基于树集成模型规则抽取的机器学习可解释性研究

雍 华[1]
YONG Hua

摘 要 随着机器学习模型在金融、医疗等高敏感领域的广泛应用，模型可解释性成为确保决策透明性和用户信任的关键。树集成模型因其高精度与内在可解释性潜力，成为平衡性能与可解释性的重要工具。文章系统综述了从树集成模型中抽取规则以增强模型可解释性的研究进展，重点介绍了树集成模型的基本情况和从树集成模型中抽取规则以增强机器学习模型可解释性的各类技术，旨在为可解释人工智能的研究提供理论参考，推动规则驱动决策在高风险领域的可靠落地。

关键词 树集成模型；规则抽取；可解释性

doi：10.3969/j.issn.1672-9528.2025.06.013

0 引言

在机器学习的研究和应用中，学术界和工业界的研究人员越来越关注模型的透明性和可解释性，特别是在受到实质性监管和法律约束的领域，如医疗诊断、金融风控、自动驾驶等高敏感领域，需要更深入地理解模型输出结论背后的决策机制，以便评估模型所做选择的合理性。作为平衡模型性能与可解释性的重要工具，树集成模型（如随机森林、XGBoost、LightGBM、CatBoost）凭借其高预测精度和内在可解释性潜力脱颖而出。然而，树集成模型的多树叠加结构仍可能导致全局解释的复杂性，而规则抽取技术（Rule Extraction）通过将模型转化为人类可理解的逻辑规则（如“if-then”形式），为破解这一难题提供了关键路径。近年来，规则抽取技术的研究逐步深入，其核心目标是从树集成模型中提取简洁、高覆盖率的规则集，以透明化模型决策逻辑。例如，在金融领域，规则抽取可将风险评估模型转化为“资产负债率 >70% →高风险”的直观规则，帮助金融机构验证模型合规性；在医疗领域，规则集可揭示疾病预测的关键特征组合，辅助医生制定诊疗方案。

本文聚焦于树集成模型的规则抽取技术，旨在系统梳理其研究进展、技术方法与评价方法，并探讨其与可解释性需求的深度关联。具体而言，本文将从以下维度展开：

1. 宁夏大学前沿交叉学院 宁夏中卫 755000

[13] 樊礼谦，焦文海，孟轶男．基于 Kalman 滤波的导航星座集中式守时算法研究 [J]. 时间频率学报，2023,46(1):21-31.
[14] CHANG G B. Robust kalman filtering based on mahalanobis distance as outlier judging criterion[J]. Journal of geodesy, 2014,88:391-401.
[15] LI W, LIN X, LI S D, et al. 2022. Robust autocovariance least-squares noise covariance estimation algorithm[J]. Measurement, 2022,187: 110331.
[16] IRSHAD M , VEMULA N K , DEVARAPALLI R ,et al. An optimized integral performance criterion based commercial PID controller design for boost converter[J].Journal of electrical engineering, 2024, 75(4):258-267.
[17] YANGA C , BANB L .Vibration control of piezoelectric flexible manipulator based on machine vision and improved PID[C/OL]// 2021 IEEE International Conference on Power Electronics, Computer Applications(ICPECA).Piscataway:IEEE,2021[2024-05-25].https://ieeexplore.ieee.org/document/9362620.DOI: 10.1109/ICPECA51329.2021.9362620.

【作者简介】

程林（1997—），男，四川成都人，硕士，工程师，研究方向：GNSS 时间同步和自适应卡尔曼滤波研究，email:chenglin971226@163.com。

郑鑫（1990—），通信作者（email:18123315728@163.com），男，黑龙江齐齐哈尔人，硕士，工程师，研究方向：信号处理研究。

（收稿日期：2025-02-18 修回日期：2025-06-11）

（1）剖析树集成模型的可解释性基础，明确规则抽取的必要性。

（2）对主流的规则抽取方法（如 RuleFit、SKOPE Rules）进行较为系统的综述。

（3）从预测准确性和可解释性方面梳理针对规则抽取技术的评价方法；通过整合理论方法与实际应用，本文期望为可解释机器学习的研究提供系统性参考，推动规则驱动决策在高风险场景中的可靠落地。

1 树集成模型

决策树作为一种经典的机器学习模型，因其直观的树形结构和易于理解的决策逻辑，长期被视为可解释性机器学习的标杆 [1]。决策树通过递归分割特征空间生成树形结构，每个内部节点表示特征测试，叶节点表示测试结果，用户可沿树的分支路径追溯决策逻辑。决策树通过模拟人类决策过程，以“if-then”规则的形式提供透明化的推理路径，成为平衡模型性能与可解释性的重要工具。另外，通过计算特征在树节点分裂中的贡献度（如信息增益、基尼不纯度减少量），可以量化特征对预测的影响。

然而，决策树在实际应用时，虽然可解释性较强，但由于过拟合和不一致的预测会导致对新样本的预测性能不佳。部分研究表明，将多个单一分类器组合成一个集成分类器，形成集成学习模型，可以获得比任何单一分类器更好的分类性能 [2]，成为在偏差和方差之间进行权衡的最实用方法之一 [3-4]。在集成学习中，Bagging、Boosting 和 Stacking[5] 是 3 种基本技术。Bagging 通过自助采样生成多个训练子集，并行训练基学习器（如决策树），并采用投票或平均法集成结果，核心思想是降低方差，适用于高方差模型（如随机森林）。Boosting 则采用串行训练，后续模型专注于修正前序模型的预测误差，通过加权投票整合结果（如 AdaBoost、梯度提升树），核心是降低偏差，但可能因过度拟合噪声而增加方差。Stacking 通过训练元模型整合多个异构基学习器（如 SVM、神经网络）的预测结果，通常分两阶段：基模型生成预测特征，元模型基于这些特征进行最终决策，虽计算成本较高，但能融合不同模型的优势。三者均通过多样性提升泛化能力，但 Bagging 侧重并行降方差，Boosting 强调串行纠偏，Stacking 则聚焦异质模型的多层次融合。在研究和实际应用中，将多棵决策树集成到一起可以解决决策树过拟合等问题，产生了树集成模型，如随机森林 [6] 或梯度提升树 [7]，这显著提高了分类和回归任务的性能和鲁棒性，同时这类模型在分类、回归和排序任务中表现出色，应用广泛。

表 1 典型的树集成模型规则抽取方法对比

方法	核心贡献	局限性	适用场景
RuleFit	首开 LASSO 稀疏规则选择先河	深层树下规则冗余度高	低维数据分类 / 回归
NodeHarvest	二次规划优化规则权重，实现性能 - 可解释性均衡	仅支持二分类	二分类回归任务
SIRUS	经验分位数约束提升规则鲁棒性	多类分类不可用	鲁棒性要求高的场景
FIRE	树状规则结构增强可解释性	分类扩展未实现，规则数量不可控	回归任务优先
Skope-rules	精度 - 召回率约束提升规则质量	计算效率低	高召回需求场景
DefragTrees	贝叶斯优化规则区域形状与数量	高复杂度规则下失效	低复杂度规则集生成
inTrees	统计驱动剪枝，支持分类与回归	回归需离散化，引入参数敏感性	通用分类任务

2 树集成模型规则抽取

从经过训练的树集合中抽取规则可以避免探索不切实际的大型搜索空间，并且被认为是解释随机森林等树集成模型的最先进事后方法 [8]。规则抽取技术旨在从树集成模型中提取人类可理解的规则，从而增强模型的可解释性。该方法依靠生成基于规则集的预测器来获得树集成模型的预测能力 [9]。简言之，规则集（或规则列表）在可解释性方面可以匹敌决策树 [10]，因为都由易于理解的“if-then”语句形式的逻辑模型组成 [11]。决策树和规则集之间的一个有趣联系涉及从以前训练的树集成模型中抽取规则。这样，用于选择规则的搜索空间将限制为树集成模型中的分支节点集，这些节点表示用于拆分训练数据和进行预测的逻辑条件。根据技术实现方式，现有方法可分为基于规则生成的直接抽取法、基于后处理的规则优化方法、基于模型重构的间接抽取法以及基于规则度量的剪枝与选择方法。

基于规则生成的直接抽取法通过遍历决策树的路径生成逻辑规则（IF-THEN 形式），并结合集成模型的特点对规则进行筛选与优化。例如，Friedman 和 Popescu 提出的 RuleFit 算法从树集成模型的叶子节点生成规则，并将规则转化为二进制特征，通过 LASSO 回归筛选关键规则，最终生成一个透明的线性模型 [12]。然而，RuleFit 在处理深层树或高度相关特征时表现不佳 [13]。Meinshausen 提出的 NodeHarvest 算法则通过二次规划优化规则权重，旨在生成稀疏且可解释的规则集 [14]。尽管其在回归任务中表现良好，但对多类分类问题的支持有限。

基于后处理的规则优化方法通过对原始规则集进行剪枝、合并与优化，进一步提升规则的可解释性与实用性。Bénard 等人 [15-16] 提出的 SIRUS 算法通过限制随机森林的分割节点为特征的经验分位数，生成简化的规则集。SIRUS 在分类和回归任务中均表现出良好的预测性能与鲁棒性，但其分类变体无法处理多类分类问题。Liu 和 Mazumder 提出的 FIRE 算法通过引入非平滑融合惩罚，生成具有树状结构的规则集。FIRE 在回归任务中表现优异，但其对分类任务的扩展尚未得到充分研究。

基于模型重构的间接抽取法通过构建代理模型或知识蒸馏技术，将树集成模型的决策逻辑映射为显式规则。例如，BATrees 通过生成更多标记数据点来拟合决策树，但其生成的树可能过于复杂，降低了可解释性 [17]。Skope-rules 从装袋集成模型中提取规则，并通过计算规则相似性简化规则集，同时引入精度和召回率约束以提高预测性能 [18]。DefragTrees 利用贝叶斯模型选择方法优化规则的数量、形状和位置，生成简化的“等效”模型 [19]。然而，随着规则复杂性的增加，其简化效果可能受限。

基于规则度量的剪枝与选择方法通过度量与剪枝规则，消除冗余或不必要的规则，从而降低模型复杂度。例如，inTrees 算法通过统计检验筛选关键规则，最终生成简化的树集成学习器 [20]。尽管 inTrees 在分类任务中表现良好，但其在回归任务中的应用需通过离散化实现，可能引入额外复杂性。TE2Rules 算法通过关联规则挖掘方法提取特征交互规则，但其适用范围有限，仅适用于 GBDT 模型 [21]。

为全面评估现有规则抽取方法的性能，表 1 对比了典型方法的优势与局限性。RuleFit 支持稀疏规则选择并兼容线性模型，但其规则冗余度高，可解释性受限。NodeHarvest 计算效率高且生成稀疏规则集，但仅支持二分类问题。SIRUS 在预测性能和对数据扰动的鲁棒性方面表现优异，但无法处理多类分类问题。FIRE 生成树状结构规则，显著提升了可解释性，但其对分类任务的扩展有限。inTrees 支持规则剪枝并降低模型复杂度，但在回归任务中需通过离散化处理。Skope-rules 引入精度与召回率约束以简化规则集，但其规则生成效率较低。DefragTrees 利用贝叶斯优化规则数量与形状，但随着规则复杂性增加，其简化效果可能受限。

通过对以上文献和技术方法进行总结，规则抽取的通用框架通常包括 4 个步骤：规则生成、规则度量、规则修剪和规则集构建。在规则生成阶段，一个集成的树模型通常由多个决策树构成，每棵树的根节点到叶子节点的每一条路径都表示一条决策规则，通过遍历树集成模型的决策路径提取逻辑规则。例如，RuleFit 从叶子节点生成规则，而 SIRUS 通过限制分割节点生成简化规则。在规则度量阶段，基于规则长度、规则频率、误差、支持度、置信度或信息增益等指标评估规则的重要性。例如，inTrees 通过统计检验筛选关键规则。在规则修剪阶段，通过合并相似规则或剔除低贡献规则降低复杂度，并选择一组相关的非冗余规则。例如，Skope-rules 通过计算规则相似性简化规则集。在规则集构建阶段，结合优化算法（如二次规划、LASSO 回归）选择最优规则子集，构建一个可解释的规则学习器，确保分类性能与可解释性均衡，用于决策和解释。例如，NodeHarvest 通过二次规划优化规则权重。

3 树集成模型规则抽取的评价

在树集成模型规则抽取领域，大部分研究工作都集中在探索规则抽取技术和方法，以提高预测任务的可解释性，并尽量降低预测的准确性。但是在实际应用中，对给定数据集和给定应用场景的机器学习任务采用哪种规则抽取方法最合适仍然不清楚，需要进行评估来量化各种规则抽取方法的质量，以确定所提供的可解释性是否以及在多大程度上实现了既定的目标，并比较可用的规则抽取方法，为特定任务提出最佳的解释 [22]。但是，对于规则抽取方法的评估还没有形成统一的体系和评估指标，背后的原因是可解释性本质上是一个主观概念，解释的质量取决于解释场景、解释方法和用户需求，很难将其标准化 [23]。

通常情况下，对于树集成模型规则抽取的评估可分为主观评估和客观评估两类。

主观评估指标方面，主要关注人类对所抽取规则集的主观认识程度，主要对规则的理解程度、信任程度、满意程度以及用户体验等，其中大部分可以通过李克特量表来进行衡量，尤其是在评估信任程度时 Jian 等人 [24] 提出的自动化信任度评估问卷被广泛使用 [25]，然而，Cui 等人 [26] 研究发现主观评估存在很多的误导，即用户使用主观评价好的机器学习系统时不一定会有好的表现和性能，所以对可解释人工智能主观评估的结果可能无法预测使用其进行实际决策任务的效果。目前针对主观评估研究的正式成果还比较少，可以看出可解释人工智能领域的研究需要心理学、人机交互等领域的研究者共同参与 [27]。

客观评估指标方面，现有的研究主要从规则集的可解释性和规则集的解释保真度两个方面来评估从树集成模型中抽取出的规则集的质量。

Lakkaraju 等人 [28] 为量化规则集的可解释性，提出了 4 个核心评价指标：规则集规模（Size）、规则长度（Length）、覆盖率（Coverage）与规则重叠度（Overlap）。这些指标通过结构化度量框架，系统性地刻画了规则集的解释性本质。规则集规模通过限制规则数量以降低认知负荷，规则集包含

的规则越少，用户越能快速理解特定类别标签对应的决策边界。规则长度指单条规则的前件数量，直接影响其可读性，尽管逻辑表达式天然具备可解释性[29]，但过长的规则（如包含冗余条件）会显著削弱人类理解效率。覆盖率衡量规则集对数据空间的描述完整性，高覆盖率确保规则集能够解释大部分样本的预测逻辑，避免关键决策场景的遗漏。规则重叠度评估不同规则在特征空间中的独立性，高度重叠的规则会迫使模型依赖“打破平局”机制（tie-breaking）进行预测，这种隐式决策逻辑将严重损害可解释性。因此，最小化规则重叠有助于构建清晰、互斥的决策边界。该框架表明，可解释性不仅依赖于规则集的简洁性（规模与长度），更需通过高覆盖率和低重叠度实现决策逻辑的透明性与一致性。这一理论为后续可解释规则生成算法的设计提供了重要指导原则。

在以往的研究中，黑盒模型与从中抽取的规则集（代理模型）之间的差异通常被称为解释保真度或完整性，规则集的预测结果与黑盒模型的预测结果非常吻合，那么这个模型就被认为是高度忠实的。Johansson 等人[30]提出，规则抽取方法应该用来完成两个不同的解释任务。第一项任务是理解黑盒模型所做单个数据点预测背后的基本推理思路，第二项任务是提取可解释的模型，作为更容易理解的预测器。针对前一项任务的方法以保真度为基础进行评估，而针对后一项任务的方法则以样本外准确度为主要标准。关于从树集合中提取规则方法的相关工作属于第二类。事实上，很多研究的结果是根据预测性能进行比较的，而忽略了保真度的测量。事实上，可解释性和保真度都是获得有价值的解释所必需的[31]。然而，它们可能会发生冲突，因为开发准确的代理模型通常需要更复杂的模型，这会降低其可解释性。因此，可解释性和保真度之间的权衡开始发挥作用，解释方法不应根据这种权衡中的单个点来评估[32]。正如 Messalas 等人[33]首次讨论的那样，可以确定两种类型的保真度。第一种被称为外部保真度，是指替代模型的预测与不透明模型的预测的一致程度。外部保真度的常见度量包括计算两个模型之间的不一致，通常表示为无法解释的方差的分数或通过均方误差表示。虽然外部保真度评估了代理模型和不透明模型的预测之间的一致性，但并不能保证代理模型的决策过程忠实于原始过程。这方面由内部保真度捕获，而内部保真度显然很难评估[34]。评估内部保真度的与模型无关的方法涉及测量代理模型和不透明模型提供的特征重要性排名之间的一致性。例如，这可以通过根据两个模型确定的特征重要性对数据集的特征进行排序，然后计算两个排序之间的等级相关性（例如，Kendall 的 τ）来完成[35]。目前，缺乏特定于模型的方法来评估应用于树集成的规则提取技术的内部保真度。

4 结语

树集成模型规则抽取技术作为连接机器学习模型与人类可解释性需求的重要桥梁，近年来在方法创新与应用拓展方面取得了显著进展。尽管现有方法在金融风控、医疗诊断等领域展现了良好的应用潜力，但仍面临准确性 - 可解释性权衡、规则泛化能力不足以及动态适应性有限等挑战。未来，随着可解释人工智能研究的深入，规则抽取技术有望在提升树集成模型透明度的同时，进一步推动其在高风险领域的应用。

参考文献：

[1]SILVA A P D. Optimization approaches to supervised classification[J]. European journal of operational research, 2017, 261(2): 772-788.

[2] BAUER E, KOHAVI R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants[J]. Machine learning, 1999, 36: 105-139.

[3] BREIMAN L. Bagging predictors[J]. Machine learning, 1996, 24: 123-140.

[4] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of computer and system sciences, 1997, 55(1): 119-139.

[5] SCHWENKER F. Ensemble methods: foundations and algorithms [book review][J]. IEEE computational intelligence magazine, 2013, 8(1): 77-79.

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

[7] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.

[8] ARIA M, CUCCURULLO C, GNASSO A. A comparison among interpretative proposals for random forests[J]. Machine learning with applications, 2021, 6: 100094.

[9] DI TEODORO G, MONACI M, PALAGI L. Unboxing tree ensembles for interpretability: a hierarchical visualization tool and a multivariate optimal re-built tree[J]. EURO journal on computational optimization, 2024, 12: 100084.

[10] YANG H Y, RUDIN C, SELTZER M. Scalable bayesian rule lists[EB/OL].(2024-04-03)[2024-05-22].https://doi.org/10.48550/arXiv.1602.08610.

[11] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models

instead[J]. Nature machine intelligence, 2019, 1(5): 206-215.

[12] FRIEDMAN J H, POPESCU B E. Predictive learning via rule ensembles[EB/OL].(2008-11-11)[2024-12-13].https://doi.org/10.48550/arXiv.0811.1679.

[13] LIU B, MAZUMDER R. FIRE: an optimization approach for fast interpretable rule extraction[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.NewYork:ACM, 2023: 1396-1405.

[14] MEINSHAUSEN N. Node harvest[J]. The annals of applied statistics, 2010: 2049-2072.

[15] BENARD C, BIAU G, VEIGA S D, et al. SIRUS: stable and interpretable rule set for classification[EB/OL]. (2020-12-16)[2024-06-22].https://doi.org/10.48550/arXiv.1908.068522021.

[16] BENARD C, BIAU G, VEIGA S D, et al. Interpretable rand om forests via rule extraction[EB/OL].(2021-02-10)[2024-05-19].https://doi.org/10.48550/arXiv.2004.14841.

[17] VIDAL T, SCHIFFER M, PACHECO T, et al. Born-again tree ensembles[C]//International conference on machine learning. NewYork: ACM, 2020: 9743-9753.

[18] SENDI N, ABCHICHE-MIMOUNI N, ZEHRAOUI F. A new transparent ensemble method based on deep learning[J]. Procedia computer science, 2019, 159: 271-280.

[19] BOLOGNA G, HAYASHI Y. A comparison study on rule extraction from neural network ensembles, boosted shallow trees, and SVMs[J]. Applied computational intelligence and soft computing, 2018, 2018(1): 4084850.

[20] DENG H T. Interpreting tree ensembles with inTrees[J]. International journal of data science and analytics, 2019, 7(4): 277-287.

[21] LAL G R, CHEN X T, MITHAL V. TE2Rules: extracting rule lists from tree ensembles[EB/OL].(2024-01-23)[2024-05-26].https://doi.org/10.48550/arXiv.2206.14359.

[22] ZHOU J L, GANDOMI A H, CHEN F, et al. Evaluating the quality of machine learning explanations: a survey on methods and metrics[J]. Electronics, 2021, 10(5): 593.

[23] CARVALHO D V, PEREIRA E M, CARDOSO J S. Machine learning interpretability: a survey on methods and metrics[J]. Electronics, 2019, 8(8): 832.

[24] JIAN J Y, BISANTZ A M, DRURY C G. Foundations for an empirically determined scale of trust in automated systems[J]. International journal of cognitive ergonomics, 2000, 4(1): 53-71.

[25] PAYROVNAZIRI S N, CHEN Z Y, RENGIFO-MORENO P, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review[J]. Journal of the american medical informatics association, 2020, 27(7): 1173-1185.

[26] CUI X C, LEE J M, HSIEH J P A. An integrative 3C evaluation framework for explainable artificial intelligence[C]//AMCIS 2019 Proceeding. NewYork: ACM, 2019:15-17.

[27] 孔祥维 , 唐鑫泽 , 王子明 . 人工智能决策可解释性的研究综述 [J]. 系统工程理论与实践 ,2021,41(2):524-536.

[28] LAKKARAJU H, BACH S H, LESKOVEC J. Interpretable decision sets: a joint framework for description and prediction[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. NewYork: ACM, 2016:1675-1684.

[29] CRUZ N, BARATGIN J, OAKSFORD M, et al. Bayesian reasoning with ifs and ands and ors[J]. Frontiers in psychology, 2015, 6: 192.

[30] JOHANSSON U, SONSTROD C, LOFSTROM T, et al. Rule extraction with guarantees from regression models[J]. Pattern recognition, 2022, 126: 108554.

[31] GILPIN L H, BAU D, YUAN B Z, et al. Explaining explanations: an overview of interpretability of machine learning[C]//2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). Piscataway: IEEE, 2018: 80-89.

[32] ZHOU J L, GANDOMI A H, CHEN F, et al. Evaluating the quality of machine learning explanations: a survey on methods and metrics[J]. Electronics, 2021, 10(5): 593.

[33] MESSALAS A, KANELLOPOULOS Y, MAKRIS C. Model-agnostic interpretability with shapley values[C]//2019 10th international conference on information, intelligence, systems and applications (IISA). Piscataway:IEEE,2019: 1-7.

[34] ZHOU Z H. Rule extraction: using neural networks or for neural networks?[J].Journal of computer science and technology, 2004, 19(2): 249-253.

[35] VELMURUGAN M, OU-YANG C, SINDHGATTA R, et al. Through the looking glass: evaluating post hoc explanations using transparent models[J]. International journal of data science and analytics, 2023: 1-21.

【作者简介】

雍华（1990—），男，宁夏中卫人，硕士，助教，研究方向：金融科技、可解释机器学习、神经网络。

（收稿日期：2025-02-22 修回日期：2025-06-11）