校准解释：包含不确定性信息和反事实

Helena L¨ofstr¨om[a,b,*], Tuwe L¨ofstr¨om[c], Ulf Johansson[c], Cecilia S¨onstr¨od[c]

[a] *Jönköping International Business School, Jönköping University, Sweden*
[b] *Department of Information Technology, University of Borås, Sweden*
[c] *Department of Computing, Jönköping University, Sweden*

arXiv:2305.02305v3 [cs.AI] 4 Nov 2023

摘要

虽然针对人工智能模型的局部解释能够为个体预测提供洞见，例如特征重要性，但它们饱受不稳定性等问题的困扰。特征权重的不可靠性常常源于校准不佳的机器学习模型所导致的偏差，使这些挑战更加严重。此外，在可解释人工智能（XAI）中，特征重要性不确定性这一关键方面仍基本未得到解决。本文提出的新型特征重要性解释方法称为校准解释（CE），旨在直接应对这些问题。CE建立在Venn-Abers的基础之上，不仅能够校准底层模型，还能提供可靠的特征重要性解释，并对特征权重给出精确定义。CE通过处理输出不确定性，超越了传统解决方案。它通过为特征权重和模型的概率估计提供不确定性量化来实现这一点。此外，CE与模型无关，具有易于理解的条件规则，并能够生成嵌入不确定性量化的反事实解释。对25个基准数据集进行评估的结果凸显了CE的有效性，使其成为一种快速、可靠、稳定且稳健的解决方案。

关键词：可解释人工智能、特征重要性、校准解释、Venn-Abers、不确定性量化、反事实解释

1. 引言

许多基于人工智能的系统都有一项核心任务：预测建模。在这一过程中，机器学习（ML）算法会利用历史样本创建模型，将输入映射到目标，以便预测新的输入。用于基于人工智能的决策支持的预测模型通常并非为实现透明性而设计。尽管它们应用于医学或国防等关键场景，却往往只能给出一个可能的结果（David Gunning，2017；Ribeiro 等人，2016），这可能导致误用（用户的依赖程度高于适当水平）或弃用（用户的依赖程度低于适当水平）（Alvarado-Valencia 与 Barrero，2014；Bu˛cinca 等人，2020年）。

由于缺乏透明性，这类模型的预测往往需要解释。在可解释人工智能（XAI）领域，目标是创建相关方法，帮助人类用户判断何时应信任预测、何时不应信任，例如医学诊断中出现错误预测的情况（Marx 等，2023）。解释应揭示底层模型的优势与弱点，以说明模型未来将如何表现（David Gunning，2017；Dimanov 等人，2020年）。

解释主要分为两类：局部解释，介绍个体预测背后的原因；全局解释，则提供有关模型整体行为的信息（Guidotti 等人，2018b；Moradi 和 Samwald，2021；Martens 和 Foster，2014）。局部

*通讯作者。*Email addresses:* helena.lofstrom@ju.se (Helena L¨ofstr¨om), Tuwe.lofstrom@ju.se (Tuwe L¨ofstr¨om),Ulf.johansson@ju.se (Ulf Johansson), Cecilia.sonstrod@ju.se (Cecilia S¨onstr¨od)

解释存在一些缺点，例如不稳定性：实例的细微变化可能导致解释大不相同，并给评估解释质量带来问题（Slack 等，2021；Rahnama 和 Bostr¨om，2019）。

事后解释可以是事实性（特征值会对预测产生这种影响），也可以是反事实（要改变实例的特征值，最少需要做出哪些改动才能改变预测）（Mothilal 等，2020；Guidotti，2022；Wachter 等，2017）。反事实解释被认为对人类非常友好，因为它们反映了人类的推理方式（Molnar，2022）。

当今最先进的解释方法还有一个局限：它们对模型不确定性和可靠性的洞见有限。尽管早期学者已指出，不确定性信息对决策至关重要（例如，参见（Parmigiani，2002）），但直到近些年，不确定性估计才被视为解释中不可或缺的一部分，有助于提高底层模型的透明度（Bhatt 等，2021；Slack 等，2021；Antor´an 等，2020）。Zhou 等（2021a）指出，对于人工智能用户而言，首要问题并非模型的不透明性或复杂性，而是输出不确定性。Phillips-Wren 和 Adya（2020）将用户不确定性列为影响决策质量的四个压力因素之一。因此，准确的不确定性估计能够提升可解释人工智能中解释的质量和实用性。尽管不确定性被视为透明决策中的关键因素，但对于复杂问题，获取校准的不确定性估计的成本很高（Bhatt 等，2021；Marx 等，2023）。

在局部解释方法中，大多数机器学习模型输出的概率估计通常被用作衡量各类别可能性的指标。然而，众所周知，机器学习模型往往校准不佳，这意味着其概率估计与预测正确的实际概率并不一致（Van Calster 等，2019年；Grushka-Cockayne 等，2017年）。为解决这些问题，人们开发了诸如 Venn-Abers（VA）（Vovk & Petej，2012年）之类的校准方法。VA 会为每个预测生成一个概率区间，通过正则化可将其汇总为校准概率估计，以便与其他校准方法或底层模型的概率估计进行比较。

使用 VA 进行决策时，需要注意的是，该技术还会为每个类别提供区间，以量化概率估计中的不确定性；从解释视角来看，这一点很有价值。区间的宽度反映了模型的不确定性：区间越窄，概率估计中的不确定性越低；区间越宽，概率估计中的不确定性越高。

本文的一个核心假设是，使用校准不佳的模型的解释来支持决策，其价值值得怀疑，因为校准不佳的模型无法反映现实。使用此类模型的解释最终很可能导致弃用，因为这些解释与现实不符，会产生误导。本文还假设，将解释表示为条件规则有助于理解，因为这能明确说明解释适用的条件，使其易于理解。此外，明确定义特征权重相对于模型预测的含义也很重要。本文适用于二分类，但侧重于解释概率估计，而非类别标签，因为概率估计包含有关预测的更丰富信息。量化特征权重的不确定性，是帮助用户深入理解预测的重要工具。最后，从决策角度来看，能够轻松探索事实解释和反事实解释也很有益。

本文介绍了一种全新的局部特征重要性解释方法，称为校准解释。校准解释具有以下主要特点：

1. 作为构建解释的前提，底层模型经过校准，以确保与现实保持一致。VA 为校准概率估计提供不确定性估计，并将相关信息传达给用户。2. 校准解释方法提供了具有明确含义的特征权重，这些含义与模型的校准概率估计相关。特征权重表示该特征对于解释预测的重要性。3. 特征权重附带不确定性量化，其定义与底层模型的校准概率估计精确相关。

4. 条件规则经过定义，使其能够结合实例值进行直观解读。5. 校准解释支持基于模型校准概率估计的事实性解释和反事实解释。

本文结构如下：下一节回顾与解释方法、校准和 VA 预测器相关的基本概念。第 3 节介绍本文的主要贡献，即定义校准解释。第 4 节介绍事实解释和反事实解释。第 5 节介绍两项实验的设置和结果。最后，本文先进行讨论，再作总结。

2. 背景

2.1. 解释的期望特性

在可解释人工智能中创建高质量解释，取决于其所要实现的目标，而这些目标可能各不相同。例如，评估用户对解释界面的评价，与评估解释是否准确反映底层模型，是两回事（L¨ofstr¨om 等，2022）。尽管如此，事后解释方法仍有一些普遍值得追求的特性：

- 解释方法必须能够可靠地反映底层模型，这一点至关重要；它与解释方法应对底层模型具有高度忠实度这一理念密切相关（Slack 等，2021；Murdoch 等，2019；Carvalho 等，2019a；Arrieta 等，2020；Das & Rad，2020；Chromik & Schuessler，2020；Hoffman 等，2018；Doshi-Velez & Kim，2017；Adadi & Berrada，2018；Wang 等，2019）。从这个意义上说，可信解释必须具有能够准确对应概率估计的特征权重，而概率估计又必须准确反映现实；也就是说，模型必须校准良好（Bhatt 等，2021）。

- 稳定性 和鲁棒性 是解释方法的另外两项关键特征（Dimanov 等人，2020年；Agarwal 等人，2022年；Alvarez-Melis 和 Jaakkola，2018年；Chromik 和 Schuessler，2020年）。稳定性指解释的一致性；同一个实例和模型在多次运行中应产生相同的解释（Slack 等人，2021；Carvalho 等人，2019b）。Zhou 等人（2021年）指出，稳定性是任何解释技术的基本特征之一；作者还写道，不稳定的方法被认为不可靠，因为它无法深入揭示模型的工作方式（Slack 等人，2021；Moradi 和 Samwald，2021；Hoffman 等人，2018；Carvalho 等人，2019b；Adadi 和 Berrada，2018；Wang 等人，2019；Mueller 等人，2019；Agarwal 等人，2022年）。相比之下，鲁棒性指的是解释方法即使在实例或其环境经受小幅扰动时，仍能产生一致结果的能力（Dimanov 等人，2020年；Doshi-Velez 和 Kim，2017年）。

因此，可解释人工智能中的解释方法应具备可靠、稳定和稳健的基本特征。

两种最成熟的局部特征重要性技术是局部可解释模型无关解释（LIME）（Ribeiro 等人，2016年）和 SHapley 加性解释（ SHAP ）（ Lundberg 和 Lee，2017 年）。LIME 使用代理模型，通过在局部使用加权线性模型来尝试解释每个实例。它对输入数据进行扰动，并观察其对模型预测的影响，从而创建局部忠实的解释。每个特征解释由一个条件和一个权重组成。该技术对扰动选择较为敏感，可能导致不稳定性和鲁棒性下降。多项研究已强调 LIME 中的不稳定性问题（Rahnama & Bostr¨om，2019年；Zhou 等人，2021年；Agarwal 等人，2022年）。

SHAP 值基于合作博弈论，通过考虑所有可能的特征组合，提供了一种统一的方法来解释每个特征对模型预测的影响。与 LIME 不同，SHAP 不提供任何规则或条件。相反，对任何特定实例的解释都是相对于基线而言的，这可能会带来解释方面的挑战（Molnar，2022）。另一方面，

这也使得可以汇总这些解释，从而获得全局解释。计算 SHAP 值的计算量很大，尤其是对于高维数据，且复杂性会随着特征数量的增加而上升。因此，要有效实现这一过程，可能需要使用专用库。

虽然 LIME 和 SHAP 已成为使用最广泛、知名度最高的两种解释方法，但也有其他解释方法旨在弥补不同方面的不足，例如：局部解释锚点（ANCHOR）（Ribeiro 等人，2018年）、模型无关监督局部解释（MAPLE）（Plumb 等人，2018年）以及局部规则解释（LORE）（Guidotti 等人，2018a）。ANCHOR 用于识别局部规则或条件，并侧重于找出最显著的特征及其阈值。MAPLE 是一种监督式邻域方法，结合了局部线性模型和决策树集成的思想。LORE 旨在生成可解释、易于人类理解的规则，通过反事实解释特定预测为何会产生。模型无关反事实解释（MACE）（Karimi 等人，2020年）是另一种反事实方法，尝试找出能够产生期望预测的特征集，同时使其与描述该实例的原始特征集之间的距离最小。Slack 等人（2021）提出了一种有趣的方法，其框架基于朴素贝叶斯，并通过在 SHAP（BayesSHAP）和 LIME（BayesLIME）中实例化该方法，生成带有不确定性信息且校准良好的解释。

2.2. 校准与 *Venn-Abers* 预测器

概率预测器会生成一个预测类别标签，并给出所有可能标签上的概率分布。为了确定预测概率分布的有效性，需要根据实际标签的观测结果进行统计检验。正如 Vovk 等人（2005）所讨论的，要使概率预测具有有效性通常是不可能的。然而，本文重点关注校准，其定义如下：

$$p(c \mid p^c) \approx p^c, \tag{1}$$

其中， p^c 表示特定类别标签 c的概率估计。校准良好的模型所生成的预测概率与观测准确率相符。这意味着，如果模型为某个特定标签分配了 0.9 的概率估计，那么该标签应有约 90% 的概率是正确的。校准良好的模型还会为每个实例和类别生成校准良好的概率估计。然而，许多预测模型生成的概率估计校准不佳（Van Calster 等，2019）。当模型校准不佳时，可以使用带标签数据中的一个独立部分（称为校准集），应用外部校准方法来调整预测概率。

Venn 预测器是一种概率预测器（Vovk 等，2004），针对每个标签输出多概率预测。这些多概率预测会转换为概率区间，区间大小表示对估计结果的置信度。

在归纳式范恩预测（Lambrou 等，2015）中，底层模型用于根据*Venn* 分类体系将校准实例划分为若干类别。对于每个类别，落入该类别的测试实例的估计概率，是该类别所有校准实例中各类别标签的相对频率。通过将测试实例纳入计算来确保有效性。由于测试实例的正确标签未知，因此会尝试每个可能的标签，从而得到多概率概率分布。

使用 Venn 预测器时，选择合适的分类体系可能颇具挑战。一种替代方法是*Venn-Abers* 预测器（VA）（Vovk & Petej，2012），其分类体系通过保序回归自动优化。保序回归是一种非参数回归技术，它通过拟合分段常数函数，使函数单调递增或递减。VA预测器与评分分类器搭配使用；由于 VA预测器属于 Venn 预测器，因此继承了其有效性保证。

要定义 VA 校准器，设 $Z_T = \{z_1, \ldots, z_n\}$，其中 $n = l + q$，为一个训练集。每个实例$z_i = (x_i,\ y_i)$由两部分组成；即一个 对象 x_i 和一个 标签 y_i。由于校准需要一个校准集，因此训练集被划分为适当的训练集 Z_q和校准集 $Z_l = \{z_1, \ldots, z_l\}$。训练一个评分分类器，使其在预测对象 x时输出一个预测分数 $s(x)$。值越高

表示对正类的信念越强，即类别标签为 1。要从评分分类器获得预测类别标签，需要将分数与固定的阈值 t 进行比较。当 $s(x) > t$时，预测结果取值为 1，否则为 0。对评分分类器使用 VA 时，阈值不是固定值 t。相反，会根据若干个真实目标已知的预测分数拟合一个保序回归模型，从而得到一个递增函数 $g(s(x))$（Zadrozny & Elkan，2001年），该函数被解释为 x 的标签为 1 的概率，即校准器。

因此，归纳式 VA 预测器会为测试对象 x[1] 及其对应的未知标签 y 生成如下多概率预测：

1. 令 g_0和 g_1分别为 $\{(s(x_1)\ y_1)\ldots,(s(x_l)\ y_l),(s(x),0)\}$和 $\{(s(x_1)\ y_1)\ldots,(s(x_l)\ y_l),(s(x)$ ，1)}。2. 令 $y=1$ 的概率区间为 $[g_0(s(x))\ g_1(s(x))]$（下文简称 $[p_0,p_1]$）。3. 当需要常规概率估计时，按照 Vovk & Petej (2012) 的建议，将概率区间 $[p_0,p_1]$ 聚合为单个校准概率估计，得到一个正则化值：

$$p=\frac{p_1}{1-p_0+p_1} \tag{2}$$

总之，VA 会生成一个校准概率估计（正则化 p-值）和一个概率区间 $[p_0,p_1]$。

3. 校准解释的定义

本节介绍一种名为校准解释（CE）的新型特征重要性解释方法。所提出的方法的核心在于认识到，特征权重必须具有直观含义，才能便于理解。CE 中的事实性解释由底层模型给出的校准预测构成，并附有不确定性区间以及一组事实性特征规则；每条规则均由带有不确定性区间的特征权重和事实性条件构成，涵盖该特征的实例值。反事实解释仅包含一组反事实特征规则；每条规则均由带有不确定性区间的预测估计和反事实条件构成，涵盖该特征的替代实例值。预测估计表示正类的概率估计。

以下是 CE 工作原理的概要说明：假设存在一个使用适当的训练集 Z_q训练得到的评分分类器，我们希望为测试对象 x 生成局部解释。使用 VA 对底层模型进行校准，以处理 x，从而获得概率区间 $[p_0,p_1]$ 和校准概率估计 p。使用 VA 对对象 x的轻微扰动版本进行估计，得到概率区间（$[p'_{0.f},p'_{1.f}]$）和校准概率估计（p'_f）；系统地每次改变一个特征（详见下文）。要获得特征 f的特征权重（及不确定性区间），计算 p与所有 p'_f（及 $[p'_{0.f},p'_{1.f}]$）的平均值[2]之间的差。

$$w_f=p-\frac{1}{|V_f|-1}\sum p'_f, \tag{3}$$

$$w_0^f=p-\frac{1}{|V_f|-1}\sum p'_{0.f}, \tag{4}$$

$$w_1^f=p-\frac{1}{|V_f|-1}\sum p'_{1.f}. \tag{5}$$

特征权重的精确定义是：原始测试对象 x 上的校准概率估计，与扰动后的对象上得到的校准概率估计（平均值）之间的差值

[1]为方便起见，为减少冗余，省略测试实例中的索引 $n+1$。[2]排除 p（以及 $[p_0,p_1]$），即 x 上的 VA 结果。

x 的各个版本。上限和下限的定义类似，使用扰动后的 x 各个版本的概率区间计算。只要使用相同的测试对象、底层模型和校准集，得到的解释也会相同。

More 形式上，通过以下步骤为测试对象 x生成事实性解释 :

1. 使用 VA 获取 x的概率区间 $[p_0, p_1]$ 和校准概率估计 p 。2. 将所有特征分为类别特征 C 和数值特征 N。为数值特征定义离散化器，以确定阈值以及小于或大于条件（$\leq, >$），适用于这些特征[3]。

3. 对每个特征 $f \in C$:

(a) 遍历所有可能的类别值 $v \in V_f$，每次将 x_f 的特征值替换为一个值，从而创建扰动后的实例 $x'_f = v$。(b) 计算并存储扰动后实例的概率区间 $[p'_{0.f}, p'_{1.f}]$和校准概率估计p'_f。使用公式 (3)、(4) 和 (5) 计算权重。(c) 使用特征 f、值 v 和恒等条件（$=$）定义事实性条件。

4. 对于每个特征 $f \in N$:

(a) 使用离散化器的阈值，找出围绕 x_f 的特征值、最接近的下限或上限阈值 t。将校准集中特征 f 的所有可能特征值分为两组 V_f，两组由 t 分隔。(b) 在每组中，提取代表 25th、 50th 和 75th 百分位数的百分位值 pv 。遍历 pv 中的值，逐一将其中一个值替换为 x_f 的特征值，从而创建扰动后的实例 $x'_{f_{pv}}$。计算并存储该扰动后实例的概率区间 $[p'_{0.f_{pv}}, p'_{1.f_{pv}}]$ 和校准概率估计 $p'_{f_{pv}}$。对组内所有扰动后实例取平均值，为每组生成一个概率区间$[p'_{0.f}, p'_{1.f}]$ 和校准概率估计 p'_f。使用公式 (3)、(4) 和 (5) 计算权重。(c) 使用阈值 t 和特征 f 定义一个事实性条件。使用 $\leq$ 或 $>$ 条件，使事实性条件覆盖该实例值。

讨论能够产生明确且可解释规则的规则逻辑至关重要。对于类别特征，唯一符合逻辑的事实性规则是基于恒等条件的规则。对于数值特征，则可以构建多种不同的规则。由于特征权重是 x 的校准概率估计与某一组内扰动后的实例的平均校准概率估计之间的差值，因此，如果特征规则可以是一个区间 规则，即采用 $0 <$ 特征 f ≤ 2 的形式，就会模糊特征权重的含义。原因在于，通常可以合理地假设，低于区间的取值（特征 f ≤ 0）所对应的概率可能与高于区间的取值（特征 f > 2）所对应的概率存在显著差异，从而使两者的平均值难以解释。因此，只有与特征值距离最近的阈值才会被用于构建规则。

该方法在最坏情况下的计算复杂性由以下因素决定:

- 训练底层模型并计算校准集上的概率估计。这项工作只需完成一次，其复杂性由底层模型决定。
- 初始化 VA 一次，其中涉及排序和预先计算一些值，最坏情况下耗时$O(l \log l)$，即取决于校准集的大小 l（Petej，2018年）。
- 该方法逐个处理特征，并针对类别特征遍历所有类别值，或针对数值特征遍历所有百分位数样本。每次迭代都会调用 VA，其复杂性为 $O(\log l)$，用于生成校准后的概率区间（Petej，2018年）。因此，该方法的复杂性为 $O(|F| i \log l)$，其中 $|F|$ 是特征数量， i 是

[3]这是通过使用 LIME 中 `EntropyDiscretizer` 类别的一个二元子类来实现的。

每个特征的迭代次数（即值或样本的数量）。通常，特征数量远少于实例数量，因此 $|F| << l$。值或样本的数量通常也较少（数值特征的默认样本数为 3）。

- 在调用 VA 时，也会再次调用底层模型以获取概率估计。该估计由底层模型确定。

根据定义，反事实解释利用 VA 对底层模型进行校准，从而生成校准良好的预测并量化不确定性。反事实解释利用 VA 的方式，也使该方法能够量化特征权重的不确定性。对于类别特征使用相等规则、对于数值特征使用二元规则，因此根据规则和实例值来解读特征权重既直观又明确。LIME 和 SHAP 依赖于生成大量扰动后的实例；与之不同，反事实解释依赖于校准集，而校准集必须取自与测试实例相同的分布。反事实解释在校准概率估计时，会利用从校准集中获得的有关模型的见解。依赖于与测试对象来自相同分布的校准集，有助于生成稳健的解释，因为底层分布在测试数据和校准数据之间保持一致，始终能够对解释进行平滑处理。根据校准集的大小（该校准集用于为每个特征训练少量保序回归模型），与 LIME 和 SHAP 等现有方法相比，生成校准良好的解释在大多数情况下速度较快。

3.1. 反事实校准解释

使用上述反事实解释的定义，可以轻松生成反事实规则。使用反事实校准解释（CCE）时，主要区别在于第 3c 步和第 4c 步中定义的条件是反事实条件，而不是事实性条件。这意味着，分类反事实条件（第 3c 步）使用 $\neq$ 条件。数值反事实条件（第 4c 步）定义为 $\leq$-条件和 $>$-条件，但所使用的条件会排除实例值。由于数值特征使用了非二元离散化器，数值反事实条件通常允许同时形成反事实 $\leq$-条件和 $>$-条件[4]。概率区间 $[p'_{0.f}, p'_{1.f}]$ 按照上述第 3 步和第 4 步中所述的反事实解释流程定义。为每个替代值定义一个反事实规则是非常简单的，每条规则的预期概率区间已由 $[p'_{0.f}, p'_{1.f}]$定义。公式（3）中定义的特征权重仅用于根据影响对反事实规则进行排序，即到p的距离。

在分类中，反事实解释通常指一种能够提供改变预测类别所需最小变更集合的方法。CE 和 CCE 以概率估计的形式提供解释，因此，为连续目标定义反事实解释更具相关性。因此，CCE 属于反事实解释，因为它会生成规则，描述由反事实条件引发的变化。由于这些解释基于概率估计而非类别标签，因此，对概率估计（或概率区间）的任何调整都可视为一种替代结果，即使它不会改变预测类别标签。需要注意的是，尽管 CCE 展示了对预测至关重要的特征值变化，但它可能不会明确指出数值特征的真正反事实（即导致变化的最小值）。CCE 强调对模型预测影响最大的变化，从而为模型的决策过程提供更细致的视角。

*3.2. calibrated-explanations*软件包

所提出的方法 calibrated-explanations[5] 以开源 Python 软件包的形式发布。此处介绍的版本为 $v0.2.3$。该软件包可以通过以下方式安装:

- pip install calibrated-explanations

[4]LIME 中所有现有的离散化器均可用于反事实解释。默认使用 `EntropyDiscretizer` 类别。[5] https://github.com/Moffran/calibrated explanations

•conda install -c conda-forge calibrated-explanations

要解释已训练的模型，可以使用 CalibratedExplainer，并将模型和校准集作为输入。要创建一组校准解释，可以使用 explain factual() 或 explain counterfactual() 方法，并将待解释的测试对象作为输入。或者，也可以使用 WrapCalibratedExplainer，并将尚未训练的模型作为输入，以集成的方式进行训练（使用 fit(train X, train y) 函数）、校准并创建解释。

`explain factual()` 或 `explain counterfactual()` 方法会输出一个解释集合（CalibratedExplanations），其中包含一组 FactualExplanation 或 CounterfactualExplanation 对象。可以使用 `plot all()` 方法绘制整个个体解释集合。若要访问单个解释，可以对该集合使用 `get explanation(index)`。`plot explanation()` 可用于绘制单个解释，而 `plot explanation(index)` 则可用于绘制集合中的指定解释。在绘制 FactualExplanation 时，所有绘图函数均接受一个额外参数 `uncertainty=True`，用于生成不确定性图。

有关如何使用该软件包的更多详细信息，请参阅文档或 notebook 文件夹中的示例。所使用的 Venn-Abers 实现是 Ivan Petej 编写的 venn-abers 软件包（Vovk 等，2015；Vovk & Petej，2014）。

4. 校准解释的呈现

可以从不同角度分析反事实解释（CE）和校准解释（CCE）。本文使用三种不同类型的图来展示事实解释和反事实解释。前两种用于事实解释：

- 常规解释，提供校准良好的解释，但不包含任何不确定性信息。这些解释可直接与其他特征重要性解释技术（如 LIME）进行比较。
- 不确定性解释通过提供校准良好的解释（包括不确定性区间），同时突出特征的重要性以及与其估计重要性相关的不确定性程度。通过在绘图函数中添加参数 uncertainty=True 来可视化不确定性。

这些图表的灵感来源于 LIME。尤其是，LIME 中的规则一直被视为解释中的有价值信息。基于上述原因，CE 会生成带有事实性解释的二元规则。

第三种图是反事实图，展示使用其他特征值时概率估计会受到怎样的影响。

绘制 CE 和 CCE 解释时，用户可以选择限制要显示的规则数量。对于事实解释，规则数量等于特征数量，因为每个特征对应一条规则。然而，对于反事实解释，CCE 会尽可能生成更多反事实规则，并按照影响力排序，从影响力最大的规则开始显示。

4.1. 事实性校准解释

4.1.1. 不考虑不确定性

常规 CE解释在多个方面与 LIME解释相似。首先，每个特征都有对应的条件规则，用于确定权重的定义条件（写在图的左侧）。其次，每条规则都有对应的特征权重，其定义方式是：将该权重与违反规则条件时得到的概率相加，结果即为该规则所涵盖实例的概率。因此，可以明确无误地理解规则的含义及其产生的权重。

图1展示了糖尿病数据集中的一个常规CE图，正类的校准概率估计约为0.6，这表明该患者应被归类为糖尿病患者 （顶部的两条柱状图对此进行了说明）。主图展示了特征权重，这些权重以0处的黑色实线为中心；该线对应于正类的校准概率估计 p ，并由*P(y*糖尿病患者) 柱中的实线表示

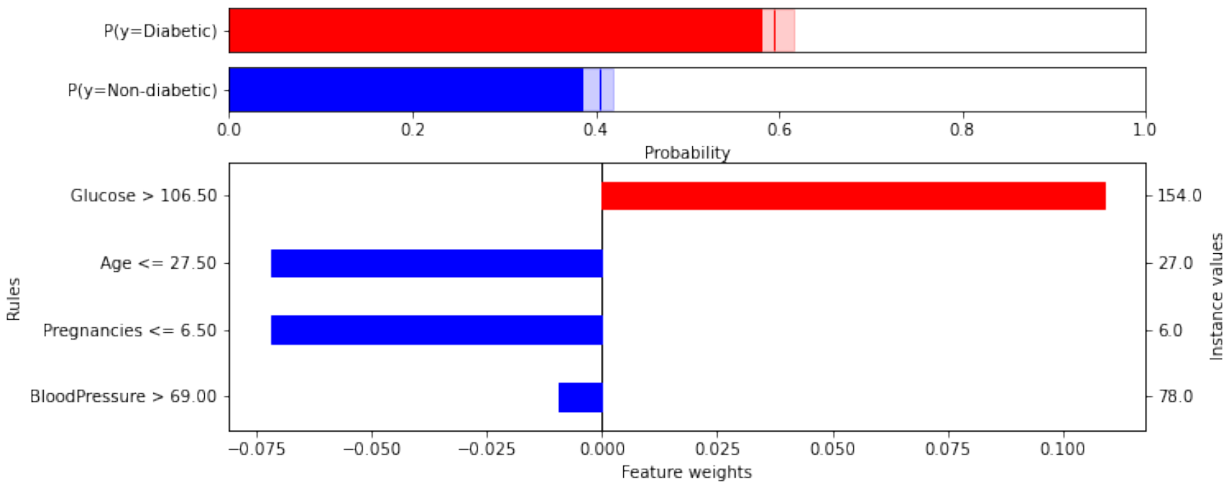

图1：来自 Diabetes 数据集、带有连续特征值的常规 CE图。事实性条件显示在特征图左侧，实际特征值显示在右侧。

在*P(y=*糖尿病患者*)* 柱中。在各项特征中，有一项特征明显促使患者被判定为糖尿病患者，而其余三项特征则促使患者被判定为非糖尿病患者。在此示例中，特征葡萄糖 对预测的影响最大，只要该特征值高于106.5，其权重为0.11，倾向于判定为糖尿病患者。年龄 和妊娠次数 分别是影响最大的第二和第三项特征；年龄较低以及妊娠次数少于6.5，都会使预测倾向于非糖尿病患者，影响约为0.07。血压 是被识别为会影响预测的第四项特征，同样使预测倾向于非糖尿病患者。由于各项权重指向相反方向，因此尽管概率估计相当接近，这一模棱两可的预测仍然可以得到解释。

对于特征具有类别值且目标为二元值（民主党或共和党）的数据集，如图2所示，规则看起来有所不同。在给定示例中，规则要么等于0，要么等于1，其中特征值由规则覆盖。该数据集取自1984年美国国会投票结果，其中0表示该国会议员对一项法律提案投了反对票，1表示投了赞成票。在图中，顶部特征 *physician-fee-freeze* 最能解释为什么该国会议员被预测为共和党。几乎所有特征都支持正类，这解释了为什么该预测对正类具有较高的概率估计。

4.1.2. 带有不确定性信息

由于 CE 会为权重创建不确定性区间，因此可以从不确定性中获取信息。当通过将参数 uncertainty= True 设为 True 来显示不确定性时，CE 会在图中添加区间，以展示预测的不确定性。较窄的区间表示估计值或权重具有较高的确定性，而较宽的区间则表示估计值或权重具有较高的不确定性。此外，特征区间还展示了特征对预测估计的影响。图3展示了与图1中相同实例的解释，并加入了不确定性信息。

底层模型的概率区间既通过顶部概率条中的浅蓝色或浅红色表示，也通过图表主体部分的浅灰色区域表示。类似地，特征权重以实色显示，从代表校准概率估计的黑线延伸至权重区间的下界；区间下界与上界之间则以浅色显示。规则按照校准预测的特征权重排序。蓝色条表示该特征支持非糖尿病类别，红色条表示支持糖尿病类别。不同特征权重的不确定性程度略有差异，其中最不重要的特征不确定性最大。

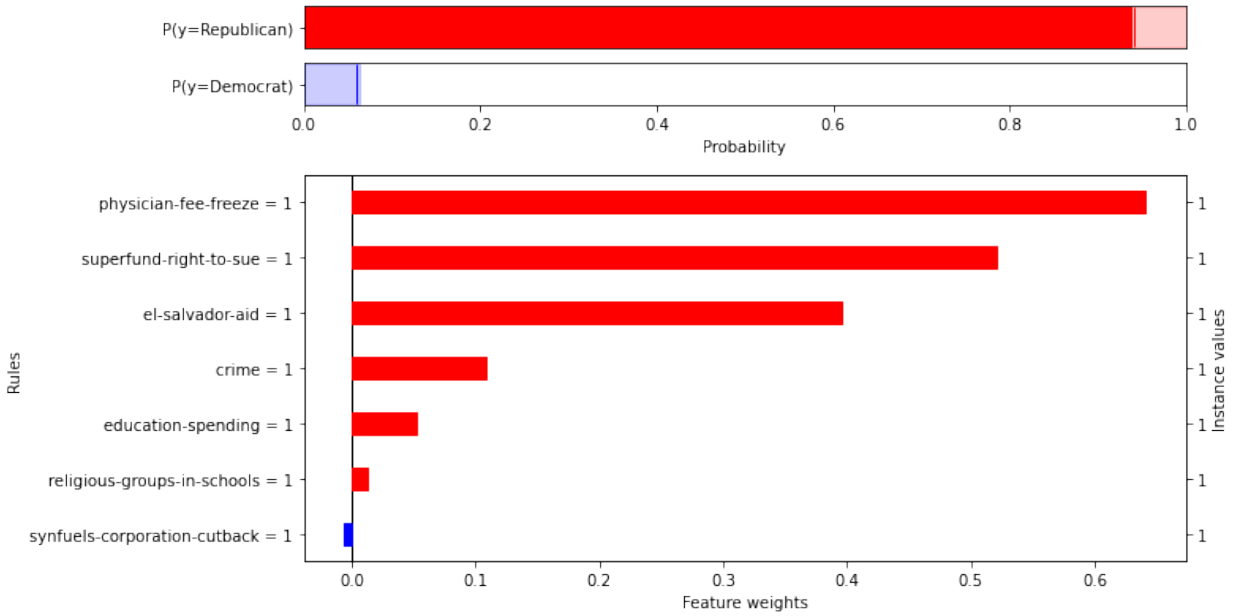

图2：Vote 数据集中的类别型特征值常规反事实解释图。

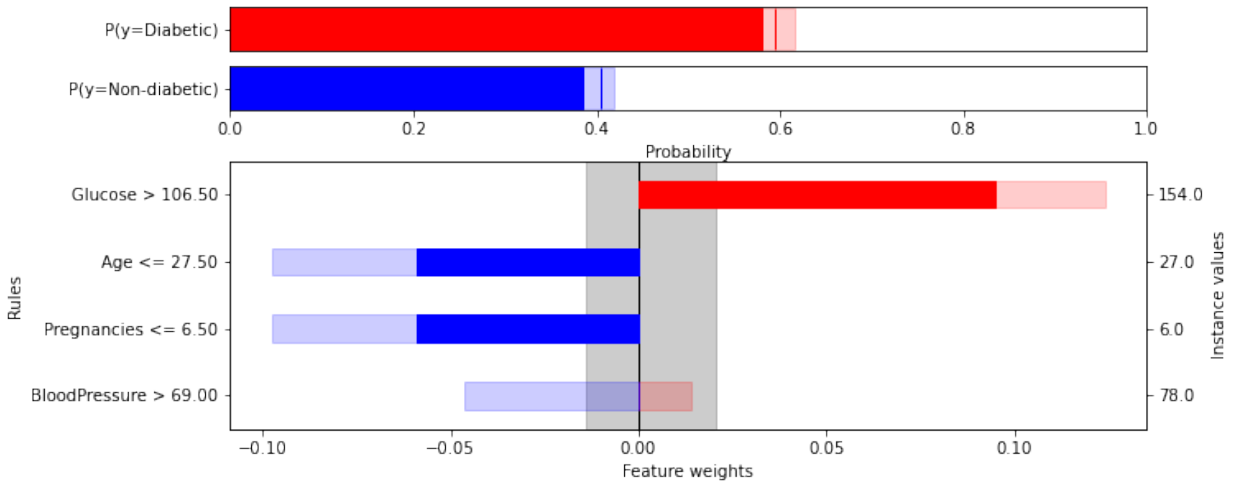

图3：Diabetes 数据集中的区间不确定性 CE 图。

血压的不确定性区间甚至涵盖了原始概率估计，表明该特征是否会影响预测尚不确定。

图4展示了不同的情况：不患肝脏疾病的概率介于两个类别之间。权重指向两个方向，不确定性区间也存在很大差异。mcv 和$sgot$ 特征的上界相同，而$sgot$ 的下界低得多，表明其不确定性程度大得多。

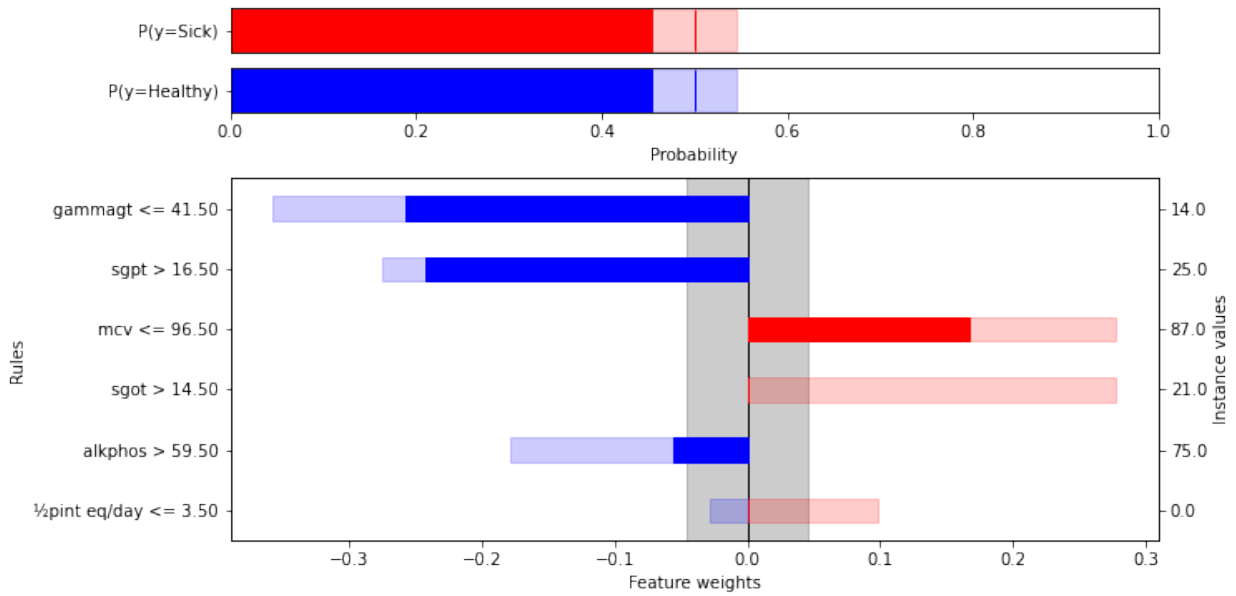

图4：来自 Liver 数据集的带区间不确定性 CE 图。

4.2. 反事实校准解释

在 CCE 中，图表显示的不是特征权重，而是聚焦于 VA 概率区间。每条规则都展示了这样得到的替代 VA 概率区间：将特征值扰动为反事实规则条件所涵盖的值。数值特征最多只能导致

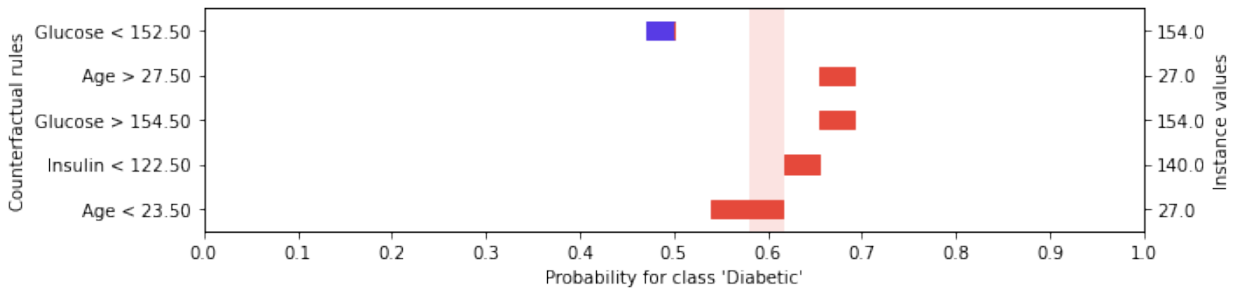

图5：来自 Diabetes 数据集的带连续特征值的 CCE 反事实图（与图3中的实例相同）。反事实条件显示在左侧，特征值显示在右侧。

分为两条反事实规则（位于特征值周围的阈值之上或之下），而对于每个替代类别值，都会创建一条反事实规则。在图中，仅显示会影响概率估计的反事实规则。

与不确定性图一样，反事实图显示，葡萄糖 和年龄 这两个特征对预测的影响最大。然而，在图中，规则表明，即使仅小幅降低葡萄糖 特征（葡萄糖 < 152.5），也可能使概率估计发生足够大的变化，从而改变模型的预测。较高的年龄 和葡萄糖 会增加患者为糖尿病患者的可能性。

在图6中，所选实例与图2中的相同，国会议员是共和党的预测结果明确。

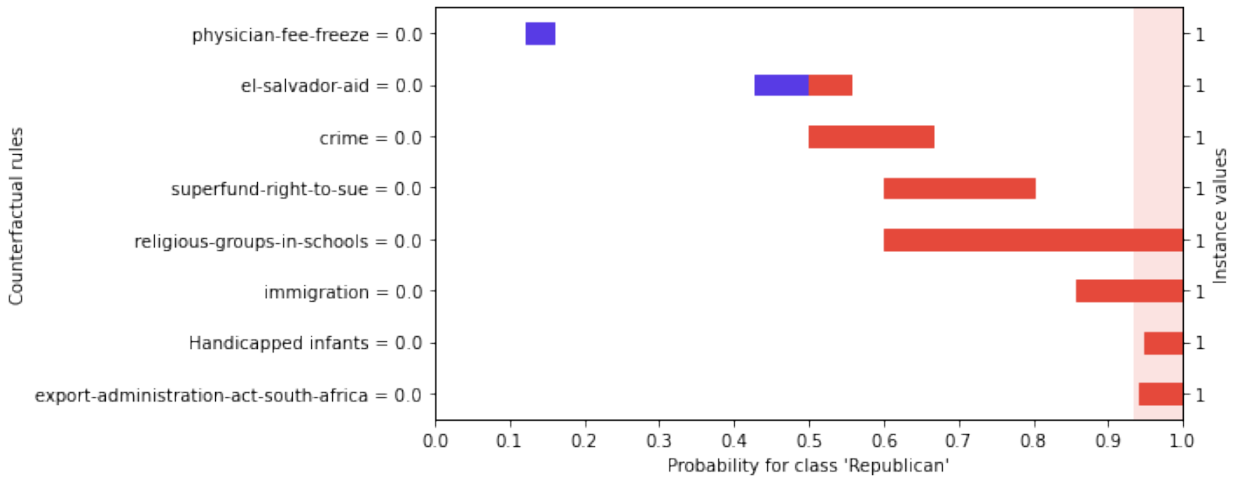

图6：数据集 Vote 中包含类别特征值的 CE 反事实图。

。这个实例有一些值得指出的有趣特征。首先，正如图2中所见，医师费冻结 特征对预测非常重要，按照反事实规则更改其值会改变预测。第二条反事实规则会大幅降低确定性，但无法明确指出模型会将其预测为民主党 还是共和党。其余几条反事实规则也会显著影响概率和不确定性，但不会改变预测。

5. 实验

5.1. 实验设置

为了验证 CE 的一些论断，我们进行了两项实验。这两项实验评估了该方法的稳定性和鲁棒性。稳定性通过以下实验进行评估：针对每个数据集，使用相同的模型、校准集和测试集进行解释，每个数据集重复 30 次。唯一的变化来源是随机种子。鲁棒性通过以下实验进行评估：在训练集和校准集被随机重新采样后，重新训练并解释一个新模型。该实验针对每个数据集运行 30 次，所有运行均使用相同的测试集。实验使用了一个包含 20 个分层实例的测试集（确保两个类别得到同等代表）。采用这种方式测量鲁棒性，是为了避免推断出与所解释测试实例不属于同一分布的扰动后的对象。作为与鲁棒性结果的比较，在同一个测试集上计算了每个模型的概率估计。我们预期，一个稳定且鲁棒的解释方法应当使特征权重具有较低的方差。实验使用了随机森林和 xGBoost，并对事实解释和反事实解释进行了评估。

使用了25个二分类问题，这些问题可从 UCI 代码仓库（Dua & Graff，2017年）或 PROMISE 软件工程代码仓库（Sayyad Shirabad & Menzies，2005）公开获取。数据集的特征见表1，其中$\#inst.$ 表示实例数，$\#attrib.$表示输入属性数。唯一值少于10个的属性被视为类别属性。

Name	#实例数	#属性数	名称	#实例	#归因
疝气病	357	60	kc2	369	22
creditA	690	43	kc3	325	40
糖尿病	768	9	肝脏	341	7
German数据集	955	28	pc1req	104	9
Haberman	283	4	pc4	1343	38
HeartC数据集	302	23	声纳	208	61
HeartH数据集	293	21	spect	218	23
Hearts	270	14	spectf	267	45
hepati	155	20	输血	502	5
iono	350	34	井字棋	958	28
je4042	270	9	vote	517	17
je4243	363	9	wbc	463	10
kc1	1192	22			

表1：数据集特征

用于运行实验的代码位于代码仓库的评估文件夹中，并附有补充材料。本文的实验文件命名为 Classification [..] stab rob。 - - -

5.2. 结果

表2中的结果是对30次运行和20个实例所测得的稳定性和鲁棒性的方差取平均值。方差按每个实例分别测量，并基于30次运行中最重要特征的特征权重计算；最重要特征定义为最常出现最高特征权重绝对值的特征。平均方差是对20个实例的方差取平均所得。之所以使用最重要特征，是因为该特征最有可能被用于决策，同时也是预期变化最大的特征（这是由于其权重的绝对值最高）。模型列表示校准前鲁棒性实验中底层模型概率估计的平均方差。

	稳定性				鲁棒性				
	xGB		RF		xGB		RF		
数据集	CE	CCE	CE	CCE	反事实解释	CCE	模型	反事实解释	CCE 模型
疝气病	3.9e-35	3.9e-35	7.8e-33	7.8e-33	0.015	0.015	0.015	0.017	0.017 0.010
creditA	1.5e-34	1.5e-34	1.5e-32	1.5e-32	0.022	0.022	0.011	0.015	0.015 0.009
糖尿病	1.2e-33	1.2e-33	5.0e-33	5.0e-33	0.017	0.017	0.014	0.015	0.015 0.013
German数据集	1.2e-34	1.2e-34	9.6e-34	9.6e-34	0.003	0.003	0.015	0.005	0.005 0.014
Haberman	5.2e-34	5.2e-34	1.2e-33	1.2e-33	0.010	0.010	0.016	0.011	0.011 0.016
HeartC数据集	0.0e+00	0.0e+00	2.7e-33	2.7e-33	0.012	0.012	0.017	0.012	0.012 0.016
HeartH数据集	2.6e-33	2.6e-33	4.9e-33	4.9e-33	0.017	0.017	0.017	0.011	0.011 0.016
HeartS数据集	1.6e-33	1.6e-33	4.3e-33	4.3e-33	0.019	0.019	0.018	0.017	0.017 0.017
hepati	1.9e-34	1.9e-34	5.2e-33	5.2e-33	0.022	0.022	0.018	0.014	0.014 0.017
iono	4.1e-35	4.1e-35	3.6e-33	3.6e-33	0.028	0.028	0.017	0.013	0.013 0.016
je4042	1.5e-33	1.5e-33	6.4e-33	6.4e-33	0.018	0.018	0.017	0.015	0.015 0.017
je4243	5.3e-34	5.3e-34	5.0e-33	5.0e-33	0.010	0.010	0.018	0.010	0.010 0.017
kc1	5.4e-34	5.4e-34	1.7e-33	1.7e-33	0.011	0.011	0.017	0.009	0.009 0.017
kc2	7.7e-34	7.7e-34	5.0e-34	5.0e-34	0.018	0.018	0.018	0.007	0.007 0.017
kc3	1.4e-35	1.4e-35	5.7e-34	5.7e-34	0.007	0.007	0.017	0.005	0.005 0.017
肝脏	2.2e-33	2.2e-33	8.7e-33	8.7e-33	0.025	0.025	0.017	0.030	0.030 0.017
pc1req	7.7e-35	7.7e-35	1.8e-33	1.8e-33	0.014	0.014	0.018	0.014	0.014 0.017
pc4	3.7e-34	3.7e-34	2.5e-33	2.5e-33	0.009	0.009	0.017	0.007	0.007 0.017
声纳	1.8e-34	1.8e-34	2.2e-33	2.2e-33	0.019	0.019	0.017	0.009	0.009 0.017
spect	0.0e+00	0.0e+00	1.5e-33	1.5e-33	0.007	0.007	0.017	0.006	0.006 0.016
spectf	4.6e-34	4.6e-34	4.2e-34	4.2e-34	0.006	0.006	0.016	0.005	0.005 0.016
输血 2.8e-34		2.8e-34	2.9e-33	2.9e-33	0.009	0.009	0.017	0.008	0.008 0.017
ttt	0.0e+00	0.0e+00	4.2e-32	4.2e-32	0.017	0.017	0.016	0.025	0.025 0.016
vote	0.0e+00	0.0e+00	6.5e-33	6.5e-33	0.014	0.014	0.016	0.011	0.011 0.016
wbc	5.7e-33	5.7e-33	2.5e-33	2.5e-33	0.021	0.021	0.016	0.018	0.018 0.015
平均值	7.6e-34	7.6e-34	5.4e-33	5.4e-33	0.015	0.015	0.016	0.012	0.012 0.016

表2：稳定性与鲁棒性

如表2所示，事实性解释 CE（CE）和反事实解释 CE（CCE）的稳定性实际上均为0，说明该方法按定义具有稳定性。极其微小的变化源于不同的随机种子，其影响可以忽略不计。

鲁棒性同样较低，尽管平均方差鲁棒性明显高于稳定性。鲁棒性与作为参考的概率估计的方差相当。这表明，该方法对于校准集和模型的变化等扰动具有相当好的鲁棒性。显然，由于该方法解释的是模型的校准概率估计，因此可以预期，该方法会对概率估计的变化较为敏感。

6. 讨论

6.1. 校准解释的特征

良好解释方法最重要的特征之一，是它必须具有可靠性。对于可解释人工智能方法而言，可靠性通常是指其对被解释模型的忠实度。然而，如果模型本身不可靠，即校准不佳、因而无法准确反映现实，那么高忠实度更可能带来危害而非益处。因此，要生成可靠的解释，解释必须对一个校准良好的模型具有较高忠实度。此外，可靠性还涉及不确定性这一层面。要使解释真正可靠，它必须能够揭示模型何时以较高置信度进行预测，以及何时以较低置信度进行预测。这一点超越了模型具有较高或较低概率估计的情况，因为概率估计可能很高，但模型可能没有见过许多相似案例，因而对该预测的置信度较低，反之亦然。所提出的方法的主要优势在于，它在可靠性的全部三个层面上都提供了较高的可靠性。

- 对模型进行校准，以确保其准确反映现实。这包括对模型概率估计的不确定性进行量化。
- 这些解释具有很高的忠实度，其特征权重是根据模型的校准概率估计精确定义的。特征权重也经过校准，以确保忠实度。
- 该方法对特征权重进行不确定性量化，从而深入了解每个权重的置信程度。对于反事实解释，该方法则对估计的概率区间进行不确定性量化。

稳定性 如前所述，是解释方法的一个关键方面，也是 LIME 主要因缺乏稳定性而受到批评的地方（参见例如 Rahnama & Bostr¨om (2019年)）。该方法在设计上具有稳定性，这一点在实验中得到了清楚的体现。只要使用相同的校准集和模型，通过 CE 或 CCE 为某个实例生成的解释就具有稳定性。

另一个重要特征是对变化的鲁棒性 。只要特征规则涵盖输入值中的扰动，且模型预测受到的影响不超过轻微程度，这些解释在设计上就具有鲁棒性，即保持一致。显然，如果扰动对模型预测造成很大影响，那么在此情境下就不能将其视为小扰动，因为它会对模型产生重大影响，因此，为了确保该方法可靠，也应对解释产生重大影响。依赖与测试对象来自相同分布的校准集，有助于生成鲁棒的解释，因为共享的底层分布始终会对这些解释进行平滑处理。评估该方法受校准变化和模型预测变化影响程度的实验表明，该方法会根据底层模型预测变化的幅度，相应地更新其特征权重。

6.2. 校准解释与最先进方法的比较

本文提出的方法能够生成校准良好的局部解释，并量化模型概率估计和特征权重的不确定性。如表3所示，只有 BayesSHAP 和 BayesLIME 同样具备生成校准解释和量化不确定性的能力

方法	局部	全局	事实性	事实性	反事实 校准	反事实 无关	模型不确定性 定量	
CE	X	-	X	X	X	X	X	
LIME	X	-	X	-	-	X	-	
SHAP	X	X	X	-	-	X	-	
MAPLE	X	-	X	-	-	X	-	
BayesLIME	X	-	X	-	X	X	X	
BayesSHAP	X	X	X	-	X	X	X	
ANCHOR	X	-	X	-	-	X	-	
MACE	X	-	-	X	-	X	-	
LORE	X	-	-	X	-	X	-	

表3：反事实解释与事实性及反事实最先进解释方法的特征（Ribeiro 等人，2016年；Lundberg 和 Lee，2017年；Plumb 等人，2018年；Ribeiro 等人，2018年；Slack 等，2021；Guidotti 等人，2018a；Karimi 等人，2020年）。

量化。这些方法是将贝叶斯框架应用于 SHAP 和 LIME 的实例化，而非新方法。

最先进方法中没有一种能够同时生成事实解释和反事实解释（见表3），这进一步凸显了反事实解释的全面性。值得指出的是，所有反事实方法都无法生成校准良好的解释，也无法进行不确定性量化。

7. 结论

本文介绍了校准解释 （反事实解释），这是一种新颖的解释方法，不仅能够校准底层模型，还能在解释中提供不确定性信息。每个特征的权重由其对正类校准概率估计的贡献来定义，而解释则通过直观明了的条件规则呈现。

反事实解释（CE）并不止于事实解释；它还能够自然地生成反事实规则，并深入揭示每种反事实结果的不确定性。该方法以开源软件包的形式提供，可通过 pip 或 conda 安装，并支持事实解释和反事实解释的绘图。

在25个基准数据集上开展了两项实验，以评估该方法的鲁棒性和稳定性。在评估中，反事实解释（CE）展现出卓越的性能：在运行时间方面效率出色，凭借内在的校准和不确定性量化能力而具有可靠性，能够通过一致的校准集和模型保持稳定性，并且能够抵御小幅扰动。从本质上说，校准解释体现了高质量解释方法所具备的典型特征。

反事实解释既能生成事实解释，也能生成反事实解释，与其他最先进方法相比，这一点相当独特。论文研究的其他方法都无法像反事实解释那样提供如此丰富的特性：它们要么侧重于事实解释，要么侧重于反事实解释。除反事实解释外，只有一种方法（包含两种实例化）同时纳入了不确定性信息和校准解释。

展望未来，后续工作包括开展真实场景评估，以了解用户感知，并支持图像和文本数据。改进当前可用的图表，并引入更多可视化方法以增强规则洞见，也是重点工作。

致谢

Helena L¨ofstr¨om 是瑞典博罗斯大学数字零售工业研究生院（INSiDR）的博士生，其研究由瑞典知识基金会资助，资助编号20160035。作者感谢瑞典知识基金会和工业合作伙伴为延雪平大学（瑞典）的知识密集型产品实现 SPARK 研究与教育环境提供资金支持。项目：AFAIR 资助编号20200223，以及 PREMACOP 资助编号20220187。

利益冲突

作者声明，我们没有需要披露的利益冲突。

写作过程中对生成式 AI 和人工智能辅助技术的声明

在准备本作品期间，作者使用了 ChatGPT、Bard 和 Bing Chat，以获取有关如何改进和简化描述与句子的建议。使用这些工具/服务后，作者根据需要审阅并编辑了内容，并对出版物的内容承担全部责任。

参考文献

Adadi, A., & Berrada, M. (2018年). 深入黑箱：可解释人工智能（XAI）综述. *IEEE Access*, *6*, 52138–52160.

Agarwal, C., Saxena, E., Krishna, S., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., & Lakkaraju, H. (2022年). OpenXAI：迈向模型解释的透明评估. *arXiv*预印本 *arXiv:2206.11104*, .

Alvarado-Valencia, J. A., & Barrero, L. H. (2014). 判断性预测中的依赖、信任与启发式。*Computers in Human Behavior*，*36*，102–113。

Alvarez-Melis, D., & Jaakkola, T. S. (2018). 关于可解释性方法的鲁棒性。*arXiv* 预印本 *arXiv:1806.08049*，. Antor´an, J., Bhatt, U., Adel, T., Weller, A., & Hern´andez-Lobato, J. M. (2020). 找到线索：一种解释不确定性估计的方法。*arXiv* 预印本 *arXiv:2006.06848*，. Arrieta, A. B., D´ıaz-Rodr´ıguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garc´ıa, S., Gil- L´opez, S., Molina, D., Benjamins, R. 等 (2020). 可解释人工智能（xai）：概念、分类体系、机遇与挑战——迈向负责任的人工智能。*Information Fusion*，*58*，82–115。Bhatt, U., Antor´an, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melanˆcon, G., Krishnan, R., Stanley, J., Tickoo, O. 等 (2021). 不确定性作为一种透明性形式：衡量、传达和利用不确定性。载于*2021*年*AAAI/ACM*人工智能、伦理与社会会议论文集（第401–413页）。Buˆcinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). 在评估可解释人工智能系统时，代理任务和主观测量可能会产生误导。载于第*25*届智能用户界面国际会议论文集 （第454–464页）。Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019a). 机器学习可解释性：方法与指标综述。*Electronics*，*8*，832。网址：http://dx.doi.org/10.3390/electronics8080832。doi:10.3390/electronics8080832。Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019b). 机器学习可解释性：方法与指标综述。*Electronics*，*8*，832。Chromik, M., & Schuessler, M. (2020). 面向可解释人工智能中黑箱解释的人类受试者评估分类体系。载于*ExSS-ATEC@ IUI*。Das, A., & Rad, P. (2020). 可解释人工智能（xai）的机遇与挑战：综述。*arXiv* 预印本 *arXiv:2006.11371*，.

David Gunning (2017). 可解释人工智能。网页。网址：https://www.darpa.mil/attachments/XAIProgramUpdate.pdf 美国国防高级研究计划局。Dimanov, B., Bhatt, U., Jamnik, M., & Weller, A. (2020). 你不该相信我：学习能够对多种解释方法隐藏不公平性的模型。人工智能与应用前沿：*ECAI 2020*，. Doshi-Velez, F., & Kim, B. (2017). 迈向一门严谨的可解释机器学习科学。*arXiv* 预印本 *arXiv:1702.08608*，. Dua, D., & Graff, C. (2017). UCI机器学习知识库。网址：http://archive.ics.uci.edu/ml。Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendahl Jr, K. C. (2017). 过拟合且过度自信的预测集成。《管理科学》，*63*，1110–1130。Guidotti, R. (2022). 反事实解释及其发现方法：文献综述与基准测试。《数据挖掘与知识发现》，（第1–55页）。Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018a). 黑箱决策系统的局部规则解释。arXiv:1805.10820。Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018b). 黑箱模型解释方法综述。《*ACM*计算调查》*(CSUR)*，*51*，1–42。Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). 可解释人工智能的度量指标：挑战与前景。*arXiv* 预印本 *arXiv:1812.04608*，. Karimi, A.-H., Barthe, G., Balle, B., & Valera, I. (2020). 面向重大决策的模型无关反事实解释。载于人工智能与统计国际会议（第895–905页）。PMLR。Lambrou, A., Nouretdinov, I., & Papadopoulos, H. (2015). 归纳式范恩预测。《数学与人工智能年刊》，*74*，181–201。L¨ofstr¨om, H., Hammar, K., & Johansson, U. (2022). 解释方法质量评估标准的元综述。载于 J. De Weerdt, & A. Polyvyanyy (编)，智能信息系统（第55–63页）。Cham：施普林格国际出版公司。Lundberg, S. M., & Lee, S.-I. (2017). 一种解释模型预测的统一方法。载于第*31*届神经信息处理系统国际会议论文集（第4768–4777页）。Martens, D., & Foster, P. (2014). 解释数据驱动的文档分类。*MIS Quaterly*，*38*，73–100。Marx, C., Park, Y., Hasson, H., Wang, Y., Ermon, S., & Huan, L. (2023). 但你确定吗？一种面向可解释人工智能的不确定性感知视角。载于人工智能与统计国际会议（第7375–7391页）。PMLR。Molnar, C. (2022). 可解释机器学习。（第2版）。Leanpub。网址：https://christophm.github.io/interpretable-ml-book。Moradi, M., & Samwald, M. (2021). 使用置信项集对黑箱分类器进行事后解释。《应用专家系统》，*165*，113941。Mothilal, R. K., Sharma, A., & Tan, C. (2020). 通过多样化反事实解释来解释机器学习分类器。载于*2020*年公平性、问责制与透明度会议论文集（第607–617页）。

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G.（2019年）。人类—人工智能系统中的解释：文献元综述、关键思想与出版物概要，以及可解释人工智能参考书目。*arXiv*预印本 *arXiv:1902.01876*。Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B.（2019年）。可解释机器学习的定义、方法与应用。美国国家科学院院刊，*116*，22071–22080。Parmigiani, G.（2002）。测量复杂决策分析模型中的不确定性。《医学研究统计方法》，*11*，513–537。Petej, I.（2018）。具有有效性保证的小规模和大规模概率分类器。伦敦大学皇家霍洛威学院博士论文。Phillips-Wren, G., & Adya, M.（2020年）。压力下的决策：信息过载、时间压力、复杂性与不确定性的作用。《决策系统杂志》，*29*，213–225。Plumb, G., Molitor, D., & Talwalkar, A. S.（2018）。模型无关监督局部解释。神经信息处理系统进展，*31*。Rahnama, A. H. A., & Bostr¨om, H.（2019年）。关于LIME框架中数据偏移与标签偏移的研究。*arXiv*预印本 *arXiv:1910.14421*。Ribeiro, M. T., Singh, S., & Guestrin, C.（2016年）。"我为什么应该相信你？"：解释任何分类器的预测。载于第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集KDD '16（第1135–1144页）。美国纽约州纽约市：美国计算机协会。doi:10.1145/2939672.2939778。Ribeiro, M. T., Singh, S., & Guestrin, C.（2018年）。Anchors：高精度模型无关解释。载于*AAAI*人工智能会议论文集。第32卷。Sayyad Shirabad, J., & Menzies, T.（2005）。PROMISE软件工程数据库库。加拿大渥太华大学信息技术与工程学院。Slack, D., Hilgard, A., Singh, S., & Lakkaraju, H.（2021）。可靠的事后解释：对可解释性中的不确定性进行建模。神经信息处理系统进展，*34*，9391–9404。Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., & Steyerberg, E. W.（2019年）。校准：预测分析的致命弱点。《*BMC*医学》，*17*，1–7。Vovk, V., Gammerman, A., & Shafer, G.（2005）。《随机世界中的算法学习》。柏林、海德堡：施普林格出版社。Vovk, V., & Petej, I.（2012）。Venn-Abers 预测器。*arXiv*预印本 *arXiv:1211.0025*。Vovk, V., & Petej, I.（2014）。Venn-Abers 预测器。arXiv:1211.0025。Vovk, V., Petej, I., & Fedorova, V.（2015）。有和无有效性保证的大规模概率预测器。arXiv:1511.00213。Vovk, V., Shafer, G., & Nouretdinov, I.（2004）。自校准概率预测。载于神经信息处理系统进展（第1133–1140页）。Wachter, S., Mittelstadt, B., & Russell, C.（2017年）。无需打开黑箱的反事实解释：自动化决策与《通用数据保护条例》。*Harv. JL & Tech.*，*31*，841。

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y.（2019年）。设计理论驱动、以用户为中心的可解释人工智能。载于*2019*年*CHI*计算机系统人因大会论文集CHI ’19（第1–15页）。美国纽约州纽约市：美国计算机协会。URL：https://doi.org/10.1145/ 3290605.3300831。doi:10.1145/3290605.3300831。Zadrozny, B., & Elkan, C.（2001年）。从决策树和朴素贝叶斯分类器中获得校准概率估计。载于第*18*届国际机器学习大会论文集（第609–616页）。Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A.（2021年）。评估机器学习解释的质量：关于方法与指标的综述。*Electronics*，*10*，593。Zhou, Z., Hooker, G., & Wang, F.（2021年）。S-LIME：用于模型解释的稳定化LIME。载于第*27*届*ACM SIGKDD*知识发现与数据挖掘会议论文集（第2429–2438页）。