软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(10):2293−2311 [doi: 10.13328/j.cnki.jos.004510]

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

熵加权多视角协同划分模糊聚类算法*

蒋亦樟, 邓赵红, 王 骏, 钱鹏江, 王士同

(江南大学 数字媒体学院,江苏 无锡 214122)

通讯作者: 邓赵红, E-mail: dzh666828@aliyun.com

摘 要: 当前,基于协同学习机制的多视角聚类技术存在如下两点不足:第一,以往构造的用于各视角协同学习的逼近准则物理含义不明确且控制简单;第二,以往算法均默认各视角的重要性程度是相等的,缺少各视角重要性自适应调整的能力.针对上述不足:首先,基于具有良好物理解释性的 Havrda-Charvat 熵构造了一个全新的异视角空间划分逼近准则,该准则能有效地控制异视角间的空间划分相似程度;其次,基于香农熵理论提出了多视角自适应加权策略,可有效地控制各视角的重要性程度,提高算法的聚类性能;最后,基于 FCM 框架提出了熵加权多视角协同划分模糊聚类算法(entropy weight-collaborative partition-multi-view fuzzy clustering algorithm,简称 EW-CoP-MVFCM).在模拟数据集以及 UCI 数据集上的实验结果均显示,所提算法较之已有多视角聚类算法在应对多视角聚类任务时具有更好的适应性.

关键词: 多视角聚类;协同学习;Havrda-Charvat 熵;香农熵;模糊 C 均值聚类

中图法分类号: TP181

中文引用格式: 蒋亦樟,邓赵红,王骏,钱鹏江,王士同.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

英文引用格式: Jiang YZ, Deng ZH, Wang J, Qian PJ, Wang ST. Collaborative partition multi-view fuzzy clustering algorithm using entropy weighting. Ruan Jian Xue Bao/Journal of Software, 2014,25(10):2293−2311 (in Chinese). http://www.jos.org.cn/1000-9825/4510.htm

Collaborative Partition Multi-View Fuzzy Clustering Algorithm using Entropy Weighting

JIANG Yi-Zhang, DENG Zhao-Hong, WANG Jun, QIAN Peng-Jiang, WANG Shi-Tong

(School of Digital Media, Jiangnan University, Wuxi 214122, China)

Corresponding author: DENG Zhao-Hong, E-mail: dzh666828@aliyun.com

Abstract: There are two weaknesses of current multi-view clustering technologies based on collaborative learning. Firstly, the approximation-criteria of collaborative learning between each view is not clear for its physical meaning and is too simple to control the approximation-performance. Secondly, the existing algorithms assume that the significance of each view is equal, which is obviously inappropriate from the viewpoint of adaptively adjusting the importance of each view. In order to overcome the above shortcomings, a novel approximation-criteria of cluster partition based on the Havrda-Charvat entropy is proposed to control the similarity of cluster partition between each view. Then, an adaptive weighting strategy for each view based on the theory of Shannon entropy is presented to control the significance of each view and enhance the performance of the clustering algorithm. Finally, the collaborative partition multi-view fuzzy clustering algorithm using entropy weighting (EW-CoP-MVFCM) is provided. As demonstrated by extensive experiments in simulation data and UCI benchmark dataset, the proposed new algorithm shows the better adaptability than the classical algorithms on the multi-view clustering problems.

* 基金项目: 国家自然科学基金(61170122, 61272210, 61202311, 61300151); 江苏省自然科学基金(BK2009067, BK2012552, BK20130155); 中央高校基本科研业务费专项资金(JUSRP21128, JUDCF13030); 教育部新世纪优秀人才支持计划(NCET-12-0882); 江苏省 2013 年度普通高校研究生科研创新计划(CXZZ13_0760)

收稿时间: 2012-08-23; 定稿时间: 2013-09-27

Key words: multi-view clustering; collaborative learning; Havrda-Charvat entropy; Shannon entropy; fuzzy C-means

在实际生产或生活中经常会遇到同源异构的数据.所谓同源异构数据,即指数据的来源及采样的对象是一致的,但采样的视角(特征空间组成)存在一定的差异性.比如在测量人的血液时,可以从不同的测量指标(特征)切入对人的健康进行分析,目的在于最终从不同视角分析出趋于一致的更可靠的结果.由此诞生了一个新的技术领域——多视角技术,该技术的特色在于:通过对同一对象的不同特征所构造的数据样本进行分析,并利用各视角间的协同又称作交互式的处理模式找出视角间的相似性成分,进而得到一个趋于一致性的全局决策结果.由于该技术全面地考虑了被研究对象在各个视角下所存在的特征信息,因而在求同存异的指导思想下,所得到的决策结果要比传统的仅基于单一视角特征空间所得到的决策结果更为全面、可靠.这也使得近年来,多视角技术在机器学习等领域内得到了广泛的关注与应用,特别是在分类与聚类方面该技术均有着不俗的表现[1−4].本文研究的重点将主要集中于多视角聚类领域.

正如我们熟知的那样,传统的聚类分析方法有 *K*-means[5,6],Fuzzy C-means(FCM)[7,8],MEC[9,10]及 PCM[11,12]等,均是围绕单一视角的聚类分析方法.那么,当上述算法遇到多视角聚类任务时,普遍的做法是先独立地考虑每一视角样本,将各视角样本看成一个独立的聚类任务进行处理,在获取每一视角对应的聚类结果后,再利用集成学习机制[13,14]选择一个合适的集成学习策略将多个聚类结果进行集成,进而得到最终的聚类结果.但此种人为地将每一视角割裂开来进行独立分析的多视角策略,极可能因某一视角下聚类结果存在明显的偏差或各个视角间的聚类结果差异性较大,造成集成学习所获取的全局划分结果较差又或者造成算法的性能不稳定.

因上述问题的存在,经过研究发现:将多视角技术引入传统的聚类分析方法,使得各视角在聚类过程中协同学习,被认为是一个有效的解决方案.近几年,基于上述策略提出了一些有效的多视角聚类方法,比较经典的工作有文献[15−18].其中,文献[15]从概率的角度基于 EM 算法提出了可用于解决多视角问题的协同聚类算法——Co-EM 算法,并通过文本类的数据集证明了该算法的有效性;文献[16]则首次在 FCM 算法中采用了协同聚类的思想,通过对各视角间的模糊划分进行控制,构造了划分协同控制函数,得到了 Co-FC 算法,该算法在多个数据集上均表现出了一定的优势;文献[17]针对高维数据处理较为困难的问题,提出了将高维数据通过一定方法投影到不同的低维空间上,再利用多视角聚类技术对这些低维的数据集进行分析,最终得到一个全局性的划分结果;在 2009 年,文献[18]参考了文献[16]的协同思想,同样以经典 FCM 算法作为基础模型,进一步提出了两种不同的多视角协同划分方法,并依此构造了全新的多视角模糊聚类算法 Co-FKM.

虽然近几年多视角聚类技术得到了一定的发展,特别是基于经典 FCM 框架提出了一些有效的方法,如多视角 Co-FKM 算法[18],但通过对已有多视角聚类算法的研究可以发现,以往的模型普遍存在以下两点不足:

1) 不同视角间的协同学习准则,如空间划分逼近准则,过于简单且物理解释性较弱,这使得该类准则并不能充分探究出各视角间的关系,如相似性程度;
2) 已有的多视角聚类算法通常未能考虑到各视角重要性所存在的差异,只是简单地默认各视角的重要性是均衡的,这使得已有算法的聚类性能及适应能力还有待提高.

针对上述第 2)点不足,需要指出的是,目前各算法所采用的视角重要性均衡化的处理方式与实际应用所面临的情况是不符的.例如,在实际应用中,各视角由于采样的特征空间不同,造成了各视角样本并不一定均具有良好的聚类特性,因而各视角的重要性很自然是不同的.大量的事实表明:类别划分并不明确的特征空间所对应的视角对整个多视角聚类分析任务而言并无太大的帮助,甚至会起负作用,因而在聚类过程中应尽量弱化此种视角的影响力.

针对上述挑战,本文首先基于 Havrda-Charvat 熵构造了全新的异视角空间划分逼近准则,利用熵良好的信息表征特性来控制各视角间的空间划分结果,使得各视角在协同学习时能够找到尽可能多的相似性信息,进而使得各视角的空间划分结果尽量一致,以获取一个较为稳定且更为全面的全局性空间划分结果.进一步地,本文引入香农熵和可能性概率权值对各视角进行了自适应加权处理,从而在聚类过程中有效地调节各视角间的权重关系,使得空间划分最为明确的视角所占的比重尽可能地大,同时使空间划分较模糊的视角所占比重尽可能

地小,最终获取最优的空间划分结果.特别地,根据最终获取的概率权重,可以得到针对该多视角任务的最佳聚类视角,即,权重最大值所对应的视角.

通过引入上述技术,本文在经典 FCM 框架下探讨了具体的多视角聚类分析新算法,即,熵加权多视角协同划分模糊聚类算法(entropy weight-collaborative partition-multi-view fuzzy clustering algorithm,简称 EW-CoP-MVFCM).由于该新目标函数在保证各视角样本空间划分趋于一致的前提下,更进一步凸显了最佳聚类视角的作用,从而使得其获取的空间划分结果较之以往的算法更为准确、合理.因而,该算法对整个聚类任务而言获取的结果具备更好的全局决策价值,同时,最佳视角的获取也为后期对研究聚类对象的全局性空间划分决策提供了一种新的集成策略.因此,通过上述两大机制进行多视角协同聚类,将确保本文算法在应对现实世界复杂多变的聚类任务时较之以往方法具有更好的数据分析与决策能力,亦增强了算法的适应性.

1 相关工作

在数据挖掘领域内,经常会面对同源异构的数据样本,即,具备多视角特性的样本集合,对于此类样本的聚类任务,一般可采用以下两种策略:

1) 利用经典的单视角聚类算法,如 FCM 算法,对每一个视角样本进行单独的聚类分析,而后,通过集成学习技术将每一视角对应的模糊划分结果经过某种集成策略形成一个全局性的决策解;
2) 利用当前较流行的多视角聚类算法,如文献[18]中提出的基于模糊聚类技术的 Co-FKM 算法,该算法通过所构造的不同视角间的空间逼近准则,使得每一视角在单独聚类过程中还需考虑其他视角的影响,通过找出各视角间的相似性,以使聚类结果尽可能地保持一致,最终根据集成学习策略得到一个全局性的决策解.

在上述两种多视角聚类技术方案中,对于最终的全局空间划分结果均依赖于集成学习机制,该技术的目的在于,选定某种集成规则,将各视角下的聚类结果有效集成以获取全局性的结果.

另外,与多视角具有一定相似性的聚类技术还有如下几个方向.

1) 多任务聚类

该方法将多个具有一定关联性的聚类任务通过多任务学习框架进行协同式学习,经典的工作有文献[19]中的 LSSMTC 算法.若我们将各视角看作单一任务,则在一些情况下可使用多任务聚类技术来处理多视角聚类任务.然而,这两类聚类技术在本质上是存在差异的:多任务聚类算法考虑的是任务之间的相似性,不同任务对应不同的聚类对象;而多视角聚类是不同视角对应相同的聚类对象,这使得多任务聚类技术用于多视角聚类通常不能得到理想的效果.另外,在多任务聚类中,对于各任务的样本特征数(维数)通常需要保持一致,这使得在面对各视角样本维数不一致的聚类任务时,该方法无法使用.

2) 组合聚类

该方法源于多任务的思想,它通过将多个任务组合成一个任务进行整体性的空间划分.较为有效的方法有文献[19]中的组合 *K*-means 算法(CombKM),该算法的思想同样也适用于多视角聚类,在利用该算法对多视角数据进行聚类分析时,考虑到多视角样本本质上是同源不同特征的组合形式,因此只需将这些不同特征下的样本合并成一个整体样本即可.但如此做法必定会破坏各视角的独立性特征,进而影响最终的空间划分结果.

3) 基于样本与特征空间的协同聚类

该方法在考虑对样本聚类的同时还引入了对特征划分的考虑,而对特征的划分结果本质上对应于多视角样本中每一视角对应的特征组合,该领域内较成功的工作有文献[20]中的 Co-clustering 算法.在使用该算法处理多视角聚类任务时,同样采用组合聚类的策略,先合并各视角样本再进行聚类分析.由于该算法采用了与组合聚类相同的策略,虽然其考虑了特征空间的划分,但该划分相对比较粗糙,因此利用该算法处理多视角聚类任务时效果也不甚理想.此外,该算法在计算时采用了矩阵求逆的运算,这使其在处理大样本或高维数据时常常会受到硬件条件的制约而不再适用.

除了如上的相关工作,本文将重点介绍与分析两种与本文算法有着紧密关联的多视角模糊聚类算法.首先

介绍单视角 FCM 算法和集成学习策略相结合的多视角模糊聚类技术,然后介绍一种经典的基于协同学习技术的多视角聚类方法 Co-FKM.

1.1 单视角FCM算法处理多视角问题

给定数据样本 $\boldsymbol{X}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}\in R^{N\times D}$($N$ 为样本总量,D 为样本维数),$\boldsymbol{U}=[\mu_{ij}]\in R^{C\times N}$ 为样本 $\boldsymbol{X}$ 上的模糊划分矩阵(C 表示所需聚类的类别总数),$\boldsymbol{V}=[\boldsymbol{v}_i]\in R^{C\times D}$ 为样本的类中心.

根据上述定义,经典的 FCM 算法目标函数可表示为如下形式:

$$\begin{cases} J_{\mathrm{FCM}}(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}^{m}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2 \\ \text{s.t. } \mu_{ij}\in[0,1] \text{ and } \sum_{i=1}^{C}\mu_{ij}=1, 1\leqslant j\leqslant N \end{cases} \tag{1}$$

其中,$\boldsymbol{v}_i=(v_{i1},\ldots,v_{ik})$为第 i 类的中心点;μ_{ij} 表示第 j 个样本属于 i 类的隶属度,其中,模糊指数 m 必须满足 $m>1$;$\boldsymbol{x}_j$ 表示第j个样本点.根据相关的优化理论,通过拉格朗日的极值求解方法可以得到类中心 $\boldsymbol{v}_i$ 以及隶属度μ_{ij}的迭代表达式:

$$\boldsymbol{v}_i=\frac{\sum_{j=1}^{N}\mu_{ij}^{m}\boldsymbol{x}_j}{\sum_{j=1}^{N}\mu_{ij}^{m}} \tag{2}$$

$$\mu_{ij}=\frac{1}{\sum_{k=1}^{C}\left[\frac{\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2}{\|\boldsymbol{x}_j-\boldsymbol{v}_k\|^2}\right]^{\frac{1}{m-1}}} \tag{3}$$

根据以上两式迭代优化终止后,所获取的隶属度矩阵 $\boldsymbol{U}$ 在去模糊化之后得到空间划分矩阵,根据该矩阵可获取每一个样本 $\boldsymbol{x}_j$ 所对应的类别信息.

通过上述优化策略最终可以获得模糊划分矩阵 $\boldsymbol{U}$,那么在面对多视角聚类任务时,若不考虑各视角间的相似性,则最直接的方法,即,将每一视角对应的样本集代入到目标函数公式(1)中得到各自的空间划分矩阵 $\boldsymbol{U}_k$ ($1<k<K$),k 表示视角个数.然而,为了得到全局性的决策解,此时需要引入集成学习技术,将各视角下的$[\boldsymbol{U}_1,\ldots,\boldsymbol{U}_K]$整合成一个具有全局描述能力的空间划分矩阵 $\tilde{\boldsymbol{U}}$. 该种方法的工作原理如图 1 所示.

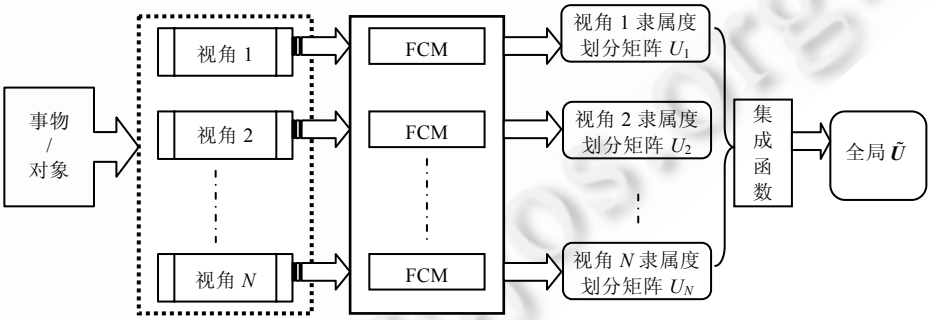


Fig.1 Principle of processing the multi-view clustering task by using the traditional FCM algorithm

图 1 传统 FCM 算法处理多视角聚类任务的工作原理图

根据图 1 所示之工作原理可以发现,该多视角聚类技术因人为地将各视角割裂开来进行独立的聚类分析,未能考虑到聚类过程中各视角间的协同学习(挖掘相似性成分).虽然最后引入了集成学习,策略得到了全局性的空间划分结果,但是由于各视角并无协同学习,其视角与视角之间的关联性(相似性)未得到合理的运用,这极易造成最终获取的全局性空间划分结果因某一视角存在严重误分而导致算法性能的下降.

1.2 Co-FKM算法

为了解决单一视角算法面对多视角聚类任务时遇到的难题,文献[18]以经典 FCM 算法作为多视角聚类模型构建的基础,提出了多视角模糊聚类算法(Co-FKM).该算法的主要贡献在于:通过引入协调各视角间空间划分的隶属度约束项 $\Delta_k=\eta\frac{1}{K-1}\sum_{\substack{k'=1,\\k'\neq k}}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij,k'}^m-\mu_{ij,k}^m)\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{i,k}\|^2$,保证在算法收敛时得到的各视角空间划分结果尽可能地一致,从而实现聚类过程中各视角的协同学习.Co-FKM 算法的目标函数可表示如下:

$$J_{\text{Co-FKM}}(\boldsymbol{U},\boldsymbol{V})=\sum_{k=1}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}[\mu_{ij,k}^m\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{i,k}\|^2+\eta\Delta_k] \tag{4-1}$$

$$\begin{cases}\Delta_k=\eta\dfrac{1}{K-1}\displaystyle\sum_{k'=1,k'\neq k}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij,k'}^m-\mu_{ij,k}^m)\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{i,k}\|^2\\ \text{s.t. }\mu_{ij,k}\in[0,1]\text{ and }\displaystyle\sum_{i=1}^{C}\mu_{ij,k}=1,1\leqslant j\leqslant N,1\leqslant k\leqslant K\end{cases} \tag{4-2}$$

将公式(4-2)代入公式(4-1),经过化简最终可得到如下的目标函数:

$$J_{\text{Co-FKM}}(\boldsymbol{U},\boldsymbol{V})=\sum_{k=1}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}[\tilde{\mu}_{ij,k,\eta}\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{i,k}\|^2] \tag{5}$$

公式(5)中,$\tilde{\mu}_{ij,k,\eta}=(1-\eta)\mu_{ij,k}^m+\frac{\eta}{K-1}\sum_{k'=1,k'\neq k}^{K}\mu_{ij,k'}^m$,其中,参数$\eta$为调控各视角隶属度划分的协同学习参数,而$\tilde{\mu}_{ij,k,\eta}$则表示为当前视角下的隶属度$\mu_{ij,k}^m$与其余各视角隶属度$\mu_{ij,k'}^m$之间的加权平均值.

根据经典FCM算法的优化策略,同样通过拉格朗日极值求解方法,可以得到隶属度$\mu_{ij,k}$以及中心点$\boldsymbol{v}_{i,k}$的优化迭代公式如下:

$$\boldsymbol{v}_{i,k}=\frac{\sum_{j=1}^{N}\tilde{\mu}_{ij,k,\eta}\boldsymbol{x}_{j,k}}{\sum_{j=1}^{N}\tilde{\mu}_{ij,k,\eta}},i=1,2,...,C \tag{6}$$

$$\mu_{ij,k}=\frac{1}{\displaystyle\sum_{h=1}^{C}\left[\frac{(1-\eta)d_{ij,k}^2+\frac{\eta}{K-1}\sum_{\substack{k'=1,\\k'\neq k}}^{K}d_{ij,k'}^2}{(1-\eta)d_{hj,k}^2+\frac{\eta}{K-1}\sum_{\substack{k'=1,\\k'\neq k}}^{K}d_{hj,k'}^2}\right]^{\frac{1}{m-1}}},i=1,2,...,C;j=1,2,...,N \tag{7}$$

在上述迭代策略的基础上,最终可以获取各视角下对应的模糊隶属度划分矩阵.那么,为了求得一个具备全局性考量的模糊划分标准,文献[18]利用各视角下所获取的模糊隶属度的几何平均值来体现整体的划分结果,其具体的表达如下所示:

$$\hat{\mu}_{ij}=\sqrt[K]{\prod_{k\in K}\mu_{ij,k}} \tag{8}$$

利用式(8)所得结果,在去模糊化之后,即可获取用以表征整个事物/对象的空间划分结果,为后期进一步对事物进行决策提供了有力的空间划分参考.

多视角模糊聚类算法的工作机制如图 2 所示.

虽然上述算法成功地在每一视角的聚类过程中融入了不同视角间的空间划分逼近准则,实现了各视角间的协同学习,使得该算法在面对多视角聚类任务时较之传统单视角集成聚类技术展示出了更为有效的聚类性能,但其依然存在诸多问题,其中迫切需要改进的即为本文引言部分所述的两点不足.本文将针对这两个不足,在下一节给出一种全新的多视角模糊聚类方法.

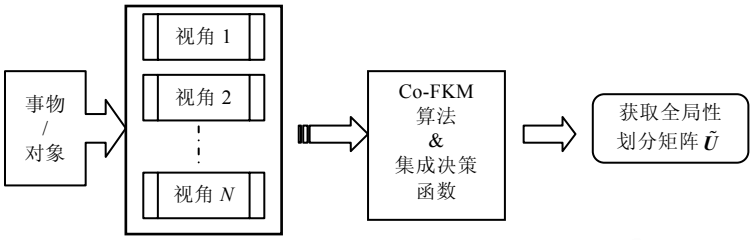


Fig.2 Principle of processing the multi-view clustering task by using Co-FKM algorithm

图 2 Co-FKM 算法处理多视角聚类任务的工作原理图

2 熵加权多视角协同划分模糊聚类算法(EW-CoP-MVFCM)

针对前面所述的当前多视角聚类方法所存在的两点不足,本文基于熵理论引入如下两大新技术以对当前方法进行有效改进.

- 首先,提出了基于 Havrda-Charvat 熵的异视角空间划分逼近准则.如我们所知,在信息论中,熵理论经常被用来描述随机变量的不确定性,而模糊聚类算法中的模糊隶属度 μ_{ij}^m 也被用以描述样本点 x_j 归属的不确定性.因此,本文选用 Havrda-Charvat 熵构造出新的异视角空间划分逼近准则,试图利用熵理论寻找出各视角间的最大相似性成分,从而在提高算法性能的同时也从熵的角度给予了异视角空间划分逼近准则新的物理意义;
- 其次,提出了基于香农熵的多视角自适应加权策略.鉴于以往的算法对视角的重要性程度存在弱化的问题,本文提出了熵加权的多视角聚类技术,通过给每一个视角加以权重,从而在迭代优化过程中自适应地找出最佳视角,同时获取最优的模糊划分结果.为了对该权重进行有效调控,本文引入香农熵理论对各视角的重要性程度进行掌控.

通过上述两大改进策略的共同作用,即得到了本文所提到的熵加权多视角协同划分模糊聚类算法(entropy weight-collaborative partition-multi-view fuzzy clustering algorithm,简称 EW-CoP-MVFCM),该算法具有更好的样本适应性,相应的工作原理如图 3 所示.另外,在第 2.1 节、第 2.2 节还将分别对上述两大策略的构造及其物理意义进行详细的解释与说明.

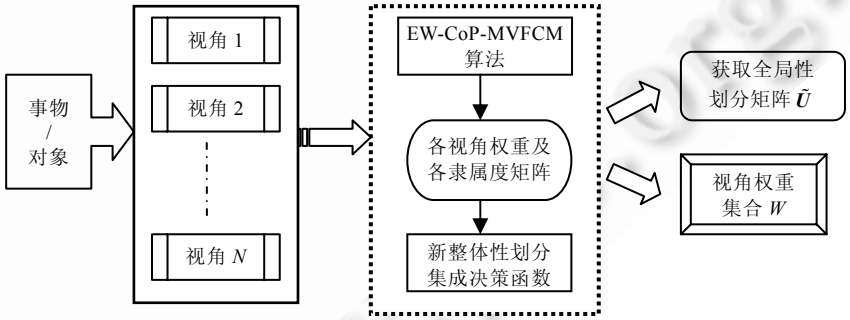


Fig.3 Principle of processing the multi-view clustering task by using EW-CoP-MVFCM algorithm

图 3 EW-CoP-MVFCM 算法处理多视角聚类任务的工作原理图

2.1 基于Havrda-Charvat熵的异视角空间划分逼近准则

由于现有的多视角聚类算法对模糊隶属度 μ_{ij}^m 的控制过于简单且控制项的物理含义也不明确,因此,本文利用熵常被用以描述随机变量的不确定性程度的性质,并根据模糊聚类算法中模糊隶属度 μ_{ij}^m 也被用以描述

样本点$\boldsymbol{x}_j$隶属于第i类空间的模糊程度的理论,自然地将熵理论引入到模糊聚类算法中,用来控制并更好地解释模糊隶属度μ_{ij}^m.经过上述的研究与分析,本文构造了基于第k视角下的样本点$\boldsymbol{x}_{j,k}$的 Havrda-Charvat 熵[21],其具体定义如下:

$$I^{(m)}(\boldsymbol{x}_{j,k})=\frac{1}{1-2^{1-m}}\sum_{i=1}^{C}(\mu_{ij,k}^m-1) \tag{9}$$

很明显,若将模糊隶属度矩阵$\boldsymbol{U}_k=[\mu_{ij,k}^m]$看作概率矩阵,那么当满足$\sum_{i=1}^{C}\mu_{ij,k}=1$的约束条件时,$I^{(m)}(\boldsymbol{x}_{j,k})$的值为 0.这直观地说明了该视角的样本集$\boldsymbol{X}_k$中的样本点$\boldsymbol{x}_{j,k}$隶属于各个划分的不确定性达到最小;也即说明,当目标函数取得最优解时,该视角下的隶属度$\mu_{ij,k}$的 Havrda-Charvat 熵相应地也取得最小值.同时,样本集$\boldsymbol{X}_k$中的各样本点所具有的 Havrda-Charvat 熵总量亦达到最小.

虽然公式(9)可以保证划分的不确定性达到最小,但还只限于单视角框架,为了将其拓展至多视角领域,本文参照文献[18]的相关策略,将公式(9)改造为如下形式:

$$H=\frac{1}{1-2^{1-m}}\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij,k}^m-1)+\eta\frac{1}{1-2^{1-m}}\frac{1}{K-1}\sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k'=1,k'\neq k}^{K}[(\mu_{ij,k'}^m-1)-(\mu_{ij,k}^m-1)] \tag{10-1}$$

将上式整理后可得:

$$H=(1-\eta)\frac{1}{1-2^{1-m}}\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij,k}^m-1)+\eta\frac{1}{1-2^{1-m}}\frac{1}{K-1}\sum_{k'=1,k'\neq k}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij,k'}^m-1) \tag{10-2}$$

观察公式(10-2)可以发现:在保证相异的视角空间(即对应于每一个视角)下隶属度μ_{ij}对应的 Havrda-Charvat 熵达到最小值的前提下,通过参数η的取值,可以有效调控当前视角与其余各视角隶属度划分的权重关系(η需满足 $0<\eta<1$ 且一般取$\eta=(K-1)/K$),从而得到当前视角下的隶属度加权平均值.该策略能够促使各视角的隶属度划分尽可能地趋向一致,从而获取更具全局性的空间划分结果.

2.2 基于香农熵的多视角自适应加权策略

本节针对现有的多视角聚类算法中,几乎默认了同一种前提假设,即,每一视角对最终聚类结果的贡献是均等的.实际上,由于某些视角往往存在严重的空间重叠现象而不具可分性,对于这些视角而言,若在整体的聚类过程中给予其与其他具备较好可分性的视角同等的重要性程度,而这显然是不合乎逻辑的.合理的处理方式应是:根据视角的可划分特性给予其不同的重要性程度,这样所得到的整体性聚类结果才能达到最佳.但人为地给定重要性权重显然也是不合理的,为此,本文基于香农熵理论[9,22]提出了具备多视角技术特征的自适应加权项,再引入视角权重系数w_k,并根据$\sum_{k=1}^{K}w_k=1$且$w_k\geqslant 0$的条件重新构造了目标函数式.

此外,为了对视角权重系数w_k加以调控,本文在$\sum_{k=1}^{K}w_k=1$且$w_k\geqslant 0$的条件下,将视角权重看作概率分布,则可用香农熵表示为

$$f(w_k)=-\sum_{k=1}^{K}w_k\ln w_k \tag{11}$$

通过上述手段引入熵技术,使得目标函数在达到最优的同时熵尽可能地大,这也便是经典的极大熵原理.而熵的极大化又会导致各概率分量尽可能地趋于均衡[9],而目标函数的极大化则使得视角权重更加趋向于某一分量即某一视角,从而将最具聚类效果(空间划分最为鲜明)的视角凸显出来.由于在各个视角上引入了自适应极大熵加权的概念,这使得该算法能够有效地降低那些聚类特性较差的视角对算法优化迭代时的干扰,从而获取更为理想的空间划分结果,同时也增强了算法的有效性.

2.3 EW-CoP-MVFCM算法

根据第 2.1 节和第 2.2 节有关异视角空间划分逼近准则及多视角自适应加权策略的定义,本文仍以经典 FCM 算法框架为基础,重新构造了具备多视角聚类能力的目标函数如下:

$$\begin{cases} J_{\text{EW-CoP-MVFCM}}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})=\sum_{k=1}^{K}w_k\left[\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij,k}^{m}\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{i,k}\|^2+\lambda_1\Theta_k\right]+\lambda_2\sum_{k=1}^{K}w_k\ln w_k \\ \Theta_k=(1-\eta)\dfrac{1}{1-2^{1-m}}\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij,k}^{m}-1)+\eta\dfrac{1}{1-2^{1-m}}\dfrac{1}{K-1}\sum_{\substack{k'=1,\\k'\neq k}}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij,k'}^{m}-1) \\ \text{s.t. } \mu_{ij,k}\in[0,1],\sum_{i=1}^{C}\mu_{ij,k}=1 \text{ and } \sum_{k=1}^{K}w_k=1,1\leqslant j\leqslant N,1\leqslant k\leqslant K \end{cases} \tag{12}$$

上式中,共包含两个部分.

- 第 1 部分为基于 Havrda-Charvat 熵理论的多视角协同聚类部分:

$$T_1(\boldsymbol{U},\boldsymbol{V},\boldsymbol{X})=\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij,k}^{m}\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{i,k}\|^2+\lambda_1\Theta_k,$$

  其本质是:通过异视角空间划分逼近准则找出各视角间尽可能多的相似性成分,最终使得各视角的空间划分结果趋于一致;

- 第 2 部分为基于极大熵理论的自适应视角加权部分:

$$T_2(\boldsymbol{U},\boldsymbol{V},\boldsymbol{X},\boldsymbol{W})=\sum_{k=1}^{K}w_kT_1(\boldsymbol{U},\boldsymbol{V},\boldsymbol{X})+\sum_{k=1}^{K}w_k\ln w_k,$$

  利用此部分,可自适应调控各视角在多视角聚类过程中的权重关系,最终在算法达到最优时,根据视角的权重矩阵 $\boldsymbol{W}$ 进而获得最佳视角划分结果.

其中,协同学习参数η的取值参照文献[18],取$\eta=(K-1)/K$,而正则化平衡因子λ_1 及λ_2 的取值则参照文献[23],利用网格寻优获取.

同样,为获取具备全局特性的空间划分结果,本文摒弃了文献[18]中的集成策略,即公式(8),根据新获取的视角划分权重矩阵 $\boldsymbol{W}=[w_1,w_2,\ldots,w_K]$重新定义了新的集成学习方法,即全局性加权视角空间划分集成法,具体形式如下:

$$\tilde{\boldsymbol{U}}=\sum_{k=1}^{K}w_k\boldsymbol{U}_k \tag{13}$$

因最终迭代优化终止时,最佳视角所占的比重往往较大(最佳视角即最易划分的视角),因此,由此得到的空间划分结果较之以往的算法更为理想.

2.3.1 各参数优化推导

本节将利用经典的优化理论,通过拉格朗日极值求解的相关策略,对目标函数公式(12)进行优化,为了得到目标函数的最优解,本节给出如下几个定理.

定理 1. 在给定第 k 个视角的隶属度划分矩阵 $\boldsymbol{U}_k$ 以及视角权重矩阵 $\boldsymbol{W}_k$ 后,目标函数公式(12)取得极值时应满足如下必要条件:

$$\boldsymbol{v}_{i,k}=\frac{\sum_{j=1}^{N}\mu_{ij,k}^{m}\boldsymbol{x}_{j,k}}{\sum_{j=1}^{N}\mu_{ij,k}^{m}} \tag{14}$$

证明:利用给定的第 k 个视角的隶属度划分矩阵 $\boldsymbol{U}_k$ 以及视角权重矩阵 $\boldsymbol{W}_k$,对目标函数 $J_{\text{EW-CoP-MVFCM}}(V_k)$求偏导,并令 $\dfrac{\partial J_{\text{EW-CoP-MVFCM}}}{\partial V_k}=0$,可得 $\boldsymbol{v}_{i,k}=\dfrac{\sum_{j=1}^{N}\mu_{ij,k}^{m}\boldsymbol{x}_{j,k}}{\sum_{j=1}^{N}\mu_{ij,k}^{m}}$.至此,定理 1 得证.证毕. □

定理 2. 在给定第 k 个视角的类中心矩阵 $\boldsymbol{V}_k$ 以及视角权重矩阵 $\boldsymbol{W}_k$ 后,目标函数公式(12)取得极值时应满足如下必要条件:

$$\mu_{ij,k}=\frac{1}{\left[\sum_{h=1}^{C}\frac{\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{i,k}\|^2+\frac{1-\eta}{1-2^{1-m}}\lambda_1}{\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{h,k}\|^2+\frac{1-\eta}{1-2^{1-m}}\lambda_1}\right]^{\frac{1}{m-1}}},\ i=1,2,...,C;j=1,2,...,N \tag{15}$$

证明:利用给定第 k 个视角的类中心矩阵 $\boldsymbol{V}_k$ 以及视角权重矩阵 $\boldsymbol{W}_k$,根据约束条件 $\sum_{i=1}^{C}\mu_{ij,k}=1$,引入拉格朗日乘子 $\alpha_{j,k}$,得到相应的优化目标函数式如下:

$$J_{\text{EW-CoP-MVFCM}}(\mu_{ij,k},\alpha_{j,k})=\sum_{k=1}^{K}w_k\left[\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij,k}^{m}\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{i,k}\|^2+\lambda_1\Theta_k\right]+\lambda_2\sum_{k=1}^{K}w_k\ln w_k-\sum_{k=1}^{K}\sum_{j=1}^{N}\alpha_{j,k}\left(\sum_{i=1}^{C}\mu_{ij,k}-1\right) \tag{16}$$

上式中,$\Theta_k=(1-\eta)\frac{1}{1-2^{1-m}}\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij,k}^{m}-1)+\eta\frac{1}{1-2^{1-m}}\frac{1}{K-1}\sum_{k'=1,k'\neq k}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij,k'}^{m}-1)$.

分别对 $\mu_{ij,k},\alpha_{j,k}$ 求偏导,并令偏导数为 0.

根据 $\frac{\partial J_{\text{EW-CoP-MVFCM}}}{\partial\mu_{ij,k}}=0$ 可得:

$$\mu_{ij,k}=\left[\frac{\alpha_{j,k}}{w_k\left(m\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{i,k}\|^2+\frac{1-\eta}{1-2^{1-m}}\lambda_1 m\right)}\right]^{\frac{1}{m-1}},i=1,2,...,C;j=1,2,...,N \tag{17}$$

令 $\frac{\partial J_{\text{EW-CoP-MVFCM}}}{\partial\alpha_{j,k}}=0$ 可得:

$$\sum_{i=1}^{C}\mu_{ij,k}=1 \tag{18}$$

联立公式(17)及公式(18),消去 $\alpha_{j,k}$,即可得到:

$$\mu_{ij,k}=\frac{1}{\left[\sum_{h=1}^{C}\Sigma\frac{\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{i,k}\|^2+\frac{1-\eta}{1-2^{1-m}}\lambda_1}{\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{h,k}\|^2+\frac{1-\eta}{1-2^{1-m}}\lambda_1}\right]^{\frac{1}{m-1}}}. \qquad \square$$

定理 3. 在给定第 k 个视角的类中心矩阵 $\boldsymbol{V}_k$ 以及隶属度划分矩阵 $\boldsymbol{U}_k$ 后,目标函数公式(12)取得极值时应满足如下必要条件:

$$w_k=\frac{\exp\left(\frac{-\left[\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij,k}^{m}\|\boldsymbol{x}_{j,k}-\boldsymbol{v}_{i,k}\|^2+\lambda_1\Theta_k\right]}{\lambda_2}\right)}{\sum_{k'=1}^{K}\exp\left(\frac{-\left[\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij,k'}^{m}\|\boldsymbol{x}_{j,k'}-\boldsymbol{v}_{i,k'}\|^2+\lambda_1\Theta_{k'}\right]}{\lambda_2}\right)},i=1,2,...,C;j=1,2,...,N \tag{19}$$

证明:仿照定理 2 的证明过程,利用拉格朗日极值求解方法及约束项 $\sum_{k=1}^{K}w_k=1$, 即可证得定理 3. $\square$

2.4 EW-CoP-MVFCM算法描述

根据第 2.3 节相关优化理论及迭代公式推导所获取的参数学习规则,本节将给出 EW-CoP-MVFCM 算法的

具体步骤如下:

输入:多视角样本集 $View=\{view_1,\ldots,view_K\}$共 K 个视角($1\leqslant k\leqslant K$),而任意一个视角对应的样本集为 $view_k=\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N\}$,聚类数目 $C(2\leqslant C<N)$,迭代阈值ε,模糊指数 m,迭代次数f,最大迭代次数 L,协同学习参数η,正则化平衡因子λ_1,λ_2;

输出:各视角聚类中心点 $\boldsymbol{v}_{i,k}$,全局性的模糊划分矩阵 $\boldsymbol{U}$,各视角权重 w_k.

Step 1. 初始化随机产生中心点集 $\boldsymbol{v}_i$,随机产生归一化的模糊隶属度矩阵μ_{ij},随机产生归一化的视角权重 w_k;

Step 2. 根据公式(14)更新各视角下的中心点 $\boldsymbol{v}_{i,k}$;

Step 3. 根据公式(15)更新各视角下的隶属度$\mu_{ij,k}$;

Step 4. 根据公式(19)更新各视角的权重 w_k;

Step 5. 如果$\|J^{k+1}-J^k\|<\varepsilon$或者 $f>K$,则算法结束,跳出循环;否则,跳回 Step 2;

Step 6. 算法收敛后,输出各视角下的聚类中心点 $\boldsymbol{v}_{i,k}$、各视角下的模糊隶属度$\mu_{ij,k}$ 以及各视角权重 w_k;

Step 7. 根据 Step 6 所获取的各视角权重 w_k 及各视角下的模糊隶属度$\mu_{ij,k}$,根据公式(13)可获取具备全局特性的模糊空间划分矩阵 $\tilde{\boldsymbol{U}}$. 另外,根据 $\max(\boldsymbol{W})$可找出最佳聚类视角.

3 EW-CoP-MVFCM 全局收敛性分析

在文献[24]中,经典的 FCM 算法的收敛性早已得到证明.而本文提出的 EW-CoP-MVFCM 算法虽然依旧采用经典 FCM 算法的结构框架,但由于引入了具有多视角特性的异视角空间划分逼近准则以及多视角自适应加权策略,使得该算法的收敛性有必要进行再次的证明与分析.

为了对本文算法的收敛性进行判断,我们根据文献[25,26]的相关收敛性理论可知,若需对一种算法的全局收敛性进行判断,只需确定该算法是否满足 Zangwill 收敛性定理的 3 个充要条件即可.为了便于后续进一步对本文算法的收敛性进行判断,这里首先给出 Zangwill 的收敛性定理如下.

引理 1(Zangwill 收敛性定理). 假设点 $\boldsymbol{z}^{(0)}\in V$,由点-集映射 $A:V\to P(V)$定义的迭代算法产生迭代序列$\{\boldsymbol{z}^{(t)}\}$,解集$\Omega\subset V$.如果:

(1) $\{\boldsymbol{z}^{(t)}\}\subset\Gamma\subseteq V$,其中,$\Gamma$表示为一个紧集;

(2) 存在连续函数 $f:V\to\boldsymbol{R}$,满足:

(a) 若 $\boldsymbol{z}\notin\Omega$,则对于任意 $\boldsymbol{y}\in A(\boldsymbol{z})$,$f(\boldsymbol{y})<f(\boldsymbol{z})$;

(b) 若 $\boldsymbol{z}\in\Omega$,算法终止,或对于任意 $\boldsymbol{y}\in A(\boldsymbol{z})$,$f(\boldsymbol{y})\leqslant f(\boldsymbol{z})$;

(3) 若 $\boldsymbol{z}\notin\Omega$,映射 A 在 $\boldsymbol{z}$ 处是闭的,

则算法终止于解集Ω,或$\{\boldsymbol{z}^{(t)}\}$的任一收敛子序列的极限是解集Ω的一个点.

满足上述引理 1 中条件(2)的函数 f 被称为是 A 在 V 上的 Zangwill 函数.

3.1 EW-CoP-MVFCM全局收敛性证明

本节将主要对 EW-CoP-MVFCM 算法的收敛性进行验证,同时为了方便证明,我们首先给出如下几个定义及相关的一些定理,具体描述可参见下文.

定义 1. 设 $A:V\to P(V)$表示为一个点-集映射,那么在点$\boldsymbol{x}^*\in\boldsymbol{X}$处,若有$\{\boldsymbol{x}^{(t)}\}\subset\boldsymbol{X}$,$\boldsymbol{x}^{(t)}\to\boldsymbol{x}^*$,$\boldsymbol{y}^{(t)}\in A(\boldsymbol{x}^{(t)})$,$\boldsymbol{y}^{(t)}\to\boldsymbol{y}^*$,使得$\boldsymbol{y}^*\in A(\boldsymbol{x}^*)$,则可称点-集映射 A 在$\boldsymbol{x}^*\in\boldsymbol{X}$处是闭的;如果它在$\boldsymbol{X}$中的每一点是闭的,则称点-集映射 A 在$\boldsymbol{X}$上是闭的.

定义 2a. 在 k 视角下,定义映射关系 $G_{1,k}:\boldsymbol{M}_{W_k}\times\boldsymbol{M}_{U_k}\to\boldsymbol{R}^{cd_k}$:

$$G_{1,k}(W_k,\boldsymbol{U}_k)=\boldsymbol{V}_k=(\boldsymbol{v}_{1,k},\boldsymbol{v}_{2,k},\ldots,\boldsymbol{v}_{c,k})^T \tag{20}$$

其中有类中心 $\boldsymbol{v}_{i,k}=(v_{i1,k},v_{i2,k},\ldots,v_{id,k})^T\in\boldsymbol{R}^d$,$i=1,2,\ldots,c$,可通过公式(14)求得.

定义 2b. 在 k 视角下,定义映射关系 $G_{2,k}:\boldsymbol{M}_{W_k}\times\boldsymbol{R}^{cd_k}\to\boldsymbol{M}_{U_k}$:

$$G_{2,k}(W_k,\boldsymbol{V}_k)=\boldsymbol{U}_k \tag{21}$$

其中有模糊矩阵 $\boldsymbol{U}_k=[\mu_{ij,k}]_{c\times n}$,可通过公式(15)计算而得.

定义 2c. 在 k 视角下,定义映射关系 $G_{3,k}:\boldsymbol{R}^{cd_k}\times\boldsymbol{M}_{U_k}\to\boldsymbol{M}_{W_k}$:

$$G_{3,k}(\boldsymbol{V}_k,\boldsymbol{U}_k)=W_k \tag{22}$$

其中有视角权重值 W_k,k=1,2,...,K,可通过公式(19)求得.

定义 3. 在 k 视角下,定义映射关系 $G_{m,k}:\boldsymbol{M}_{W_k}\times\boldsymbol{R}^{cd_k}\times\boldsymbol{M}_{U_k}\to\boldsymbol{M}_{W_k}\times\boldsymbol{R}^{cd_k}\times\boldsymbol{M}_{U_k}$,其集合表述形式如下:

$$G_{m,k}(W_k,\boldsymbol{V}_k,\boldsymbol{U}_k)=\{(W_k^*,\boldsymbol{V}_k^*,\boldsymbol{U}_k^*)\,|\,\boldsymbol{V}_k^*=G_{1,k}(W_k,\boldsymbol{U}_k),W_k^*=G_{2,k}(\boldsymbol{V}_k^*,\boldsymbol{U}_k),\boldsymbol{U}_k^*=G_{3,k}(W_k^*,\boldsymbol{V}_k^*)\}.$$

定义 4. 如果存在 $(W_k^{(t)},\boldsymbol{V}_k^{(t)},\boldsymbol{U}_k^{(t)})\in G_{m,k}(W_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{U}_k^{(t-1)})$, t 表示当前的迭代次数,t=2,3,....这里,$(W_k^{(1)},\boldsymbol{V}_k^{(1)},\boldsymbol{U}_k^{(1)})$ 为 $\boldsymbol{M}_{W_k}\times\boldsymbol{R}^{cd_k}\times\boldsymbol{M}_{U_k}$ 中的任意值,则序列 $\{(W_k^{(t)},\boldsymbol{V}_k^{(t)},\boldsymbol{U}_k^{(t)})\}$ 称为 EW-CoP-MVFCM 算法的优化迭代序列.

定理 4. 设 k 视角下的数据样本 $\boldsymbol{X}_k$={$\boldsymbol{x}_{1,k},\boldsymbol{x}_{2,k},\ldots,\boldsymbol{x}_{n,k}$},定义 $\Phi(\boldsymbol{V}_k)=f_{m,k}(W_k^*,\boldsymbol{V}_k,\boldsymbol{U}_k^*)$ 为 $\Phi:\boldsymbol{R}^{cd_k}\to\boldsymbol{R}$ 上的函数,其中,$W_k^*\in\boldsymbol{M}_{W_k}$,$\boldsymbol{U}_k^*\in\boldsymbol{M}_{U_k}$, m>1,0<η<1,λ_1>0,λ_2>0 为定值.当且仅当 $\boldsymbol{V}_k=G_{1,k}(W_k^*,\boldsymbol{U}_k^*)$ 时,$\boldsymbol{V}_k$ 为 Φ在视角 k 上的关于 $\boldsymbol{R}^{cd_k}$ 的全局最小值点.

定理 5. 设 k 视角下的数据样本 $\boldsymbol{X}_k$={$\boldsymbol{x}_{1,k},\boldsymbol{x}_{2,k},\ldots,\boldsymbol{x}_{n,k}$},定义 $\Lambda(\boldsymbol{U}_k)=f_{m,k}(W_k^*,\boldsymbol{V}_k^*,\boldsymbol{U}_k)$ 为 $\Lambda:\boldsymbol{M}_{U_k}\to\boldsymbol{R}$ 上的函数,其中,$W_k^*\in\boldsymbol{M}_{W_k}$,$\boldsymbol{V}_k^*\in\boldsymbol{R}^{cd_k}$, m>1,0<η<1,λ_1>0,λ_2>0 为定值.当且仅当 $\boldsymbol{U}_k=G_{2,k}(W_k^*,\boldsymbol{V}_k^*)$ 时,$\boldsymbol{U}_k$ 为 Λ在视角 k 上的关于 $\boldsymbol{M}_{U_k}$ 的全局最小值点.

定理 6. 设 k 视角下的数据样本 $\boldsymbol{X}_k$={$\boldsymbol{x}_{1,k},\boldsymbol{x}_{2,k},\ldots,\boldsymbol{x}_{n,k}$},定义 $\Omega(W_k)=f_{m,k}(W_k,\boldsymbol{V}_k^*,\boldsymbol{U}_k^*)$ 为 $\Omega:\boldsymbol{M}_{W_k}\to\boldsymbol{R}$ 上的函数,其中,$\boldsymbol{V}_k^*\in\boldsymbol{R}^{cd_k}$,$\boldsymbol{U}_k^*\in\boldsymbol{M}_{U_k}$, m>1,0<η<1,λ_1>0,λ_2>0 为定值.当且仅当 $W_k=G_{3,k}(\boldsymbol{V}_k^*,\boldsymbol{U}_k^*)$ 时,W_k 为 Ω在视角 k 上的关于 $\boldsymbol{M}_{W_k}$ 上的全局最小值点.

定理 7(EW-CoP-MVFCM 收敛定理). 设 k 视角下的数据样本 $\boldsymbol{X}_k$={$\boldsymbol{x}_{1,k},\boldsymbol{x}_{2,k},\ldots,\boldsymbol{x}_{n,k}$}包含至少 c 个独立的点,c<n,定义:

$$\begin{aligned}\vartheta_k=\{(W_k^*,\boldsymbol{V}_k^*,\boldsymbol{U}_k^*)\boldsymbol{M}_{W_k}\times\boldsymbol{R}^{cd_k}\times\boldsymbol{M}_{U_k}\,|\,&f_{m,k}(W_k^*,\boldsymbol{V}_k^*,\boldsymbol{U}_k^*)<f_{m,k}(W_k^*,\boldsymbol{V}_k^*,\boldsymbol{U}_k),\boldsymbol{U}_k^*\neq U_k;\\&f_{m,k}(W_k^*,\boldsymbol{V}_k^*,\boldsymbol{U}_k^*)<f_{m,k}(W_k,\boldsymbol{V}_k^*,\boldsymbol{U}_k^*),W_k^*\neq W_k;\\&f_{m,k}(W_k^*,\boldsymbol{V}_k^*,\boldsymbol{U}_k^*)<f_{m,k}(W_k^*,V_k,\boldsymbol{U}_k^*),\boldsymbol{V}_k^*\neq V_k\}\end{aligned} \tag{23}$$

为 EW-CoP-MVFCM 算法对应的最优化问题的第 k 个子解集:

$$\begin{cases}\min J_{\text{EW-CoP-MVFCM}}=\min\sum_{k=1}^{K}f_{m,k}(W_k,\boldsymbol{V}_k,\boldsymbol{U}_k)\\ \text{s.t. }\boldsymbol{V}_k\in\boldsymbol{R}^{cd_k},\boldsymbol{U}_k\in\boldsymbol{M}_{U_k},W_k\in\boldsymbol{M}_{W_k},\end{cases}$$

其中,关于 $\min J_{\text{EW-CoP-MVFCM}}$ 的详细计算策略见公式(12).令 $(W_k^{(0)},\boldsymbol{V}_k^{(0)},\boldsymbol{U}_k^{(0)})$ 为当前第 k 视角下,根据映射 $G_{m,k}$ 进行迭代的起始解,且满足 $W_k^{(0)}\in\boldsymbol{M}_{W_k}$,$\boldsymbol{V}_k^{(0)}\in\boldsymbol{R}^{cd_k}$,$\boldsymbol{U}_k^{(0)}=G_{2,k}(W_k^{(0)},\boldsymbol{V}_k^{(0)})$, 则优化迭代过程所产生的序列对应于第 k 个视角可表示为 $\{(W_k^{(t)},\boldsymbol{V}_k^{(t)},\boldsymbol{U}_k^{(t)})\}$, 其中,$t$=1,2,....该序列终止于解集 ϑ_k 中的一个点 $(W_k^*,\boldsymbol{V}_k^*,\boldsymbol{U}_k^*)$, 或存在上述序列相应的一个子序列同样收敛于解集 ϑ_k 中的一个点.

证明:为了证明 EW-CoP-MVFCM 算法的全局收敛性,即需证得 EW-CoP-MVFCM 算法满足引理 1 的三大条件即可,所以本证明部分将分别从这三大条件的满足性入手,给出本文算法的收敛性分析.

(a) 紧集

根据凸包的定义,设 k 视角下的数据样本存在:

$$conv(\boldsymbol{X}_k)=\{\boldsymbol{x}_k\in R^{d_k}\,|\,\boldsymbol{x}_k=\textstyle\sum_{j=1}^{N}\alpha_{j,k}\boldsymbol{x}_{j,k};\sum_{j=1}^{N}\alpha_{j,k}=1;\alpha_{j,k}\geqslant 0;\boldsymbol{x}_j\in\boldsymbol{X}_k\subset R^{d_k}\}$$

为数据集 $\boldsymbol{X}_k$ 上的凸包.根据公式(14)的相关表述易知:

$$\boldsymbol{v}_{i,k}^{(t)}\in conv(\boldsymbol{X}_k),i=1,2,...,c,\ k=1,2,...,K,\ t=1,2,...$$

进而可得 $\boldsymbol{V}_k$=($\boldsymbol{v}_{1,k},\boldsymbol{v}_{2,k},\ldots,\boldsymbol{v}_{c,k}$)∈$[conv(\boldsymbol{X}_k)]^T$.因此,显然存在:

$$(W_k^{(t)},\boldsymbol{V}_k^{(t)},\boldsymbol{U}_k^{(t)})\in\boldsymbol{M}_{W_k}\times[conv(\boldsymbol{X}_k)]^T\times\boldsymbol{M}_{U_k}\subset\boldsymbol{M}_{W_k}\times\boldsymbol{R}^{cd_k}\times\boldsymbol{M}_{U_k}.$$

而 $\boldsymbol{M}_{W_k}, conv(\boldsymbol{X}_k), \boldsymbol{M}_{U_k}$ 均为存在上下界的闭集,因此根据紧集的定义可知,它们均为紧集;

而 $\boldsymbol{M}_{W_k} \times [conv(\boldsymbol{X}_k)]^T \times \boldsymbol{M}_{U_k}$ 亦满足紧集定义,视为紧集.

(b) 连续性

对于第 k 视角,则有 $f_{m,k}$ 为 $G_{m,k}$ 在 $\boldsymbol{X}_k$ 上的 Zangwill 函数.由于 $\boldsymbol{V}_k^{(t+1)} = G_{1,k}(W_k^{(t)}, \boldsymbol{U}_k^{(t)})$,根据定理 4,有:

$$f_{m,k}(W_k^{(t)}, \boldsymbol{V}_k^{(t+1)}, \boldsymbol{U}_k^{(t)}) \leqslant f_{m,k}(W_k^{(t)}, \boldsymbol{V}_k^{(t)}, \boldsymbol{U}_k^{(t)});$$

同理,根据定理 5,有:

$$f_{m,k}(W_k^{(t)}, \boldsymbol{V}_k^{(t+1)}, \boldsymbol{U}_k^{(t+1)}) \leqslant f_{m,k}(W_k^{(t)}, \boldsymbol{V}_k^{(t+1)}, \boldsymbol{U}_k^{(t)});$$

根据定理 6,可得如下不等式:

$$f_{m,k}(W_k^{(t+1)}, \boldsymbol{V}_k^{(t+1)}, \boldsymbol{U}_k^{(t+1)}) \leqslant f_{m,k}(W_k^{(t)}, \boldsymbol{V}_k^{(t+1)}, \boldsymbol{U}_k^{(t+1)}).$$

根据上述 3 个不等式,可得到如下不等关系:

$$f_{m,k}(W_k^{(t+1)}, \boldsymbol{V}_k^{(t+1)}, \boldsymbol{U}_k^{(t+1)}) \leqslant f_{m,k}(W_k^{(t)}, \boldsymbol{V}_k^{(t)}, \boldsymbol{U}_k^{(t)}) \tag{24}$$

故,在第 k 视角下的 $f_{m,k}$ 函数是 $G_{m,k}$ 在 $\boldsymbol{X}_k$ 上的 Zangwill 函数.

(c) 闭映射

根据定义 2 的有关描述,对于第 k 个视角,显然存在如下两个连续映射:

- $G_{1,k} : \boldsymbol{M}_{W_k} \times \boldsymbol{M}_{U_k} \to \boldsymbol{R}^{cd_k}$;
- $G_{3,k} : \boldsymbol{R}^{cd_k} \times \boldsymbol{M}_{U_k} \to \boldsymbol{M}_{W_k}$.

为了证明 $G_{2,k} : \boldsymbol{M}_{W_k} \times \boldsymbol{R}^{cd_k} \to \boldsymbol{M}_{U_k}$ 为闭映射,不妨设定如下假设条件:

1) $(W_k^{(t)}, \boldsymbol{V}_k^{(t)}) \to (W^*, \boldsymbol{V}^*), t \to \infty$;
2) $\boldsymbol{U}_k^{(t)} = G_{2,k}(W_k^{(t)}, \boldsymbol{V}_k^{(t)}), t \to \infty$;
3) $\boldsymbol{U}_k^{(t)} \to \boldsymbol{U}_k^*, t \to \infty$.

根据:

$$\mu_{ij,k}^* = \lim_{t\to\infty} \mu_{ij,k}^{(t)} = \lim_{t\to\infty} \frac{1}{\left(\sum_{h=1}^{c} \frac{w_k^{(t)} \left(\| \boldsymbol{x}_{j,k} - \boldsymbol{v}_{i,k}^{(t)} \|^2 + \frac{1-\eta}{1-2^{1-m}} \lambda_1 \right)}{w_k^{(t)} \left(\| \boldsymbol{x}_{j,k} - \boldsymbol{v}_{h,k}^{(t)} \|^2 + \frac{1-\eta}{1-2^{1-m}} \lambda_1 \right)} \right)^{\frac{1}{m-1}}} = \frac{1}{\left(\sum_{h=1}^{c} \frac{\lim\limits_{t\to\infty} \left(\| \boldsymbol{x}_{j,k} - \boldsymbol{v}_{i,k}^{(t)} \|^2 + \frac{1-\eta}{1-2^{1-m}} \lambda_1 \right)}{\lim\limits_{t\to\infty} \left(\| \boldsymbol{x}_{j,k} - \boldsymbol{v}_{h,k}^{(t)} \|^2 + \frac{1-\eta}{1-2^{1-m}} \lambda_1 \right)} \right)^{\frac{1}{m-1}}} \tag{25}$$

又有:

$$\begin{aligned} \lim_{t\to\infty} \left(\| \boldsymbol{x}_{j,k} - \boldsymbol{v}_{i,k}^{(t)} \|^2 + \frac{1-\eta}{1-2^{1-m}} \lambda_1 \right) &= \lim_{t\to\infty} (\| \boldsymbol{x}_{j,k} - \boldsymbol{v}_{i,k}^{(t)} \|^2) + \frac{1-\eta}{1-2^{1-m}} \lambda_1 \\ &= (\| \boldsymbol{x}_{j,k} - \lim_{t\to\infty} \boldsymbol{v}_{h,k}^{(t)} \|^2) + \frac{1-\eta}{1-2^{1-m}} \lambda_1 \\ &= (\| \boldsymbol{x}_{j,k} - \boldsymbol{v}_{h,k}^* \|^2) + \frac{1-\eta}{1-2^{1-m}} \lambda_1 > 0 \end{aligned} \tag{26}$$

进而可得:

$$\boldsymbol{U}_k^* = G_{2,k}(W_k^*, \boldsymbol{V}_k^*) \tag{27}$$

因此,我们可以得到: $G_{2,k} : \boldsymbol{M}_{W_k} \times \boldsymbol{R}^{cd_k} \to \boldsymbol{M}_{U_k}$ 为一闭映射;

同时,易证得 $G_{m,k} : \boldsymbol{M}_{W_k} \times \boldsymbol{R}^{cd_k} \times \boldsymbol{M}_{U_k} \to \boldsymbol{M}_{W_k} \times \boldsymbol{R}^{cd_k} \times \boldsymbol{M}_{U_k}$ 亦为闭映射.

综合条件(a)~条件(c)的证明,因每一视角下的 $f_{m,k}(W_k, \boldsymbol{V}_k, \boldsymbol{U}_k)$ 均为 Zangwill 函数,即,均为凸函数,因此,由其求和所得的 EW-CoP-MVFCM 算法亦满足 Zangwill 收敛定理,具备全局收敛性.至此,定理 7 得证. □

4 实验研究

4.1 实验设置

为了验证本文方法对多视角聚类任务的有效性及自适应视角权重划分的效果,本节将分别通过人工合成数据集以及 UCI 标准数据库数据(http://archive.ics.uci.edu/ml)对 EW-CoP-MVFCM 算法进行分析与评估.有关人工合成数据以及 UCI 标准数据库数据的详细描述分别于第 4.2 节和第 4.3 节中给出.此外,为对本文所提 EW-CoP-MVFCM 算法聚类性能做出评判,将于第 4.2 节及第 4.3 节给出与当前最新的多视角模糊聚类算法 Co-FKM[18]、基于多任务学习框架的 LSSMTC 算法[19]、基于多任务的组合 *K*-means 算法(CombKM)[19]及基于样本与特征空间协同聚类的 Co-clustering 算法[20]在人工集与标准集上进行性能比较的聚类结果,并对此结果进行适当的分析与解释.

为了公正地对各聚类算法的聚类性能做出合理的评价,本文采用如下两种评价指标进行算法的性能分析.

1) 归一化互信息(normalized mutual information,简称 NMI)[6,23]:

$$NMI=\frac{\sum_{i=1}^{C}\sum_{j=1}^{C}N_{i,j}\log N\cdot N_{i,j}/N_i\cdot N_j}{\sqrt{\sum_{i=1}^{C}N_i\log N_i/N\cdot\sum_{j=1}^{C}N_j\log N_j/N}} \tag{28}$$

其中,$N_{i,j}$ 表示第 i 个聚类与类 j 的契合程度,N_i 表示第 i 个聚类所包含的数据样本量,N_j 表示类 j 所包含的数据样本量,而 N 表示整个数据样本的总量大小;

2) 芮氏指标(rand index,简称 RI)[23,27]:

$$RI=\frac{f_{00}+f_{11}}{N(N-1)/2} \tag{29}$$

其中,f_{00} 表示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} 则表示数据点具有相同的类标签并且属于同一类的配对点数目,而 N 表示整个数据样本的总量大小.

以上两种方法,其取值范围均为[0,1],且均随着数值的增大,显示出算法的性能更为优越.

在本文实验部分,各可调参数的设置遵循文献[18,23]的计算策略,具体见表 1.本文的实验结果均为 10 次算法完整运行下的均值及方差所组成.

Table 1 Definition and settings of the related notations

表 1 参数定义和设置

参数	取值策略
模糊指数 m^*	$m=\frac{\min(N,D-1)}{\min(N,D-1)-2}$,其中,$N$ 表示样本大小,D 表示维数
协同学习参数 η	$\eta=(K-1)/K$,其中,K 为视角数
正则化平衡各视角隶属度划分因子 λ_1	利用网格搜索法进行遍历寻优,寻优范围:[0.01,10]
正则化平衡各视角权重因子 λ_2	利用网格搜索法进行遍历寻优,寻优范围:[0.1,100]

* 需注意的是,利用此公式需满足 $D>3$,在 $D\leqslant 3$ 时,易产生分母为 0 或为负值的情况,则此模糊指数寻参公式失效.在此公式失效时,本文依照文献[23]的寻参策略,采用网格搜索法进行遍历寻优,其寻优区间在[1.1,3]之间

实验环境:实验硬件平台为 Intel Core i5-2410M×2 CPU,其主频为 2.3GHz,内存为 2GB.编程环境为 MATLAB 7.0.

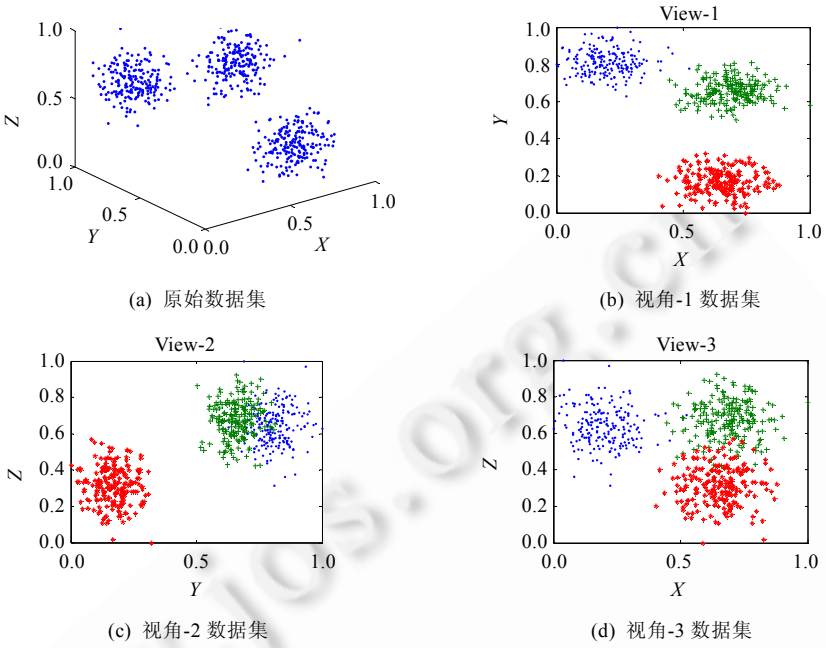
4.2 人工合成数据集实验

为了充分验证本文方法的多视角特性及自适应视角权重划分的效果,在人工合成数据集部分,本文首先构造具有 3 维特性的数据集 $A(x,y,z)$用来表示待分析的事物/对象,而后,根据投影的原则选取了 3 个视角,分别为:

- *view*_1=(x,y),由 A 事物向特征 x 与特征 y 方向投影得到;
- *view*_2=(y,z),由 A 事物向特征 y 与特征 z 方向投影得到;

- *view*_3=(*x*,*z*),由 *A* 事物向特征 *x* 与特征 *z* 方向投影得到.

对于每一个视角而言都包含 3 类样本,每一类样本由 200 个点对构成共 600 个样本.有关对象 *A* 的示意图及各视角下的样本组成如图 4 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

Fig.4 Original dataset and the corresponding datasets under different

图 4 原始数据集及其在不同视角下的对应数据集

观察图 4 可以发现:对于事物 *A* 从不同的方位所投影得到的 3 个视角样本,其中只有视角 1 具备非常清晰的聚类特性,而视角 2 及视角 3 的样本类别之间均存在一定程度的重叠;特别是视角 2,其中两类的重叠性非常高.如若使用现有的多视角技术对此类聚类任务进行分析,所得到的聚类结果必定会受到视角 2 及视角 3 的影响,从而破坏了整体的可分性.为了验证上述观点,本文在该人工合成数据集上进行了相应的实验验证并与有关算法进行了比较,实验结果如表 2 和图 5 所示.

根据表 2 和图 5 的结果分析,可得到如下结论.

1) 与非多视角聚类算法比较.

对表 2 的前 3 列结果进行分析可以发现,无论是多任务的 LSSMTC 算法和 CombKM 算法,还是基于样本空间与特征空间协同分析的 Co-clustering 算法,聚类结果均表现一般.分析其原因主要在于:针对多任务聚类算法而言,虽然该类算法具备多任务协调工作的能力,但由于其关注的重点局限于样本本身,而多视角聚类任务其本质就是将一个事物 *A* 通过不同的特征空间抽样得到不同的样本集合,其重点考虑的是各视角下的特征因素,因此,多任务算法得到了一般的聚类效果也是可以理解的;而对于具备样本空间与特征空间协同分析能力的 Co-clustering 算法而言,该算法仍是基于单视角框架下的聚类技术,因此在处理数据时只可将多个视角合并为一个样本集合再进行处理,而原本的每一视角在此样本集合中则变成了此整体样本的特征子集.由于视角 2 和视角 3 为较难聚类的视角,它们的存在使得 Co-clustering 算法在面对此类聚类任务时效果也并不理想;较之上述算法,无论是经典的 Co-FKM 算法还是本文的 EW-CoP-MVFCM 算法,因均采用了多视角技术,即,各视角的空间划分尽可能地逼近,从而使得上述两种多视角聚类算法获得了更好的聚类效果.这也印证了多视角聚类技术的有效性;

2) 与多视角聚类算法比较.

通过对表 2 的后两列进行分析可知,本文的方法明显优于 Co-FKM 算法.分析其原因主要在于,本文前几节一直强调的视角重要性不均等的理论.观察图 4 可清晰地看出,视角 1 具备非常优秀的空间划分特性,而视角 2 和视角 3 的空间划分特性并不明显,特别是在类别的区分面上存在严重的交叠现象,因此,当 Co-FKM 这类假设空间划分重要性均等的算法遇到上述各视角重要性不均衡的聚类任务时,算法的有效性必定会受到一定的影响;而本文的 EW-CoP-MVFCM 方法因采用了独有的多视角自适应加权技术,使得优化终止时最佳视角的权重突出,如图 5 所示.由于视角 1 具备很好的空间划分特性,因此在熵加权技术的帮助下,我们最终获取的视角权重系数也与真实的情况趋于一致,从而获取了更为合理的聚类效果,提高了算法的适用性.

Table 2 Comparison of performance indices NMI & RI of several algorithms on the synthetic datasets

表 2 各种算法在模拟数据集上的性能比较

Dataset	Index	LSSMTC	CombKM	Co-Clustering	Co-FKM	EW-CoP-MVFCM
A	NMI-mean	0.746 8	0.895 7	0.652 7	0.922 9	**1**
	NMI-std	0.012 6	0.167 9	0	0.124 3	**0**
	RI-mean	0.900 3	0.916 6	0.722 4	0.968 4	**1**
	RI-std	0.006 4	0.134 2	1.17e-016	0.051 3	**0**

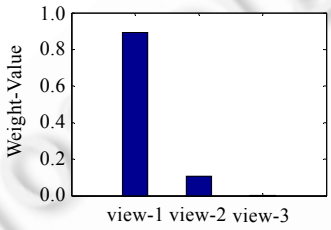


Fig.5 Weight of each views for target A with EW-CoP-MVFCM algorithm

图 5 EW-CoP-MVFCM 算法获取的事物 A 各视角权重情况

4.3 UCI数据集实验分析

为了对 EW-CoP-MVFCM 算法的聚类性能及实际应用价值作进一步的探讨与分析,本节将分两部分对 EW-CoP-MVFCM 算法进行探讨:(1) 基于 UCI 数据库中的经典数据集 IRIS,对本文方法寻找各视角权重划分并找出最佳视角的能力做出进一步的评判,具体实验设置见第 4.3.1 节;(2) 利用 UCI 数据库中的经典多视角数据集 Multiple Features、Image Segmentation 及 Water Treatment Plant 对本文所提算法的性能做出更加充分的评价,同时与相关算法进行比对分析.本节所用数据的基本信息见表 3.

Table 3 Classic UCI datasets

表 3 UCI 经典数据集

Dataset	Description	Size	Dimension	Cluster
IRIS	Class of IRIS plant	150	4	3
Multiple features (MF)	Handwritten numerals represented by multiple features	2 000	649	10
Image segmentation (IS)	Outdoor images	2 310	19	7
Water treatment plant (WTP)	The dataset came from the daily measures of sensors in an urban waste water treatment plant	527	38	13

4.3.1 IRIS 实验分析

由于 IRIS 每一维(属性特征)对应的样本分布具有很好的最佳视角解释性,其每一维特征对应的数据分布如图 6 所示.基于上述原因,本文选用此数据集,并将其每一维特征看作一个视角,即,将原本的 IRIS 数据样本拆分成 4 个视角样本集合,从而利用本文算法进行聚类分析,其聚类结果如图 7 和表 4 所示.

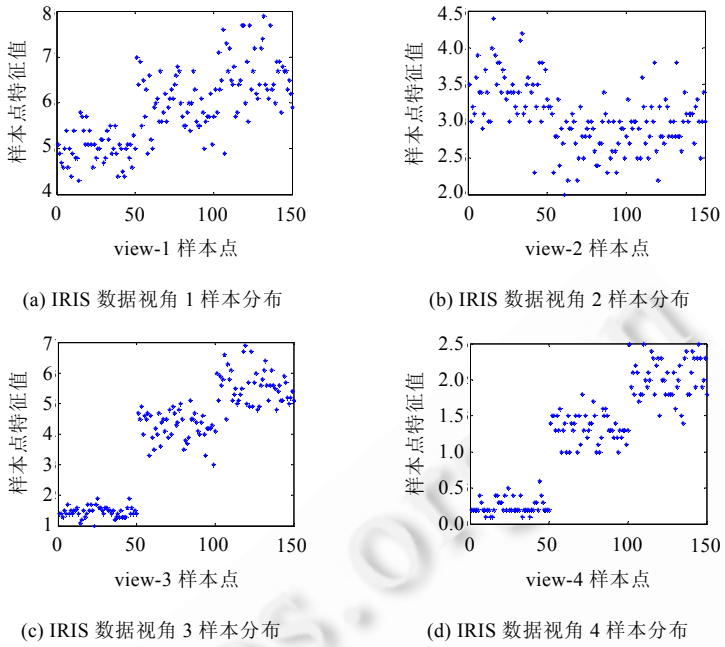


Fig.6 Datasets of each views for IRIS

图 6 IRIS 各视角数据直观图

Table 4 Comparison of performance indices NMI & RI of several algorithms on the IRIS datasets

表 4 各种算法在 IRIS 数据集上的性能比较

Dataset	Index	LSSMTC	CombKM	Co-Clustering	Co-FKM	EW-CoP-MVFCM
IRIS	NMI-mean	0.726 0	0.741 2	0.758 2	0.836 6	**0.871 1**
	NMI-std	0.047 0	0.053 7	1.17e-016	1.48e-016	**0.012 8**
	RI-mean	0.887 9	0.863 6	0.879 7	0.934 1	**0.951 9**
	RI-std	0.021 9	0.051 1	0	0	**0.005 3**

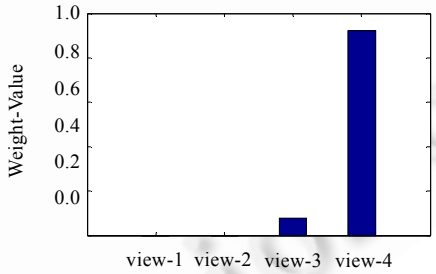


Fig.7 Weight of each views for IRIS with EW-CoP-MVFCM algorithm

图 7 EW-CoP-MVFCM 算法获取的 IRIS 各视角权重情况

根据表 4 及图 7 的结果可以发现,其与人工合成数据集的实验结果是一致的.同样,由于本文算法有效地利用了最佳视角,使得最佳视角的权重达到了最大化,从而获取了更为合理的空间划分结果.从 NMI 和 RI 两大评价指标上也可以看出,本文算法较之其余的几种方法显得更为有效.同时,图 7 的视角权重比更加有力地证明了:本文算法能够通过新目标函数寻找到最优的视角权重划分结果,并与真实的视角重要性程度保持一致.

4.3.2 多视角真实数据集实验分析

本节将采用经典机器学习数据库 UCI 中的 3 种具备多视角特性的数据集:1) MF 数据集;2) IS 数据集;

3) WTP 数据集.通过利用上述数据集,考察本文所提 EW-CoP-MVFCM 算法在处理真实多视角聚类任务时其性能的有效性.为了对这 3 种数据集所包含的视角有更为直观的印象,本文将给出这三大数据集的各视角特征组成,详见表 5.同时,针对这 3 种真实数据集的算法性能比对结果可见表 6.

Table 5 Description of MF-dataset, IS-dataset and WTP-dataset and the construction of each view

表 5 MF,IS 及 WTP 数据集简介及两者的各视角组成

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

Table 6 Comparison of performance indices NMI & RI of several algorithms on the datasets of MF, IS and WTP

表 6 各种算法在 MF,IS 及 WTP 数据集上的性能比较

Dataset	Index	NMI-mean	NMI-std	RI-mean	RI-std
MF	LSSMTC	–	–	–	–
	CombKM	0.688 3	0.033 7	0.899 1	0.011 9
	Co-clustering	0.702 4	0.028 7	0.910 4	0.012 6
	Co-FKM	**0.786 8**	0.071 5	0.941 7	0.022 6
	EW-CoP-MVFCM	0.782 6	**0.035 2**	**0.948 3**	**0.012 2**
IS	LSSMTC	–	–	–	–
	CombKM	0.481 5	0.028 3	0.798 1	0.017 6
	Co-clustering	0.518 2	0.010 5	0.818 9	0.020 8
	Co-FKM	0.542 2	0.035 3	0.829 3	0.019 5
	EW-CoP-MVFCM	**0.581 7**	**0.036 2**	**0.840 8**	**0.025 5**
WTP	LSSMTC	–	–	–	–
	CombKM	0.176 1	0.012 1	0.704 2	0.003 7
	Co-clustering	0.202 9	0.010 3	0.705 1	0.006 1
	Co-FKM	0.198 6	0.014 0	0.705 6	0.004 4
	EW-CoP-MVFCM	**0.212 8**	**0.005 8**	**0.707 7**	**0.002 0**

由于 LSSMTC 算法需要保证各聚类任务的维数一致,因此在面对 MF,IS 及 WTP 等各视角维数均不相等的样本时便无法使用.另外,通过观察其余各算法对 MF 数据集的聚类结果可以发现,基于多视角的 Co-FKM 及本文的算法有着较大的聚类优势,但是,因 MF 数据并无任何视角存在明显的可分性,即,各视角的重要性程度较均衡,上述原因使得本文所提算法与 Co-FKM 算法的聚类结果从 NMI 及 RI 两大指标的均值上看较为接近,而从方差上分析依旧是本文的方法较为稳定.究其原因在于:本文所采用的基于 Havrda-Charvat 熵构造的异视角空间划分逼近准则具有良好的物理解释性,能够充分挖掘视角间的相似性成分,这一特性保证了本文方法每次运行得到的结果偏差不大,较为稳定;而对于 IS 数据集,本文方法优势明显,进一步说明了 EW-CoP-MVFCM 算法的有效性;最后,通过对 WTP 数据集的实验结果分析,同样也可得到与上述两大数据集一致的结论.综上,通过在真实多视角数据集上的实验与分析,我们可以得到一个较明确的结论,即,多视角聚类算法在处理具备多视角特性的聚类任务时一般均优于非多视角聚类算法,而具有更强协同学习能力及视角选择性的 EW-CoP-MVFCM 算法又要优于以往的多视角聚类算法.至此,本文算法的性能得到了充分的验证与肯定.

5 结 论

本文基于多视角聚类技术,在经典 FCM 算法的基础上引入了基于 Havrda-Charvat 熵构造的异视角空间划分逼近准则.由于该准则由熵理论构造得到,因而具有良好的物理解释性,这使得本文方法在运用该准则时能够更好地找出视角间的相似性成分,进而得到更具指导意义的全局性空间划分结果.此外,本文另一大贡献在于重新审视了视角的重要性程度,摒弃了以往默认的各视角重要性一致的结论.通过对极大熵相关理论的理解,提出了基于香农熵的多视角自适应加权策略,并成功地将该策略引入到最新的多视角模糊聚类技术中.在新获取的目标函数取得最优解时,我们进一步根据各视角最终所占的权重关系评价出视角的重要性程度,藉此提出了一种全新的多视角集成方法,即,全局性加权视角空间划分集成法.通过在模拟数据集以及 UCI 真实数据集的实验结果,均显示出了本文方法较之以往多视角算法及相关算法具有更好的聚类性能.但因本文仍然采用经典的 FCM 框架,对于欧氏距离的使用,致使本文算法在面对高维多视角聚类问题时其算法性能将面临一定的考验,该问题的解决将是今后研究的重点.

References:

[1] Li GX, Chang KY, Hoi SCH. Multi-View semi-supervised learning with consensus. IEEE Trans. on Knowledge and Data Engineering, 2012,24(11):2040−2051. [doi: 10.1109/TKDE.2011.160]

[2] Li GX, Hoi SCH, Chang KY. Two-View transductive support vector machines. In: Parthasarathy S, Liu B, Goethals B, Pei J, Kamath C, eds. Proc. of the 10th SIAM Int'l Conf. on Data Mining. 2010. 235−244. http://epubs.siam.org/doi/abs/10.1137/1.9781611972801.21

[3] Bickel S, Scheffere T. Multi-View clustering. In: Proc. of the 4th IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society, 2004. 19−26. [doi: 10.1109/ICDM.2004.10095]

[4] Jain AK, Murty MN, Flynn PJ. Data clustering: A review. ACM Computing Surveys, 1999,31(3):264−323. [doi: 10.1145/331499.331504]

[5] Yu S, Tranchevent LC, Liu XH, Glanzel W, Suykens JAK, De Moor B, Moreau Y. Optimized data fusion for kernel *k*-means clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012,34(5):1031−1039. [doi: 10.1109/TPAMI.2011.255]

[6] Jing LP, Ng MK, Huang JZ. An entropy weighting *k*-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Trans. on Knowledge and Data Engineering, 2007,19(8):1026−1041. [doi: 10.1109/TKDE.2007.1048]

[7] Zhu L, Chung FL, Wang ST. Generalized fuzzy C-means clustering algorithm with improved fuzzy partitions. IEEE Trans. on Systems Man and Cybernetics: Part B, 2009,39(3):578−591. [doi: 10.1109/TSMCB.2008.2004818]

[8] Hall LO, Goldgof DB. Convergence of the single-pass and online fuzzy C-means algorithms. IEEE Trans. on Fuzzy Systems, 2011,19(4):792−794. [doi: 10.1109/TFUZZ.2011.2143418]

[9] Deng ZH, Wang ST, Wu XS, Hu DW. Robust maximum entropy clustering algorithm RMEC and its outlier labeling. Engineering Science, 2004,6(9):38−43 (in Chinese with English abstract).

[10] Karayiannis NB. MECA: Maximum entropy clustering algorithm. In: Proc. of the 3rd IEEE Conf. on Fuzzy Systems. Orlando: IEEE Press, 1994. 630−635. [doi: 10.1109/FUZZY.1994.343658]

[11] Krishnapuram R, Keller JM. A possibilistic approach to clustering. IEEE Trans. on Fuzzy Systems, 1993,1(2):98−110. [doi: 10.1109/91.227387]

[12] Krishnapuram R, Keller JM. The possiblistic means algorithms: Insights and recommendation. IEEE Trans. on Fuzzy Systems, 1996,4(3):385−393. [doi: 10.1109/91.531779]

[13] Asur S, Parthasarathy S, Ucar D. An ensemble framework for clustering protein interaction networks. Bioinformatics, 2007,23(13):i29−i40. [doi: 10.1093/bioinformatics/btm212]

[14] Wang HJ, Shan HH, Banerjee A. Bayesian cluster ensembles. Statistical Analysis and Data Mining, 2011,4(1):54−70. [doi: 10.1002/sam.10098]

[15] Yamanishi Y, Vert JP, Kanehisa M. Protein network inference from multiple genomic data: A supervised approach. Bioinformatics, 2004,20(1):i363−i370. [doi: 10.1093/bioinformatics/bth910]

[16] Pedrycz W. Collaborative fuzzy clustering. Pattern Recognition Letter, 2002,23(14):1675−1686. [doi: 10.1016/S0167-8655(02)00130-7]

[17] Chaudhuri K, Kakade SM, Livescu K, Sridharan K. Multi-View clustering via canonical correlation analysis. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning. New York: ACM Press, 2009. 1−8. [doi: 10.1145/1553374.1553391]

[18] Cleuziou G, Exbrayat M, Martin L, Sublemontier JH. CoFKM: A centralized method for multiple-view clustering. In: Proc. of the 9th IEEE Int'l Conf. on Data Mining (ICDM 2009). 2009. 752−757. [doi: 10.1109/ICDM.2009.138]

[19] Gu Q, Zhou J. Learning the shared subspace for multi-task clustering and transductive transfer classification. In: Proc. of the 9th IEEE Int'l Conf. on Data Mining. 2009. 159−168. [doi: 10.1109/ICDM.2009.32]

[20] Gu Q, Zhou J. Co-Clustering on manifolds. In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2009. 359−368. [doi: 10.1145/1557019.1557063]

[21] Havrda JH, Charvat F. Quantification methods of classification processes: Concepts of structural α-entropy. Kybernetica, 1967,3(1):30−35.

[22] Wang XZ, An SF. Research on learning weights of fuzzy production rules based on maximum fuzzy entropy. Journal of Computer Research and Development, 2006,43(4):673−687 (in Chinese with English abstract). [doi: 10.1360/crad20060416]

[23] Deng ZH, Choi KS, Chung FL, Wang ST. Enhanced soft subspace clustering integrating within-cluster and between-cluster information. Pattern Recognition, 2010,43(3):767−781. [doi: 10.1016/j.patcog.2009.09.010]

[24] Bezdek JC, Hathaway RJ, Sabin MJ, Tucker WT. Convergence theory for fuzzy C-means: Counterexamples and repairs. IEEE Trans. on Systems Man and Cybernetics, 1987,17(5):873−877. [doi: 10.1109/TSMC.1987.6499296]

[25] Zangwill W. Convergence conditions for nonlinear programming algorithms. Management Science, 1969,16(1):1−13. [doi: 10.1287/mnsc.16.1.1]

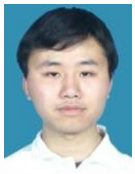
[26] Luenberger DG, Ye YY. Linear and Nonlinear Programming. 3rd ed., New York: Springer-Verlag, 2008. 201−208. [doi: 10.1007/978-0-387-74503-9]

[27] Liu J, Mohammed J, Carter J, Ranka S, Kahveci T, Baudis M. Distance-Based clustering of CGH data. Bioinformatics, 2006,22(16):1971−1978. [doi: 10.1093/bioinformatics/btl185]

附中文参考文献:

[9] 邓赵红,王士同,吴锡生,胡德文.鲁棒的极大熵聚类算法 RMEC 及其例外点标识.中国工程科学,2004,4(9):38−45.

[22] 王熙照,安素芳.基于极大模糊熵原理的模糊产生式规则中的权重获取方法研究.计算机研究与发展,2006,43(4):673−687. [doi: 10.1360/crad20060416]



蒋亦樟(1988－),男,江苏无锡人,博士生,CCF 学生会员,主要研究领域为模式识别,智能计算.
E-mail: jyz0512@163.com

邓赵红(1981－),男,博士,副教授,CCF 高级会员,主要研究领域为智能计算,系统建模.
E-mail: dzh666828@aliyun.com

王骏(1978－),男,博士,副教授,CCF 会员,主要研究领域为智能计算,数据挖掘.
E-mail: wangjun_sytu@hotmail.com

钱鹏江(1979－),男,博士,副教授,主要研究领域为智能计算,模式识别.
E-mail: qianpjiang@126.com

王士同(1964－),男,博士,教授,博士生导师,主要研究领域为模式识别,人工智能.
E-mail: wxwangst@yahoo.com.cn