计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700213

基于差异性度量的基础聚类三支过滤算法

梁 伟[1,2] 段晓东[1] 徐健锋[1,3,4]
1 南昌大学软件学院 南昌 330047
2 华南理工大学软件学院 广州 510006
3 同济大学电子与信息工程学院 上海 201804
4 泰豪软件股份有限公司 南昌 330096
(416627317007@email.ncu.edu.cn)

摘 要 基础聚类成员预处理是聚类集成算法中的一个重要研究步骤。众多研究表明,基础聚类成员集合的差异性会影响聚类集成算法性能。当前聚类集成研究围绕着生成基础聚类和优化集成策略展开,而针对基础聚类成员的差异性度量及其优化的研究尚不完善。文中基于Jaccard相似性提出一种基础聚类成员差异性度量指标,并结合三支决策思想提出了基础聚类成员差异性三支过滤方法。该方法首先设定基础聚类成员的三支决策的初始阈值$\alpha^{(0)}$和$\beta^{(0)}$,然后计算各个基础聚类成员的差异性度量指标,进而实施三支决策。其决策策略为:当基础聚类成员的差异性度量指标小于指定阈值$\alpha^{(0)}$时,删除该基础聚类成员;当基础聚类成员的差异性度量指标大于指定阈值$\beta^{(0)}$时,保留该基础聚类成员;当基础聚类成员的差异性度量指标大于$\alpha^{(0)}$且小于$\beta^{(0)}$时,该基础聚类成员被归入三支决策边界域等待进一步判断。当结束一轮三支决策后,算法将重新计算三支决策阈值$\alpha^{(1)}$和$\beta^{(1)}$并对上轮三支决策边界域重新进行三支决策,直至没有基础聚类成员被归入三支决策边界域或达到指定迭代次数。对比实验表明基础差异性度量的基础聚类三支过滤方法能够有效地提升聚类集成效果。

关键词:基础聚类过滤;三支决策;三支优化;聚类集成;差异性度量
中图法分类号 TP18

Three-way Filtering Algorithm of Basic Clustering Based on Differential Measurement

LIANG Wei[1,2], DUAN Xiao-dong[1] and XU Jian-feng[1,3,4]
1 School of Software, Nanchang University, Nanchang 330047, China
2 School of Software Engineering, South China University of Technology, Guangzhou 510006, China
3 College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
4 Tellhow Software Co., LTD, Nanchang 330096, China

Abstract The pre-processing of basic clustering members is an important research step in the ensemble clustering algorithm. Numerous studies have shown that the difference in the set of basic clustering members affects the performance of the ensemble clustering. The current ensemble clustering research revolves around the generation of basic clustering and the integration of basic clustering, while the differential measurement and optimization of basic clustering members are not perfect. Based on Jaccard's similarity, this study proposes a measurement for the differential of basic clustering members and constructs a differential three-way filtering method for basic clustering members by introducing the three-way decisions idea. This method first sets the initial thresholds $\alpha^{(0)}$ and $\beta^{(0)}$ of the three-way decisions for basic clustering members and then calculates the differential of each basic clustering member to implement the three-way decisions. Its decision strategy is: when the differential metric of the basic clustering member is less than the specified threshold $\alpha^{(0)}$, the basic clustering member will be deleted; when the differential metric of the basic clustering member is greater than the specified threshold $\beta^{(0)}$, the basic clustering member will be retained; and when the differential metric of the basic clustering member is greater than $\alpha^{(0)}$ and less than $\beta^{(0)}$, the basic clustering member will be added into the boundary domain of the three-way decisions, and boundary domains will be further judged by the three-way decisions with new thresholds. After completing a round of the three decisions, the algorithm recalculates thresholds of the three-way decisions and remakes the three-way decisions on boundary domains of the three-way decisions remained in the last round until no

到稿日期:2020-07-31 返修日期:2020-08-30 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61763031);江西省自然科学基金资助项目(20202BAB202018)
This work was supported by the National Natural Science Foundation of China(61763031) and Jiangxi Provincial Natural Science Foundation (20202BAB202018).
通信作者:徐健锋(jianfeng_x@ncu.edu.cn)

basic clustering member is added to boundary domains of the three-way decisions or the specified number of iterations is reached. The comparative experiment shows that the differential measurement three-way filtering method for basic clustering can effectively improve the performance of ensemble clustering.

Keywords　Basic clustering filtering, Three-way decision, Three-way optimization, Clustering ensemble, Differential measurement

1　引言

聚类集成通过融合不同版本的基础聚类来实现无监督学习任务[1]。聚类集成的算法框架由基础聚类生成和聚类成员集成这两个关键步骤组成，这两个关键步骤决定了聚类集成算法的性能。首先，聚类集成算法需要产生一组合适规模的基础聚类成员集合；其次，通过高效的集成策略（一致性函数）聚合这组成员集合。集成策略能尽可能地提取各个聚类成员的信息，从而生成一个能代表所有成员的一致性结果[2]。当前，对聚类集成问题的研究也主要围绕着这两个步骤展开。

基础聚类成员的产生主要有两种思路：一是采用同一聚类方法设置不同的初始参数[3-4]或采用不同聚类方法作用于同一数据集[5-6]，从而获得一组基础聚类成员；二是对数据集进行多种非等值变形，进而得到一组基础聚类成员，变形包括数据集投影[7-9]和采样操作处理[10-12]等。

一致性函数集成的本质就是通过有效的组合策略将基础聚类成员高效地组合起来。可以用作一致性函数的相关技术主要有超图[13]、信息论[14]、关联矩阵[15]、投票[16]等。

此外，一些研究发现[17-20]，基础聚类成员的质量、规模和差异性会在一定程度上影响聚类集成性能。同时，数据集规模过大也会增加集成过程的时间消耗。基于上述原因，集成算法的性能未能得到很好的发挥，聚类集成算法性能仍有一定的提升空间。为了得到更好的聚类集成算法效果，文献[7，13，19，21-22]等分别从基础聚类成员生成、一致性函数等方面进行了研究并获得了显著的成果。文献[20]在现有的聚类集成两步框架结构上扩展了基础聚类成员预处理步骤，形成了3步结构的算法框架。通过质量过滤这一新加入的步骤，文献[20]在一定程度上消除了基础聚类成员的质量对整体算法性能的影响，提升了算法的性能。

虽然上述研究提升了聚类集成的效果，但是对基础聚类成员的差异性的探索较少。有研究表明，基础聚类成员之间的差异性是影响聚类集成性能的关键因素[23]。因此，为了进一步研究基础聚类成员差异性对聚类集成算法的影响，本文基于Jaccard相似性[24]提出了一种基础聚类成员差异性度量指标，并且结合三支决策思想提出了基础聚类成员过滤算法。

2　相关技术研究

传统聚类集成的算法流程主要由基础聚类的生成和基础聚类的集成这两个主要步骤构成。文献[19，20，25]等均进一步丰富了已有的算法框架。而文献[20]提出的聚类集成框架由基础聚类的生成、基础聚类预处理和基础聚类的集成这3个主要步骤组成。

如图1所示，生成M个基础聚类就是对数据D执行M次聚类，M次聚类结果构成的基础聚类集合（成员集合）可记为$\Pi=\{C^1,C^2,\cdots,C^M\}$，其中$C^i$表示$\Pi$中的第$i$个基础聚类成员。$|\Pi|$表示成员集合$\Pi$的基数，含义为$\Pi$中基础聚类成员数量。$\Pi$中任意基础聚类成员都由若干个类簇构成，以成员$C^i$为例，记为$C^i=\{c_1^i,c_2^i,\cdots,c_{n^i}^i\}$，其中$c_j^i$表示成员$C^i$中的第$j$个类簇。

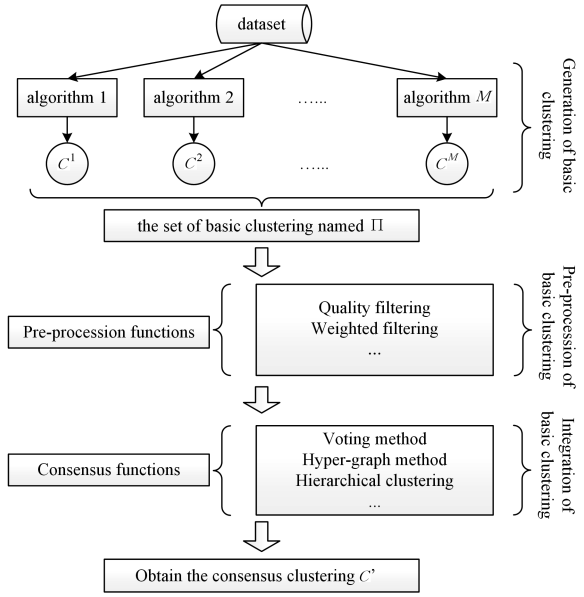


图1　聚类集成三步算法框架

Fig. 1　Framework of three-step algorithm of clustering ensemble

生成基础聚类的关键是最大化每个基础聚类的多样性[26]。近几十年来，在基础聚类生成的步骤中，相关学者已经提出了各种方法来增加基础聚类成员的内部多样性。这些方法基本可以分为以下两类：(1)基础聚类的同类集，即使用不同初始参数的单一聚类算法来生成一组基础聚类成员；(2)基础聚类的异构集，即使用不同的聚类算法生成一组基础聚类成员。文献[27]中详细记录了有关基础聚类生成的步骤。此外，文献[28-29]旨在从数据集中提取信息。文献[30]基于初始聚类的子集提出了一个新的框架，该框架可以获得更多隐信息，从而提高聚类效率。

基础聚类预处理是使用合适的预处理方法对基础聚类集合进行预处理。其预处理过程可以表示为$\Pi^*=g(\Pi)$，其中函数g为预处理方法，Π^*为预处理后的基础聚类成员集合。文献[20]提出了一种基于信息熵方法的基础聚类质量度量指标类簇平均熵，并通过类簇平均熵构建三支决策模型，从而解决了基础聚类成员质量影响聚类集成算法性能的问题。

基础聚类集成过程是一致性函数f对预处理后的结果的集中过程。一致性函数f可由多种方法或技术构成，基于相似矩阵[15]和图论[13]的集成方法的研究热度最高。CA矩阵(Co-association matrix)是相似矩阵法的典型代表，通过计算任意两元素在同一类簇中出现的次数来判定二者之间的离散程度，由CA矩阵演化的研究有文献[7，13，19，20，22]。基于图论的方法一般将基础聚类成员提供的信息集转化为图

$G=\{V,E\}$，并利用图分割技术确定一致性结果。文献[3]提出的基于聚类的相似度划分算法(CSPA)、超图划分算法(HGPA)和基于元聚类的算法(MCA)是3种典型的基于图划分的聚类集成方法。文献[31]提出的一种基于二部图的算法，利用了数据点和类簇之间的相似性。文献[21]对文献[31]提出的算法进行了改进，在一致性函数中引入了与类簇相关联的权重向量。文献[32]提出了一种基于随机游动的类簇相似性快速传播的聚类集成方法，该方法通过将基础类簇作为图节点来构建类簇的相似图，并利用类簇相似性 Jaccard 系数计算初始边权重。

在聚类集成算法研究中，很多理论表明[26,33-34]，当基础聚类器输出的基础聚类成员不一致时(即聚类成员有差异时)，才能得到较佳的聚类集成结果；当基础聚类成员之间的差异在一定范围内增大时，即当基础聚类成员之间关联度越低时，集成学习的优势越突出。但是如何定义和度量差异性却没有一个统一的标准。文献[33]中利用规范化互信息(Normalized Mutual Information，NMI)度量聚类集成的不一致性，用图显示一对聚类间的 NMI 与其平均准确度的关系，得出的结论是：当一对基础聚类成员之间的 NMI 分布较宽(即聚类成员的不一致性较大)，且成员集合的平均准确度相对较高时，聚类集成的性能越好[33]。

Hadjitodorov 等在 ARI(Adjusted Rand Index)的基础上提出4种度量聚类成员差异性的方法，通过实验分析发现，差异性适中的基础聚类成员集合获得的聚类集成结果比差异性最高的基础聚类成员集合的更好，证明了当基础聚类成员集合的整体差异性超过某一阈值强度时，聚类集成的质量反而会下降[34]。

Luo 等在文献[26]中提出了4种差异性度量方法 CEBDM(conditional entropy based diversity meeasure)，DFBDM(double fault based diversity measure)，IRABDM(inter-rater agreement based diversity measure)，CFDBDM(coincident failure diversity based diversity measure)，并将这4种度量方法和基于互信息、信息熵以及 ARI 的3种度量方法进行了对比，得到了不同的差异性度量方法之间的性能关系[26]。

上述研究一方面证明了基础聚类成员差异性会对聚类集成性能带来影响，另一方面也从多种角度对差异性度量进了行研究。但是当前研究的结论仅仅停留在定性层面，并未得出一个明确的量化结论。同时，虽然当前少有针对基础聚类成员的差异化过滤研究，但是已有相关学者开展了大规模数据集下的聚类集成预处理研究。因此，本文将基于 Jaccard 相似性对基础聚类差异性进行评估，并引入三支决策思想，提出一种基于聚类成员差异性的三支决策过滤方法。

3 基于 Jaccard 相似性的基础聚类成员差异性度量

基于传统的二元关系相关性计算方法的 Jaccard 系数，本节构建了一种新的聚类成员差异性度量方法。在任意一组随机变量中，变量之间相互独立，因此根据 Jaccard 系数的使用条件，Jaccard 相似性的定义如下所示：

$$sim(S,T)=\frac{|S\cap T|}{|S\cup T|} \tag{1}$$

其中，S 和 T 为两个独立的对象特征集合。

由于聚类中不同类簇间相互独立，基于上述理论，一对相互独立的类簇间的相似性定义如下。

定义1 对于一对相互独立的类簇 c_s，c_t，如果 c_s 与 c_t 不属于同一聚类，那么类簇 c_s，c_t 的相似性可以通过 Jaccard 相似性求得，计算公式如下所示。

$$j(c_s,c_t)=\frac{|c_s\cap c_t|}{|c_s\cup c_t|} \tag{2}$$

由定义1可得，任意两个类簇间的相似程度可通过上述计算获得。由于基础聚类中的类簇相互独立，对于一组相互独立的随机变量，其联合概率可通过累加获得。这意味着，两个相互独立的基础聚类成员的相似性可由基础聚类的类簇相似性求得。具体内容如定义2所示。

定义2 给定基础聚类成员 $C^i=\{c_1^i,c_2^i,\cdots,c_{k^i}^i\}$，如果有聚类成员 $C^m=\{c_1^m,c_2^m,\cdots,c_{k^m}^m\}$ 与其相互独立，那么聚类成员 C^m 较 C^i 的相似性计算公式如下所示。

$$Sim(C^i,C^m)=\frac{1}{k^i\times k^m}\sum_{s=1}^{k^i}\sum_{t=1}^{k^m}j(c_s^i,c_t^m) \tag{3}$$

其中，$j(c_s^i,c_t^m)$ 为不同成员中两个类簇间的 Jaccard 相似性度量。

由于聚类相似性 $sim(C^i,C^m)$ 的取值范围为(0,1)，聚类成员间差异化程度可由相似性计算，具体计算公式如下所示。

$$CDM(C^i,C^m)=1-sim(C^i,C^m) \tag{4}$$

由定义2可得，如果指定基础聚类成员 C^i 作为基准聚类，那么成员集合 Π 中的基础聚类成员将以 C^i 的类簇标签为标准，计算出与基准聚类 C^i 的差异程度。由于基础聚类之间相互独立，成员集合中所有的基础聚类成员对基准聚类 C^i 的差异性评估，可以理解为基准聚类 C^i 在成员集合 Π 中的差异化程度。因为聚类集成的无监督学习特性，所以基础聚类成员很难直接提取特征标签。因此，为了衡量基础聚类成员的差异性程度，我们通过分配指定基础聚类成员的类簇标签来作为差异性衡量基准。同时，基于上述分析，由于基准聚类 C^i 的差异化程度可以用来评估类簇 C^i 本身，基础聚类 C^i 在成员集合 Π 中的差异化程度，即评估基础聚类成员差异化程度的聚类差异化度量($CDM_\Pi(C^i)$)，具体定义和计算公式如下所示。

定义3 给定基础聚类成员 $C^i=\{c_1^i,c_2^i,\cdots,c_k^i\}$，其在成员集合 Π 中的差异化程度 $CDM_\Pi(C^i)$ 可由 Π 中成员对 C^m 的平均差异度评估获得，计算公式如下所示。

$$CDM_\Pi(C^i)=\sum_{m=1}^{M}CDM(C^i,C^m) \tag{5}$$

其中，$CDM(C^i,C^m)$ 为成员集合 Π 中任意一对聚类成员的差异性度量。

定义2利用 Jaccard 相似性计算了两个独立变量之间的相似度，进而计算了两个聚类成员的差异度 $CDM(C^i,C^j)$。按照定义3，成员集合 Π 中的所有成员均可计算差异化程度。由于 Jaccard 反映了两个独立变量之间的相似性，$sim(C^i,C^j)$ 值越大，基础聚类成员 C^i 和 C^j 之间的相似性越高。而当聚类成员差异性 $CDM(C^i,C^m)\in(0,1)$ 时，$CDM(C^i,C^m)$ 的数值越大，基础聚类成员 C^i 和 C^j 之间的差异性越大。因此，基础聚类成员在成员聚类 Π 中的差异化度量 $CDM_\Pi(C^i)$ 取值越大，基础聚类成员在当前成员集合 Π 中的差异化程度就越大。

4　基于差异性度量的基础聚类成员三支过滤算法

基础聚类成员的差异性会对聚类集成性能带来影响，而基于基础聚类成员差异度的成员过滤是提升聚类集成性能的一种新思路。由于基础聚类成员差异度求解是一个典型的不确定性问题，而三支决策思想[35]是一种典型的求解不确定性问题的思想，将其用于基础聚类成员的差异度求解不失为一种好的思路。三支决策思想是粗糙集理论[36]和粒计算理论[37]发展过程中诞生出的一种重要思想。相较于传统的二支决策（正/负决策）而言，三支决策增加了延迟决策作为不能准确做出正负决策时的决策行为，这3种决策合理地解释了粗糙集理论中的3个决策域（正域、负域和边界域）的划分行为。

三支决策的基本思想是通过评价函数 $\lambda(x)$ 对某个对象集合 D 中的元素 $x\in D$ 来进行不确定程度的度量。当 $\lambda(x)>\alpha$ 时，x 被划分到集合 D 的正决策域 $POS_{(\alpha,\beta)}(x)$；当 $\lambda(x)<\beta$ 时，x 被划分到集合 D 的负决策域 $NEG_{(\alpha,\beta)}(x)$；当 $\beta<\lambda(x)<\alpha$ 时，x 被划分到集合 D 的延迟决策域 $BND_{(\alpha,\beta)}(x)$。其中 (α,β) 为三支决策阈值，通常设定为 $0\leqslant\beta<\alpha\leqslant1$。基于三支决策，全域 D 可以被划分为3个不相交的区域：

1) $POS_{(\alpha,\beta)}(D)=\{x\in D|\lambda(x)\geqslant\alpha\}$

2) $BND_{(\alpha,\beta)}(D)=\{x\in D|\beta<\lambda(x)<\alpha\}$

3) $NEG_{(\alpha,\beta)}(D)=\{x\in D|\lambda(x)\leqslant\beta\}$

因此，基于上述三支决策思想，用于求解基础聚类成员 C^m 差异度的三支决策评估函数 $\lambda(C^m)$ 可以根据下文进行重新定义。

定义4　对于成员集合 Π 中的任一基础聚类成员 C^m，其在成员集合 Π 中的差异性三支决策评价函数计算公式如下所示。

$$\lambda(C^m)=\frac{1}{1+e^{-\varepsilon\times CDM_{\Pi}(C^m)}} \tag{6}$$

其中，$CDM_{\Pi}(C^m)$ 为基础聚类成员 C^m 在成员集合 Π 中的差异性评估；ε 调节了 $\lambda(C^m)$ 在区间(0,1)上分布的均匀程度。

定义4使用Sigmoid函数改进了基础聚类成员差异性度量 $CDM_{\Pi}(C^m)$，解决了差异性程度在区间(0,1)上分布不均匀的问题。当基础聚类成员 C^m 在成员集合 Π 中的差异性较小时，$\lambda(C^m)$ 值较小；当基础聚类成员 C^m 在成员集合 Π 中的差异性增大时，差异性三支评估函数 $\lambda(C^m)$ 也随之增大；二者在定义域单调递增。

根据上述基础聚类差异性度量以及三支决策思想，本文提出了基于差异性度量的基础聚类三支过滤方法，具体内容如下所示。

首先设定基础聚类差异性度量三支决策的阈值为 $0<\alpha^{(0)}<\beta^{(0)}<1$，对成员集合 Π 中任一基础聚类成员 C^m（$C^m\in\Pi$），使用差异性三支决策函数 $\lambda(C^m)$ 进行如下三支划分。

1）若 $\alpha\leqslant\lambda(C^m)$，将聚类成员 C^m 划分到三支决策的正域区间，记为 $C^m\in POS_{(\alpha^{(0)},\beta^{(0)})}(\Pi)$；

2）若 $\lambda(C^m)<\beta$，将聚类成员 C^m 划分到三支决策的负域区间，记为 $C^m\in NEG_{(\alpha^{(0)},\beta^{(0)})}(\Pi)$；

3）若 $\beta\leqslant\lambda(C^m)<\alpha$，将成员 C^m 划分到三支决策的边界域区间，记为 $C^m\in BND_{(\alpha^{(0)},\beta^{(0)})}(\Pi)$。

经过上述划分后得出结论：$POS_{(\alpha^{(0)},\beta^{(0)})}(\Pi)$ 集合中的基础聚类成员在成员集合 Π 中差异化程度较高；$NEG_{(\alpha^{(0)},\beta^{(0)})}(\Pi)$ 集合中的基础聚类成员在成员集合 Π 中差异化程度较低；$BND_{(\alpha^{(0)},\beta^{(0)})}(\Pi)$ 集合中的基础聚类成员在成员集合 Π 中差异化程度无法确定，需要进行进一步判断。

因此本研究将保留 $POS_{(\alpha^{(0)},\beta^{(0)})}(\Pi)$ 域中的所有基础聚类成员，删除 $NEG_{(\alpha^{(0)},\beta^{(0)})}(\Pi)$ 域中的所有基础聚类成员，同时更新三支决策阈值 $(\alpha^{(1)},\beta^{(1)})$，以满足 $0<\beta^{(1)}<\beta^{(0)}<\alpha^{(0)}<\alpha^{(1)}<1$，进而继续对 $BND_{(\alpha^{(0)},\beta^{(0)})}(\Pi)$ 域中的基础聚类成员重新进行三支决策，以生成新的三支决策区间 $POS_{(\alpha^{(1)},\beta^{(1)})}(\Pi)$，$BND_{(\alpha^{(1)},\beta^{(1)})}(\Pi)$，$NEG_{(\alpha^{(1)},\beta^{(1)})}(\Pi)$。

迭代运行上述步骤直至 $|BND_m(\Pi)|=0$（即三支决策的边界域中无成员加入）或 $m=1$（不能继续细化粒度），其中，$m\in\{1,2,3,\cdots,R\}$，R 为迭代次数。最终，由所有三支决策的正域区间的成员组合形成的成员集合 Π 便是成员差异化程度最大的基础聚类成员集合。

基于上述面向基础聚类成员的三支决策过滤思想可获得算法1，具体内容如下所示。

算法1　基于差异性度量的基础聚类三支过滤算法（CDF3WD）

输入：成员集合 $\Pi=\{C^1,C^2,\cdots,C^M\}$，迭代次数 R，初始阈值 $\alpha^{(0)},\beta^{(0)}$

输出：成员集合 Π^*

1. 基于定义1－定义3计算 Π 中所有基础聚类成员的差异化程度
2. 使用定义4构建评估函数 λ
3. for t = 0,1,2,…,R－1 / * 对 Π 中基础聚类成员进行过滤迭代 * /
4. 将阈值 $\alpha^{(t)},\beta^{(t)}$ 代入评估函数 λ
5. for m = 1,2,…,|Π| / * 对所有成员三支决策 * /
6. if $\lambda(C^m)\geqslant\alpha^{(t)}$ then
7. 把成员 C^m 归属到 $POS_{(\alpha^{(t)},\beta^{(t)})}(\Pi)$ 域
8. else if $\beta^{(t)}\leqslant\lambda(C^m)<\alpha^{(t)}$ then
9. 把成员 C^m 归属到 $BND_{(\alpha^{(t)},\beta^{(t)})}(\Pi)$ 域
10. else then
11. 把成员 C^m 归属到 $NEG_{(\alpha^{(0)},\beta^{(0)})}(\Pi)$ 域
12. end if
13. end for
14. if $|BND_{(\alpha^{(t)},\beta^{(t)})}(\Pi)|=0$ then
15. for n = 0,1,2,…,t
16. 把 $POS_{(\alpha^{(n)},\beta^{(n)})}(\Pi)$ 中所有聚类成员加入 Π^*
17. end for
18. break
19. end if
20. 计算新阈值 $\alpha^{(t)},\beta^{(t)}$
21. end for
22. for n = 0,1,2,…,R－1
23. 把 $POS_{(\alpha^{(n)},\beta^{(n)})}(\Pi)$ 中所有聚类成员加入 Π^*
24. end for
25. 输出成员集合 Π^*

1) http://archive.ics.uci.edu/ml

5 实验与结果

本节将使用8组现实世界数据集对所提算法进行实验，并使用两种聚类算法的性能评估指标对聚类集成进行评估。本节所有实验环境为python3.6，Windows10操作系统，Intel i7-7820X 8核心16线程的处理器，32 G安装内存，IDE环境为pyCharm professional 2017.11。

5.1 实验准备

5.1.1 实验数据集

本节对比实验数据集来自UCI机器学习数据仓库[1)]。共计8个公开的真实数据集参与实验，分别是SET1：image segmentation(IS)，SET2：optical digit recognition(ODR)，SET3：texture，SET4：steel plates faults(SPF)，SET5：letter recognition(LR)，SET6：Multiple Features(MF)，SET7：pen digit(PD)，SET8：glass。数据集的具体内容如表1所列。

表1 数据集描述

Table 1 Description of datasets

No.	Dataset	Object	Attribute	Class
SET1	IS	2310	19	7
SET2	ODR	5620	64	10
SET3	texture	5500	40	11
SET4	SPF	1941	27	7
SET5	LR	20000	16	26
SET6	MF	178	13	3
SET7	PD	10992	16	10
SET8	glass	214	9	6

5.1.2 评估指标

为了合理地评估算法性能，本节实验部分使用两种常用的聚类算法性能评估标准即F检验[38]和标准化互信息[39](NMI)。

1)F检验(F-measure)

F-measure是一种准确度评估指标。其基于数据集的真实聚类C^g中各类簇标签来计算准确度P以及召回率R，并以此来进一步算出F-检验值。聚类集成结果C'中类簇C_j'相对真实聚类C^g中类簇C_i^g的准确率和召回率计算公式如下所示：

$$P(c_i^g, c_j') = \frac{|c_i^g \bigcap c_j'|}{|c_j'|} \tag{7}$$

$$P(c_i^g, c_j') = \frac{|c_i^g \bigcap c_j'|}{|c_i^g|} \tag{8}$$

其中，c_i^g为真实聚类C^g中拥有的第i类标签的类簇；c_j'为聚类结果C'中的第i个类簇。

类簇C'中匹配真实类簇c_i^g的元素的百分比计算公式如下所示：

$$FM(c_j') = \frac{2 \times P(c_i, c_j') \times R(c_i, c_j')}{P(c_i, c_j') + R(c_i, c_j')} \tag{9}$$

聚类结果C'相对真实聚类C^g的准确度计算公式如下所示：

$$FM(C') = \frac{\sum_{j=1}^{k} FM(c_j')}{k} \tag{10}$$

2)标准化互信息(NMI)

NMI是一种常用的聚类评估标准。NMI检验可以提供一组随机变量间的信息互通指数(关联程度)。对于一组聚类结果，如果C^g为反应数据集真实聚类的结果，C'为待分析的一组聚类结果中的一个基聚类结果，那么NMI指标得分的计算公式如下所示：

$$NMI(C', C^g) = \frac{\sum_{i=1}^{n'} \sum_{j=1}^{n^g} n_{ij} \log \frac{n_{ij} n}{n_i' n_j^g}}{\sqrt{\sum_{i=1}^{n'} n_i' \log \frac{n_i'}{n} \sum_{j=1}^{n^g} n_i^g \log \frac{n_i^g}{n}}} \tag{11}$$

其中，n'为基聚类C'中的类簇数量；n^g为真实聚类C^g中的类簇数量；n_{ij}为元素出现在聚类C'中的第i个类簇中的个数；n_j^g为元素出现在真实聚类C^g中的第j个类簇中的个数；另外，n_{ij}为基聚类C'中的第i个类簇和真实聚类C^g中的第j个类簇中的公共元素个数。

5.2 定性实验分析

本部分从不同角度设计了对比实验以分析不同算法在数据集上的表现。首先，对比实验定性分析了本文提出的CDF3WD，通过选用8组数据集中的4组进行定性实验分析，以验证CDF3WD对提升聚类集成性能的有效性。其次，对比实验在8个数据集上进行定量分析，通过对比已有的集中差异性度量方法来综合分析算法性能。

(1)定性实验分析的实验设计

实验选用了两种经典聚类集成算法。在基础聚类成员生成方面，统一使用k均值方法为基础聚类器，使用不同参数方式生成M个基础聚类成员。在算法选取方面，EAC和HGPA[40]作为两种经典聚类集成算法的代表，因此将上述两种算法作为对比实验的基准算法组。

使用本文提出的三支过滤算法CDF3WD优化EAC和HGPA，得到CDF3WD-EAC，CDF3WD-HGPA算法。参与定性实验的几种算法的详细描述如表2所列。

表2 参与定性实验的方法

Table 2 Methods for qualitative experiment

Model	Name	Description
1	EAC	使用CA矩阵的层次凝练方法
2	HGPA	使用超图划分的聚类集成方法
3	CDF3WD-EAC	CDF3WD优化的EAC方法
4	CDF3WD-HGPA	CDF3WD优化的HGPA方法

在进行定性实验分析前，为方便开展实验，本文进行了以下准备：1)实验使用k均值算法在SET 1-4数据集上各预生成了150组基础聚类成员；2)设置聚类集成算法相关参数，如设置成员集合基数$M=15$；3)将BCF3WD迭代次数R设置为5。具体实验方案如下所示。

实验将Model1-4分别在各个数据集上运行20轮，每种算法取20轮平均值作为性能得分，最终按数据集进行性能对比。

(2)综合定量分析的实验设计

为验证本文提出的差异性指标以及CDF3WD方法的有效性，本节实验选用已有的差异性指标以及几种高效能聚类集成方法进行了对比实验。参与对比实验的差异性度量方法有CEBDM和DFBDM。参与实验的几种聚类集成方法有EAC[39]，LWEA[17]，LWGP[17]。另外，为了便于进行实验，在SET 1-8数据集上使用模拟退火算法进行阈值参数的最优化实验。

使用本文提出的三支过滤算法 CDF3WD 优化 EAC，LWEA，LEGP，得到 CDF3WD-EAC，CDF3WD-LWEA 以及 CDF3WD-LWGP 算法。同时，使用 CEBDM，DFBDM 优化得到 CEBDM-EAC，DFBDM-EAC，CEBDM-LWEA，DFBDM-LWEA，CEBDM-LWGP，DFBDM-LWGP。参与实验的几种算法的详细描述见表 3。

表 3　参与综合实验的方法

Table 3　Methods for comparison experiment

Model	Name	Description
1	EAC	使用 CA 矩阵的层次凝练方法
2	LWEA	使用局部加权的层次凝练方法
3	LWGP	使用局部加权的超图划分方法
4	CDF3WD-EAC	CDF3WD 优化的 EAC 方法
5	CDF3WD-LWEA	CDF3WD 优化的 LWEA 方法
6	CDF3WD-LWGP	CDF3WD 优化的 LWGP 方法
7	CEBDM-EAC	CEBDM 优化的 EAC 方法
8	CEBDM-LWEA	CEBDM 优化的 LWEA 方法
9	CEBDM-LWGP	CEBDM 优化的 LWGP 方法
10	DFBDM-EAC	DFBDM 优化的 EAC 方法
11	DFBDM-LWEA	DFBDM 优化的 LWEA 方法
12	DFBDM-LWGP	DFBDM 优化的 LWGP 方法

对比实验设计如下：1）对比几种优化后的聚类集成算法之间的性能得分；2）基于聚类集合基数 M 分析几种参与对比试验的聚类集成算法对聚类集成效果的影响。

5.3　综合定量实验分析

5.3.1　定量实验分析

为验证 CDF3WD 对聚类集成算法的提升效果，本节将对比两种经典聚类集成算法 EAC，HGPA 与 CDF3WD 优化后的 CDF3WD-EAC 以及 CDF3WD-HGPA 的算法性能得分。实验分别使用 CDF3WD-EAC，CDF3WD-HGPA 算法进行 100 次实验，并计算每种算法 100 次实验的 NMI 得分的均值以作为该算法的算法性能得分，同时分别执行 100 次基础 EAC，LWEA 算法，并分别计算这两种算法 100 次实验的 NMI 得分的均值以作为基准参考。

实验结果如图 2 所示。与 EAC，HGPA 这类基准聚类集成算法相比，CDF3WD-EAC，CDF3WD-HGPA 算法在所有测试数据集上的平均 NMI 得分均有大幅度的提升，特别是 CDF3WD-EAC 方法在所有测试数据集上都获得了最好的平均 NMI 得分。

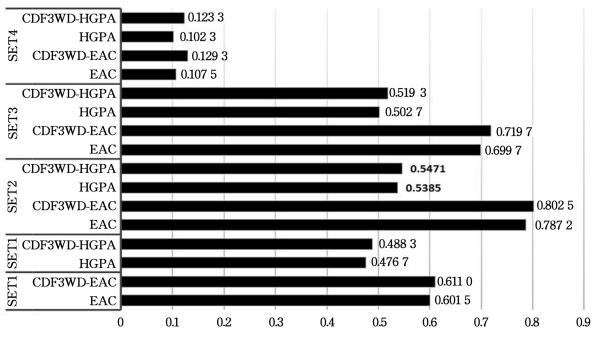


图 2　聚类集成算法在数据集上 20 轮实验的平均得分

Fig. 2　Average NMI score of clustering ensemble methods in 20 runs

在对聚类集成算法不友好的 SET4：SPF 数据集上，CDF3WD-EAC 以及 CDF3WD-HGPA 算法表现仍然优于基准聚类集成算法，其效果明显高于基准 EAC 和基准 HGPA 算法。CDF3WD-EAC 以及 CDF3WD-HGPA 在 SET2 数据集上的 NMI 均值得分提升效果最为显著，较基准 EAC 和基准 HGPA 分别提升了 1.53%和 0.86%。另外，在其他数据集上，CDF3WD-EAC 和 CDF3WD-HGPA 算法的效果较基准 EAC 和基准 HGPA 算法的也有明显的提升。实验证明，CDF3WD 能够提升聚类集成的效果，进一步证明了成员集合差异性对聚类集成算法性能有影响。

5.3.2　综合实验分析

1）与其他聚类集成方法对比

为进一步评估 CDF3WD 重构差异成员集合的性能及其提升聚类集成的效果，本节引入两种差异度评估方法 CEBDM 和 DFBDM 进行横向对比。实验使用一种经典聚类集成方法和两种新型聚类集成方法作为基准算法，分别使用 3 种方法结合基准方法构建优化方法。因此基准方法组和优化方法组如下。

基准组：EAC，LWEA，LWGP；

优化组：CDF3WD-EAC，CDF3WD-LWEA，CDF3WD-LWGP，CEBDM-EAC，CEBDM-LWEA，CEBDM-LWGP，DFBDM-EAC，DFBDM-LWEA，DFBDM-LWGP。

CEBDM 和 DFBDM 优化基准聚类的具体操作为：计算差异性度量，进而为聚类成员加权，最后代入集成算法计算结果。

参与实验的 12 种聚类集成方法按照一类方法、一个分组的方式进行实验，具体操作为：如果 EAC 方法为一个分组，那么 EAC 为基准算法，xxx-EAC 为优化算法组，其中 xxx 为 CDF3WD，CEBDM 和 DFBDM 这 3 种方法中的一种。

实验设定所有方法在相同的基聚类基数 M 下进行，且每种算法获得的聚类结果簇数 k 保持一致。为更好地评估性能，实验基础聚类成员将由基础聚类器统一生成。

实验使用 F 检验和 NMI 两种评估标准对实验算法评估性能得分，首先每种算法分别执行 50 次，然后计算每种算法 50 次的平均 NMI 得分和 F 检验得分。实验参数 M 设定为 20，迭代次数 R 设定为 5。12 种方法的详细实验结果如表 4 所列。可以看出，在对 8 个基准数据集的 50 轮测试中，因为 CDF3WD，CEBDM 和 DFBDM 这 3 种方法调整了成员集合 Π 的差异化程度，所以在一定程度上消除了一致性基础聚类成员带来的噪声影响。因此，在传统算法 EAC 分组中，CDF3WD，CEBDM 和 DFBDM 算法较基准 EAC 算法在性能上均有一定程度的提升。CDF3WD 在本分组中的优化性能表现较为稳定，在各个数据集上较基准 EAC 算法的性能均有所提升。其中在数据集 SET6 上，CDF3WD 的平均优化效果最好，达到了 1.06%的（F 检验）提升；SET4 数据集次之。在 LWEA 和 LWGP 分组中，由于 LWEA 和 LWGP 算法采用了局部权重策略，一定程度上降低了聚类成员的一致性带来的噪声影响。由于 CEBDM 和 DFBDM 在本实验中同样使用了加权策略，在 LWEA 和 LWGP 分组中对基准算法的优化效

果一般。由于CDF3WD使用了三支过滤策略，较加权策略效果更明显，在数据集上表现较基准算法均有提升，特别是在LWGP分组上，CDF3WD优化的算法性能较基准聚类明显提升。在数据集SET3上，CDF3WD-LWGP较基准LWGP提升了1.16%。因此，从表中分析可得，CDF3WD通过重构成员集合，提升了聚类集成性能。同时，在与其他差异性评估方法的横向对比中，CDF3WD性能更稳定，能够更好地提升聚类集成性能。

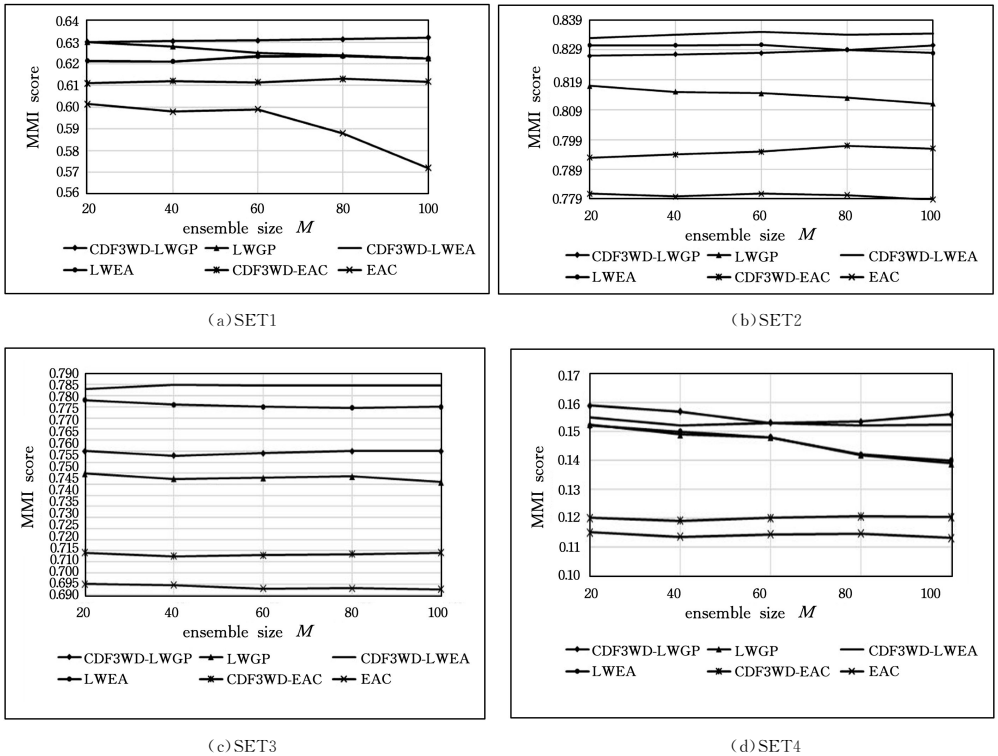
表4 聚类集成方法在数据集上的性能得分

Table 4 Performance score of various methods on each dataset

Dataset	measure	EAC	CDF3WD-EAC	CEBDM-EAC	DFBDM-EAC	LWEA	CDF3WD-LWEA	CEBDM-LWEA	DFBDM-LWEA	LWGP	CDF3WD-LWGP	CEBDM-LWGP	DFBDM-LWGP
SET1	NMI	0.5993	0.6099	**0.6112**	0.6108	0.6222	**0.6227**	0.6218	0.6213	0.6298	**0.6312**	0.6186	0.6192
	F-M	0.5711	**0.5882**	0.5864	0.5823	**0.5931**	0.5918	0.5926	0.5925	0.5995	**0.6013**	0.5815	0.5823
SET2	NMI	0.7815	**0.7927**	0.7824	0.7857	0.8302	**0.8338**	0.8301	0.8296	0.8164	**0.8264**	0.8162	0.8157
	F-M	0.7542	0.7579	0.7585	**0.7591**	0.8083	**0.8101**	0.8079	0.8072	0.7877	**0.7922**	0.7868	0.7870
SET3	NMI	0.6946	**0.7089**	0.7034	0.7085	0.7783	**0.7824**	0.7776	0.7771	0.7437	**0.7553**	0.7431	0.7426
	F-M	0.6714	0.6721	**0.6783**	0.6762	0.7582	**0.7639**	0.7579	0.7581	0.7392	**0.7412**	0.7386	0.7379
SET4	NMI	0.1148	0.1205	0.1209	**0.1213**	**0.1518**	**0.1556**	0.1507	0.1513	0.1532	**0.1591**	0.1516	0.1523
	F-M	0.0886	0.1021	**0.1027**	0.0926	0.1358	**0.1490**	0.1351	0.1345	0.1246	**0.1361**	0.1237	0.1243
SET5	NMI	0.4321	0.4423	0.4403	**0.4497**	0.4574	**0.4592**	0.4571	0.4567	0.4513	**0.4551**	0.4493	0.4486
	F-M	0.4147	0.4203	0.4165	**0.4217**	0.4352	**0.4361**	0.4343	0.4341	0.4377	**0.4428**	0.4263	0.4252
SET6	NMI	0.5979	**0.6035**	0.6012	0.5987	0.6597	**0.6609**	0.6585	0.6601	**0.6826**	0.6860	0.6821	0.6819
	F-M	0.5827	**0.5933**	0.5854	0.5839	**0.6461**	0.6312	0.6447	0.6454	**0.6732**	0.6542	0.6627	0.6625
SET7	NMI	0.6980	**0.7069**	0.7046	0.7037	**0.7695**	0.7693	0.7692	0.7694	0.7751	**0.7762**	0.7615	0.7587
	F-M	0.6717	**0.6801**	0.6738	0.6746	**0.7601**	0.7599	0.7595	0.7598	**0.7568**	0.7536	0.7419	0.7336
SET8	NMI	0.1752	**0.2176**	0.2150	0.2113	0.2217	**0.2344**	0.2127	0.2119	0.2233	**0.2247**	0.2004	0.2016
	F-M	0.1857	**0.2263**	0.2199	0.2214	0.2327	**0.2348**	0.2153	0.2141	**0.2177**	0.2169	0.2092	0.2113

2)成员集合 Π 的基数 M 对聚类集成方法的影响

本节评估基础聚类成员集合 Π 的基数 M 对CDF3WD-EAC，CDF3WD-LWEA，CDF3WD-LWGP以及基准EAC，LWEA，LWGP算法的影响。参与实验的6种聚类集成方法可以分成两组，其中CDF3WD-EAC，CDF3WD-LWEA，CDF3WD-LWGP为CDF3WD优化的对照组算法，EAC，LWEA，LWGP为基准算法组。本部分取SET 1-4共4个数据集进行实验，设定5组不同的基数 $M(M=20,40,60,80,100)$，在不同的基数 M 下6个算法分别进行20次实验，然后计算两种评估指标的平均得分。6种算法在不同数据集上的表现如图3所示。



(a)SET1 (b)SET2

(c)SET3 (d)SET4

图3 成员集合 Π 的基数 M 对聚类集成方法的影响

Fig. 3 Influence of ensemble size M on different clustering ensemble methods

图3中，随着集成规模 M 的增大，实验组中的3种算法在不同数据集上的性能总体呈上升趋势，但是在基数 M 较大时，3种优化的算法较基准算法的趋势明显不同。实验证明，随着基础聚类集合的增大，本文提出的CDF3WD由于删除了更多具有相似性的聚类成员，从而消除了一致性聚类的噪声带来的性能影响。随着基础聚类集合的增大，本文提出的CDF3WD优化的CDF3WD-EAC，CDF3WD-LWEA，CDF3WD-LWGP获得了较基准算法更好的稳定性。

结束语　本文通过对聚类集成的基础聚类成员进行差异性度量，提出了一种三支决策过滤方法。通过Jaccard相似性来构建基础聚类成员间的差异性度量，并使用三支决策评估函数重新定义差异性度量，实现了差异性度量的不确定性求解，进而提出了基于差异性度量的基础聚类三支过滤方法。该方法构建了合适差异程度的基础聚类成员集合，进而提升了聚类集成算法的性能。多组对比实验证明基于差异性度量的基础聚类三支过滤方法能够有效提升聚类集成效果。

三支决策思想拥有良好的不确定性问题处理能力。使用三支决策构建基础聚类成员的预处理，不仅将成员集合的差异性求解问题简化为不确定性求解问题，也进一步丰富了基础聚类成员预处理的研究框架。实验证明，三支决策思想对基础聚类成员差异性度量求解的效果显著，对聚类集成性能提升的效果明显。

参 考 文 献

[1] HUANG D, LAI J H, WANG C D. Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis[J]. Neurocomputing, 2015, 170: 240-250.

[2] ZHOU Z H. Ensemble Methods-Foundations and Algorithms [M]. Taylor&Francis, 2013, 81(3): 470-470.

[3] TOPCHY A, JAIN A K, PUNCH W. Clustering ensembles: models of consensus and weak partitions[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(12): 1866-1881.

[4] STREHL A, GHOSH J. Cluster ensembles: A knowledge reuse framework for combining multiple partitions. [J]. Journal of Machine Learning Research, 2002, 3(12): 583-617.

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

[6] FERN X Z, BRODLEY C E. Random projection for high dimensional data clustering: A cluster ensemble approach[C]// Proceedings of 20th International Conference on Machine Learning. 2003: 186-193.

[7] APPROACH C E, FERN X Z, BRODLEY C E. Random Projection for High Dimensional Data Clustering[C]// Twentieth International Conference on International Conference on Machine Learning. AAAI Press, 2003.

[8] MINAEIBIDGOLI B, TOPCHY A, PUNCH W F. Ensembles of partitions via data resampling[C]// International Conference on Information Technology: Coding&Computing. IEEE Computer Society, 2004.

[9] DUDOIT S, FRIDLYAND J. Bagging to improve the accuracy of a clustering procedure[J]. Bioinformatics, 2003, 19(9): 1090-1099.

[10] YANG Y, JIANG J. Hybrid sampling-based clustering ensemble with global and local constitutions[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(5): 952-965.

[11] ZHOU P, DU L, SHI L, Wang H, et al. Learning a robust consensus matrix for clustering ensemble via kullback-leibler divergence minimization[C]// Proc. the 25th International Joint Conference on Artificial Intelligence. 2015.

[12] YU Z, LUO P, YOU J, et al. Incremental semi-supervised clustering ensemble for high dimensional data clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(3): 701-714.

[13] YU Z, LI L, LIU J, et al. Adaptive noise immune cluster ensemble using affinity propagation[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(12): 3176-3189.

[14] FROUZAN R, SAMAD N, HAMID P, et al. Dibversity Based Cluster Weighting In Cluster Ensemble: An Information Theory Approach. [J]. Artificial Intelligence Review, 2019, 52(2): 1341-1368.

[15] WANG T. CA-Tree: A hierarchical structure for efficient and scalable coassociation-based cluster ensembles[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2011, 41(3): 686-698.

[16] TUMER K, AGOGINO A K. Ensemble clustering with voting active clusters[J]. Pattern Recognition Letters, 2008, 29(14): 1947-1953.

[17] HUANG D, WANG C D, LAI J H. Locally Weighted Ensemble Clustering[J]. IEEE Transactions on Cybernetics, 2016, 48(5): 1460-1473.

[18] HONG Y, YUN C, PAWAN L, et al. A three-way cluster ensemble approach for large-scale data[J]. International Journal of Approximate Reasoning, 2019, 115: 32-49.

[19] KANG Q, LIU S Y, ZHOU M C, et al. A weight-incorporated similarity-based clustering ensemble method based on swarm intelligence[J]. Knowledge Based Systems, 2016, 104(Jul): 156-164.

[20] LIANG W, ZHANG Y J, XU J F, et al. Optimization of Basic Clustering for Ensemble Clustering: An Information-Theoretic Perspective[J]. IEEE Access, 2019, 7: 179048-179062.

[21] HUANG D, WANG C, PENG H, et al. Enhanced ensemble clustering via fast propagation of cluster-wise similarities[J]. IEEE Trans. Syst. Man, Cybern. , Syst. 2019, 11: 1-12.

[22] PARVIN H AND MINAEI-BIDGOLI B. A clustering ensemble framework based on selection of fuzzy weighted clusters in a locally adaptive clustering algorithm[J]. Pattern Anal. Appl. , 2015, 18(1): 87-112.

[23] SONG J H. Research on clustering integration algorithm [D].

Harbin: Harbin Engineering University, 2015.

[24] NIWATTANAKUL S, SINGTHONGCHAI J, NAENUDORNE, et al. Using of Jaccard Coefficient for Keywords Similarity [C] // Iaeng International Conference on Internet Computing & Web Services. International Association of Engineers, 2013.

[25] IAM-ON N, BOONGEON T, GARRETT S, et al. A Link-Based Cluster Ensemble Approach for Categorical Data Clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(3): 413-425.

[26] LUO H L, KONG F S, LI Y X. An Analysis of Diversity Measures in Clustering Ensembles[J]. Chinese Journal of Computers, 2007, 30(8): 1315-1324.

[27] NATTHAKAN I, GARRETT S. LinkCluE: A MATLAB package for link based cluster ensembles[J]. Stat. Softw., 2010, 36(9): 1-36.

[28] PARVIN H, MINAEI-BIDGOLI B. A clustering ensemble framework based on elite selection of weighted clusters[J]. Adv. Data Anal. Classification, 2013, 7(2): 181-208.

[29] YU Z, LUO P, YOU J, et al. Incremental semi-supervised clustering ensemble for high dimensional data clustering[J]. IEEE Trans. Knowl. Data Eng. 2016, 28(3): 701-714.

[30] FERN X, BRODLEY C. Solving cluster ensemble problems by bipartite graph partitioning[C] // Proc. Int. Conf. Mach. Learn., 2004: 36.

[31] DOMENICONI C, AL-RAZGAN M. Weighted cluster ensembles: Methods and analysis[J]. ACM Trans. Knowl. Discovery Data, 2009: 2-17.

[32] HUANG D, LAI J, WANG C. Robust ensemble clustering using probability trajectories[J]. IEEE Trans. Knowl. Data Eng., 2016, 28(5): 1312-1326.

[33] GREENE D, TSYMBAL A, BOLSHAKOVA N, et al. Ensemble Clustering in Medical Diagnostics[C] // 17th IEEE Symposium on Computer-Based Medical Systems, 2004 (CBMS 2004). IEEE, 2004.

[34] HADJITODOROV S T, KUNCHEVA L I, TODOROVA L P. Moderate diversity for better cluster ensembles[J]. Information Fusion, 2006, 7(3): 268-275.

[35] YAO Y. Decision-theoretic rough set models[C] // International Conference on Rough Sets and Knowledge Technology. Springer-Verlag, 2007: 1-12.

[36] QIAN Y H, ZHANG H, SANG Y L, et al. Multi-granulation decision theoretic rough sets[J]. International Journal of Approximate Reasoning, 2014, 55(1): 225-237.

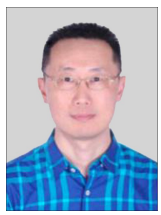
[37] MIAO D, XU F, YAO Y, et al. Set theory description of particle calculation[J]. Journal of Computer, 2012, 35 (2): 351-363.

[38] ABUALIGAH L M, KHADER A T, AL-BETAR M A, et al. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering[J]. Expert Systems with Applications, 2017, 84: 24-36.

[39] STREHL A, GHOSH J. Cluster ensembles: A knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2003, 3(12): 583-617.

[40] LU Z, PENG Y, IP H H S. Combining multiple clusterings using fast simulated annealing. [J]. Pattern Recognition Letters, 2011, 32(15): 1956-1961.

LIANG Wei, born in 1993, Ph.D candidate, is a student member of China Computer Federation. His main research interests include machine learning, granular computing, three-way decision and ensemble clustering.



XU Jian-feng, born in 1973, Ph.D candidate, professor, is a member of China Computer Federation. His main research interests include data mining, rough set, three-way decision and machine learning.