文章编号: 1001-0920 (2016) 04-0593-08 **DOI:** 10.13195/j.kzyjc.2015.0057

基于划分自适应融合的多视角模糊聚类算法

邓赵红, 张丹丹, 蒋亦樟, 刘解放, 王士同

(江南大学 数字媒体学院，江苏 无锡 214122)

摘 要: 针对多视角聚类任务如何更好地实现视角间的合作之挑战, 提出一种新的视角融合策略. 该策略首先为每个视角设置一个划分, 然后通过自适应学习获取一个融合权重矩阵对每个视角的划分进行自适应融合, 最终利用视角集成方法得到全局划分结果. 将上述策略应用到经典的FCM (Fuzzy C-means) 模糊聚类框架, 提出相应的多视角模糊聚类算法. 在模拟数据集和UCI数据集上的实验结果均显示, 所提出的算法较几种相关聚类算法在应对多视角聚类任务时具有更好的适应性和更好的聚类性能.

关键词: 数据划分；自适应融合；视角集成；自适应学习；模糊聚类

中图分类号: TP18 **文献标志码:** A

Multi-view fuzzy clustering algorithm based on partition adaptive-fusion

DENG Zhao-hong, *ZHANG Dan-dan*, *JIANG Yi-zhang*, *LIU Jie-fang*, *WANG Shi-tong*

(School of Digital Media，Jiangnan University，Wuxi 214122，China．Correspondent：DENG Zhao-hong，E-mail: dzh666828@aliyun.com)

Abstract: How to effectively realize the view fusion in the multi-view clustering is an important challenge. A new view fusion strategy is proposed for this task. Firstly, the partition matrix of each view is setup. And then the adaptive-fusion in each view partition is made based on the multiple view partitions by an adaptive-fusion weighting matrix. Finally, the global partition is obtained by using an integration approach. This strategy is integrated with the classic fuzzy C-means clustering framework, and the corresponding multi-view clustering algorithm is presented. Experimental studies are carried out on the synthetic and UCI real-world multi-view datasets. The experimental results show that the proposed algorithm outperforms several existing related algorithms in the adaptative abilities and clustering performance.

Keywords: data partition；adaptive-fusion；view integration；adaptive learning；fuzzy clustering

0 引 言

随着数据采集技术的发展, 数据的复杂性越来越高, 复杂数据给传统聚类技术带来许多挑战. 数据的复杂性包括很多方面, 如: 数据集的大小、数据属性特征的复杂性、时间性或可扩展性等. 观察复杂数据集时, 可以通过多个视角诠释, 即得到多视角数据. 多视角数据是指一些事物或对象从不同属性空间有多个视角或特征组, 它是从不同角度集合多种类型衡量的结果. 例如一个银行客户数据集, 可以被分为表示客户的人口信息的人口统计视角、显示有关客户账户信息的账户视角和描述客户消费行为的消费视角. 研究表明, 由于传统的聚类方法 K-means[1-2]、Fuzzy C-means (FCM)[3-4]、MEC[5-6]、PCM[7-8]等主要是针对单一视角数据的聚类方法, 当它们面对多视角数据时, 只能孤立地对各视角样本进行独立地聚类分析以获取每一视角下的聚类结果, 然后使用集成学习机制[9-10]将每一视角下的聚类结果进行统一, 最终获取全局意义下的聚类结果. 但是, 人为地将多视角数据分解为多个单视角数据进行处理会因某一视角的聚类结果不佳或不同视角聚类结果存在明显差异而影响最终获取的全局划分结果. 近年来, 针对多视角数据, 人们在模式识别领域提出了一种既从全局考虑各视角间共性, 又从局部考虑各视角间差异性的新学习技术, 即多视角学习技术. 该技术在聚类分析时通过有效地利用各视角间的共性和差异性来指导各视角的学习, 最终获取具备全局意义的一致性聚类结果.

收稿日期: 2015-01-13；**修回日期:** 2015-04-28.

基金项目: 国家自然科学基金面上项目(61170122)；江苏省杰出青年基金项目(BK20140001)；新世纪优秀人才支持计划项目(NCET120882).

作者简介: 邓赵红(1981–), 男, 副教授, 博士, 从事人工智能、神经模糊计算的研究；王士同(1964–), 男, 教授, 博士生导师, 从事人工智能、模式识别、神经模糊计算等研究.

基于多视角学习技术的聚类方法[11-17], 其有效性近年来得到了一定的关注和探讨. 文献 [11] 基于EM 算法提出了协同聚类算法, 即 CO-EM 算法, 可用于解决多视角聚类问题. 文献 [12] 同样基于协同的思想提出了一种双视角谱聚类算法, 该算法通过最小化视角间聚类结果的差异性对每一视角分别进行聚类, 以期得到较为一致的聚类结果. 文献 [13] 利用图论的知识提出了多视角谱聚类算法. 文献 [14] 基于经典的 K-means 算法提出一种双层自动加权 K-means 聚类算法 (TW-k-means 算法). 在已有的多视角聚类算法中, 基于划分的聚类算法近年来受到了较多关注. 文献 [15] 提出了 CoFC 算法, 该算法第一次在经典的 FCM 框架中通过给各个视角一个模糊划分, 采用协同的思想达到一致的结果. 文献 [16] 的 Co-FKM 是 CoFC 算法后又一个基于 FCM 框架的多视角模糊聚类算法, 该算法利用新的多视角协同划分策略, 在改进多视角协同学习的同时, 提出了一种集成学习策略以获取最终的划分结果. 相关文献的实验表明, Co-FKM 算法较 EM 框架下的多视角算法和其他一些相关算法具有一定的优势. 同时, 文献 [17] 基于 FCM 框架提出多视角模糊聚类 Co-FCM 算法及其增强的版本 WV-Co-FCM, 它们也属于基于划分的多视角聚类算法.

此外, 与多视角学习技术具有一定相关性的聚类技术还有多任务聚类技术、组合聚类技术和基于样本与特征空间的协同聚类技术等.

多任务聚类技术通过多任务学习框架将多个具有一定关联性的聚类任务进行协同学习, 如 LSSMTC 算法[18]. 当人们在处理多视角聚类任务时, 如果将各视角看作单一聚类任务, 则可以使用该聚类技术. 但是, 该技术在处理聚类任务时是针对不同的聚类对象, 且每个聚类对象作为单一任务处理, 而多视角数据聚类任务是不同视角下的同一个对象. 这种本质上的差异使得该技术在处理多视角聚类任务时无法得到理想的效果, 此外, 多任务聚类技术通常要求处理的各任务数据样本维数相等, 这使得其无法处理各视角样本维数不等的多视角聚类任务. 组合聚类技术将多个任务组合成一个任务再对该整体进行聚类分析, 如组合 K-means 算法 (CombKM)[18]. 该算法在处理多视角数据时, 由于多视角样本是同一个对象不同属性的组合形式, 只需要将这些不同属性下的样本组合为一个整体再进行处理即可. 然而, 这种做法可能会使得原来对象不同属性下的独立性遭到破坏, 从而使最终得到的全局聚类结果不理想. 人们基于样本与特征空间的协同聚类技术提出了对特征的划分, 其中多视角样本中不同视角数据样本的特征组合对应于特征划分的结果, Co-clustering 算法[19]就是一个典型的协同聚类算法. 在使用该算法处理多视角聚类任务时, 采用与组合聚类相同的做法, 也是先将各视角下的样本组合为一个整体再进行聚类处理. 这种情况下同时对特征进行划分, 得到的结果相对粗糙, 故而该技术应用于多视角聚类得到的聚类结果也不够理想.

从以上进展可知, 多视角学习技术已成为聚类的研究热点. 特别地, Co-FKM 方法[16]在已有的多视角聚类方法中显示出了较好的有效性和可扩展性, 其以经典的 FCM 算法为框架来实现多视角模糊聚类. Co-FCM 方法通过给每个视角分配一个模糊划分, 提出一种集成学习策略得到具有全局意义的划分结果, 但在实现各视角的协调学习过程中, 不能使各视角达到自动协调, 同时, 在处理多视角数据时需要人为控制各视角间的关系, 不利于实际的应用.

鉴于以上问题, 本文提出了基于划分自适应融合的多视角模糊聚类方法. 该方法通过自适应调整划分的权重矩阵, 使得各视角间能够自适应学习, 在处理多视角数据集时能更加灵活地融合各个视角的聚类结果, 从而可以期望能得到更好的聚类效果. 在模拟数据集和 UCI 数据集上的实验结果均显示, 所提出的算法较几种相关聚类算法在应对多视角聚类任务时具有更好的适应性和更好的聚类性能.

1 Co-FCM 算法

对于本文探讨的多视角聚类, 为了便于理解和描述, 首先给出如下定义.

定义 1 (视角) 为观测某个对象的特定角度. 对于同一对象, 不同的视角可提供不同的局部信息, 本文假定不同视角得到的信息是同构的.

一个多视角样本集

$$X = \{\mathrm{view}_1, \mathrm{view}_2, \cdots, \mathrm{view}_K\}$$

存在 K 个视角, 第 k 个视角的样本集可以表示为

$$\mathrm{view}_k = \{x_{1,k}, x_{2,k}, \cdots, x_{N,k}\}.$$

对该多视角数据利用 Co-FKM 聚类, 根据某种相似性度量, 每一视角均被聚类成 C $(2 \leqslant C \leqslant n)$ 个子类, 第 k 个视角的各类中心用矩阵 $Z_k = [z_{1,k}, z_{2,k}, \cdots, z_{c,k}]$ 表示, 第 k 个视角的划分可用矩阵 $U_k = [u_{ij,k}]$ 表示, 其中每一项满足如下约束:

$$\begin{cases} u_{ij,k} \in [0,1],\ 1 \leqslant i \leqslant C,\ 1 \leqslant j \leqslant N; \\ \displaystyle\sum_{i=1}^{C} u_{ij,k} = 1,\ 1 \leqslant k \leqslant K. \end{cases}$$

具体地, Co-FKM 算法的目标函数为

$$P(U,Z)=\sum_{k=1}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}\tilde{u}_{ij,k,\eta}\|x_{j,k}-z_{i,k}\|^2. \quad (1)$$

其中

$$\tilde{u}_{ij,k,\eta}=(1-\eta)u_{ij,k}^m+\frac{\eta}{K-1}\sum_{k'=1,k'\neq k}^{K}u_{ij,k'}^m, \quad (2)$$

参数 η 为调控各视角划分重要性的参数, $\tilde{u}_{ij,k,\eta}$ 表示当前视角下的划分隶属度 $u_{ij,k}^m$ 与其余各视角划分隶属度 $u_{ij,k'}^m$ 的加权融合.

根据式 (1), 通过拉格朗日极值求解方法, 可得到划分隶属度及类中心的优化迭代公式为

$$z_{i,k}=\frac{\sum_{j=1}^{N}\tilde{u}_{ij,k,\eta}x_{i,k}}{\sum_{j=1}^{N}\tilde{u}_{ij,k,\eta}},\ i=1,2,\cdots,C; \quad (3)$$

$$u_{ij,k}=\frac{1}{\sum_{h=1}^{C}\Big[\dfrac{(1-\eta)d_{ij,k}^2+\dfrac{\eta}{K-1}\sum_{k'=1,k'\neq k}^{K}d_{ij,k'}^2}{(1-\eta)d_{hj,k}^2+\dfrac{\eta}{K-1}\sum_{k'=1,k'\neq k}^{K}d_{hj,k'}^2}\Big]^{1/(m-1)}},$$

$$i=1,2,\cdots,C,\ j=1,2,\cdots,N. \quad (4)$$

利用式 (3) 和 (4) 最终可获得各视角下对应的模糊划分隶属度矩阵. 为了得到具有全局特性的模糊隶属度划分, 文献 [16] 在得到各视角下的模糊隶属度划分后, 利用几何均值的集成方法得到具有全局特性的划分结果, 其表达式为

$$\widehat{U}=\sqrt[K]{\prod_{k=1}^{K}U_k}. \quad (5)$$

由上述描述可知, Co-FKM 算法在处理多视角聚类任务时, 与传统的单视角聚类算法相比, 可以使得不同视角的聚类过程相互融合, 进而实现不同视角间的协调学习, 最终得到较有效的聚类性能. 然而, 上述聚类方法的划分融合策略依然存在一定的不足: 在实现视角划分融合时, 各视角划分的重要性必需人工设置, 即利用式 (2) 中的人工设置参数 η 来调控 $\tilde{u}_{ij,k,\eta}$ 中各视角间协同学习的程度, 因此具有较大的局限性. 由此分析可知, 如何实现各视角划分的自适应融合是一项非常有意义的工作, 下节针对此问题作了相应的探讨.

2　基于划分自适应融合的多视角模糊聚类算法

2.1　基于划分自适应融合的多视角模糊聚类算法

根据前节对目前基于划分融合的多视角聚类算法的局限性分析, 本文提出一种具有划分自适应融合能力的多视角模糊聚类方法, 其目标函数为

$$\begin{aligned}&P(U,Z,W)=\\&\sum_{k=1}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{t=1}^{K}w_{k,t}u_{ij,t}^m d_{ij,k}^2+\\&\gamma\sum_{k=1}^{K}\sum_{t=1}^{K}w_{k,t}\log(w_{k,t})=\\&\sum_{k=1}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}\tilde{u}_{ij,k}d_{ij,k}^2+\gamma\sum_{k=1}^{K}\sum_{t=1}^{K}w_{k,t}\log(w_{k,t}),\\&\sum_{i=1}^{C}u_{ij,k}=1,\ u_{ij,k}\in[0,1],\\&1\leqslant j\leqslant N,\ 1\leqslant k\leqslant K;\\&\sum_{t=1}^{K}w_{k,t}=1,\ w_{k,t}\in[0,1],\ 1\leqslant k\leqslant K.\end{aligned} \quad (6)$$

其中: $d_{ij,k}^2=\|x_{j,k}-z_{i,k}\|^2$; $U_k=[u_{ij,k}]$ 为第 k 个视角对应的划分矩阵; $\tilde{u}_{ij,k}=\sum_{t=1}^{K}w_{k,t}u_{ij,t}^m$ 为划分融合项, 实现了不同视角划分在第 k 个视角聚类任务中的视角融合, 此时每个视角划分的重要性通过权值 $w_{k,t}$ 来体现; $w_{k,t}$ 为第 t 个视角的划分在第 k 个视角聚类中的重要程度; 所有的视角划分权重构成了视角融合权重矩阵 W.

为了视角划分权重 W 的自适应学习, 式 (6) 引入了香浓熵正则化项. 香浓熵作为一种不确定度量, 在聚类方面得到了有效的利用, 如经典的极大熵聚类算法即是基于香浓熵的. 对不确定性进行划分时, 在没有任何先验信息的情况下一般认为熵达到极大时的划分最优. 另一方面, 当有其他信息可用时, 期望在可用信息上得到的不确定划分和极大熵情况下得到的划分间进行折衷. 基于上述原理, 极大化香浓熵在聚类、特征加权等方面得到了广泛的使用, 例如文献 [2, 5, 6, 22-23] 都引入了极大熵技术来开发相应的算法. 本文中, 引入该项可以使得在多视角聚类过程中各视角划分对应的权重得到自适应调控, 从而使得目标函数达到最优时获得最佳多视角融合的聚类效果. 目标函数中参数 γ 的取值可人工设定或利用网格寻优策略得到.

最后, 根据获取的各视角的划分利用如下集成方

法得到最终具备全局特性的空间划分矩阵[16]:

$$\bar{U} = \sqrt[K]{\prod_{t=1}^{K} U_t}. \tag{7}$$

2.2 目标函数的优化

通过如下策略对式(6)分解得到3个子问题:

问题 P_1: 固定 $U=\hat{U}$ 和 $W=\hat{W}$, 然后解决子问题 $P(\hat{U}, Z, \hat{W})$;

问题 P_2: 固定 $Z=\hat{Z}$ 和 $W=\hat{W}$, 然后解决子问题 $P(U, \hat{Z}, \hat{W})$;

问题 P_3: 固定 $U=\hat{U}$ 和 $Z=\hat{Z}$, 然后解决子问题 $P(\hat{U}, \hat{Z}, W)$.

对3个子问题进行优化求解可得到聚类的学习规则.

1) 问题 P_1 的解决方案由定理1给出.

定理1 假设 $U=\hat{U}$ 和 $W=\hat{W}$ 固定, $P(\hat{U}, Z, \hat{W})$ 最小化时的必要条件为

$$Z_{i,k} = \frac{\sum\limits_{j=1}^{N}\sum\limits_{t=1}^{K} w_{k,t} u_{ij,t}^{m} x_{j,k}}{\sum\limits_{j=1}^{N}\sum\limits_{t=1}^{K} w_{k,t} u_{ij,t}^{m}}. \tag{8}$$

证明 利用给定的隶属度划分矩阵 $\hat{U}$ 和视角融合权重矩阵 $\hat{W}$, 对目标函数求偏导, 并令 $\partial P/\partial Z=0$, 可得

$$Z_{i,k} = \frac{\sum\limits_{j=1}^{N}\sum\limits_{t=1}^{K} w_{k,t} u_{ij,t}^{m} x_{j,k}}{\sum\limits_{j=1}^{N}\sum\limits_{t=1}^{K} w_{k,t} u_{ij,t}^{m}}. \qquad \square$$

2) 问题 P_2 的解决方案由定理2给出.

定理2 假设 $Z=\hat{Z}$ 和 $W=\hat{W}$ 固定, $P(U, \hat{Z}, \hat{W})$ 最小化时的必要条件为

$$u_{ij,t} = \frac{1}{\sum\limits_{h=1}^{C}\left[\dfrac{\sum\limits_{k=1}^{K} w_{k,t} d_{ij,k}^{2}}{\sum\limits_{k=1}^{K} w_{k,t} d_{hj,k}^{2}}\right]^{1/(m-1)}}. \tag{9}$$

证明 相对于各个视角的划分隶属度 $u_{ij,t}$ 最小化目标函数式(6), 由于存在一组约束条件

$$\sum_{i=1}^{C} u_{ij,t} = 1,$$

$$u_{ij,t} \in [0,1],\ 1 \leqslant j \leqslant N,\ 1 \leqslant t \leqslant K,$$

可建立如下的拉格朗日函数:

$$L(u_{ij,t}, \lambda_1) = P(U, \hat{Z}, \hat{W}) + \lambda_1\Big(\sum_{i=1}^{C} u_{ij,t} - 1\Big).$$

分别对 $u_{ij,k}$、λ_1 求导并使得导数为零, 得到

$$\frac{\partial L}{\partial \lambda_1} = \sum_{i=1}^{C} u_{ij,t} - 1 = 0,$$

$$\frac{\partial L}{\partial u_{ij,t}} = m\sum_{t=1}^{K} w_{k,t} u_{ij,t}^{m-1} d_{ij,k}^{2} + \lambda_1 = 0.$$

化简后得到

$$u_{ij,t} = \frac{1}{\sum\limits_{h=1}^{C}\left[\dfrac{\sum\limits_{k=1}^{K} w_{k,t} d_{ij,k}^{2}}{\sum\limits_{k=1}^{K} w_{k,t} d_{hj,k}^{2}}\right]^{1/(m-1)}}. \qquad \square$$

3) 问题 P_3 的解决方案由定理3给出.

定理3 假设 $U=\hat{U}$ 和 $Z=\hat{Z}$ 固定, $P(\hat{U}, \hat{Z}, W)$ 最小化的必要条件为

$$w_{k,t} = \frac{\exp\left[\dfrac{-\sum\limits_{i=1}^{C}\sum\limits_{j=1}^{N} u_{ij,t}^{m} d_{ij,k}^{2}}{\gamma}\right]}{\sum\limits_{T=1}^{K}\exp\left[\dfrac{-\sum\limits_{i=1}^{C}\sum\limits_{j=1}^{N} u_{ij,T}^{m} d_{ij,k}^{2}}{\gamma}\right]}. \tag{10}$$

证明 相对于视角权重 $w_{k,t}$ 来最小化目标函数式(6), 由于存在一组约束条件

$$\sum_{t=1}^{K} w_{k,t} = 1,\ w_{k,t} \in [0,1],\ 1 \leqslant k \leqslant K,$$

可以建立如下拉格朗日函数:

$$L(w_{k,t}, \lambda_2) = P(\hat{U}, \hat{Z}, W) + \lambda_2\Big(\sum_{t=1}^{K} w_{k,t} - 1\Big).$$

分别对 $w_{k,t}$、λ_2 求导数并使得其为零, 得到

$$\frac{\partial L}{\partial \lambda_2} = \sum_{t=1}^{K} w_{k,t} - 1 = 0,$$

$$\frac{\partial L}{\partial w_{k,t}} =$$

$$\sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij,k}^{m} d_{ij,k}^{2} + \gamma(\log(w_{k,t}) + 1) + \lambda_2 = 0.$$

进而得到

$$w_{k,t} = \frac{\exp\left[\dfrac{-\sum\limits_{i=1}^{C}\sum\limits_{j=1}^{N} u_{ij,t}^{m} d_{ij,k}^{2}}{\gamma}\right]}{\sum\limits_{h=1}^{K}\exp\left[\dfrac{-\sum\limits_{i=1}^{C}\sum\limits_{j=1}^{N} u_{ij,h}^{m} d_{ij,k}^{2}}{\gamma}\right]}. \qquad \square$$

2.3 算法描述

根据第2.2节推导的参数学习规则, 给出所提出算法的具体步骤如下.

输入:多视角样本集

$$X = \{\text{view}_1, \text{view}_2, \cdots, \text{view}_K\}$$

共 K 个视角, 其中 $\text{view}_k = \{x_{1,k}, x_{2,k}, \cdots, x_{N,k}\}$; 聚类类别 $C(2 \leqslant C \leqslant N)$; 迭代阈值 ε; 模糊指数 m; 迭代次数 l; 参数 γ.

输出: 全局性的模糊划分矩阵 $\bar{U}$, 各视角聚类中心点 $z_{i,k}$, 视角融合权重矩阵 $W = \{w_{k,t}\}$.

Step 1: 随机产生各视角的模糊隶属度矩阵 $u_{ij,t}$ $(1 \leqslant t \leqslant K)$, 随机产生视角融合权重矩阵 $W = \{w_{k,t}\}$.

Step 2: 根据式 (8) 更新各视角下的中心点 $z_{i,k}$.

Step 3: 根据式 (9) 更新各视角下的隶属度 $u_{ij,t}$.

Step 4: 根据式 (10) 更新视角融合权重矩阵 $W = \{w_{k,t}\}$.

Step 5: 如果 $\|P^{l+1} - P^l\| < \varepsilon$, 则算法停止迭代循环, 否则返回 Step 2.

Step 6: 算法收敛后, 输出各视角下的模糊隶属度.

Step 7: 根据 Step 6 所获取的各视角下的模糊隶属度 $u_{ij,t}$, 利用式 (7) 获取具备全局特性的模糊空间划分矩阵 $\bar{U}$.

2.4　时间算法复杂度和收敛性分析

算法的时间复杂度描述如下: 对于本文算法, 其时间复杂度为

$$O(TK + TKNC + TKC).$$

其中: T 为算法迭代的总次数, K 为数据集的视角个数, N 为数据集的大小, C 为类别数. 实际上, 本文算法与经典的 FCM 算法的时间复杂度相差不大.

由相关收敛性理论[20-21]可知, 本文算法也满足 Zangwill 收敛性定理的条件. 类似于许多基于交替迭代的算法 (如 FCM 和 MEC), 本文算法是局部收敛的, 针对不同的初始化收敛于不同的局部最优解.

3　实验分析

为了验证本文方法对多视角聚类任务的有效性, 分别对人工合成数据集和 UCI 标准多视角数据进行实验分析和评估. 为了对本文算法的聚类性能作出合理的评判, 给出与相关聚类算法的性能比较. 实验中采用的聚类算法有双层自动加权聚类算法 TW-k-means[14]、多视角模糊聚类算法 WV-CoFCM[17]、Co-FKM[16]、基于多任务的组合 K-means 算法 (CombKM)[18]、基于多任务学习框架的 LSSMTC 算法[18]和基于样本与特征空间协同聚类的 Co-clustering 算法[19].

采用以下两种评价指标对各类算法的聚类性能进行评估:

1) 归一化互信息 (NMI)[2,22-23], 有

$$\text{NMI} = \frac{\sum_{i=1}^{C}\sum_{j=1}^{C} N_{i,j} \log \frac{N N_{i,j}}{N_i N_j}}{\sqrt{\sum_{i=1}^{C} N_i \log \frac{N_i}{N} \sum_{j=1}^{C} N_j \log \frac{N_j}{N}}}. \tag{11}$$

其中: $N_{i,j}$ 为第 i 个聚类与类 j 的契合程度, N_i 为第 i 个聚类所包含的数据样本量, N_j 为类 j 所包含的数据样本量, N 为整个数据样本的总量大小.

2) 芮氏指标 (RI)[22-23], 有

$$\text{RI} = \frac{f_{00} + f_{11}}{N(N-1)/2}. \tag{12}$$

其中: f_{00} 为数据点具有不同的类标签并且属于不同类的配对点数目, f_{11} 为数据点具有相同的类标签并且属于同一类的配对点数目, N 为整个数据样本的总量大小.

以上两种评价指标的取值范围均为 $[0, 1]$, 取值越接近于 1, 表示算法的聚类性能越好. 在实验部分, 各可调参数的设置采用网格搜索法确定, 具体网格见表 1, 所显示的评价指标均为最优参数下运行 20 次得到的均值和方差.

表 1　参数网格搜索设置

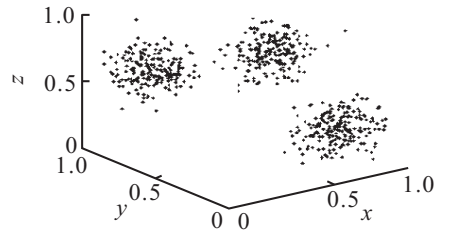
算法	参数寻优范围
LSSMTC	正则化参数 λ : $\{0.25, 0.5, 0.75\}$; 转换空间 $l = d$, 其中 d 为特征数
CombKM	无
Co-clustering	正则化参数 λ : $\{0.1, 1, 10, 100, 500, 1\,000\}$; 正则化参数 μ : $\{0.1, 1, 10, 100, 500, 1\,000\}$; 特征类别 $m = \lfloor d/2 \rfloor$, 其中 d 为特征数, $\lfloor\rfloor$ 表示下取整
TW-k-means	正则化参数 λ : $\{1, 2, \cdots, 30\}$; 正则化参数 η : $\{10, 20, \cdots, 120\}$
WV-CoFCM	模糊指数 m : $\{1.05, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2\}$; 协同学习参数 η : $[0, (K-1)/K]$, 其中 K 为视角数; 正则化参数 λ : $\{1\text{e}^{-7}, 1\text{e}^{-6}, 1\text{e}^{-5}, 1\text{e}^{-4}, 1\text{e}^{-3}, 1\text{e}^{2}, 1\text{e}^{-1}, 1\text{e}^{1}, 1\text{e}^{2}, 1\text{e}^{3}, 1\text{e}^{4}, 1\text{e}^{5}, 1\text{e}^{6}, 1\text{e}^{7}\}$
Co-FKM	模糊指数 m : $\{1.05, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2\}$; 协同学习参数 η : $[0, (K-1)/K]$, 其中 K 为视角数
本文方法	模糊指数 m : $\{1.05, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2\}$; 正则化参数 γ : $\{2^{-12}, 2^{-11}, \cdots, 2^{12}\}$

表 2 各种算法在模拟数据集 D 上的性能比较

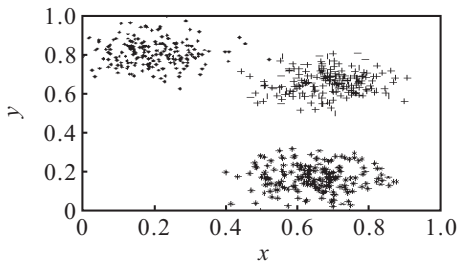
评价指标&算法	CombKM	LSSMTC	Co-clustering	TW-k-means	WV-CoFCM	Co-FKM	本文算法
NMI-mean	0.947 9	0.632 3	0.930 5	0.982 6	1	0.984 7	**1**
NMI-std	0.127 3	0.233 0	0.142 5	0.007 6	0	0.041 0	**0**
RI-mean	0.958 3	0.829 0	0.944 5	0.986 1	1	0.994 2	**1**
RI-std	0.101 7	0.113 2	0.113 9	0.062 1	0	0.016 1	**0**

3.1 人工合成数据集实验

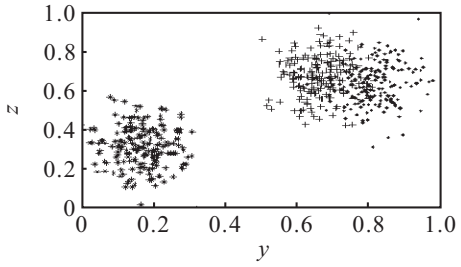
为了充分验证本文方法的效果, 在人工合成数据集部分构建了具有 3 维特性的数据集 D_0, 如图 1(a) 所示. 该数据集包含 3 类数据, 每一类由 200 个样本点构成, 从不同的视角观测数据集 D_0, 得到包含 3 个视角的多视角数据 D, D 的每一视角包含了原始数据的 2 个特征, 所构造的多视角数据在不同视角的分布如图 1(b) ～图 1(d) 所示.



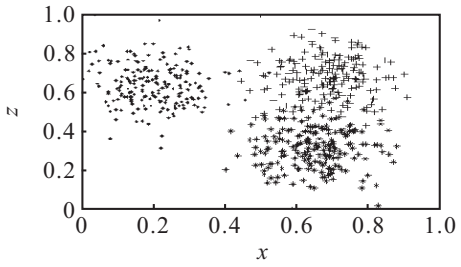
(a) 原始数据集



(b) 视角 1 数据集



(c) 视角 2 数据集



(d) 视角 3 数据集

图 1 多视角模拟数据集 D

将本文算法与其他相关算法在数据集 D 上进行相应的实验, 得到对应的评价指标的数据, 结果如表 2 所示.

对表 2 的结果进行分析可见, 无论是组合任务的 CombKM 算法和多任务的 LSSMTC 算法, 还是基于样本空间与特征空间协同分析的 Co-clustering 算法, 与几种多视角聚类算法相比, 聚类结果均表明几种多视角聚类算法的聚类性能优于非多视角聚类算法的聚类性能. 比较本文算法、Co-FKM、TW-k-means 和 WV-coFCM 可知, 本文算法与 WV-CoFCM 方法在该模拟数据集上的聚类结果相当. 对比各算法在 NMI 和 RI 指标上的均方差可发现, WV-CoFCM 算法和本文算法显示了更好的稳定性. 此外, 本文算法较 CoFKM 和 WV-CoFCM 算法, 不需要人工设置各视角的权重, 各视角间的权重可自适应学习得到.

实验结果表明, 具备划分自适应融合能力的本文算法较需人工设定划分融合程度的 Co-FKM 算法和 WV-CoFCM 算法, 在多视角学习时能够获取更佳的视角权重, 最终得到了更好的多视角聚类效果.

3.2 UCI 真实数据集实验

选择经典机器学习数据库 UCI 中的具备多视角特性的数据集 (MF 数据集、IS 数据集和 WTP 数据集) 来进一步进行性能评估, 3 个数据集的相关信息如表 3 所示. 由于 LSSMTC 算法本身的局限性, 它需要在多视角数据中各视角数据维数相等的情况下才能使用, 因此, LSSMTC 算法无法处理 MF、IS 和 WTP 等各视角维数均不相等的数据样本. 本部分未能给出 LSSMTC 算法在各种数据集上的结果, 其他算法在这些数据集上的性能对比结果见表 4.

对于表 4 的实验结果, 观察各对比算法的聚类性能可知, 基于多视角的 TW-k-means 算法、WV-CoFCM 算法、Co-FKM 算法和本文算法在各多视角真实数据集上, 较其他几种方法均体现出了较好的聚类性能. 同时, 对比本文算法与其他几种多视角的聚类算法, 无论从 NMI 还是 RI 指标来看, 都体现了一定的优势. 虽然在前节的模拟数据集上, 本文算法与 WV-CoFCM 方法相当, 但在表 4 所示的各个真实数据集上, 本文算法显示出更好的聚类性能.

表 3　MF、IS 和 WTP 数据集相关信息介绍

Dataset	View	Composition of Each View	Dimension	Size
Multiple features(MF)	Mfeat-fou view	76 Fourier coefficients of the character shapes	76	2 000
	Mfeat-fac view	216 profile correlations	216	
	Mfeat-kar view	64 Karhunen-Love	64	
	Mfeat-pix view	240 pixel averages in 2*3 windows	240	
	Mfeat-zer view	47 Zernike moments	47	
	Mfeat-mor view	6 morphological variables	6	
Image segmentation(IS)	Shape view	9 features about the shape information of the 7 images	9	2 310
	RGB view	10 features about the RGB values of the 7 images	10	
Water treatment plant(WTP)	Input view	The first 22 features describing different input conditions	22	527
	Output view	The 23th-29th features describing the output demands	7	
	Performance input view	The 30th-34th features describing the performance input demands	5	
	Global performance Input view	The 35th-38th features describing the global performance input demands	4	

表 4　各种算法在 MF、IS 和 WTP 数据集上的性能比较

数据集	算法	NMI		RI	
		mean	std	mean	std
MF	LSSMTC				
	CombKM	0.753 8	0.048 5	0.936 0	0.017 5
	Co-clustering	0.759 4	0.033 7	0.936 7	0.936 7
	TW-k-means	0.837 9	0.044 3	0.946 8	0.024 0
	WV-CoFCM	0.832 0	0.040 5	0.956 4	0.014 0
	Co-FKM	0.829 7	0.029 8	0.956 2	0.011 4
	本文算法	**0.842 1**	**0.029 9**	**0.961 1**	**0.012 0**
IS	LSSMTC				
	CombKM	0.612 0	0.026 3	0.855 1	0.021 5
	Co-clustering	0.610 5	0.019 0	0.864 5	0.012 4
	TW-k-means	0.617 2	0.028 8	0.867 5	0.017 1
	WV-CoFCM	0.624 4	0.020 1	0.879 7	0.010 9
	Co-FKM	0.618 3	0.020 2	0.879 7	0.008 5
	本文算法	**0.625 5**	**0.010 2**	**0.880 4**	**0.006 1**
WTP	LSSMTC				
	CombKM	0.190 0	0.011 9	0.704 7	0.006 1
	Co-clustering	0.192 4	0.013 2	0.704 4	0.004 9
	TW-k-means	0.181 0	0.016 1	0.703 9	0.008 5
	WV-CoFCM	0.196 4	0.012 2	0.707 4	0.004 1
	Co-FKM	0.198 0	0.008 7	0.708 2	0.004 2
	本文算法	**0.201 9**	**0.009 0**	**0.709 3**	**0.002 5**

综上, 无论是在人工还是真实多视角数据集上, 多视角聚类算法较非多视角聚类算法都相对具有优势. 本文提出的新的多视角聚类方法较几种已有的基于视角划分融合的多视角聚类算法显示了进一步的性能优势.

4　结　论

本文在经典 FCM 算法框架上, 通过引入视角划分自适应融合技术, 提出了基于划分自适应融合的多视角模糊聚类算法. 在视角融合权重矩阵的自适应学习下, 使得各个视角之间协调作用更加灵活, 同时各个视角可以自适应学习, 进而达到更好的聚类效果. 实验研究表明, 无论是在模拟数据集还是 UCI 真实数据集上的聚类结果, 均显示出所提出方法具有更好的聚类性能. 虽然方法的有效性得到了有效的验证, 其仍面临一定的考验. 例如, 所提出方法以经典的 FCM 算法为框架, 使用的是欧氏距离, 这使得其在面对高维多视角聚类问题时面临维数灾难问题. 如何解决该问题将会是今后研究的重点. 此外, 类似于多数无监督学习算法, 最优参数的选取是一个重要的问题, 不同的实际应用中需要不同范围的参数值. 由于最优参数通常是由应用决定的, 对于一个确切的应用, 可通过先验知识或可用的有效标记数据集决定所采用算法需要的合适的参数范围, 集成学习等策略可以在一定程度上避免最优参数的选取, 拟在后续工作中作进一步深入研究.

参考文献(References)

[1] Yu S, Tranchevent L C, Liu X H, et al. Optimized data fusion for kernel k-means clustering[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 34(5): 1031-1039.

[2] Jing L P, Ng M K, Huang J Z. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data[J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(8): 1026-1041.

[3] Zhu L, Chung F L, Wang S T. Generalized fuzzy C-means

clustering algorithm with improved fuzzy partitions[J]. IEEE Trans on Systems Man and Cybernetics, 2009, 39(3): 578-591.

[4] Hall L O, Goldgof D B. Convergence of the single-pass and online fuzzy C-means algorithms[J]. IEEE Trans on Fuzzy Systems, 2011, 19(4): 792-794.

[5] 邓赵红, 王士同, 吴锡生, 等. 鲁棒的极大熵聚类算法RMEC及其例外点标识[J]. 中国工程科学, 2004, 4(9): 38-45.
(Deng Z H, Wang S T, Wu X S, et al. Robust maximum entropy clustering algorithm RMEC and its outlier labeling[J]. Engineering Science, 2004, 6(9): 38-45.)

[6] Karayiannis N B. MECA: Maximum entropy clustering algorithm[C]. IEEE Conf on Fuzzy Systems. Orlando: IEEE Press, 1994: 630-635.

[7] Krishnapuram R, Keller J M. A possiblistic approach to clustering[J]. IEEE Trans on Fuzzy Systems, 1993, 1(2): 98-110.

[8] Krishnapuram R, Keller J M. The possiblistic means algorithms: Insights and recommendation[J]. IEEE Trans on Fuzzy Systems, 1996, 4(3): 385-393.

[9] Asur S, Parthasarathy S, Ucar D. An ensemble framework for clustering protein interaction networks[J]. Bioinformatics, 2007, 23(13): 29-40.

[10] Wang H J, Shan H H, Banerjee A. Bayesian cluster ensembles[J]. Statistical Analysis and Data Mining, 2011, 4(1): 54-70.

[11] Bickel S, Scheffer T. Multi-view clustering[C]. The 4th IEEE Int Conf on Data Mining. Brighton: ICDM, 2004: 822-833.

[12] deSa V R. Spectral clustering with two views[C]. Proc of the 24th Int Conf on Machine learning Workshop on Learing. New York: ICML, 2005: 20-27.

[13] Zhou D, Burges C J C. Spectral clustering and transductive learning with multiple views[C]. Proc of the 24th Int Conf on Machine learning. New York: ICML, 2007: 1159-1166.

[14] Chen X, Xu X, Huang J Z, et al. TW-k-means: Automated two-level variable weighting clustering algorithm for multiviewdata[J]. IEEE Trans on Knowledge and Data Engineering, 2013, 25(4): 932-944.

[15] Pedrycz W. Collaborative fuzzy clustering[J]. Pattern Recognition Letter, 2002, 23(14): 1675-1686.

[16] Cleuziou G, Exbrayat M, Martin L, et al. CoFKM: A centralized method for multiple-view clustering[C]. Proc of the 9th IEEE Int Conf on Data Mining. Miami FL: IEEE, 2009: 752-757.

[17] Jiang Y, Chung F L, Wang S, et al. Collaborative fuzzy clustering from multiple weighted views[J]. IEEE Trans on Cybernetics, 2015, 45(4): 688-701.

[18] Gu Q, Zhou J. Learning the shared subspace for multi-task clustering and transductive transfer classification[C]. Proc of the 9th IEEE Int Conf on Data Mining. Miami FL: IEEE, 2009: 159-168.

[19] Gu Q, Zhou J. Co-Clustering on manifolds[C]. Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ICML, 2009: 359-368.

[20] Zangwill W I. Convergence conditions for nonlinear programming algorithms[J]. Management Science, 1969, 16(1): 1-13.

[21] Luenberger D G, Ye Y. Linear and nonlinear programming[M]. Springer Science & Business Media, 2008: 183-214.

[22] Deng Z H, Choi K S, Chung F L, et al. Enhanced soft subspace clustering integrating within-cluster and between-cluster information[J]. Pattern Recognition, 2010, 43(3): 767-781.

[23] 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[J]. 软件学报, 2014, 25(10): 2293-2311.
(Jiang Y Z, Deng Z H, Wang J, et al. Collaborative partition multi-view fuzzy clustering algorithm using entropy weighting[J]. J of Software, 2014, 25(10): 2293-2311.)

(责任编辑：郑晓蕾)