带特征选择的综合因果多目标反事实解释方法

刘金平,汤浩楠,李兴旺,徐鹏飞*,袁晟玮
(湖南师范大学信息科学与工程学院,湖南长沙 410081)

摘 要: 随着复杂机器学习模型应用扩展,各行业对模型可解释性的需求剧增. 反事实解释是重要的事后可解释方法,但传统方法常将多目标合并为单目标优化,导致权重分配困难且难以调和目标冲突,也因忽略因果关系使生成的反事实样本不现实. 此外,现有方法在高维、冗余、噪声数据下存在计算效率低、预测精度下降及全局解释不足等问题. 为此,本文提出综合因果多目标反事实解释方法(Comprehensive Causal multi-objective counterfactual Explanation with Feature Selection,CCE-FS). 该方法首先基于最大互信息系数筛选关键特征以提升预测精度和全局解释力,然后将反事实搜索转化为多目标优化问题,有效平衡多目标关系. 同时引入领域因果关系约束,确保反事实样本现实合理. CCE-FS还提供可视化特征效应分析,增强用户理解并揭示模型偏见. Statlog数据集实验表明,CCE-FS通过特征选择显著提高了反事实样本的有效性、正常性、稀疏度,并使连续特征接近度提升46.3%. 在Adult-Income和COMPAS数据集上的验证进一步证明,CCE-FS在因果一致性、数据分布合理性和连续特征邻近度方面均优于现有方法,展现了更强的解释与应用潜力.

关键词: 反事实解释;多目标优化;特征选择;因果关系;最大互信息系数;可视化特征效应
基金项目: 国家自然科学基金(No.62371187)
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2025)06-1805-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20241166
第二十七届中国科协年会学术论文

Comprehensive Causality Multi-Objective Counterfactual Explanation with Feature Selection

LIU Jin-ping, TANG Hao-nan, LI Xing-wang, XU Peng-fei*, YUAN Sheng-wei
(*College of Information Science and Engineering*, *Hunan Normal University*, *Changsha*, *Hunan* 410081, *China*)

Abstract: The widespread adoption of complex machine learning models across diverse industries has significantly increased the demand for model interpretability. The counterfactual explanation is a crucial post-hoc explanation method. However, traditional approaches often combine multiple objectives into a single objective optimization problem, leading to difficulties in weight assignment and reconciling conflicting objectives. Furthermore, existing methods also suffer from low computational efficiency, degraded prediction accuracy, and insufficient global explanations when dealing with high-dimensional, redundant, and noisy data. To address these issues, this article proposes a comprehensive causal multi-objective counterfactual explanation method with feature selection (CCE-FS). CCE-FS first employs the maximal information coefficient (MIC) to select key features, thereby enhancing prediction accuracy and global explanatory power. It then formulates the counterfactual search as a multi-objective optimization problem, effectively balancing the relationships between multiple objectives. Domain-specific causal relationships are incorporated as constraints to ensure the generated counterfactuals are realistic and plausible. Additionally, CCE-FS provides visual feature effect analysis to enhance user understanding and reveal potential model biases. Experiments conducted on the Statlog dataset demonstrate that CCE-FS significantly improves the validity, normality, and sparsity of counterfactual samples through feature selection, achieving a 46.3% enhancement in proximity for continuous features. Further validation on the Adult-Income and COMPAS datasets confirms that CCE-FS outperforms existing methods in causal consistency, data distribution reasonableness, and proximity of continuous features. These results highlight CCE-FS's superior explanatory capabilities and greater application potential.

收稿日期:2024-12-27;修回日期:2025-06-14;责任编辑:覃怀银
*通讯作者:徐鹏飞

Key words: counterfactual explanations; multi-objective optimization; feature selection; causal relationship; maximal information coefficient; visualization of feature effects

Foundation Item(s): National Natural Science Foundation of China (No.62371187)

1 引言

随着大数据和人工智能的迅猛发展,机器学习已广泛应用于工业故障监测[1]、医疗诊断[2]、辅助驾驶[3]等领域并取得显著成效. 然而,多数机器学习模型如同“黑匣子”,内部决策机制缺乏透明度,限制了其在医学诊断、金融信贷等高解释性需求场景中的可信应用[4]. 因此,可解释性机器学习日益受到重视[5],其中反事实解释作为一种事后解释技术,因其直观易懂的优势备受关注[6]. 其核心思想是将反事实样例融入解释过程[7],通过对原始数据实例施加微小扰动,生成一个与之相似但预测结果不同的新实例.

目前,反事实解释的研究呈现多样化发展趋势,相关研究主要围绕相关核心质量维度开展,其中生成反事实样例的可行性、多样性以及数据合理性[8~13]成为学界重点关注对象. 尽管这些方法已取得显著进展,但仍面临诸多亟待解决的技术瓶颈与挑战:

(1)在反事实样本生成中,研究者普遍采用的是将多个子目标加权聚合的优化策略,各子目标权重设定是技术难点,现有方法多依赖研究者经验赋值[14],易导致目标失衡. 此外,各子目标间常存在内在冲突[15],进一步加剧了自动平衡这些目标的难度.

(2)生成的反事实样本的有效性高度依赖对现实因果机制的精准建模. 尽管已有研究者如Duong等人[16]尝试利用结构因果模型(Structural Causal Model, SCM)来构建可操作的反事实样本,但多数算法未将因果因素纳入核心框架,导致解释效力受限.

(3)在高维冗余数据场景下,当前的反事实解释方法面临严峻考验[17]. 特征维度激增不仅推高计算复杂度、降低预测精度,还会削弱全局解释能力. 例如,SHAP(SHapley Additive exPlanations)方法[18]在处理大规模数据时计算成本呈指数级增长. 此外,不相关或冗余特征使优化算法陷入局部最优,严重影响样本质量[19]. 尽管已有研究者探索了特征重要性问题[20],但多数方法未充分考虑特征选择,现有方案多停留于经验性筛选层面.

针对上述挑战,本文提出一种带特征选择的综合因果多目标反事实解释方法(Comprehensive Causality multi-objective counterfactual Explanation with Feature Selection, CCE-FS). CCE-FS在数据预处理阶段引入基于最大互信息系数(Maximal Information Coefficient, MIC)的特征选择方法,精准剔除冗余与干扰特征,有效降低反事实样本优化计算复杂度,同时提升模型的预测精度及解释的全局性. 在核心优化环节,采用多目标优化策略消除权重主观赋值问题,同时嵌入领域因果约束,确保生成样本符合现实因果逻辑. 本文主要贡献如下:

(1)构建多目标因果解释体系. 将传统基于单目标优化的反事实样本搜索范式转化为多目标优化问题,引入相对影响度(Relative Impact Metric, RIM)量化特征对模型输出的影响,将因果约束深度融入优化过程,保障样本生成的因果合理性.

(2)引入特征智能筛选机制. 通过MIC智能度量特征与目标变量的关联强度,自动剔除冗余特征,显著提升高维数据场景下反事实样本质量,降低计算开销,避免局部最优陷阱,增强解释效能.

(3)设计可视化后解释方案. 利用加速无关模型解释技术(Accelerated Model-agnostic Explanations, AcME)量化特征效应,提供可视化手段直观呈现黑盒模型决策逻辑,助力用户识别模型潜在偏见,提升事后可解释性.

所提出的CCE-FS反事实解释方法能为金融信贷、医疗诊断等领域决策提供更科学、合理的解释支持,帮助决策者理解预测依据,做出更明智的决策.

2 理论基础

本节简要介绍MIC、多目标优化算法NSGA-II以及AcME方法的基本原理.

2.1 MIC

MIC是由Reshef等人[21]提出的用于衡量变量间统计依赖性的重要指标. 相较于传统的皮尔逊相关系数等只能捕捉线性关系的度量方法,MIC可全面量化任意复杂、非线性关系. 其通过系统遍历数据在不同分辨率下的网格划分,计算每种配置下变量的互信息,取最大值并标准化得到MIC值:

$$\mathrm{MIC}=\max_{x,y\ \text{such that}\ xy<B}\left(\frac{\max_G I_G}{\log\min\{x,y\}}\right) \tag{1}$$

其中,I_G是网格G上的互信息,B是样本大小n的函数,通常取$B=n^{0.6}$.

MIC能灵活量化变量间复杂非线性的关系,不受特定函数形式的限制. MIC取值范围为0~1,0表示变量独立,1则表示存在完美的函数关系. 由于其评估关系的通用性和等价性(即对于不同类型但具有相似噪声水平的关系给予相似的分数),MIC已成为特征选择领域中的一个强有力的工具.

2.2 NSGA-II与AcME

本文采用NSGA-II[22]算法优化反事实样本搜索. NSGA-II核心优势在于快速非支配排序与拥挤度距离

策略,提升收敛速度并维持解集多样性,有效避免局部最优. 算法核心流程包括:评估种群中个体适应度;快速非支配排序结合拥挤度计算,筛选非支配解集;执行选择、交叉、变异操作生成新的个体,合并新种群;迭代逼近帕累托最优解集.

AcME[23]是高效的特征重要性估计方法. 其核心是通过分位数值替换(如Q1、Q2、Q3)基线向量中的特定特征,构建变分位数矩阵,对比矩阵各行与基线预测的标准化差异,量化特征值变化对模型决策的影响.

3　CCE-FS

本节详细介绍本文所提出的CCE-FS模型,包括其整体框架、优化目标组成、基于遗传算法的反事实样本搜索以及如何使用特征效应来提高可解释性.

3.1　系统模型

图1展示了CCE-FS的整体的框架,它包括数据预处理、反事实实例生成和解释三个主要阶段,每个模块都针对现有问题提供了新的方案.

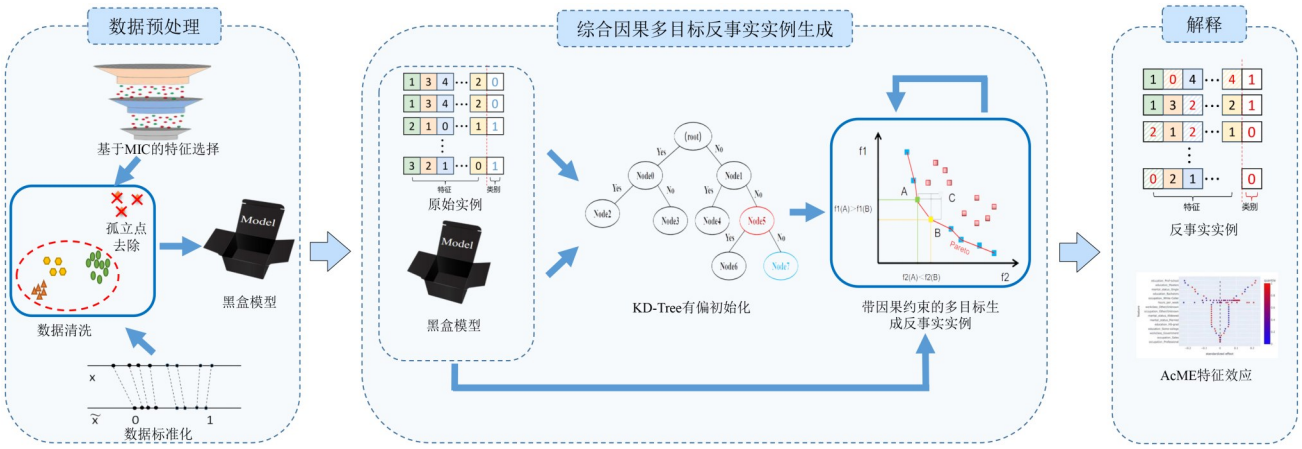


图1　CCE-FS整体框架

数据预处理. 依次执行缺失值修复、MIC特征筛选、数据标准化及类别编码. 通过MIC量化特征与目标变量的关联强度,剔除冗余信息;标准化消除量纲影响,编码处理使数据适配模型输入要求.

反事实实例生成. 基于KD-Tree建初始样本集,引入因果约束确保变量关系合理性,采用多目标优化算法迭代优化,提升样本的现实可行性.

解释分析. 结合AcME方法量化特征效应,直观呈现特征对模型预测的影响机制,为复杂决策提供可解释支持.

3.2　基于MIC的特征选择

本文以MIC为核心度量,评估特征与目标变量间的统计依赖关系. 筛选原则是优先保留MIC值显著较高的特征. 这些高MIC值特征认定为影响目标变量的关键因素,可有效提升模型预测精度与解释效能. 针对MIC值相同的特征,仅保留其一,避免模型过拟合,增强泛化能力与解释清晰度. 在超参数选择中,本文采用原始MIC方法[21]默认的网格分辨率α和裁剪参数c,平衡统计显著性、计算效率与复杂关系捕捉能力. 去除冗余特征不仅避免了模型因信息重复而过拟合,还提升了泛化能力,确保模型精简高效,并增强解释的清晰度.

3.3　目标函数设计

本文构建的多目标优化模型包含四个显式目标项与一个隐式因果约束项,具体设计如下:

预测输出损失. 旨在量化反事实样本预测值与期望预测值之间的偏差. 本文采用Hinge损失量化反事实样本预测值与期望输出的偏差,聚焦预测值跨越阈值(如0.5)的优化,避免严格二值化约束,其计算公式为

$$\text{hinge\_yloss}=\max(0,1-y^{*}y^{*}) \tag{2}$$

其中,y表示正确的类别标签,y^{*}为反事实样本的预测输出概率.

邻近度损失. 基于逆权重欧式距离衡量原始样本与反事实样本的相似度,突出关键特征影响[24,25],其计算方法如下:

$$\text{dist}(x_i,x_i^{*})=\sum_{j=1}^{n}\sqrt{w_j(x_{ij}-x_{ij}^{*})^2} \tag{3}$$

其中,w_j为由AcME方法估计出的权重倒数(越重要特征的w_j越小),f_j为权重系数,x_{ij}表示原始样本x_i的第j个特征,x_{ij}^{*}表示反事实样本x_i的第j个特征.

稀疏度. 量化特征变化数量,稀疏度的高低直接反映了两者在特征层面的差异程度[26],采用指数函数进行度量.

相对影响度. 本文提出RIM指标全面评估特征变化对模型的相对影响. 对于连续特征,通过相对变化差异与逆权重乘积衡量;对分类特征,则基于变化特征数量加权计算,即

$$\mathrm{RIM}_i=\begin{cases} d_{\text{real count}}\cdot\sum_{j=1}^{d_{\text{count}}}\Pi_{x_{ij}\neq x_{ij}^*}\cdot\left|\dfrac{x_{ij}-x_{ij}^*}{x_{ij}}\right|\cdot w_j, \text{if } x_{ij} \text{ is numerical} \\ d_{\text{real real}}\cdot\sum_{j=1}^{d_{\text{cat}}}\Pi_{x_{ij}\neq x_{ij}^*}\cdot w_j, \text{if } x_{ij} \text{ is categorical} \end{cases} \tag{4}$$

其中，$d_{\text{real count}}$表示连续特征中实际变化的特征数，$d_{\text{real real}}$表示分类特征中实际变化的特征数.

合理性. 反事实样本的合理性要求其尽可能符合原始数据集的分布特征，避免生成脱离数据分布的离群样本. 本文采用了经典的局部离群因子(Local Outlier Factor，LOF)作为反事实样本的合理性衡量指标，通过以下三步评估样本离群程度：

$$\mathrm{rd}_k(o,p)=\max\left(d_k(o),d(o,p)\right) \tag{5}$$

其中，$d_k(o)$代表邻近点o的第k距离，$d(o,p)$表示数据点o到数据点p的距离.

$$\mathrm{lrd}_k(p)=\frac{\left|N_k(p)\right|}{\sum_{o\in N_k(p)}\mathrm{rd}_k(o,p)} \tag{6}$$

其中，$N_k(p)$是指数据点p的第k邻域，用来表征点p的局部邻域密集程度.

$$\mathrm{lof}_k(p)=\frac{1}{\left|N_k(p)\right|}\sum_{o\in N_k(p)}\frac{\mathrm{lrd}_k(o)}{\mathrm{lrd}_k(p)} \tag{7}$$

则通过对比点p与邻域点的局部密度，量化其离群程度.

因果关系. 将因果逻辑转化为等式或不等式约束. 例如，人种在现实世界中不能改变，故关于人种的因果约束条件可以设置为等式$x_i-x_i^*=0$.

最终，模型将上述目标整合为多目标最小化任务，在满足因果约束的前提下求解帕累托最优解.

$$\min_x o(x)=\min_x\left(\text{hinge\_loss}(\cdot),\text{dist}(\cdot),\Pi(\cdot),\text{RIM}(\cdot),\text{lof}(\cdot)\right) \tag{8}$$

3.4 反事实样本优化搜索

本文提出一种有偏优化的初始种群构建策略，通过细致调整各类交叉与变异算子，以更精准地适应问题的特征. 具体通过引入KD-Tree算法，为每个原始样本生成一组潜在空间邻近但是预测结果相反的样本作为初始种群. 这种精心设计的初始化策略，使得算法能够聚焦于原始样本的邻近区域，从而提供更为精确的反事实样本集合.

在算子设计方面，针对不同特征类型采用差异化策略. 数值特征采用模拟二进制交叉(Simulated Binary Crossover，SBX)与多项式变异(Polynomial Mutation，PM)算子，类别特征则采用均匀交叉(Uniform Crossover，UX)结合随机选择变异的方法，以适配不同特征类型的运算需求.

考虑到数据集规模与特征维度的差异性，本文提出根据数据复杂度动态调整种群大小与迭代次数的策略，具体参数设置如表1所示.

表1 种群大小与迭代次数设置

数据集	种群大小	迭代次数
Statlog	50	30
Adult-Income	500	80
COMPAS	300	50

3.5 特征效应可视化

为增强用户对于黑盒模型以及反事实样本的信任和理解，本文提出一种基于可视化特征效应的反事实样本和黑盒预测模型事后可解释性增强方法. 通过估计特征效应，使用户能够明确各特征对模型预测输出的影响方向，进而深入理解黑盒模型的内部运作机制，显著增强模型的可靠性和可解释性. 此外，通过特征效应提供的信息，用户能够在生成反事实样本之前就了解到每个特征对于模型预测输出的实际影响方向，而不仅仅依赖于反事实解释所提供的假设场景，还能揭示模型可能存在的偏见，进一步增强用户对于反事实样本的判断能力. 在实际应用中，本文采用AcME方法进行特征效应评估.

4 实验验证与结果分析

4.1 数据集与预处理

本文基于加州大学欧文分校(University of California Irvine，UCI)的Statlog、Adult-Income和COMPAS三个公开数据集开展实验. 这些数据集涵盖金融、社会经济和司法领域，具有广泛的代表性，在相关研究中有着广泛的引用. 表2汇总了这三个数据集的基本信息和预测模型准确率.

表2 数据集基本信息和预测模型准确率

数据集	样本数	连续特征数	分类特征数	预测模型准确率
Statlog	1 000	7	13	0.803
Adult-Income	32 561	6	8	0.835
COMPAS	7 214	1	4	0.686

Statlog(German Credit Data). 该数据集用于个人信用状况评估，判断借款人属于低风险还是高风险类别. 数据集共包括1 000个样本，20个特征，因特征丰富且规模适中，已成为金融借贷领域研究中的重要基准数据集[27].

Adult-Income. 该数据集记录了人口普查数据，其主要任务是预测个人年收入是否超过5万美元. 数据集共包含32 561个样本，14个特征，在社会经济领域研究中具有很高的代表性.

COMPAS. 该数据集是用于犯罪再犯风险预测的数据集,共包含7 214个样本,每个样本由5个特征构成. 该数据集是研究算法公平性与可解释性,尤其是司法决策辅助场景的常用数据集.

由于上述数据集同时包含连续特征和分类特征,为了确保能在机器学习算法中使用这些特征,本文使用了独立编码将分类特征转换为二进制向量. 此外,为了消除特征之间不同量纲对模型的影响,本文对数据进行了标准化处理.

4.2　对比方法

为评估CCE-FS框架生成的反事实样本质量,本文采用了当前该领域内广泛应用的代表性方法进行对比分析. 具体方法如下:

SingleCF. 基于Wachter等人[7]工作,将原始样本与反事实样本的模型预测差异(y-loss)和相似度(距离)组合后进行优化,生成单一的反事实样本.

Dice. Diverse Counterfactual Explanations(Dice)由Mothilai等人[10]提出,借助DPP引入多样性度量,扩展损失函数以兼顾反事实样本的合理性与多样性.

No-Dice. 该方法是对Dice方法的扩展[10]. 通过将多样性目标的权重设置为0,忽略多样性因素,通过综合考虑Hinge损失,评估生成反事实样本的质量.

NICE. 最近实例反事实解释(Nearest Instance Counterfactual Explanations,NICE)由Brughmans等人[28]提出,专为结构化数据设计,利用最近异类邻居信息加速搜索,通过迭代引入领域特征值生成解释,具备模型不可知性、高效生成及用户需求适配性.

CCE-FS-C1. 该方法是本文中提出的CCE-FS方法的变体,删除了预处理步骤中的特征选择,用于评估特征选择对反事实样本生成的影响.

CCE-FS-C2. 作为对CCE-FS方法的另一变体,移除了优化目标中的RIM指标,以评估RIM指标对模型性能的实际贡献.

4.3　评价指标

本文选取有效性、稀疏度和邻近度等广泛认可的指标,并基于LOF算法提出新指标"LOF正常性比率".

有效性(Val). 用于评估反事实样本被模型正确预测为期望类别的比例,计算方法如下:

$$\text{Validity}=\frac{1}{S}\sum_{i=1}^{S}\Pi\left(f(x_i^*)=y_i^*\right) \tag{9}$$

其中,S表示需要生成反事实样本的原始样本的总数,x_i^*表示反事实样本,y_i^*表示期望类别,$\Pi(\cdot)$为指示函数用来判断反事实样本预测输出是否与期望输出一致(一致时为1,否则为0).

稀疏度(Spa). 用来衡量反事实样本相较于原始样本发生了变化的特征数量,分连续与分类特征计算:

$$\text{Con\_Spa}=\frac{1}{S}\sum_{i=1}^{S}\left(1-\frac{1}{d_{\text{cont}}}\sum_{j=1}^{d_{\text{cont}}}\Pi_{x_{ij}^*\neq x_{ij}}\right) \tag{10}$$

$$\text{Cat\_Adj}=\frac{1}{S}\sum_{i=1}^{S}\left(1-\frac{1}{d_{\text{cat}}}\sum_{j=1}^{d_{\text{cat}}}\Pi_{x_{ij}^*\neq x_{ij}}\right) \tag{11}$$

其中,d_{cont}和d_{cat}分别表示连续和分类特征的数量.

邻近度(Adj). 指原始样本与反事实样本之间的距离度量. 其中,类别邻近度(Cat_Adj)直接使用分类特征稀疏度来衡量,如式(11)所示. 对于连续特征邻近度(Con_Adj),沿用Wachter等人[7]的距离标准,计算公式为

$$\text{Con\_Adj}=\frac{1}{S}\sum_{i=1}^{S}\left(-\frac{1}{d_{\text{cont}}}\sum_{j=1}^{d_{\text{cont}}}\frac{\left|x_{ij}^*-x_{ij}\right|}{\text{MAD}_j}\right) \tag{12}$$

其中,MAD_j表示第j个连续特征的中值绝对偏差.

LOF正常性比率(Lof_S). 基于LOF算法评估反事实样本与原始数据集分布的一致性,计算方式如下:

$$\text{Lof\_S}=\frac{1}{S}\sum_{i=1}^{S}\left(\Pi(f(x_i^*)=1)\right) \tag{13}$$

4.4　不同特征选择方法对比

为深入探讨不同特征选择技术对模型性能的影响,本文首先在Statlog数据集上开展实验,采用了多种特征选择方法,包括基于MIC的过滤法、基于随机森林的包裹法,以及基于L1正则化项的嵌入法.

表3显示了不同特征选择方法在Statlog数据集上使用非线性预测模型时的性能. 所有方法在有效性和正常性比率上均得满分1,验证了CCE-FS方法生成有效反事实样本的能力,并保持了与原始数据分布的高度一致性.

表3　不同特征选择方法对比

方法	Val	Con_Spa	Cat_Spa	Adj	Lof_S
CCE-FS-C1	**1**	**0.387**	0.716	−594	**1**
MIC	**1**	0.314	0.747	**−302**	**1**
Random_forests	**1**	0.356	0.717	−596	**1**
Embedded_L1	**1**	0.324	**0.756**	−390	**1**

注:加粗数据表示最优结果.

不同方法在连续和分类稀疏度上差异较小. 具体而言,未进行特征选择的CCE-FS展现了最佳性能,而MIC、Random_forests和Embedded_L1的连续稀疏度分别为0.314、0.356和0.324. 类似地,MIC、Raandom_forests、Embedded_L1的分类稀疏度分别为0.747、0.717、0.756,同样仅表现出微小差异. 这表明各方法在特征变化数量上相近,但策略有细微差别.

而连续邻近度指标差异显著. 基于MIC特征选择的方法得分最高,远超未特征选择的CCE-FS-C1. 因此,使用MIC进行特征选择的方法在保持反事实样本与原始样本连续特征相似性方面具有显著优势,从而

在特征选择阶段就为生成高质量的反事实样本提供了保障.

4.5 无因果约束的性能评估

为评估因果约束的有效性与必要性,进一步在Adult-Income和COMPAS数据集上开展无因果约束的实验. 表4展示了使用以非线性预测模型为基准预测模型时,CCE-FS与SingleCF、NICE、Dice和No-Dice方法及它们的变种方法在两个数据集上的表现,其中最优结果以粗体表示.

分析Adult-Income数据集的实验结果可见:多数方法在生成单个反事实样本时都表现出了高度有效性,除NICE方法外,其余方法有效性指标都达到了1. 然而,在连续特征稀疏度方面,不同的方法差异显著. 在未经过事后的连续特征增强的方法连续特征稀疏度普遍为0.2~0.3,经事后增强后提升至0.6左右. 相比之下,本文提出的CCE-FS连续特征稀疏度达到0.594. 尽管略显逊色,但其性能仍明显高于未进行事后增强的方法,并优于近期提出的NICE方法.

表4 无因果约束的性能评估

数据集	方法	Val	Con_Spa	Cat_Spa	Con_Adj	Lof_S
Adult-Income	SingleCF	**1**	0.601	0.763	–0.744	0.615
	Dice	**1**	0.279	0.765	–0.868	0.630
	No-Dice	**1**	0.262	0.765	–0.840	0.628
	NICE	0.960	0.440	0.784	–1.374	**0.790**
	Dice-Spare	**1**	0.607	0.765	–0.747	0.634
	No-Dice-Spare	**1**	**0.619**	0.765	–0.710	0.618
	CCE-FS	**1**	0.594	**0.807**	**–0.663**	0.700
COMPAS	SingleCF	0.982	0.509	0.701	–1.693	0.360
	Dice	**1**	0.398	0.717	–1.733	0.394
	No-Dice	0.903	0.093	0.898	–3.280	0.661
	NICE	0.920	0.240	**0.915**	–1.288	**0.800**
	Dice-Spare	**1**	**0.534**	0.716	–1.651	0.385
	No-Dice-Spare	0.903	0.201	0.898	–3.201	0.646
	CCE-FS	**1**	0.484	0.882	**–0.912**	0.760

注:变种方法是指以Spare为后缀的扩展方法,主要针对已经生成的反事实样本进行事后的连续特征稀疏性加强.

就反事实样本的合理性而言,CCE-FS模型在Adult-Income数据集上表现出色. 具体来说,CCE-FS的Lof_S为0.700,仅次于NICE且优于其他方法,表明其样本更符合原始数据分布.

此外,从特征邻近度的角度看,CCE-FS仍展现出明显优势. 该方法连续特征接近度为–0.663,显著优于其他方法,更好维持了特征相似性.

类似地,在COMPAS数据集上,CCE-FS仍然表现出卓越的有效性,其有效性指标仍保持为1. 在连续特征稀疏度上,CCE-FS的性能超越了Dice、No-Dice、NICE以及No-Dice-Spare方法,但低于经过事后的连续特征稀疏度加强操作方法,这一表现与CCE-FS在Adult-Income数据集上的表现基本一致. 至于分类特征稀疏度,CCE-FS的性能与No-Dice接近,优于其他方法的性能.

综合来看,CCE-FS在Adult-Income和COMPAS数据集上展现出了稳定且可靠的性能. 在确保反事实样本有效性的同时,CCE-FS生成的反事实样本表现出了较为稳定的稀疏度,并在连续特征邻近度和数据合理性方面得到显著提升. 这一系列的实验结果表明,CCE-FS在生成反事实样本时,不仅能够保持模式的一致性和有效性,还能在各方面综合考虑,展现出其在实际应用中的可行性和优越性.

4.6 有因果约束的性能评估

为进一步验证CCE-FS在多个因果关系约束下的表现,本文分别对Adult-Income和COMPAS数据集进行了深入实验. 基于现实世界的因果关系,制定了一系列因果约束,并根据约束数量进而划分为一级和二级两个不同的约束等级,具体内容见表5和表6. 在这两个不同等级的因果约束条件下的实验结果详见表7与表8,其中最优结果以粗体表示.

从表7中可以清晰发现,将因果约束从一级调整到二级时,CCE-FS在Adult-Income数据集上的如下性能指标得到了显著的提升:连续特征稀疏度从0.540提高到0.772,表明在二级约束条件下生成的反事实样本中连续特征更趋稀疏;类别特征稀疏度从0.740增加到0.814,这说明在二级约束条件下,生成的反事实样本的类别特征变化量更小;LOF正常性比率从0.700增加到了0.861,验证了二级约束下反事实样本与原始数据分布的匹配度更高.

值得注意的是,表7中连续特征接近度从最初的

表5　Adult-Income数据集上不同等级的因果约束

等级	特征	因果约束
一级	年龄	不能减小
	种族	不能改变
二级	性别	不能改变
	年龄	短期内不能改变
	教育程度	短期内不能改变
	婚姻状况	未婚-0,已婚-1,丧偶-2,分居-2,离异-2,只能保持不变或按升序增加

表6　COMPAS数据集上不同等级的因果约束

约束等级	特征	因果约束
一级	—	—
二级	性别	不能改变
	种族	不能改变
	年龄	进行分段,并将每个阶段指定为不可变阶段

表7　Adult-Income数据集不同因果约束下CCE-FS性能表现

等级	Val	Con_Spa	Cat_Spa	Con_Adj	Lof_S
一级	**1**	0.540	0.740	**–0.698**	0.700
二级	**1**	**0.772**	**0.814**	–0.893	**0.861**

表8　COMPAS数据集不同因果约束下CCE-FS性能表现

等级	Val	Con_Spa	Cat_Spa	Con_Adj	LoF_S
一级	**1**	**0.440**	0.865	**–0.950**	0.760
二级	**1**	0.020	**0.960**	–2.689	**0.871**

–0.698降低到了–0.893,这是由于二级约束将年龄设为不可变特征,仅有一个连续特征“每周工作时长”是可变的. 因此在使用NSGA-II算法进行优化时,会优先选择基于该特征扰动的反事实样本,导致其变化幅度增大.

类似地,COMPAS数据集实验结果如表8所示. 表8的结果表明:在分类特征稀疏度和LOF正常性比率两个指标上,二级因果约束下的CCE-FS方法的性能均超过了一级因果约束下的性能. 但连续特征的稀疏度和邻近度有所降低,这是因为该数据集只包含一个连续特征且其权重远高于分类特征.

综上,合理引入现实因果关系约束可提升反事实样本与真实数据的吻合度,既体现了CCE-FS对因果约束的强适应性,也凸显了实际应用中科学配置因果约束的重要性.

4.7　RIM指标验证

本文在传统的使用指示函数统计稀疏度的基础上提出的RIM,旨在全面量化特征变化(尤其是连续特征)对模型的相对影响. 为了验证RIM指标的有效性,使用CCE-FS方法在Adult-Income数据集上进行了对比实验,实验结果如表9所示,其中最优结果以粗体表示.

连续特征接近度:无论在一级还是二级因果约束,CCE-FS方法生成的反事实样本相对于CCE-FS-C2方法

表9　RIM指标的验证

等级	方法	Val	Con_Spa	Cat_Spa	Con_Adj	LoF_S
一级	CCE-FS	**1**	**0.540**	0.740	**–0.698**	0.700
	CCE-FS-C2	**1**	0.485	**0.753**	–0.869	**0.740**
二级	CCE-FS	**1**	**0.772**	**0.814**	**–0.893**	**0.861**
	CCE-FS-C2	**1**	0.718	0.802	–1.269	**0.861**

生成的反事实样本,其连续特征接近度都显著提升. 连续特征稀疏度:类似地,CCE-FS方法相较于CCE-FS-C2方法生成的反事实样本,其分类特征稀疏度也都得到明显提升.

综合来说,使用RIM指标后,分类特征稀疏度和LOF正常性比率的表现相对稳定的同时,连续特征的稀疏度和接近度均得到明显提升. 这表明,RIM指标在考虑特征变化幅度对模型的相对影响时,能够在保持分类特征稳定性的同时,使模型对连续特征的小幅度变化更加敏感,从而有效提升反事实样本的质量.

4.8　特征效应可视化和反事实样本分析

基于前几个小节的实验,CCE-FS的性能已经得到充分验证. 该方法能够在不同的因果约束等级下生成一系列既符合数据分布又符合现实世界因果关系的反事实样本(如表10所示). 下文将对这些反事实样本进行定量的分析.

表10详细展示了CCE-FS在Adult-Income数据集不同因果等级下生成的8个反事实样本,从中可归纳出如“教育程度提升可促进个人年收入超5万”等规律. 为了进一步增强黑盒预测模型与反事实样本的可解释性,本文基于AcME方法来估计特征效应,这两方面的知识有助于用户理解模型的预测机制,如图2所示. 分析图2,可得到以下结论:

(1)特征重要性序列. 排名前三的特征分别是专科教育程度、硕士教育程度、未婚婚姻状况,后续依次为本科教育程度、白领职位等,这些结果为用户提供了对模型中各特征重要性的直观参考.

(2)特征效应方向. 特征按影响分为正负两组. 例如,专科教育程度、硕士教育程度等特征具有正向影响. 用户可据此分析反事实样本的特征变化,提升解释的可靠性,同时识别模型潜在偏见.

综上,反事实样本与特征效应的结合,既深化用户

表 10 反事实样本

约束等级	年龄/岁	每周工作时长/h	性别	种族	婚姻状况	教育程度	工作类型	职位
一级	38	60	男	白	未婚	副学士→专科	政府部门	白领
	46→47	40	男	白	已婚	高中	政府部门	蓝领→白领
	69	24	男	白	已婚	高中→硕士	私人部门	销售→专职
	45	14→22	女	白	已婚	高中→硕士	私人部门	专职
二级	23	40	女	白	未婚→已婚	本科	私人部门	白领
	41	55→60	女	白	离异	本科	私人部门	白领
	66	40→50	男	白	已婚	高中	政府部门	蓝领→白领
	34	45→75	男	白	已婚	高中	私人部门→自雇	蓝领→白领

注:“→”符号表示特征发生了变化,指向的具体值为反事实样本的数值.

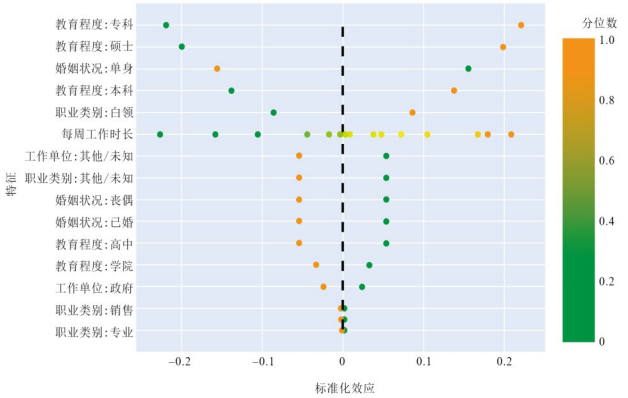


图 2 基于 AcME 的特征权重和对 Adult-Income 数据集的影响

对黑盒模型的理解与信任,又增强反事实解释的可解释性,为可解释机器学习提供了使模型决策更透明的新分析视角.

5 结论

本文提出了一种新颖的带特征选择的综合因果多目标反事实解释框架——CCE-FS. 预处理阶段通过特征选择优化多特征数据集处理性能,降低计算复杂度,提升预测准确率与解释全局性;此外,CCE-FS将传统单一优化搜索转化为多目标优化问题,引入因果关系约束确保反事实样本符合因果逻辑,并结合 AcME 方法可视化特征效应,增强黑盒模型与反事实样本的可解释性. 在 Statlog、Adult-Income 和 COMPAS 上的实验表明,相较于同类方法,CCE-FS 生成的反事实样本在保持高度有效性的同时,显著提升了数据分布合理性和连续特征邻近度. 总之,通过提供符合因果逻辑的解释,CCE-FS 可增强用户对模型预测结果的信任,推动人工智能技术在高解释性需求的领域的应用和推广. 然而,尽管 CCE-FS 方法取得显著进展,但该方法在优化过程中,LOF 算法的 K 值和阈值选择受数据集属性影响显著. 因此,未来的工作将在以下两方面展开:一是探索基于数据集聚类特性探索 LOF 参数自适应调整方法;二是将特征因果关系深度融入优化目标,超越传统因果约束,提升反事实解释的深度与广度.

参考文献

[1] 蔡美玲, 罗迪, 肖敬日, 等. 连续与离散变量协同分析的非平稳非高斯工业过程异常检测[J]. 电子学报, 2024, 52(10): 3291-3300.
CAI M L, LUO D, XIAO J R, et al. Continuous and discrete variables-concurrent analysis-based nonstationary and non-Gaussian industrial process anomaly detection[J]. Acta Electronica Sinica, 2024, 52(10): 3291-3300. (in Chinese)

[2] 刘金平, 吴娟娟, 张荣, 等. 基于结构重参数化与多尺度深度监督的COVID-19胸部CT图像自动分割[J]. 电子学报, 2023, 51(5): 1163-1171.
LIU J P, WU J J, ZHANG R, et al. Toward automated segmentation of COVID-19 chest CT images based on structural reparameterization and multi-scale deep supervision[J]. Acta Electronica Sinica, 2023, 51(5): 1163-1171. (in Chinese)

[3] 苏越阳, 姚迪, 毕经平. 基于噪声标签重加权的车辆轨迹异常检测方法[J]. 电子学报, 2025, 53(1): 182-192.
SU Y Y, YAO D, BI J P. A vehicle trajectory anomaly detection method based on noise label re-weighting[J]. Acta Electronica Sinica, 2025, 53(1): 182-192. (in Chinese)

[4] NAZEMI A, FABOZZI F J. Interpretable machine learning for creditor recovery rates[J]. Journal of Banking & Finance, 2024, 164: 107187.

[5] CHEN V, YANG M Y, CUI W B, et al. Applying interpretable machine learning in computational biology-pitfalls, recommendations and opportunities for new developments[J]. Nature Methods, 2024, 21(8): 1454-1461.

[6] KARIMI A H, BARTHE G, SCHÖLKOPF B, et al. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations[J]. ACM Computing Surveys, 2023, 55(5): 1-29.

[7] WACHTER S, MITTELSTADT B, RUSSELL C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR[J]. Harvard Journal Of Law & Technology, 2017, 31: 841.

[8] AUGUSTIN M, BOREIKO V, CROCE F, et al. Diffusion visual counterfactual explanations[J]. Advances In Neural Information Processing Systems, 2022, 35: 364-377.

[9] KENNY E M, KEANE M T. On generating plausible counterfactual and semi-factual explanations for deep learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(13): 11575-11585.

[10] MOTHILAL R K, SHARMA A, TAN C H. Explaining machine learning classifiers through diverse counterfactual explanations[C]//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York: ACM, 2020: 607-617.

[11] POYIADZI R, SOKOL K, SANTOS-RODRIGUEZ R, et al. FACE: Feasible and actionable counterfactual explanations[C]//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, 2020: 344-350.

[12] 向许, 于洪, 张晓霞, 等. IsomapVSG-LIME: 一种新的模型无关解释方法[J]. 智能系统学报, 2023, 18(4): 841-848.
XIANG X, YU H, ZHANG X X, et al. IsomapVSG-LIME: A novel local interpretable model-agnostic explanations[J]. CAAI Transactions on Intelligent Systems, 2023, 18(4): 841-848. (in Chinese)

[13] KANAMORI K, TAKAGI T, KOBAYASHI K, et al. DACE: Distribution-aware counterfactual explanation by mixed-integer linear optimization[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Freiburg: IJCAI, 2020: 2855-2862.

[14] CHO S H, SHIN K S. Feature-weighted counterfactual-based explanation for bankruptcy prediction[J]. Expert Systems with Applications, 2023, 216: 119390.

[15] CHEN L S, FERNANDO H, YING Y M, et al. Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance[J]. Advances In Neural Information Processing Systems, 2024, 36: 1-53.

[16] DUONG T D, LI Q, XU G D. Achieving counterfactual fairness with imperfect structural causal model[J]. Expert Systems with Applications, 2024, 240: 122411.

[17] KUMAR I E, VENKATASUBRAMANIAN S, SCHEIDEGGER C, et al. Problems with Shapley-value-based explanations as feature importance measures[EB/OL]. (2022-06-30)[2025-06-30]. https://arxiv.org/abs/2002.11097v2.

[18] WANG H J, LIANG Q X, HANCOCK J T, et al. Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods[J]. Journal of Big Data, 2024, 11(1): 44.

[19] MAKAROVA A, SHEN H, PERRONE V, et al. Overfitting in Bayesian optimization: An empirical study and early-stopping solution[C]//The 2nd Workshop on Neural Architecture Search. Washington DC: ICLR, 2021: 1-15.

[20] FIGUEROA BARRAZA J, LÓPEZ DROGUETT E, RAMOS MARTINS M. FS-SCF network: Neural network interpretability based on counterfactual generation and feature selection for fault diagnosis[J]. Expert Systems with Applications, 2024, 237: 121670.

[21] RESHEF D N, RESHEF Y A, FINUCANE H K, et al. Detecting novel associations in large data sets[J]. Science, 2011, 334(6062): 1518-1524.

[22] DEB K, PRATAP A, AGARWAL S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182-197.

[23] DANDOLO D, MASIERO C, CARLETTI M, et al. AcME: Accelerated model-agnostic explanations: Fast whitening of the machine-learning black box[J]. Expert Systems with Applications, 2023, 214: 119115.

[24] KEANE M T, SMYTH B. Good Counterfactuals and Where to Find Them: A Case-based Technique for Generating Counterfactuals for Explainable AI (XAI)[M]//Case-Based Reasoning Research and Development. Cham: Springer International Publishing, 2020: 163-178.

[25] YANG H R, CHEN H X, ZHANG S X, et al. Generating counterfactual hard negative samples for graph contrastive learning[C]//Proceedings of the ACM Web Conference 2023. New York: ACM, 2023: 621-629.

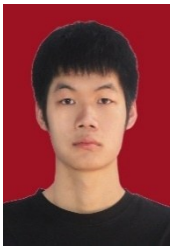
[26] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python[J]. Journal of Machine Learning Research, 2011, 12: 2825-2830.

[27] MARCEAU L, QIU L L, VANDEWIELE N, et al. A comparison of deep learning performances with other machine learning algorithms on credit scoring unbalanced data [EB/OL]. (2020-02-25)[2025-06-30]. https://arxiv.org/abs/1907.12363v2.

[28] BRUGHMANS D, LEYMAN P, MARTENS D. NICE: An algorithm for nearest instance counterfactual explanations[J]. Data Mining and Knowledge Discovery, 2024, 38(5): 2665-2703.

作者简介

刘金平 男,1983年9月出生于湖南省洞口县. 现为湖南师范大学信息科学与工程学院教授、博士生导师. 主持、参与国家和省部级科研课题10余项,发表国际高水平SCI期刊论文60余篇,中文权威期刊论文20余篇,申请国家发明专利40余项,已授权20项.
E-mail: lip@phunnu.edu.cn



汤浩楠 男,2002年10月出生于湖南省长沙市. 现为湖南师范大学软件工程专业硕士研究生. 主要研究方向为可解释性机器学习.
E-mail: TangHaonan@hunnu.edu.cn

李兴旺 男,2002年5月出生于湖南省娄底市. 现为湖南师范大学软件工程专业硕士研究生. 主要研究方向为医学图像分析.
E-mail: 202420294253@hunnu.edu.cn

徐鹏飞 男,1977年9月出生于湖南省岳阳市. 现为湖南师范大学信息科学与工程学院副教授. 主要研究方向为智能信息处理.
E-mail: xupf@hunnu.edu.cn

袁晟玮 男,2000年8月出生于湖南省常德市. 现为湖南师范大学软件工程专业硕士研究生. 主要研究方向为工业过程变量预测.
E-mail: ysw714@hunnu.edu.cn