软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(7):2447−2463 [doi: 10.13328/j.cnki.jos.006586]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于最小不满足核的随机森林局部解释性分析*

马舒岑[1,2], 史建琦[1,2], 黄滟鸿[1,2], 秦胜潮[3], 侯 哲[4]

[1](国家可信嵌入式软件工程技术研究中心(华东师范大学), 上海 200062)

[2](华东师范大学 软件工程学院, 上海 200062)

[3](深圳大学 计算机与软件学院, 广东 深圳 518060)

[4](School of Information and Communication Technology, Griffith University, Brisbane 4111, Australia)

通信作者: 史建琦, E-mail: jqshi@sei.ecnu.edu.cn; 黄滟鸿, E-mail: yhhuang@sei.ecnu.edu.cn

摘 要: 随着机器学习在安全关键领域的应用愈加广泛, 对于机器学习可解释性的要求也愈加提高. 可解释性旨在帮助人们理解模型内部的运作原理以及决策依据, 增加模型的可信度. 然而, 对于随机森林等机器学习模型的可解释性相关研究尚处于起步阶段. 鉴于形式化方法严谨规范的特性以及近年来在机器学习领域的广泛应用, 提出一种基于形式化和逻辑推理方法的机器学习可解释性方法, 用于解释随机森林的预测结果. 即将随机森林模型的决策过程编码为一阶逻辑公式, 并以最小不满足核为核心, 提供了关于特征重要性的局部解释以及反事实样本生成方法. 多个公开数据集的实验结果显示, 所提出的特征重要性度量方法具有较高的质量, 所提出的反事实样本生成算法优于现有的先进算法; 此外, 从用户友好的角度出发, 可根据基于反事实样本分析结果生成用户报告, 在实际应用中, 能够为用户改善自身情况提供建议.

关键词: 机器学习可解释性; 特征重要性; 反事实样本; 形式化方法; 逻辑推理

中图法分类号: TP311

中文引用格式: 马舒岑, 史建琦, 黄滟鸿, 秦胜潮, 侯哲. 基于最小不满足核的随机森林局部解释性分析. 软件学报, 2022, 33(7): 2447–2463. http://www.jos.org.cn/1000-9825/6586.htm

英文引用格式: Ma SC, Shi JQ, Huang YH, Qin SC, Hou Z. Minimal-unsatisfiable-core-driven Local Explainability Analysis for Random Forest. Ruan Jian Xue Bao/Journal of Software, 2022, 33(7): 2447−2463 (in Chinese). http://www.jos.org.cn/1000-9825/6586.htm

Minimal-unsatisfiable-core-driven Local Explainability Analysis for Random Forest

MA Shu-Cen[1,2], SHI Jian-Qi[1,2], HUANG Yan-Hong[1,2], QIN Sheng-Chao[3], HOU Zhe[4]

[1](National Trusted Embedded Software Engineering Technology Research Center (East China Normal University), Shanghai 200062, China)

[2](Software Engineering Institute, East China Normal University, Shanghai 200062, China)

[3](College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China)

[4](School of Information and Communication Technology, Griffith University, Brisbane 4111, Australia)

Abstract: With the broader adoption of machine learning (ML) in security-critical fields, the requirements for the explainability of ML are also increasing. The explainability aims at helping people understand models' internal working principles and decision basis, which adds their realibility. However, the research on understanding ML models, such as random forest (RF), is still in the infant stage. Considering the strict and standardized characteristics of formal methods and their wide application in the field of ML in recent years, this work leverages formal methods and logical reasoning to develop a machine learning interpretability method for explaining the prediction of RF. Specifically, the decision-making process of RF is encoded into first-order logic formula, and the proposed approach is centered

* 基金项目: 国家重点研发计划(2019YFB2102602)

本文由"智能系统的分析和验证"专题特约编辑明仲教授、张立军教授和秦胜潮教授推荐.

收稿时间: 2021-09-05; 修改时间: 2021-10-14; 采用时间: 2022-01-10; jos 在线出版时间: 2022-01-28

around minimal unsatisfiable cores (MUC) and local interpretation of feature importance and counterfactual sample generation method are provided. Experimental results on several public datasets illustrate the high quality of the proposed feature importance measurement, and the counterfactual sample generation method outperforms the state-of-the-art method. Moreover, from the perspective of user friendliness, the user report can be generated according to the analysis results of counterfactual samples, which can provide suggestions for users to improve their own situation in real-life applications.

Key words: explainable machine learning; feature importance; counterfactual sample; formal method; logical reasoning

如今，机器学习的应用已十分普遍，它被广泛应用于如自动驾驶[1]、医疗[2]、智慧政府[3]等安全敏感性较高的领域. 然而，机器学习模型仍被看作是一个黑盒模型，理解模型内部的运作原理十分困难，从而阻碍了它的长足发展[4]. 由此，机器学习的可解释性便尤为重要. 下面通过一个例子来反映机器学习可解释性的重要之处：机器学习常被应用于医学诊疗，机器学习模型接收人的各项体征作为输入进行预测，预测值可作为疾病诊断的参考之一，而医生需要相信该模型，并提前了解该模型的运作机制，才能够得心应手地使用模型；此外，当患者被确诊为某种疾病时，使用可解释性方法进行分析，能够得到模型中对疾病诊断起决定性作用的重要特征，例如血压特征之于心脏病、血糖特征之于糖尿病；进一步地，使用可解释性方法分析如何通过改善这些重要特征值使得病情有所好转，分析结果可为制定治疗方案提供具有价值的参考. 由此可见，模型的预测结果能够回答“是什么”的原始问题，而对模型的解释能够回答“为什么”的问题，从而帮助使用者理解模型的决策过程，提高模型的可信度；此外，从“为什么”的视角出发能够解决更多相关的问题，例如模型修正、样本结果矫正等.

在此，本文聚焦于机器学习的可解释性问题[5,6]. 如上文所述，其中一种常用的解释方法为探究在模型预测时起重要作用的特征，即特征重要性(feature importance)[7]分析，它涵盖了两方面：局部解释和全局解释. 局部解释用于分析单个样本的重要特征，而全局解释用于分析模型的重要特征. 然而，大多数现有的分析特征重要性的工具，例如 Anchor[8]和 LIME[9]，运算时将复杂的模型转化为近似的简单模型，并利用统计方法分析特征对预测的影响；近似模型和统计方法意味着这些工具仍将机器学习模型视为黑盒，并未真正深入模型内部去探究模型运作的内核，了解模型内部的逻辑.

反事实分析(counterfactual analysis)作为另一种洞察模型内部行为的方法，旨在研究如何通过改变样本值使得预测结果改变为理想值. 在此，本文聚焦于反事实样本的生成，以此反映机器学习模型的可靠性. 模型预测的不理想样本在实际中往往意味着事故的发生，如贷款失败、交通事故以及上文中提及的疾病确诊等，通过生成反事实样本，可以从当前事故中分析成因，吸取经验，从而为逆境转势提供具有参考性的建议.

集成树(ensemble tree)模型，如随机森林(random forest)[10]以及提升方法(boosting)[11]，将决策树(decision tree)[12]加以组合，以提高计算和泛化能力，作为近年来新兴起的、高度灵活的机器学习算法，在机器学习领域具有不凡的表现，尤其是在处理电子表格、大型数据库等结构化数据[13]方面具有广泛的应用前景. 然而，现有的可解释性方法大都聚焦于神经网络[14−17]，针对集成树模型的可解释性方法仍尚未成熟. 这是因为当前大多数的反事实分析方法以梯度计算为核心，而树属于离散结构，并不具备梯度这一特性，从而限制了反事实分析在集成树模型方面的应用.

决策树的语义，尤其是以二元决策图形式出现的变体，已在逻辑和形式化方法领域得到了充分理解[18−20]，因此，形式化方法尤其适用于树模型的分析；同时，由严谨的逻辑推理生成的模型解释能够极大弥补近似和统计方法的不足. 基于上述考虑以及当前机器学习可解释性的发展情况，针对随机森林模型，本文提出了基于形式化方法的可解释方法.

本文的主要贡献如下：

(1) 局部解释方面，使用谓词逻辑将随机森林对于单个样本的决策过程编码为一阶逻辑公式，并利用 SMT 求解器分析，利用 SMT 求解器产生的最小不满足核提供局部诱因解释；

(2) 反事实分析方面，利用最小不满足核改进现有的反事实样本生成算法. 与现有算法相比，相同时间内，利用改进算法能够生成距离原样本更近的反事实样本；

(3) 根据反事实分析结果为用户设计了一份易于实现的用户报告, 旨在为如何通过改变特征值将不理想样本预测转为理想结果提供建议;

(4) 通过公开数据集的实验结果以及一个案例分析验证了本文的贡献.

本文第 1 节介绍相关工作. 第 2 节介绍背景知识. 第 3 节、第 4 节详细介绍基于形式化方法的特征重要性和反事实分析方法. 第 5 节展示实验结果以及案例分析结果. 第 6 节为本文的总结与展望. 本文聚焦随机森林二分类器的可解释性方法, 该方法可推广至多分类器. 方法的工作流程如图 1 所示, 共分为 3 部分: 随机森林分类器的逻辑编码(第 3.1 节)、局部解释——样本的局部诱因解释(第 3.2 节)、反事实分析——生成最优反事实样本(第 4 节).

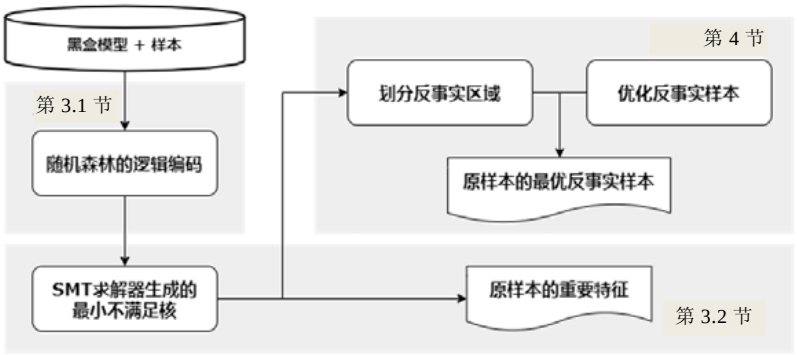


图 1 基于形式化的机器学习可解释性方法流程图

1 相关工作

1.1 特征重要性分析

近年来, 机器学习在各应用领域取得巨大突破的同时, 也由于各类事故的发生而饱受非议, 例如优步的自动驾驶系统误判导致的交通事故、脸书用户资料被窃取用于舆论操控等. 在这样的背景下, 可信的机器学习应运而生, 而可解释性作为其中一个分支, 也受到了学术界的极大关注. 可解释性被定义为向人们解释或呈现可理解的术语的能力[21], 具体而言, 可解释性解释了模型内在的规则, 验证了模型可靠性. 前文提到, 可解释性大致分为局部解释和全局解释. 局部解释通常以输入样本为导向, 通过分析输入样本的每一维特征对模型最终决策结果的贡献来实现[21], 换言之, 特征重要性为局部解释最常用的方法. 在这方面, Ribeiro 等人提出了两个著名的局部解释工具, 其中一个是 LIME[9]: 通过在某个原样本周围的采样生成扰动样本集, 并使用原模型对其进行预测, 根据扰动样本及其预测结果在特定样本周围训练一个易于解释的模型, 例如 Lasso 或者线性模型; 同时, 该解释模型根据扰动样本与原样本之间的距离进行相应的加权, 特征的权重表明了特征的重要性. 之后, 同样基于样本扰动, Ribeiro 等人提出了 Anchor[8]: 根据扰动样本的预测情况, 提取一组满足预设置信度的 IF-THEN 谓词规则来代表模型的预测规则, 扰动样本中满足该谓词规则的特征值被称作“锚点”, 当这些“锚点”特征值固定不变而其他特征值变化时, 预测结果也保持不变. 这两种方法通过构建新的模型或规则来逼近样本周围的局部边界. 而特征重要性的全局解释提供了对整个模型的预测结果具有重要影响的特征. 在这方面, Štrumbelj 等人[22]提出了基于博弈论的 Shapley 方法. 考虑到除特征本身之外, 特征之间的交互作用也会对预测结果产生影响, Shapley 方法通过计算特征对于其余特征集合价值边际贡献的加权和反映特征对于预测结果的影响. 同样考虑到特征交互作用对于预测结果的影响, Henelius 等人[23]利用随机化方法寻找一组尽可能大的特征集合, 该集合的稳定能够最大限度地保证预测值的稳定. 本文特征重要性度量方法进行了局部解释, 通过使用逻辑编码深入模型分析, 解释了模型内部的运作规则.

1.2 反事实样本

在介绍反事实样本之前，不得不先提及对抗样本的概念. 对抗样本(adversarial sample)最初由 Szegedy 等人[14]提出，指的是使模型预测精确度降低的扰动样本，因此，对抗样本经常被用来对模型进行攻击. 近年来，学术界挖掘到了对抗样本在机器学习可解释性方面的潜力；在可解释性领域，对抗样本被称为反事实样本[4]，生成反事实样本的过程[24]指的是尽可能小幅度改变原样本特征值，从而使其预测值变为预定义的输出的过程，而可解释之处在于，样本值如何改变在现实中往往可以为逆境如何转为顺境提供相应的建议和方法，例如贷款失败后，用户如何通过提升自身的条件进而改进提交的用户信息，使得下一次贷款成功，以及房东如何适当放宽租房条件，使得租客愿意支付更高的租金等.

在这方面，Ignatiev 等人[25]从理论上介绍了可解释性与对抗样本之间的对偶关系；Poyiadzi 等人[26]提出了 FACE 的反事实分析工具，通过研究样本之间的密度加权最短路径，找到使得样本转为理想样本的最佳方法；Watcher 等人[27]将问题转为计算损失最小的反事实样本，而损失包括该最佳反事实样本与原样本的距离，以及反事实样本预测值与原样本标签之间的距离；张培歆等人[28]提出了轻量级、可扩展的 Adversarial Discrimination Finder 算法：利用梯度和聚类方法计算深度神经网络的对抗样本，并将其视为个体歧视样本，以此对神经网络的公平性进行验证. 大多数反事实样本生成算法仅适用于神经网络，正如前文提到的那样，集成树模型的离散结构限制了反事实分析在集成树模型领域的发展. 特别地，Tolomei 等人[29]提出了针对树模型的反事实分析：找到所有叶子值为理想预测结果的路径；对于一条路径上涉及的特征阈值稍加扰动，使其靠近或远离该特征所有特征值的平均值，观察经扰动后生成的新样本的预测值是否为理想预测结果，若是,则将其作为候选者；之后，在所有候选者中选择最佳反事实样本. 该方法能够深入树模型内部而进行相应的分析，可惜的是，该方法对于所有样本均使用同一种度量，使得方法的针对性较低. 本文提出的反事实分析方法能够为不同的样本量身打造，具有更高的精确度.

1.3 形式化方法在可信机器学习方面的应用

对于可信机器学习的验证，使用形式化方法是一个极佳的选择. 形式化方法以严格数学规范、设计和验证为核心，将可信机器学习涉及的安全性[30]、隐私性[31]、公平性[32]、鲁棒性和可解释性[33]等性质转化为形式化规范和约束，通过模型检查和定理证明等形式化技术对这些性质进行验证. 近年来，形式化方法已在验证机器学习可信性质方面取得了许多成就. 在验证鲁棒性方面，Ehlers 等人[34]将前馈神经网络(feedforward neural network)近似成线性模型，并将其编码为 SMT 实例，以便于 SAT 求解器识别并选择网络中的特定节点，基于该方法可验证噪声模型的鲁棒性性质；聂超群等人[19]通过将树的鲁棒性性质编码为逻辑公式，通过 SMT 求解器验证了局部鲁棒特征重要性；杨鹏飞等人[35]利用抽象解释方法 DeepPoly 分析深度神经网络的局部鲁棒性，并用线性规划识别 DeepPoly 中存在的虚假区域，以此对抽象解释进行细化，提高分析的精度. 在安全性方面，在假设输入为多面体集合的前提下，向为明等人[36]将以修正线性单元(*ReLU*)为激活函数的神经网络安全性验证问题，转化为集合可达性验证问题；Tran 等人[37]将神经网络的输入假设为星集，并利用两种可达性分析算法对安全性的可达性进行分析. 在公平性方面，Ghosh 等人[38]提出了基于随机可满足性(stochastic satisfiability)的框架 Justicia，并将其实例化为多分类和偏差环节算法，以回答包括统计平等、平等几率等不同维度的公平性度量问题. 在可解释性方面，Bride 等人[18]开发了 Slias 工具，他们将随机森林转化为逻辑公式的形式，并通过提取该公式的最大可满足核(maximum satisfiable core)来分析模型全局的决策过程；Shih 等人[39]将贝叶斯网络编译为有序决策图(ordered decision diagrams)，以解释对决策具有影响的最小特征子集. 鉴于上述形式化方法能够深入模型分析的特性，本文将充分利用形式化，对机器学习模型进行不同方面的可解释性的研究.

2 背景知识

本文聚焦于随机森林分类器，下面将对随机森林分类器的基本原理进行介绍.

如文献[40]所述, 用于分类的数据集由一组形如($\boldsymbol{x}$,y)的样本组成, 其中, $\boldsymbol{x}$ 表示输入向量(本文中加粗表示向量), 通常被称为特征值; y 表示输入的样本标签. 令 X^d 表示特征空间, Y^n 表示输出空间.

决策树 t 由内部节点和终端节点组成, 终端节点通常被称为叶子节点, 而内部节点又被称为决策节点. 本文关注的是二叉树结构的决策树, 其中, 每个决策节点具有两个后继节点, 分为称为左子节点和右子节点. 令 N 表示决策节点的集合, L 表示叶子节点的集合. 每个决策节点接纳了特征空间 X^d 的一个子集, 根节点接纳了 X^d 本身; 对于每个节点 n (除根节点 $n_0\in N$ 外), 其前驱节点为 $n_p\in N$, 且该前驱节点与一个特征值 x_i-阈值 η_{n_p} 公式相对应: 当 $x_i\leqslant\eta_{n_p}$ 成立时, n 为 n_p 的左子节点, 且节点 n_p 对应的 X^d 子集中满足该公式的样本去往该左子结点 n; 反之, n 为 n_p 的右子节点, 且节点 n_p 对应的 X^d 子集中满足该公式的样本去往该右子节点 n. 叶子节点 $l\in L$ 最终接纳了满足其所有父节点对应的特征值-阈值公式的一部分输入, 而叶子节点的值 v_l 表示这些输入对应标签出现的概率, 即 $v_l=(p_1,\ldots,p_n)$.

随机森林分类器 $\hat{f}$ 为 k 棵二叉决策树的集合, 即 $\hat{f}=\{t_1,...,t_k\}$. 随机森林分类器的输出由决策树投票抉择: 将 $\boldsymbol{x}$ 输入随机森林分类器后, 各决策树输出该样本到达的叶子节点值 $t(\boldsymbol{x})=v_l=(p_1,\ldots,p_n)$, 即各类的概率集合, 随机森林分类器采取所有概率平均值的最大值作为输出, 表明该样本的预测结果为概率平均值最大值对应的类. 令 $t_j^i(\boldsymbol{x})$ 表示第 j 棵树输出的第 i 类的概率, 且满足 $\sum_{i=1}^{n}t_j^i(\boldsymbol{x})=1$, 由此, 随机森林的输出为

$$\hat{f}(\boldsymbol{x})=\frac{1}{k}\arg\max_i\sum_{j=1}^{k}t_j^i(\boldsymbol{x}).$$

3 基于最小不满足核的特征重要性分析

本节将介绍用于特征重要性分析的新方法: 基于最小不满足核为局部解释(针对样本)设计了方法框架.

3.1 随机森林的形式化编码

为便于逻辑分析, 本文将随机森林模型编码为一阶逻辑公式. 从决策树中的一条路径开始: 路径由一个叶子节点 l 和若干决策节点 $n\in N_l$ 组成, 其中, N_l 表示根节点 n_0 与叶子节点 l 之间节点的集合. 包含叶子节点 l 的路径的形式化定义如下.

定义 1(决策树中一条路径的形式化定义).

$$\pi(l)::=\bigwedge_{n\in N_l}\begin{pmatrix}L_{p_n}=n\rightarrow x_i\leqslant\eta_{p_n}\\ R_{p_n}=n\rightarrow x_i>\eta_{p_n}\end{pmatrix}\wedge(w=v_l),$$

其中, $\pi(l)$表示从根节点到叶子节点的路径以及叶子节点的值. 当决策节点 n 为其前驱节点 p_n 的左子节点 L_{p_n} 时, 表明 p_n 对应的特征值-阈值公式满足 $x_i\leqslant\eta_{n_p}$; 反之, 节点 p_n 对应的特征值-阈值公式满足 $x_i>\eta_{n_p}$. w 表示叶子节点值的约束变量[20]. 基于此, 决策树可形式化定义如下.

定义 2(决策树的形式化定义).

$$\Pi(t)::=\bigvee_{l\in L}\pi(l),$$

其中, $\Pi(t)$表示决策树为所有路径公式$\pi(l)$的析取范式. 进一步地, 随机森林分类器的形式化编码由所有树的合取范式以及随机森林最终的输出组成, 因此, 随机森林分类器的形式化编码 $R(\boldsymbol{x})$表示为定义 3.

定义 3(随机森林分类器的形式化定义).

$$R(\boldsymbol{x})::=\bigwedge_{j=1}^{k}\Pi(t_j)\wedge\left(output=\frac{1}{k}\arg\max_i\sum_{j=1}^{k}t_j^i(\boldsymbol{x})\right).$$

3.2 局部分析:样本的诱因解释

在文献[19]中, 作者使用形式化方法验证了随机森林的鲁棒性: 将随机森林处理单个样本的鲁棒性性质编码为合取范式(conjunctive normal form, CNF 公式), 并根据 SMT (satisfiability modulo theories)求解器分析该

性质后提供的最小不满足核(minimal unsatisfiable core, MUC)[41]解释该样本的鲁棒性程度. 受此启发, 本文提出一种将随机森林处理单个样本的决策过程转化为一阶逻辑公式的方法, 相应地, 将该逻辑公式交给 SMT 求解器分析, 利用求解器提供的最小不满足核来解释单个样本的重要特征. 最小不满足核的形式化定义如下.

定义 4(最小不满足核). 令 F 为一个 CNF 公式, F_C 为 F 中子式的集合. 如果 $S\subseteq F_C$ 同时满足以下条件, 则称 S 为 F 的最小不满足核.

(1) F 是不可满足的;

(2) S 是不可满足的;

(3) 不存在任何 $S''\subset S$ 为不可满足的.

最小不满足核以集合的形式存在. 当 CNF 公式不可满足时, SMT 求解器中的 DPLL 引擎[42]回溯其分析过的状态搜索最小不满足核. 下面给出一个 CNF 公式的例子.

例 1:

$$\varphi=\omega_1\wedge\omega_2\wedge\omega_3\wedge\omega_4=(x>1)\wedge(y<0)\wedge(y>x)\wedge(y>-3).$$

该公式不可满足: 前 3 个子公式相互矛盾. 然而, 他们中的任意两个却是不矛盾的. 因此, 由上述定义可得, 该公式的最小不满足核为$\{\omega_1,\omega_2,\omega_3\}$. 值得注意的是: 不可满足的逻辑公式可能具有不止一个最小不满足核; 所有最小不满足核的地位相同. 因此, 为了提高计算效率, 本文仅取其中一个最小不满足核.

介绍完最小不满足核的定义, 接着, 将随机森林的决策过程编码为 CNF 公式.

定义 5(随机森林分类器决策过程公式).

$$\Phi_0::=R(\boldsymbol{x})\wedge(\boldsymbol{x}=\boldsymbol{x}^{org})\wedge(output=y^{org}).$$

决策过程可概括为: 在随机森林模型 $R(\boldsymbol{x})$中, 将特征值赋值给输入即 $\boldsymbol{x}=\boldsymbol{x}^{org}$, 模型经计算后输出它的预测结果 $output$. 当 $output$ 与样本自身的标签 y^{org} 相等时, 该决策过程是精确的; 同时, 公式Φ_0为可满足的, 而可满足的公式Φ_0能够解释该精确的决策过程.

根据文献[43], 对样本预测值起重要作用的特征被定义为局部诱因解释(abductive explanation).

定义 6(局部诱因解释). 给定样本的输入 $\boldsymbol{x}^{org}$, y^{org} 及随机森林决策模型Φ_0, 当$\Phi_0(\boldsymbol{x}^{org},y^{org})$为 True 时, 可表示为 $\boldsymbol{x}^{org},y^{org}\models\Phi_0$, 表明随机森林输出了精确的预测. 在此情况下, 若存在特征值子集$\boldsymbol{\delta}\subseteq\boldsymbol{x}^{org}$ 使得$\boldsymbol{\delta},y^{org},\neg\Phi_0\models\bot$, 那么特征值子集$\boldsymbol{\delta}$为该样本的局部诱因解释, 表明$\boldsymbol{\delta}$中的特征值对应特征为重要特征.

联系上文可知, 最小不满足核适合描述样本的局部诱因解释. 为了获取决策模型$\neg\Phi_0$, 本文将决策过程公式进行了重构, 如下所示:

$$\Phi_1::=R(\boldsymbol{x})\wedge(\boldsymbol{x}=\boldsymbol{x}^{org})\wedge(output\neq y^{org}).$$

当决策过程精确时, 模型本应输出的预测结果为 y^{org}, 那么期望预测结果 y^{org} 与实际预测结果 $output$ 的不相等使得公式Φ_1不可满足. 由此, Φ_1等价于$\neg\Phi_0$, 即$\Phi_1\equiv\neg\Phi_0$. 进一步地, $output$ 由 $\boldsymbol{x}^{org}$ 映射得来, 对于不同样本的 $\boldsymbol{x}^{org}$, 随机森林模型 $R(\boldsymbol{x})$不变, 因此, 公式Φ_1的不可满足性可最终归因于赋值过程 $\boldsymbol{x}=\boldsymbol{x}^{org}$, 符合局部诱因解释的定义. 当 SMT 求解器解析决策过程公式Φ_1时, DPLL 引擎被设置成仅可追溯 $\boldsymbol{x}=\boldsymbol{x}^{org}$ 中的子式, 每个子式表示相应特征值的赋值过程; 返回的最小不满足核为若干赋值子式的集合, 而子式中特征值对应的特征即为对公式Φ_1 的不可满足性影响较大的特征, 即为对精确的决策过程起决定性作用的特征; 局部诱因解释的结果最终归纳为重要特征的集合.

将决策过程毫无保留地转化为逻辑公式交付 SMT 求解器分析, 能够从返回的最小不满足核中提取对于样本决策有重要影响的特征作为局部诱因解释; 同时, 由于舍去了离散特征值以及使用统计方法的步骤, 解析过程的精确性得以保证.

4 基于最小不满足核的反事实分析

基于统计和近似的局部解释无法进行拓展以解决其他的相关问题, 例如, 样本的关键特征在作为对于预测有重要影响因素的同时, 也可能暴露该样本的缺陷: 样本的关键特征值被稍加修改后, 能够改变样本原本

的预测结果. 从理论的角度而言, 反事实样本(counterfactual sample)指代这些修改后能够使模型输出不同预测结果的样本. 本节展示了如何充分利用先前编码逻辑公式, 生成符合期望的最优反事实样本.

4.1 划分反事实区域

在文献[44]中, 作者将电路抽象为逻辑公式, 并用求解器进行分析, 当该公式为不可满足时, 表明电路中存在异常; 通过反复修正不可满足性公式中的最小不满足核, 即逐步修改矛盾子公式中变量的值, 消除子公式之间的矛盾, 公式最终转化为可满足的. 相应地, 电路恢复正常. 这给了本文一个启发: 在最小不满足核的指导下, 可以通过修改样本中的特征值, 使得相应的预测结果变为期望值.

首先, 根据公式$\boldsymbol{\Phi}_1$得到的最小不满足核, 能够知道样本的哪些特征对预测结果的影响最大; 而后, 试图通过修改这些重要特征值, 使得决策过程公式变为可满足的, 即 $y^{org}\neq output$ 可满足, 决策模型最终输出的结果与样本自身的标签不同. 不同于电路模型中值仅存在“0”和“1”两种情况, 随机森林模型中的特征值可变化范围大, 通过逐步修改特征值使得不可满足的决策过程公式转变为可满足将带来巨大的工作量. 因此, 本文转而检查在临近$\boldsymbol{x}^{org}$的区域内是否存在与$\boldsymbol{x}^{org}$特征值相近的反事实样本. 对于二分类数据集来说, 通常定义负面的分类结果为“0”, 正面的分类结果为“1”. 对于这些负面的分类结果, 希望通过修改特征值使得分类结果变为正面, 即期望生成的正面结果 *output* 与负面样本本身的标签 y^{org} 不相等. 因此, 决策过程公式可转化为如下形式.

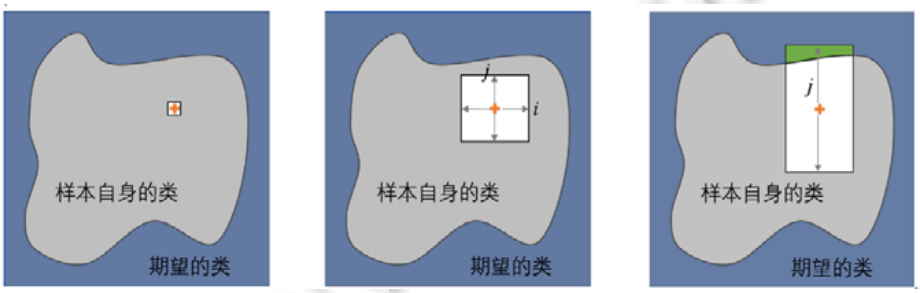
定义 7(反事实区域划分公式).

$$\boldsymbol{\Phi}_1::=R(\boldsymbol{x})\wedge(\boldsymbol{x},\boldsymbol{x}^{org},\boldsymbol{\tau})\wedge(output\neq y^{org}),$$

其中,

$$\sigma(\boldsymbol{x},\boldsymbol{x}^{org},\boldsymbol{\tau})::=\bigwedge_{i=1}^{d}|x_i-x_i^{org}|\leqslant\tau_i.$$

搜索范围$\boldsymbol{\tau}$代表 $\boldsymbol{x}^{org}$ 周围的区域, τ_i 代表 x_i^{org} 方向上的搜索范围, 公式$\boldsymbol{\Phi}_2$的可满足性意味着搜索范围内存在反事实样本, 将其输入到随机森林模型 $R(\boldsymbol{x})$内进行计算之后, 模型的输出 *output* 与样本自身的标签不同. τ_i 初始化为 0, 之后, 根据最小不满足核的情况逐步变化: 当最小不满足核中不存在特征 f_i 时, 表明该特征当前对于预测结果的影响不大, τ_i 保持不变, 即当前不在特征 f_i 的方向上搜索; 否则, τ_i 按既定步长增大. 此外, 当最小不满足核中存在实际生活中不可人为改变的特征时, 例如年龄、性别等, τ_i 同样保持不变. 搜索范围经过一次扩大后, 可能仍未足够大到能够覆盖到反事实样本所在区域. 换言之, 公式$\boldsymbol{\Phi}_2$仍不可满足. 因此, 持续扩大搜索范围, 即不断增加τ_i的值, 直至寻找到存在于$\boldsymbol{x}^{org}$不远处的反事实样本, 使得公式$\boldsymbol{\Phi}_2$变为可满足. 此时, 搜索区域不仅覆盖了原样本所属类的区域, 同时覆盖到了期望类的区域. 我们将期望类与搜索范围相交的区域称为“反事实区域”(如图 2 所示).



(a) (b) (c)

(a) 起始, 搜索范围仅覆盖了的原样本所在的一点区域;
(b) 在这一轮, 特征 f_i 和特征 f_j 在当前的最小不满足核中, τ_i 和 τ_j 相应增大;
(c) 在这一轮, 特征 f_i 不在当前的最小不满足核中, 因此 τ_i 保持不变. 暗色区域为尚未搜索到的区域, 白色区域为已搜索到的区域, 绿色区域为反事实区域

图 2 最小不满足核指引下的搜索过程

4.2 优化反事实样本

在获得反事实区域之后，本文希望在反事实区域中找到离原样本最近的反事实样本，即最优反事实样本. 本文采用零阶优化方法(zero order optimization, Opt-Counterfactual)[45]寻找最优反事实样本，零阶优化适用于随机森林这样的离散模型. 反事实区域中存在着无数反事实样本，原样本到任意反事实样本的向量称为搜索向量，某一搜索向量对应方向上的最优反事实样本生成目标方程如下:

$$g(\boldsymbol{\theta})=\arg\min_{\lambda>0}\left(\hat{f}\left(x^{org}+\lambda\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\right)\neq y^{org}\right),$$

其中,

$$\boldsymbol{\theta}=\boldsymbol{x}^{con}-\boldsymbol{x}^{org}.$$

$\hat{f}$ 为随机森林分类器; $\boldsymbol{\theta}$表示从原样本 $\boldsymbol{x}^{org}$ 到反事实样本 $\boldsymbol{x}^{con}$ 的搜索向量; λ为它们之间的距离，即搜索距离，单位搜索向量为搜索方向. $g(\boldsymbol{\theta})$试图得到$\boldsymbol{\theta}$方向上距离原样本最近的反事实样本. 在所有各自方向上的最优反事实样本中，本文希望得到全局最优的反事实样本:

$$\min_{\boldsymbol{\theta}} g(\boldsymbol{\theta}).$$

具体而言，在零阶优化方法中，随机无梯度方法(randomized gradient-free)被用来优化给定的搜索向量. 梯度的计算方法为

$$\hat{g}=\frac{g(\boldsymbol{\theta}')-g(\boldsymbol{\theta})}{\beta}\cdot\boldsymbol{u},$$

其中,

$$\boldsymbol{\theta}'=\boldsymbol{\theta}+\beta\boldsymbol{u}.$$

在$\boldsymbol{\theta}$方向上增加平滑系数β和单位随机高斯向量 $\boldsymbol{u}$ 作为微小扰动，得到扰动向量$\boldsymbol{\theta}'$，再根据计算出两个向量之间的梯度 $\hat{g}$，该梯度被用来对搜索向量$\boldsymbol{\theta}$进行优化 $\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}-\eta\hat{g}$，$\eta$为既定步长. 值得一提的是: 这里梯度的计算基于随机森林的预测结果，并不依赖随机森林内部结构，正如前文提到的那样，树结构没有梯度.

在运行零阶优化方法时，需要输入一个初始搜索向量$\boldsymbol{\theta}_0$; 对于该向量，本文希望它对应的反事实样本已经足够接近原样本，以便于接下来使用随机无梯度方法对该向量进行更精细的优化. 为此，给定的一个反事实样本 $\boldsymbol{x}^{con}$，对其进行细粒度搜索和二分搜索，使其逆着搜索方向尽可能地靠近原样本. 算法 1 展示了两个搜索方法的全过程，其中, α为升降系数, ε为使算法停止运算的最大误差. 在第 1 阶段，通过细粒度搜索，使得反事实样本按照微小步长α沿着搜索方向移动，最终落于$[\boldsymbol{x}^{org}+v_{in}\boldsymbol{\theta},\boldsymbol{x}^{org}+v_{out}\boldsymbol{\theta}]$区间内; 第 2 阶段，通过二分搜索，使得反事实样本逼近反事实区域和样本类所在区域的分界线.

算法 1. 二分搜索和细粒度搜索算法 *Fine-binary*$(\hat{f},\boldsymbol{\theta},\boldsymbol{x}^{org},\boldsymbol{x}^{con})$.

输入: 随机森林分类器 $\hat{f}$，搜索向量$\boldsymbol{\theta}$, 原样本 $\boldsymbol{x}^{org}$ 和反事实样本 $\boldsymbol{x}^{con}$.

输出: $\boldsymbol{\theta}$方向上反事实样本与原样本之间的距离 v_{out}.

1: $\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$
2: $v_{out}\leftarrow\|\boldsymbol{\theta}\|$, $v_{in}\leftarrow\|\boldsymbol{\theta}\|$
3: **while** $\hat{f}(\boldsymbol{x}^{org}+v_{in}\boldsymbol{\theta})=\hat{f}(\boldsymbol{x}^{con})$ **do**
4: $v_{out}\leftarrow v_{in}$, $v_{in}\leftarrow v_{out}(1-\alpha)$
5: **endwhile**
6: **while** $v_{out}-v_{in}>\varepsilon$ **do**
7: $v_{mid}\leftarrow(v_{out}+v_{in})/2$
8: **if** $\hat{f}(\boldsymbol{x}^{org}+v_{mid}\boldsymbol{\theta})=\hat{f}(\boldsymbol{x}^{con})$ **then**
9: $v_{out}\leftarrow v_{mid}$
10: **else**

11: $v_{in}\leftarrow v_{mid}$

12: **endif**

13: **endwhile**

算法 2 描述了生成最优反事实样本的全过程, 其中, κ_i 为扩大搜索范围 τ_i 的搜索步长, $MUC^2_{x^{org}}$ 表示由公式 $\Phi_2(x^{org},y^{org})$生成的最小不满足核, T 为零阶优化的执行次数, $*x^{con}$ 为最终的最优反事实样本. 值得一提的是: 特征值可能并不处于同一个数量级上, 例如年龄(25)和年收入(300 000 元), 单位随机高斯向量对于搜索向量来说扰动可能导致各特征值的变化幅度千差万别, 使搜索向量的优化过程变得不可控制. 为了尊重数据的原始性, 本文不在数据预处理时将特征值处理为同一数量级, 而是设置数量级控制向量 $\boldsymbol{\mu}$, 使得高斯向量对于向量各方向上的扰动效果保持一致.

算法 2. x^{org} 的最佳反事实样本生成算法.

输入: 随机森林分类器 $\hat{f}$, 随机森林的 CNF 公式 $R(x)$, 样本特征 x^{org}, 样本标签 y^{org} 和特征集合 F.

输出: 最优反事实样本$*x^{con}$, $\Phi_1::=R(x)\wedge(x,x^{org},\tau)\wedge(output\neq y^{org})$.

1: $\tau\leftarrow 0$

2: $\Phi_2\leftarrow R(x)\wedge\sigma(x,x^{org},\tau)\wedge(output\neq y^{org})$

3: **while** $UNSAT=solver(\Phi_2)$ **do**

4: **forall** $f_i\in F$ **and** $f_i\in MUC^2_{x^{org}}$ **do**

5: $\tau_i\leftarrow\tau_i+\kappa_i$

6: **endfor**

7: **endwhile**

8: $floor\leftarrow x^{org}-\tau$, $ceil\leftarrow x^{org}+\tau$

9: 区间内随机生成样本并将其归纳到集合 X 中

10: $\lambda_{\min}\leftarrow 0$

11: **forall** $x\in X$ **and** $\hat{f}(x)\neq y^{org}$ **do**

12: $\theta\leftarrow x-x_{org}$

13: $\lambda\leftarrow Fine\text{-}binary(\hat{f},\theta,x^{org},x)$

14: **if** $\lambda_{\min}>\lambda$ **then**

15: $\lambda_{\min}\leftarrow\lambda$, $\theta_0\leftarrow\theta$

16: **endif**

17: **endfor**

18: **for** t=0,1,2,…,T **do**

19: $\theta_t'\leftarrow\theta_t+\beta u_t\cdot\mu$

20: $g(\theta_t')\leftarrow Fine\text{-}binary(\hat{f},\theta_t',x^{org},x)$, $g(\theta_t)\leftarrow Fine\text{-}binary(\hat{f},\theta_t,x^{org},x)$

21: $\hat{g}\leftarrow\frac{g(\theta_t')-g(\theta_t)}{\beta}\cdot u_t$

22: $\theta_{t+1}=\theta_t-\eta_t\hat{g}$

23: **endfor**

24: $*x^{con}\leftarrow x^{org}+g(\theta_T)||\theta_T||$

应用该算法时应注意: 首先, 在细粒度和二分搜索过程中对向量进行缩放时, 若某向量分量对应的特征从未出现在最小不满足核中, 则该分量的值保持不变, 有助于缩小特征空间, 提高搜索效率; 其次, 步长 κ_i 也需要根据特征值的数量级而变化, 且将其设置得尽可能小, 以保证反事实区域的可用性和精确性.

5 实 验

本节介绍实验结果.

5.1 数据集与实验设置

我们对 7 个 UCI 数据集进行实验: credit (二分类), breast (二分类), heart (二分类), adult (二分类), german (二分类), diabete (二分类)和 MNIST (多分类), 同时还选择了两个有关贷款的数据集: lending (二分类)和 bank loan (二分类). MNIST 数据集专门用于可视化局部诱因解释和反事实分析. 每个数据集被分成两个子集: 80%用于训练, 20%用于测试.

本节还将通过贷款分析案例展示反事实分析在实际中的应用, 即反事实分析如何为那些贷款失败的人提供建议, 帮助他们下一次贷款成功: 使用反事实分析, 能够为如何有针对性地更改提交的用户信息提供建议; 同时, 从用户友好的角度出发, 用户信息需尽可能小幅度地修改, 使得用户实际上不必为此付出过多代价, 而反事实分析提供的结果正好适用于该情景.

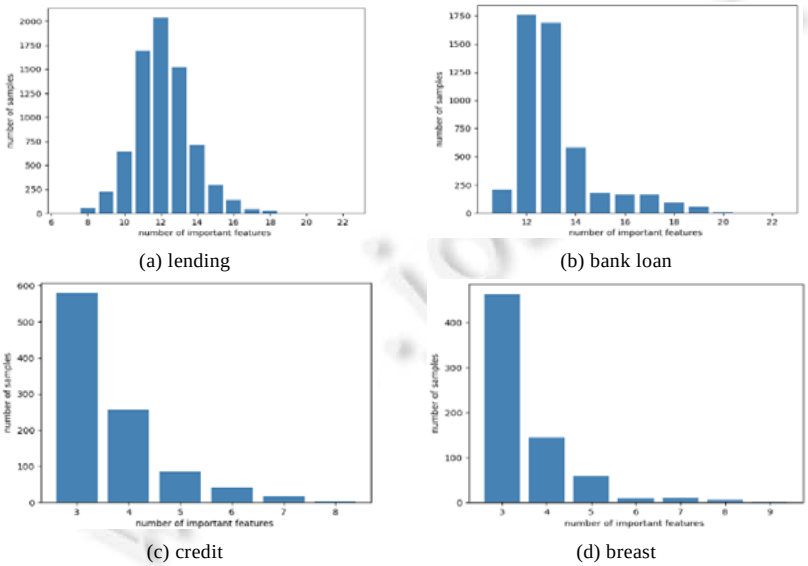
5.2 实验结果

5.2.1 局部分析

针对每个数据集, 统计了其中每个测试样本局部诱因解释中的重要特征数: 若某样本的特征为(年龄,性别,年收入,贷款金额), 其重要特征为(年收入,贷款金额), 那么该样本的重要特征数为 2. 统计结果总结于表 1, 表中包含计算各样本重要特征的平均时长、出现次数最多的重要特征数即众数、重要特征数的平均值、数据集的总特征数、平均数/总特征数作为特征利用率以及模型精确度. 图 3 展示了各数据集中所有测试样本重要特征数的分布情况.

表 1 样本的重要特征数统计结果

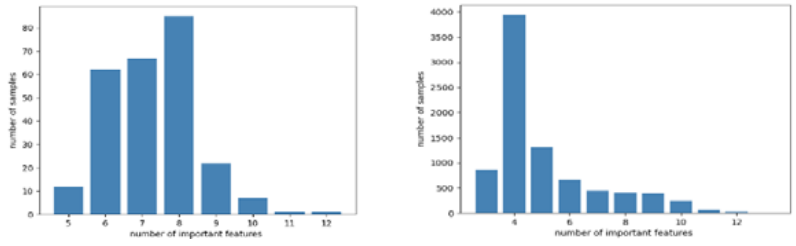
数据集	lending	bank loan	credit	heart	breast	adult	german	diabete	MNIST
平均时长(s)	16.23	1.98	14.37	0.68	0.53	4.12	2.32	0.41	70.41
众数	12	3	12	8	3	4	8	5	104
平均数	12	3	13	7	4	5	10	5	105
特征总数	30	11	23	13	9	13	20	8	784
特征利用率(%)	40	28	57	54	44	38	50	63	13
模型精确度(%)	94.19	98.30	82.58	77.78	97.86	85.58	73.50	79.22	93.37



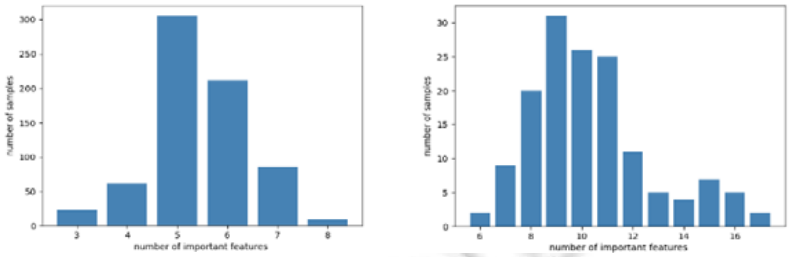
(a) lending (b) bank loan

(c) credit (d) breast

图 3 测试样本的重要特征数分布



(e) heart (f) adult



(g) german (h) diabete

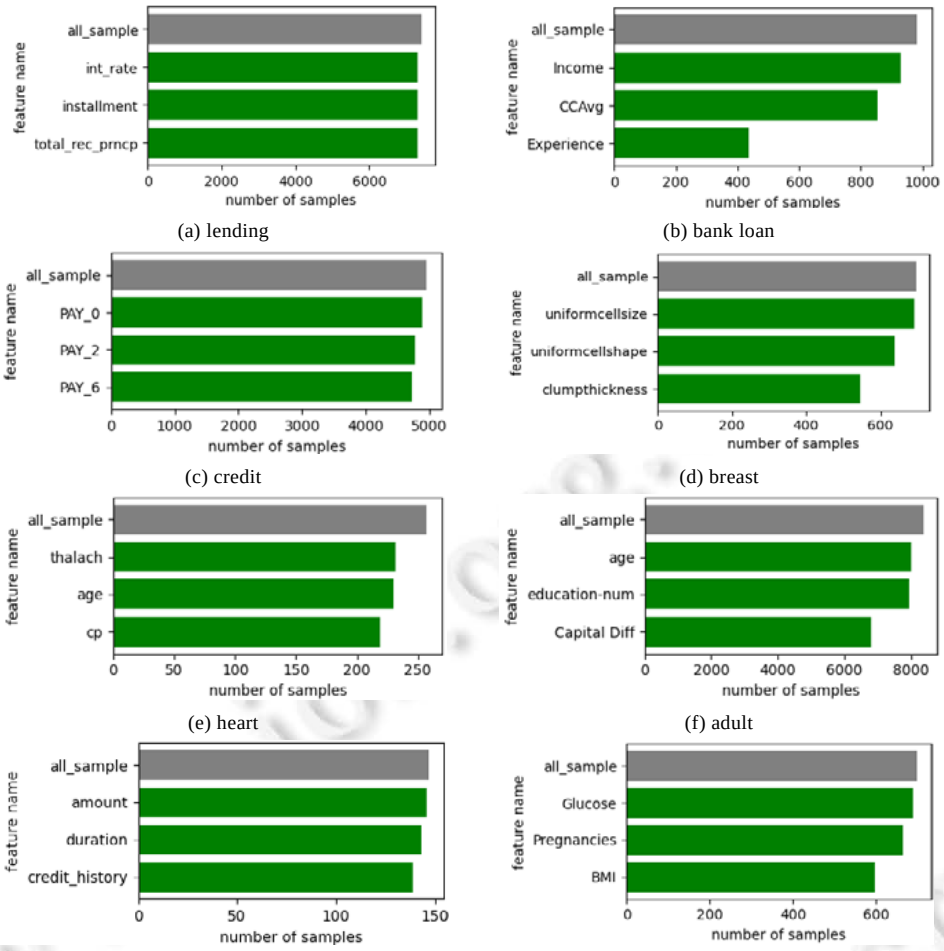
图 3 测试样本的重要特征数分布(续)

此外, 本文还统计了所有最小不满足核中各特征出现的频数(frequency of features in minimal unsatifiable core, FFMUC): 若现有 3 个 MUC $\{f_1,f_2,f_5\}$, $\{f_1,f_2\}$, $\{f_2,f_3,f_6\}$, 那么特征 f_1 的 FFMUC 为 2, f_2 的 FFMUC 为 3, 图 4 展示了 8 个二分类的数据集中 FFMUC 排名前三的特征分布, 图 5 利用 MNIST 数据集展示了局部诱因解释的可视化结果, “x”标记了重要特征所在位置.

由表 1 可知: 重要特征数的众数和平均值近乎相等; 特征利用率的最大值为 57%, 最小值为 13%, 表明在各数据集中确实存在能作为该数据集“标识”的特征, 且占总特征数的小部分. 图 3 同样证实了众数与平均值近乎相等, 且特征利用率往往低于 50%. 图 4 中最上方的数据条表明参与运算的样本数, 其余的数据条表明该特征出现的频数. 以 bank loan 数据集为例: bank loan 数据集中共有 982 个样本参与运算, 而有 928 个样本的最小不满足核中包含了 Income 这一特征, 说明 Income 特征在模型的全局预测中占据着重要的地位. 此外, 由图可知, 排名前三的特征出现在了绝大部分样本的最小不满足核中, 表明了对模型预测的重要影响. 从图 5 可知: 不同数字的手写图像的重要特征不同, 且相同数字不同手写风格的图像的重要特征也不尽相同. 这反映了基于最小不满足核的局部解释方法的可行性和灵活性, 针对不同的样本, 它给出的解释并非千篇一律. 此外, 在局部诱因解释中, 重要特征分布在相应数字的形状周围, 且位于角落的特征通常对于预测的影响很小, 这证明了局部诱因解释是符合常识的, 并且是有意义的. 表格类的数据的局部诱因解释是易于理解的, 例如图 4(g) german 中所展示的贷款金额、贷款周期、信用历史, 而图像类的局部诱因解释不易于理解, 如图 5 所示: 其尚未勾勒出某一图形的轮廓, 也不存在规律. 图像类的局部诱因解释的意义更在于为后续的反事实分析服务. 总之, 该方法在保证模型精确度的同时, 能够提供较高质量的解释.

为了探究局部诱因解释在全局解释方面的可拓展性, 根据特征 FFMUC 的排名进行特征选择, 并将选择后的数据集交给原模型进行预测, 观察模型精确度的变化. 特征选择的常用策略为: 将去除重要特征的数据集输入模型进行预测, 以此观察模型精确度的下降程度. 然而, 这种策略存在弊端: 前后的特征分布产生了变化, 使得并不能完全确定模型精确度的变化归因于特征分布的变化, 还是归因于被删除的重要特征. 因此, 该策略并不能达到特征选择原本的目的[46]. 在此, 本文依次对当前选中查看的特征的值加上高斯噪声, 以此消除该特征值对于预测结果的影响, 同时维护了再训练前后特征分布的一致性.

图 6 展示了 4 个二分类的 UCI 数据集在根据 FFMUC 的排名前十的特征进行特征选择后, 模型精确度的变化情况. 由图可知, 总体情况来看, 随着特征值的逐步扰动, 模型精度不断下降, 表明了最小不满足核在全局解释方面的可拓展性.



(a) lending (b) bank loan

(c) credit (d) breast

(e) heart (f) adult

(g) german (h) diabete

图 4 数据集中 FFMUC 排名前三的特征分布

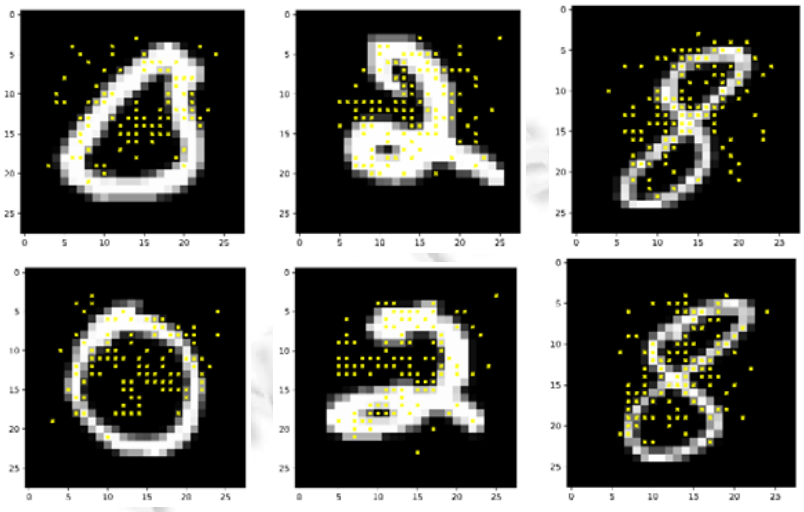


图 5 MNIST 测试样本的重要特征示意图

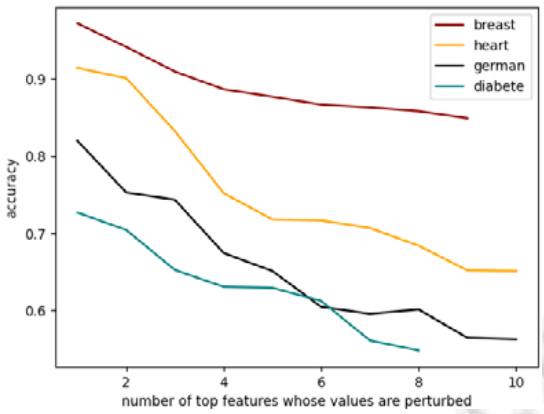


图 6　前 N 个特征值扰动后模型的预测精度变化

5.2.2　反事实分析

本文记录并比较了 Opt-Counterfactual 和基于最小不满足核的 Opt-Counterfactual (MUC-Counterfactual)方法生成最优反事实样本的总时长. 该总时长包括利用细粒度和二分搜索确认初始搜索向量的搜索时间以及优化反事实样本的优化时间, 其中, MUC-Counterfactual 的搜索时间包括划分反事实区域的时间和之后确定初始搜索向量的时间.

同时, 最优反事实样本与原样本之间距离的平均值也用于评价算法的性能, 即 $\frac{1}{m}\sum_{j=1}^{m}*\boldsymbol{x}_j^{con}-\boldsymbol{x}_j^{org}$. 两种算法中相同的部分设置了相同的参数, 以保证比较的公平性.

由表 2 可知, 两个算法的运行时间相近. 然而, MUC-Counterfactual 生成的反事实样本离原样本更近, MUC-Counterfactual 具有更高的效率. 这是由于 Opt-Counterfactual 在优化搜索向量时, 试图在向量每个可能的方向上都进行优化, 而基于最小不满足核的指导, MUC-Counterfactual 能够具有目的性地在个别方向上进行优化, 虽然前期牺牲了一部分时间划分反事实区域, 但后期因此提高了优化效率.

表 2　反事实样本实验结果

数据集	算法	搜索时长(s)	优化时长(s)	总时长(s)	距离
lending	Opt-Counterfactual	36.77	9.92	46.69	1 4943.58
	MUC-Counterfactual	39.97	7.95	47.92	**8 312.92**
bank loan	Opt-Counterfactual	25.15	12.46	37.61	25.46
	MUC-Counterfactual	27.92	10.55	38.48	**7.75**
credit	Opt-Counterfactual	101.122	59.09	160.22	23 522.53
	MUC-Counterfactual	74.77	11.89	86.66	**628.82**
heart	Opt-Counterfactual	3.69	4.22	7.92	8.38
	MUC-Counterfactual	5.98	3.87	9.86	**7.55**
breast	Opt-Counterfactual	31.32	23.64	54.97	5.66
	MUC-Counterfactual	11.83	18.44	30.28	**5.14**
adult	Opt-Counterfactual	76.45	31.26	107.71	4 089.73
	MUC-Counterfactual	16.19	9.88	26.06	**16.46**
german	Opt-Counterfactual	152.00	35.56	187.56	6.78
	MUC-Counterfactual	13.12	10.70	23.82	**1.51**
diabete	Opt-Counterfactual	10.55	23.05	33.60	20.70
	MUC-Counterfactual	23.16	21.12	44.29	**18.32**
MNIST	Opt-Counterfactual	4.33	65.45	69.78	387.35
	MUC-Counterfactual	93.29	4.91	98.2	**36.53**

图 7 显示了反事实分析的可视化结果, 分别为原始图像、基于 MUC-Counterfactual 方法生成的反事实样本以及基于 Opt-Counterfactual 方法生成的反事实样本; 对图像的细微扰动通常难以被人眼察觉, 然而从图中可得知, Opt-Counterfactual 生成一个具有可见阴影的反事实样本, 而 MUC-Counterfactual 生成的反事实样本几乎与原始样本相同. 另外, 由于 Opt-Counterfactual 的初始反事实样本是从原始数据集中挑选的, 当数据集

中样本量较少时，并不能为 Opt-Counterfactual 挑选到好的初始反事实样本提供条件；相比之下，反事实区域这一连续空间中包含了无数的反事实样本，为 MUC-Counterfactual 挑选好的初始样本提供了条件.

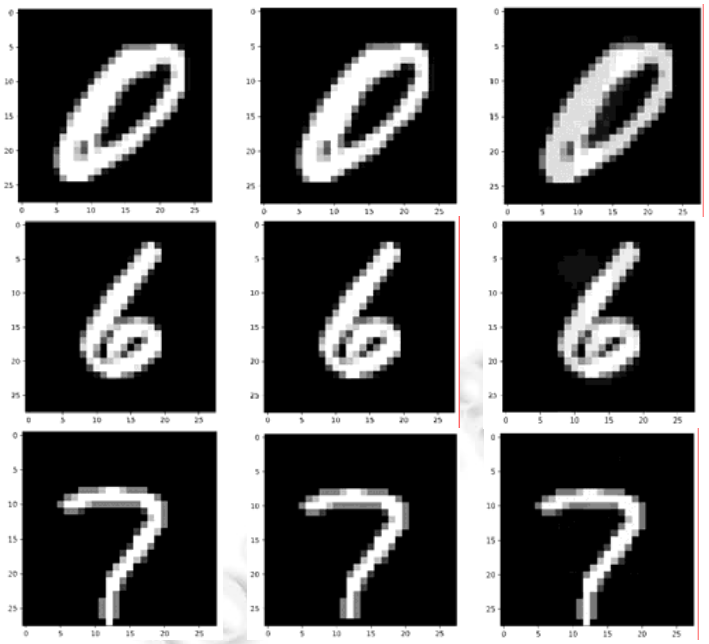


图 7　反事实分析的可视化结果

5.2.3　案例分析

最后将展示反事实分析如何为贷款失败的用户提供建议，以帮助他们下一次贷款成功.

以用户 No.405415 为例，根据他提供的用户资料可知他的基本信息：他已经在佛罗里达州的 Marine Max 工作了 9 年，年收入 48 000 美元且租有一房；该用户希望贷款 18 200 美元，为期 60 个月，以此偿还他的信用卡支付以及其他的高息贷款. 根据他提交的所有用户信息，银行利用机器学习模型进行预测，来评价是否能够同意他此次的贷款. 根据模型的评价结果，银行决定拒绝了他这次的贷款. 在这种情况下，从为用户考虑的角度出发，银行可利用本文提出的反事实分析为用户找到贷款成功的方法，即用户需要改进哪些信息使得银行下次的评价结果为同意贷款；并可根据反事实分析的结果为用户设计了一个报告，报告中罗列了有待改进的用户信息以及其改进程度，例如贷款额(用户信息)减少约 0.7% (改进程度)，为用户下一次的贷款成功提供具有价值和针对性的建议. 图 8 和图 9 展示了基于 MUC-Counterfactual 和 Opt-Counterfactual 方法提供的报告. 相比之下，基于 MUC-Counterfactual 提供的建议不涉及对用户信息的大幅度更改，并且更易于实现. 由此可见，基于 MUC-Counterfactual 提供的用户报告更有助于客户改进其个人信息，使他们在下一次贷款中获得成功.

亲爱的用户No.405415，为帮助您成功申请贷款，以下建议供您参考：
减少约0.7%的贷款额；
提高1.6%的利息；
月分期付款减少约0.1%；
增加0.18%的收入

图 8　基于 MUC-Counterfactual 反事实样本结果的银行贷款信息改进建议报告

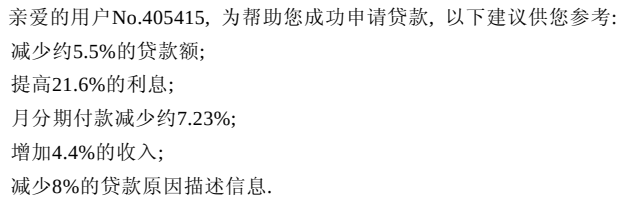


图 9 基于 OPT-Counterfactual 反事实样本结果的银行贷款信息改进建议报告

6 总结与展望

本文研究了基于形式化方法的机器学习模型可解释性问题, 重点聚焦于机器学习模型中极具发展前景而可解释性应用尚未成熟的随机森林分类器模型. 首先, 将随机森林分类器对于单个样本的决策过程巧妙编码为一阶逻辑公式, 并交付 SMT 求解器分析, 根据 SMT 求解器提供的最小不满足核来解释样本的重要特征, 将其作为局部诱因解释. 进一步地, 本文提出了一种新的反事实分析方法: 依据生成的最小不满足核划分原样本的反事实区域, 并在其中寻找最优反事实样本.

实验结果表明, 基于形式化的局部诱因解释方法能够提供高质量、直观的特征相关解释. 同时, 探究了局部诱因解释在全局解释方面的拓展性; 在相同的时间内, 与现有的反事实分析方法相比, 基于最小不满足核的反事实分析能够生成更接近原样本的反事实样本. 在案例分析中, 本文展示了基于最小不满足核的反事实分析的实际适用性: 从用户友好的角度出发, 根据最优反事实样本生成的报告, 能够为负面样本转为正面样本提供具有针对性且易于实现的建议.

在未来的研究中, 编码的随机森林决策过程公式可被应用于更多性质的验证当中, 例如隐私性、安全性等. 此外, 可将最小不满足核挖掘到的模型内部的逻辑信息与现有全局解释方法相结合, 例如 Shapley Values, 更大程度地发挥它的潜能.

References:

[1] Tian Y, Pei K, Jana S, *et al*. DeepTest: Automated testing of deep-neural-network-driven autonomous cars. In: Proc. of the 40th Int'l Conf. on Software Engineering. 2018. 303–314. [doi: 10.1145/3180155.3180220]

[2] Chen M, Hao Y, Hwang K, *et al*. Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 2017, 5: 8869–8879. [doi: 10.1109/ACCESS.2017.2694446]

[3] Alexopoulos C, Lachana Z, Androutsopoulou A, *et al*. How machine learning is changing e-government. In: Proc. of the 12th Int'l Conf. on Theory and Practice of Electronic Governance. 2019. 354–363. [doi: 10.1145/3326365.3326412]

[4] Molnar C. Interpretable machine learning. 2020. https://christophm.github.io/interpretable-ml-book/index.html

[5] Vilone G, Longo L. Explainable artificial intelligence: A systematic review. arXiv: 2006.00093, 2020.

[6] Arrieta AB, Díaz-Rodríguez N, Del Ser J, *et al*. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020, 58: 82–115.

[7] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 2018, 6: 52138–52160. [doi: 10.1109/ACCESS.2018.2870052]

[8] Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. 2018. 1527–1535.

[9] Ribeiro MT, Singh S, Guestrin C. Why should I trust you?" Explaining the predictions of any classifier. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2016. 1135–1144. [doi: 10.1145/2939672.2939778]

[10] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5–32.

[11] Schapire RE. A brief introduction to boosting. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. 1999. 1401–1406. [doi: 10.1109/CICC.1996.510579]

[12] Safavian S, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans. on Systems, Man, and Cybernetics, 1991, 21(3): 660–674. [doi: 10.1109/21.97458]

[13] Yu G, Yuan J, Liu Z. Unsupervised random forest indexing for fast action search. In: Proc. of the CVPR 2011. IEEE, 2011. 865–872. [doi: 10.1109/CVPR.2011.5995488]

[14] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv: 1412.6572, 2014.

[15] Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 2574−2582. [doi: 10.1109/CVPR.2016.282]

[16] Papernot N, McDaniel P, Jha S, *et al*. The limitations of deep learning in adversarial settings. In: Proc. of the 2016 IEEE European Symp. on Security and Privacy (EuroS&P). IEEE, 2016. 372−387. [doi: 10.1109/EuroSP.2016.36]

[17] Zhang H, Zhou H, Miao N, *et al*. Generating fluent adversarial examples for natural languages. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. 5564−5569.

[18] Bride H, Cai CH, Dong J, *et al*. Silas: A high-performance machine learning foundation for logical reasoning and verification. Expert Systems with Applications, 2021, 176(1): Article No.114806. [doi: 10.1016/j.eswa.2021.114806]

[19] Nie C, Shi J, Huang Y. VARF: Verifying and analyzing robustness of random forests. In: Proc. of the Int'l Conf. on Formal Engineering Methods. Cham: Springer, 2020. 163−178.

[20] Einziger G, Goldstein M, Sa'ar Y, *et al*. Verifying robustness of gradient boosted models. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2019. 2446−2453.

[21] Ji SL, Li JF, Du TY, Li B. Survey on techniques, applications and security of machine learning interpretability. Jisuanji Yanjiu yu Fazhan/Computer Research and Development, 2019, 56(10): 2071−2096 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2019.20190540]

[22] Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 2014, 41(3): 647−665. [doi: https://doi.org/10.1007/s10115-013-0679-x]

[23] Henelius A, Puolamäki K, Boström H, Asker L, Papapetrou P. A peek into the black box: Exploring classifiers by randomization. Data Mining and Knowledge Discovery, 2014, 28(5): 1503−1529.

[24] Pan WW, Wang XY, Song ML, Chen C. Survey on generating adversarial examples. Ruan Jian Xue Bao/Journal of Software, 2020, 31(1): 67−81 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5884.htm [doi: 10.13328/j.cnki.jos.005884]

[25] Ignatiev A, Narodytska N, Marques-Silva J. On relating explanations and adversarial examples. Advances in Neural Information Processing Systems, 2019, 32: 15883−15893.

[26] Poyiadzi R, Sokol K, Santos-Rodriguez R, *et al*. FACE: Feasible and actionable counterfactual explanations. In: Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society. 2020. 344−350.

[27] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. arXiv:1711.00399, 2017.

[28] Zhang P, Wang J, Sun J, *et al*. White-box fairness testing through adversarial sampling. In: Proc. of the 42nd ACM/IEEEInt'l Conf. on Software Engineering. 2020. 949−960. [doi: 10.1145/3377811.3380331]

[29] Tolomei G, Silvestri F, Haines A, *et al*. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In: Proc. of the 23rd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2017. 465−474. [doi: 10.1145/3097983.3098039]

[30] Li XJ, Wu GW, Yao L, Zhang WZ, Zhang B. Progress and future challenges of security attacks and defense mechanisms in machine learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 406−423 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6147.htm [doi: 10.13328/j.cnki.jos.006147]

[31] Liu RX, Chen H, Guo RY, Zhao D, Liang WJ, Li CP. Survey on privacy attacks and defenses in machine learning. Ruan Jian Xue Bao/Journal of Software, 2020, 31(3): 866−892 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5904.htm [doi: 10.13328/j.cnki.jos.005904]

[32] Liu WY, Shen CY, Wang XF, Jin B, Lu XJ, Wang XL, Zha HY, He JF. Survey on fairness in trustworthy machine learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(5): 1404−1426 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6214.htm [doi: 10.13328/j.cnki.jos.006214]

[33] Hua YY, Zhang DX, Ge SM. Research progress on interpretability of deep learning model. Journal of Cyber Security, 2020, 5(3): 1−12 (in Chinese with English abstract). [doi: 10.19363/J.cnki.cn10-1380/tn.2020.05.01]

[34] Ehlers R. Formal verification of piece-wise linear feed-forward neural networks. In: Proc. of the Int'l Symp. on Automated Technology for Verification and Analysis. Cham: Springer, 2017. 269−286.

[35] Yang P, Li R, Li J, *et al*. Improving neural network verification through spurious region guided refinement. In: Tools and Algorithms for the Construction and Analysis of Systems. 2021. 389−408. [doi: 10.1007/978-3-030-72016-2_21]

[36] Xiang W, Tran HD, Johnson TT. Reachable set computation and safety verification for neural networks with relu activations. arXiv: 1712.08163, 2017.

[37] Tran HD, Lopez DM, Musau P, *et al*. Star-based reachability analysis of deep neural networks. In: Proc. of the Int'l Symp. on Formal Methods. Cham: Springer, 2019. 670−686.

[38] Ghosh B, Basu D, Meel KS. Justicia: A stochastic SAT approach to formally verify fairness. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2021, 35(9): 7554–7563.

[39] Shih A, Choi A, Darwiche A. A symbolic approach to explaining Bayesian network classifiers. arXiv: 1805.03364, 2018.

[40] Zhang G, Hou Z, Huang Y, *et al*. Extracting optimal explanations for ensemble trees via logical reasoning. arXiv: 2103.02191, 2021.

[41] Clarke EM, *et al*. eds. Handbook of Model Checking. Cham: Springer, 2018.

[42] Marques-Silva J, Lynce I, Malik S. Conflict-driven Clause Learning SAT Solvers. Handbook of Satisfiability. IOS Press, 2021. 133–182.

[43] Ignatiev A, Narodytska N, Asher N, *et al*. On relating 'Why?' and 'Why Not?' explanations. arXiv: 2012.11067, 2020.

[44] Sülflow A, Fey G, Bloem R, *et al*. Using unsatisfiable cores to debug multiple design errors. In: Proc. of the 18th ACM Great Lakes Symp. on VLSI. 2008. 77–82. [doi: 10.1145/1366110.1366131]

[45] Cheng M, Le T, Chen PY, *et al*. Query-efficient hard-label black-box attack: An optimization-based approach. arXiv:1807.04457, 2018.

[46] Hooker S, Erhan D, Kindermans PJ, *et al*. A benchmark for interpretability methods in deep neural networks. arXiv: 1806.10758, 2018.

附中文参考文献:

[21] 纪守领, 李进锋, 杜天宇, 李博. 机器学习模型可解释性方法、应用与安全研究综述. 计算机研究与发展, 2019, 56(10): 2071–2096. [doi: 10.7544/issn1000-1239.2019.20190540]

[24] 潘文雯, 王新宇, 宋明黎, 陈纯. 对抗样本生成技术综述. 软件学报, 2020, 31(1): 67–81. http://www.jos.org.cn/1000-9825/5884.htm [doi: 10.13328/j.cnki.jos.005884]

[30] 李欣姣, 吴国伟, 姚琳, 张伟哲, 张宾. 机器学习安全攻击与防御机制研究进展和未来挑战. 软件学报, 2021, 32(2): 406–423. http://www.jos.org.cn/1000-9825/6147.htm [doi: 10.13328/j.cnki.jos.006147]

[31] 刘睿瑄, 陈红, 郭若杨, 赵丹, 梁文娟, 李翠平. 机器学习中的隐私攻击与防御. 软件学报, 2020, 31(3): 866–892. http://www.jos.org.cn/1000-9825/5904.htm [doi: 10.13328/j.cnki.jos.005904]

[32] 刘文炎, 沈楚云, 王祥丰, 金博, 卢兴见, 王晓玲, 查宏远, 何积丰. 可信机器学习的公平性综述. 软件学报, 2021, 32(5): 1404–1426. http://www.jos.org.cn/1000-9825/6214.htm [doi: 10.13328/j.cnki.jos.006214]

[33] 化盈盈, 张岱墀, 葛仕明. 深度学习模型可解释性的研究进展. 信息安全学报, 2020, 5(3): 1–12. [doi: 10.19363/J.cnki.cn10-1380/tn.2020.05.01]

马舒岑(1997－), 女, 硕士, 主要研究领域为形式化方法, 机器学习可解释性.

秦胜潮(1974－), 男, 博士, 教授, 主要研究领域为软件理论与形式化方法, 软件工程, 程序语言.

史建琦(1984－), 男, 博士, 副研究员, 博士生导师, 主要研究领域为工业软件, 可信人工智能, 嵌入式控制系统.

侯哲(1988－), 男, 博士, 讲师, 博士生导师, 主要研究领域为自动推理, 形式化验证, 机器学习, 区块链.

黄滟鸿(1986－), 女, 博士, 副研究员, 主要研究领域为可信计算, 形式化建模与验证, 高可信嵌入式控制软件.