第 55 卷　第 4 期
2019 年 7 月

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

Vol. 55, No. 4
July, 2019

DOI:10.13232/j.cnki.jnju.2019.04.009

基于样本对加权共协关系矩阵的聚类集成算法

王　彤[1],魏　巍[1,2*],王　锋[1,2]

(1.山西大学计算机与信息技术学院,太原,030006;
2.山西大学计算智能与中文信息处理教育部重点实验室,太原,030006)

摘　要:聚类集成的目标是通过集成多个聚类结果来提高聚类算法的稳定性、鲁棒性以及精度.近些年,聚类集成受到了越来越多的关注.现有的集成聚类通常平等地对待所有基聚类,而不考虑它们的重要度.虽然学者们已经在这一方面做出了一些努力,例如使用加权策略来改进共协关系矩阵,但无论是给基聚类加权还是对类重要度评价时都忽略了样本对于其所在类贡献的差异.为此,提出了基于样本对加权共协关系矩阵的聚类集成算法,该算法利用k-means算法产生多个基聚类结果,然后对于其中的每个类再利用k-means算法产生多个小类,并计算去掉样本对所在的小类后类的不确定性变化的程度来评价该样本对的重要度,最后通过层次聚类算法得到聚类结果.在六个UCI数据集上的实验结果表明,基于样本对加权共协关系矩阵的聚类集成算法的性能优于三种经典的基于共协关系矩阵的聚类集成算法.

关键词:聚类,聚类集成,共协矩阵,加权策略
中图分类号:TP391　　**文献标识码:**A

Sample pairwise weighting co-association matrix based ensemble clustering algorithm

Wang Tong[1], Wei Wei[1,2*], Wang Feng[1,2]

(1. School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, China;
2. Key Laboratory of Computation Intelligence and Chinese Information Processing, Ministry of Education, Shanxi University, Taiyuan, 030006, China)

Abstract: The goal of clustering ensemble is to improve the stability, robustness and accuracy of the final clustering results by integrating multiple clustering results. In recent years, clustering ensemble has attracted more and more attention. One limitation of most existing clustering ensemble methods is that they generally treat all base clustering equally, regardless of their importance. Although scholars have made some efforts in this aspect, for example, the weighted strategy is used to improve the co-association matrix. However, they ignore the difference in the contribution of samples to the classes they belong to when either weighting the base clustering or evaluating the class importance. Therefore, sample pairwise weighting co-association matrix based ensemble clustering algorithm is proposed. The algorithm firstly uses the k-means algorithm to generate multiple base partition results and multiple small classes for each class. The importance of the sample to the class is evaluated by calculating the change degree of uncertainty of the

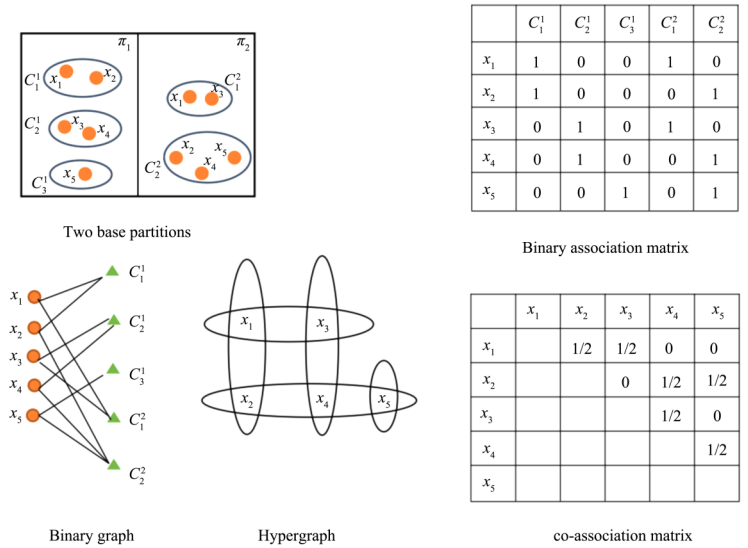
基金项目:国家自然科学基金(61772323,61303008,61603229,61502288),山西省高等教育机构科技创新项目(2016111)
收稿日期:2019－05－14
*通讯联系人,E-mail:weiwei@sxu.edu.cn

class after removing the subclass of the sample pairwise. Finally, the final clustering result can be obtained through the hierarchical clustering algorithm. Experimental results on six UCI data sets show that the performance of sample pairwise weighting co-association matrix based clustering ensemble algorithm is superior to the three classical clustering ensemble algorithms based on co-association matrix.

Key words: clustering, clustering ensemble, co-association matrix, weighted strategy

聚类分析是数据挖掘的一个基本问题,它能够根据相似性将一组未知分布的数据进行分类,同时也能够发现隐藏在数据中的结构.其中,数据聚类在数据挖掘以及机器学习的领域仍然是一个十分重要且具有挑战性的问题.数据聚类其实就是将物理或抽象样本集合分成由类似的样本组成的多个类的过程,并能够达到同一类中样本相似度高、不同类间样本相似度低的目的.在过去的十几年中,研究人员提出了很多聚类算法,有基于划分的方法[1-2]、基于层次的方法、基于密度的方法[3-4]、基于网格的方法[5]以及基于模型的方法.由于每个算法都有其特有的优化标准,对于特定数据结构以及类的形状才能够有很好的性能,也就是说没有一种单一算法能够适用于任意的数据结构及类型,这也是聚类集成出现的原因.

聚类集成根据多个聚类结果找到一个新的数据划分,这个划分在最大程度上共享了所有输入的聚类结果对数据集的聚类信息.相较于使用单一算法得到聚类结果,集成聚类的目标在于融合多个聚类结果以得到更优、更鲁棒聚类.常用的聚类集成的方法可分为基于图的方法、基于关系矩阵的方法、基于二部图的方法.基于图的方法是用图分割技术来挖掘数据的类结构,如超图[6]、Link[7]等.基于关系矩阵的方法是通过关系矩阵反映两两样本间的关系,然后再基于该矩阵设计集成算法,如共协关系矩阵[8].基于二部图的方法是通过构造样本与类的关联矩阵,利用图聚类方法(如谱聚类算法[9])进行集成.图1举例说明了两个基聚类的四种表现形式:超图、共协关系矩阵、二部图和二部关联矩阵.超图[6]中顶点代表数据样本,一个闭合的曲线是一个超边,代表一个基聚类.在共协关系矩阵[8]中,矩阵的行和列都是数据



	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2
x_1	1	0	0	1	0
x_2	1	0	0	0	1
x_3	0	1	0	1	0
x_4	0	1	0	0	1
x_5	0	0	1	0	1

	x_1	x_2	x_3	x_4	x_5
x_1		1/2	1/2	0	0
x_2			0	1/2	1/2
x_3				1/2	0
x_4					1/2
x_5					

图1　两个基聚类的四种表示方法

Fig. 1　Four representations of two base partitions

样本个数,对应的值是两个样本在整个集成中出现在一个类中的平均频率. 二度图[10]中顶点代表数据样本和类,数据样本和类之间的线代表数据样本属于这个类. 在二度关联矩阵中,矩阵的行是类的个数,列是数据样本的个数,若数据样本属于该类,对应元素是1,否则为0.

聚类集成一般包括三个步骤:基聚类的生成、基聚类的加权和通过一致性函数得到最后的划分. 基聚类的生成主要有几种方法:(1)通过使用相同聚类算法,每次运行都设置不同的参数和随机初始化生成基聚类;(2)使用不同的聚类算法,如k-means、层次聚类等生成基聚类;(3)使用同样的算法在数据集的不同样本集合上聚类生成基聚类;(4)在数据集的不同特征子集或在数据集的不同子空间的投影上聚类生成基聚类. 这些方法可以单独使用,也可以任意组合. 本文采用的是第一种方法生成基聚类.

集成聚类算法中通过对基聚类进行加权可以更好地平衡各基聚类对最终聚类结果的贡献,得到性能更好的集成聚类算法,再通过一致性函数来得到最后的划分. 最常用的方法有k-means[1],PAM(Partitioning Around Medoid)[11]以及层次聚类的单连接、平均连接等方法. 对于共协关系矩阵来说,任何一种聚类算法都可用于产生最终的聚类结果.

在过去的几年中,学者们提出许多聚类集成的方法[8,12-18]. Fred and Jain[8]提出Evidence Accumulation Clustering(EAC)算法,构建共协关系矩阵,若一对样本出现在一个类中,值为1,否则为0,最后通过层次聚类的方法得到最后的聚类结果. 这一方法平等地对待基聚类、类以及样本,但忽略了它们不同的贡献. 许多学者为了改善这一情况,通过改进共协关系矩阵从而更大程度地利用聚类结果所给的信息,得到最后的聚类结果. 其中,有通过评价基聚类的重要度来改进共协关系矩阵的一些方法,Yang and Chen[12]通过融合三个不同的评价聚类的指标来给基聚类一个权值,从而得到加权的共协关系矩阵. Nanda and Pujari[13]也提出了一个加权聚类集成的方法,通过计算样本和类中心的距离与类中心之间的距离的差值来评价一个基聚类的重要度,给基聚类一个权值,从而得到加权的共协关系矩阵. 这一类方法造成在一个划分中所有被分到同一个类中样本对的值都是相同的结果,忽略了类与样本对于基聚类的贡献. 为了改善这一问题,学者们又提出通过评价类的重要度来改进共协关系矩阵的一些方法. Zhong et al[14]通过计算样本之间的距离来评价一个类的可靠度. Huang et al[15]用熵来评价一个类的不确定程度从而得到类的重要度来给每对样本一个权值,但这样会使出现在同一个类中的所有样本给定的权值都是一样,也就是说在同一个类中的每个样本都被平等对待,显然,这样不能精准评价样本的贡献程度.

因此,本文提出了一种基于样本对加权共协关系矩阵的聚类集成算法,利用k-means算法对类进一步分类,产生若干个小类,通过考虑去掉所要计算的该对样本所在的小类之后类的不确定程度的变化来更加精准地评价每个样本对于所在的类的贡献程度.

1 相关知识

设$X=\{x_1,x_2,\cdots,x_n\}$是样本的集合,其中,

$$x_i=\{x_{i1},x_{i2},\cdots,x_{id}\}$$

d为维度,n是数据样本的个数. 聚类集成就是在数据集X上重复地产生M个聚类结果,即:

$$\prod=\{\pi^1,\pi^2,\pi^3,\cdots,\pi^M\}$$

其中,

$$\pi^i=\{C_1^i,C_2^i,C_3^i,\cdots,C_j^i,\cdots,C_{k_i}^i\}$$

是第i个划分,C_j^i代表第i个基聚类的第j个类,k_i代表第i个划分中类的个数.

Fred and Jain[8]最先提出共协关系矩阵(***CM***矩阵),其定义如下:

$$\boldsymbol{CM}=\{m_{ij}\}_{n\times n} \tag{1}$$

其中,

$$m_{ij}=\frac{1}{M}\sum_{m=1}^{M}\delta_{ij}^m$$

$$\delta_{ij}^{m}=\begin{cases}1, & Cls^{m}(x_i)=Cls^{m}(x_j)\\ 0, & \text{otherwise}\end{cases}$$

$Cls^{m}(x_i)$代表样本x_i在基聚类π^{m}中所属于的类,δ_{ij}^{m}代表样本x_i和x_j在基聚类π^{m}中是否在同一个类中.若在一个类中,值为1,否则,值为0.

2　基于样本对加权的共协关系矩阵

传统的共协矩阵中,根据各样本对是否处于同一个类赋予其对应的矩阵元素0或1,各个样本对被同等对待.事实上,各样本对中的样本所处类的情况和各样本对其所在类的影响程度都不尽相同,平等地对待每个样本对有可能导致共协关系矩阵不能很好地反映各个基聚类的情况,进而影响到聚类集成的效果.为了解决这个问题,本文提出了一种样本对重要性度量,并基于该重要性度量定义新的共协矩阵.

由于样本对的重要性通常由各样本所处类的情况以及样本在其所在类中的地位决定,因此首先利用信息熵[15]给出关于基聚类的不确定性度量的定义:

$$H^{m}(C_t)=-\sum_{k=1}^{k_m}p(C_t,C_k^{m})\log_2 p(C_t,C_k^{m}) \quad (2)$$

其中,

$$p(C_t,C_k^{m})=\frac{|C_t\cap C_k^{m}|}{|C_t|}$$

是$C_t\cap C_k^{m}$类C_t与类C_k^{m}中相同样本的个数,$|C_t|$是类中样本的个数.显然,$p(C_t,C_k^{m})$的取值范围是[0,1],$H^{m}(C_t)$的取值范围是[0,∞].显然,$p(C_t,C_k^{m})$越大,$H^{m}(C_t)$越小,也就是C_t相对于π^{m}这个基聚类不确定性度量越小.

此外,由于基聚类之间是独立的,因此,C_t相对于所有基聚类的不确定性度量定义为:

$$H^{\Pi}(C_t)=\sum_{m=1}^{M}H^{m}(C_t) \quad (3)$$

由式(3)容易得出H^{m}的取值范围是[0,∞],H^{m}值越大,类C_t的不确定性就越大.

一个类相对于基聚类的不确定性实际上是由该类中样本造成的,容易想到每个样本对于类的不确定性度量的影响程度不同.为此,本文基于将一个样本从某个类中去掉后该类不确定性变化的程度,给出这个样本的权重.但是在实际应用中,由于单个样本对于整个类的影响过小,导致不同样本之间的权重差异过小.为此,利用k-means算法将每个类划分成若干个小类(其中,k-means算法的k值,根据类的规模设置),然后以这些小类为最小单位,通过从类中去除这些小类引起的类不确定性度量变化的程度评价样本对的重要性.

根据上述分析给出样本x_i和x_j重要性度量的定义:

$$H^{\Pi}(C_t,x_i,x_j)=H^{\Pi}(C_t)-H^{\Pi}(C_t|(x_i,x_j)) \quad (4)$$

其中,

$$H^{\Pi}(C_t|(x_i,x_j))=\sum_{m=1}^{M}H^{m}(C_t|(x_i,x_j))$$

$$H^{m}(C_t|(x_i,x_j))=-\sum_{m=1}^{k_m}p((C_t|(x_i,x_j)),(C_k^{m}|(x_i,x_j)))\log_2 p((C_t|(x_i,x_j)),(C_k^{m}|(x_i,x_j)))$$

$C_t|(x_i,x_j)$代表类C_t去掉x_i和x_j所在的小类中的所有样本之后的类.若x_i和x_j被分在同一个小类里,就去掉一个小类,被分到不同的小类中,就去掉两个小类.

由于去掉了与x_i和x_j在同一小类中的样本,类C_t的不确定性度量会有所减小,只不过有的减少得多,有的减少得少,所以由式(4)可以看出$H^{\Pi}(C_t,x_i,x_j)$的取值范围是[0,∞],并且若$H^{\Pi}(C_t,x_i,x_j)$越大,则从C_t中去掉样本x_i和x_j所在的小类后,C_t的不确定性减小的程度就越大.也就是说,样本x_i和x_j所在小类C_t导致的不确定性变小的程度很大,因此样本对x_i和x_j的权值应该很小.基于这样的分析,给出样本对x_i和x_j的权重如下:

$$w(C_t,x_i,x_j)=\exp(-H(C_t,x_i,x_j)/(\theta\times M)) \quad (5)$$

其中,参数$\theta>0$,是用来调节样本对的不稳定性对于最后的权值的影响.

由式(5)可以看出权值$w(C_t,x_i,x_j)$的取值范围是[0,1],样本对导致类不确定性变化的程度越大,给的权值就越低.图2给出了不同的参

数θ对于权值的影响,可以看到当$\theta<1$时,随着$H^{\Pi}(C_t,x_i,x_j)$的增加$w(C_t,x_i,x_j)$会显著地减少,当$\theta>4$时,随着$H^{\Pi}(C_t,x_i,x_j)$的增加$w(C_t,x_i,x_j)$减少的速度会很缓慢.

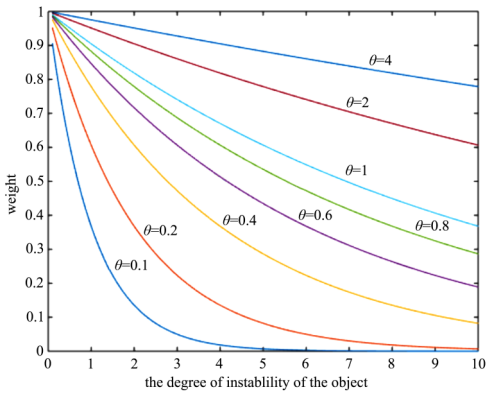


图 2 对于不同参数,权值与样本对不稳定程度之间的相关性

Fig. 2 For different parameters, the correlation between weight and the degree of instability of the object

根据式(5),可以定义基于样本加权的共协关系矩阵$\boldsymbol{WCM}$为:

$$\boldsymbol{WCM}=\{m_{ij}\}_{n\times n} \tag{6}$$

其中,

$$m_{ij}=\frac{1}{M}\sum_{m=1}^{M}w_{ij}^{m}\delta_{ij}^{m}$$

$$w_{ij}^{m}=w(Cls(x_i),x_i,x_j)$$

$$\delta_{ij}^{m}=\begin{cases}1, & Cls^{m}(x_i)=Cls^{m}(x_j)\\ 0, & \text{otherwise}\end{cases}$$

其中w_{ij}^{m}代表样本x_i和x_j对所在类C_t的权值.

3 算法描述

基于第2节给出的基于样本对加权的共协关系矩阵设计了基于加权共协矩阵的聚类集成算法OWEC(Sample pairwise Weighting co-association matrix based Ensemble Clustering algorithm),该算法的详细描述如下:

算法1 基于样本对加权共协关系矩阵的聚类集成算法

输入:数据集$X=\{x_1,x_2,\ldots,x_n\}$

小类的个数$const_k$

聚类结果中类的个数K

输出:聚类结果π^*

步骤1. 利用k-means算法得到M个基聚类结果.

步骤2. 对于每一个在$\prod$集合中的类,用k-means算法得到$const_k$个小类.

步骤3. 利用式(6)计算基聚类集合中所有的类的不确定性度量.

步骤4. 计算样本x_i和x_j以及类C_t所对应的权值.

步骤5. 计算并得到加权的共协关系矩阵.

步骤6. 利用层次聚类算法得到最后的聚类结果.

4 实验与结果

4.1 数据集 为了验证提出的改进聚类集成算法的有效性,选取六个UCI数据集来进行实验,数据集的详细描述如表1所示.

表1 数据集

Table 1 Datasets

Datasets	Objects	Dimensions	Classes
Iris	150	4	3
Wine	178	13	3
Column_3C	310	6	3
Glass	214	9	6
Musk	476	167	2
Seeds	210	7	3

4.2 评价指标 使用调整兰德系数(Adjusted Rand Index,ARI)和聚类精度(Clustering Accuracy,CA)来评价聚类算法的性能.

假设$\pi'=\{\pi'_1.....\pi'_{K_s}\}$是真实的基聚类结果,$\pi^G=\{\pi_1^G.....\pi_{k_t}^G\}$是实验得到的基聚类结果,其中,$k_s$和$k_t$是$\pi'$和$\pi^G$的类别个数.

ARI通过考虑在两个基聚类中,样本对是否在相同的类来计算得到,其计算公式如下[19]:

$$ARI(\pi',\pi^G)=\frac{2(N_{00}N_{11}-N_{01}N_{10})}{(N_{00}+N_{01})(N_{01}+N_{11})(N_{00}+N_{10})(N_{10}+N_{11})} \tag{7}$$

其中,N_{11}代表在基聚类π'和π^G中都出现在相同的类的样本对的个数,N_{00}代表在基聚类π'和π^G

中,都出现在不同的类的样本对的个数,N_{01}代表在基聚类π^G中出现在相同的类而在基聚类π'出现在不同的类的样本对的个数,N_{10}代表在基聚类π^G中出现在不同的类但在基聚类π'出现在相同类的样本对的个数.

*CA*的计算公式如下[20]:

$$CA(\pi',\pi^G)=\frac{1}{N}\sum_{i=1}^{k_t}|\pi_i^G\cap \mathrm{mode}(\pi_i^G,\pi')| \quad (8)$$

其中,

$$\mathrm{mode}(\pi_i^G,\pi')=\mathrm{argmax}_{\pi_j'\in\pi'}|\pi_i^G\cap\pi_j'|$$

4.3　实验和结果　通过三组实验验证OWEC的性能.第一组实验研究参数的取值变化对提出方法性能的影响,第二组实验将OWEC算法与EAC[8],LWEA(Locally Weighted Evience Accumulation)[15]和LWGP(Locally Weighted Graph Partitioning)[15]算法进行比较.第三组实验研究集成规模对提出算法性能的影响.

首先,先分析参数对OWEC算法性能的影响.对于参数的每一个值,分别运行OWEC算法20次,求得平均的*ARI*和*CA*值.其中M=30,用k-means算法随机初始化产生基聚类,k是固定的,是每个数据集真实的k值.用k-means算法产生小类,k是固定的根据具体的数据集来变化,其中Iris,Wine,Seeds数据集设置$const_k$=10,Column_3C数据集设置$const_k$=20,Glass数据集设置$const_k$=3,Musk数据集设置$const_k$=30.结果如表2和表3所示.从结果上来看,对于Iris数据集,θ的取值不论是怎样变化,*ARI*和*CA*值的变化都不明显,而对于其余的五个数据集,θ在0.1~1时,*ARI*和*CA*值的变化不明显,因此θ对于我们提出的算法的影响并不大,建议在[0.2,1]范围内选取.在之后的实验中,θ的值设置为0.3.

接下来,将本文提出的OWEC算法与EAC[8],LWEA[15]与LWGP[15]算法进行比较.

表2　对于不同的参数OWEC算法的平均性能(*ARI*)

Table 2　Average performance of OWEC algorithm for different parameters(*ARI*)

Datasets	θ											
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	2.0	4.0
Iris	0.6249	0.6141	0.6183	0.6167	0.6107	0.5985	0.6003	0.6170	0.6124	0.6023	0.5937	0.5976
Wine	0.3238	0.3680	0.3721	0.3796	0.3794	0.3703	0.3498	0.3511	0.3595	0.3372	0.3201	0.3394
Column_3C	0.3616	0.4057	0.4403	0.4234	0.4320	0.4196	0.4066	0.3800	0.4495	0.4206	0.4186	0.4226
Glass	0.2732	0.3005	0.3357	0.3115	0.3094	0.3503	0.3145	0.3388	0.3038	0.3388	0.3038	0.7303
Musk	0.1687	0.2914	0.2592	0.2639	0.2511	0.2644	0.2949	0.2876	0.2949	0.2355	0.2086	0.2444
Seeds	0.3631	0.4283	0.4573	0.4575	0.4615	0.4861	0.4789	0.4352	0.5461	0.5153	0.4891	0.4681

表3　对于不同的参数OWEC算法的平均性能(*CA*)

Table 3　Average performance of OWEC algorithm for different parameters(*CA*)

Datasets	θ											
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	2.0	4.0
Iris	0.8567	0.8490	0.8603	0.8543	0.8480	0.8410	0.8443	0.8467	0.8560	0.8477	0.8463	0.8430
Wine	0.6646	0.6820	0.6713	0.6904	0.6483	0.6789	0.6646	0.6612	0.6840	0.6621	0.6590	0.6671
Column_3C	0.7152	0.7471	0.7690	0.7563	0.7571	0.7426	0.7450	0.7410	0.7687	0.7519	0.7518	0.7556
Glass	0.6993	0.7008	0.7098	0.7178	0.7136	0.7203	0.7231	0.7376	0.7210	0.7376	0.7210	0.3318
Musk	0.6818	0.7569	0.7420	0.7291	0.7232	0.7255	0.7450	0.7561	0.7561	0.7243	0.7143	0.7255
Seeds	0.7140	0.7593	0.7731	0.7743	0.7702	0.7864	0.7807	0.7624	0.8188	0.7993	0.7881	0.7793

对于每一个测试的方法,都是用 k-means 算法固定的 k 值随机初始化产生基聚类,k 值是数据集的真实类别个数. 实验中基聚类个数 M=100. 为保证实验结果的有效性每个算法运行 20 次,得到平均 ARI 值和 CA 值,并且为了实验的公平公正,每一次运行,OWEC 与现有的三个算法都是在相同的基聚类上进行的. 其中,对于 LWEA 算法和 LWGP 算法选取参数 θ=0.4,OWEC 算法选取参数 θ=0.3. 实验结果如表 4 所示. OWEC 算法在 Iris, Glass, Seeds, Musk 数据集上的 ARI 值和 CA 值得到了明显的提升,在 Wine, Column_3C 数据集上的性能虽然没有明显的提升,但是也是比所比较的方法好的. 说明 OWEC 还是有很大优势的.

表 4　不同的算法在不同数据集上的平均性能(ARI,CA)

Table 4　Average performance of different algorithms over different data sets (ARI,CA)

Datasets	ARI				CA			
	OWEC	EAC	LWEA	LWGP	OWEC	EAC	LWEA	LWGP
Iris	0.6324	0.4394	0.4634	0.4400	0.8527	0.6533	0.7527	0.7413
Wine	0.3632	0.3045	0.3331	0.3295	0.6607	0.6348	0.6354	0.6399
Column_3C	0.4320	0.3054	0.3963	0.3675	0.7571	0.6935	0.7326	0.6944
Glass	0.3581	0.2092	0.2701	0.2520	0.7290	0.6782	0.6449	0.6729
Musk	0.2876	0.0291	0.1785	0.1687	0.7561	0.5924	0.6893	0.6890
Seeds	0.5016	0.1317	0.2522	0.2385	0.7943	0.5286	0.5952	0.5886

进一步分析 OWEC 算法以及 EAC, LWEA 和 LWGP 算法在不同集成规模下的性能差异. 对每一个集成大小 M,在每个数据集上分别运行每个方法 20 次,计算出平均性能(ARI 和 CA 值),其中 LWEA,LWGP 算法参数仍然取 0.4,OWEC 方法参数取 0.3. 不同集成规模 M 下不同数据集中各个方法的平均性能 ARI 和 CA 值如图 3 和图 4 所示,可以看出随着集成规模的变化,OWEC 算法在各个数据集上达到最好的 ARI 和 CA 值,整体的性能最好.

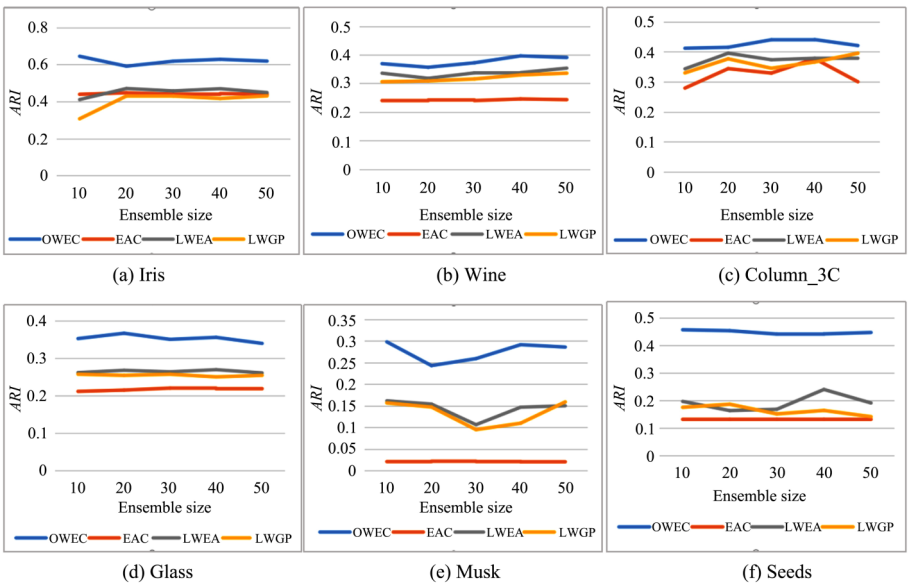


图 3　在不同集成规模下不同的算法在各个数据集上的平均性能(ARI)

Fig. 3　Average performance of different algorithms on each data set under different ensemble size (ARI)

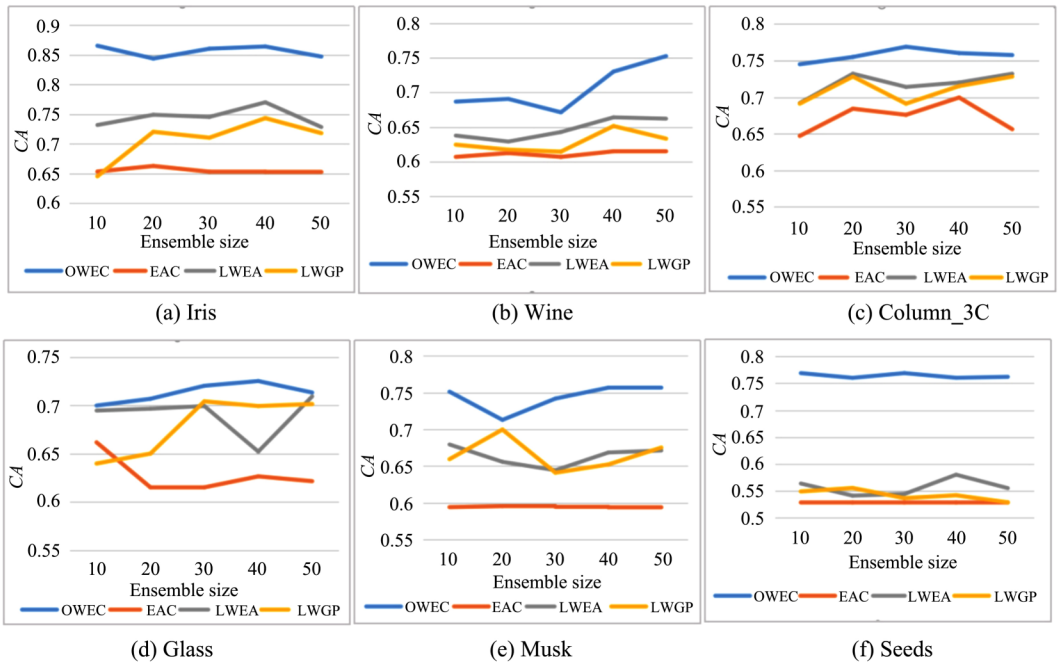


图4　在不同集成规模下不同的算法在各个数据集上的平均性能(*CA*)

Fig. 4　Average performance of different algorithms on each data set under different ensemble size (*CA*)

5　结束语

本文提出了基于样本对加权共协关系矩阵的聚类集成算法,该算法首先利用*k*-means算法得到多个基聚类结果,然后利用熵计算样本对于类的影响程度,得到权值后的共协关系矩阵,最后利用层次聚类算法来得到最终的集成聚类结果.由于该算法在构建共协关系矩阵时考虑了不同样本对对于所在类的重要程度,因此提出的共协关系矩阵能够更好地刻画各基聚类的情况.在UCI数据集上的实验结果表明,提出的算法在调整兰德系数(*ARI*)和聚类精度(*CA*)两个指标上都在一定程度上优于对比算法.

参考文献

[1] Van Ham F. Using Multilevel call matrices in large software projects//IEEE Symposium on Information Visualization 2003. Seattle, WA, USA:IEEE,2003:227－232.

[2] Furnas G F. Generalized fisheye views// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York,NY,USA:ACM,1986:16－23.

[3] Schaffer D, Zao Z P, Greenberg S, et al. Navigating hierarchically clustered networks through fisheye and full‐zoom methods. ACM Transaction on Computer‐Human Interaction, 1996,3(2):162－188.

[4] 贾云得,吕宏静,刘万春.鱼眼变形立体图像恢复稠密深度图的方法.计算机学报,2000,23(12):1332－1336. (Jia Y D,Lu H J,Liu W C. Fish-eye lens camera stereo vision for dense depth map recovery. Chinese Journal of Computers, 2000, 23 (12):1332－1336.)

[5] Van Hee K, Sidorova N, Voorhoeve M. Soundness and separability of workflow nets in the stepwise refinement approach//Proceedings of the 24th International Conference on Application and Theory of Petri Nets. Springer Berlin Heidelberg, 2003,2679:337－356.

[6] Strehl A, Ghosh J. Cluster ensembles‐a knowledge reuse framework for combining

multiple partitions. The Journal of Machine Learning Research, 2003, 3: 583－617.

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

[8] 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.

[9] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: analysis and an algorithm// Proceedings of the 14th International Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2002: 849－856.

[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: 36－43.

[11] Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis. Hoboken: John Wiley and Sons Inc, 1990.

[12] Yang Y, Chen K. Temporal data clustering via weighted clustering ensemble with different representations. IEEE Transactions on Knowledge and Data Engineering, 2010, 23(2): 307－320.

[13] Nanda A, Pujari A K. Weighted co-clustering based clustering ensemble// 3rd National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics. Hubli, India: IEEE Press, 2011: 46－49.

[14] Zhong C M, Yue X D, Zhang Z H, et al. A clustering ensemble: Two-level-refined co-association matrix with path-based transformation. Pattern Recognition, 2015, 48(8): 2699－2709.

[15] Huang D, Wang C D, Lai J H. Locally weighted ensemble clustering. IEEE Transactions on Cybernetics, 2018, 48(5): 1460－1473.

[16] 黄栋, 王昌栋, 赖剑煌等. 基于决策加权的聚类集成算法. 智能系统学报, 2016, 11(3): 418－425. (Huang D, Wang C D, Lai J H, et al. Clustering ensemble by decision weighting. CAAI Transactions on Intelligent Systems, 2016, 11(3): 418－425.)

[17] Berikov V, Pestunov I. Ensemble clustering based on weighted co-association matrices: error bound and convergence properties. Pattern Recognition, 2017, 63: 427－436.

[18] Nazari A, Dehghan A, Nejatian S, et al. A comprehensive study of clustering ensemble weighting based on cluster quality and diversity. Pattern Analysis and Applications, 2019, 22(1): 133－145.

[19] Rand W M. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 1971, 66(336): 846－850.

[20] Nguyen N, Caruana R. Consensus clusterings// Proceedings of the 7th IEEE International Conference on Data Mining. Omaha, NE, USA: IEEE, 2007: 607－612.

(责任编辑 杨可盛)