第55卷第3期
2023年6月

南京航空航天大学学报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 55 No. 3
Jun. 2023

DOI:10.16356/j.1005-2615.2023.03.019

结构与纯度结合的新型决策树分裂准则

杜　斐[1,2]，陈松灿[1,2]

（1.南京航空航天大学计算机科学与技术学院，南京 211106；
2.工信部模式分析和机器智能重点实验室，南京 211106）

摘要：决策树（Desision tree，DT）生长关键步骤的分裂或分叉准则通常根据纯度和误分类误差等实现，分裂生长分为轴平行和非轴平行方式。这些分裂准则一般与数据内在结构（如类别是否是多簇或单簇组成）无关。为了弥补这一缺失，本文提出了两种混合分裂准则，分别用加权和两步法将同类内的节点间距（Between-node margin within the same class，BNM）和同一节点内的类紧性（Within-class compactness and between-class separation in the same inner node，CSN）与纯度度量相结合。由于传统决策树以贪婪方式生长，仅能确定出当前的一个局部最优分裂点，为改善这个缺点，本文首先根据纯度确定出前k个候选分裂点，然后通过最大化BNM和最小化CSN确定最终的分裂点，不仅缓和了纯度上的局部最优性，而且引入了数据结构的全局性，因此能较大程度地改进后代节点的分裂，增强树的泛化性和可解释性。将上述两种分裂准则组合还可以进一步提升性能。在21个标准验证数据集上的比较结果表明：新准则下的决策树不仅提高了预测性能、降低了复杂性，而且相比于其他采用混合分裂准则的DTs更具竞争力。

关键词：决策树；分裂准则；全局结构；纯度；数据结构

中图分类号：TP181　　**文献标志码：**A　　**文章编号：**1005-2615(2023)03-0534-10

Novel Splitting Criteria for Decision Trees with Combination of Structure and Purity

DU Fei[1,2]，*CHEN Songcan*[1,2]

（1. College of Computer Science and Technology，Nanjing University of Aeronautics & Astronautics，Nanjing 211106，China；
2. MIIT Key Laboratory of Pattern Analysis and Machine Intelligence，Nanjing 211106，China）

Abstract: As a critical part of desision tree（DT）growth，its nodes can be split to grow by either axis- or non-axis-aligned way based on such splitting criteria as purity and misclassification error. However，these have nothing to do with the geometric structure of data，e.g. multicentric data or single-center data. In order to compensate for this，two splitting criteria are proposed by combining the between-class margin in the same inner node（BCM）and the between-node margin within the same class（BNM）respectively with the purity measure in weighting and the two-step method. Unlike traditional greedy growth of DT which only finds the current locally optimal splitting point，the proposed method first selects the top-k purity splitting nodes，then determines the optimal one by maximizing BCM and minimizing BNM. Since not only alleviating purity-based local optimality but also considering global structures of the data，our method greatly improves the division of descendant nodes and generalization of the formed trees，while enhancing the interpretability. In addition，two aforementioned splitting criteria can be combined to further boost the performance. The comparison results on 21 benchmark datasets show an improvement in predictive performance of new trees with reduction in

基金项目：国家自然科学基金（62076124）。
收稿日期：2022-06-07；**修订日期：**2022-10-10
通信作者：陈松灿，男，教授，博士生导师，E-mail：s.chen@nuaa.edu.cn。

引用格式：杜斐，陈松灿．结构与纯度结合的新型决策树分裂准则[J]．南京航空航天大学学报，2023，55(3)：534-543.
DU Fei，CHEN Songcan. Novel splitting criteria for decision trees with combination of structure and purity[J]. Journal of Nanjing University of Aeronautics & Astronautics，2023，55(3)：534-543.

complexity, while also are competitive with many other DTs using hybrid splitting criteria.

Key words: decision tree; split criterion; global structure; purity; data structure

作为一种监督学习算法，决策树本质上是通过一系列决策规则对数据进行分类的预测模型。其简单的树形结构接近于人类在面临决策问题时的一种自然的处理机制[1]，决策规则易于理解和解释，实现相对简单。在航空航天领域，决策树常用于故障检测和诊断[2-3]以及对环境、设备的在线监测[3]。与其他机器学习算法相比，决策树具有以下优势：(1)决策树具有可解释性，其决策规则便于用户理解检测和诊断的结果；(2)与神经网络相比，决策树不需要设置参数的先验假设，也不需要优化参数，因此时间复杂度较低[4]，在大型数据集上，它能够以较短的时间完成训练并获得很好的预测性能；(3)决策树可以处理多分类任务，如在多故障环境中进行训练[3]；(4)决策树能够拟合非线性和高维数据[5]。

决策树的分裂准则在很大程度上可以决定树的生长方向和性质，是很活跃的研究领域[6-7]。经典的分裂准则通常用分裂后子节点的纯度来表示分裂点的判别能力，比如：ID3[8]、C4.5[9]、分类回归树(Classification and regression tree，CART)[10-12]。ID3和C4.5都是基于信息熵的启发式算法，分别用信息增益和增益比定义分裂准则，CART根据基于基尼系数的启发式思想，沿减小节点不纯度最显著的分裂点进行分裂。但这些分裂准则仅使用了基于频率的纯度度量，只能贪心地选择当前纯度最优的分裂点，如果能够进一步挖掘数据的附加信息，也许决策树可以做出对整棵树生长更好的选择。目前仅有少部分针对该问题的研究，例如，Segment+C4.5[7]、SPES[13]、SPCE[14]、CV-ID3及FR-ID3[4]。然而，上述方法往往只考虑了单一属性上的数据分布，无法同时考虑数据的所有维度。此外，在最新的研究中，dGMML-DT根据数据的结构设计出了一个严格的凸目标函数，通过对其优化求解，找到最佳分裂特征[15]，这说明数据的结构对于决策树分裂是有效的判别信息。因此，本文提出了两种基于数据全局结构的混合分裂准则，与以往的研究不同的是，本文的方法从整体上考虑簇与簇之间的关系，分裂点满足纯度需求，且在一定程度上尊重簇间结构，从而提高分裂的可解释性和泛化性。

在多簇数据中，一个类有多个簇中心，基于纯度的分裂准则划分多簇数据集并不理想，极有可能会将一个簇切分开，令树的分裂生长更复杂，有过拟合的风险。对于簇中心不同的两个同类簇A、B，尽管类标签相同，但在当前节点的数据分布上可能会更容易分开，因此，希望在此次分裂时能将A、B分离，即不同节点内的同类间距尽可能大。在决策树的生长过程中，不同分支的节点空间是互相独立的，所以这次划分可能会使A、B在各自节点空间中的后续划分更简单，从而降低树的复杂性。对于划分到同个节点的不同类，应使类内更加紧凑、类间尽量分离，这相当于假定划分到同个节点中的每一个类都为一个样本簇，直观上，希望节点内的不同样本簇满足簇内相似性高且簇间相似度低的特性[1]，因此，最佳分裂点应当最大化子节点中簇内紧凑性及簇间离散度。

决策树的分裂生长通常选择在当前情况下令不同类尽可能分离的分裂点，是一个贪心的局部最优策略。从优化的角度来看，贪心搜索往往不能找到全局最优解，而束搜索没有选择当前局部最优，却可能找到比其更接近全局最优的解[16-17]。同样，在决策树中，有些分支虽无法提升此时的泛化能力，甚至导致泛化能力暂时下降，但在其基础上的后续划分却有可能使性能显著提高[1]。本文方法并不绝对贪心，它不要求在一次划分中尽可能分离不同类，而是根据纯度选择局部k优的k个候选分裂点，再利用全局结构信息得到最佳分裂点。

本文分别用加权和两步法把两种度量同类内的节点间距(Between-node margin within the same class，BNM)和同一节点内的类紧性(Within-class compactness and between-class separation in the same inner node，CSN)与纯度度量结合，提出了两种同时考虑全局结构和纯度的混合分裂准则，具体地，它们表示在判别能力较好的几个候选分裂点中，选择不同节点内同类间距最大和同一节点内类内紧凑、类间分离的分裂点，能在一定程度上顺应数据分布划分，降低树的复杂度和预测错误率。此外，本文又组合了上述两种分裂准则，以进一步提高性能。

1　混合分裂准则

决策树通常根据判别能力选择最佳分裂点，而对于两个及两个以上判别能力相似或相同的候选分裂点，无法确定哪个更合适[7]。混合分裂准则能够从多角度选择分裂点，进而诱导出一个更好、更小的树。

1.1 Segment+C4.5

Wang等[7]提出了实例片段(Segment of examples)的概念,即:节点内的样本s按某一属性A_j进行排序后,根据类标签切割得到的几个子队列。定义节点中的片段数为节点中按不同属性排序得到的最小片段数,有

$$\mathrm{Seg}\#(s)=\min_j\left|\mathrm{Seg}(s,A_j)\right| \tag{1}$$

式中$\left|\mathrm{Seg}(s,A_j)\right|$表示在$s$中按照$A_j$排列得到的片段数。Seg#($s$)衡量了该属性上类排列的复杂程度。片段数越少意味着在纯度相似的候选点中,该分裂点的判别能力更强。Segment+C4.5是一种以两步法结合纯度和片段数两种度量的混合分裂准则,首先选出接近最佳纯度的k个候选分裂点,然后取候选分裂点中片段数最小的分裂点作为最佳分裂点$x_{j^*i^*}$,即

$$x_{j^*i^*}=\arg\min_{x_{ij}\in x}\left[\frac{|s_1|}{|s|}\left|\mathrm{Seg}(s_1;\,x_{ji})\right|+\frac{|s_2|}{|s|}\left|\mathrm{Seg}(s_2;\,x_{ji})\right|\right] \tag{2}$$

式中:|*|表示该数据集的样本数;s_1、s_2分别为左右子节点;$\left|\mathrm{Seg}(s_1;\,x_{ij})\right|$表示在$s$中按照$x_{ij}$排列得到的片段数。

Segment+C4.5不仅有效地提高了泛化性,同时缩小了树的大小。

1.2 SPES

Yan等[13]在Segment+C4.5的基础上,引入加权因子来平衡预期片段数和纯度,即:将纯度和片段数以加权的方式结合,有

$$\mathrm{HM}=\arg\max\ \alpha\times F(s_p)+(1-\alpha)\times S(E(\mathrm{seg})) \tag{3}$$

式中:s_p为分裂点;seg为最小片段数;α为加权因子;$S(E(\mathrm{seg}))$表示归一化后的最小片段数;$F(s_p)$表示该点分裂后的纯度度量。

当$\alpha>0.5$,最佳分裂点的选择更倾向于频率,反之,更偏向片段信息。这样构建的统一模型可以扩展到任何单变量决策树。此外,他们提出了两种归一化方法规范预期片段数$S(E(\mathrm{seg}))$,使其和分裂性能$F(s_p)$具有相同的数量级。

1.3 SPCE

Yan等[14]综合训练数据的分布和排列信息,提出了一种基于分裂点校正矩阵的混合分裂准则,用截断矩阵的形式实现了结合片段数、分裂比率和Hellinger距离的两步法。矩阵中每行包含Hellinger距离hd_i、子节点片段数之和$\mathrm{segments}_i$以及分裂点的分裂比率ratio_i。首先根据Hellinger距离对矩阵降序排序得到k行截断矩阵,然后在截断矩阵中选择$\mathrm{segments}_i$最小、ratio_i最大的分裂点作为最佳分裂点。其中,分裂比率定义为:沿分裂点a_i切分得到两个子节点的片段数比值,有

$$\mathrm{Ratio}(s,a_i,T)=\begin{cases}\dfrac{\mathrm{Seg}(s_2;\,a_i)}{\mathrm{Seg}(s_1;\,a_i)} & \mathrm{Seg}(s_1;\,a_i)>\mathrm{Seg}(s_2;\,a_i)\\ \dfrac{\mathrm{Seg}(s_1;\,a_i)}{\mathrm{Seg}(s_2;\,a_i)} & \mathrm{Seg}(s_1;\,a_i)\leqslant\mathrm{Seg}(s_2;\,a_i)\end{cases} \tag{4}$$

除此之外,Shi等[4]考虑了数据中不同属性的聚类价值和故障率,提出了两种基于高斯混合模型的改进ID3算法:CV-DT和FR-DT。Jaworski等[18]分别将基于错误分类的i型准则和基于基尼系数的i型及ii型准则结合,提出了两种混合分裂准则。

近两年,混合分裂准则的研究仍有进展。Yan等[19]针对决策树在平衡或大致平衡的二分类数据集中训练时经常出现的局部类不平衡问题,提出了一种基于现有多个分裂准则的自适应算法,以局部类不平衡比ir作为权重因子,平衡各个分裂准则之间的重要性。Barata等[20]提出了一个公平的混合分裂准则SCAFF,用一个可调参数权衡分裂点的分类性能和公平性。Rahman等[21]提出了一种用于增量学习的混合分裂准则,即:改进分离轴定理(improved Separating axis theorem,iSAT),根据已知类和未知类混合的新批次数据找到两个轴平行最小边界框(Axis aligned minimum bounding box,AABB),在这两个AABB上,分别考虑基于熵的分裂准则和基于分离轴定理(Separating axis theorem,SAT)的分裂准则。

本文的方法和上述混合分裂准则的对比情况如表1所示。混合分裂准则以某种方式结合混合信息1和混合信息2,多角度衡量分裂点的判别能力,选择最佳分裂点。Segment+C4.5、SPES、SPCM都是从某属性上类排列的复杂程度中获取信息,CV-ID3考虑了某属性上的类间多样性和类内相似性,FR-ID3从聚类复杂度和先验知识中获取信息,AdaDT为了解决局部类不平衡问题,用权重因子权衡纯度和类概率分布的重要性,iSAT同时考虑了新数据中纯度和结构的信息。值得注意的是,上述混合分裂准则往往是从单一属性中得到的混合信息2。本文的方法与其不同,它从数据全局结构中获取混合信息2,考虑了各个维度的综合信息,避免了单一维度带来的片面影响,树的泛化性更强。此外,局部类不平衡问题经常存

表 1　混合分裂准则对比

Table 1　Comparison of hybrid splitting criteria

混合分裂准则	结合方式	混合信息 1	混合信息 2	混合信息 2 来源维度	混合信息 2 来源
Segment+C4.5	两步法	信息增益率	最小片段数	1	类排列
SPES	加权	信息增益率	最小片段数	1	类排列
SPCM	两步法	Hellinger距离	片段数、分裂比率	1	类排列
CV-ID3	乘积	信息增益	聚类有效性	1	结构
FR-ID3	和	聚类数量	故障率	1	频率
AdaDT	加权	信息增益率	Hellinger距离	1	类概率分布
iSAT	选择式	信息增益	SAT	1	结构
SCAFF	加权	AUC(Area under curve)	敏感特征的 AUC	$m_{\text{敏感}}$	公平性
BNM结合纯度	两步法	基尼系数	不同子节点内同类间距	m	结构
CSN结合纯度	加权	基尼系数	同一节点内类紧性	m	结构
BNM+CSN结合纯度	两步法、加权	基尼系数	不同子节点内同类间距、同一节点内类紧性	m	结构

在于决策树的生长过程中,本文的树也可以用AdaDT中的ir赋予分裂准则不同的重要性,避免树中的局部类不平衡。SCAFF是针对有偏数据的敏感属性设计出的公平分裂准则,iSAT是应用于增量学习的分裂准则,在实验部分,由于iSAT和SCAFF应用场景的特殊性,本文没有与其进行对比实验。

2　基于结构和纯度的混合分裂准则

基于结构的决策树类似于采用“自顶而下”拆分策略的层次聚类,决策树的向下划分类似于将一组粗粒度聚类簇拆分为更多细粒度聚类簇的过程[22-23]。树中的分裂节点把当前的特征空间划分为更小的几个特征子空间。划分后,同一节点内的不同类为不同聚类簇,不同节点内的同个类也是不同聚类簇,根据“物以类聚”的思想,同一簇的样本应尽可能相似,不同簇的样本应尽可能不同,所以最佳划分应当最大化不同节点内的同类间距和同一节点内的类间距,最小化同一节点内的类内离散度。

本文在二叉树中实现基于结构和纯度的混合分裂准则,其目标是找到当前节点中的最佳分裂点(Optimal splitting point ,OSP)。

2.1　BNM结合纯度

现假设当前样本集合D按分裂点a进行分裂得到子节点D_{l}、D_{r},其中$a=(t,v)$,表示沿属性t上的v分裂。令$X_{i,\text{l}}^{(c)}$、$X_{i,\text{r}}^{(c)}$分别表示左右子节点中第c类的第i个样本,Cl_{s}表示子节点中相同类的个数,$\boldsymbol{m}_{\text{l}}^{(c)}$、$\boldsymbol{m}_{\text{r}}^{(c)}$分别表示左右子节点中第$c$类的均值中心。

需注意,在每次分裂前,需要对父节点D进行归一化。本文用样本簇均值中心之间的欧几里德距离(Euclidean distance)[24]来表示同类间距。BNM的公式为

$$\text{BNM}(D,a)=\frac{\sum_{c=1}^{Cl_{\text{s}}}(\boldsymbol{m}_{\text{l}}^{(c)}-\boldsymbol{m}_{\text{r}}^{(c)})^{\text{T}}(\boldsymbol{m}_{\text{l}}^{(c)}-\boldsymbol{m}_{\text{r}}^{(c)})}{Cl_{\text{s}}} \tag{5}$$

由于不同分裂点的Cl_{s}不同,所以需要对同类间距求平均以避免偏心多分类的分裂点。此外,考虑到式(5)也会偏向于特征空间内更边缘的分裂点,即:将父节点划分为一个极度小的单类子节点和一个极度大的多类子节点。因此,BNM应增加一个惩罚项,在同类间距尽可能大的同时,令两个子节点内不同类之间的最短距离尽可能小,即有

$$\begin{aligned}\text{BNM}(D,a)=&\frac{\sum_{c=1}^{cl_{\text{s}}}\left(\boldsymbol{m}_{\text{l}}^{(c)}-\boldsymbol{m}_{\text{r}}^{(c)}\right)^{\text{T}}\left(\boldsymbol{m}_{\text{l}}^{(c)}-\boldsymbol{m}_{\text{r}}^{(c)}\right)}{cl_{\text{s}}}-\\&\left(\frac{\sum_{p=1}^{Cl_{\text{nl}}}\min\left(\|X_{i,\text{l}}^{(p)}-X_{i,\text{l}}^{(!=p)}\|_2^2\right)}{Cl_{\text{nl}}}+\right.\\&\left.\frac{\sum_{q=1}^{Cl_{\text{nr}}}\min\left(\|X_{i,\text{r}}^{(q)}-X_{i,\text{r}}^{(!=q)}\|_2^2\right)}{Cl_{\text{nr}}}\right)\end{aligned} \tag{6}$$

式中:$\|X_{i,\text{l}}^{(p)}-X_{i,\text{l}}^{(!=p)}\|_2$、$\|X_{i,\text{r}}^{(q)}-X_{i,\text{r}}^{(!=q)}\|_2$表示不同类之间的欧几里德距离;$Cl_{\text{nl}}$、$Cl_{\text{nr}}$表示左右子节点内的类别数。在实际应用中,实现式(6)的时间复杂度非常大,所以将两个节点中不同类之间的最短距离近似表示为$\dfrac{\sum_{c=1}^{Cl'}\min(|X_i^{(c)}-\text{thr}|+|\text{thr}-X_i^{(!=c)}|)}{Cl'}$,即:在该维度上,不同类和分裂阈值thr之间的最小差值和,其中:Cl'表示两个子节点间不同类的个数,$X_i^{(c)}$表示节点内第c类的第i个样本。则式(6)可以近似表示为

$$\mathrm{BNM\_}(D,a)=\frac{\sum_{c=1}^{Cl_s}\left(\boldsymbol{m}_{\mathrm{l}}^{(c)}-\boldsymbol{m}_{\mathrm{r}}^{(c)}\right)^{\mathrm{T}}\left(\boldsymbol{m}_{\mathrm{l}}^{(c)}-\boldsymbol{m}_{\mathrm{r}}^{(c)}\right)}{Cl_s}-\left(\frac{\sum_{p=1}^{Cl_{\mathrm{nl}}}\min\left(\left|X_{i,\mathrm{l}}^{(p)}-\mathrm{thr}\right|+\left|\mathrm{thr}-X_{i,\mathrm{l}}^{(!=p)}\right|\right)}{Cl_{\mathrm{nl}}}+\frac{\sum_{q=1}^{Cl_{\mathrm{nr}}}\min\left(\left|X_{i,\mathrm{r}}^{(q)}-\mathrm{thr}\right|+\left|\mathrm{thr}-X_{i,\mathrm{r}}^{(!=q)}\right|\right)}{Cl_{\mathrm{nr}}}\right)\tag{7}$$

然后将其与纯度度量加权结合

$$G(D,a)=\frac{|D^{\mathrm{l}}|}{|D|}G(D^{l})+\frac{|D^{\mathrm{r}}|}{|D|}G(D^{\mathrm{r}})\tag{8}$$

$$G\_\mathrm{BNM}(D,a)=\omega_1 G(D,a)+\omega_2\mathrm{BNM\_}(D,a)\tag{9}$$

式中：$G(D^{\mathrm{l}})$、$G(D^{\mathrm{r}})$为子节点的纯度；$G(D,a)$为父节点的纯度；$|D|$、$|D^{\mathrm{l}}|$、$|D^{\mathrm{r}}|$分别表示对应数据集的样本数；ω_1、ω_2为加权因子。最后，找到使$G_\mathrm{BNM}(D,a)$最大的最佳分裂点(Optimal splitting point，OSP)为

$$\mathrm{OSP}=\arg\max_a G\_\mathrm{BNM}(D,a)\tag{10}$$

$\mathrm{BNM_}(D,a)$越大，子节点中同类样本属于同个簇的可能性越小，在结构上，此次划分的可靠性越强。当同类间距极小时，该划分极有可能是沿簇中心分裂，从而导致树的过度复杂。因此，选择$G_\mathrm{BNM}(D,a)$最大的分裂点，是在一定纯度的基础上，综合划分的结构可靠性做出的选择。

2.2 CSN结合纯度

本文用类内离散度和类间离散度的比值度量节点内类紧性[25]，比值越小说明类内越紧、类间越分散。对节点内的样本进行归一化后，节点内类紧性CSN的公式为

$$\mathrm{CSN}(D)=\frac{\sum_{k=1}^{K}\sum_{i=1}^{N_k}(\boldsymbol{X}_i^{(k)}-\boldsymbol{m}^{(k)})^{\mathrm{T}}(\boldsymbol{X}_i^{(k)}-\boldsymbol{m}^{(k)})}{\mathrm{BCM}(D)}\tag{11}$$

式中：$\boldsymbol{m}^{(k)}$表示该节点内第k类样本的均值中心；$\boldsymbol{X}_i^{(k)}$表示节点内第k类的第i个样本；N_k表示该节点内第k类的样本数；K表示节点内的类别数；$\mathrm{BCM}(D)$为节点内的类间距(Between-class margin in the same inner node，BCM)，对于二分类任务，直接计算两类均值中心之间的距离为

$$\mathrm{BCM}(D)=(\boldsymbol{m}^{(0)}-\boldsymbol{m}^{(1)})^{\mathrm{T}}(\boldsymbol{m}^{(0)}-\boldsymbol{m}^{(1)})\tag{12}$$

对于多分类任务，根据ECOC(Error-correcting output code)编码将其拆分为d个二分类任务，根据式(12)计算二分类任务的BCM之和作为总体类间距，即

$$\mathrm{BCM}(D)=\sum_{i=1}^{d}\mathrm{BCM}\left(D,\mathrm{ECOC}[i]\right)\tag{13}$$

由于不同子节点内样本数不同，且样本数越多，子节点的影响越大[1]，所以计算多个子节点的类紧性时需要对不同的分支赋予权重，即

$$\mathrm{CSN}(D,a)=\frac{|D^{\mathrm{l}}|}{|D|}\mathrm{CSN}(D^{\mathrm{l}})+\frac{|D^{\mathrm{r}}|}{|D|}\mathrm{CSN}(D^{\mathrm{r}})\tag{14}$$

本文尝试用两步法结合纯度和每次划分的$\mathrm{CSN}(D,a)$。首先，根据式(8)计算出纯度度量上的最佳分裂点OSP^*为

$$\mathrm{OSP}^*=\arg\max_a G(D,a)\tag{15}$$

然后把纯度接近OSP^*的前k个分裂点放进候选点集合CP中，根据式(14)计算其类紧性。最后，结合纯度和类紧性的最佳分裂点OSP表示为

$$\mathrm{OSP}=\arg\min_{a\in\mathrm{CP}}\mathrm{CSN}(D,a)\tag{16}$$

从判别能力的角度看，基于CSN和纯度的决策树是一个局部k优的算法，选出的OSP不仅具备较好的判别能力，而且在一定程度上遵循数据内在结构，决策规则更容易理解。

2.3 BNM+CSN结合纯度

BNM关注同类数据中的簇间差异性，CSN关注类间差异性及类内相似性，因此，将二者与纯度结合可以从簇间关系和类间关系上引入全局结构的知识。算法1描述了分别以加权和两步法结合BNM和CSN作为混合分裂准则的决策树构造算法。首先在当前节点中，根据各个特征的取值确定第一候选分裂点集合CP^1，对其中每个分裂点计算G_BNM，并取最大的k个分裂点作为第二候选分裂点集合CP^2，计算CSN，选择CSN最小的分裂点作为最佳分裂点。树内设置超参数：最大深度max_depth和最小叶内节点数min_leaf_sample，非叶节点根据得到的最佳分裂点进行分裂，并递归地进行此过程，直到满足递归停止条件：深度≥max_depth或节点内样本数≤min_leaf_sample。

算法1 结合BNM和CSN的决策树构造算法

输入：数据集D，最大深度max_depth，最小叶内样本数min_leaf_sample，候选点的个数k

输出：一个二分类决策树DT

(1) build_tree(X)

(2) if X中样本标签全部相同 or

　　样本数≤min_leaf_sample then

(3)　　令X为叶子节点，

　　并根据X中最多数的类别赋予其标签

(4) end if

(5) if当前深度＜max_depth：

(6)　　根据式(9)计算X各个分裂点的G_BNM

(7)　　将各个分裂点的G_BNM从大到小排序

(8)　取前k个分裂点组成候选点集合CP
(9)　for CP中的每个候选点CP_i do
(10)　　根据式(14)计算CP_i的CSN
(11)　　最小CSN的分裂点为最优分裂点OSP
(12)　end for
(13)　根据OSP分裂X，得到X_l，X_r
(14)　if $X_l <$ min_leaf_sample or
　　$X_r <$ min_leaf_sample then
(15)　　令X为叶子节点，
　　并根据X中最多数的类别赋予其标签
(16)　else
(17)　　build_tree(X_l)
(18)　　build_tree(X_r)
(19)　end if
(20) end if
(21) return build_tree(D)

2.4　时间复杂度分析

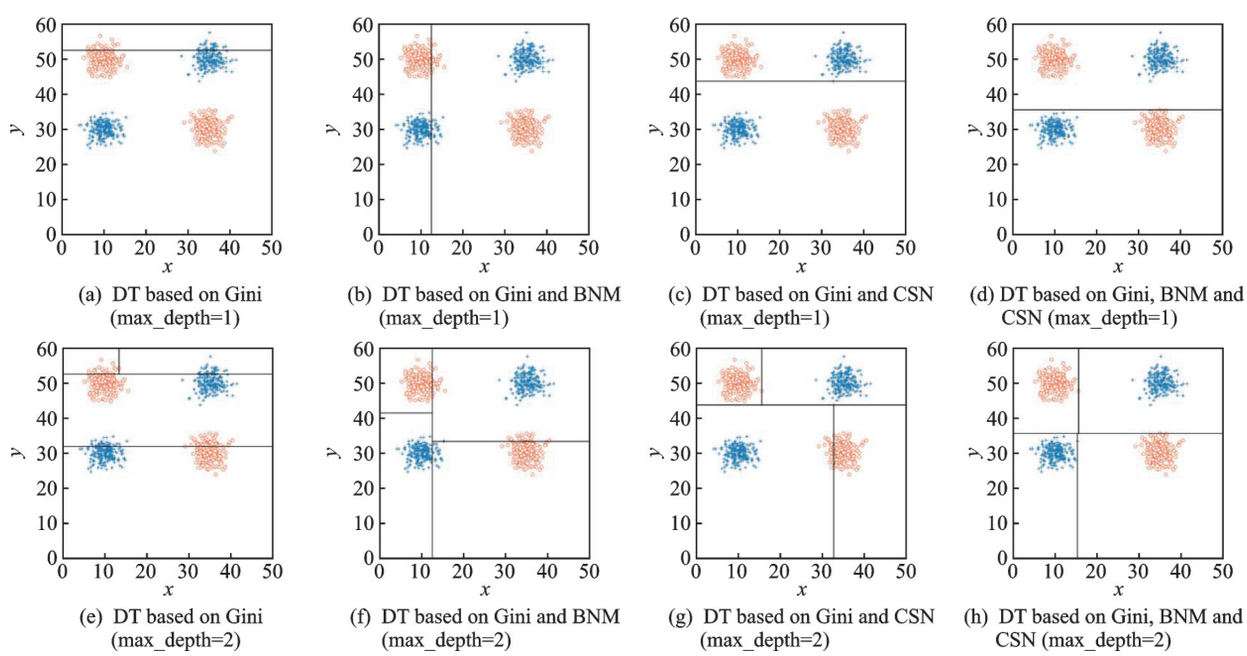
假设当前节点P中有N个样本，特征维度为m，特征i有α_i个第一候选分裂点，则第一候选点的总数CCP为$O\left(\sum_{i=1}^{m}\alpha_i\right)$。为了得到最佳分裂点，首先对样本的每个特征值进行排序，寻找第一候选分裂点，该过程复杂度为$O(mN\log N)$。然后，计算每个候选点的纯度度量，并保存最大值，该过程复杂度为$O((N+1)\cdot CCP)$。因此，CART中单个非叶节点的时间复杂度为：$O(mN\log N+(N+1)\cdot CCP)$。

计算每个候选点的BNM的复杂度为$O(N\cdot CCP)$，BNM结合纯度的时间复杂度为：$O(mN\log N+(2N+1)\cdot CCP)$。为了结合CSN，需要对当前得到的纯度进行排序，选择最大的K个候选点，计算其CSN，因此CSN结合纯度的时间复杂度为：$O(mN\log N+(N+1)\cdot CCP+CCP\log CCP+N\cdot K+K)$，BNM+CSN结合纯度的时间复杂度为：$O(mN\log N+(2N+1)\cdot CCP+CCP\log CCP+N\cdot K+K)$。通常情况下$K\ll N$，CSN结合纯度的时间复杂度近似于CART，BNM+CSN结合纯度的时间复杂度近似于BNM结合纯度。在下文的实验结果可看出，结合BNM的分裂准则往往能降低树高，减小非叶节点数，因此，通过BNM结合纯度或BNM+CSN结合纯度构造决策树的时间复杂度往往和CART相当，甚至更小。

3　实　　验

3.1　异或问题

经典的异或问题是一个多簇数据集上的二分类任务。实验用4个正态分布的二维聚类簇模拟异或问题，其中每个数据集各包含200个样本，空心圆和“+”分别代表两类数据。在实验中，分别用基尼系数、BNM结合基尼系数、CSN结合基尼系数和BNM+CSN结合基尼系数诱导出的4种决策树在生成的异或数据集上训练，结果见图1。如图1(a)所示，仅根据基尼系数划分异或数据集并不理想，由于无法找到一次分离开两类数据的分裂点，该树容易将同个簇划分到两个特征子空间中。如图1(e,i,m)所示，在前次划分的基础上，后续分裂把聚类簇不断地切分成细碎的区域，分裂结果在结构上是混乱无序的，因此其可解释性并不强。在面对异或问题时，人类往往通过两次判断就可以得到正确结果。如图1(d,h,l,p)所示，本文在用



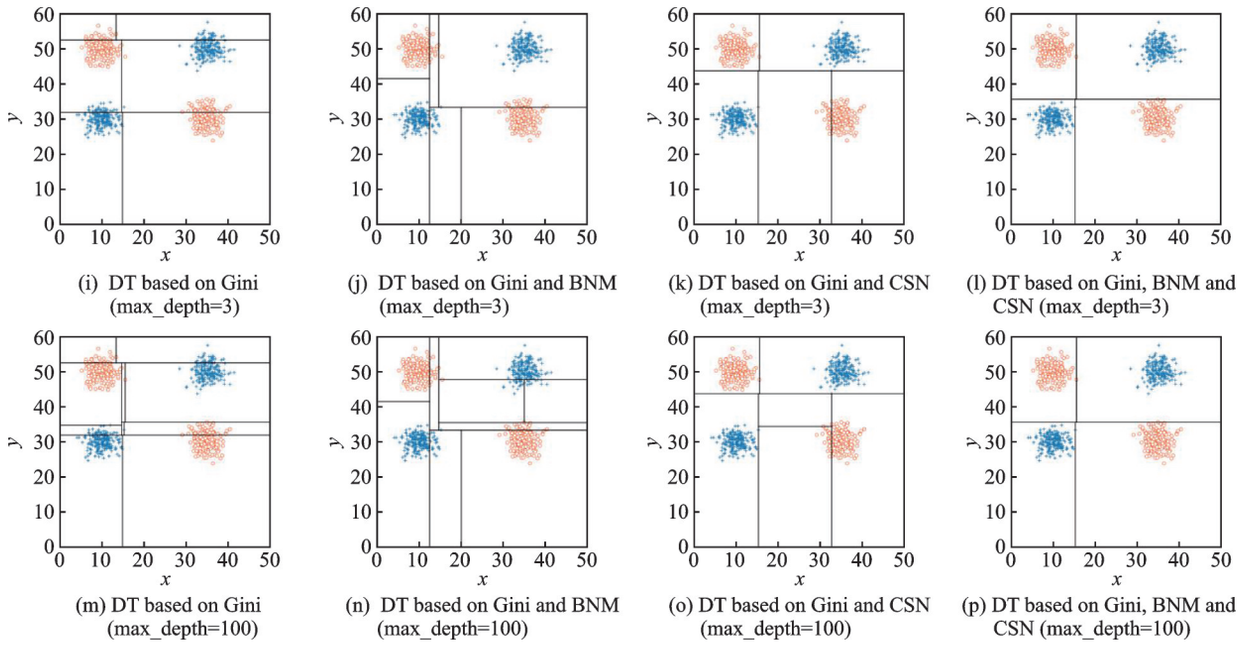


图1 在异或问题上不同分裂准则诱导出的决策树

Fig.1 Decision tree induced by different splitting criteria on XOR problem

BNM+CSN结合基尼系数诱导出的决策树中实现了这一情况，每一次划分都保证了簇的完整性，所以该树最小，可解释性最强，符合人类的思维方式。图中的切线紧贴聚类簇边界，并非在不同簇之间的正中位置，这是因为实验设置的候选切点是每个属性上的属性值，而非属性值的两两平均。和图1(a)相比，图1(b,c)中单独用BNM结合基尼系数和CSN结合基尼系数诱导出的切线更靠近簇边界，因此其生成的树更小，可解释性更强。

3.2 对比实验

3.2.1 实验设置

在实验中，本文从UCI[26]和KEEL[27]机器学习数据库中选取了21个二进制数据集，其维度和样本数多样化，表2包含了所选数据集的详细信息：样本数量(♯Insts)和特征数量(♯Ftrs)。同时，为了便于计算BNM和CSN，数据集中的属性全都为数值型。

下文用基尼系数(Gini)衡量不纯度，用“+Gini”表示结合纯度，节点内先进行Min-max归一化，再计算其BNM或CSN。为了评估此方法，本文与近年来提出的混合分裂准则进行了对比实验，即：segment+C4.5、SPES、SPCE。令min_leaf_samples=2，即：当叶内样本数≤2时，停止分裂该节点。此外，对于segment+C4.5、SPES、SPCE、CSN+Gini、BNM+CSN+Gini，设置候选分裂点的个数$k=\{2,3,5,7,10,15,20,30\}$，并选择最好的作为最佳参数。实验进行了5折交叉验证，最终结果为多个验证集的均值。本文分别用泛化性和模型复杂度评价决策树的性能，泛化性通过测试精度(acc)衡量，模型复杂度通过树深(depth)和叶子节点数(leaf_nodes)体现。

3.2.2 实验结果及分析

表2总结了使用最佳参数的不同混合分裂准则和C4.5、仅基于Gini的决策树的5折交叉验证的测试精度。易见，BNM+CSN+Gini在12个数据集中表现出最佳分类性能，BNM+Gini在4个数据集中表现了最佳性能，这表明本文的方法优于其他方法。用↑和↓表示本文的树相对于仅基于Gini的决策树是否提高了精度，其中，BNM+Gini和CSN+Gini分别在18和11个数据集上提高了精度，BNM+CSN+Gini只对1个数据集的精度造成了损害，易见BNM+Gini、CSN+Gini和BNM+CSN+Gini对分类性能都有所提升。这说明，本文提出的3种方法往往比仅用Gini诱导出的决策树具有更好的泛化性，同时结合类内和类间簇间关系的BNM+CSN+Gini在最多数据集上展示了最佳分类性能。这说明纯度最高的分裂点不一定是最佳分裂点，考虑全局结构和纯度的分裂准则可以诱导出泛化性更好的决策树。

表3总结了最佳参数下决策树的树深和叶子节点数。显然，对于不同数据集，树深和叶子节点数相差很大，由于基于Gini的决策树和C4.5之间的比较结果很明确，表格中只列出了前者的树深和叶子节点数。在后3列中，用√表示本文的树相对于仅用Gini诱导出的决策树降低了模型复杂度。在18个数据集中，BNM+Gini和BNM+CSN+

表 2　测试精度的对比结果

Table 2　Comparison results of test accuracy

Dataset	数据集的详细信息		C4.5	Gini	Segment+C4.5	SPES	SPCE	BNM+Gini	CSN+Gini	BNM+CSN+Gini
	#Insts	#Ftrs								
diabetes	768	8	0.681 0	0.723 9	0.720 0	0.713 5	0.703 1	0.748 7↑	0.722 6↓	0.736 9↑
yeast_banlance	1 484	8	0.659 7	0.701 5	0.670 5	0.685 3	0.686 0	0.721 0↑	0.683 3↓	0.724 4↑
Diabetic	1 762	19	0.624 7	0.628 2	0.634 2	0.630 8	0.626 4	0.648 1↑	0.617 7↓	0.630 8↑
Algerian_Forest_Fires	244	10	0.959 1	0.954 8	0.971 3	0.971 3	0.979 5	0.959 0↑	0.971 3↑	0.971 3↑
data_banknote_authentication	1 372	4	0.974 5	0.946 1	0.981 8	0.978 1	0.989 8	0.988 4↑	0.945 4↓	0.990 5↑
adult	32 561	108	0.819 4	0.839 8	0.818 6	0.820 0	0.817 0	0.854 6↑	0.834 9↓	0.852 9↑
Raisin_Dataset	900	7	0.808 9	0.817 8	0.812 2	0.821 1	0.823 3	0.855 6↑	0.824 4↑	0.860 0↑
KnuggetChase3	194	39	0.762 3	0.824 8	0.778 3	0.819 7	0.788 9	0.819 8↓	0.840 2↑	0.840 2↑
sonar	208	60	0.667 9	0.702 1	0.711 5	0.706 4	0.793 0	0.683 2↓	0.774 0↑	0.793 1↑
waveform1_R	5 000	21	0.827 0	0.836 8	0.832 2	0.839 4	0.853 0	0.849 2↑	0.845 8↑	0.855 4↑
waveform0_R	5 000	21	0.798 2	0.818 6	0.803 6	0.812 0	0.826 0	0.826 8↑	0.822 4↑	0.827 6↑
waveform2_R	5 000	21	0.826 4	0.857 2	0.838 6	0.837 4	0.859 4	0.867 4↑	0.857 8↑	0.863 6↑
appendicitis	106	7	0.782 3	0.848 5	0.782 3	0.791 8	0.820 8	0.858 0↑	0.867 5↑	0.867 5↑
australian	690	14	0.805 8	0.852 2	0.820 3	0.813 0	0.843 5	0.865 2↑	0.789 9↓	0.789 9↓
german	1 000	20	0.669 0	0.697 0	0.688 0	0.694 0	0.691 0	0.736 0↑	0.700 0↑	0.742 0↑
bupa	345	6	0.623 2	0.655 1	0.637 7	0.637 7	0.687 0	0.666 7↑	0.649 3↓	0.672 5↑
banana	5 300	2	0.872 5	0.877 4	0.872 8	0.872 8	0.870 8	0.887 5↑	0.873 0↓	0.887 2↑
ecoil2	336	7	0.887 1	0.928 5	0.925 5	0.913 5	0.934 5	0.931 4↑	0.925 5↓	0.952 3↑
ecoil3	336	7	0.904 7	0.907 6	0.937 5	0.940 4	0.934 6	0.916 6↑	0.916 6↑	0.931 5↑
divorce	170	54	0.952 9	0.964 7	0.970 6	0.970 6	0.976 5	0.970 6↑	0.964 7↓	0.970 6↑
Occupancy_estimation	10 129	16	0.998 3	0.998 0	0.998 2	0.998 2	0.999 4	0.995 7↓	0.998 3↑	0.999 0↑

表 3　模型复杂度的对比结果

Table 3　Comparison results of model complexity

Dataset	Gini		BNM+Gini		CSN+Gini		BNM+CSN+Gini	
	depth	leaf_nodes	depth	leaf_nodes	depth	leaf_nodes	depth	leaf_nodes
diabetes	12.0	53.2	10.4√	30.0	10.0√	74.8	10.0√	34.8
yeast_banlance	22.0	128.4	13√	26.2	16.4√	187.2	11.6√	85.6
Diabetic	16.2	98.0	9.8√	39.2	14.6√	139.0	16.6√	106.8
Algerian_Forest_Fires	2.2	3.4	2.2	3.2	3.0	4.2	3.2	4.6
data_banknote_authentication	6.2	15.6	6.0√	20.0	6.4	14.8	6.0√	17.2
adult	41.8	1 216.6	24.8√	269.0	39.8	1 442.8	19.8√	128.0
Raisin_Dataset	10.4	43.6	2.8√	4.4	11.0	58.6	1.0√	2.0
KnuggetChase3	3.8	6.2	3.4√	6.0	5.0	11.6	3.2√	5.6
sonar	5.0	10.0	5.0	11.8	4.6√	12.4	4.4√	15.2
waveform1_R	16.4	151.2	12.8√	74.4	14.2√	153.2	13.0√	132.2
waveform0_R	14.0	167.6	13.4√	99.2	11.2√	209.6	11.2√	109.6
waveform2_R	15.8	145.0	12.6√	107.2	13.8√	156.2	12.4√	110.0
appendicitis	2.4	4.2	1.8√	3.2	2.0√	3.2	2.0√	3.6
australian	10.4	33.2	8.8√	22.8	12.0	55.0	8.2√	18.8
german	13.6	78.2	12.2√	39.2	12.8√	103.0	12.0√	49.4
bupa	9.2	39.0	7.8√	21.0	8.2√	43.2	8.0√	23.2
banana	22.8	243.2	14.2√	102.8	17.0√	355.4	15.4√	126.0
ecoil2	6.2	14.2	4.2√	8.2	5.8√	14.2	4.4√	9.4
ecoil3	6.4	11.6	4.6√	6.8	6.0√	14.8	6.8	15.2
divorce	1.4	2.4	1.4	2.8	1.6	2.8	1.2√	2.4
Occupancy_estimation	4.4	6.8	3.0√	4.6	5.0	8.8	8.4	12.4

Gini降低了树高,CSN+Gini降低了13个数据集的复杂度,这说明考虑全局结构的混合分裂准则能够降低模型复杂度,提高树的可解释性。BNM+Gini、BNM+CSN+Gini分别在8和11个数据集上取得最小树高,CSN+Gini仅在一个数据集上取得最低的模型复杂度,在多数数据集上,BNM能更

好地帮助模型利用数据的结构信息。然而,BNM的可解释性建立在多簇数据的簇间差异性的基础上,所以BNM在单簇数据集中很难起作用,如:在表2、3中,对于sonar数据集,BNM没能降低模型复杂度,甚至损害了模型的准确率。BNM也可以减少树的叶子节点数。在german数据集中,树高仅降低1,叶子节点数减小为原始模型的50%,这说明BNM对叶子节点数的影响,不只是因为树高的降低,也因为BNM诱导出的决策树尊重簇间关系,减小了过拟合的风险。

3.3 消融实验

为了进一步了解簇间度量的有效性,本文进行了消融实验,BNM及CSN是否被用于结合Gini作为每个分裂准则,实验进行五折交叉验证,选取最佳参数并对21个数据集的实验结果求均值,如表4所示。Gini结合BNM或CSN都能够提高准确率,当结合BNM和CSN时,模型同时综合了纯度信息和同类簇间信息及异类簇间信息,获得了最佳性能。此外,Gini结合BNM或CSN都可以降低模型复杂度,其中,Gini结合BNM、Gini结合BNM和CSN的效果显著,这是因为BNM利用了兄弟节点间的特征空间独立性和多簇数据的簇间分布差异,分隔开特征空间中差异较大的同类簇,从而提高了可解释性。

表4 消融实验

Table 4 Ablation experiments

分裂准则		acc	depth	leaf_nodes
BNM	CSN			
×	×	0.839 1	11.6	117.7
√	×	0.848 6	8.3	43.0
×	√	0.843 3	10.5	145.9
√	√	0.856 8	8.5	48.2

3.4 参数分析

为了进一步研究BNM、CSN对分类性能的贡献,本文统计了模型参数的最佳取值,主要关注结合BNM和纯度的加权因子ω_1、ω_2以及结合CSN的候选点个数k。令$\omega_1=1$,图2(a)给出了BNM和BNM+CSN+Gini取最佳测试精度时ω_2的分布直方图,其中,横坐标Weight_value表示最佳参数k的取值($k\in[0.002\ 5\quad 0.10]$),纵坐标Number of weight表示在对比实验中最佳参数ω_2取该值的个数。图2(b)给出了CSN和BNM+CSN+Gini取最佳测试精度时k的分布直方图,其中,横坐标K_value表示最佳参数k的取值($k\in[2\quad 30]$),纵坐标Number of k表示在对比实验中最佳参数k取该值的个数。在BNM+Gini模型中,在0.01到0.03之间的ω_2取值占所有数据集的62%;BNM+CSN+Gini的ω_2取值有81%集中于[0.005, 0.03]。并且如图2(b)所示,在CSN+Gini和BNM+CSN+Gini模型中,取$k=2$的数据集最多。实验结果表明,ω_2和k取值往往比较小,这说明在最佳分裂点的选择中,纯度信息仍然是至关重要的,数据的全局结构信息作为纠正过高纯度的辅助手段,通过选择判别能力强且利于后续划分的分裂点,能够诱导出泛化性和可解释性更好的决策树。

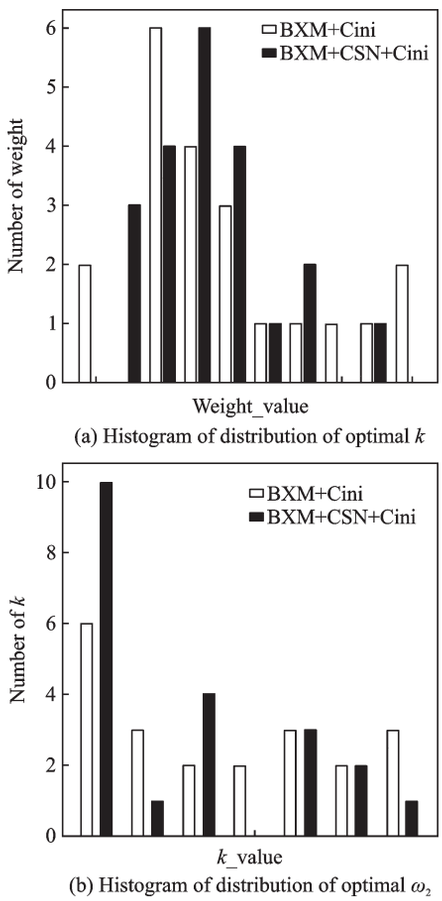


图2 最佳参数的分布直方图

Fig.2 Histogram of distribution of optimal parameters

4 结 论

本文提出了3种纯度结合全局结构的混合分裂准则,从同类间距和类紧性两方面获取簇间关系信息,在满足一定纯度要求的候选分裂点中选择最佳分裂点。本文将其应用于模拟的二维异或问题,和CART相比,该树的划分更接近人类的思维方式,可解释性更强。此外,为了检验其有效性,本文在21个标准验证数据集上与其他混合分裂准则进行了对比实验,实验结果表明,该树具有更好的泛化性和更低的模型复杂度。然而,由于距离度量的局限性,本文的方法只能计算连续数据的BNM和CSN,对于离散数据可用独热(one-hot)编码转化为连续数据,但在决策树模型中,高维离散数据的

one-hot编码易产生切分不平衡现象。BNM是针对多簇数据集设计的度量，在单簇数据中往往不能发挥很好的作用。此外，考虑到特征相关性和特征重要性也是决策树生长的重要部分，为了进一步研究这一课题，未来将结合目前的工作，将其应用到决策树的分裂生长中。

参考文献：

[1] 周志华.机器学习[M].北京：清华大学出版社，2016.
ZHOU Zhihua. Machine learning [M]. Beijing: Tsinghua University Press, 2016.

[2] NOR A K M, PEDAPATI S R, MUHAMMAD M. Reliability engineering applications in electronic, software, nuclear and aerospace industries: A 20 year review (2000—2020)[J]. Ain Shams Engineering Journal, 2021, 12(3): 3009-3019.

[3] MIN Z, KE W, YANG W, et al. Online condition diagnosis for a two-stage gearbox machinery of an aerospace utilization system using an ensemble multi-fault features indexing approach[J]. Chinese Journal of Aeronautics, 2019, 32(5): 1100-1110.

[4] SHI J, HE Q, WANG Z. GMM clustering-based decision trees considering fault rate and cluster validity for analog circuit fault diagnosis[J]. IEEE Access, 2019, 7: 140637-140650.

[5] BERTONI A, HALLSTEDT S I, DASARI S K, et al. Integration of value and sustainability assessment in design space exploration by machine learning: An aerospace application [J]. Design Science, 2020. DOI: https://doi.org/10.1017/dsj.2019.29.

[6] MYLES A J, FEUDALE R N, LIU Y, et al. An introduction to decision tree modeling[J]. Journal of Chemometrics: A Journal of the Chemometrics Society, 2004, 18(6): 275-285.

[7] WANG R, KWONG S, WANG X Z, et al. Segment based decision tree induction with continuous valued attributes[J]. IEEE Transactions on Cybernetics, 2014, 45(7): 1262-1275.

[8] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[9] ROSS QUINLAN J. C4.5: Programs for machine learning[J]. Mach Learn, 1993, 16(3): 235-240.

[10] BREIMAN L, FRIEDMAN J H, OLSHEN R A, et al. Classification and regression trees (CART)[M]. Belmont, CA, USA: Wadsworth International Group, 1984: 358-361.

[11] TIMOFEEV R. Classification and regression trees (CART) theory and applications[D]. Berlin: Humboldt University, 2004.

[12] BREIMAN L, FRIEDMAN J H, OLSHEN R A, et al. Classification and regression trees[M]. New York: Routledge Press, 2017.

[13] YAN J, ZHANG Z, XIE L, et al. A unified framework for decision tree on continuous attributes [J]. IEEE Access, 2019, 7: 11924-11933.

[14] YAN J, ZHANG Z, LIN K, et al. A hybrid scheme-based one-vs-all decision trees for multi-class classification tasks [J]. Knowledge-Based Systems, 2020, 198: 105922.

[15] LI D, CHEN S. A novel splitting criterion inspired by geometric mean metric learning for decision tree[C]// Proceedings of 2022 the 26th International Conference on Pattern Recognition (ICPR).[S.l.]: IEEE, 2022: 4808-4814.

[16] OW P S, MORTON T E. Filtered beam search in scheduling[J]. The International Journal of Production Research, 1988, 26(1): 35-62.

[17] MEISTER C, VIEIRA T, COTTERELL R. If beam search is the answer, what was the question? [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).[S.l.]: Association for Computational Linguistics (ACL), 2020.

[18] JAWORSKI M, DUDA P, RUTKOWSKI L. New splitting criteria for decision trees in stationary data streams[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 29(6): 2516-2529.

[19] YAN J, ZHANG Z, DONG H. AdaDT: An adaptive decision tree for addressing local class imbalance based on multiple split criteria [J]. Applied Intelligence, 2021, 51(7): 4744-4761.

[20] PEREIRA BARATA A, TAKES F W, JAAP VAN DEN HERIK H, et al. Fair tree classifier using strong demographic parity [EB/OL]. (2021-10-10). http://10.48550/arxiv.2110.09295.

[21] RAHMAN M G, ISLAM M Z. Adaptive decision forest: An incremental machine learning framework[J]. Pattern Recognition, 2022, 122: 108345.

[22] MURTAGH F, CONTRERAS P. Algorithms for hierarchical clustering: An overview[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012, 2(1): 86-97.

[23] JOHNSON S C. Hierarchical clustering schemes[J]. Psychometrika, 1967, 32(3): 241-254.

[24] DANIELSSON P E. Euclidean distance mapping[J]. Computer Graphics and Image Processing, 1980, 14(3): 227-248.

[25] BALAKRISHNAMA S, GANAPATHIRAJU A. Linear discriminant analysis—A brief tutorial[J]. Institute for Signal and Information Processing, 1998, 18: 1-8.

[26] DUA D, GRAFF C. UCI machine learning repository [EB/OL]. (2017-05-10) [2022-04-08]. http://archive.ics.uci.edu/ml.

[27] ALCALÁ-FDEZ J, SANCHEZ L, GARCIA S, et al. KEEL: A software tool to assess evolutionary algorithms for data mining problems[J]. Soft Computing, 2009, 13(3): 307-318.

（编辑：刘彦东）