arXiv:2301.02499v1 [stat.ML] 6 Jan 2023

使用 Pearl反事实方法评估反事实解释

Bevan I. Smith南非约翰内斯堡威特沃特斯兰德大学机械、工业与航空工程学院 bevan.smith@wits.ac.za

摘要

反事实解释（CEs）是用于生成替代情景的方法，能够产生不同的理想结果。例如，如果预测某位学生无法通过一门课程，反事实解释就可以为该学生提供其他途径，使其有望被预测为能够通过课程。反事实解释的应用非常广泛。然而，目前反事实解释是基于机器学习模型生成的，而这些模型未必会考虑数据中真实的因果结构。这样做可能会使反事实解释结果产生偏差。本研究拟采用 Judea Pearl 的反事实计算方法来检验反事实解释；令人意外的是，迄今为止，反事实解释领域的文献中尚未见到这种做法。此外，我还在三种不同的因果结构上评估这些反事实解释，以展示真实的底层因果结构如何影响生成的反事实解释。本研究提出了一种使用 Pearl的方法评估反事实解释的方法。研究结果表明（尽管样本量有限），有百分之三十的反事实解释与使用 Pearl的方法计算出的结果相矛盾。这表明我们不能盲目信任反事实解释；在直接使用原始机器学习模型计算反事实之前，了解真实的因果结构至关重要。

关键词 反事实解释 · 结构因果模型（SCM） · 反事实

1 背景

本研究提出了一种使用 Pearl反事实方法 [1, 2] 验证反事实解释的方法。反事实解释（CEs）是快速发展的可解释机器学习领域的一部分，该领域旨在解释机器学习模型为何做出其预测 [3, 4]。如果模型做出的预测并不理想，那么了解其做出预测的原因，就能帮助我们生成反事实解释，为我们提供其他理想结果。例如，如果预测某名学生会不及格，我们可以生成反事实解释，建议该学生如何改变某些特征，以提高其通过概率。反事实解释（CE）“描述了能够将预测改变为预定义输出的特征值的最小变化”[3]。显然，我们可以看到，反事实解释前景广阔，应用范围也十分广泛，例如提高通过率、建议患者改变哪些方面以改善健康状况、建议客户采取哪些措施以获得银行贷款，等等。然而，主要问题在于，反事实解释及其生成算法尚未得到适当验证

[5]。为什么需要对它们进行验证或测试？这是因为，使用 DiCE 和 WACH 等各种优化算法生成的反事实解释，[6, 4], 是基于使用原始数据训练的模型生成的 [3]。这会带来问题，因为机器学习（ML）模型并不关心特征之间的因果关系和因果结构，只关注相关性 [1]。机器学习模型的设计目标是做出准确的预测，而不考虑因果性。因此，如果生成反事实解释的算法基于这些机器学习模型，那么它们生成的反事实解释会与通过真实结构因果模型生成的结果相同吗？我们已经知道，在拟合简单线性回归模型时，如果不考虑因果结构，系数就会产生偏差 [7]。例如，在回归模型中遗漏混杂特征，或在模型中纳入碰撞结构。如果在回归模型中不考虑因果关系会导致问题，那么使用现有的、已训练好的机器学习模型来生成反事实解释，问题又会有多严重呢？本研究旨在解决以下问题：反事实

解释是使用机器学习模型生成的，而这些模型未必会考虑数据中真实的潜在因果结构。这可能导致对反事实的估计出现错误。

为了解决这个问题，我提议使用 Judea Pearl[1, 2]研究中提出的反事实方法。该方法的详细介绍见第 5 节。不过，其核心思想是：要计算反事实，我们首先需要一个结构因果模型（SCM）及其相关图，用于描述真实的数据生成过程。这样，我们便可以根据数据之间真实的因果关系计算反事实，无论这种关系存在于输入特征之间，还是输入特征与输出特征之间。利用机器学习模型计算反事实解释时，未必会考虑因果关系，而 Pearl的方法则会考虑。我提议使用该方法来评估和检验现有的反事实解释方法。

本文的主要贡献如下：

- 我提出了一种使用 Pearl 开发的反事实方法评估反事实解释的方法。迄今为止，这种方法竟然还未应用于反事实解释文献中。借助这种方法，我们可以评估反事实解释，了解它们是否值得信赖。
- 使用这种方法，我展示了当前方法生成的反事实解释与通过Pearl的方法计算出的结果之间存在何种冲突，以及不同的因果结构会如何影响结果。这提醒我们，在生成反事实解释之前，理解真实的底层因果结构至关重要。

2 反事实解释

近年来，反事实解释（CE）在机器学习文献中备受关注 [4, 3, 8, 5]。反事实解释旨在操控模型的输入，使输出变为期望的结果。生成反事实解释时，需要求解一个优化问题：对输入特征空间进行扰动，使其尽可能接近原始特征空间，同时能够产生不同且理想的输出 [8, 3]。因此，这是一个旨在尽可能少地扰动输入空间，以改变输出的优化问题。

2.1 Wachter等人的方法

Wachter等人提出的方法就是一种反事实解释方法 [4]。其损失函数如下：

$$L(x, x', y', \lambda) = \lambda \cdot (f(x') - y')^2 + d(x, x'), \tag{1}$$

其中， x 和 x' 分别是原始输入空间和反事实输入空间， y' 是反事实输出， λ 是一个调节参数。值得注意的是， f 指的是基于数据训练的原始机器学习模型，该模型不一定会考虑数据中的因果关系。 L 旨在平衡以下两者：最小化原始输入与反事实 x 之间的距离，以及预测的 f（x'）与原始 y' 输出之间的距离。

2.2 Mothilal等人的方法

Mothilal等人开发了另一种反事实解释方法 [6] ，也是本研究采用的方法。该方法称为 DiCE：**多**样性**反**事实**解**释。他们的优化方法基于以下损失函数：

$$C(x) = \underset{c_1,...,c_k}{\arg\min} \frac{1}{k} \sum_{i=1}^{k} yloss(f(c_i), y) + \frac{\lambda_1}{k} \sum_{i=1}^{k} dist(c_i, x) - \lambda_2 dpp_diversity(c_1, ..., c_k), \tag{2}$$

其中， c_i 是反事实解释（CE），k 是反事实解释的总数，*f*(.) 指基于数据训练的机器学习模型，*yloss*(.) 是一种用于最小化反事实预测与期望结果之间距离的度量，*dist*(.) 是反事实输入 c_i 与实际输入 *x*之间的距离，*dpp_*多样性(.) 用于最大化多样性，最后，lambda 值用于平衡损失函数的三个部分。本研究的一个重要方面是说明上述反事实解释方法使用的是基于数据训练的机器学习模型。

2.3 反事实解释的特征

反事实解释需要具备一些重要特征，以下是我着重强调的几点：

• 可行性/合理性：指反事实解释符合现实，不会小于或大于原始数据中观察到的值 [9]。这意味着反事实解释必须来自一个可能世界 [4]。
• 可行动：特征必须能够修改、能够改变。不可行动的特征包括年龄、性别、种族等。 [9]。

• 因果性：我会在讨论部分（第 8 节）中对此进行更详细的讨论。这一特征要求反事实解释中保持因果关系。不过，我对这一要求持不同意见。

3 结构因果模型

结构因果模型（SCM）指的是真实的数据生成过程，并支持反事实分析 [5, 1]。SCM 与称为有向无环图（DAG）的图形因果模型相关，如图 1 右侧所示。考虑一组特征 X_1，.....， X_n，称为内生特征，其中每个特征 X_i 都通过确定性函数 f_i 生成，该函数取决于其原因（或父节点，PA）。 PA_i 指的是 X_i的父节点（已知原因），而 U_i 指的是 X_i [10]的某个随机未解释原因，也称为外生特征（模型之外的因素）。每个特征都对应有向无环图中的一个顶点。

$$X_i := f_i(PA_i, U_i) \qquad (i = 1, ..., n) \tag{3}$$

图 1 展示了基于机器学习模型生成反事实（左图）与基于真实结构因果模型有向无环图生成反事实（右图）之间差异的一些直观解释 [11]。结构因果模型会依据真实因果结构生成反事实，而反事实解释则基于原始机器学习模型生成 f，该模型不一定会考虑因果结构。问题是：反事实解释算法会生成与通过结构因果模型生成的反事实相同的结果吗？

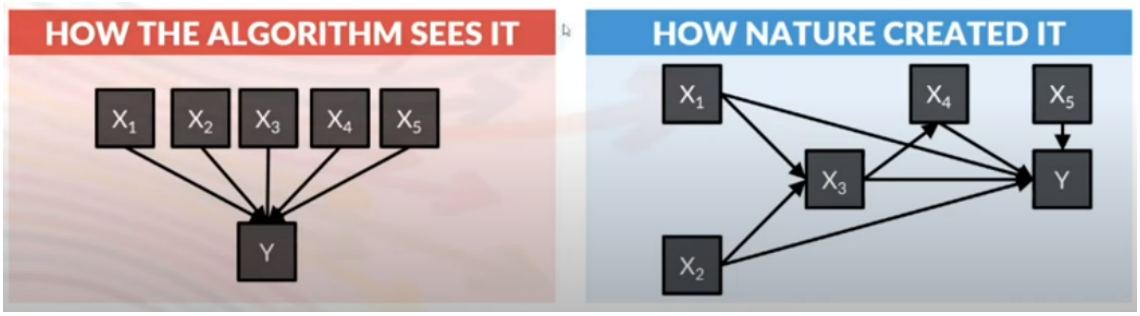


图 1：机器学习模型（左）与结构因果模型（右） [11]。

4 因果结构

在第 5 节介绍珀尔反事实方法（PCM）之前，我先介绍因果推断中的三种重要因果结构：链（或中介变量）、分叉（共同原因）和碰撞结构。本研究将在使用 PCM 评估反事实解释时运用这三种结构。根据 Pearl 的说法，这三种类型是因果结构的基本构件，“使我们能够检验因果模型、发现新模型、评估干预的效果，等等” [2]。本文将通过有向无环图和结构因果模型来呈现这些结构。

4.1 链（中介变量）

链或中介变量有向无环图如图 2 所示。链的一个例子是火灾（X）、烟雾（Z）和警报器（y）系统。火灾不会直接导致警报器响起，但必须先产生烟雾，再由烟雾触发警报器。在这种情况下，火灾与警报器之间不存在直接的因果路径。当然，在其他情况下，X 与 y 之间可能存在因果路径。这里我们将范围限定在基本链上。该模型告诉我们，如果控制了Z，就会阻断X 与y 之间的因果流，从而无法测量X 对y的真实影响。也就是说，在给定Z的条件下，X与y 条件独立。实际而言，这意味着如果我们将X、Z 和y 纳入模型，就等同于控制Z，并阻断X 与y之间的路径。图 2 的结构因果模型（SCM）见公式 4 和 5。这里我们引入外生变量U，它们对于后文介绍的 Pearl的方法至关重要。Z 是X的函数

以及 U_z，其中X 是一个内生特征（例如火灾）， U_z 指的是所有导致Z、但我们无法逐一纳入考虑的外部特征。后文提到的所有U 特征也是如此。

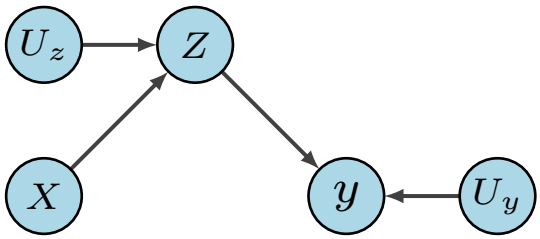


图 2：展示X 与Z 以及y之间链式路径的有向无环图。因果流从X 指向Z，再指向y。

$$Z = f(X, U_z) \tag{4}$$
$$y = f(Z, U_y) \tag{5}$$

4.2 分叉（共同原因）

图 3 展示了一个表示分叉或共同原因因果结构的有向无环图。实际上，它表示一种伪相关：给定Z时，X 与y 条件独立。也就是说，在这种情况下，要获得X 对y 的真实因果效应，就需要对Z进行条件化（控制）。图 3 对应的结构因果模型见方程 6 至 8。这种特定因果结构的一个例子是，X 指鲨鱼袭击，y 指冰淇淋销售。我们可能会获得数据显示这两个事件相关：鲨鱼袭击增加时，冰淇淋销售也会增加。然而，解释这种伪相关的是Z，它指的是夏季。鲨鱼袭击和冰淇淋销售都与季节条件相关。

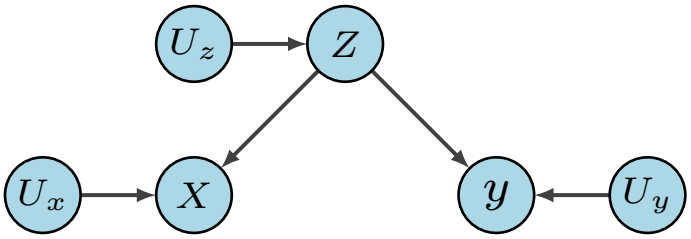


图 3：显示共同父节点混杂的有向无环图，其中Z 是X 和y的父节点，但在这种情况下，X 与y 之间没有因果流。也称为伪相关。

$$X = f(Ux) \tag{6}$$
$$Z = f(Uz) \tag{7}$$
$$y = f(Z, Uy) \tag{8}$$

4.3 碰撞结构

本研究采用的最终因果结构是碰撞结构，如图4中的有向无环图所示，X 与y之间不存在因果路径。碰撞结构的一个例子是：X 是洒水器，y 是降雨，Z 是湿草地。可以说，每个特征都处于开启或关闭状态。因此，Z 是洒水器和降雨的函数。例如，如果 Z = 是干燥的（关闭），那么洒水器和降雨都

（X）都必须处于关闭状态。然而，如果Z = 是湿润的，那么X 或y 处于开启状态，或者二者都处于开启状态。这里需要注意的重要一点是，X 和y 在边际上相互独立；也就是说，在不以任何变量为条件时，它们相互独立。例如，是否降雨与洒水器是否开启无关。然而，当我们以Z为条件时，洒水器和降雨就会条件相关。这意味着，以草地为条件会造成一种错觉，仿佛原本相互独立的特征之间存在因果关系。例如，如果我们知道草地是湿的（Z = 开启），而且正在下雨（y = 开启），那么我们就能立即知道洒水器的状态（X =关闭）。两个相互独立的变量（X 和y）在给定z后会变得条件相关。这与链和分叉的情况相反；在链和分叉中，以Z 为条件会使我们估计出X 与y之间真实的因果关系。碰撞结构的结构因果模型见方程9至10。

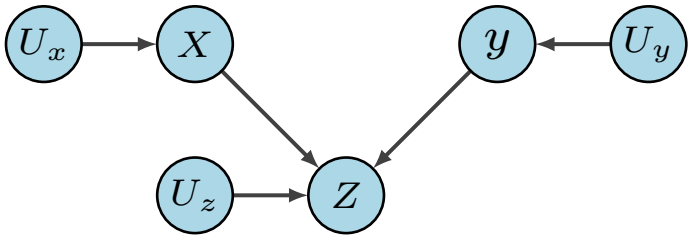


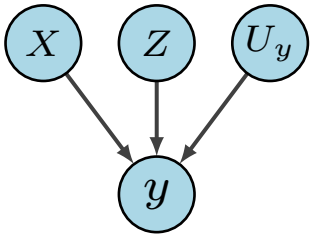
图 4：显示 X、Z 和 y 之间存在碰撞路径，但 X 与 y之间不存在因果路径的有向无环图。

$$Z = f(X, y, U_z) \tag{9}$$

$$X = f(U_x) \qquad y = f(U_y) \tag{10}$$

4.4 结构因果模型与机器学习的区别

从这三种基本因果结构中需要注意的关键一点是：结构因果模型会考虑因果结构（即数据生成过程），而机器学习模型则从X、y 和Z的联合分布中采样，并不考虑因果结构。显然，这三种类型截然不同。例如，机器学习模型的有向无环图和线性模型可能如图5和方程11所示。问题是，如果我们从机器学习模型生成反事实解释，它们会与接下来所示的、使用Pearl的方法从真实结构因果模型生成的反事实解释相同吗?



F图 5：显示监督式机器学习模型如何基于数据进行训练的有向无环图。

$$y = f(X, Z, Uy) \tag{11}$$

5 Pearl 反事实方法（PCM）

现在我们来介绍珀尔反事实方法（PCM）。本节将介绍如何使用结构因果模型生成反事实。该方法基于Judea Pearl的研究，相关内容见《为什么之书》第8章 [2]以及《统计因果推断：入门》第4章 [1]。该方法包含三个步骤：溯因、行动和预测，具体如下。

5.1 溯因：计算外生变量，U

考虑具有已知真实因果结构的数据。该结构包含一个有向无环图和结构因果模型。例如，考虑图 3 中的有向无环图和结构因果模型，其中我们可以看到由共同父节点导致的伪相关。结构因果模型表示为 $Z = f(U_z)$以及$y = f(Z, U_y)$。PCM 的初始目标是计算外生变量 U_z 和U_y的值。这些变量指的是对于某个个体而言影响Z 和y的所有未测量变量。这些U 变量也称为噪声变量，用于描述某个个体或观测的世界。我们首先需要计算这些变量，然后才能扰动输入特征以计算反事实。这些变量之所以需要首先计算，是因为它们描述了某个个体必须在事实世界和反事实世界中保持不变的“情境”。请回想，反事实背后的思想是保持其他所有条件不变，只改变一个特征。这里的其他条件指的是外生变量。

为了在本例中计算 U_z 和U_y ，我们从感兴趣的数据中选取一个个体（例如患者、学生等），并将观测到的事实特征输入结构因果模型。例如，假设我们通过观察测得X、Z 和y 。我们将该个体的Z 输入第一个结构因果模型$Z = f(U_z)$，并计算出U_z。要计算U_y，则将该个体测得的y 和Z 输入第二个结构因果模型$y = f(Z, U_y)$。这样，我们就得到了所选个体的外生变量。这一步称为溯因，计算依据是实际测量数据。

5.2 行动：对反事实变量进行干预

在第二步（称为行动）中，我们输入新的反事实特征，例如$Z=z$。这意味着Z 不再是U_z 的函数，因此我们删除了从U_z 指向Z 的箭头，并移除了Z 与其父节点U_z之间的关系。参见图6。应用反事实时，我们会移除特征与其父节点之间的关系，无论父节点是内生特征还是外生特征。这也称为对Z应用珀尔do算子。这是计算反事实的关键步骤。现在，我们得到了一个经过修改（或手术式改动）的模型，并用它来计算反事实结果。

5.3 预测

在最后一步，即预测步骤中，我们使用新的 $Z = z$特征值，以及步骤 1 中计算出的 U 变量，并计算新的 $y = f(Z=z, U_y)$。结果部分将通过更复杂数据的详细示例，介绍如何执行此操作。

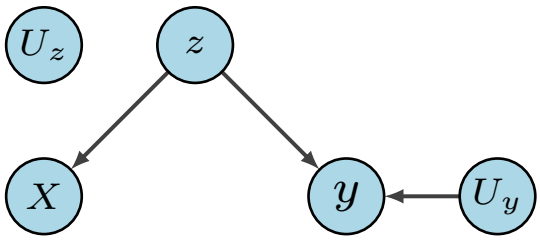


图6：显示共同父节点混杂的有向无环图，但现在删除了 U_z 与 z 之间的边，表明一个独立于其父节点的反事实量。

6 方法论

本研究用于生成反事实解释的反事实解释方法是 DiCE ML [6]。选择该方法是因为它可作为 Python 软件包直接使用，且易于实现。以下介绍使用结构因果模型评估反事实解释的工作流程。另请参见图7。

1. 根据上述讨论的三种类型之一选择因果结构，并基于结构因果模型生成数据集。2. 使用这些数据训练监督学习模型。该模型用于 DiCE 反事实解释生成过程。3. 在数据集中确定一个案例/单元。4. 使用 DiCE 生成反事实解释，DiCE 会利用前一步训练好的模型。这些反事实解释会使输出切换为相反的类别。5. 现在，单独按照第 5.1 节所述的 Pearl 方法进行溯因，以计算所有外生特征。6. 按照 Pearl的行动方法（第 5.2 节），将 DiCE 反事实解释输入结构因果模型。7. 使用 Pearl的预测方法估计输出类别（第 5.3 节）。

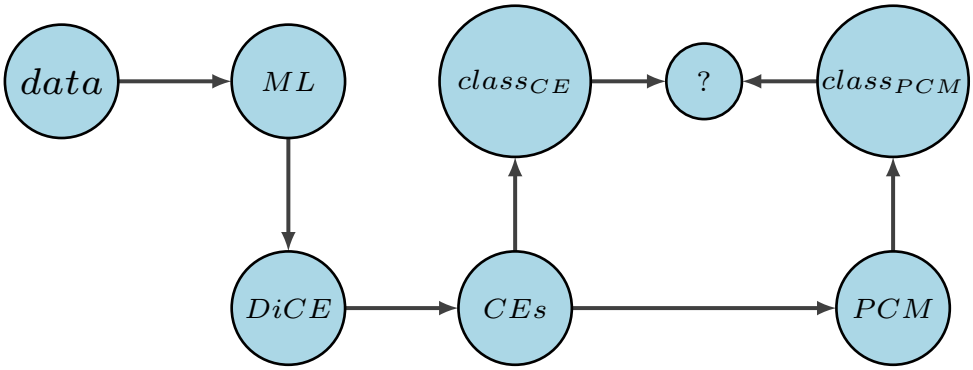


图7：方法论流程：生成 DiCE 反事实解释，将反事实解释输入 PCM，并比较输出类别。

6.1 实验

本节详细介绍基于上述三种因果结构开展的实验，但为了更贴近现实生活，所用数据集的复杂度有所提高。表1所示的数据集包含七个特征，其设计大致参考了描述修读某门课程的大学生的特征。

表1：本研究所用特征的描述。

特征	描述	统计
x1	成绩	μ=50, σ=5
x2	Age	μ=20, σ=1
x3	成绩	μ=45, σ=6
x4	性别	p = 0.6
x5	助学金	p = 0.3
x6	成绩	μ=70, σ=5
x7	成绩	μ=50, σ=5

6.2 实验 1：链式因果结构

实验 1 的有向无环图和结构因果模型基于链式因果结构（第4.1节），如图 8 和公式 12 至 19 所示。请注意，在结构因果模型中，y 不是 x_7 的函数，因此不会在预测中使用它；然而

机器学习模型会在训练和预测中使用 x_7。在有向无环图中，为简洁和节省空间，X 指代所有未由 x_7 和 x_3表示的特征，即 x_1、x_2、x_4、x_5、x_6。这些特征直接指向 y，因此被视为 y 的父节点。此外，U 指代所有指向X的外生变量，即 U_1、U_2、U_4、U_5、U_6。在结构因果模型中，β 的系数从 0 到 8 任意选定如下：0.4、0.6、0.4、0.6、0.7、0.4、0.4、0.3、0.7。

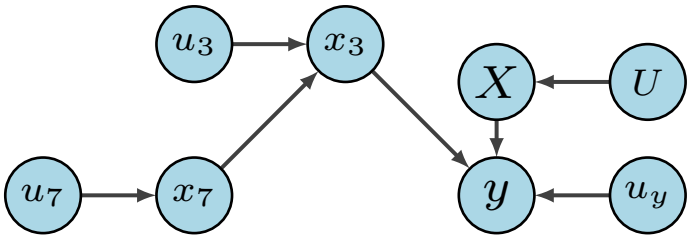


图 8：基于链（中介变量）因果结构的实验 1 因果结构

e.

$$x_1 = u_1 \tag{12}$$
$$x_2 = u_2 \tag{13}$$
$$x_3 = \beta_1 x_7 + u_3 \tag{14}$$
$$x_4 = u_4 \tag{15}$$
$$x_5 = u_5 \tag{16}$$
$$x_6 = u_6 \tag{17}$$
$$x_7 = u_7 \tag{18}$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + U_y \tag{19}$$

6.2.1 实验 2：分叉（共同原因）因果结构

第二个实验基于第 4.2 节中的分叉结构。有向无环图和结构因果模型呈现在图 9 以及方程 20 和 27 中。实验 1 中，因果路径从 x_7 流向 x_3，再流向 y；而在实验 2 中，x_3 同时是 x_7 和 y 的原因。除此之外，其他内容均与实验 1 相同。

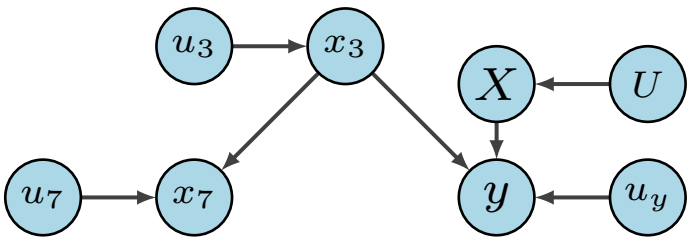


图 9：基于分叉（共同原因/父节点）的实验 2 因果结构。

$$x_1 = u_1 \tag{20}$$
$$x_2 = u_2 \tag{21}$$
$$x_3 = u_3 \tag{22}$$
$$x_4 = u_4 \tag{23}$$
$$x_5 = u_5 \tag{24}$$
$$x_6 = u_6 \tag{25}$$
$$x_7 = \beta_8 x_3 + u_7 \tag{26}$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta 4 x_4 + \beta 5 x_5 + \beta_6 x_6 + u_y \tag{27}$$

6.3 实验 3：碰撞因果结构

基于图 10 和方程 28 至 35 所示碰撞因果结构的实验 3 的有向无环图和结构因果模型。再次注意，x_7 与 y 之间不存在直接的因果路径，并且 x_3 不是 y 的父节点。因此，结构因果模型不将 x_7 或 x_3 视为 y 的父节点，也不会在计算 y 时将其纳入其中（见方程 35）。然而，在机器学习模型中，x_7 和 x_3 很可能会被纳入其中。这再次凸显了机器学习建模与结构因果模型建模之间的区别。

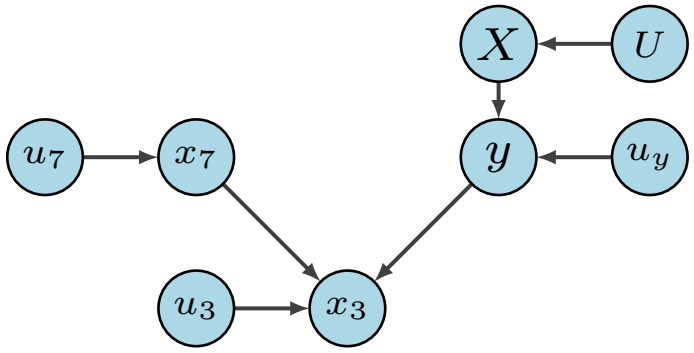


图 10：实验 3 的因果结构基于碰撞因果结构。

$$x_1 = u_1 \tag{28}$$
$$x_2 = u_2 \tag{29}$$
$$x_3 = \beta_1 x_7 + u_3 \tag{30}$$
$$x_4 = u_4 \tag{31}$$
$$x_5 = u_5 \tag{32}$$
$$x_6 = u_6 \tag{33}$$
$$x_7 = u_7 \tag{34}$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + u_y \tag{35}$$

6.4 局限性与假设

在本研究的所有反事实解释计算和评估中，我都假设所有特征均具有可操作性，且所有反事实值均合理。本研究的目的并非探讨可操作性和合理性，而是研究因果性。然而，在现实生活中，所有特征都具有可操作性且所有反事实都合理的情况极不可能发生。作此假设是为了简化研究。

7 结果

本节介绍两个主要结果。首先，展示使用 DiCE 反事实解释的 PCM**方法**（第 7.1 节）。其次，展示该方法在来自三种因果结构的 30 个示例上的**结果**（第 7.2 节）。

7.1 Pearl 反事实方法（PCM）

7.1.1 实验1：链式因果结构

按照第 6 节概述的步骤，我们首先：

步骤1： 通过对表1中的每个特征进行采样来生成数据集，并根据公式12至19中的结构因果模型生成数据。图11的前五行展示了生成数据的一个样本。请注意，y 是数值，类别为二元值 0 或 1。DiCE 只能处理类别值的更改，因此我将输出 y 转换为类别：将大于或等于均值的值设为 1，将低于均值的值设为 0。此方法在之后的所有结果中均保持一致。

	x1	x2	x3	x4	x5	x6	x7	y	class
0	47.362877	21.002521	17.223274	1.0	0.0	71.937835	44.392533	76.494094	0.0
1	46.651160	19.471523	21.829535	0.0	0.0	74.260529	52.435793	79.836456	1.0
2	52.134719	19.182215	21.486748	0.0	1.0	71.614470	55.552280	80.557391	1.0
3	49.471940	21.605364	22.491716	0.0	0.0	72.944226	56.659063	81.226121	1.0
4	47.020510	22.175850	16.103378	1.0	0.0	66.486686	41.943270	73.765417	0.0

图 11：步骤 1 中生成的数据的前五行。

步骤 2： 接下来，在 DiCE 操作所用的数据上训练一个逻辑回归模型。由于输出是一个类别，因此使用逻辑回归。任何能够执行分类的监督学习模型都可以用于此处。

步骤 3： 选择数据集中任意一个你感兴趣的个体，并据此计算反事实。该个体及 DiCE 计算出的两个反事实解释如图 12 所示。请注意，原始类别为 1，而反事实将类别切换为 0。

Query instance (original outcome : 1)

	x1	x2	x3	x4	x5	x6	x7	class
0	54.404495	20.252169	18.953856	0.0	0.0	77.000465	51.353214	1

Diverse Counterfactual set (new outcome: 0.0)

	x1	x2	x3	x4	x5	x6	x7	class
0	54.414494	20.252169	18.953855	0.0	0.0	69.196125	53.364685	0
1	39.938810	20.252169	18.953855	1.0	0.0	77.000468	51.363214	0

图 12：为实验 1 随机选取的个体，显示在顶部。在此次 DiCE 计算中，生成了两个 DiCE 反事实解释，显示在底部的两行中。

步骤 4 和 5： 接下来进行溯因，即使用图 12 所示该个体的实际特征值计算外生变量。使用图 12 顶行中的值，我重新列出公式 28 至 35，但这次代入了具体数值，以计算U 值，结果如表 2 所示。

$$54.5 = \boldsymbol{u_1} \tag{36}$$
$$20.2 = \boldsymbol{u_2} \tag{37}$$
$$18.9 = 0.6 \cdot 53.3 + \boldsymbol{u_3} \tag{38}$$
$$0.0 \ = \boldsymbol{u_4} \tag{39}$$
$$0.0 \ = \boldsymbol{u_5} \tag{40}$$
$$69.2 = \boldsymbol{u_6} \tag{41}$$
$$53.3 = \boldsymbol{u_7} \tag{42}$$
$$81.7 = 0.4 + 0.6 \cdot 54.5 + 0.4 \cdot 20.2 + 0.6 \cdot 18.9 + 0.7 \cdot 0 + 0.4 \cdot 0 + 0.4 \cdot 77.0 + \boldsymbol{u_y} \tag{43}$$

表 2：实验 1 的外生变量。

u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_y
54.4	20.2	-1.58	0	0	77.0	51.4	-1.58

步骤 6： 在此示例中，我选择反事实 0：即图 12 中反事实的第一行。在这里可以看到，DiCE 更改了两个特征 $x_6 = 69.2$ 和 $x_7 = 53.4$，以获得不同的输出类别 0。其余特征保持不变。因此，应用 Pearl 的方法时，如果某个特征发生变化，我们就删除其父节点，并输入新的反事实值。这是因为，对某个特征进行干预时，我们会引入一种新的机制来确定这些特征的状态，而该机制不再“听从”其父节点 [2]。

现在，我们将上述内容（即外生变量和反事实解释）应用于下方所示的本实验结构因果模型。最终，我们试图计算结构因果模型的输出 y，并将其与 DiCE 反事实解释的输出 y 进行比较。

$$x_1 = u_1 = 54.4 \tag{44}$$
$$x_2 = u_2 = 20.2 \tag{45}$$
$$x_3 = 0.6 \cdot 53.3 + (-1.58) \tag{46}$$
$$x_4 = u_4 = 0.0 \tag{47}$$
$$x_5 = u_5 = 0.0 \tag{48}$$
$$x_6 = CE_6 = 69.2 \tag{49}$$
$$x_7 = CE_7 = 53.3 \tag{50}$$
$$y = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \beta_3 x_3 + \beta_4 u_4 + \beta_5 u_5 + \beta_6 CE_6 + u_y = 79.08 \tag{51}$$

输出 y = 79.08 i 大于均值 79.07。因此，将 DiCE 计算出的反事实解释应用于 PCM 时，我们得到类别 1，而 DiCE 对该个体预测的类别为 0。

7.1.2 实验 2 - 分叉因果结构

按照上述相同步骤，对于分叉因果结构，我们生成数据集（步骤 1）、训练分类模型（步骤 2）、选择一个个体/案例并生成 DiCE 反事实解释（步骤 3 和 4，见图 13），计算外生变量（步骤 4 和 5），确定反事实值并应用行动（步骤 6），最后计算结构因果模型的输出并与 DiCE 的输出进行比较（步骤 7）。

图 13 展示了所选个体以及两个 DiCE 反事实解释。

Query instance (original outcome : 1)

	x1	x2	x3	x4	x5	x6	x7	class
0	53.248295	19.537193	51.998009	0.0	0.0	76.005814	21.638533	1

Diverse Counterfactual set (new outcome: 0.0)

	x1	x2	x3	x4	x5	x6	x7	class
0	34.350138	19.537193	51.998009	1.0	0.0	76.015812	21.638533	0
1	53.258294	19.537193	24.591432	0.0	0.0	76.015812	15.370225	0

图 13：为分叉因果结构选定的个体及生成的两个 DiCE 反事实解释。

表3展示了计算得到的外生变量。

表3：实验 2 的外生变量。

u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_y
53.2	19.5	51.9	0	0	76.0	0.84	0.84

如果我们再次在示例中使用 CE_0，从图 13 可以看出，特征 $x_1 = 34.4$ 和 $x_4 = 1.0$ 与原始特征不同：它们是反事实。因此，我们会特别关注它们，删除其原始因果机制，并用新值替换；在结构因果模型中，其余特征保持不变。

$$y = \beta_0 + \beta_1 \cdot CE_1 + \beta_2 u_2 + \beta_3 u_3 + \beta 4 \cdot CE_4 + \beta_5 u_5 + \beta_6 u_6 + u_y = 91.9 \tag{52}$$

输出 y = 91.6 小于均值 93.6。因此，将 DiCE 计算出的反事实解释应用于 PCM 时，我们得到类别 0，这与 DiCE 预测的类别相同。

7.1.3 实验 3 - 碰撞因果结构

最后一个示例针对碰撞因果结构，我们再次按照之前相同的六个步骤进行。这里不再重复细节，只展示结果。图 14 显示了选定的个体和 DiCE 生成的反事实，表 4 展示了计算得出的U 变量。

Query instance (original outcome : 0)

	x1	x2	x3	x4	x5	x6	x7	class
0	44.859966	20.429888	58.347687	0.0	1.0	76.730919	44.246429	0

Diverse Counterfactual set (new outcome: 1.0)

	x1	x2	x3	x4	x5	x6	x7	class
0	44.859967	20.429888	61.960862	0.0	1.0	76.730918	44.256429	1
1	44.859967	20.429888	58.357685	0.0	1.0	79.906119	44.256429	1

图 14: 为实验 3 选定的个体，以及针对碰撞因果结构生成的两个 DiCE 反事实解释 .

表 4：实验 3 的外生变量。

u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_y
44.8	20.4	0.43	0	1	76.7	44.3	0.43

如果我们再次在图 14 的这个示例中使用 CE_0，就会发现只有特征 $x_3 = 61.9$ 与原始值不同。然而，由于碰撞结构的因果结构，x_3 不是 y 的父节点，因此会被忽略。因此，输出 y 的计算如下：

$$y = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \beta_4 u_4 + \beta_5 u_5 + \beta_6 u_6 + u_y = 67.0 \tag{53}$$

输出 y = 67.0 大于均值 66.84。因此，将 DiCE 计算出的反事实解释应用于 PCM 时，我们得到类别 1，这与 DiCE 预测的类别相同。

7.2 因果结构的影响

第 7.1 节介绍了通过 DiCE 计算反事实解释的逐步方法，然后将这些值用于 PCM，以计算反事实输出。此处旨在介绍使用 PCM 的方法。本节仍采用相同的方法，旨在了解使用 PCM 时，不同的因果结构是否会产生不同的反事实结果。为此，我们将该方法应用于 30 个 DiCE 反事实解释，每种因果结构各 10 个。

7.2.1 链（中介变量）

表5展示了选取5个个体的结果。DiCE为每个个体生成两个反事实解释，共计十个反事实解释。每个反事实解释都输入链式结构因果模型，并按照上述步骤执行了6个步骤。“类别”列指原始个体和反事实解释所属的类别。“类别（PCM）”列表示使用Pearl反事实方法预测的类别。可以看到，在十次比较中，有三次PCM预测的类别与DiCE方法不同。表中以灰色标出这些情况。

表5：链式因果结构的结果

Item	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y	类别	类别（PCM）	y（PCM）
单元***1***	54.4	20.2	18.9	0	0	77.0	51.3	81.7	1	-	
U	54.4	20.2	-1.58	0	0	77.0	51.4	-1.58	-	-	
CE_0	54.4	20.2	18.9	0	0	**69.2**	**53.4**	-	0	1	79.1
CE_1	**39.9**	20.2	18.9	**1**	0	77.0	51.3	-	0	0	77.8
第***2***单元	52.5	18.7	15.9	0	0	80.5	40.1	80.9	1	-	
U	52.5	18.7	-0.16	0	0	80.5	40.1	-0.16	-	-	
CE_0	52.5	18.7	15.9	0	0	**73.1**	40.1	-	0	0	77.5
CE_1	**41.4**	18.7	15.9	0	0	**70.9**	40.1	-	0	0	70.4
第***3***单元	57.5	21.3	23.2	0	1	61.9	58.6	82.3	1	-	
U	57.5	21.3	-0.2	0	1	61.9	58.6	-0.2	-	-	
CE_0	57.5	21.3	**19.0**	0	1	61.9	58.6	-	0	1	79.8
CE_1	57.5	21.3	**15.1**	0	**0**	61.9	58.6	-	0	0	77.1
第***4***单元	54.6	20.2	19.9	0	0	66.2	54.1	77.9	0	-	
U	54.6	20.2	-1.7	0	0	66.2	54.1	-1.7	-	-	
CE_0	**62.7**	20.2	19.9	0	0	66.2	54.1	-	1	1	82.8
CE_1	**62.7**	**18.5**	19.9	0	0	66.2	54.1	-	1	1	82.1
第***5***单元	50.4	20.4	25.6	0	1	67.5	58.8	83.7	1	-	
U	50.4	20.4	2.1	0	1	67.5	58.8	2.1	-	-	
CE_0	50.4	20.4	**21.7**	0	1	67.5	58.8	-	0	1	80.9
CE_1	**35.3**	20.4	25.6	0	1	67.5	58.8	-	0	0	74.2

7.2.2 分叉因果结构

表6展示了使用分叉因果结构时十个反事实解释的结果。灰色行标出了PCM方法生成的类别与DiCE相反的情况。这种情况在十次比较中出现了两次。

表6：分叉因果结构的结果。

Item	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y	类别	类别（PCM）	y（PCM）
单元***1***	47.5	19.53	50.6	0	0	68.5	19.9	94.1	0	-	
U	47.5	19.53	50.6	0	0	68.5	-0.3	-0.3	-	-	
CE_0	47.5	19.53	**51.7**	**1**	0	68.5	19.9	-	1	1	95.5
CE_1	47.5	19.53	**55.8**	0	0	68.5	19.9	-	1	1	97.0
第***2***单元	56.7	18.9	42.4	1	0	63.9	15.9	92.6	1	-	
U	56.7	18.9	42.4	1	0	63.9	-1.11	-1.11	-	-	
CE_0	**51.2**	**20.5**	42.4	1	0	63.9	15.9	-	0	0	89.9
CE_1	**62.4**	18.9	**28.2**	1	0	63.9	15.9	-	0	0	87.5
第***3***单元	52.3	18.7	43.1	1	0	64.4	18.6	92.9	1	-	
U	52.3	18.7	43.1	1	0	64.4	1.36	1.36	-	-	
CE_0	**53.5**	**22.1**	43.1	1	0	64.4	18.6	-	0	1	95.0
CE_1	52.3	**20.3**	43.1	1	0	64.4	18.6	-	0	0	93.6
第***4***单元	46.6	20.2	52.9	1	0	72.2	20.4	96.6	1	-	
U	46.6	20.2	52.9	1	0	72.2	-0.8	-0.8	-	-	
CE_0	**36.4**	**22.3**	52.9	1	0	72.2	20.4	-	0	0	91.7
CE_1	**38.8**	20.2	52.9	1	0	72.2	20.4	-	0	0	92.3
第***5***单元	54.4	21.1	46.7	0	0	74.9	19.9	100.7	1	-	
U	54.4	21.1	46.7	0	0	74.9	19.9	2.1	-	-	
CE_0	**36.4**	21.1	46.7	0	0	**70.1**	19.9	-	0	0	87.9
CE_1	**46.1**	21.1	46.7	0	0	74.9	19.9	-	0	1	95.7

7.2.3 碰撞因果结构

表 7 展示了针对碰撞因果结构生成的十个反事实解释的 PCM 结果。可以看到，其中五个（50%）DiCE 反事实解释产生的输出与 PCM 输出类别相冲突。这些结果以灰色显示。

表 7：碰撞因果结构的结果

Item	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y	类别	类别（PCM）	y（PCM）
单元***1***	47.3	21.1	57.4	1	0	59.7	48.9	62.3	0	-	
U	47.3	21.1	0.48	1	0	59.7	48.9	0.48	-	-	
CE_0	47.3	21.1	57.4	1	0	**85.2**	48.9	-	1	1	72.5
CE_1	47.3	**20.7**	**69.5**	1	0	59.7	48.9	-	1	0	62.1
第***2***单元	46.8	19.7	58.9	0	1	68.9	51.9	64.0	0	-	
U	46.8	19.7	-0.3	0	1	68.9	51.9	-0.3	-	-	
CE_0	46.8	19.7	**69.7**	0	1	68.9	51.9	-	1	0	64.05
CE_1	**61.4**	19.7	58.9	0	1	68.9	51.9	-	1	1	72.7
第***3***单元	43.3	19.4	56.7	0	0	75.3	51.4	62.7	0	-	
U	43.3	19.4	-1.5	0	0	75.3	51.4	-1.5	-	-	
CE_0	**46.9**	19.4	**63.1**	0	0	75.3	51.4	-	1	0	64.8
CE_1	**60.5**	**16.8**	56.7	0	0	75.3	51.4	-	1	1	72.0
第***4***单元	45.5	19.9	63.8	0	1	66.3	59.7	64.1	0	-	
U	45.5	19.9	1.5	0	1	66.3	59.7	1.5	-	-	
CE_0	45.5	19.9	63.8	0	1	66.3	47.17	-	1	1	73.1
CE_1	45.5	19.9	63.8	0	1	66.3	48.6	-	1	1	73.1
第***5***单元	43.6	20.4	63.9	0	1	77.3	55.3	67.4	1	-	
U	43.6	20.4	1.3	0	1	77.3	55.3	1.3	-	-	
CE_0	43.6	20.4	**54.8**	0	1	77.3	55.3	-	0	1	67.4
CE_1	43.6	20.4	**61.5**	0	**0**	77.3	55.3	-	0	1	67.4

8 讨论

本节讨论研究结果，尤其是该方法与当前文献的比较及其契合之处。

8.1 Pearl的反事实方法

Verma等人 [5] 指出，反事实解释文献当前面临的一项挑战是缺乏使用因果模型（结构因果模型）来指导反事实解释。Mahajan等人 [8] 表示，反事实解释中的可行性（即来自一个可能世界）从根本上说是一个因果概念，而且保持因果关系十分重要。也就是说，如果我们改变一个特征，就必须考虑其他特征与该特征之间的因果关系。他们举例说明，反事实解释可能会建议将教育程度提高到硕士学位，以获得不同的期望结果。据他们所述，如果我们改变教育程度，那么由于年龄与教育程度存在因果联系，个人的年龄也需要改变。为保持因果约束，他们提出应基于结构因果模型约束特征间距离。Riccardo Guidotti [9] 也将因果性视为反事实解释的一项要求：要使反事实解释合理且可执行，就应保持特征之间的因果关系。这两篇论文的观点都有道理。我们希望使用数据中的真实因果结构。这甚至是本文的一项主要贡献。然而，在处理反事实时，其中存在一些细微之处。根据 Pearl 的观点，当我们对某个特征进行干预并执行 do演算时，会通过删除该特征与其父节点之间的关系，操纵并改变数据中的原始因果结构。这与 Mahajan等人 [8]和 Guidotti [9]所要求的内容略有不同。确实，我们希望利用数据中的真实因果结构；然而，如上文 Pearl的行动方法所述（见第 5.2 节），反事实特征不再受其父节点约束。因此，如果我们将 Pearl的方法应用于 Mahajan等人的例子：假设在原始因果结构中，年龄是教育程度的父节点，而反事实解释建议提高教育程度，那么由于我们现在正对教育程度进行干预，它便不再受周围因素的影响，而只成为我们干预的函数。年龄不再是教育程度的原因，这两个特征也不再受到因果约束。然而，这与 Mahajan等人略有不同，因为后者旨在保持因果关系。

此外，保留反事实变量的因果结构，难道不只是维持了相同的数据分布吗？请记住，反事实本质上并不属于该分布。当然，我们的特征仍必须具有可行性，也就是说，它们必须来自现实世界。然而，维持反事实与其父节点之间的因果结构，似乎会让问题停留在 Pearl因果层级的第 1 层：关联 [2]。反事实的定义就是保持其他所有因素不变，只改变该特征。如果其他因素也按照原始数据生成过程随之改变，那么这似乎就不是反事实。不过，我承认，如果我误解了上述文献的内容，我愿意接受指正。

读者或许会问，既然我似乎并不在意因果结构中的约束，为何还要如此强调使用结构因果模型？请记住，**只有反事实特征会从其父节点中删除**；其他输入特征仍保留其因果结构。此外，反事实特征与输出之间的因果关系也不受影响。确实，我们必须通过让数据保持在现实世界中来确保可行性，但不能通过保留反事实特征与其父节点之间的因果关系来做到这一点。

8.2 因果结构的影响

第二个目标是利用 PCM 考察不同的因果结构如何影响使用 DiCE 生成的反事实解释。首先，在 30 个不同的结果中，有 10 个（33%）显示反事实解释的输出与 PCM 的输出存在冲突。尽管这并不能得出确切结论，但结果表明 DiCE 生成的反事实解释可能会给出错误估计（假设 PCM 正确）。为获得更具说服力的结果，我建议针对 500 到 1000 个估计值开展模拟研究。当前面临的挑战是，至少目前每次 PCM 计算都需要手动完成，因此要重复数百次非常困难。未来的工作可以编写代码来模拟这类研究。
另一项重要结果是，碰撞因果结构中的反事实与 DiCE 生成的反事实解释有 50% 的冲突。链结构有 10 例中的 3 例冲突，分叉结构有 10 例中的 2 例，而碰撞结构有 10 例中的 5 例冲突。同样，这些结果并非定论，但它们表明，因果结构的类型可能会显著影响所生成的真实反事实解释。机器学习模型只是将所有特征都纳入模型，而真实的结构因果模型不会纳入所有特征。事实上，在统计预测模型中纳入碰撞结构会导致因果估计出现不必要的偏差[7] ，从而可能“破坏因果分析”[12]。这些结果表明，仅仅使用整个数据集训练模型，就可能给反事实估计引入不必要的误差。在盲目生成反事实解释之前，了解潜在的因果结构至关重要。

8.3 因果发现

本研究使用了具有已知真实因果结构的模拟数据。读者可能会质疑，这种方法是否只适用于模拟数据；如果面对真实数据且不知道因果结构，我们该如何使用 PCM 验证反事实解释？这时就需要进行因果发现，也就是从真实数据中发现真实因果结构 [13]。对这种方法的进一步说明超出了本研究的范围。不过，我认为，若要大幅增强我们对 DiCE 等方法生成的反事实解释的信任度，我们必须先对数据进行因果发现，尽可能确定真实因果结构，然后将反事实解释输入真实结构上的 PCM 方法，以此进行验证。真实数据集可能比我在本研究中创建的数据集复杂得多。这部分留待未来研究。

8.4 总结与未来工作

在文章开篇，我指出评估（验证）反事实解释（CE）至关重要。一个主要原因是，反事实解释由机器学习模型生成，而这些模型未必会考虑数据中的因果结构。这可能导致预测和反事实估计出现误差。反事实解释正受到数据领域的广泛关注；如果我们不能深入理解反事实解释的生成方式，那么它们充其量也毫无价值。

本研究结果表明，用于生成反事实解释的机器学习方法与 Pearl 的方法之间存在一些冲突。这让我们不禁怀疑，盲目由机器学习模型生成的反事实解释究竟有多值得信赖。不同的因果结构表现出不同程度的冲突。

接下来该怎么做？在生成反事实解释之前，我们必须尽最大努力了解真实的因果结构，并据此调整模型。因果发现对反事实解释的未来发展至关重要。未来的工作将包括：在具有已知因果结构的真实数据上，使用反事实解释开展 PCM。

我也认为，验证反事实解释的终极方法，是将其应用于真实数据，并在一段时间后测量实际输出。当然，这极具挑战性。

为了更有把握地确认这些结果，最好在数百个反事实解释上进行模拟，并将其应用于真实数据。

参考文献

[1] Judea Pearl、Madelyn Glymour 和 Nicholas P. Jewell。《统计学中的因果推断：入门》。Wiley，2016年。[2] Judea Pearl 和 Dana Mackenzie。《为什么之书》。Basic Books，2018年。[3] Christoph Molnar。《可解释机器学习：让黑箱模型变得可解释的指南》。2019年。[4] Sandra Wachter、Brent Mittelstadt 和 Chris Russel。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》。《arXiv》预印本 arXiv:1711.00399，2017年。[5] Sahil Verma、John P. Dickerson 和 Keegan Hines。机器学习的反事实解释：重新审视挑战。arXiv:2106.07756v1，2021年。[6] Ramaravind K. Mothilal、Amit Sharma 和 Chenhao Tan。通过多样化反事实解释说明机器学习分类器。2020年公平性、问责制与透明度会议论文集，2020年。[7] M. Luque-Fernandez、M.A.and Sc homaker、D. Redondo-Sanchez、M. Jose Sanchez Perez、A. Vaidya 和 M.E. Schnitzer。教育札记：非传染性疾病流行病学数据分析中的悖论性碰撞器效应：可复现的示例与网络应用。《国际流行病学杂志》，48:640–653，2019年。[8] Divyat Mahajan、Chenhao Tan 和 Amit Sharma。在机器学习分类器的反事实解释中保留因果约束。arXiv:1912.03277v3，2020年。[9] Riccardo Guidotti。反事实解释及其寻找方法：文献综述与基准测试。《数据挖掘与知识发现》，2022年。[10] Bernhard Schölkopf。面向机器学习的因果性。arXiv:1911.10500v2，2019年。[11] Peter Tennant。因果推断与有向无环图导论，2022年。[12] Stephen L. Morgan 和 Christopher Winship。反事实与因果推断：社会研究的方法与原理。《剑桥大学出版社》，2007年。[13] Clark Glymour、Kun Zhang 和 Peter Spirtes。基于图模型的因果发现方法综述。《遗传学前沿》10:524.doi: 10.3389/fgene.2019.00524，2019年。