第 57 卷　第 2 期
2021 年 3 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 57, No. 2
Mar., 2021

DOI:10.13232/j.cnki.jnju.2021.02.003

基于信息熵加权的聚类集成算法

邵长龙,孙统风,丁世飞*

(中国矿业大学计算机科学与技术学院,徐州,221116)

摘　要:聚类集成的目的是通过集成多个不同的基聚类来生成一个更好的聚类结果,近年来研究者已经提出多个聚类集成算法,但是目前仍存在的局限性是这些算法大多把每个基聚类和每个簇都视为同等重要,使聚类结果很容易受到低质量基聚类和簇的影响.为解决这个问题,研究者提出一些给基聚类加权的方法,但大多把基聚类看作一个整体而忽视其中每个簇的差异.受到信息熵的启发,提出一种基于信息熵加权的聚类集成算法.算法首先对每个簇的不稳定性进行衡量,然后提出一种基于信息熵的簇评价指标,进而从簇层面进行加权,在对加权矩阵进行划分后得到最终的聚类结果.该算法有两个主要优点:第一,提出了一个有效的簇评价性指标;第二,从比基聚类层面更细化的簇层面进行加权.一系列的实验证明了该算法的有效性和鲁棒性.

关键词:聚类集成,聚　类,簇层面加权,信息熵

中图分类号:TP399　　**文献标识码:**A

Ensemble clustering based on information entropy weighted

Shao Changlong, Sun Tongfeng, Ding Shifei*

(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China)

Abstract: The purpose of ensemble clustering is to generate a better clustering result by integrating multiple different base clustering. In recent years, researchers have proposed multiple ensemble clustering algorithms. However, the current limitation is that most of these algorithms regard each base clustering and each cluster as equally important, which makes the clustering results susceptible to low-quality base clusterings and clusters. In order to solve this problem, researchers have proposed some methods to weight the base clustering, but most of these methods regard the base clustering as a whole, and ignore the difference of the clusters. In this paper, we are inspired by information entropy and propose an ensemble clustering algorithm based on weighted information entropy. This algorithm first measures the uncertainty of each cluster, then proposes a cluster evaluation index based on information entropy, and then weights it from the cluster level. After dividing the weighting matrix, the final clustering result is obtained. The algorithm in this paper has two main advantages. First, it proposes an effective cluster evaluation index. Second, it calculates the weights from the cluster level that is more refined than from the base cluster level. A series of experiments have proved the effectiveness and robustness of the proposed algorithm.

Key words: ensemble clustering, clustering, cluster-level weighted, information entropy

聚类是一种无监督的机器学习技术,通过计算数据对象间的相似度把数据集分成若干个簇,使在相同簇的对象有较高的相似度,不同簇的对象则差异较大[1].目前聚类已被运用在各种领域:在图像处理领域,Cong et al[2]基于超像素谱聚类提出了一种图像分割算法,在计算复杂度、处理时

基金项目:国家自然科学基金(61976216,61672522)

收稿日期:2021－01－11

* 通讯联系人,E-mail:dingsf@cumt.edu.cn

间和整体分割效果方面都取得了实质性的改善. 在认知计算领域,Saini et al[3]提出一种基于认知计算的多目标自动文档聚类技术,实验结果证明该方法优于传统方法. 在医学诊断领域,Thanh et al[4]提出一种新型聚类算法用于医学诊断中的推荐系统.

过去几十年,大量聚类算法被开发出来,但当前的聚类算法仍然存在一些问题,例如:聚类结果很大程度上取决于参数和初始化;聚类结果不够鲁棒等. 为了解决这些问题,研究人员提出聚类集成算法. 与使用算法生成单个聚类结果的传统方法不同,聚类集成是将多个不同的聚类结果集成在一起以生成更好聚类结果的过程. 聚类集成算法的有效性吸引了许多研究者,并提出了许多相关的算法. 聚类集成的具体内容介绍如下:

给定一个具有 n 个数据对象的数据集 $X=\{x_1,x_2,\cdots,x_n\}$,对此数据集 X 使用 m 次聚类算法,得到 m 个聚类结果 $P=\{p_1,p_2,\cdots,p_m\}$. 其中 p_i $(i=1,2,\cdots,m)$ 为第 i 个聚类算法得到的聚类结果,又称为聚类成员或基聚类. 具体地,基聚类的生成有三种方法:(1)使用一种聚类算法,每次运行时随机设置不同的参数并随机初始化;(2)使用不同的聚类算法产生不同的基聚类;(3)将数据集进行采样得到不同的子集,再对子集进行聚类,从而得到不同的基聚类.

每个基聚类包含若干个簇,记作 $p_i=\{C_i^1,C_i^2,\cdots,C_i^j\}$,其中 j 是基聚类 p_i 里包含簇的个数. 所谓聚类集成就是对集合 P 通过一致性函数(又叫共识函数)T 进行合并,得到数据集 X 的最终聚类结果 P^*. 聚类集成的具体流程如图 1 所示.

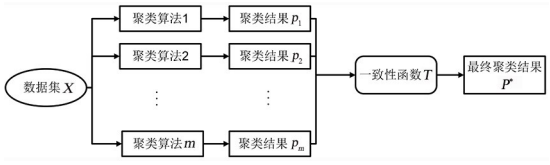

图 1 聚类集成流程示意图

Fig. 1 The flowchat of ensemble clustering

总体来说,现有的聚类集成算法可分三个大类,即基于共关联矩阵的算法、基于图分区的算法和基于中值聚类的算法.

基于共关联矩阵的算法根据数据点与数据点之间在相同簇中共现的频率得到一个共关联矩阵,并以该矩阵作为相似度矩阵,采用层次聚类的算法得到最终的结果. Fred and Jain[5]首次提出共关联矩阵的概念,并据此设计了证据集积累聚类算法. Li et al[6]将基聚类的多尺度特征纳入考虑,提出一种针对密度聚类的集成方法. Rathore et al[7]利用随机投影对高维数据进行降维,并利用共关联矩阵设计了一种针对模糊聚类的聚类集成算法. Zhong et al[8]认为删除共关联矩阵值较小的项可以提高聚类效果,并猜想那些项之中可能包含着大量噪声.

基于图分区的算法将聚类集成的信息构成一个图结构,再利用图分割算法将图分割成若干块,进而得到最终的聚类结果. Strehl and Ghosh[9]将基聚类的每个簇都看作一个超边缘,构造了三种超图结构,对其进行图分割得到最终的聚类结果. Fern and Brodley[10]将基聚类构造成二部图,其中对象和簇都表示为图节点,并用 Ncut 算法[11]对其进行分割. Huang et al[12]提出一种针对大规模数据的基于采样的谱聚类算法,并设计了一个二部图对其进行聚类集成.

基于中值聚类的算法将聚类集成问题建模成一个最优化问题,其优化目标是寻找一个与所有基聚类最相似的聚类结果,这个聚类结果被视为所有基聚类的中值点. 这个问题已经被证明是一个 NP 难问题[13],所以在全局聚类空间里寻找最优解在较大的数据集上是不可行的. 为此,Cristofor and Simovici[14]提出利用遗传算法求聚类集成的近似解,其中聚类被视为染色体. Wu et al[15]提出一种效用函数,将聚类集成问题转化到基于 k-means 建立的框架中解决. Huang et al[16]将聚类集成问题化为二元线性规划问题,并通过因子图模型进行求解.

尽管聚类集成取得了一些进展,但目前的研究仍然存在局限性,比如:这些算法大多把每个基聚类和每个簇都视为同等重要,使得聚类结果容易受到低质量基聚类的影响.

在聚类集成中,基聚类的质量在集成过程中起至关重要的作用,低质量的基聚类可能严重影响聚类结果. 为了避免低质量基聚类的影响,研

究者已经开展了一些工作,其中比较可行的方法是设计一个评价标准来评估基聚类,并在集成过程中针对不同质量的基聚类进行加权以增强集成性能;但这些方法多是将每个基聚类视为一个整体,并为每个基聚类分配权重,而不考虑其内部簇的多样性. 比如:Yu et al[17]将重点放在聚类集成中的基聚类选择上,根据评价指标从基聚类集合中仅选择部分基聚类进行集成,并将基聚类视为特征. 这样可以使用合适的特征选择技术来执行基聚类选择,然而同一基聚类中的不同簇可能具有不同的稳定性,有必要纳入考虑. 为此,Huang et al[18]提出一种局部加权的聚类集成方法,将簇不稳定性整合到局部加权方案中以提高共识性能. 本文对此进行改进,提出一种基于信息熵加权的聚类集成算法(Information Entropy Weighted Ensemble Clustering,IEWEC),消除原方法的参数并对具体计算过程进行了改造. IEWEC改进了Huang et al[18]提出的集成驱动聚类指数,并结合信息熵和Jaccard系数提出了基于信息熵的簇评价方法,通过此方法对簇稳定性进行评估,然后在生成共协矩阵的过程中根据评估结果进行加权,最后将Ncut算法[19]当作共识函数以得到最终结果.

本文的主要贡献:

(1)结合信息熵的概念和Jaccard系数提出一种衡量簇稳定性的评价标准,称为基于信息熵的簇评价指标.

(2)提出的评价指标从簇层面进行加权,避免了其他方法仅考虑基聚类层面而忽视簇层面差异的弊端.

1　相关工作

1.1　信息熵　1948年,Shanon提出信息熵的概念来解决对信息的量化度量问题. 信息熵$H(X)$对于离散随机变量X的形式定义如下:

$$H(X)=-\sum_{x\in X}p(x)\log_2 p(x) \tag{1}$$

其中,$p(x)$代表离散型随机变量的各个情况发生的概率.

在信息论中,信息熵是衡量信息源不确定性的量度,而本研究的目的是衡量簇的不稳定性,故可以考虑仿造熵的形式给出对簇不稳定性的衡量指标. Jaccard系数可以反映两个簇之间的相似程度,用它替换信息熵中的概率分布可以衡量一个簇与基聚类中其他簇的相似度,而一个与其他基聚类中的簇更相似的簇更稳定,因为这说明这个簇在多次基聚类过程中更大程度保持着原有的结构,进而可以衡量每个簇的稳定程度,然后根据簇的稳定程度进行加权.

1.2　集成驱动聚类指数　2017年,Huang et al[18]提出集成驱动聚类指数(Ensemble-Driven Cluster Index,*ECI*)作为评估簇不稳定性的指标,其详细过程如下:

首先,借助基聚类里所有的簇衡量一个簇的不稳定性,方法如下:

$$\begin{aligned} H(C_i)&=-\sum_{j=1}^{n}p(C_i,C_j)\log_2 p(C_i,C_j)\\ p(C_i,C_j)&=\frac{|C_i\cap C_j|}{|C_i|}\end{aligned} \tag{2}$$

其中,n为基聚类中簇的总个数,$|.|$代表集合中点的个数.

得到每个簇的不稳定性后,集成驱动聚类指数(*ECI*)定义如下:

$$ECI(C_i)=\mathrm{e}^{-\frac{H(C_i)}{\theta\cdot M}} \tag{3}$$

其中,M为基聚类的个数,$\theta>0$为参数,建议值的范围是$[0.2,1]$,并在实验中设置为0.4.

Huang et al[18]以集成驱动聚类指数作为评价簇稳定性的指标对簇进行加权,设计了两种共识函数,并用实验证明了此算法的优越性. 此指标后来被用在多个算法中,例如:Huang et al[20]将集成驱动聚类指数和MCLA(Meta-Clustering Algorithm)算法相结合,提出LWMC(Locally Weighted Meta-Clustering)算法,效果比原算法更好. He and Huang[21]结合MCLA算法、集成驱动聚类指数和随机游走算法,提出MC^{3}LR(Meta-Cluster Based Consensus Clustering with Local Weighting and Random Walking)算法,不仅提升了聚类效果,还减少了原算法的时间复杂度.

尽管Huang et al[18]的方法有诸多优点,但在实际应用中,由于参数θ的存在使聚类结果受参数的影响很大,而参数θ的确定却是非常困难的.

尽管通过大量实验可以确定参数的最佳范围,但得到一个固定的值仍然很困难. 为此,本研究提出一种新的加权指标,不需要参数也能取得较好的聚类结果,后文的实验也证明了这一点.

2 本文算法(IEWEC)

2.1 基于信息熵的簇评价指标 为了评估基聚类中每个簇的稳定性,仿造信息熵的形式,利用基聚类中其他簇的信息对其进行衡量. 方法如下:

$$H(C_i) = -\sum_{j=1}^{n} p(C_i, C_j)\log_2 p(C_i, C_j) \tag{4}$$

其中,n为基聚类中簇的总个数.

将两个簇之间的Jaccard系数当作$p(C_i, C_j)$,Jaccard系数可以用来衡量两个簇之间的相似度. 具体如下:

$$p(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \tag{5}$$

其中,$|.|$代表集合中点的个数.

通过上面的方法可以对基聚类中每个簇的稳定性进行衡量,接下来提出一种基于信息熵的簇评价指标(Information Entropy Index,*IEI*),具体如下:

$$IEI(C_i) = \frac{\max(H) - H(C_i)}{\max(H) - \min(H)} \tag{6}$$

直观上,*IEI*指标反映了簇C_i里的点在其他基聚类中仍分在同一个簇内的可能性,*IEI*越大说明簇C_i里的点越有可能在其他基聚类中分到同一个簇中.

下一节利用*IEI*指数加权构建共协矩阵,以增强集成效果.

2.2 构建加权的共协矩阵 在得到每个簇的*IEI*指数后,利用它形成一个加权的共协矩阵$\boldsymbol{S}$,具体方法如下:

$$S(i,j) = \frac{1}{M}\sum_{m=1}^{M} w_i^m \cdot l \tag{7}$$

其中,M为基聚类的个数.

$$w_i^m = IEI(Cls(o_i)) \tag{8}$$

$$l = \begin{cases} 1, & \text{if } Cls(o_i) = Cls(o_j) \\ 0, & \text{其他} \end{cases} \tag{9}$$

其中,$Cls(o_i)$为对象o_i所在的簇.

通过上面这种方式可以构建一个加权的共协矩阵$\boldsymbol{S}$. 矩阵$\boldsymbol{S}$基于簇层面加权,充分考虑每个簇不同的稳定性,这有利于共识函数得到更好的聚类结果.

2.3 通过图分割得到最终结果 把得到的加权共协矩阵$\boldsymbol{S}$看作一个无向图. 具体来说,$N \times N$的共协矩阵可以看成一个有N个点的无向图,矩阵上的值就是无向图边的权值,对这个无向图进行图分割就能得到最终结果. 在图分割算法的选择上,因为归一化割(Normalized Cut,Ncut)是有效并且鲁棒的图分割算法,所以选择它作为本研究的图分割算法[10]. 归一化割是谱聚类的一种,其基本思想是定义一个割准则,该准则考虑了不同簇之间的总相异性和簇内的总相似度. 选用归一化割作为本算法的共识函数,通过对加权共协矩阵进行归一化割得到最终结果.

综上所述,IEWEC算法具体逻辑如下:

算法 基于信息熵加权的聚类集成算法

输入:数据集X,聚类数目K

1. 对数据集X运用聚类算法生成m个基聚类$P = \{p_1, p_2, \cdots, p_m\}$.

2. 仿造信息熵计算方法,通过式(4)和式(5)计算基聚类中每个簇的稳定性H.

3. 通过式(6),根据基聚类中每个簇的稳定性H计算基于信息熵的簇评价指标*IEI*.

4. 通过式(7)、式(8)和式(9),利用*IEI*指标形成一个加权的共协矩阵$\boldsymbol{S}$.

5. 将矩阵$\boldsymbol{S}$看成一个无向图,用Ncut算法对其进行图分割得到最终结果Label.

输出:最终结果Label

3 实验结果与分析

在多个数据集上进行实验,并与现有的若干聚类集成算法进行对比以验证本文算法的有效性,然后通过实验研究聚类集成规模对聚类结果的影响.

3.1 实验条件 实验均在MATLAB R2016a中实现. 配置如下:Windows7 64位操作系统,Intel i5双核1.7 GHz中央处理器,8 G内存.

首先生成包含M个基聚类的聚类集合,称M为聚类集成规模,并固定聚类集成规模M=100.

使用k-means算法生成基聚类,并在区间$\left[2,\sqrt{N}\right]$中随机选取k-means的聚类个数k. 然后将本文算法与其他聚类集成算法进行比较,并进一步测试在不同聚类集成规模下本文算法的聚类表现.

3.2　数据集　使用UCI(University of California Irvine)机器学习数据库中的八个数据集作为实验数据集,表1给出了数据集的详细信息.

表1　实验所用UCI数据集的属性

Table 1　The attributes of the UCI datasets used in experiments

数据集	样本数	类别数	特征数
Balancescale	625	3	4
Blood	748	2	4
Cardiotocograph (CTG)	2126	10	21
Dermatology	336	6	33
Ionosphere	351	2	34
Seeds	210	3	7
Wine	178	3	13
Zoo	101	7	16

3.3　对比算法　使用五种聚类集成算法与IEWEC算法进行对比:Evidence Accumulation Clustering(EAC)[5],Hybird Bipartite Graph Formulation(HBGF)[11],Weighted Connected-Triple(WCT)[21],Weighted Evidence Accumulation Clustering(WEAC)[22],Locally Weighted Evidence Accumulation(LWEA)[18].

为了实验的公平性,对比算法均采用相同的共识函数,即用Ncut算法作为共识函数.实验中如果对比算法有参数,则按照原始文献中的参考值进行设置.

3.4　评价标准　选取ARI(Adjusted Rand Index)和NMI(Normalized Mutual Information)来衡量聚类结果.

ARI是一种聚类评估算法[24],通过计算样本点对位于同一类簇和不同类簇的数目来度量两个聚类结果之间的相似程度,其计算式如下:

$$ARI=\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \tag{10}$$

其中,a表示在真实和实验情况下都属于同一个簇的点对数目,b表示在真实情况下属于同一个簇而在实验情况下不属于同一个簇的点对数目,c表示在真实情况下不属于同一个簇而在实验情况下属于同一个簇的点对数目,d表示在真实和实验情况下都不属于同一个簇的点对数目.ARI的取值范围为$[-1,1]$,值越大表明和真实结果越吻合,即聚类效果更好.

NMI是一种常见的聚类有效性外部评价指标[25],从信息论的角度评估了两个聚类结果的相似性.设实验结果为X,真实结果为Y,则其计算式如下:

$$NMI(X,Y)=\frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{11}$$

其中,$I(X,Y)$表示X和Y之间的互信息,$H(X)$和$H(Y)$表示X和Y的熵.NMI的取值范围为$[0,1]$,值越大表明和真实结果的共享信息越多,即聚类效果越好.

3.5　与其他聚类集成算法的对比实验　进行聚类集成算法的对比实验.在每个数据集中,每个聚类集成算法均运行20次,每次运行都根据3.1所述随机生成基聚类,得到聚类结果后计算相应的聚类评价标准的均值及其标准差.实验结果如表2和表3所示,表中的黑体字表示指标最高的值,括号内为标准差.

从表2和表3可以看出,在八个数据集上进行聚类实验,IEWEC算法实验结果的ARI仅在Blood上略逊色,在其他七个数据集上都是最高的;其NMI仅在CTG上略逊色,而且数值相差不大,在其他七个数据集上也都是最高的.

综上所述,IEWEC算法的总体性能优于其他方法.

3.6　在不同聚类规模下的鲁棒性　本节研究不同的聚类集成规模对IEWEC算法聚类结果的影响.在八个数据集上取不同数量的基聚类进行集成,基聚类规模为$[10,100]$,以10递增;基聚类的生成设置与3.1相同.

图2展示了不同聚类集成规模下IEWEC算法在八个数据集上的ARI指标.可以看出,在大部分数据集上,IEWEC算法仅有小幅波动,并在集成规模达到50以后逐渐趋于平稳.

图3展示了不同聚类集成数目下IEWEC算

表 2 不同算法的 *ARI* 指标比较

Table 2 The *ARI* of IEWEC and other algorithms

	EAC	HBGF	WCT	WEAC	LWEA	IEWEC
Balancescale	0.133 (0.060)	0.122 (0.054)	0.125 (0.054)	0.131 (0.066)	0.136 (0.103)	**0.138 (0.052)**
Blood	0.069 (0.016)	**0.077 (0.005)**	0.066 (0.005)	0.038 (0.016)	0.031 (0.017)	0.071 (0.025)
CTG	0.130 (0.006)	0.137 (0.008)	0.125 (0.004)	0.126 (0.006)	0.137 (0.007)	**0.142 (0.008)**
Dermatology	0.723 (0.040)	0.721 (0.052)	0.725 (0.071)	0.719 (0.014)	0.722 (0.103)	**0.730 (0.054)**
Ionosphere	0.150 (0.007)	0.164 (0.010)	0.168 (0.010)	0.149 (0.006)	0.169 (0.003)	**0.173 (0.008)**
Seeds	0.664 (0.071)	0.653 (0.055)	0.658 (0.055)	0.588 (0.046)	0.628 (0.055)	**0.696 (0.042)**
Wine	0.364 (0.025)	0.353 (0.051)	0.360 (0.001)	0.362 (0.024)	0.336 (0.079)	**0.374 (0.002)**
Zoo	0.677 (0.017)	0.703 (0.011)	0.704 (0.005)	0.683 (0.020)	0.700 (0.011)	**0.709 (0.008)**

表 3 不同算法的 *NMI* 指标比较

Table 3 The *NMI* of IEWEC and other algorithms

	EAC	HBGF	WCT	WEAC	LWEA	IEWEC
Balancescale	0.102(0.046)	0.107(0.045)	0.101(0.047)	0.105(0.051)	0.108(0.088)	**0.111(0.042)**
Blood	0.014(0.002)	0.016(0.002)	0.018(0.004)	0.012(0.001)	0.020(0.002)	**0.025(0.006)**
CTG	0.262(0.005)	**0.280(0.007)**	0.265(0.004)	0.261(0.005)	0.277(0.008)	0.268(0.006)
Dermatology	0.820(0.001)	0.874(0.011)	0.872(0.015)	0.874(0.014)	0.873(0.024)	**0.876(0.029)**
Ionosphere	0.115(0.004)	0.125(0.006)	0.127(0.005)	0.114(0.003)	0.127(0.003)	**0.129(0.004)**
Seeds	0.671(0.041)	0.669(0.034)	0.670(0.035)	0.628(0.028)	0.653(0.030)	**0.689(0.031)**
Wine	0.409(0.030)	0.416(0.036)	0.420(0.008)	0.412(0.030)	0.400(0.057)	**0.430(0.003)**
Zoo	0.756(0.007)	0.757(0.006)	0.759(0.003)	0.756(0.007)	0.758(0.006)	**0.761(0.004)**

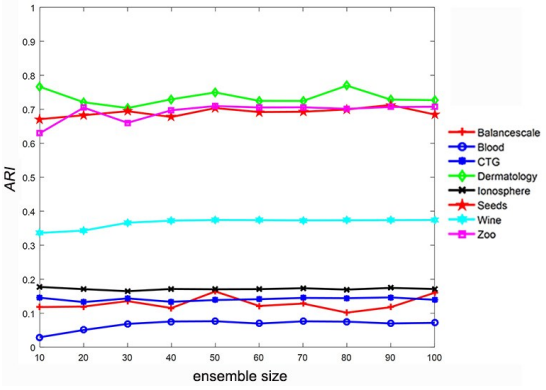

图 2 不同聚类集成规模下 IEWEC 算法的 *ARI* 值

Fig. 2 *ARI* values of IEWEC algorithm under different clustering ensemble sizes

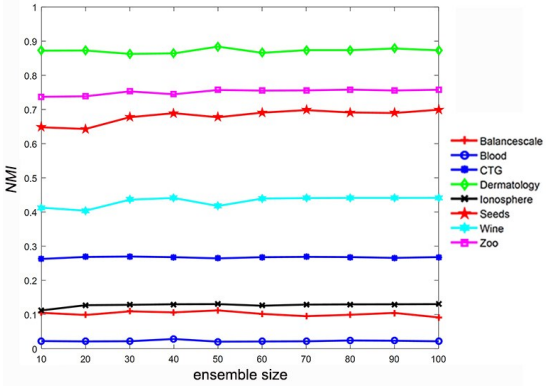

图 3 不同聚类集成规模下 IEWEC 算法的 *NMI* 值

Fig. 3 *NMI* values of IEWEC algorithm under different clustering ensemble sizes

法在八个数据集上的 *NMI* 指标. 可以看出,算法在所有数据集上的 *NMI* 均十分平稳,变化不大.

综上,聚类集成规模对于 IEWEC 算法影响不大. 在大多数数据集中 IEWEC 算法可以依靠较少的基聚类得到较稳健的结果,鲁棒性较好.

4 结　论

本文提出一种基于信息熵的聚类集成算法.首先,对每个簇的稳定性进行度量,为此设计了一种基于信息熵的簇评估标准,可以衡量一个簇里的点在其他基聚类中被分到一起的可能性;然后利用评估结果对共协矩阵进行加权,最后通过图划分得到最终结果.该算法有效地考虑了每个簇的不稳定性,避免了其他聚类集成算法只从基聚类层面进行加权的弊端.实验证明了此算法的有效性和鲁棒性.未来的工作是进一步探索聚类集成过程中基聚类的隐藏信息,并通过这些隐藏信息来提高聚类性能.

参考文献

[1] Ding S F, Jia H J, Du M J, et al. A semi-supervised approximate spectral clustering algorithm based on HMRF model. Information Sciences, 2018, 429: 215－228.

[2] Cong L, Ding S F, Wang L J, et al. Image segmentation algorithm based on superpixel clustering. IET Image Processing, 2018, 12(11): 2030－2035.

[3] Saini N, Saha S, Bhattacharyya P. Automatic scientific document clustering using self-organized multi-objective differential evolution. Cognitive Computation, 2019, 11(2): 271－293, doi: 10.1007/s12559-018-9611-8.

[4] Thanh N D, Ali M, Son L H. A novel clustering algorithm in a neutrosophic recommender system for medical diagnosis. Cognitive Computation, 2017, 9(4): 526－544.

[5] Fred A L N, Jain A K. Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835－850.

[6] Li E R, Li Q Y, Geng Y A, et al. Ensemble clustering using maximum relative density path// 2018 IEEE International Conference on Big Data and Smart Computing. Shanghai, China: IEEE, 2018, doi: 10.1109/BigComp.2018.00036.

[7] Rathore P, Bezdek J C, Erfani S M, et al. Ensemble fuzzy clustering using cumulative aggregation on random projections. IEEE Transactions on Fuzzy Systems, 2018, 26(3): 1510－1524.

[8] Zhong C M, Hu L Y, Yue X D, et al. Ensemble clustering based on evidence extracted from the co-association matrix. Pattern Recognition, 2019, 92: 93－106.

[9] Strehl A, Ghosh J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 2003, 3: 583－617.

[10] Fern X Z, Brodley C E. Solving cluster ensemble problems by bipartite graph partitioning// Proceedings of the 21st International Conference on Machine Learning. New York, NY, USA: ACM, 2004, doi: 10.1145/1015330.1015414.

[11] Shi J B, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888－905.

[12] Huang D, Wang C D, Wu J S, et al. Ultra-scalable spectral clustering and ensemble clustering. IEEE Transactions on Knowledge and Data Engineering, 2019, 32(6): 1212－1226, doi: 10.1109/TKDE.2019.2903410.

[13] Topchy A, Jain A K, Punch W. Clustering ensembles: models of consensus and weak partitions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1866－1881.

[14] Cristofor D, Simovici D. Finding median partitions using information-theoretical-based genetic algorithms. Journal of Universal Computer Science, 2002, 8(2): 153－172.

[15] Wu J J, Liu H F, Xiong H, et al. K-means-based consensus clustering: a unified view. IEEE Transactions on Knowledge and Data Engineering, 2014, 27(1): 155－169.

[16] Huang D, Lai J H, Wang C D. Ensemble clustering using factor graph. Pattern Recognition, 2016, 50: 131－142.

[17] Yu Z W, Li L, Gao Y J, et al. Hybrid clustering solution selection strategy. Pattern Recognition, 2014, 47(10): 3362－3375.

[18] Huang D, Wang C D, Lai J H. Locally weighted ensemble clustering. IEEE Transactions on Cybernetics, 2017, 48(5): 1460－1473.

[19] Li Z G, Wu X M, Chang S F. Segmentation using superpixels: a bipartite graph partitioning approach// 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012:789－796.

[20] Huang D, Wang C D, Lai J H. LWMC: a locally weighted meta-clustering algorithm for ensemble clustering//International Conference on Neural Information Processing. Springer Berlin Heidelberg, 2017:167－176.

[21] He N N, Huang D. Meta-cluster based consensus clustering with local weighting and random walking// International Conference on Intelligent Science and Big Data Engineering. Springer Berlin Heidelberg, 2019:266－277.

[22] Iam-On N, Boongoen T, Garrett S M, et al. A link-based approach to the cluster ensemble problem. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(12):2396－2409.

[23] Huang D, Lai J H, Wang C D. Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis. Neurocomputing, 2015, 170:240－250.

[24] Santos J M, Embrechts M. On the use of the adjusted rand index as a metric for evaluating supervised classification//International Conference on Artificial Neural Networks. Springer Berlin Heidelberg, 2009: 175－184.

[25] Vinh L T, Lee S, Park Y T, et al. A novel feature selection method based on normalized mutual information. Applied Intelligence, 2012, 37(1): 100－120.

(责任编辑 杨可盛)