控制与决策
Control and Decision

基于代理模型的XAI可解释性量化评估方法

李瑶, 王春露, 左兴权, 黄海, 丁忆宁, 张修建

引用本文:
李瑶,王春露,左兴权,黄海,丁忆宁,张修建. 基于代理模型的XAI可解释性量化评估方法[J]. 控制与决策, 2024, 39(2): 680−688.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.0592

基于代理模型的XAI可解释性量化评估方法

李 瑶[1,3], 王春露[1,3], 左兴权[2,3†], 黄 海[2], 丁忆宁[2,3], 张修建[4,5]

(1. 北京邮电大学 网络空间安全学院，北京 100876；2. 北京邮电大学 计算机学院，北京 100876；
3. 可信分布式计算与服务教育部重点实验室，北京 100876；4. 北京航天计量测试技术研究所，北京 100076；
5. 国家市场监管重点实验室(人工智能计量测试与标准)，北京 100076)

摘 要: 可解释人工智能(explainable artificial intelligence, XAI)近年来发展迅速,已出现多种人工智能模型的解释技术,但是目前缺乏XAI可解释性的定量评估方法. 已有评估方法大多需借助用户实验进行评估,这种方法耗时长且成本高昂. 针对基于代理模型的XAI,提出一种可解释性量化评估方法. 首先,针对这类XAI设计一些指标并给出计算方法,构建包含10个指标的评估指标体系,从一致性、用户理解性、因果性、有效性、稳定性5个维度来评估XAI的可解释性;然后,对于包含多个指标的维度,将熵权法与TOPSIS相结合,建立综合评估模型来评估该维度上的可解释性;最后,将该评估方法用于评估6个基于规则代理模型的XAI的可解释性. 实验结果表明,所提出方法能够展现XAI在不同维度上的可解释性水平,用户可根据需求选取合适的XAI.

关键词: 可解释人工智能；可解释性评估；评估模型；代理模型；规则模型；定量评估

中图分类号: TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.0592

引用格式: 李瑶, 王春露, 左兴权, 等. 基于代理模型的XAI可解释性量化评估方法 [J]. 控制与决策, 2024, 39(2): 680-688.

Quantitative evaluation method for interpretability of XAI based on surrogate model

LI Yao[1,3], *WANG Chun-lu*[1,3], *ZUO Xing-quan*[2,3†], *HUANG Hai*[2], *DING Yi-ning*[2,3], *ZHANG Xiu-jian*[4,5]

(1. School of Cyberspace Security，Beijing University of Posts and Telecommunications，Beijing 100876，China；
2. School of Computer Science，Beijing University of Posts and Telecommunications，Beijing 100876，China；
3. Key Laboratory of Trustworthy Distributed Computing and Service of Ministry of Education，Beijing 100876，China；4. Beijing Aerospace Institute for Metrology and Measurement Technology，Beijing 100076，China；5. Key Laboratory of Artificial Intelligence Measurement and Standards for State Market Regulation，Beijing 100076，China)

Abstract: Explainable artificial intelligence (XAI) is growing rapidly in recent years and many interpretability techniques have emerged, but there is a lack of quantitative evaluation approaches for XAI's interpretability. Most of existing evaluation methods rely on users' experiments, which is time-consuming and costly. Aiming at the surrogate model-based XAI, we propose a quantitative evaluation approach for the XAI's interpretability. Firstly, we devise some indices for this kind of XAI and give their computational method, and construct an index system with 10 quantitative indices to evaluate the XAI's interpretability from five dimensions, namely consistency, user comprehension, causality, effectiveness and stability. For the dimension with multiple indices, a comprehensive evaluation model is established by combining the entropy weight method with TOPSIS to evaluate the XAI's interpretability in the dimension. The proposed approach is applied to the evaluation of the interpretability of 6 XAIs based on the rule surrogate model. Experimental results show that the approach can demonstrate the XAI's interpretability in different dimensions, and users can choose suitable XAI according to their needs.

Keywords: explainable artificial intelligence；interpretability evaluation；evaluation model；surrogate model；rule model；quantitative evaluation

收稿日期: 2022-04-12；**录用日期:** 2022-10-10.
责任编委: 唐加福.
†通讯作者. E-mail: zuoxq@bupt.edu.cn.
*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

0 引 言

近年来,随着人工智能的快速发展,人工智能模型越来越复杂,人们很难理解以深度学习为代表的人工智能模型进行决策所依据的逻辑. 只有当人们理解人工智能模型的决策机理,才能信任其作出的决策,进而将其应用于一些重要领域,如信用评分、医疗保健、自动驾驶、军事应用等.

从可解释性的角度,人工智能模型可分为透明盒(transparent box)模型和黑盒(black box)模型[1]. 透明盒模型的决策机理是透明的,本身具有可解释性,如决策树模型、规则模型、线性模型等. 而黑盒模型的决策机理不透明,人们很难了解其决策过程,如神经网络、支持向量机等. 针对黑盒模型,近年来出现了多种解释方法,如LIME[2]、SHAP[3]、Salience maps[4]等. 这些解释方法将黑盒模型变为可解释的人工智能(explainable artificial intelligence, XAI)[5].

XAI的可解释性依赖于这些解释方法. 若其能够很好地解释黑盒模型,则XAI具有良好的可解释性;否则,难以保证XAI的可靠性和可信性. 因此,XAI的可解释性评估具有重要意义,近年来成为研究热点. 而XAI的可解释性涉及多领域的交叉[6]和人的主观感受,其评估具有很大挑战性.

由于XAI的可解释性与人的主观感受有关,当前评估方法大多依赖于人的主观评价[7-9],通过设计用户实验和量表来定性或定量评估XAI解释性. 这种评估方法存在成本高、耗时长等缺点. 已有少量工作研究了XAI可解释性的客观评估方法[10-11].

在实际应用中,一个黑盒模型可用多种解释方法对其解释. 即使利用一种解释方法对黑盒模型进行解释,由于解释方法的不稳定,对该黑盒模型也会产生多个解释性不尽相同的个体XAI(individual explainable artificial intelligence). 个体XAI是指利用解释方法对黑盒模型进行解释时所生成的具体XAI,如基于决策树代理模型的解释方法对同一黑盒模型进行多次解释,会产生不同的决策树模型,每个决策树代理模型与黑盒模型构成的个体XAI的解释性水平均不尽相同. 因此,除了评估解释方法的解释性,也需评估个体XAI的可解释性. 然而,当前的客观评估方法大多针对解释方法进行评估,缺乏针对个体XAI解释性评估研究. 此外,已有客观评估方法大多用于评估黑盒模型的局部解释性,缺乏对黑盒模型的全局解释性进行客观评估的研究.

本文针对基于全局代理模型的XAI,提出一种XAI可解释性的多维度量化评估方法(evaluation method for interpretability of XAI based on surrogate model, EMXS). 代理模型是一种典型的解释方法,它利用黑盒模型输入和输出构建可解释的机器学习模型(代理模型)[12],实现对黑盒模型的解释. 代理模型不受限于黑盒的内部结构和实现原理,是一种应用广泛的模型无关的解释方法.

本文所提出评估方法旨在评估个体XAI的可解释性,而非评估解释方法,其目的是帮助用户选取解释性好的XAI. 针对个体XAI,本文构建包含10个指标的指标体系,从一致性、用户理解性、因果性、有效性、稳定性5个维度来评估其可解释性. 其中,用户理解性和有效性维度各包含多个指标,本文引入综合评估模型来融合这些指标对其进行量化评估. 将所提出评估方法用于评估6个基于规则全局代理模型的XAI,验证该方法的有效性.

本文的主要内容如下.

1) 针对基于全局代理模型的XAI,提出从一致性、用户理解性、因果性、有效性、稳定性5个维度来量化评估可解释性的方法.

2) 针对基于规则全局代理模型的XAI,构建其可解释性评估指标体系,其中除包含通用的指标外,还包含针对规则模型设计的指标:矛盾率、平均绝对因果效应、充分性、规则有效性、规则项有效性、稳定性.

3) 针对用户理解性和有效性维度,引入综合评估模型来融合多个评估指标对其进行量化评估. 目前还没有利用综合评估模型来评估XAI可解释性的研究.

4) 将所提出方法用于评估6个基于规则全局代理模型的XAI的可解释性,并对评估结果进行对比分析,验证所提出评估方法的有效性.

1 相关工作

目前,人工智能可解释性评估还没有形成标准的、普遍认可的体系[13].

1.1 XAI的代理模型解释技术

根据美国国防部高级研究计划局(defense advanced research projects agency, DARPA)对XAI解释方法的分类[14],解释方法可分为深度解释、可解释模型、模型归纳. 基于代理模型的解释方法属于模型归纳方法[15].

代理模型解释方法通过获取黑盒模型的可解释代理模型来理解黑盒模型的决策机制和逻辑. 目前已有很多基于代理模型的解释方法. 一些学者研究了黑盒模型的局部代理模型,在样本周围构建具有局部保真度的可解释代理模型来解释模型针对样本的决策机理[16]. LIME[2]是一种典型的局部代理模型

方法,其构建局部线性模型对复杂黑盒模型进行拟合. Ribeiro等[2]进一步提出了一种基于规则的局部代理模型Anchor[16]. 此外,还有LORE[17]、LEMNA[18]等局部代理模型.

一些学者研究了黑盒模型的全局代理模型. 全局代理模型是指训练一个可解释机器学习模型来拟合整个黑盒模型的决策行为[14],以帮助用户理解黑盒内部的整体决策逻辑. 针对深度神经网络模型,文献[19]利用决策树建立其全局代理模型. 此外,一些学者建立了基于规则模型[20]、逻辑回归模型[21]等全局代理模型.

1.2 XAI的可解释性评估方法

可解释性评估方法主要分为2类:以人为中心的评估、客观评估[22]. 以人为中心的评估是指通过用户或专家的反馈对解释性进行评估,客观评估是指用客观的评估指标对解释性进行量化评估.

当前大多数可解释性评估方法基于用户反馈进行评估. 一项重要工作是Hoffman等[23]通过建立XAI解释过程的概念模型,从解释的优良、用户满意度、用户理解性、用户信任和依赖以及好奇心的影响等方面对可解释性进行评估,并对用户实验设计给出了具体的建议和示例. 文献[24]基于医生的反馈评估了LIME在临床领域的可解释性,考虑了医生对模型预测的接受程度、LIME提供的解释与医生对模型预测的解释的一致性以及医生对LIME的信任和依赖程度等因素. 文献[8]提出了一种可解释人工智能设计和评估框架,从用户心理模型、用户信任和信赖、解释有用性和满意度、人机任务性能、计算测度5个方面对可解释性进行评价. 文献[25]通过在MTurk (amazon mechanical turk)众包平台召集志愿者,从用户理解性、用户对模型不确定性的识别、用户信任3个方面对可解释性进行评估.

用户实验方法能够评估XAI的解释性,但是存在评估成本高、耗时长等不足. 用户实验所基于的假设是实验能够恰当地捕捉到用户表现与解释间的复杂动态联系,且好的解释能够提升用户的表现. 然而,文献[9]通过一项涉及3 800名参与者的研究表明:清晰、详细的解释反而会损害用户表现.

相比以人为中心的评估方法,可解释性的客观量化评估方法还较少[22]. 局部解释技术方面:文献[26]针对基于局部代理模型的解释方法,对解释方法的稳定性、鲁棒性和解释内容的有效性进行了量化评估;文献[10]针对显著性图解释技术,从忠诚度、解释方法的定位能力、敏感性和稳定性4个方面对可解释性进行了量化评估;文献[11]设计了2组实验来评估LIME的局部解释的充分性和显著性. 全局解释技术方面:文献[27]利用决策树全局代理模型对CNN进行了解释,通过对CNN进行控制和调节来计算代理模型的特征信息增益、特征稀疏性、特征完整性、决策树的预测准确性、完整性5个方面对可解释性进行量化评估,是一种针对特定黑盒模型的评估方法.

一些研究关注特定黑盒模型或应用场景中XAI可解释性的量化评估. 文献[28]研究恶意软件检测和漏洞发现领域可解释性量化评估方法. 还有一些工作对解释技术的特定性能进行量化评估,如文献[29]采用基于扰动的方法设计了2个量化指标对解释方法的敏感性和保真度进行评估.

综上,已有客观评估方法大多针对解释方法进行评估,缺乏针对个体XAI的解释性评估方法. 已有方法大多对局部解释进行评估,缺乏对全局解释的评估. 一些评估方法只针对特定黑盒模型和应用场景进行评估. 不同于以上量化评估方法,本文针对个体XAI,提出一种基于全局代理模型的XAI可解释性量化评估方法,为用户选择满足其需求的XAI提供依据.

2 基于规则代理模型的XAI

代理模型解释方法是一种通用解释方法,与黑盒模型种类和内部机理无关,通过构建黑盒模型的局部或全局代理模型来解释黑盒模型的决策依据和机理. 代理模型为透明盒模型,与黑盒模型组成XAI. 规则模型与人类的决策方式相似,具有简单、直观、易于理解的特点[30],是一种典型的代理模型,目前已出现多种基于规则代理模型的解释技术,包括Anchor[16]、LORE[17]、MUSE[20]等. 为此,本文针对基于规则代理模型的XAI,研究其可解释性量化评估方法.

下面以分类问题为例,介绍黑盒模型的规则代理模型的构建过程. 假设训练黑盒模型的数据集D为$D=\{(x_i,y_i)|x_i\in X,y_i\in Y\}$. 其中:$(x_i,y_i)$为第$i$个样本,$i\in\{1,2,\dots,n\}$;$x_i$为输入黑盒模型的特征向量;$X$为输入空间;$y_i$为样本标签;$Y$中包含$N$个类别标签. 利用数据集$D$构建黑盒模型$B$. 对于每个样本$x$,假设黑盒模型的预测结果$B(x)=y'$,则得到黑盒模型的输入和输出数据集合:$D'=\{(x,y')|x\in X,y'=B(x)\}$. 利用$D'$训练得到一个规则代理模型$S_B$,用其拟合黑盒模型,使得$S_B$的预测结果与黑盒模型$B$的预测结果尽可能一致.

规则代理模型S_B主要有2种形式,即规则集和规则列表. 规则集由相互独立的一组规则组成. 规则

列表中的每条规则按照IF-ELSE的形式组织.

规则代理模型S_B包含一个规则集合R,通过这个规则集合来解释整个黑盒模型的决策行为. 对于黑盒模型的每个输入产生的决策行为,代理模型可提供一条或多条规则对其进行解释. 规则集R的第i条规则r_i可表示为(q_i, c_i). 其中:q_i为由r_i中的多个规则项(rule item)组成的规则项集,c_i为规则r_i预测的类别标签. 令$C=\{c_1, c_2, \ldots, c_{N'}\} \subseteq Y$为规则代理模型的输出空间. C中包含N'个类别标签且$N' \leqslant N$. 对于一个样本(x, y),假设其被解释为代理模型S_B中的第i个规则r_i,则可表示为$S_B(x, y)=r_i$.

3　XAI可解释性的量化评估方法

针对基于全局代理模型的个体XAI,根据美国国家标准技术研究院(national institute of standards and technology, NIST)提出的可解释人工智能需要具备4个原则[31],本文从一致性、用户理解性、有效性、因果性、稳定性5个维度对XAI可解释性进行评估. 根据文献[32]中的指标体系构建的6个基本原则,筛选各维度的指标并设计新指标,形成包括10个指标的指标体系,用于量化评估XAI在5个维度上的可解释性. 所提出方法的评估过程如图1所示. 评估对象XAI包括黑盒模型和规则代理模型. 首先,从XAI中提取5个维度的指标. 其中,用户理解性和有效性维度包含多个指标,利用综合评估模型融合这些指标. 综合评估模型首先利用熵权法获取指标的权重,然后利用TOPSIS综合评价方法得到综合评估结果. 最终得到该XAI在5个维度上的可解释性评估结果.

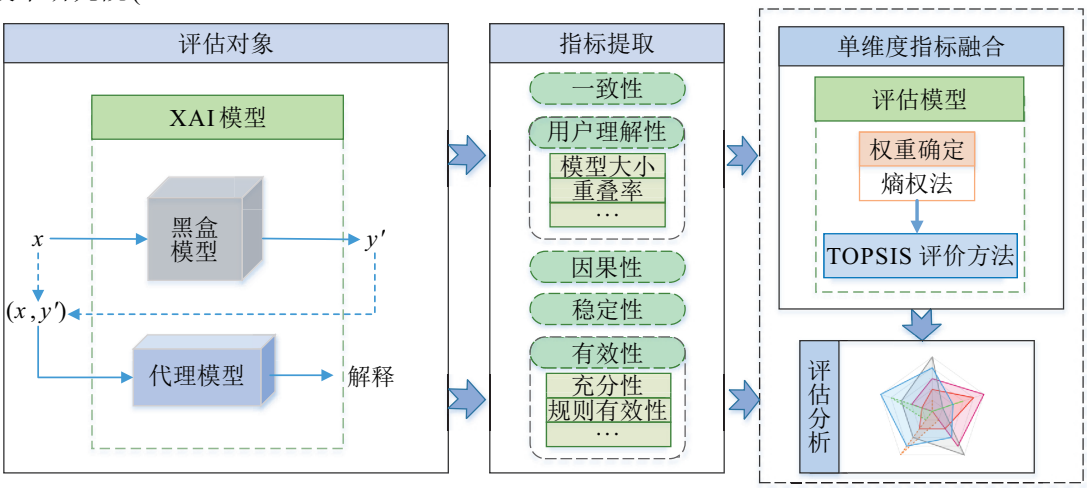

图 1　XAI可解释性的评估过程

3.1　可解释性评估指标体系

本文针对基于全局代理模型的XAI,构建的评估指标体系如图2所示.

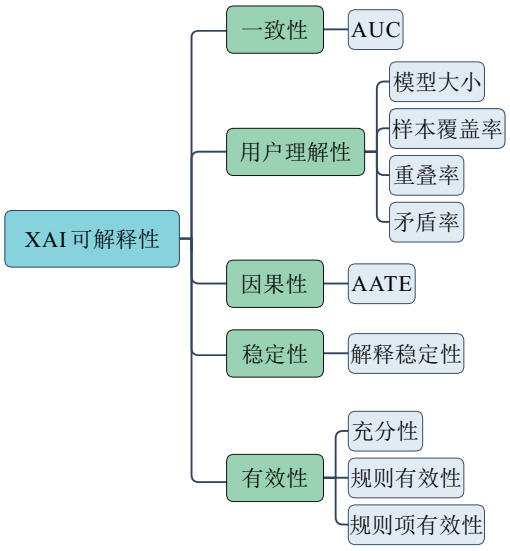

图 2　XAI可解释性评估指标体系

具体指标如下.

1) 一致性.

一致性是代理模型拟合黑盒模型的程度,这是代理模型对黑盒模型进行解释的基础. 若代理模型与黑盒模型没有足够的一致性,则其不能有效地解释黑盒模型. 当前普遍采用代理模型和黑盒模型的预测结果偏差来评估一致性,指标包括accuracy、F1分数、AUC分数、RMSE等[1,19]. 考虑到accuracy和F1不适用于数据集中类别分布不均衡情况,对于二分类问题,本文采用AUC分数作为一致性指标;对于多分类问题,采用宏平均AUC(macro-AUC)指标.

2) 用户理解性.

用户理解性是指用户对解释的理解程度. 代理模型的复杂程度会影响用户对模型决策机制的理解,代理模型需具有合理的规模才能提供有效的解释[6]. 对于规则代理模型的复杂性,目前的研究主要从模型大小、最大长度、总长度、平均长度等方面进行

评估[21],其中,模型大小指的是规则代理模型中规则的数量,是较广泛采用的评估指标. 因此,本文采用模型大小对解释的复杂性进行量化评估.

此外,解释的明确程度会影响用户对解释的理解,明确的解释为黑盒模型的各种行为提供清晰的解释[20]. 关于规则代理模型的明确程度,目前的研究主要从样本覆盖率、标签覆盖率、重叠率等方面进行评估[16]. 本文采用样本覆盖率、重叠率2个常用的量化指标,并提出矛盾率指标共3个指标来对解释的明确程度进行量化评估.

样本覆盖率是指规则模型对样本的覆盖率. 当一个样本的特征部分能够匹配到至少一条规则时,称该样本被规则覆盖. 令D'_i为被规则r_i覆盖的样本集合,表示为$D'_i=\{(x,y')|x\text{被}r_i\text{覆盖},y'\in C\}$,则被所有规则覆盖的数据集合表示为$D_c=\bigcup\limits_{r_i\in R}D'_i$. 本文用数据集中被所有规则覆盖的样本占总样本的比例来度量样本覆盖率,有

$$I_c(R)=\frac{|D_c|}{|D'|}. \tag{1}$$

若一个样本被多个规则覆盖,则该样本可被多个规则解释,会导致解释的不唯一. 本文利用被多条规则覆盖的样本数占被规则模型覆盖的所有样本数的比例来度量重叠率,有

$$O_c(R)=\frac{|\{(x,y)|(x,y)\in D'_i,(x,y)\in D'_j,i\neq j\}|}{|D_c|}. \tag{2}$$

当一个样本被多个规则覆盖时,若这些规则得到不同的预测结果,则这些规则会产生相互矛盾的解释,从而影响用户对规则模型解释的信任. 本文用规则模型覆盖的所有样本中,得到矛盾预测结果的样本所占比例来度量矛盾率,有

$$C_c(R)=\frac{|\{(x,y)|(x,y)\in D'_i,(x,y)\in D'_j,c_i\neq c_j\}|}{|D_c|}, \tag{3}$$

其中c_i为规则模型中规则r_i预测的类别标签.

综上,针对基于规则代理模型的XAI,本文通过模型大小、样本覆盖率、重叠率、矛盾率4个指标来量化评估XAI的用户理解性.

3) 因果性.

因果性是指生成的解释与黑盒模型预测的因果关系水平. 因果关系是指解释中包含的特征是预测结果的原因,如气温升高会导致冰淇淋店的用电量和销售额的增加,气温与销售额有因果关系;若对销售额进行预测,则希望生成的解释能够提取与销售额有因果关系的特征(气温),而不希望提取那些与销售额相关却无因果关系的特征(用电量)[33]. 代理模型提取越多与预测结果有因果性的特征,其解释性越好. 为了评估解释的因果性,目前的研究工作主要通过计算ATE (average treatment effect)、CATE (conditional average treatment effect)等指标进行评估[34-35].

对于规则代理模型,在其他规则项保持不变的条件下,对规则代理模型所用到的每个规则项分别进行扰动,若预测结果发生改变,则该规则项与预测结果有因果效应,本文采用常用的ATE来计算规则项的因果效应. 对于规则代理模型用到的每个规则项T,本文用T-learner算法[36]计算T的ATE. 该算法包括2个步骤:①基于树的方法估计T分别取0和1时预测结果改变的条件期望;②计算2个条件期望的差值作为T的ATE.

本文利用所有规则项因果效应的绝对值的均值AATE (average absolute treatment effect)来评价该规则代理模型的整体因果效应.

4) 稳定性.

稳定性是指XAI对样本进行解释时生成解释的稳定性. 对于同一样本或相似的样本,若生成具有较大差异的解释,则会影响用户对XAI的信任. 如自动驾驶领域中,若在行驶情况没有发生明显变化时,XAI向用户提供几种不同的解释,则用户会对自动驾驶系统失去信任[10]. 在已有的解释方法评估中,对于同一样本用解释方法多次重复解释,通过多次解释的相似性来评估解释方法的稳定性[17,26]. 解释的相似性通常利用Jaccaed系数进行评估[17]. 对于基于规则代理模型的个体XAI,同一样本会匹配同一个规则. 为此,针对个体XAI,本文通过对样本施加扰动前后,个体XAI对样本进行解释的相似性来评估其解释的稳定性. 其中,解释相似性利用Jaccard系数进行评估.

针对规则代理模型,本文从覆盖的样本中随机抽取20%的样本进行扰动,通过计算样本扰动前后所匹配规则中的规则项集合间的Jaccard系数的平均值作为解释的稳定性. 令D_{stab}为随机抽取的样本集合,假设其中一个样本(x,y)经过扰动后的样本为(x_{pert},y). 其中:$x_{\text{pert}}=x\pm\text{rand}(\boldsymbol{e}\cdot k)$,rand( )为随机函数,向量$\boldsymbol{e}$为样本各特征值的值域大小(domain size),即取值上界与下界的差值,k为扰动幅度. 对于D_{stab}中的一个样本(x,y),设其匹配的规则为$r_i=S_B(x,y)$,扰动后的样本(x_{pert},y)匹配的规则为$r_j=S_B(x_{\text{pert}},y)$,若样本匹配多条规则,则从这些规则中

选择对所覆盖样本进行预测时准确率最高的规则作为该样本匹配的规则. 解释的稳定性表示为

$$S_c(R)=\frac{1}{|D_{\text{stab}}|}\sum_{\substack{(x,y)\in D_{\text{stab}}\\ r_i=S_B(x,y)\\ r_j=S_B(x_{\text{pert}},y)}}\frac{|q_i\bigcap q_j|}{|q_i\bigcup q_j|},\tag{4}$$

其中q_i为规则模型中规则r_i的规则项集.

5) 有效性.

有效性是指解释能够准确地反映黑盒的决策逻辑. 有效的解释应充分包含黑盒模型预测时所依据的信息. 在图像分类领域,文献[11]将LIME解释技术生成的图像解释输入至黑盒模型后,通过观察预测结果是否改变来评估图像解释的充分性. 受文献[11]启发,本文根据规则代理模型中样本匹配的规则来构造新样本,输入至黑盒模型,通过观察预测结果是否改变来评估解释的充分性.

针对规则代理模型,所提出的充分性指标计算方法如下.

① 在每条规则覆盖的样本集合中,随机选取10%的样本作为实验样本,令$D_{\text{suf}}(r_i)$为规则r_i覆盖的实验样本集合;

② 基于每个实验样本(x,y)产生新样本(x_{mut},y),使得x_{mut}中,样本(x,y)所匹配的规则中各规则项对应的特征项的值保持不变,其余特征项的值为0;

③ 将样本(x_{mut},y)输入至黑盒模型,得到预测结果$B(x_{\text{mut}})$;

④ 计算该规则构建的新样本中,输入黑盒模型后与代理模型预测结果一致的样本所占的比例,作为该规则的充分性;

⑤ 将各条规则的充分性进行加权组合,作为规则代理模型的充分性,其中,每条规则的权重为该规则覆盖的样本数量与代理模型覆盖的所有样本数量的比值,用向量$\boldsymbol{W}_R$表示各规则的权重.

解释的充分性表示为

$$E_c(R)=\sum_{r_i\in R}w_i\cdot\frac{\sum\limits_{(x,y)\in D_{\text{suf}}(r_i)}\text{eql}(c_i,B(x_{\text{mut}}))}{|D_{\text{suf}}(r_i)|}.\tag{5}$$

其中: c_i为规则r_i的类别标签; $w_i\in\boldsymbol{W}_R$为第i个规则的权重; eql( )函数为

$$\text{eql}(c_i,c_j)=\begin{cases}1,\ c_i=c_j;\\ 0,\ c_i\neq c_j.\end{cases}\tag{6}$$

为评估解释中特征的有效性,文献[26]通过依次移除样本中解释包含的特征来构建新样本,观察黑盒模型对其预测结果是否改变来计算特征的有效性. 受文献[26]的启发,本文提出规则有效性、规则项有效性2个指标,分别评估代理模型中规则和规则项的有效性,计算方法如下.

将原测试集中样本标签替换为黑盒模型预测的标签构成新测试集,依次移除每条规则/规则项,构成新的规则模型. 若新规则模型在新测试集上的预测精度下降,则表明该规则/规则项是有效的. 代理模型中,有效的规则占全部规则的比例作为规则有效性指标. 每条规则的规则项有效性指标值为该规则中有效规则项占全部规则项的比例,将各条规则的规则项有效性指标值与规则权重$\boldsymbol{W}_R$进行加权求和,作为代理模型的规则项有效性.

综上,本文通过计算充分性、规则有效性、规则项有效性3个指标来量化评估XAI的有效性.

3.2 综合评估模型

XAI的用户理解性和有效性维度中包含多个指标. 为了量化评估XAI在这2个维度上的解释性,本文建立了综合评估模型. 分别将用户理解性和有效性维度的多个指标值输入评估模型,模型输出为各XAI在这2个维度上的可解释性排序.

对于每个维度,本文利用熵权法来确定各指标的权重. 然后,将指标权重与TOPSIS评价方法相结合,建立综合评估模型,对解释性进行综合评估.

4 实验与分析

为验证EMXS的有效性,本文在真实数据集上训练黑盒模型,然后利用基于规则的全局代理模型解释技术对其进行解释,生成多个XAI模型,每个XAI模型由1个黑盒模型和1个可解释的代理模型组成. 对所生成的XAI进行评估和分析,以帮助用户结合实际需求选取解释性好的XAI.

4.1 数据集

采用UCI公开数据库[37]中的典型标准数据集建立黑盒模型. 威斯康辛乳腺癌数据集(breast cancer)包含了威斯康辛州的569个乳腺癌病例数据,每个病例包括30个生理指标数据和1个恶性/良性类别数据.

4.2 XAI模型

在数据集上训练2个黑盒模型,即支持向量机(SVM)和多层感知机(MLP). MLP包含3个隐层,每层神经元数量分别为100、50、50. 将数据集随机划分为2部分: 70%的数据作为训练集用于训练黑盒模型,30%的数据作为测试集.

对训练后的黑盒模型,建立其规则代理模型. 选取3个流行的规则提取方法来建立黑盒模型的3个

规则代理模型. 3 种规则提取方法为: 1) SBRL (scalable Bayesian rule lists)[38]; 2) BRS (Bayesian rule sets)[39]; 3) MDL (MDL-based rule lists)[40]. SBRL 和 MDL 产生 1 组结构化的规则列表, 而 BRS 生成由 1 组互不相关的规则组成的规则集. 2 种黑盒模型与 3 种规则提取方法得到的规则代理模型组合为 6 个 XAI, 即 SBRL-SVM、BRS-SVM、MDL-SVM、SBRL-MLP、BRS-MLP、MDL-MLP.

4.3 实验结果

对于乳腺癌数据上的 6 个 XAI 模型, 提取每个 XAI 的指标如表 1 所示. 稳定性指标与扰动程度相关, 后文将单独分析. 由表 1 可见, 6 个 XAI 模型的一致性指标均在 90 % 以上. 其中: BRS-SVM 的一致性水平最高, 达到 96.8 %; MDL-MLP 具有较低因果性, 表明其解释难以体现真正影响黑盒决策的原因. 用户理解性和有效性维度包含多个指标, 综合评估模型的评估结果如表 2 所示. 其中: BRS 规则提取算法相关的 XAI 的用户理解性较低, 这是因为 BRS 生成的规则集中各规则相互独立, 一个样本可能被多条规则覆盖而具有较高重叠率; MDL-SVM 的有效性较低, 这是因为其充分性较低且代理模型中包含无效规则和大量无效规则项.

表 1 实验中 6 个 XAI 的可解释性指标值

XAI	一致性	用户理解性				因果性	有效性		
		模型大小	样本覆盖率	重叠率	矛盾率		充分性	规则有效性	规则项有效性
MDL-SVM	0.949	4	0.556	0.0	0.0	0.166	0.315	0.75	0.170
MDL-MLP	0.922	3	0.450	0.0	0.0	0.148	0.818	1.0	1.0
BRS-SVM	0.968	3	0.596	0.450	0.0	0.401	0.902	1.0	0.775
BRS-MLP	0.912	4	0.708	0.515	0.0	0.237	0.083	1.0	1.0
SBRL-SVM	0.946	5	0.725	0.0	0.0	0.184	0.892	0.4	0.70
SBRL-MLP	0.924	6	0.871	0.0	0.0	0.298	0.779	1.0	1.0

表 2 用户理解性和有效性的评估结果

XAI	用户理解性				有效性			
	D^+	D^-	C	次序	D^+	D^-	C	次序
MDL-SVM	0.086	0.216	0.714	2	0.185	0.071	0.280	6
MDL-MLP	0.088	0.249	0.738	1	0.018	0.223	0.924	1
BRS-SVM	0.170	0.172	0.502	5	0.034	0.221	0.866	3
BRS-MLP	0.195	0.124	0.389	6	0.182	0.151	0.454	5
SBRL-SVM	0.116	0.201	0.634	3	0.097	0.197	0.672	4
SBRL-MLP	0.168	0.204	0.549	4	0.027	0.217	0.889	2

XAI 的稳定性与对样本的扰动幅度相关, 将扰动幅度 k 从 0 逐步递增至 50 %, 计算各 XAI 的稳定性值, 如图 3 所示. 随着扰动幅度的增加, XAI 的稳定性逐步降低. 扰动程度小于 10 % 时, 各 XAI 的稳定性没有明显差异; 当扰动程度大于 10 % 时, MDL-MLP 的稳定性较高, MDL-SVM 次之, SBRL-MLP 与 BRS-MLP 稳定性相当, SBRL-SVM 具有较低稳定性.

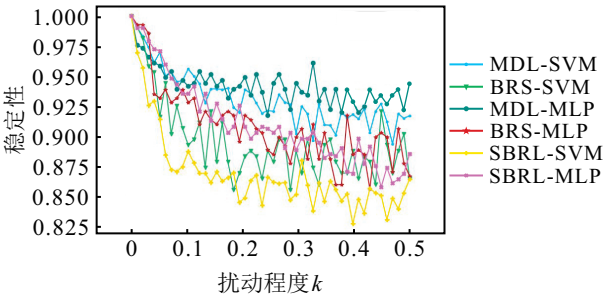

图 3 各 XAI 的稳定性

本文对各 XAI 在各维度上的解释性次序进行 min-max 标准化处理, 使得排序结果映射至 [0, 1] 区间, 用雷达图展示, 如图 4 所示. BRS-MLP 和 SBRL-SVM 在各维度上的解释性明显劣于其他 XAI 模型. MDL-SVM 在稳定性、用户理解性、一致性维度上解释性较好, 而在有效性和因果性维度上解释性较差. MDL-MLP 在稳定性、有效性、一致性维度上表现最好, 但是在因果性和用户理解性维度上表现较差. SBRL-MLP 具有较高的有效性、因果性、稳定性以及中等水平的用户理解性和一致性. BRS-SVM 有最好的用户理解性、因果性, 而其稳定性和一致性较差.

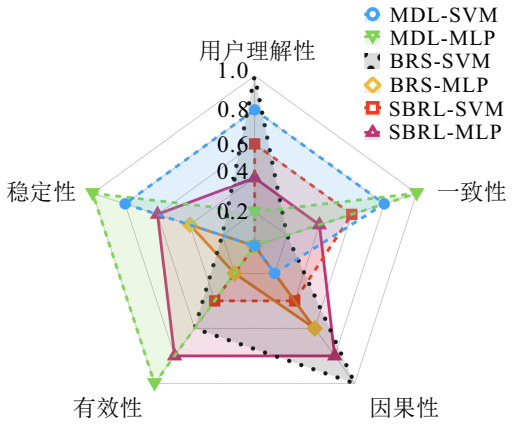

图 4 6 个 XAI 的可解释性雷达图

5 结 论

当前人工智能可解释性客观评估方法大多针对解释方法进行评估, 缺乏针对个体 XAI 的评估方法.

本文针对基于全局代理模型的个体XAI,提出了一种XAI可解释性的多维度量化评估方法(EMXS). 首先,针对这类XAI,设计了矛盾率、平均绝对因果效应、充分性、规则有效性、规则项有效性、稳定性指标,并将设计的指标与已有指标相结合,形成包含10个指标的指标体系,从一致性、用户理解性、因果性、有效性、稳定性5个维度来评估其可解释性;然后,利用熵权法、TOPSIS综合评价方法建立综合评估模型来融合用户理解性和有效性维度上的多个指标;最后,将EMXS用于评估6个XAI的可解释性. 实验结果表明,EMXS能够有效地评估XAI的可解释性水平,为人们选取符合需求的XAI提供依据.

未来研究工作包括: 1) 进一步研究基于其他代理模型的XAI可解释性评估方法; 2) 结合个体XAI评估方法与面向解释方法的评估方法,对XAI可解释性进行更全面的评估.

参考文献(References)

[1] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models[J]. ACM Computing Surveys, 2018, 51(5): 1-42.

[2] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier[C]. Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining. NewYork, 2016: 1135-1144.

[3] Lundberg S M, Lee S-I. A unified approach to interpreting model predictions[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems. Cambridge, 2017: 4768-4777.

[4] Smilkov D, Thorat N, Kim B, et al. SmoothGrad: Removing noise by adding noise[J/OL]. 2017, arXiv: 1706.03825.

[5] van Lent M, Fisher W, Mancuso M. An explainable artificial intelligence system for small-unit tactical behavior[C]. Proceedings of the National Conference on Artificial Intelligence. San Jose, 2004: 900-907.

[6] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)[J]. IEEE Access, 2018, 6: 52138-52160.

[7] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning[J/OL]. 2017, arXiv: 1702.08608.

[8] Mohseni S, Zarei N, Ragan E D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems[J]. ACM Transactions on Interactive Intelligent Systems, 2021, 11(3/4): 1-45.

[9] Poursabzi-Sangdeh F, Goldstein D G, Hofman J M, et al. Manipulating and measuring model interpretability[C]. Proceedings of the CHI Conference on Human Factors in Computing Systems. New York, 2021: 1-52.

[10] Li X H, Shi Y H, Li H Y, et al. An experimental study of quantitative evaluations on saliency methods[C]. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York, 2021: 3200-3208.

[11] Shah S S, Sheppard J W. Evaluating explanations of convolutional neural network image classifications[C]. International Joint Conference on Neural Networks. Barcelona, 2020: 1-8.

[12] Rai A. Explainable AI: From black box to glass box[J]. Journal of the Academy of Marketing Science, 2020, 48(1): 137-141.

[13] 纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096.
(Ji S L, Li J F, Du T Y, et al. Survey on techniques, applications and security of machine learning interpretability[J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096.)

[14] Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program[J]. AI Magazine, 2019, 40(2): 44-58.

[15] 成科扬, 王宁, 师文喜, 等. 深度学习可解释性研究进展[J]. 计算机研究与发展, 2020, 57(6): 1208-1217.
(Cheng K Y, Wang N, Shi W X, et al. Research advances in the interpretability of deep learning[J]. Journal of Computer Research and Development, 2020, 57(6): 1208-1217.)

[16] Ribeiro M T, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, 2018: 1527-1535.

[17] Guidotti R, Monreale A, Giannotti F, et al. Factual and counterfactual explanations for black box decision making[J]. IEEE Intelligent Systems, 2019, 34(6): 14-23.

[18] Guo W, Mu D, Xu J, et al. Lemna: Explaining deep learning based security applications[C]. Proceedings of the ACMSIGSAC Conference on Computer and Communications Security. New York, 2018: 364-379.

[19] Schaaf N, Huber M, Maucher J. Enhancing decision tree based interpretation of deep neural networks through L1-orthogonal regularization[C]. The 18th IEEE International Conference on Machine Learning and Applications. Boca Raton, 2019: 42-49.

[20] Lakkaraju H, Kamar E, Caruana R, et al. Faithful and customizable explanations of black box models[C]. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, 2019: 131-138.

[21] Tian Y, Liu G J. MANE: Model-agnostic non-linear explanations for deep learning model[C]. IEEE World Congress on Services. Beijing, 2020: 33-36.

[22] Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence[J]. Information Fusion, 2021, 76: 89-106.

[23] Hoffman R R, Mueller S T, Klein G, et al. Metrics for explainable AI: Challenges and prospects[J/OL]. 2018, arXiv: 1812.04608.

[24] Barr K N, Blomberg T, Liu J T, et al. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models[C]. IEEE the 33rd International Symposium on Computer-Based Medical Systems. Rochester, 2020: 7-12.

[25] Wang X R, Yin M. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making[C]. The 26th International Conference on Intelligent User Interfaces. New York, 2021: 318-328.

[26] Fan M, Wei W Y, Xie X F, et al. Can we trust your explanations? Sanity checks for interpreters in android malware analysis[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 838-853.

[27] Fan L X, Liu C, Zhou Y H, et al. Interpreting and evaluating black box models in a customizable way[C]. IEEE International Conference on Big Data. Atlanta, 2021: 5435-5440.

[28] Warnecke A, Arp D, Wressnegger C, et al. Evaluating explanation methods for deep learning in security[C]. IEEE European Symposium on Security and Privacy. Genoa, 2020: 158-174.

[29] Yeh C K, Hsieh C Y, Suggala A, et al. On the (in) fidelity and sensitivity of explanations[C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, 2019: 10967-10978.

[30] 周志杰, 曹友, 胡昌华, 等. 基于规则的建模方法的可解释性及其发展[J]. 自动化学报, 2021, 47(6): 1201-1216.
(Zhou Z J, Cao Y, Hu C H, et al. The interpretability of rule-based modeling approach and its development[J]. Acta Automatica Sinica, 2021, 47(6): 1201-1216.)

[31] Phillips P J, Hahn C A, Fontana P C, et al. Four principles of explainable artificial intelligence[J]. NIST Interagency/Internal Report, National Institute of Standards and Technology, Gaithersburg, MD, DOI: 10.6028/NIST.IR.8312,2020.

[32] 彭张林, 张爱萍, 王素凤, 等. 综合评价指标体系的设计原则与构建流程[J], 科研管理, 2017, 38(S1): 209-215.
(Peng Z L, Zhang A P, Wang S F, et al. Designing principles and constructing processes of the comprehensive evaluation indicator system[J]. Science Research Management, 2017, 38(S1): 209-215.)

[33] Guo R C, Cheng L, Li J D, et al. A survey of learning causality with data: Problems and methods[J]. ACM Computing Surveys, 2021, 53(4): 1-37.

[34] Tian L, Alizadeh A A, Gentles A J, et al. A simple method for estimating interactions between a treatment and a large number of covariates[J]. Journal of the American Statistical Association, 2014, 109(508): 1517-1532.

[35] Hahn P R, Murray J S, Carvalho C M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)[J]. Bayesian Analysis, 2020, 15(3): 965-1056.

[36] Künzel S R, Sekhon J S, Bickel P J, et al. Metalearners for estimating heterogeneous treatment effects using machine learning[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(10): 4156-4165.

[37] Dua D, Graff C. UCI machine learning repository[D]. Irvine: University of California, 2019.

[38] Yang H Y, Rudin C, Seltzer M. Scalable Bayesian rule lists[C]. Proceedings of the 34th International Conference on Machine Learning. New York, 2017: 3921-3930.

[39] Wang T, Rudin C, Doshi-Velez F, et al. A Bayesian framework for learning rule sets for interpretable classification[J]. J Mach Learn Res, 2017, 18(1): 2357-2393.

[40] Proença H M, van Leeuwen M. Interpretable multiclass classification by MDL-based rule lists[J]. Information Sciences, 2020, 512: 1372-1393.

作者简介

李瑶 (1996−), 女, 硕士生, 从事可解释人工智能、可解释性评估方法等研究, E-mail: 1934545502@qq.com;

王春露 (1969−), 女, 教授, 硕士, 从事人工智能、信息安全等研究, E-mail: wangcl@bupt.edu.cn;

左兴权 (1971−), 男, 教授, 博士, 从事智能优化、人工智能、数据挖掘、智能交通等研究, E-mail: zuoxq@bupt.edu.cn;

黄海 (1979−), 男, 讲师, 博士, 从事数据挖掘、机器学习等研究, E-mail: hhuang@bupt.edu.cn;

丁忆宁 (1998−), 男, 硕士生, 从事人工智能可解释性、人工智能鲁棒性等研究, E-mail: dingyn@bupt.edu.cn;

张修建 (1984−), 男, 高级工程师, 从事人工智能安全可信计量技术的研究, E-mail: zxjwell@163.com.