ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 3,May 2026,pp. 780－794
DOI:10. 16337/j. 1004-9037. 2026. 03. 012

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于谷检测和三支集成选择的随机森林聚类方法

谭　诚，李金玉，张文斌，杜明晶

（江苏师范大学人工智能与计算机学院江苏省高校教育智能技术重点实验室，徐州 221116）

摘　要：随机森林聚类作为一种无监督学习方法，虽在高维复杂数据处理中鲁棒性强，但面临负类样本引入导致原始数据区分度弱及噪声决策树干扰聚类效果的问题。针对上述问题，本文提出一种基于谷检测和三支集成选择的随机森林聚类（Random forest clustering based on valley detection and three-way ensemble selection，VDTES-RFC）方法。首先，利用谷检测技术寻找潜在分裂点生成训练数据，并计算Gini指数确定最优分裂点以完成分类森林训练；其次，将决策树视为基聚类器并提取其相似度矩阵，采用三支集成选择策略优选高质量决策树组成新森林；最后，使用共识函数整合相似度矩阵得到最终聚类结果。实验结果表明，该方法有效提升了聚类的准确性与鲁棒性，实现了效率与性能的双优化。

关键词：决策树；随机森林聚类；谷检测；聚类集成选择；三支决策

中图分类号：TP181　　**文献标志码：**A

引用格式：谭诚，李金玉，张文斌，等. 基于谷检测和三支集成选择的随机森林聚类方法[J]. 数据采集与处理，2026，41(3)：780-794. TAN Cheng，LI Jinyu，ZHANG Wenbin，et al. Random forest clustering based on valley detection and three-way ensemble selection[J]. Journal of Data Acquisition and Processing，2026，41(3)：780-794.

引　　言

随机森林概念由Breiman于2001年提出[1]，其由多个决策树组成，每棵决策树通过数据的重采样和随机选择的特征子集训练而成。由于采样和特征选择的随机性，从而使随机森林具有较好的鲁棒性。同时，随机森林能够提供特征重要性估计且易于调参，这对于理解数据结构和进行特征工程非常有帮助。目前，随机森林已成为数据科学领域中最受欢迎的机器学习算法之一[2-3]。鉴于随机森林鲁棒性好、泛化能力强等优点，在聚类分析中，也逐步引起了研究者的关注。聚类分析是一种无监督学习技术[4-5]，不需要预先定义的标签或类别，旨在将数据点分组为簇，簇内的点尽可能相似，簇间的点尽可能不同。常用的聚类算法有K-means、层次聚类、DBSCAN等。近年来，随机森林聚类将随机森林高鲁棒、强泛化的能力融入到聚类分析问题中，此研究方向得到广泛关注[6]。尽管该方法在理论层面展现出诸多优势，但在实际应用中仍面临以下挑战。(1)现有随机森林聚类方法一般使用标准分类森林构造随机森林进行聚类。其间，为实现聚类任务，往往需要人为地引入负类标签。此举让随机森林区分正类和负类时表现较强，而在原始数据上的区分度表现较弱。(2)现有随机森林聚类方法通常采用全集成策略，即，通过创建多棵决策树，将它们简单组合构成用于聚类的随机森林，此策略存在一定的局限性。由于随机森林中的每棵决策树是基于重采样的数据子集和随机选择的特征子集构建，可能会有部分决策树质量较差。这些质量较差的决策树在集成后会对随机森林聚类效果产生不良影响，从而降低聚类的准确性和鲁棒性。

针对上述挑战，本文设计了一种基于谷检测和三支集成选择的随机森林聚类方法（Random forest

基金项目：国家自然科学基金（62006104）；江苏师范大学研究生科研与实践创新基金（2025XKT1457）。
收稿日期：2025-06-15；**修订日期：**2025-07-03

clustering based on valley detection and three-way ensemble selection, VDTES-RFC)。该方法的主要贡献包括两个方面:(1)针对随机森林训练时,负类样例可能导致聚类结果不理想问题,本文提出了基于谷检测的潜在分裂点优化策略,该策略完全依赖原始数据的固有分布,减少了负类样例的影响,从而有效提升了模型对原始数据的区分能力。(2)针对训练后随机森林结果中多样性差,部分决策树质量低,导致聚类结果不理想的问题,本文设计了一种基于三支决策选择策略的聚类集成框架。通过引入三支决策优化了随机森林集成时的质量和多样性,提高了聚类结果的性能和准确性。

1　相关工作

1.1　随机森林相似性度量方法

随机森林聚类中,相似性度量是非常重要的环节。最早,Shi等[7]提出了随机森林中提取对象间相似性度量的思想。该方法认为,若两个对象最终位于同一叶子节点,它们在决策树路径上的所有划分均做出相同的响应,则可以认为它们是相似的。在此基础上,Zhu等[8]对该相似性度量方法进行了扩展,提出了一种更加宽松的相似性度量方法。Zhu等认为,尽管两个对象没有落在相同叶子节点,它们仍然可能具有较高相似性,特别是那些经过更多次树分裂的对象。另外,因为基于共同路径的相似性度量过于严格,导致未充分利用随机森林中包含的有益信息。为此,Bicego等[9]提出进一步考虑树深层路径上分开对象间的相似性度量方法,即,通过考察更深层的树路径来判断对象间的相似性。在相似性计算框架不断演进的同时,另一个影响聚类结果的关键因素是:用于聚类的随机森林本身的集成质量,其直接决定了最终度量结果的优劣。一个理想的聚类森林,其内部的决策树应兼具高准确性和高多样性。为此,已有研究致力于通过筛选和优化基分类器来提升森林的整体性能。例如,有学者提出通过保留分类效果好的决策树并减少它们之间的相关性来构建更优的随机森林[10]。然而,这类方法通常依赖于有监督的指标来评估和筛选决策树,这在无监督的聚类任务中难以直接应用。因此,如何在没有外部标签指导的情况下,有效评估并剔除森林中的噪声决策树,以提升聚类效果,成为了一个亟待解决的问题。

1.2　聚类集成选择策略

聚类集成技术旨在通过筛选和优化,融合多个基聚类器的输出,探索具有准确率高、鲁棒性强和稳定性好的聚类森林,而不需要直接依赖于特征信息。然而,若聚类森林中包含低质量的成员,简单地将多个基聚类器组合可能会降低最终解决方案的整体质量。为此,研究人员提出了聚类集成选择的概念,其核心在于挑选出一组既具有高质量又保持多样性的基聚类器子集。与直接将所有基聚类器成员合并的聚类集成不同,聚类集成选择策略通过筛选出最优的基聚类器子集,构建一个规模更小但性能更优的聚类集成模型。

在聚类集成选择中,研究者需考虑兼顾基聚类器的质量和多样性,从而有效提高聚类效果和性能[11]。聚类集成选择的方法主要有两种:(1)使用调整兰德指数(Adjusted rand index, ARI)作为标准,通过适当的基聚类器挑选策略选取基聚类器子集。例如,使用中位数多样性或重采样技术选择基聚类器子集[12-13]。(2)将归一化互信息熵(Normalized mutual information, NMI)作为判定基准,通过相应的基聚类器挑选策略选取基聚类器子集。例如,Metaxas等[14]通过计算NMI之和来衡量质量和多样性,再对基聚类器的结果进行再次聚类,并选择每个簇内NMI之和最大的基聚类器。同时,部分研究者也提出了一些其他的策略,如聚类组合[15]、扩展证据累积聚类[16]以及结合迁移学习与聚类集成选择,使用多目标自进化过程优化聚类成员选择[17]。此外,Lu等[18]提出了一种考虑协方差多样性度量,根据协方差的差异选择基聚类器子集的策略。Yu等[19]提出了在加权共识函数基础上结合不同特征的选择方法。

除了上述基于特定指标(如ARI、NMI)进行筛选的策略,研究者们还从其他角度探索了更精巧的

选择框架。例如,一些工作将集成选择问题建模为图模型,其中每个基聚类器是图中的一个节点,它们之间的相似度作为边权重,通过寻找图中的最优子结构来确定最终的集成成员[20]。然而,此类方法虽然在理论上较为完备,但其构建的图模型规模与数据点和总簇数直接相关,当集成规模庞大时,图划分过程的计算开销巨大,可扩展性面临挑战。另一些研究则引入了多目标优化的思想,将集成的"质量"与"多样性"视为两个相互冲突的目标,利用多目标进化算法来寻找一组最优的聚类器子集[21]。这类方法虽然能有效权衡质量与多样性,但其进化过程通常需要大量迭代,计算成本高昂,且算法性能对种群大小、交叉变异率等众多超参数的设定十分敏感,因此设计一种更高效、鲁棒且决策过程更简洁的集成选择策略,仍然是该领域一个值得探索的方向。

2 VDTES-RFC方法

2.1 VDTES-RFC基本框架

针对随机森林聚类的诸多挑战,本文提出了VDTES-RFC方法,本节对该方法进行详细介绍,全面阐述其设计思想和关键步骤。从整体框架和基本流程两个维度对所提方法进行阐述,明确方法的基本结构和各步骤之间的联系。

本文所提方法VDTES-RFC的框架如图1所示,整体框架分为两个阶段:随机森林训练阶段和聚类阶段。(1)随机森林训练阶段。在此阶段中,首先利用谷检测技术从原始数据中选取潜在的分裂点,这些分裂点能够有效提升对原始数据的区分能力,其次利用边缘随机采样技术生成负类数据,并将其与原始数据组合成训练数据,然后通过数据重采样的方法训练每棵决策树,继而完成标准随机森林模型训练。(2)聚类阶段。在此阶段中,首先利用标准随机森林的决策树提取数据集样本间的相似性信息,并使用谱聚类算法得到初步聚类结果,然后采用三支集成选择策略进行随机森林集成选择,从标准随机森林中选择出部分决策树组成新随机森林,最后利用共识函数整合所选决策树的相似度矩阵,应用聚类算法得到最终聚类结果。

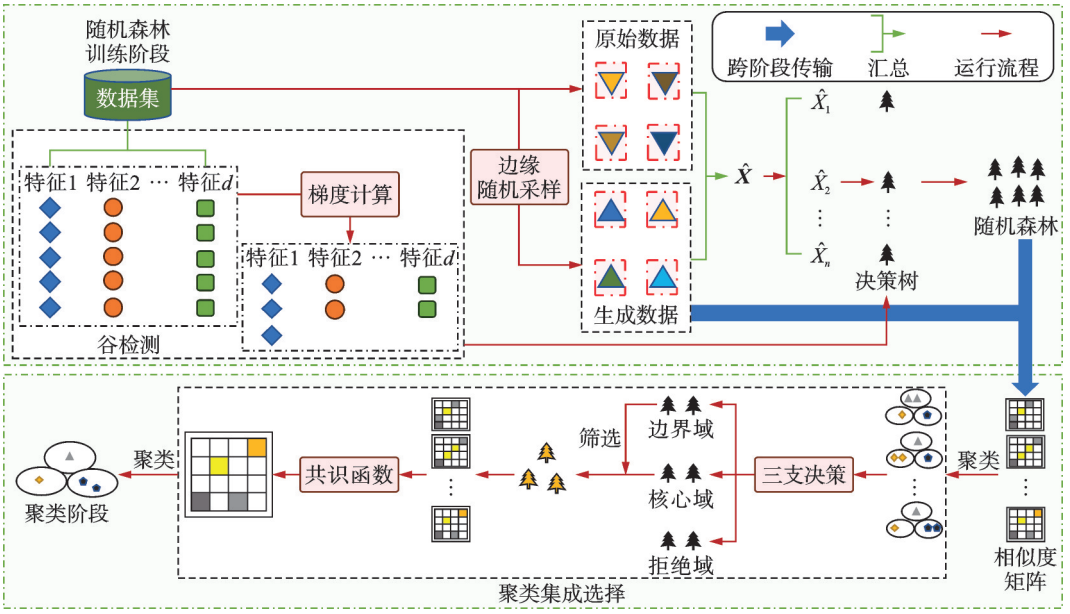

图1 VDTES-RFC的框架

Fig.1 Framework of VDTES-RFC

2.2　随机森林训练阶段

2.2.1　基于谷检测的分裂点优化

在标准随机森林生成过程中,VDTES-RFC采用谷检测技术探索原始数据集中潜在的特征分裂点。谷检测技术常用于描述系统或数据的波动、变化和分布等特征,尤其是存在极值(最大值和最小值)波动的情况。通过对谷检测结果的分析可以更深入理解系统数据和更好地预测系统行为,进而识别系统潜在的规律或异常,并进行有效的优化或干预。在机器学习和数据挖掘领域,谷检测技术用来描述数据集的分布特性,特别是在聚类分析和特征选择中广泛应用。谷检测的常用方法包括差分法、滑动窗口法等。通常,滑动窗口法适合复杂噪声、宽谷或需提取多特征的分析场景;而差分法特别适用于实时性要求高、陡峭谷的检测。为了提高性能,本文采用了差分法进行谷检测。

差分法谷检测在每个候选特征上,通过检测数据分布中的谷点(即局部最小值)来确定可能的分割位置,从而有效地捕捉特征值变化的关键转折点,为后续节点分裂提供可靠的分裂依据。如在原始数据中,当某个点的前后差分符号相反,且该点的值低于其邻域内的其他点时,可以认为该点是一个谷值。即对于特征值序列$X=\{x_1,x_2,\cdots,x_n\}$,差分法计算相邻值的梯度变化$d_i=x_{i+1}-x_i$,当$d_{i-1}<0$且$d_i>0$时,判定x_i为局部极小值。差分法谷检测通过计算特征值的梯度变化,精确地定位数据分布中的局部最小值。

2.2.2　训练标准分类森林

聚类过程通常是一种无监督学习过程。随机森林应用于聚类分析时,通常无法进行真正有意义标签下的学习。针对此问题,本文采用了一种基于伪标签的标准分类随机森林训练策略。首先,从数据集的边缘分布中随机采样创建生成样本,然后,把生成样本与原始数据结合,形成训练集$\hat{\boldsymbol{X}}$,原始数据标记为正例,生成样本标记为负例。在具有伪标签的数据集$\hat{\boldsymbol{X}}$上训练随机森林模型。

随机森林训练时,首先在训练集$\hat{\boldsymbol{X}}$上采用重采样技术生成多个数据子集,然后在这些数据子集上生成相应决策树。如此,既增加了决策树的多样性,又有效避免了过拟合。

为了增强决策树对原始数据的区分能力,VDTES-RFC方法采用潜在分裂点优选替代传统决策树分裂点选择策略。分裂时,首先针对每个潜在分裂点,计算其划分后左右两个子集的基尼指数,然后选择基尼指数最小的分裂点作为最优分裂点,并在该节点进行数据划分,将样本分配至左右子节点。决策树生成时经过潜在分裂点搜索和优选,不断生成新节点,直至决策树完全生长。

最后,集成多棵决策树,构建标准随机森林。

2.3　聚类阶段

2.3.1　基于路径的相似度信息提取

训练完标准分类随机森林后,如何计算决策树中数据点之间的相似度成为了一个至关重要的问题。传统的随机森林聚类通过计算决策树中不同数据点的共同路径长度来衡量它们的相似性:共同路径越长,相似性越高;反之,相似性越低。然而,共同路径相似计算方法存在一定的局限。其仅考虑了共同路径上的特征差异,而忽略了其他特征差异对相似性的潜在贡献,可能导致相似性度量不准确。

如图2所示,图中展示了3个样本:$\boldsymbol{x}_1(0.4,0.4)$、$\boldsymbol{x}_2(0.6,0.4)$和$\boldsymbol{x}_3(0.7,0.8)$。数据集先以阈值$T_1=0.5$选择第一个特征进行分割,然后以阈值$T_2=0.5$选择第

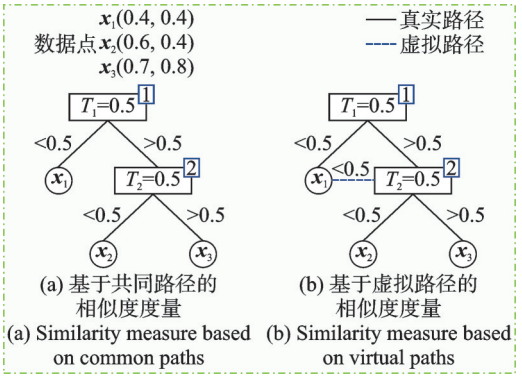

图2　部分路径图

Fig.2　Partial path graph

二个特征再进行分割。如图2(a)所示，$\boldsymbol{x}_1$和$\boldsymbol{x}_2$以及$\boldsymbol{x}_1$和$\boldsymbol{x}_3$公共路径长度为(0)，依照共同路径相似性计算方法，$\boldsymbol{x}_1$和$\boldsymbol{x}_2$以及$\boldsymbol{x}_1$和$\boldsymbol{x}_3$的相似度相同。然而，当观察3个样本的第二个特征时，发现$\boldsymbol{x}_1$和$\boldsymbol{x}_2$相应特征的值均小于0.5，而$\boldsymbol{x}_3$的相应特征的值大于0.5。显然，3个样本相似度一样的结果是不合理的。

本文采用了Bicego[9]提出的方法作为相似度的度量。该方法通过将数据点与另一个数据点路径上每个节点的阈值进行比较来生成虚拟路径，从而实现更准确的数据点之间相似度计算。如图2(b)所示，在计算$\boldsymbol{x}_1$和$\boldsymbol{x}_2$的相似度时，将$\boldsymbol{x}_1$与第二个特征的阈值$T_2=0.5$进行比较，产生一个左分支(因为第二个特征值小于0.5)，形成一条虚拟路径。在经过阈值$T_2=0.5$划分后，$\boldsymbol{x}_1$的虚拟路径与$\boldsymbol{x}_2$的真实路径一致(都是左分支)，说明在此相似度度量下，$\boldsymbol{x}_1$和$\boldsymbol{x}_2$之间的相似度要大于$\boldsymbol{x}_1$和$\boldsymbol{x}_3$之间的相似度。如此考虑，可以更加精准地提取样本间的相似度信息。相似度度量公式定义为

$$s(\boldsymbol{y},\boldsymbol{z})=\frac{|\boldsymbol{Y}\cap\boldsymbol{Z}|}{|\boldsymbol{Y}\cap\boldsymbol{Z}|+|\boldsymbol{Y}-\boldsymbol{Z}|+|\boldsymbol{Z}-\boldsymbol{Y}|} \tag{1}$$

式中：$\boldsymbol{y}$、$\boldsymbol{z}$分别表示两个数据样本的特征向量；$\boldsymbol{Y}$、$\boldsymbol{Z}$分别表示样本$\boldsymbol{y}$和$\boldsymbol{z}$在决策树中从根节点到各自叶子节点所经过的节点集合；$|\boldsymbol{Y}\cap\boldsymbol{Z}|$表示两样本在各自到彼此路径上共享相同分区判断的次数；$|\boldsymbol{Y}-\boldsymbol{Z}|$表示样本$\boldsymbol{y}$在到$\boldsymbol{z}$的路径上不与$\boldsymbol{z}$共享相同分区判断的次数；$|\boldsymbol{Z}-\boldsymbol{Y}|$表示样本$\boldsymbol{z}$在到$\boldsymbol{y}$的路径上不与$\boldsymbol{y}$共享相同分区判断的次数。

2.3.2 基于三支决策的动态集成选择

(1)集成选择策略

在随机森林聚类集成选择过程中，基聚类器的评估与选择直接影响到最终聚类结果的优劣。聚类本身是无监督的，没有任何外部信息，例如真实标签等信息。为评估聚类算法性能，研究者通常利用伪标签[15]等方法处理，辅助评估聚类算法聚类结果的质量。即，将聚类结果与伪标签进行比较，量化聚类结果的有效性、准确性和一致性，从而评估聚类解决方案的好坏。然而，在聚类集成选择中，这种评估方法不太合理。因为集成选择旨在利用基聚类器聚类结果本身的信息来确定最佳的聚类集成，而不是依赖于所谓的、不准确的外部标准[22]。因此，选择基聚类器本身特征和聚类结构等内部信息来评估聚类的质量，进而完成聚类集成选择，才是合理的选择。聚类集成选择中，NMI不依赖外部信息，通过信息熵直接衡量聚类间的信息共享程度，有效衡量了两个聚类结果的相似度。故本文选用NMI评估基聚类器质量，并采用归一化互信息之和(Sum of normalized mutual information，SNMI)准则进行聚类集成选择。

假设给定一个基聚类器集合$\boldsymbol{\Pi}=\{\pi_1,\pi_2,\cdots,\pi_V\}$，那么聚类集成选择应最大化如下准则[23]，有

$$\mathrm{SNMI}(\pi,\boldsymbol{\Pi})=\sum_{i=1}^{V}\mathrm{NMI}(\pi,\pi_i) \tag{2}$$

式中：$\mathrm{NMI}(\pi,\pi_i)$为基聚类器π和π_i之间的归一化互信息。

(2)三支集成选择策略

三支决策是一种处理不确定性和模糊性的决策方法。其核心思想是三分而治，即将整个问题分为3个互斥的区域：正域、负域和边界域，对不同的区域采用不同的处理方法，为复杂问题求解提供了一种简单且有效的方法。当面对信息不完整或评估结果模糊时，避免简单二分类可能导致的误判，转而通过引入“不确定”区域来延迟决策或进行更深入的分析，从而有效提高问题解空间的准确性[24-25]。

当前随机森林聚类方法，例如文献[7-9]在集成选择过程存在明显局限性。这些方法对所有决策树生成的相似度矩阵直接进行平均化处理，忽视了不同决策树对最终聚类结果的贡献差异。简单求和取平均策略可能导致关键信息的稀释，尤其是当部分决策树质量较低或冗余性较高时，会严重影响聚类结果的准确性和有效性。针对此问题，本文提出了基于三支决策思想的集成选择策略，以优化集成选择结果，提高聚类的效果。

三支集成选择策略基于三支决策思想构建动态筛选机制。首先将每棵决策树视为一个基聚类器,通过对其提取的相似度矩阵进行初步聚类。本文使用谱聚类方法进行聚类,同时计算每棵决策树与其他所有决策树的SNMI,以量化其独立贡献价值。其次,利用三支决策设定上下阈值,将所有基聚类器的集合划分为核心域、边界域和拒绝域3个子集。

在三支集成选择策略中,下阈值α和上阈值β的设定,是实现对基聚类器(决策树)进行有效划分的核心。其设定的基本依据源于三支决策理论中对“接受”、“拒绝”和“延迟决策”3个域的语义定义。

上阈值β(核心域边界)。该阈值用于识别那些确定性高、贡献明确为正的优质决策树。一个基聚类器与其他所有聚类器的SNMI值如果高于β,则意味着它与整个集成具有高度的一致性和互补性,属于“确定接受”的对象。因此,β的值应设定得相对较高,以确保核心域成员的质量。在实践中,NMI$>$0.5通常表示两个聚类结果具有显著的相似性,故本文将β的探索范围设定在0.6以上。

下阈值α(拒绝域边界)。该阈值用于过滤掉那些质量低下或与主流聚类结构显著不符的决策树。SNMI值低于α的基聚类器被认为是“确定拒绝”的噪声或冗余成员。α的设定不宜过高,以避免错误地丢弃有潜在价值的决策树。

边界域(α,β)。介于α和β之间的区域构成了边界域,或称为“不确定”区域。落入此区域的决策树其贡献既非明确为优,也非明确为劣,需要通过后续与核心域成员的互补性计算来“延迟决策”。这个边界域的存在,正是三支决策相比传统二元选择的优势所在,它为模型提供了一个缓冲和再评估的机制,从而增强了选择的鲁棒性和精确性。

综上,阈值(α,β)的选择本质上是在“确保核心集质量”与“避免错失边界集良机”之间进行权衡。对于边界域中的基聚类器,通过进一步计算其与核心域内各基聚类器的SNMI互信息,筛选出与核心域互补性较强的个体加入核心域,从而形成高质量聚类决策树集合。通过三支集成选择的动态筛选,既保留了高贡献度的核心域基聚类器,又通过动态调整边界扩展了优质候选池,从而有效保证了聚类决策树集合的质量和多样性,增强了聚类结果的泛化能力和鲁棒性。

同时,为进一步提升集成多样性,采用基于平衡K-means的分层K-means策略(Balanced K-means based hierarchical K-means, BKHK)对核心域中的决策树进行二次筛选。BKHK每次将数据迭代划分为两个簇,每个簇具有相同数量的样本,以保证平衡的分层策略。通过约束聚类子集间的分布均衡性,在保证质量的前提下最大化决策树间的差异性,最终集成高质量和多样性的聚类随机森林。三支集成选择策略融合了三支决策与平衡分层聚类,突破了传统方法对决策树等权重处理模式的局限性,实现了对基聚类器的精细化评估与自适应选择,为复杂数据场景下聚类集成提供了更有效的解决方案。

2.3.3　基于共识矩阵的最终聚类

通过三支集成选择策略筛选高质量且多样化的决策树集合,其中每棵决策树独立生成1个相似度矩阵。由于这些决策树已通过前期SNMI阈值划分与平衡分层约束,故其对应相似度矩阵既具备较高的质量,又保持了较好的多样性。然后为融合随机森林中所有异构视角决策树信息,本文通过共识函数将多个相似度矩阵整合为统一的全局矩阵,以生成最终的聚类结构。共识函数是一种将多个基聚类结果进行整合的函数,旨在综合多个聚类输出获得准确而稳定的最终聚类结果。其主要通过构建共识矩阵实现,常用于多视图学习或多模态数据融合。共识矩阵是从多个不同的相似度矩阵中提取出一个统一的相似度矩阵,从而找到一个综合的相似度评估矩阵。

本文方法前期通过谷检测和三支集成策略有效保证了随机森林中决策树的质量和多样性。同时,集成选择时,采用了SNMI准则,筛选出的决策树相互之间的重要性区别不是太大。为简化实现并提高效率,不妨假设所有决策树的相似度矩阵对最终聚类结果具有同等的重要性。于是,本文采用了简单的平均共识函数实现聚类信息融合,即对所有决策树的相似度矩阵进行逐元素求平均,生成全局相似度矩阵。具体构建共识矩阵的方法为

$$C_{ij}=\frac{1}{N}\sum_{k=1}^{N}S_{ij}^{(k)} \tag{3}$$

式中：C_{ij}表示最终共识相似度矩阵中第i行第j列的元素；$S_{ij}(k)$表示第k棵被选中决策树对应的相似度矩阵中第i行第j列的元素；N表示经三支集成选择策略筛选后保留的决策树总数。

3 实验及结果分析

3.1 实验准备

为评估VDTES-RFC的性能，本文在8个真实数据集上进行了实验，如表1所示。实验设备配置为AMD Ryzen 7 4800H with Radeon Graphic@2.90 GHz CPU，NVIDIA GeForce GTX 1650 Ti，内存16 GB，Windows 11操作系统，编程环境为Matlab 2024b。

表1 实验数据集

Table 1 Datasets used in experiments

名称	样本数量	特征数量	类别数量
Wine	175	13	3
Seeds	210	7	3
Movementlibras	360	90	15
OlivettFaces	400	28	40
Control	600	60	6
Vowel	990	13	11
Dig	1 797	64	10
MSRA	1 799	256	12

(1)数据集描述

实验中用于验证的数据集包括Wine、Seeds、Movementlibras、OlivettFaces、Control、Vowel、Dig和MSRA。表1描述了验证数据集的详细信息。其中，Wine、Seeds、Movementlibras、Control、Vowel和Dig数据集来自UCI数据库网站。Wine数据集包含意大利不同品种葡萄酒的化学分析结果；Seeds数据集包括不同种类小麦种子的测量数据；Movementlibras数据集记录了几种手部运动信号，适用于手势识别；Control数据集是一个常用于时间序列分析和模式识别的基准数据集；Vowel数据集包含元音的声学特征，用于语音识别；Dig数据集包含手写数字图像，用于数字识别。此外，还使用了纽约大学人脸识别数据集OlivettiFaces和由微软亚洲研究院收集的大量不同种族、性别、年龄和表情的人脸数据集MSRA[26]。

(2)基准方法

为全面评估VDTES-RFC的性能，本文选取5种具有代表性的随机森林聚类方法进行对比实验，这些方法在相似性度量的构建上各具特点。

m_Shi[7]是随机森林聚类的开创性工作之一，它以样本对在森林中落入相同叶节点的频率作为相似性的基础；m_Zhu2[8]和m_Zhu3[8]这两种方法是m_Shi的改进，将相似性的考量从叶节点延伸至决策路径，分别以共享路径的长度和加权路径长度作为度量标准；m_Ting[27]方法采用了一种异于路径长度的思路，通过计算能到达样本对最低共同祖先节点的其他样本比例来衡量相似度；RatioRF[9]作为一种更精细的度量方法，它综合评估一个样本在另一个样本决策路径上所有节点的划分响应，从而捕捉更深层次的关联。

(3)参数配置

本文VDTES-RFC方法中，主要有4个关键参数：决策树的数量t、被选择的决策树数量t'、三支决策下阈值α和三支决策上阈值β。实验中，参数t设置为100，与Bicego等[9]的建议一致；被选择决策树的数量t'设置为2^5；三支决策的阈值区间(α,β)设置为(0.3,0.8)。所有数据集上的聚类结果通过循环10次取平均的方式获得。

(4)评价指标

实验中，使用NMI、ARI和ACC三个指标来评估VDTES-RFC和比较方法的性能。NMI和ARI是评估聚类性能的常用方法。通常，数值越高表示聚类性能越好。NMI的范围从0到1，其中1表示完全一致；ARI的范围从−1到1，其中1表示完全一致，负值表示聚类效果较差。ACC测量正确分配样本的比例，数值越高表示准确度越高。

(5)实验细节

在本文方法中,随机森林被训练了10次,每次训练代表一个独立的起点,用于计算数据点之间的相似性。每次训练后,通过计算数据点的相似性来构建数据点之间的关系网络。在获得相似性矩阵之后,使用4种不同的算法进行聚类:Ng-Jordan-Weiss归一化版本谱聚类[28]、经典的基于距离的K-means聚类算法以及两种层次聚类算法Complete-Link和Ward-Link。通过这些算法在不同层次结构捕获数据点之间的关系,从而让聚类结果更准确。同时,本文算法的实现基于Matlab中的统计和机器学习工具箱实现。

3.2　对比实验分析

本节在表1中的数据集上对基准方法和VDTES-RFC的聚类性能进行测试,并对所有方法的聚类结果进行比较分析。在8个真实数据集上,基准方法和VDTES-RFC的ACC、ARI和NMI值如表2所示。各个数据集在对应指标下的最高得分用粗体表示。

表2　真实数据集上的方法性能比较

Table 2　Comparison of method performance on real datasets

Method	Wine			Seeds		
	ACC	NMI	ARI	ACC	NMI	ARI
VDTES-RFC	**0.930**	**0.776**	**0.793**	**0.848**	**0.645**	**0.638**
RatioRF	0.888	0.747	0.755	0.817	0.610	0.569
m_Shi	0.797	0.618	0.565	0.729	0.491	0.430
m_Zhu2	0.863	0.685	0.665	0.798	0.600	0.548
m_Zhu3	0.881	0.732	0.741	0.802	0.599	0.566
m_Ting	0.874	0.721	0.709	0.804	0.604	0.560
Method	Movementlibras			OlivettFaces		
	ACC	NMI	ARI	ACC	NMI	ARI
VDTES-RFC	**0.464**	**0.569**	**0.289**	**0.665**	**0.809**	**0.505**
RatioRF	0.442	0.538	0.267	0.663	0.788	0.494
m_Shi	0.446	0.528	0.251	0.592	0.749	0.351
m_Zhu2	0.452	0.537	0.263	0.605	0.804	0.497
m_Zhu3	0.455	0.538	0.255	0.633	0.806	0.492
m_Ting	0.434	0.531	0.254	0.609	0.765	0.424
Method	Control			Vowel		
	ACC	NMI	ARI	ACC	NMI	ARI
VDTES-RFC	**0.755**	**0.773**	**0.632**	**0.359**	**0.364**	**0.171**
RatioRF	0.656	0.749	0.576	0.289	0.301	0.119
m_Shi	0.605	0.732	0.529	0.238	0.237	0.060
m_Zhu2	0.664	0.746	0.559	0.264	0.276	0.099
m_Zhu3	0.689	0.762	0.588	0.288	0.296	0.116
m_Ting	0.702	0.742	0.586	0.278	0.276	0.108
Method	Dig			MSRA		
	ACC	NMI	ARI	ACC	NMI	ARI
VDTES-RFC	**0.675**	**0.655**	**0.546**	**0.543**	**0.617**	**0.376**
RatioRF	0.669	0.639	0.525	0.533	0.600	0.346
m_Shi	0.509	0.486	0.287	0.503	0.567	0.315
m_Zhu2	0.648	0.617	0.488	0.527	0.597	0.350
m_Zhu3	0.659	0.643	0.493	0.507	0.581	0.325
m_Ting	0.664	0.629	0.515	0.526	0.598	0.354

从实验结果看，m_Shi的聚类表现整体处于劣势。以Wine数据集为例，其ACC仅0.797，不足VDTES-RFC(0.930)的86%；在Dig数据集上ARI值0.287，较VDTES-RFC(0.546)差距显著。这是因为该方法仅以“数据对象在叶节点的共现频率”作为相似性判据，未挖掘决策树中路径分支、节点层级等结构特征，导致相似性度量仅反映局部关联，无法刻画数据对象的全局分布关系，最终限制了聚类性能。

m_Zhu2与m_Zhu3作为改进方案，通过“共同路径特征”拓展了相似性度量维度。在Seeds数据集上，m_Zhu2的ACC达0.798，较m_Shi(0.729)提升9.5%；Wine数据集的NMI从0.618(m_Shi)提升至0.685(m_Zhu2)，验证了路径信息对相似性刻画的增益。这两种方法突破了“叶节点单一维度”的局限，将决策树的分支路径、节点遍历顺序等结构特征纳入相似性计算，更全面地捕捉了数据对象的分布模式。

m_Ting的设计思路与前两者异曲同工，其以“最低共同祖先的路径长度比例”定义相似性，同样强化了决策树结构信息的利用。在Control数据集上，m_Ting的ARI为0.586，与m_Zhu2(0.559)、m_Zhu3(0.588)的表现接近；但在Movementlibras数据集上，其ARI仅0.254，略低于m_Zhu3的0.255。这表明m_Ting的路径长度比例策略在部分场景下能与“共同路径”方法达到相近效果，但对高维时序数据(如Movementlibras的90维特征)的结构信息利用仍存在优化空间。RatioRF通过深层路径匹配机制优化相似性度量，相较于m_Zhu2和m_Zhu3更注重“非叶节点层级的分离特征”提取。以Dig数据集为例，其ACC达0.669，较m_Zhu3(0.659)提升1.5%；Wine数据集的ACC为0.888，超过m_Zhu2(0.863)和m_Zhu3(0.881)，验证了深层路径信息对相似性关系的细化作用。但在Control这类含噪声的工业数据集上，RatioRF的ARI为0.576，仍低于VDTES-RFC(0.632)，反映出其集成策略对复杂噪声场景的适配性仍有优化空间。

VDTES-RFC相较于RatioRF，通过采用谷检测技术且融合三支集成选择策略，显著提升了随机森林的聚类性能，同时优化了随机森林的聚类效率，实现了性能效率双优化。对比实验结果表明，ACC、NMI和ARI的值分别提高了4.73%、3.88%和5.03%(Wine)，3.79%、5.74%和0.12%(Seeds)，4.98%、5.76%和8.24%(Movementlibras)，0.30%、2.66%和2.23%(OlivettFaces)，7.55%、1.44%和7.48%(Control)，24.22%、20.93%和43.69%(Vowel)，0.9%、2.5%和4%(Dig)，1.88%、2.83%和8.67%(MSRA)。

综上分析可知，本文方法优于其他基准方法。

3.3 参数敏感度分析

在本节中，将深入探讨所提出的VDTES-RFC在执行聚类任务时，其性能是否受到关键参数设置的影响。具体将分析4个核心参数：决策树的数量t、被选择的决策树数量t'、三支决策下阈值α和三支决策上阈值β。理论上，4个参数的选择会在一定程度上影响算法的性能。因此，本文对这些参数的敏感性进行了实验分析。

参数t决定随机森林的大小，直接关系到模型的复杂度，决策树过多可能导致过拟合，而数量不足则可能无法捕捉到数据中的复杂模式。参数敏感性分析中，其取值范围为50、75、100、125和150。参数t'决定聚类集成选择时决策树数量，分析时取值范围为2^3、2^4、2^5和2^6。设定下阈值α的取值范围为0.2、0.3、0.4和0.5和上阈值β的取值范围为0.6、0.7、0.8和0.9，形成16种不同的阈值组合(α,β)。为了全面分析参数对聚类性能的影响，使用ACC和NMI作为不同参数配置下聚类结果的评价指标。

首先评估参数t的影响。图3显示了在Vowel、MSRA、Movementlibras和Control四个数据集上不同t值时的ACC和NMI值。结果表明，VDTES-RFC的聚类性能对t的变化不敏感，具有较好稳定性，在每个实验数据集上均波动较小。

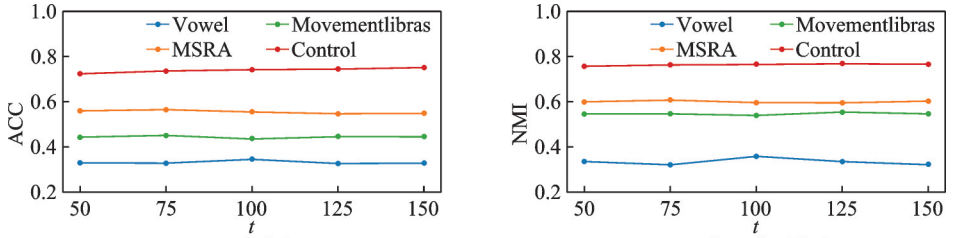

图3　不同参数t下的ACC和NMI的值

Fig.3　ACC and NMI with different value of parameter t

然后评估参数t'的影响。图4显示了不同t'值时的ACC和NMI值。虽然随着t'的增加,性能略有上升,但变化总体平稳,表明VDTES-RFC对t'的变化相对不敏感。

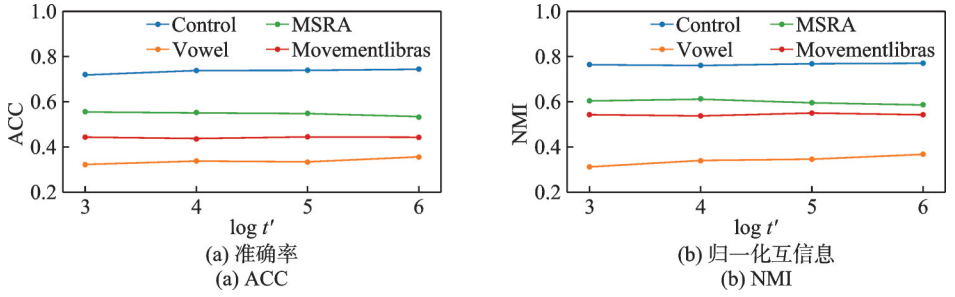

图4　不同参数t'下的ACC和NMI的值

Fig.4　ACC and NMI with different value of parameter t'

其次,表3展示了不同阈值组合对模型性能的影响,本文的参数选择是基于对这些实验结果的系统性分析。在考察下阈值α的影响时,可以发现$\alpha=0.3$是一个关键的平衡点。当α从0.2增至0.3时,模

表3　不同阈值组合的实验结果

Table 3　Results of experiments with different threshold combinations

阈值组合(α,β)		Control		Movementlibras		MSRA		Vowel	
α	β	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
0.2	0.6	0.730	0.774	0.442	0.551	0.553	0.593	0.325	0.324
	0.7	0.736	0.776	0.447	0.562	0.542	0.593	0.332	0.332
	0.8	0.731	0.753	0.472	0.576	0.548	0.595	0.323	0.324
	0.9	0.741	0.770	0.457	0.559	0.534	0.585	0.309	0.304
0.3	0.6	0.733	0.754	0.444	0.550	0.561	0.606	0.339	0.349
	0.7	0.724	0.755	0.450	0.547	0.569	0.613	0.335	0.339
	0.8	0.770	0.787	0.450	0.552	0.544	0.594	0.324	0.321
	0.9	0.734	0.771	0.442	0.547	0.547	0.586	0.325	0.323
0.4	0.6	0.728	0.761	0.451	0.552	0.559	0.605	0.327	0.326
	0.7	0.737	0.775	0.447	0.551	0.546	0.592	0.312	0.300
	0.8	0.736	0.773	0.460	0.562	0.536	0.588	0.335	0.336
	0.9	0.752	0.762	0.466	0.566	0.539	0.587	0.309	0.296
0.5	0.6	0.698	0.745	0.455	0.565	0.538	0.589	0.332	0.334
	0.7	0.715	0.749	0.436	0.537	0.541	0.586	0.334	0.328
	0.8	0.717	0.754	0.470	0.564	0.542	0.594	0.314	0.307
	0.9	0.731	0.774	0.445	0.547	0.536	0.587	0.329	0.333

型性能维持在较高水平;然而,当α继续增至0.4和0.5时,在部分数据集(如Vowel和Control)上性能则呈现下降趋势。这表明,$\alpha = 0.3$的设定既能有效剔除低质量的决策树,又能最大程度地避免将有潜在价值的决策树错误地舍弃。

因此,在确定$\alpha = 0.3$为合理的下阈值后,本文进一步评估了上阈值β的作用。在$\alpha = 0.3$的条件下,当β取值为0.8时,模型的ACC和NMI值在所有测试数据集上均表现出稳定且优异的性能,处于最优性能区间内。综上,这种分步式的参数选择策略证明了($\alpha = 0.3$,$\beta = 0.8$)的组合是基于实验数据的审慎选择,它最终确定了一个能在保证高鲁棒性的前提下,实现最优聚类效果的参数配置。

综上所述,虽然这4个参数确实会影响聚类结果,但它们对性能的影响很小,进一步表明VDTES-RFC具有较好的稳定性和较强的泛化能力。

3.4 消融实验分析

本节通过消融实验研究谷检测技术和三支集成选择机制对随机森林聚类性能的具体影响。实验中,分别对VDTES-RFC方法中谷检测技术、三支集成选择策略和二者同时进行消融。其中,去除谷检测模块的对应方法表示为VDTES-RFC-V;去除三支集成选择模块的方法表示为VDTES-RFC-E;同时去除两个模块的方法表示为VDTES-RFC-VE。为保证实验的公平,修改如上3种方法的参数配置与VDTES-RFC保持一致。实验结果如表4所示,其显示了每种方法在所有数据集上的得分,其中粗体表示最高得分。

表4 消融实验结果

Table 4 Results of ablation experiments

Method	Wine			Seeds		
	ACC	NMI	ARI	ACC	NMI	ARI
VDTES-RFC	**0.930**	**0.776**	**0.793**	**0.848**	**0.645**	**0.638**
VDTES-RFC-V	0.916	0.759	0.784	0.840	0.639	0.623
VDTES-RFC-E	0.913	0.756	0.757	0.823	0.615	0.602
VDTES-RFC-VE	0.888	0.747	0.755	0.817	0.610	0.569
Method	Movementlibras			OlivettFaces		
	ACC	NMI	ARI	ACC	NMI	ARI
VDTES-RFC	**0.464**	**0.569**	**0.289**	**0.665**	**0.809**	**0.505**
VDTES-RFC-V	0.450	0.555	0.275	0.664	0.802	0.499
VDTES-RFC-E	0.446	0.552	0.272	0.664	0.807	0.504
VDTES-RFC-VE	0.442	0.538	0.267	0.663	0.788	0.494
Method	Control			Vowel		
	ACC	NMI	ARI	ACC	NMI	ARI
VDTES-RFC	**0.755**	**0.773**	**0.632**	**0.359**	**0.364**	**0.171**
VDTES-RFC-V	0.731	0.769	0.625	0.294	0.305	0.123
VDTES-RFC-E	0.731	0.767	0.612	0.335	0.350	0.159
VDTES-RFC-VE	0.656	0.749	0.576	0.289	0.301	0.119
Method	Dig			MSRA		
	ACC	NMI	ARI	ACC	NMI	ARI
VDTES-RFC	**0.675**	**0.655**	**0.546**	**0.543**	**0.617**	**0.376**
VDTES-RFC-V	0.670	0.650	0.533	0.538	0.602	0.361
VDTES-RFC-E	0.670	0.649	0.536	0.567	0.604	0.360
VDTES-RFC-VE	0.669	0.639	0.525	0.533	0.600	0.346

消融实验结果表明,融合谷检测技术与三支集成选择策略对VDTES-RFC的性能具有关键作用,特别是三支集成选择策略。移除谷检测(VDTES-RFC-V)后,Wine、Control等数据集的ACC下降1.5%和3.28%,其原因是谷检测的缺失导致分裂点选择依赖传统策略,无法捕捉原始数据的局部极小值特征,削弱了数据划分的区分度;移除三支集成选择策略(VDTES-RFC-E)后,Vowel、MSRA等数据集的NMI下降4%和2.15%,这是由于全集成策略保留了低质量决策树,导致相似度矩阵受噪声干扰加剧;同时移除两者(VDTES-RFC-VE)时,性能下降最为显著(如Seeds数据集ARI下降12.1%),表明谷检测技术与三支集成选择策略具有协同优化效应。谷检测技术提升了每棵决策树的划分质量,而三支集成选择策略通过筛选优化提高了随机森林性能,同时也提高了随机森林的效率。两者共同保障了相似度矩阵的鲁棒性与信息完整性,从而保证VDTES-RFC具有较好的聚类效果和性能。

3.5　运行时间对比分析

本节对VDTES-RFC与各基准方法的运行时间进行了统计与分析,如表5所示。所有实验均在相同硬件环境下进行,运行时间为10次独立实验的平均值。每个数据集的最快处理速度使用粗体突出标注。

从表5的数据可以清晰地看出,本文提出的VDTES-RFC方法在计算效率上具有显著优势。与RatioRF等采用全集成策略的基准方法不同,VDTES-RFC并非利用全部100棵决策树进行最终的共识函数计算,而是通过三支决策机制筛选出一个规模更小但质量更优的决策树子集。这一步骤极大地减少了后续聚类阶段的计算量,从而在保证甚至提升性能的同时,显著降低了时间开销。因此,在8个数据集中,VDTES-RFC在6个上取得了最快的运行速度,尤其在Control和Vowel这类较为复杂的数据集上,效率提升尤为明显,例如在Control数据集上,其运行时间相比性能最接近的RatioRF减少了约32%。虽然在Dig和MSRA两个数据集上,计算最为简单的m_Shi方法运行最快,但其聚类性能在所有方法中表现最差。相比之下,VDTES-RFC在所有高性能方法中效率最高,在性能和效率之间取得了更优的平衡。综上所述,运行时间对比实验有力地证明了VDTES-RFC通过引入三支集成选择策略,实现了计算效率的显著提升,成功达成了效率与性能双优化的设计目标。

表5　运行时间

Table 5　Running time　s

数据集	m_Shi	m_Zhu2	m_Zhu3	m_Ting	RatioRF	VDTES-RFC
Wine	3.12	3.49	4.12	4.20	3.45	**3.08**
Seeds	3.41	4.05	4.73	4.89	4.00	**2.84**
Movementlibras	14.18	15.35	17.40	17.40	15.35	**14.17**
OlivettFaces	15.43	17.77	22.64	22.64	17.53	**15.41**
Control	33.35	37.37	46.11	50.64	37.65	**25.64**
Vowel	28.83	50.47	186.28	142.23	48.35	**23.81**
Dig	**22.02**	199.77	1320.11	883.30	152.37	41.76
MSRA	**81.39**	123.71	456.44	354.34	124.27	112.19

4　结束语

本文针对随机森林聚类方法中存在的负类标签依赖与噪声决策树干扰问题,提出了一种基于谷检测和三支集成选择的随机森林聚类方法VDTES-RFC。通过引入谷检测技术自动捕捉数据分布中的

潜在分裂点,降低了人为引入负类标签对原始数据区分度削弱的问题;提出三支集成选择策略,动态筛选高质量决策树,显著降低了噪声对集成结果的影响,实现了性能与效率的双优化。

本文工作为无监督聚类任务提供了新的解决思路,未来可进一步探索其在动态数据、多模态分析及其他领域中的应用潜力。

参 考 文 献:

[1] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45: 5-32.

[2] 关晓蔷,王文剑,庞继芳,等.基于空间变换的随机森林算法[J].计算机研究与发展,2021, 58(11): 2485-2499.
GUAN Xiaoqiang, WANG Wenjian, PANG Jifang, et al. Random forest algorithm based on space transformation[J]. Journal of Computer Research and Development, 2021, 58(11): 2485-2499.

[3] 张梦, 郑建宏, 刘香燕, 等. 基于集成学习的全双工中继系统安全中继选择方案研究[J]. 电子学报, 2021, 49(9): 1852-1856.
ZHANG Meng, ZHENG Jianhong, LIU Xiangyan, et al. Research on secure relay selection scheme for full-duplex relay system based on ensemble learning[J]. Acta Electronics Sinica, 2021, 49(9): 1852-1856.

[4] DALMAIJER E S, NORD C L, ASTLE D E. Statistical power for cluster analysis[J]. BMC Bioinformatics, 2022, 23(1): 205-232.

[5] 张文, 陈锦富, 蔡赛华, 等. 一种聚类分析驱动种子调度的模糊测试方法[J]. 软件学报, 2024, 35(7): 3141-3161.
ZHANG Wen, CHEN Jinfu, CAI Saihua, et al. A fuzzing method for seed scheduling driven by cluster analysis[J]. Journal of Software, 2024, 35(7): 3141-3161.

[6] BICEGO M. Dissimilarity random forest clustering[C]//Proceedings of 2020 IEEE International Conference on Data Mining (ICDM). [S.l.]: IEEE, 2020: 936-941.

[7] SHI T, HORVATH S. Unsupervised learning with random forest predictors[J]. Journal of Computational and Graphical Statistics, 2006, 15(1): 118-138.

[8] ZHU X T, CHANGE LOY C, GONG S G. Constructing robust affinity graphs for spectral clustering[C]//Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2014: 1450-1457.

[9] BICEGO M, CICALESE F, MENSI A. RatioRF: A novel measure for random forest clustering based on the Tversky's ratio model[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(1): 830-841.

[10] SUN Z, WANG G, LI P, et al. An improved random forest based on the classification accuracy and correlation measurement of decision trees[J]. Expert Systems with Applications, 2024, 237: 121549.

[11] 邵长龙,孙统风,丁世飞.基于信息熵加权的聚类集成算法[J].南京大学学报(自然科学), 2021, 57(2): 189-196.
SHAO Changlong, SUN Tongfeng, DING Shifei. Clustering ensemble algorithm based on information entropy weighting[J]. Journal of Nanjing University (Natural Sciences), 2021, 57(2): 189-196.

[12] ZHANG P, LI T, WANG G, et al. Multi-source information fusion based on rough set theory: A review[J]. Information Fusion, 2021, 68: 85-117.

[13] ZHOU P, DU L, LIU X, et al. Self-paced clustering ensemble[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(4): 1497-1511.

[14] METAXAS I M, TZIMIROPOULOS G, PATRAS I. Divclust: Controlling diversity in deep clustering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 3418-3428.

[15] MA T, YU T, WU X, et al. Multiple clustering and selecting algorithms with combining strategy for selective clustering ensemble[J]. Soft Computing, 2020, 24: 15129-15141.

[16] ABBASI S, NEJATIAN S, PARVIN H, et al. Clustering ensemble selection considering quality and diversity[J]. Artificial Intelligence Review, 2019, 52(2): 1311-1340.

[17] DONG X, YU Z, CAO W, et al. A survey on ensemble learning[J]. Frontiers of Computer Science, 2020, 14: 241-258.

[18] LU X, YANG Y, WANG H. Selective clustering ensemble based on covariance[C]//Proceedings of the 11th Multiple Classifier Systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 179-189.

[19] YU Z, LI L, GAO Y, et al. Hybrid clustering solution selection strategy[J]. Pattern Recognition, 2014, 47(10): 3362-3375.

[20] FERN X Z, BRODLEY C E. Solving cluster ensemble problems by bipartite graph partitioning[C]//Proceedings of the Twenty-First International Conference on Machine Learning. New York: ACM, 2004: 36.

[21] WANG Y, LI X, WONG K C, et al. Evolutionary multiobjective clustering algorithms with ensemble for patient stratification [J]. IEEE Transactions on Cybernetics, 2021, 52(10): 11027-11040.

[22] 李金玉, 刘静玮, 杜明晶,等. 基于聚类集成选择的随机森林聚类方法[J].计算机工程与设计,2025, 46(4): 990-996.
LI Jinyu, LIU Jingwei, DU Mingjing, et al. Random forest clustering method based on cluster ensemble selection[J]. Computer Engineering and Design, 2025, 46(4): 990-996.

[23] LIU B, XIA Y, YU P S. Clustering through decision tree construction[C]//Proceedings of the 9th International Conference on Information and Knowledge Management. New York: ACM, 2000: 20-29.

[24] 刘盾,李天瑞,李华雄.粗糙集理论: 基于三支决策视角[J].南京大学学报(自然科学),2013, 49(5): 574-581.
LIU Dun,LI Tianrui, LI Huaxiong. Rough set theory: A three-way decisions perspective[J]. Journal of Nanjing University (Natural Sciences), 2013, 49(5): 574-581.

[25] 钱进,郑明晨,周川鹏,等.多粒度三支决策研究进展[J].数据采集与处理,2024, 39(2): 361-375.
QIAN Jin, ZHENG Mingchen, ZHOU Chuanpeng, et al. Research progress on multi-granularity three-way decisions[J]. Journal of Data Acquisition and Processing, 2024, 39(2): 361-375.

[26] HE X, YAN S, HU Y, et al. Learning a locality preserving subspace for visual recognition[C]//Proceedings of Ninth IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2003: 385-392.

[27] TING K M, ZHU Y, CARMAN M, et al. Overcoming key weaknesses of distance-based neighbourhood methods using a data dependent dissimilarity measure[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: IEEE, 2016: 1205-1214.

[28] MONDAL R, IGNATOVA E, WALKE D, et al. Clustering graph data: The roadmap to spectral techniques[J]. Discover Artificial Intelligence, 2024, 4: 7-30.

作者简介:

谭诚(2002-),男,硕士研究生,研究方向:数据挖掘与随机森林聚类,E-mail: tancheng@jsnu.edu.cn。

李金玉(1994-),男,硕士研究生,研究方向:数据挖掘与随机森林聚类,E-mail: jinyu@jsnu.edu.cn。

张文斌(1976-),**通信作者**,男,讲师,研究方向:随机森林聚类与时间序列预测,E-mail: zwbwen@jsnu.edu.cn。

杜明晶(1989-),男,副教授,研究方向:粒计算与聚类分析,E-mail: dumj@jsnu.edu.cn。

(编辑:刘彦东)

Random Forest Clustering Based on Valley Detection and Three-Way Ensemble Selection

TAN Cheng, LI Jinyu, ZHANG Wenbin*, DU Mingjing

(Jiangsu Key Laboratory of Educational Intelligent Technology, School of Artificial Intelligence and Computer Science, Jiangsu Normal University, Xuzhou 221116, China)

Abstract: As an unsupervised learning method, although random forest clustering demonstrates strong robustness in processing high-dimensional and complex data, it still faces challenges such as weak discriminability of original data caused by the introduction of negative samples and the interference of noisy decision trees on clustering performance. To address these issues, this paper proposes a random forest clustering based on valley detection and three-way ensemble selection (VDTES-RFC) method. First, the valley detection technology is utilized to identify potential split points for generating training data, and the Gini index is calculated to determine the optimal split points to complete the training of the classification forest. Second, each decision tree is treated as a base clusterer to extract its similarity matrix, and a three-way ensemble selection strategy is adopted to select high-quality decision trees to construct a new forest. Finally, a consensus function is used to integrate the similarity matrices to obtain the final clustering result. Experimental results demonstrate that this method effectively improves clustering accuracy and robustness, achieving dual optimization of efficiency and performance.

Highlights:

1. The paper proposes a valley detection and three-way ensemble selection-based random forest clustering (VDTES-RFC) method to overcome original data discriminability loss and noisy decision tree interference.
2. The paper develops a dual-stage clustering scheme centered on potential split point optimization and dynamic tree filtering. It aligns Gini index-based data partitioning with similarity matrix extraction to ensure high-quality base clusterers.
3. The paper adopts a three-way ensemble selection strategy combined with a consensus function to filter high-quality decision trees. It achieves a dual optimization of clustering efficiency and performance.

Key words: decision tree; random forest clustering; valley detection; cluster ensemble selection; three-way decisions

---

Foundation items: National Natural Science Foundation of China (No.62006104); Postgraduate Research & Practice Innovation Program of Jiangsu Normal University (No.2025XKT1457).

Received: 2025-06-15; **Revised:** 2025-07-03

***Corresponding author, E-mail**: zwbwen@jsnu.edu.cn.