视图来构建投影。文献[21]学习了同时为原始问题提供多个视图的非冗余子空间,并在每个视图找到一个聚类方法。文献[22]提出一种类似将异构图像特征与图相结合的多视角光谱聚类算法。文献[23]引入共正则化谱聚类,采用共正则化框架来解决多视图聚类问题。文献[24]提出解决大规模多视图数据的多视图K-means聚类算法。文献[25]提出同时进行特征选择和多视图聚类的结构化稀疏学习方法。文献[26]分析并研究了基于采样的主动式初始中心选择方法对K-means型多视图聚类算法的影响。虽然上述算法取得了较好的性能,但其中一些算法忽略了视图多样性。文献[27]通过人为干预或先验知识实现多视图加权,但不能保证最终结果与每个视图的贡献相一致。文献[28]提出一种基于FCM的多视角模糊聚类方法CoFKM,但由于每个视图被平等对待,因此没有考虑每个视图的权重。文献[29]指出在某些视图有噪声或存在干扰的情况下,算法聚类精度可能会降低。受现有研究的启发,本文提出一种新的多视图自加权聚类算法KMFC。该算法通过学习共识隶属度矩阵进行聚类表示来挖掘多个视图的潜在共识信息,求解多个隶属度矩阵和质心矩阵,并利用视图的特定信息进一步修正共识隶属度矩阵。

1 模型建立与求解

1.1 模型建立

建立多视图聚类模型,通过引入Kullback-Leibler(KL)信息使学习的矩阵 U^* 尽可能与每个 $U^{(p)}$ 一致,从而得到如下目标函数:

$$\min \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \left\| \mathbf{x}_i^{(p)} - \mathbf{v}_k^{(p)} \right\|_2^2 \right) + \lambda \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \lg \frac{u_{ik}^{(p)}}{u_{ik}^*} \right) \quad (1)$$

在提高视图内聚类结果时,需要考虑不同视图间的聚类一致性。若直接将多个视图拼接在一起,则不利于提高聚类性能。更合理的方法是将这些视图与合适的权重 α_p ($p=1,2,\dots,m$)进行整合,并增加一个参数 q 来保持权重分布的平滑性。如果在式(1)中添加这些参数,则调整如下:

$$\begin{aligned} \min & \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \alpha_p^{(q)} \left\| \mathbf{x}_i^{(p)} - \mathbf{v}_k^{(p)} \right\|_2^2 \right) + \\ & \lambda \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \lg \frac{u_{ik}^{(p)}}{u_{ik}^*} \right) \\ \text{s.t.} & \sum_{k=1}^c u_{ik} = 1, u_{ik} \geq 0, \sum_{p=1}^m \alpha_p = 1, \alpha_p \geq 0, \\ & \sum_{k=1}^c u_{ik}^* = 1, u_{ik}^* \geq 0, i=1,2,\dots,n \end{aligned} \quad (2)$$

其中: $\mathbf{X}^{(p)} = [\mathbf{x}_1^{(p)}, \mathbf{x}_2^{(p)}, \dots, \mathbf{x}_n^{(p)}]$; $\mathbf{V}^{(p)} = [\mathbf{v}_1^{(p)}, \mathbf{v}_2^{(p)}, \dots, \mathbf{v}_c^{(p)}]$; $\mathbf{U}^{(p)} = [\mathbf{u}_1^{(p)}, \mathbf{u}_2^{(p)}, \dots, \mathbf{u}_c^{(p)}]$; $\mathbf{X}^{(p)} \in \mathbb{R}^{d_p \times n}$ 是在 p 个视图中具有 d_p 维和 n 个样本的数据矩阵,第 i 个数据点记为 $\mathbf{x}_i^{(p)} \in \mathbb{Z}^{d_p}$; $\mathbf{U}^{(p)} \in \mathbb{R}^{d_p \times c}$ 是 p 个视图的隶属度,它的第 k 个隶属度向量为 $\mathbf{u}_k^{(p)} \in \mathbb{R}^{d_p}$; 矩阵 $\mathbf{V}^{(p)} \in \mathbb{R}^{d_p \times c}$ 是 p 个视图的中心矩阵,它的第 k 个中心向量为 $\mathbf{v}_k^{(p)} \in \mathbb{R}^{d_p}$; $\mathbf{U}^* \in \mathbb{R}^{n \times c}$ 是不同视图上的公共隶属度矩阵; α_p 是第 p 个视图的权重; q 为各权重的幂指数; λ 为正则化系数。

由于每个视图都有各自的属性信息,因此KMFC算法可以使视图内信息与公共信息进行较好的拟合,其中视图内信息来自 $\mathbf{U}^{(p)}$,不同视图之间的公共信息来自共识隶属度矩阵 \mathbf{U}^* 。

1.2 模型求解

将多视图聚类模型的目标函数分为4个子问题,通过迭代交替优化方法进行求解。

1) 固定 \mathbf{V} 、 α 、 \mathbf{U}^* ,更新 \mathbf{U} ,问题式(2)可以改写如下:

$$\begin{aligned} \min & \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \alpha_p^{(q)} \left\| \mathbf{x}_i^{(p)} - \mathbf{v}_k^{(p)} \right\|_2^2 \right) + \\ & \lambda \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \lg \frac{u_{ik}^{(p)}}{u_{ik}^*} \right) \Leftrightarrow \\ \min & \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} h_{ik}^{(p)} \right) + \\ & \lambda \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \lg \frac{u_{ik}^{(p)}}{u_{ik}^*} \right) \end{aligned} \quad (3)$$

其中: $h_{ik}^{(p)} = \alpha_p^{(q)} \left\| \mathbf{x}_i^{(p)} - \mathbf{v}_k^{(p)} \right\|_2^2$ 。

式(3)的拉格朗日函数表示如下:

$$\begin{aligned} L(u_{ik}^{(p)}, \eta) &= \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} h_{ik}^{(p)} \right) + \\ & \lambda \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \lg \frac{u_{ik}^{(p)}}{u_{ik}^*} \right) + \sum_{i=1}^n \eta_i \left(\sum_{k=1}^c u_{ik} - 1 \right) \end{aligned} \quad (4)$$

其中: $\eta \geq 0$ 为约束条件下的拉格朗日乘子。

对式(4)中的 $u_{ik}^{(p)}$ 求导,设为0,得到:

$$\frac{\partial L}{\partial u_{ik}^{(p)}} = \sum_{p=1}^m h_{ik}^{(p)} + \eta_i + \lambda \sum_{p=1}^m \left(\lg \frac{u_{ik}^{(p)}}{u_{ik}^*} + 1 \right) = 0 \quad (5)$$

$$u_{ik}^{(p)} = \exp \left(\frac{-\sum_{p=1}^m h_{ik}^{(p)}}{\lambda} \right) u_{ik}^* \exp \left(\frac{-\eta_i - \lambda}{\lambda} \right) \quad (6)$$

考虑等式约束 $\sum_{k=1}^c u_{ik} = 1$,得到:

$$\begin{aligned} \sum_{k=1}^c \left[\exp \left(\frac{-\sum_{p=1}^m h_{ik}^{(p)}}{\lambda} \right) u_{ik}^* \exp \left(\frac{-\eta_i - \lambda}{\lambda} \right) \right] &= 1 \Rightarrow \\ \exp \left(\frac{-\eta_i - \lambda}{\lambda} \right) &= \frac{1}{\sum_{k=1}^c \left[\exp \left(\frac{-\sum_{p=1}^m h_{ik}^{(p)}}{\lambda} \right) u_{ik}^* \right]} \end{aligned} \quad (7)$$

结合式(6)和式(7)得到:

$$u_{ik}^{(p)} = \frac{\exp\left(\frac{-\sum_{p=1}^m h_{ik}^{(p)}}{\lambda}\right) u_{ik}^*}{\sum_{k=1}^c \left[\exp\left(\frac{-\sum_{p=1}^m h_{ik}^{(p)}}{\lambda}\right) u_{ik}^* \right]} \quad (8)$$

2) 固定 U, α, U^* , 更新 V , 省略与 V 无关的正则化项:

$$\begin{aligned} \min \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \alpha_p^{(q)} \|x_i^{(p)} - v_k^{(p)}\|_2^2 \right) + \\ \lambda \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \lg \frac{u_{ik}^{(p)}}{u_{ik}^*} \right) \Leftrightarrow \\ \min_V \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \alpha_p^{(q)} \|x_i^{(p)} - v_k^{(p)}\|_2^2 \right) \end{aligned} \quad (9)$$

使用 J 表示问题式(9)的目标函数, 得到:

$$J = \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \alpha_p^{(q)} \left(x_i^{(p)} \right)^T x_i^{(p)} + u_{ik}^{(p)} \alpha_p^{(q)} \left(v_k^{(p)} \right)^T v_k^{(p)} - 2 u_{ik}^{(p)} \alpha_p^{(q)} \left(x_i^{(p)} \right)^T v_k^{(p)} \right) \quad (10)$$

利用式(10)对 $v_k^{(p)}$ 求导, 使其为 0, 得到:

$$\frac{\partial J(v_k^{(p)})}{\partial v_k^{(p)}} = \sum_{i=1}^n \left(2 u_{ik}^{(p)} \alpha_p^{(q)} v_k^{(p)} - 2 u_{ik}^{(p)} \alpha_p^{(q)} x_i^{(p)} \right) = 0 \quad (11)$$

$$v_k^{(p)} = \frac{\sum_{i=1}^n u_{ik}^{(p)} x_i^{(p)}}{\sum_{i=1}^n u_{ik}^{(p)}} \quad (12)$$

3) 固定 U, V, U^* , 更新 α , 省略与 α 无关的正则化项:

$$\begin{aligned} \min_{\sum_{p=1}^m \alpha_p = 1, \alpha_p \geq 0} \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \alpha_p^{(q)} \|x_i^{(p)} - v_k^{(p)}\|_2^2 \right) + \\ \lambda \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \lg \frac{u_{ik}^{(p)}}{u_{ik}^*} \right) \Leftrightarrow \\ \min_{\sum_{p=1}^m \alpha_p = 1, \alpha_p \geq 0} \sum_{p=1}^m \alpha_p^{(q)} \left(\sum_{i=1}^n \sum_{k=1}^c u_{ik}^{(p)} \|x_i^{(p)} - v_k^{(p)}\|_2^2 \right) \Leftrightarrow \\ \min_{\sum_{p=1}^m \alpha_p = 1, \alpha_p \geq 0} \sum_{p=1}^m \alpha_p^{(q)} \Phi_p \end{aligned} \quad (13)$$

其中: $\Phi_p = \sum_{i=1}^n \sum_{k=1}^c u_{ik}^{(p)} \|x_i^{(p)} - v_k^{(p)}\|_2^2$

式(13)的拉格朗日函数表示如下:

$$L(\alpha, \gamma, \beta) = \sum_{p=1}^m \alpha_p^{(q)} \Phi_p - \gamma (\alpha^T \mathbf{1} - 1) - \beta^T \alpha \quad (14)$$

其中: γ, β 是拉格朗日乘子。

利用式(14)对 α_p 求导, 使其为 0, 得到:

$$\alpha_p^{q-1} = \frac{\gamma}{q \Phi_p} + \frac{\beta_p}{q \Phi_p} \quad (15)$$

问题式(13)的 KKT 条件表示如下:

$$\beta \geq 0 \quad (16)$$

$$\alpha \geq 0 \quad (17)$$

$$\alpha^T \beta = 0 \quad (18)$$

结合式(16)~式(18)得到 $\beta_p \alpha_p = 0$, 并将式(15)乘以 α_p 得到:

$$\alpha_p = \left(\frac{\gamma}{q \Phi_p} \right)^{\frac{1}{q-1}} \quad (19)$$

考虑等式约束 $\sum_{p=1}^m \alpha_p = 1$, 得到:

$$\sum_{p=1}^m \left(\frac{\gamma}{q \Phi_p} \right)^{\frac{1}{q-1}} = 1 \Rightarrow \left(\frac{\gamma}{q} \right)^{\frac{1}{q-1}} = \frac{1}{\sum_{p=1}^m \Phi_p^{\frac{1}{1-q}}} \quad (20)$$

结合式(19)与式(20)得到:

$$\alpha_p = \frac{\Phi_p^{\frac{1}{1-q}}}{\sum_{s=1}^m \Phi_s^{\frac{1}{1-q}}} \quad (21)$$

4) 固定 U, V, α , 更新 U^* , 省略与 U^* 无关的项:

$$\begin{aligned} \min_{\sum_{i=1}^n u_{ik}^* = 1, u_{ik}^* \geq 0} \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \alpha_p^{(q)} \|x_i^{(p)} - v_k^{(p)}\|_2^2 \right) + \\ \lambda \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \lg \frac{u_{ik}^{(p)}}{u_{ik}^*} \right) \Leftrightarrow \\ \min_{\sum_{i=1}^n u_{ik}^* = 1, u_{ik}^* \geq 0} \sum_{i=1}^n \sum_{k=1}^c \sum_{p=1}^m \left(u_{ik}^{(p)} \lg \frac{u_{ik}^{(p)}}{u_{ik}^*} \right) \end{aligned} \quad (22)$$

式(22)的拉格朗日函数表示如下:

$$\begin{aligned} L(u_{ik}, \xi) = \sum_{p=1}^m \sum_{i=1}^n \sum_{k=1}^c \left(u_{ik}^{(p)} \lg \frac{u_{ik}^{(p)}}{u_{ik}^*} \right) + \\ \sum_{i=1}^n \xi_i \left(\sum_{k=1}^c u_{ik}^* - 1 \right) \end{aligned} \quad (23)$$

其中: $\xi \geq 0$ 为约束条件下的拉格朗日乘子。

• 对式(23)中的 u_{ik}^* 求导, 设为 0, 得到:

$$\frac{\partial L}{\partial u_{ik}^*} = \sum_{p=1}^m \left(-\frac{u_{ik}^{(p)}}{u_{ik}^*} + 1 \right) + \xi_i = 0 \quad (24)$$

$$u_{ik}^* = \frac{1}{\xi_i} \sum_{p=1}^m u_{ik}^{(p)} \quad (25)$$

考虑等式约束 $\sum_{k=1}^c u_{ik}^* = 1$, 得到:

$$\sum_{k=1}^c \frac{1}{\xi_i} \sum_{p=1}^m u_{ik}^{(p)} = 1 \Rightarrow \xi_i = \sum_{k=1}^c \sum_{p=1}^m u_{ik}^{(p)} \quad (26)$$

结合式(25)和式(26)得到:

$$u_{ik}^* = \frac{\sum_{p=1}^m u_{ik}^{(p)}}{\sum_{k=1}^c \sum_{p=1}^m u_{ik}^{(p)}} \quad (27)$$

综上所述, 通过问题 1 的求解可更新 p 个视图的隶属度矩阵 $U^{(p)}$ 。通过问题 2 的求解更新 $v^{(p)}$ 可得到 p 个视图的聚类中心矩阵。通过问题 3 的求解可学习权值 α_p 来协调不同的视图。通过问题 4 的求解可学习一个共识隶属度矩阵 U^* 来表示不同视图之间的聚类。重复上述过程, 直到目标函数收敛。

1.3 算法流程

算法 1 KMFC 算法

输入 m 个视图的数据 $\{X^{(p)} | p = 1, 2, \dots, m\}$, $X^{(p)} \in \mathbb{R}^{d_p \times n}$, 聚类个数为 c , 参数 λ 和 q

输出 共识隶属度矩阵 $U^* \in \mathbb{R}^{n \times c}$, 聚类中心矩阵 $V^{(p)} \in \mathbb{R}^{d_p \times c}$ 和每个视图的权重 α_p

1. 初始化: 通过 K-Means 聚类方法初始化 $U^{(p)}$ 和 $V^{(p)}$, 对 p 个视图初始化 $\alpha_p = 1/m$;
2. while 不收敛 do
3. 对于每个数据点 i , 通过求解式(8)更新第 p 个视图的隶属度矩阵 $U^{(p)}$ 的第 i 行;
4. 根据式(12)更新第 p 个视图的第 k 个聚类中心向量 $v_k^{(p)}$;
5. 根据式(21)更新权重 α_p ;
6. 对于每个数据点 i , 通过求解式(27)计算共识隶属度矩阵 U^* 的第 i 行;
7. 输出 U^* , $V^{(p)}$ 和 α_p ;

2 理论分析

2.1 幂指数 q

利用参数 q 来调整权值分布, 根据式(21)可得出存在两种极端情况: 1) 当 $q \rightarrow \infty$ 时, KMFC 算法得到相等的权值, 即 $\frac{1}{c}$; 2) 当 $q \rightarrow 1^+$ 时, 设 Φ_o 为 $\{\Phi_1, \Phi_2, \dots, \Phi_o, \dots, \Phi_m\}$ 中的最小值, 将 Φ_o 代入式(21)中得到权值:

$$\lim_{q \rightarrow 1^+} \alpha_o = \lim_{q \rightarrow 1^+} \frac{1}{1 + \sum_{s \neq o} \left(\frac{\Phi_s}{\Phi_o} \right)^{\frac{1}{1-q}}} = 1 \quad (28)$$

由此可见, KMFC 算法将权重 1 赋给 Φ_o 最小的视图, 将权重 0 赋给其他视图。通过该方法可以保证 KMFC 算法在 $q > 1$ 时不存在平凡解。

2.2 收敛性分析

定理 1 在每次迭代中, 问题式(2)的目标函数值不断减小, 直到算法收敛。

证明 假设经过第 t 次迭代得到 $U^{(t)}$, $V^{(t)}$, $\alpha^{(t)}$, $U^{*(t)}$ 。在第 $t+1$ 次迭代中, 首先将 V , α 和 U^* 分别固定为 $V^{(t)}$, $\alpha^{(t)}$ 和 $U^{*(t)}$, 然后在不同的视图之间求解 $U^{(t+1)}$ 。根据问题式(8), $U^{(t+1)}$ 可由式(29)求解:

$$u_{ik}^{(t+1)(p)} = \frac{\exp \left(- \frac{\sum_{p=1}^m h_{ik}^{(t)(p)}}{\lambda} \right) u_{ik}^{*(t)}}{\sum_{k=1}^c \left[\exp \left(- \frac{\sum_{p=1}^m h_{ik}^{(t)(p)}}{\lambda} \right) u_{ik}^{*(t)} \right]} \quad (29)$$

其中: 上述目标函数和约束在 $u_i^{(p)}$ 域中是凸的。这样问题式(29)就会收敛到全局最优解。

同样地, 在第 $t+1$ 次迭代中, 将 U , α 和 U^* 分别固定为 $U^{(t)}$, $\alpha^{(t)}$ 和 $U^{*(t)}$, 根据问题式(12), $V^{(t+1)}$ 可由式(30)求解:

$$v_k^{(t+1)(p)} = \frac{\sum_{i=1}^n u_{ik}^{(t)(p)} x_i^{(p)}}{\sum_{i=1}^n u_{ik}^{(t)(p)}} \quad (30)$$

其中: 上述目标函数和约束在 $v_k^{(p)}$ 域中是凸的。这样问题式(30)就会收敛到全局最优解。

在第 $t+1$ 次迭代中, 将 U , V 和 U^* 分别固定为 $U^{(t)}$, $V^{(t)}$ 和 $U^{*(t)}$, 根据问题式(21), $\alpha^{(t+1)}$ 可由式(31)求解:

$$\alpha_p^{(t+1)} = \frac{\Phi_p^{\frac{1}{1-q}}}{\sum_{s=1}^m \Phi_s^{\frac{1}{1-q}}} \quad (31)$$

其中: 上述目标函数和约束在 α_p 域中是凸的。这样问题式(31)就会收敛到全局最优解。

在第 $t+1$ 次迭代中, 将 U , V 和 α 分别固定为 $U^{(t)}$, $V^{(t)}$ 和 $\alpha^{(t)}$, 根据问题式(27), $U^{*(t+1)}$ 可由式(32)求解:

$$u_{ik}^{*(t+1)} = \frac{\sum_{p=1}^m u_{ik}^{(t)(p)}}{\sum_{k=1}^c \sum_{p=1}^m u_{ik}^{(t)(p)}} \quad (32)$$

其中: 上述目标函数和约束在 u_i^* 域中是凸的。这样问题式(32)就会收敛到全局最优解。

显然, 问题式(2)可以分为 4 个子问题, 每个子问题都是一个凸优化问题。因此, 上述证明过程验证了算法 1 的收敛性。

3 实验与结果分析

根据聚类精度 (Accuracy, ACC)^[30]、标准化互信息 (Normalized Mutual Information, NMI)^[30]、相似性 (Jaccard)^[31]、纯度 (Purity)^[31] 4 个评价指标对 KMFC 算法进行性能评价。

3.1 数据集

在 5 个真实数据集上评估 KMFC 算法的性能:

1) COIL20 数据集^[32] 具有 1 440 张标准化图像, 包含 20 个对象, 每个对象对应 72 张图像。每张图像可以用 30 个等距投影 (Isometric, ISO)、19 个线性判别分析 (Linear Discriminant Analysis, LDA) 和 30 个邻域保持嵌入 (Neighborhood Preserving Embedding, NPE) 这 3 类异构特征来表示。

2) YALE 数据集^[33] 由 15 名受试者的 165 张图像组成, 每个受试者有 11 张图像, 对应不同的面部表情或形态。每张图像可以用 30 个 ISO、14 个 LDA 和 30 个 NPE 这 3 类异构特征来表示。

3) 3-Sources 数据集^[34] 包含 294 篇新闻文章, 涵盖商业、娱乐、健康、政治、体育和技术 6 个方面。第一视图有 3 068 个特征, 第二视图有 3 631 个特征, 第三视图有 3 560 个特征。

4) NUS-WIDE 数据集^[35] 由包含 81 个对象的 269 648 张图像组成。在实验中选择猫、牛、狗、麋鹿、鹰、马、狮子、松鼠、老虎、鲸鱼、狼、斑马 12 种动物类别。每张图像可以用 64 种颜色直方图、144 种颜色相关图、73 种边缘方向直方图、128 个小波纹理、225 个块状颜色矩和 500 袋基于 SIFT 描述的单词这 6 类低级特征来表示。

5) Prokaryotic phyla 数据集^[36] 包含 551 个原核生物种类, 包括文本数据和不同的基因组表达。本文选用的 1 个视图数据为由描述原核生物种类的文档词袋表示组成的文本数据, 另外 2 个视图数据为 2 种基因组表示^[17]。与文献[17]一样, 为降低数据集维数, 对 3 个视

图分别应用主成分分析(Principal Component Analysis, PCA)并保留解释90%方差的主成分。

数据集统计信息如表1所示。

表1 数据集统计信息
Table 1 Dataset statistics

名称	样本数	视图数	类别数	特征数
COIL20	1 440	3	20	79
YALE32	165	3	15	74
3-Sources	169	3	6	10 259
NUS-WIDE	1 600	6	8	1 134
Prokaryotic phyla	551	3	4	834

3.2 评价指标

对于ACC、NMI、Jaccard、Purity这些广泛使用的指标,值越大表示集群性能越好。

ACC表示聚类精度,定义如下:

$$A_{ACC} = \frac{\sum_{i=1}^n \delta(\tau_i, \text{map}(r_i))}{n} \quad (33)$$

其中: n 为样本点的个数; τ_i 为第 i 个样本点真实的类标签; r_i 为学习到的第 i 个样本点对应的类标签; $\delta(x, y)$ 定义为一个函数,当 $x=y$ 时 $\delta(x, y)=1$,否则为0; $\text{map}(r_i)$ 是一个映射函数,它将学习到的标签 r_i 与真实标签 τ_i 进行匹配。

NMI表示 τ_i 和 r_i 之间的相似程度,定义如下:

$$N_{NMI}(\tau_i, r_i) = \frac{\sum_{i=1}^c \sum_{j=1}^c (n_{i,j} \lg \frac{n_{i,j}}{n_i n_j})}{\sqrt{\left(\sum_{i=1}^c n_i \lg \frac{n_i}{n} \right) \left(\sum_{j=1}^c \hat{n}_j \lg \frac{\hat{n}_j}{n} \right)}} \quad (34)$$

其中: n_i 表示算法中每一类 $r_i(1 \leq i \leq c)$ 包含的样本点的个数; \hat{n}_j 表示算法中每一类 $\tau_j(1 \leq j \leq c)$ 包含的样本点的个数; $n_{i,j}$ 表示学习到的第 i 个样本点对应的类标签 r_i 和真实的类标签 τ_j 的交集中所包含的样本点的个数。

Jaccard度量有限样本集 N_1 和 N_2 之间的相似性,定义如下:

$$J_{\text{Jaccard}} = \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|} = \frac{T_{TP}}{T_{TP} + F_{FP} + F_{FN}} \quad (35)$$

其中: T_{TP} 是真阳性的数目; F_{FP} 是假阳性的数目; F_{FN} 是假阴性的数目。

Purity是正确类标签的百分比,定义如下:

$$P_{\text{Purity}} = \frac{1}{n} \sum_{i=1}^c \max_{1 \leq j \leq c} |\text{map}(r_i) \cap \tau_j| \quad (36)$$

Purity的取值范围为 $[0, 1]$,值越靠近1,性能越好。

3.3 实验设置

首先,比较KMFC算法与FCM算法和熵模糊C均值(Entropy Fuzzy C-Means, EFCM)算法的性能,以证明KMFC算法体现了多视图算法的优势。其次,将

KMFC算法与SWMC^[37]、MVASM^[38]、MVGL^[39]3种算法进行比较,其中比较算法中涉及的参数调整参照原文献进行设置并实验以显示最优结果。在KMFC算法中,参数 q 控制不同视图上权值的分配,参数 λ 调节共识隶属度矩阵的稀疏性。根据式(21)在 $[1, 30]$ 内以步长为0.01来调节 q ,在 $[0, 10]$ 内以步长0.1来调节 λ ,在3-Source、NUS-WIDE和Prokaryotic phyla数据集上设置 (q, λ) 分别为 $(1.22, 0.9)$ 、 $(1.30, 0.4)$ 和 $(1.02, 0.3)$ 时KMFC算法的评价指标变化,如图1~图3所示。

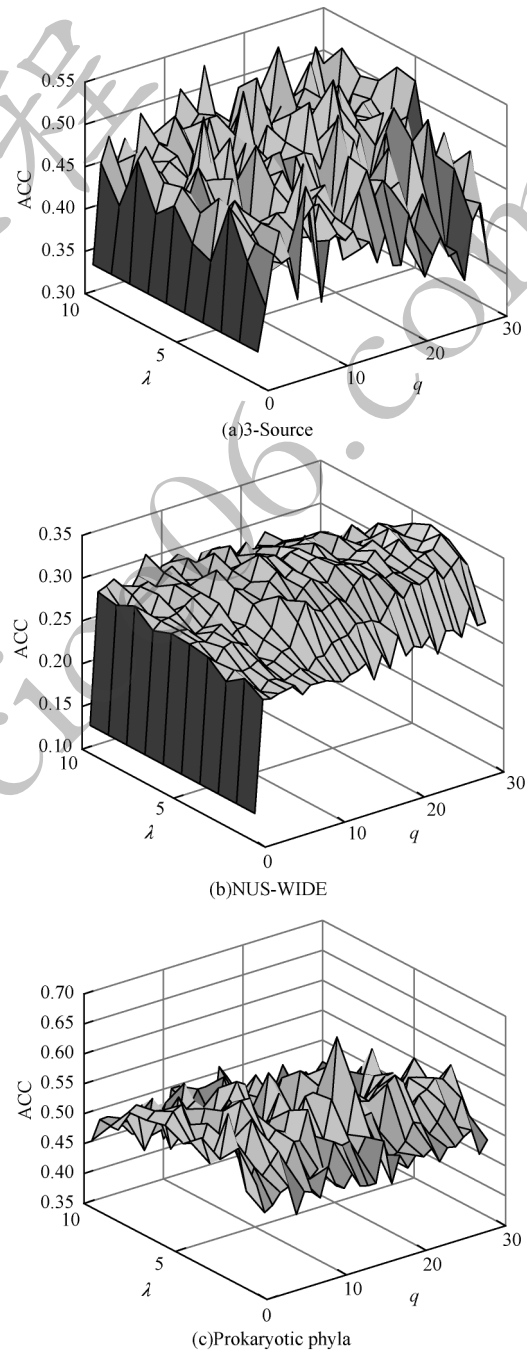


图1 KMFC算法在3-Source、NUS-WIDE和Prokaryotic phyla数据集上的ACC比较

Fig.1 ACC comparison of KMFC algorithm on 3-Source, NUS-WIDE and Prokaryotic phyla datasets

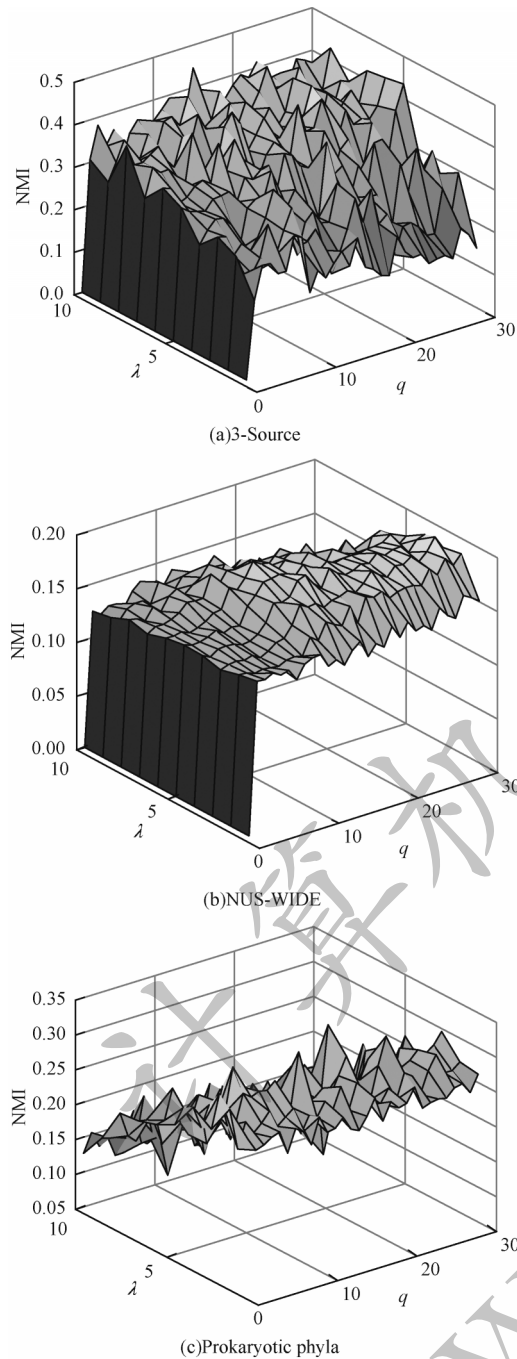


图2 KMFC算法在3-Source、NUS-WIDE和Prokaryotic phyla数据集上的NMI比较
Fig.2 NMI comparison of KMFC algorithm on 3-Source, NUS-WID and Prokaryotic phyla datasets

3.4 实验结果

在 COIL20 和 YALE32 数据集上, KMFC 和 2 种单视图聚类算法 FCM、EFCM 的 ACC、NMI、Jaccard 和 Purity 比较结果如表 2、表 3 所示, 其中括号中的数字表示算法排名。从表 2、表 3 可以看出: 在 ACC、NMI、Jaccard 和 Purity 4 个评价指标上, KMFC 算法在 COIL20 数据集上相比 FCM(2) 算法分别提高 0.584 9、0.375 5、0.692 4、0.687 4, 相比 EFCM(2) 算法分别提高 0.589 9、0.383 6、0.702 0、0.698 5; 在 YALE32 数据集上, KMFC 算法相比 FCM(2) 算法分

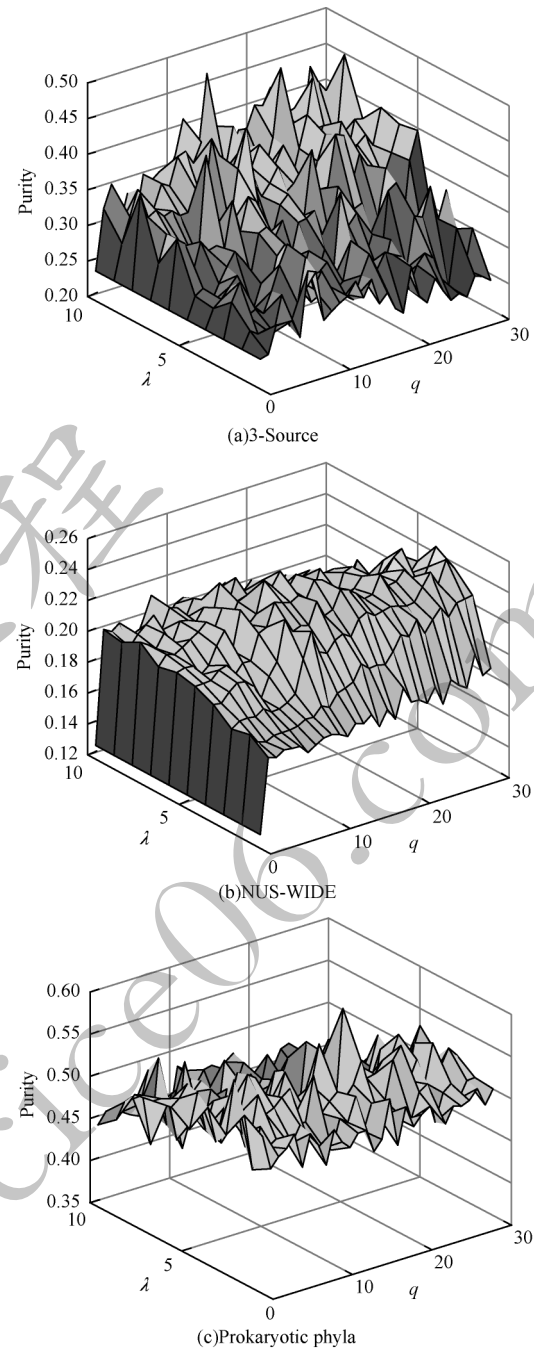


图3 KMFC算法在3-Source、NUS-WIDE和Prokaryotic phyla数据集上的Purity比较
Fig.3 Purity comparison of KMFC on 3-Source, NUS-WID and Prokaryotic phyla datasets

别提高 0.471 6、0.303 4、0.617 5、0.610 8, 相比 EFCM(2) 算法分别提高 0.506 7、0.313 4、0.628 4、0.625 3。上述结果证明了多视图学习的优越性和有效性, 并且其有利于提高聚类精度。

在 3-Source、NUS-WIDE 和 Prokaryotic phyla 数据集上不同多视图聚类算法的 ACC、NMI 和 Purity 比较结果如表 4~表 6 所示。从表 4~表 6 可以看出: 在数据集 3-Sources 上, KMFC 算法的 ACC、NMI、Purity 相比次优的 MVASM 算法分别提高 0.074 6、0.153 4、0.054 8; 在数据集 NUS-WIDE 上, KMFC 算法

相比次优的MVASM算法分别提高0.057 4、0.002 4、0.006 7;在数据集Prokaryotic phyla上,KMFC算法相比次优的MVGL算法分别提高0.060 9、0.062 6、0.000 7。KMFC算法在3个数据集上的收敛曲线如图4所示。可见,KMFC算法的聚类性能明显优于对比算法。

表2 在COIL20数据集上KMFC与2种单视图聚类算法的ACC、NMI、Jaccard和Purity比较

Table 2 Comparison of ACC,NMI,Jaccard and Purity between KMFC and two single-view clustering algorithms on COIL20 dataset

算法	ACC	NMI	Jaccard	Purity
FCM(1)	0.396 9	0.608 8	0.264 4	0.268 0
FCM(2)	0.404 0	0.621 6	0.287 4	0.294 0
FCM(3)	0.251 3	0.379 8	0.143 0	0.149 4
EFCM(1)	0.390 7	0.605 7	0.275 5	0.282 9
EFCM(2)	0.399 0	0.613 5	0.277 8	0.282 2
EFCM(3)	0.252 4	0.404 6	0.155 9	0.163 9
KMFC	0.988 9	0.997 1	0.979 8	0.981 4

表3 在YALE32数据集上KMFC与2种单视图聚类算法的ACC、NMI、Jaccard和Purity比较

Table 3 Comparison of ACC,NMI,Jaccard and Purity between KMFC and two single-view clustering algorithms on YALE32 dataset

算法	ACC	NMI	Jaccard	Purity
FCM(1)	0.333 3	0.571 4	0.260 3	0.260 3
FCM(2)	0.513 9	0.692 4	0.354 8	0.363 3
FCM(3)	0.426 7	0.625 6	0.267 2	0.267 2
EFCM(1)	0.373 3	0.593 2	0.260 1	0.260 1
EFCM(2)	0.478 8	0.681 7	0.343 9	0.348 8
EFCM(3)	0.360 0	0.561 5	0.230 2	0.230 2
KMFC	0.985 5	0.995 8	0.972 3	0.974 1

表4 在3-Source数据集上不同多视图聚类算法的ACC、NMI和Purity比较

Table 4 Comparison of ACC,NMI and Purity of different multi-view clustering algorithms on 3-Source dataset

算法	ACC	NMI	Purity
SWMC	0.360 9	0.092 3	0.402 4
MVASM	0.463 9	0.277 5	0.281 2
MVGL	0.355 0	0.080 7	0.402 4
KMFC	0.538 5	0.430 9	0.457 2

表5 在NUS-WIDE数据集上不同多视图聚类算法的ACC、NMI和Purity比较

Table 5 Comparison of ACC,NMI and Purity of different multi-view clustering algorithms on NUS-WIDE dataset

算法	ACC	NMI	Purity
SWMC	0.215 0	0.124 7	0.237 5
MVASM	0.286 0	0.182 8	0.200 1
MVGL	0.201 9	0.112 8	0.228 1
KMFC	0.343 4	0.185 2	0.244 2

表6 在Prokaryotic phyla数据集上不同多视图聚类算法的ACC、NMI和Purity比较

Table 6 Comparison of ACC,NMI and Purity of different multi-view clustering algorithms on Prokaryotic phyla dataset

算法	ACC	NMI	Purity
SWMC	0.506 4	0.032 6	0.568 1
MVASM	0.520 9	0.242 2	0.459 0
MVGL	0.542 7	0.027 1	0.575 3
KMFC	0.603 6	0.304 8	0.576 0

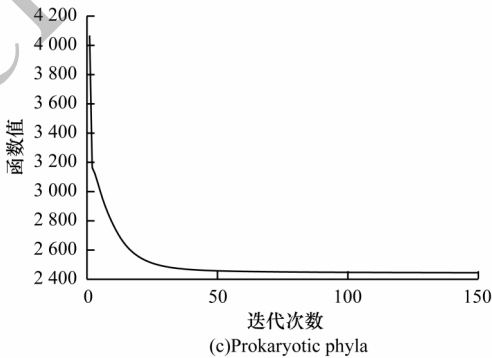
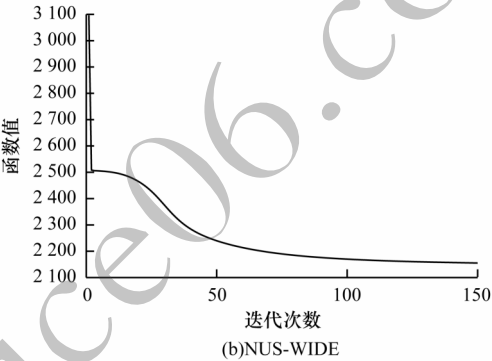
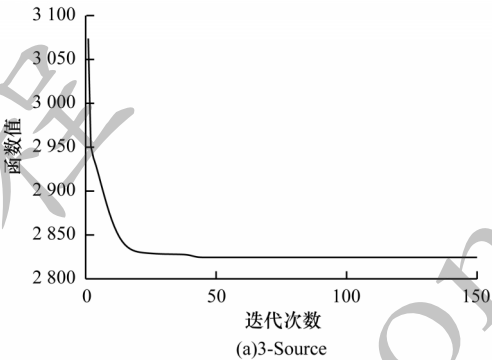


图4 在3-Sources、NUS-WIDE、Prokaryotic phyla数据集上KMFC算法的收敛曲线

Fig.4 Convergence curves of KMFC algorithm on 3-Sources,NUS-WIDE,Prokaryotic phyla datasets

4 结束语

本文提出一种新的多视图加权聚类算法,将每个视图信息及其权重进行拟合融入标准模糊C均值聚类算法,再附加一个KL信息作为模糊正则项,其中KL信息是一个视图的隶属度与其共识隶属度的比值,因此最小化KL信息会使每个视图的隶属度偏向于共识隶

属度,最终实现对共识隶属度矩阵的聚类。在多个数据集上的实验结果证明了该算法的有效性。但该算法中的权重需要引入幂指数 q 来进行调节,其细微变化即可影响算法性能,因此下一步将设计并实现融合KL信息的多视图完全自加权模糊聚类算法。

参考文献

- [1] XU C, TAO D C, XU C. A survey on multi-view learning [EB/OL]. [2021-05-11]. <https://arxiv.org/abs/1304.5634>.
- [2] WEN J, ZHANG Z, XU Y, et al. Unified embedding alignment with missing views inferring for incomplete multi-view clustering[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conference and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto, USA: AAAI Press, 2019: 5393-5400.
- [3] ZONG L L, ZHANG X C, LIU X Y, et al. Weighted multi-view spectral clustering based on spectral perturbation[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2018: 4621-4629.
- [4] JOHNSON S C. Hierarchical clustering schemes [J]. Psychometrika, 1967, 32(3): 241-254.
- [5] LEE J W T, YEUNG D S, TSANG E C C. Hierarchical clustering based on ordinal consistency [J]. Pattern Recognition, 2005, 38(11): 1913-1925.
- [6] WU K L, YANG M S. Alternative c-means clustering algorithms[J]. Pattern Recognition, 2002, 35(10): 2267-2278.
- [7] ZHANG D Q, CHEN S C. A comment on "alternative c-means clustering algorithms"[J]. Pattern Recognition, 2004, 37(2): 173-174.
- [8] TSENG P. Nearest q-flat to m points [J]. Journal of Optimization Theory and Applications, 2000, 105(1): 249-252.
- [9] WANG Y, ZHANG W J, WU L, et al. Iterative views agreement: an iterative low-rank based structured optimization method to multi-view spectral clustering[C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence. Washington D. C., USA: IEEE Press, 2016: 2153-2159.
- [10] PEDRYCZ W. Collaborative fuzzy clustering[J]. Pattern Recognition Letters, 2002, 23(14): 1675-1686.
- [11] CORNUÉJOLS A, WEMMERT C, GANÇARSKI P, et al. Collaborative clustering: why, when, what and how [J]. Information Fusion, 2018, 39: 81-95.
- [12] COSTEIRA J P, KANADE T. A multi-body factorization method for independently moving objects[J]. International Journal of Computer Vision, 1998, 29(3): 159-179.
- [13] LIU Y Y, JIAO L C, SHANG F H. An efficient matrix factorization based low-rank representation for subspace clustering[J]. Pattern Recognition, 2013, 46(1): 284-292.
- [14] WANG X B, LEI Z, SHI H L, et al. Co-referenced subspace clustering[C]//Proceedings of IEEE International Conference on Multimedia and Expo. Washington D. C., USA: IEEE Press, 2018: 1-6.
- [15] GUO X J. Robust subspace segmentation by simultaneously learning data representations and their affinity matrix[C]//Proceedings of the 24th International Conference on Artificial Intelligence. Washington D. C., USA: IEEE Press, 2015: 3547-3553.
- [16] WANG X B, GUO X J, LEI Z, et al. Exclusivity-consistency regularized multi-view subspace clustering[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 1-9.
- [17] BRBIĆ M, KOPRIVA I. Multi-view low-rank sparse subspace clustering [J]. Pattern Recognition, 2018, 73: 247-258.
- [18] ASUR S, UCAR D, PARTHASARATHY S. An ensemble framework for clustering protein-protein interaction networks[J]. Bioinformatics, 2007, 23(13): 29-40.
- [19] WANG H J, SHAN H H, BANERJEE A. Bayesian cluster ensembles[J]. Statistical Analysis and Data Mining, 2011, 4(1): 54-70.
- [20] 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, USA: ACM Press, 2009: 129-136.
- [21] NIU D L, JENNIFER G D, JORDAN M I. Multiple non-redundant spectral clustering views[C]//Proceedings of the 27th International Conference on Machine Learning. New York, USA: ACM Press, 2010: 831-838.
- [22] CAI X, NIE F P, HUANG H, et al. Heterogeneous image feature integration via multi-modal spectral clustering[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2011: 1977-1984.
- [23] KUMAR A, RAI P, DAUME H. Co-regularized multi-view spectral clustering[C]//Proceedings of the 24th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2011, 24: 1413-1421.
- [24] CAI X, NIE F P, HUANG H. Multi-view k-means clustering on big data[C]//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2013: 2598-2604.
- [25] WANG H, NIE F P, HUANG H. Multi-view clustering and feature learning via structured sparsity [J]. Journal of Machine Learning Research, 2013, 28(3): 352-360.
- [26] 洪敏, 贾彩燕, 王晓阳. K-means型多视图聚类中的初始化问题研究[J]. 计算机科学与探索, 2019, 13(4): 574-585. HONG M, JIA C Y, WANG X Y. Research on initialization of K-means type multi-view clustering[J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(4): 574-585. (in Chinese)
- [27] XIA R, PAN Y, DU L, et al. Robust multi-view spectral clustering via low-rank and sparse decomposition[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2014: 2149-2155.
- [28] CLEUZIOU G, EXBRAYAT M, MARTIN L, et al. CoFKM: a centralized method for multiple-view clustering[C]//Proceedings of the 9th IEEE International Conference on Data Mining. Washington D. C., USA: IEEE Press, 2009: 752-757.
- [29] JIANG Y Z, CHUNG F L, WANG S T, et al. Collaborative fuzzy clustering from multiple weighted views[J]. IEEE Transactions on Cybernetics, 2015, 45(4): 688-701.

(下转第150页)

(上接第121页)

- [30] CAI D, HE X, HAN J, et al. Document clustering using locality preserving indexing [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(12): 1624-1637.
- [31] CHEN G H, PAN Y, GUO M Y, et al. COMPACT: a comparative package for clustering assessment [C]//Proceedings of 2005 International Conference on Parallel and Distributed Processing and Applications. Berlin, Germany: Springer, 2005: 159-167.
- [32] NAYAR S. Columbia Object Image Library(COIL20) [EB/OL]. [2021-05-11]. https://www.researchgate.net/publication/2784735_Columbia_Object_Image_Library_COIL-100.
- [33] CAI D, HE X F, HAN J W. Using graph model for face analysis [EB/OL]. [2021-05-11]. https://www.semanticscholar.org/paper/Using-Graph-Model-for-Face-Analysis-Cai-He/19c889f2b26b785f0ac45c427339f0335b9cc514?_p2df.
- [34] GREENE D, CUNNINGHAM P. A matrix factorization approach for integrating multiple data views [M]. Berlin, Germany: Springer, 2009.
- [35] CHUA T S, TANG J H, HONG R C, et al. NUS-WIDE: a real-world Web image database from National University of Singapore [C]//Proceedings of the ACM International Conference on Image and Video Retrieval. New York, USA: ACM Press, 2009: 1-9.
- [36] BRBIĆ M, PIŠKOREC M, VIDULIN V, et al. The landscape of microbial phenotypic traits and associated genes [J]. Nucleic Acids Research, 2016, 44(21): 10074-10090.
- [37] NIE F P, LI J, LI X L. Self-weighted multiview clustering with multiple graphs [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization, 2017: 2564-2570.
- [38] HAN J W, XU J L, NIE F P, et al. Multi-view K-means clustering with adaptive sparse memberships and weight allocation [EB/OL]. [2021-05-11]. https://www.researchgate.net/publication/340572521_Multi-view_K-Means_Clustering_with_Adaptive_Sparse_Memberships_and_Weight_Allocation.
- [39] ZHAN K, ZHANG C Q, GUAN J P, et al. Graph learning for multiview clustering [J]. IEEE Transactions on Cybernetics, 2018, 48(10): 2887-2895.

编辑 陆燕菲