将预测模型与因果效应估计相联系

Jiuyong Li、Lin Liu、ZiqiXu、Ha Xuan Tran、Thuc Duy Le、Jixue Liu

南澳大学，澳大利亚阿德莱德

arXiv:2304.04566v1 [cs.LG] 10 Apr 2023

摘要

预测模型基于给定的特征进行结果预测，即估计给定特征向量时结果的条件概率。一般而言，预测模型无法估计某个特征对结果的因果效应，也就是说，当保持其他特征的值不变而改变该特征时，结果会如何变化。这是因为因果效应估计需要干预概率。然而，个性化决策、推荐和公平性计算等许多现实问题，都需要针对给定实例了解任何特征对结果的因果效应。这不同于处理变量固定的传统因果效应估计问题。本文首先探讨针对给定实例，估计任意特征（作为处理）对结果的因果效应这一挑战。理论结果自然地将预测模型与因果效应估计联系起来，并表明当满足本文确定的条件时，预测模型具有因果可解释性。本文还揭示了具有因果可解释性的模型所具备的稳健性。我们通过实验表明，在满足本文确定的条件时，各类预测模型都能够像最先进的因果效应估计方法一样准确地估计特征的因果效应。我们还展示了这类具有因果可解释性的预测模型在稳健预测和个性化决策方面的潜力。

1 引言

大多数监督学习方法旨在估计条件概率 $P(Y \mid \mathbf{X} = \mathbf{x})$，其中 $\mathbf{X}$ 称为特征（变量）集合，而 Y则表示结果（或目标变量）。人们开发了大量此类预测模型来预测结果，并且这些模型能够在各种数据集中准确地拟合条件概率。

然而，许多决策问题都需要回答关于任意特征与结果的假设性问题。如果 $X_i \in \mathbf{X}$ 从 0 变为 1， Y 会发生什么变化？例如，如果求职者拥有大学学位，他能获得这份工作吗？如果顾客收到优惠券，他会购买该产品吗？回答这类假设性问题，需要估计相应特征（例如大学学位或优惠券）对结果（例如获得工作或购买产品）的因果效应。特征（作为处理）对结果的因果效应，指的是由特征变化所导致的结果变化 [21, 29]。

因果效应估计是因果推断领域的主要研究课题，大多数因果效应估计方法都假设处理变量固定且协变量集合给定 [21, 19, 23]。已有多种条件平均处理效应（CATE）估计方法被提出，用于个性化决策 [5, 43, 36, 25, 51], ，这些方法都针对固定的处理变量和给定的协变量集合。

在许多问题情境中，处理变量并非固定，因此这些因果效应估计方法无法应用于此类问题。例如，对于一个包含人们的生活方式、饮食等（特征）以及他们六十多岁时是否发生心脏病发作（结果）的数据集，可以利用该数据集训练预测模型，根据特征值预测他们六十多岁时发生心脏病发作的风险。此外，人们可能希望了解现在应该改变哪些生活方式或饮食，以降低未来发生心脏病发作的风险。生活方式和饮食中的每个特征都可以作为处理；当它不作为处理时，则可以作为协变量。

为个体确定最佳处理方案，是个性化决策中的关键环节。在寻找个体的最佳处理方案时，会将特征交替视为处理，以评估各自的因果效应。当某个特征被视为处理时，需要确定适用于该处理与结果的协变量集合。因此，现有的因果效应估计方法并不适用，因为它假设处理变量固定，且处理与结果对应的协变量集合已知 [21, 19]。现有的处理效应估计方法需要扩展到预测性《机器学习》场景，在这种场景下，每个特征都可以作为处理，且不会预先指定固定的处理变量和协变量集合。

本文首先探讨将特征交替作为处理时，因果效应估计所面临的挑战。交替使用处理变量时，这并非对现有因果效应估计方法的直接扩展。我们通过以下示例说明这些挑战。考虑一种传统场景，其中数据集包含处理变量 T、协变量 X 和结果 Y。这意味着该数据集背后的因果图为 $T \leftarrow X \rightarrow Y$和 $T \rightarrow Y$。在本文的问题设定中，(T、 X) 都是特征，并可交替作为处理。要估计在给定 $X = x$ 的条件下， T 对 Y 的因果效应，可使用在给定 $X = x$ 的条件下， T 对 Y 的条件平均因果效应（CATE）。然而，当 X 被交替设为处理时， T 并非变量对 (X、 Y) 的协变量，而是 X 与 Y 之间的中介变量。在这种情况下，CATE 估计量不适用于估计在给定 T 的条件下， X 对 Y 的因果效应，因为将 X 设为处理时， T 也会随之改变。相反，当将 T 控制为 $T = t$ 时， X 对 Y 的受控直接效应（ CDE ）才是适当的。

如上例所示，我们面临两项挑战。首先，将交替特征作为处理时，不同特征的因果效应可能不属于同一类型。其次，为某项处理选择正确类型的因果效应需要底层因果图。因此，一般而言，如果我们不知道底层因果图，就无法通过将个体特征交替作为处理来估计其对结果的因果效应。在实践中，因果图很少是已知的。在本文中，我们将对已知因果图的要求降低为：已知数据集包含结果的所有直接原因，且不包含任何受结果影响的变量。允许存在并非直接原因的隐藏变量。这使得我们提出的方法比那些假设已知因果图、用于估计任意特征对结果的因果效应的方法更具实用性。

其次，上述解决方案的理论结果自然地将预测模型与因果效应估计联系起来。也就是说，当满足这些条件时，可以直接从预测模型中估计因果效应。这些结果表明，我们可以拥有一个具有因果可解释性的预测模型，从中可以推导出个体特征的因果效应。因果解释并不取决于预测模型的透明度，而取决于输入模型的内容。预测模型本身可以是黑箱，但每个特征对结果的影响都可以从模型中推导出来，因此该模型具有因果可解释性。

预测模型与因果解释之间的联系有助于模型解释。人们希望对预测模型进行因果解释，但通常这是不可能的。在线性回归模型中，系数常被解释为相应特征对结果的因果效应。然而，只有在已知潜在因果结构，且回归模型恰当地遵循该因果结构时，这种解释才成立。这是因为系数可能代表不同类型的因果效应，甚至可能是有偏估计 [44]。此外，即使因果解释成立，线性回归模型中的系数也表示总体层面的平均处理效应（ATE），不适用于本文讨论的个性化决策。

第三，我们进一步研究上述具有因果可解释性的模型的性质，并表明，当模型的训练环境发生变化时，该模型仍具有稳健性。

总之，本文作出以下贡献。

1. 在给定情境下估计任意特征对结果的因果效应，通常需要知道数据集背后的因果图。本文将要求从了解完整因果图，放宽为只需知道数据集包含结果的所有直接原因，且不包含受结果影响的变量。只要隐藏变量不是结果的直接原因，就允许其存在。放宽后的假设使估计任意特征对结果的因果效应能够采用更实用的方法。

2. 这项工作将预测模型与因果效应估计联系起来。当论文中指出的条件得到满足时，预测模型可用于因果效应估计，因此具有因果可解释性。我们通过实验表明，各种类型的预测模型对因果效应的估计准确度与最先进的因果效应估计方法相当。

3. 我们 a 分析并证明，具有因果可解释性的模型比其他预测模型更稳健。 ls
使用 在一个不同于模型训练环境的新环境中的所有特征。

2 图形因果模型

在本节中，我们介绍因果推断所需的背景知识。我们用大写字母表示变量，用加粗的大写字母表示变量集合。变量的取值用小写字母表示。

设 $\mathcal{G} =$ (**V**，**E**)是一个图，其中**V** $= \{V_1, \ldots, V_p\}$ 是节点集合，**E** 是边集合，表示节点之间的连线
这些节点。一个 pa ath
th π 是由不同节点组成的序列，其中每一对相邻节点在 $\mathcal{G}$ 中都相邻。一个 p

π 是有向路径，当且仅当路径上所有边的箭头都指向同一方向。相对于 V_i，(V_i, V_j) 之间的路径是后门路径，当且仅当它有一条箭头指向 V_i。给定一条路径 π，如果 V_k 上有两条边与其相接且 $V_i \to V_k \leftarrow V_j$，则 V_k 是 π 上的碰撞节点。在 $\mathcal{G}$ 中，如果存在 $V_i \to V_j$，则 V_i 是 V_j 的父节点；我们用 PA(V_j) 表示 V_j 的所有父节点组成的集合。如果存在一条从 V_i 到 V_j 的有向路径，那么 V_i 是 V_j 的祖先，而 V_j 是 V_i 的后代。

有向无环图（有向无环图）是一个不存在有向环的有向图。如果我们将节点的父节点解释为其直接原因，并将有向路径解释为一条因果路径，其中起始节点是终止节点的间接原因，则该有向无环图称为因果有向无环图。

在以下两个假设下，有向无环图与一个分布相联系。

假设1 (马尔可夫条件 [29])。 给定一个有向无环图 $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ 和 $P(\mathbf{V})$，V的联合概率分布$\mathbf{V}$， $\mathcal{G}$ 满足马尔可夫条件：对于 $\forall V_i \in \mathbf{V}$， V_i 与 V_i 的所有非后代节点独立，给定 V_i的父节点。

假设 2 (忠实性 [40])。 有向无环图 $\mathcal{G} = (\mathbf{V}$、$\mathbf{E})$ 当且仅当在 $P(\mathbf{V})$ 中呈现的每个条件独立性都由 $\mathcal{G}$ 蕴含时，有向无环图是忠实的，而 $\mathcal{G}$ 满足马尔可夫条件。

在马尔可夫条件和忠实性假设下，我们可以使用 d-分离 [29] 从有向无环图中读取 $P(\mathbf{V})$ 中变量之间的（依赖或独立）关系。

定义 1 (d-分离)。 如果且仅当满足以下条件，则有向无环图中的路径 π 被节点集 $\mathbf{S}$ d-分离：*(1)* $\mathbf{S}$ 包含路径 π 中链 $V_i \to V_k \to V_j$、 $V_i \leftarrow V_k \leftarrow V_j$ 或 $V_i \leftarrow V_k \to V_j$ 的中间节点 V_k；*(2)* 当路径 π 包含碰撞点 $V_i \to V_k \leftarrow V_j$ 时，节点 V_k 或其任何后代 V_k 均不在 $\mathbf{S}$ 中。

集合 $\mathbf{S}$ 可以是空集。如果路径包含一个碰撞点，则该路径被空集 d分离。如果路径被$\mathbf{S}$ d分离，则 $V_p \perp\!\!\!\perp V_q \mid \mathbf{S}$，其中 V_p 和 V_q 是路径的起点和终点节点。如果该路径未被任何集合d分离（包括空集），则该路径是 d连通的，并且 V_p 和 V_q 相关联。

因果（处理）效应表示因果关系的强度。为了定义因果效应，我们引入干预的概念，干预会迫使某个变量取某个值，通常用*do* 算子 [29] 表示。*do* 操作模拟了现实世界实验中的干预。例如， $do(X = 1)$ 表示对 X 进行干预，使其取值为 1。 $P(y \mid do(X = 1))$ 是干预概率。

平均处理效应量化了总体中一对变量的处理效应 .

定义 2 （平均处理效应（ATE））。 X 对 Y的平均处理效应定义为 $\mathrm{ATE}(X, Y) = P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0))$。

平均处理效应是总体中的平均处理效应。然而，某些子群体的处理效应可能与平均处理效应相差很大。条件处理效应用于捕捉子群体中的处理效应。

定义 3 （条件平均处理效应（CATE））。 设 $\mathbf{W}$ 为一组变量。给定 $\mathbf{W} = \mathbf{w}$ 时， X 对 Y的条件平均处理效应定义为 $\mathrm{CATE}(X, Y; \mathbf{W} = \mathbf{w}) = P(Y = 1 \mid do(X = 1), \mathbf{W} = \mathbf{w}) - P(Y = 1 \mid do(X = 0), \mathbf{W} = \mathbf{w})$。

作为合格的条件，$\mathbf{W}$ 通常不包含 X 或 Y 的任何后代节点。

当 X 和 Y 之间存在多条有向路径（包括路径 $X \to Y$）时，除 X 和 Y 之外，有向路径上的变量都是中介变量。受控直接效应衡量的是当中介变量 $\mathbf{Z}$ 保持在 $\mathbf{z}$ 时，经由边 $X \to Y$ 产生的直接效应。

定义 *4*（受控直接效应（控制直接效应） [30]）。将 X 与 Y 之间的中介变量设置为常数值 z（记作 Z），即 $Z = z$ 时的受控直接效应定义为：$CDE(X, Y; Z = z) = P(Y \mid do(X = 1)\ do(Z = z)) - P(Y \mid do(X = 0)\ do(Z = z))$。

要通过将干预概率（*do* 运算符）简化为正常条件概率来推断这些概率，需要因果有向无环图。可以使用后门准则、前门准则或 do 演算 [29]进行简化。

3 假设、可识别性与算法

3.1 将特征交替作为处理时的控制直接效应、条件平均处理效应与可识别性

考虑一个典型的机器学习问题设定，其中 $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}$ 是一组特征， Y 是结果。为便于说明，我们假设 $\mathbf{X}$ 和 Y 均为二元变量。不过，所有结论



(a) (b)

图1：两个示例性因果有向无环图。

本节中的结论同时适用于二元变量和连续变量，因为我们的结论表明，只要满足假设，无论变量类型如何，干预概率和观察概率都是可互换的。例如，当 Y 是连续变量时，在二元公式中，概率 $P(y \mid \mathbf{X} = \mathbf{x})$ 由 $E(y \mid \mathbf{X} = \mathbf{x})$ 取代。我们用 y 表示 $Y = 1$，用 x_i 表示 $X_i = x_i$；在上下文明确时，用 $do(x_i)$ 表示 $do(X_i = x_i)$。

令 $\mathbf{x} = (x_1, x_2, \ldots\ldots, x_m)$ 表示一个个体。要做出个性化决策，需要估计每个 $X_i \in \mathbf{X}$ 对 Y 的因果效应，同时对所有 $X_j \neq X_i$ 保持 $\mathbf{x}$ 中的其他变量值不变。这意味着每个 X_i 都交替作为处理。此设定不同于仅考虑固定处理的正常因果效应估计问题。将特征交替作为处理会给因果效应估计带来挑战。下面给出一个例子。

示例1（当交替特征作为处理时测量因果效应的挑战）。考虑如图*1(a)*所示的数据集背后的因果有向无环图。X_1 和 X_2 是 Y 的两个父节点。对于个体 $\mathbf{x} = (x_1, x_2)$，当 $X_2 = x_2$ 保持不变时，我们旨在估计 X_1 对 Y 的因果效应（X_1 是处理）；当 $X_1 = x_1$ 保持不变时，估计 X_2 对 Y 的因果效应（X_2 是处理）。

首先，让我们将 X_1 视为处理。$\mathrm{CATE}(X_1, Y;\ X_2 = x_2) = P(y \mid do(X_1 = 1)\ x_2) - P(y \mid do(X_1 = 0)\ x_2))$ 描述了在以 $X_2 = x_2$ 为条件时，X_1 对 Y 的条件平均处理效应。*CATE* 被提出用于描述因果异质性（总体中不同子群体的平均因果效应）*[1, 5, 43, 24]*。*CATE* 主要用于个性化决策 *[50, 49]* 和政策评估 *[4]*。一些研究人员使用它来近似个体处理效应 *[36, 26]*。

现在，我们将 X_2 视为处理。由于 X_1 不是 (X_2, Y) 的混杂因素，因此不能再使用 *CATE*。X_1 是 (X_2, Y) 之间的中介变量。$CDE(X_2, Y;\ X_1 = x_1) = P(y \mid do(X_2 = 1)\ do(x_1)) - P(y \mid do(X_2 = 0)\ do(x_1))$ 表示当将 X_1 设为固定值 x_1 时，X_2 对 Y 的受控直接效应。*CDE* 被提出用于因果中介分析，表示在中介变量保持不变时，处理对结果的直接效应 *[34, 28, 41, 42]*。*CDE* 已用于政策评估 *[34, 28]* 和政治解释 *[2]*。

两种不同类型的因果效应描述了我们希望捕捉的效应。其使用取决于底层因果有向无环图。当我们不知道底层因果有向无环图时，我们也不知道该使用哪一种。

两类因果效应之间的差异在于条件概率的条件集合（处理之外）：CATE使用 do 自由（观察性）条件集合，而CDE使用带有 do（干预性）的条件集合。一般而言，干预概率不同于观察概率，因此 CATE 和 CDE 也不同。

Fortuna 地说，在某些情况下，控制直接效应和条件平均处理效应具有相同的估计值。需要以下假设 n.

假设 3 (完整直接原因)**.** 结果的所有直接原因 Y 均包含在 $\mathbf{X}$ 中。

我们将在下面的观察中展示两种类型的因果效应之间的一致性。以下讨论中，我们不假定已知因果图。

观察 1（数据中 CDE 和 CATE 的估计值相同）。假设一个数据集包含变量 (X_1, X_2, Y)，其中 X_1 和 X_2 是 Y 的直接原因。在数据中，我们观察到 $X_1 \not\perp\!\!\!\perp Y$、$X_2 \not\perp\!\!\!\perp Y$ 和 $X_1 \not\perp\!\!\!\perp X_2$，其中 $\not\perp\!\!\!\perp$ 表示不独立（或相关）。根据数据，我们可以推断图*1(a)*或图*1(b)*这两个有向无环图之一代表生成该数据集的底层因果有向无环图。根据数据无法推断 X_1 和 X_2 之间的边的方向。我们的目标是在 X_2 保持为 x_2 时，估计 X_1 对 Y 的因果效应。

如果图*1(a)*是底层因果有向无环图，我们估计 条件平均处理效应$(X_1, Y;\ X_2 = x_2)$。如果图*1(b)*是底层因果有向无环图，我们估计 控制直接效应$(X_1, Y;\ X_2 = x_2)$。当 X_1 和 X_2 是 Y 的直接原因时，有向无环图*(a)*中的 条件平均处理效应$(X_1, Y;\ X_2 = x_2)$ 与有向无环图*(b)*中的 控制直接效应$(X_1, Y;\ X_2 = x_2)$ 取值相同。两个估计值分别为 $P(y \mid X_1 = 1, x_2) - P(y \mid X_1 = 0, x_2)$。

证明。令 $\mathcal{G}$ 为一个有向无环图，令 $\mathcal{G}_{X_1}$ 为从 $\mathcal{G}$ 中移除 X_1 的出边后得到的有向无环图，并令 $\mathcal{G}_{X_1 X_2}$ 为从 $\mathcal{G}$ 中移除指向 X_1 的入边和 X_2 的出边后得到的有向无环图。在以下推导中，假设 3 成立。

假设有向无环图 (a) 是底层的因果有向无环图 $\mathcal{G}$。我们进行如下推导。规则参见定理 3.4.1 [29]。

$$\begin{aligned} & P(y \mid do(x_1), x_2) \\ & = P(y \mid x_1, x_2) \quad \because Y \perp\!\!\!\perp X_1 \mid X_2 \; in \; \mathcal{G}_{\underline{X_1}} \; (Rule\ 2) \end{aligned}$$

假设 DAG (b) 是基础因果有向无环图 $\mathcal{G}$。我们得到如下推论。

$$\begin{aligned} & P(y \mid do(x_1), do(x_2)) \\ & = P(y \mid do(x_1), x_2) \quad \because Y \perp\!\!\!\perp X_2 \mid X_1 \; in \; \mathcal{G}_{\overline{X_1}\underline{X_2}} \; (Rule\ 2) \\ & = P(y \mid x_1, x_2) \quad \because Y \perp\!\!\!\perp X_1 \mid X_2 \; in \; \mathcal{G}_{\underline{X_1}} \; (Rule\ 2) \end{aligned}$$

因此，CATE(X_1, Y; $X_2 = x_2$) = $P(y \mid X_1 = 1,\ x_2) - P(y \mid X_1 = 0,\ x_2)$，且 CDE($X_1$, Y; $X_2 = x_2$) = $P(y \mid X_1 = 1,\ x_2) - P(y \mid X_1 = 0,\ x_2)$。二者具有相同的值。 □

上述观察表明，当对所有直接原因进行条件化时，干预（带有 do）与观察（do-free）是可交换的。下面的引理对此进行了说明。令 PA(Y) 表示基础因果有向无环图中 Y 的所有父节点（直接原因）。

引理 1。 *[干预与观察的可交换性]* 在假设 *1*、*2* 和 *3* 下，令 $\mathrm{PA}(Y) = \{X_1, X_2, \ldots, X_k, \ldots, X_p\}$ 其中变量的索引任意（假设 X_1 为处理），且 k 是 $[2, p]$ 中的一个值。$P(y \mid do(x_1)\ do(x_2)\ \ldots do(x_k)\ x_{k+1}, \ldots\ldots, x_p) = P(y \mid x_1,\ x_2,\ \ldots x_k,\ x_{k+1}, \ldots\ldots,\ x_p)$。

证明。 我们首先说明如何将 $do(x_k)$ 化简为 x_k。设 $\mathcal{G}$ 为数据背后的因果有向无环图，$\mathcal{G}_{X_1,\ X_2,\ \ldots\ldots,\ \overline{X_k = x_k},\ X_k}$ 为从 $\mathcal{G}$ 中移除所有指向 $X_1,\ X_2, \ldots\ldots,\ X_{k-1}$ 的入边以及从 X_k 出发的出边后得到的有向无环图。规则参见定理 3.4.1 [29]。

$$\begin{aligned} & P(y \mid do(x_1), do(x_2), \ldots do(x_k), x_{k+1}, \ldots, x_p) \\ & = P(y \mid do(x_1), do(x_2), \ldots do(x_{k-1}), x_k, x_{k+1}, \ldots, x_p) \\ & \because Y \perp\!\!\!\perp X_k \mid x_1, x_2, \ldots x_{k-1}, x_{k+1}, \ldots, x_p \\ & in \; \mathcal{G}_{\overline{X_1}, \overline{X_2}, \ldots, \overline{X_{k-1}}, \underline{X_k}} \; (Rule\ 2) \end{aligned}$$

分别对 $X_{k-1}, \ldots\ldots, X_2, X_1$ 重复上述推导，我们得到：$P(y \mid do(x_1)\ do(x_2) \ldots\ldots,\ do(x_{k-1}), x_k, \ldots\ldots, x_p) = P(y \mid x_1,\ x_2,\ \ldots x_{k-1},\ x_k, \ldots\ldots,\ x_p)$。

引理得证。 □

在因果效应估计的条件集中使用 do 或不含 do，取决于底层因果有向无环图，具体而言，取决于 Y 的父节点之间的路径。回顾我们在示例1中的讨论。当 X_1 是处理时，使用 $P(y \mid do(x_1)\ do(x_2))$，因为 X_2 是 X_1 的后代。

引理 1 确定了一种情形：由于 do anddo-free 是可交换的，父节点之间的路径不会影响条件平均处理效应和控制直接效应的估计。引理 1 表明，在 PA(Y) 内进行条件平均处理效应和控制直接效应估计是一致的。换言之，即使不知道 Y 的直接原因之间的路径，只要已知并使用所有直接原因，基于条件平均处理效应和控制直接效应估计的因果效应就是相同的。

在上述讨论中，$\mathbf{X}$ 仅包含 Y 的父节点。现在考虑一般情形，即 $\mathbf{X}$ 包含所有变量。条件平均处理效应的条件集合不允许包含 Y 的后代节点，因此条件平均处理效应不适用于一般情形。在定义 4 中，控制直接效应的控制变量被限制为中介变量，因为其动机是进行中介分析。如果将控制变量放宽为系统中除处理之外的所有变量，则可以在假设实验设置 [29] 下解释控制直接效应。当所有其他变量均被设为 $\mathbf{X} \backslash X_i = \mathbf{x}'$，且仅将 X_i 从 0 改变为 1 时，X_i 对 Y 的控制直接效应可通过 Y 的变化来测量，即 CDE(X_i, Y;$\mathbf{X} \backslash X_i = \mathbf{x}'$) = $P(y \mid do(X_i = 1)\ do(\mathbf{x}')) - P(y \mid do(X_i = 0)\ do(\mathbf{x}'))$。

关于控制直接效应的可识别性，我们有以下引理。

引理 *2*（控制直接效应的可识别性）。令 $\{X\ Y\}$ 包含系统中的所有变量，且$X' = X \backslash X_i$。x' 是 X' 的一个取值。在假设 *1*、*2*和 *3* 下。对于每个 $X_i \in PA(Y)$ 以及 $Z = PA(Y) \backslash X_i$，控制直接效应$(X_i, Y; X' = x') = P(y \mid X_i = 1, Z = z) - P(y \mid X_i = 0, Z = z)$。对于每个 $X_i \notin PA(Y)$，控制直接效应$(X_i, Y; X' = x') = 0$。

证明。 首先，令 $\mathbf{X} = \{\mathbf{X_A}, \mathbf{X_D}, \mathbf{X_I}\}$ 其中 $\mathbf{X_A}$ 包含 Y 的所有祖先节点，$\mathbf{X_D}$ 包含 Y 的所有后代节点，而 $\mathbf{X_I}$ 包含所有其他变量（即变量 $I \in \mathbf{X_I}$ 与 Y 之间不存在路径连接，I 和 Y 共享同一个后代节点，或者 I 和 Y 共享同一个祖先节点。）。

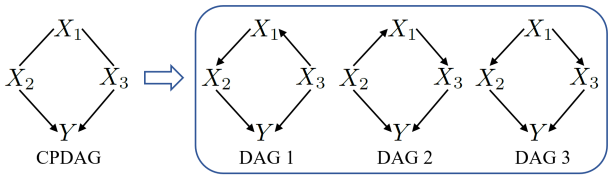

图 图 2：CPDAG（左）表示一组有向无环图（右）的等价类。只有 CPDAG 可以从数据中识别。 ta.

令 $\mathcal{G}$ 为数据背后的因果有向无环图，$\mathcal{G}_{X_i}$ 为从 $\mathcal{G}$ 中移除所有指向 X_i的入边后得到的有向无环图，$\mathcal{G}_{X_i,\mathbf{W}}$ 为从 $\mathcal{G}$ 中移除所有指向 X_i 以及指向$\mathbf{W}$中每个变量的入边后得到的有向无环图。规则指的是定理 3.4.1 [29]中的规则。

令 $\mathbf{X}'_{\mathbf{A}} = \mathbf{X}_{\mathbf{A}} \backslash X_i$ 和 $\mathbf{x}'_{\mathbf{A}}$ 为 $\mathbf{X}'_{\mathbf{A}}$ 的一个取值。$\mathbf{x}_{\mathbf{D}}$ 和 $\mathbf{x}_{\mathbf{I}}$ 分别为 $\mathbf{X}_{\mathbf{D}}$ 和 $\mathbf{X}_{\mathbf{I}}$ 的取值。 $\{\mathbf{x}'_{\mathbf{A}},\mathbf{x}_{\mathbf{D}},\mathbf{x}_{\mathbf{I}}\} = \mathbf{x}'$，其中$\mathbf{x}'$ 是 $\mathbf{X}' = \mathbf{X} \backslash X_i$ 的一个取值。

$$
\begin{aligned}
&P(y \mid do(x_i), do(\mathbf{x}')) \\
&= P(y \mid do(x_i), do(\mathbf{x}'_{\mathbf{A}}), do(\mathbf{x}_{\mathbf{D}}), do(\mathbf{x}_{\mathbf{I}})) \\
&= P(y \mid do(x_i), do(\mathbf{x}'_{\mathbf{A}}), do(\mathbf{x}_{\mathbf{D}})) \\
&\because Y \perp\!\!\!\perp \mathbf{X}_{\mathbf{I}} \mid X_i, \mathbf{X}'_{\mathbf{A}}, \mathbf{X}_{\mathbf{D}} \; in \; \mathcal{G}_{\overline{X_i},\overline{\mathbf{X}'_{\mathbf{A}}},\overline{\mathbf{X}_{\mathbf{D}}}} \; (Rule\ 3) \\
&= P(y \mid do(x_i), do(\mathbf{x}'_{\mathbf{A}})) \\
&\because Y \perp\!\!\!\perp \mathbf{X}_{\mathbf{D}} \mid X_i, \mathbf{X}'_{\mathbf{A}} \; in \; \mathcal{G}_{\overline{X_i},\overline{\mathbf{X}'_{\mathbf{A}}}} \; (Rule\ 3)
\end{aligned}
$$

基于马尔可夫假设 1，对于任意 $X_j \notin \text{PA}(Y)$ 和 $X_j \in \mathbf{X}'_{\mathbf{A}}$, $X_j \perp\!\!\!\perp Y \mid \text{PA}(Y)$。因此，对于$X_i \in \text{PA}(Y)$， $P(y \mid do(x_i)\ do(\mathbf{X}'_{\mathbf{A}} = \mathbf{x}'_{\mathbf{A}})) = P(y \mid do(x_i)\ do(\mathbf{Z} = \mathbf{z})$。

根据引理 1， $P(y \mid do(x_i)\ do(\mathbf{Z} = \mathbf{z}) = P(y \mid x_i,\ \mathbf{Z} = \mathbf{z})$。

因此，控制直接效应$(X_i,\ Y; \mathbf{X}' = \mathbf{x}') = P(y \mid X_i = 1,\ \mathbf{Z} = \mathbf{z}) - P(y \mid X_i = 0,\ \mathbf{Z} = \mathbf{z})$。

为了证明对每个 $X_i \notin \text{PA}(Y)$，控制直接效应$(X_i,\ Y; \mathbf{X}' = \mathbf{x}') = 0$。我们考虑三种情况： $X_j \in \mathbf{X}_{\mathbf{I}}$、$X_j \in \mathbf{X}_{\mathbf{A}}$ 和 $X_j \in \mathbf{X}_{\mathbf{D}}$。在这三种情况下，我们都有以下推导。

$$
\begin{aligned}
&P(y \mid do(x_i), do(\mathbf{X}' = \mathbf{x}')) = P(y \mid do(\mathbf{X}' = \mathbf{x}')) \\
&\because Y \perp\!\!\!\perp X_i \mid \mathbf{X}' \; in \; \mathcal{G}_{\overline{X_i},\overline{\mathbf{X}'}} \; (Rule\ 3)
\end{aligned}
$$

因此，控制直接效应$(X_i,\ Y; \mathbf{X}' = \mathbf{x}') = 0$

引理得证。 □

引理 2 表明，如果已知 Y 的直接原因（或父节点）集合，则可以根据数据估计 Y 上特征的控制直接效应。只有 Y 的父节点对 Y 具有非零控制直接效应，因此我们将重点关注 PA(Y) 的条件直接效应。

在某些数据集中，尤其是在具有高维性的数据集中，找出直接原因可能并不容易。人们提出了许多方法，用于从预测模型和数据中学习概念 [22, 18, 46]。原始特征可以概括为用户容易理解的概念。借助这些概念，相较于使用原始特征，更容易识别直接原因。

3.2 识别数据中的直接原因

当未知 Y 的父节点时，就需要从数据中找出它们。要从数据中找出 Y 的父节点，需要以下假设。

假设 4（不存在 Y 的后代节点）。 $\mathbf{X}$ 不包含 Y 的后代节点。或者，$\mathbf{X}$ 不包含受结果 Y 影响的变量。

从数据中只能学习到一个完全部分定向无环图（CPDAG） [38, 39] 。CPDAG 是一种表示有向无环图等价类的图，这些有向无环图编码了相同的条件独立关系。图 2 展示了 CPDAG 的一个示例。一般而言，无法从数据中学习到唯一的有向无环图，因为仅凭数据无法确定所有边的方向。需要注意的是，在某些其他假设下，例如线性函数和非高斯噪声 [37]，或加性非线性高斯噪声 [20], ，可以识别因果方向。本文不作此类假设。

在实践中，我们不假设存在已知的潜在因果有向无环图，也不假设可以从数据中学习到唯一的有向无环图。然而，给定这些假设，可以在数据中识别直接原因。

算法 1 构建用于结果预测和 CDE 估计的模型（MODE） 输入：由底层因果有向无环图生成的数据集，其中变量 (X Y) 满足假设 3 和 4。用于回答假设性问题的实例 x。参数 k 和 δ。输出：x 的预测结果，以及 CDE 最高的前 k 个特征。 1: 在 X 中发现 PA(Y)。 2: 将数据集投影到 (PA(Y) Y)。 3: 在投影后的数据集中，使用一种机器学习方法构建模型 f(PA(Y))。 4: $y_{\mathbf{x}} = f(\mathbf{x}')$ // 对 x 进行结果预测。其中 x′ 包含 x 在 PA(Y) 中的值 5: 对每个 $X_i \in$ PA(Y) 执行 6: 令 Z $=$ PA(Y)$\backslash X_i$，且 (x_i, z) 包含 x 在 PA(Y) 中的值。{情况 1：二元处理和二元结果}7: 令 $P_f(y \mid x_i, \mathbf{z})$ 表示 f(PA(Y)) 将实例 (x_i, z) 分类为 y（感兴趣的类别）的概率。 8: CDE$_{x_i}$(X_i, Y; z) $= P_f(y \mid x_i, \mathbf{z}) - P_f(y \mid x_i, \mathbf{z})$ {情况 2：连续处理和二元结果}9: CDE$_{x_i}$(X_i, Y; z) $= P_f(y \mid x_i + \delta, \mathbf{z}) - P_f(y \mid x_i, \mathbf{z})$ {情况 3：二元处理和连续结果}10: 令 $f(x_i, \mathbf{z})$ 表示 f (PA(Y)) 在输入 (x_i, z) 时的输出。 11: CDE$_{x_i}$(X_i, Y; z) $= f(x_i, \mathbf{z}) - f(x_i, \mathbf{z})$ {情况 4：连续处理和连续结果}12: CDE$_{x_i}$(X_i, Y; z) $= f(x_i + \delta, \mathbf{z}) - f(x_i, \mathbf{z})$ 13: 循环结束 14: 对所有 $X_i \in$ PA(Y) 的 CDE$_{x_i}$(X_i, Y; z) 排序 15: 返回 $y_{\mathbf{x}}$ 以及 Y 中 CDE 最高的前 k 个父节点。

–

引理 3（数据中 Y 的直接原因的可识别性）。 在假设 *1*、*2*、*3*、*4* 成立且条件独立性检验不存在统计误差的情况下，可以在数据中识别 Y 的直接原因（即 $\mathrm{PA}(Y)$）。

证明。 根据假设 1 和 2，不存在其他变量 d 将一个父节点与 Y 分离。若条件独立性检验不存在统计误差，则父节点（例如 X_i）在给定 $\mathbf{X}\backslash X_i$ 的任意子集时，均与Y 相关。根据假设 4，除父节点外，不存在具有这种性质的其他变量，因为根据马尔可夫条件，给定父节点集合后，该变量将与 Y 独立。因此，可以通过条件独立性检验识别父节点集合。

□

3.3 将预测模型与因果效应估计联系起来

已知 Y 的父节点集合时，可以通过条件概率估计条件直接效应。预测模型能够很好地建模条件概率，因此可用于估计条件直接效应。为了使用预测模型估计条件直接效应，我们作出以下假设。

假设 5（一致模型）。 模型 $Y = f$（PA(Y)）是在总体的一个无偏样本中训练得到的。P（$Y \mid \mathrm{PA}(Y)$）从模型中获得的条件概率与总体中 Y的条件概率相同。

根据假设 5，从预测模型中读取的条件概率就是总体中的条件概率。

现在，我们可以构建一个预测模型，用于估计模型中每个特征对结果 Y的控制直接效应，并根据输入实例预测结果。所提出的算法 MODE 如算法 1 所示。该模型会对输入实例的结果进行预测，并提供一个潜在处理变量列表 k，这些变量是该个体可以改变的，以产生最高的结果效应。

MODE 的第 1 行和第 2 行发现 Y的父节点，并将数据集投影到 Y和 Y的父节点上。可以使用 Hiton PC [3], 、MMPC [3] 和 PC-Select [13] 等局部因果结构学习算法，在数据中学习 PC(Y)，即 Y的所有父节点和子节点的集合。随后，根据假设 3 和 4，PA(Y) $=$ PC(Y)。

第 3 行和第 4 行使用包含 Y和 Y的父节点在内的特征构建预测模型 f(PA(Y))，并预测输入实例 x的结果。任何监督式《机器学习》模型都可用于构建该模型。该模型在结果预测中的用法与正常预测模型相同。

第 5–11行使用引理2，根据特征和结果的不同变量类型，估计 PA(Y)中每个特征对 Y的控制直接效应。控制直接效应估计中使用的条件概率由预测模型f(PA(Y)) 获得。

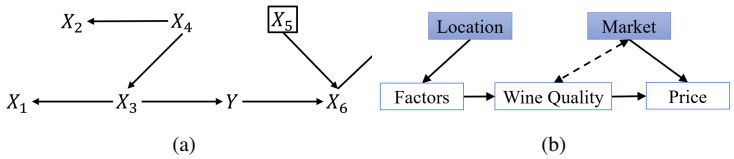

图 3: （a）示例展示了因果图中与 Y 相关的不同类型特征。方框中的 X_5 与 Y 不相关。然而，在以 X_6 为条件时，X_5 会与 Y 相关。这种关联称为虚假关联。（b）示例展示了如何在预测模型中利用虚假关联。渐变填充的方框表示未观测变量。

估计的控制直接效应可用于回答一个假设性问题：如果将 $X_i = x_i$ 更改为 x_i，即 x_i 的相反值（对于连续处理，则为 $X_i = x_i + \delta$），在其他 $\mathbf{x}$ 值保持不变的情况下，Y 会发生什么变化？

在因果推断的惯例中，$X_i = 0$ 表示未接受处理（对照），$X_i = 1$ 表示接受处理（处理组）。在假设性问题中，对照是给定值 x_i，处理组是更改后的值（例如，改为 x_i 的相反值）。因此，在该算法中，CDE包含一个下标，用于表示对照 x_{i0}。

第 14 行 并按其条件直接效应对 PA(Y) 中的特征进行排序，然后输出 $\mathbf{x}$ a 的预测结果

nd the top k fea

具有最高条件直接效应的特征。

当这些假设得到满足时，训练好的预测模型可以在无需访问数据的情况下用于控制直接效应估计（数据仅在训练阶段需要），其方式类似于对输入实例进行结果预测。

3.4 因果可解释模型的鲁棒性分析

前两个小节表明，基于所有直接原因（父节点）构建的预测模型可用于估计任一直接原因对 Y 的 CDE。因此，基于所有直接原因构建的模型具有因果可解释性。一些读者可能会认为，它的性能不如基于与 Y 相关的所有特征（包括 Y 的后代节点）构建的预测模型。我们将从预测模型的另一个方面——鲁棒性——来分析这一点。

基于PA(Y)构建的模型将比基于与 Y 相关的所有特征构建的模型更具鲁棒性。模型的鲁棒性是指模型在新环境中进行可靠预测的能力。

在特征集合$\mathbf{X}$ 和结果 Y的有向无环图中，三类特征将与 Y相关：1） Y的祖先节点，包括 Y的父节点；2） Y的后代节点；以及3） Y的祖先节点的后代节点。我们在图3(a)中展示了这三类变量。 X_3 和 X_4 属于类型1； X_6 和 X_7 属于类型2； X_1 和 X_2 属于类型3。当使用PA(Y)构建预测模型时，除PA(Y)之外的 Y 的其他祖先节点，以及 Y 的所有祖先节点的后代节点（类型2），都无助于构建该模型，因为根据马尔可夫条件（假设1），在给定PA(Y)的条件下，它们与 Y 相互独立。因此，接下来我们将讨论是否应使用 Y 的后代节点来构建预测模型。

使用 Y（类型 3）的后代节点构建预测模型时，与 Y 无关的其他变量也会被纳入模型，这可能导致模型缺乏鲁棒性。

我们首先说明，在模型中纳入 Y 的后代变量如何产生虚假关联。来看图 3(a) 中的有向无环图，其中 X_6 是 Y 的子节点。 X_5 是 X_6 的父节点。边 $Y \to X_6$ 和 $X_5 \to X_6$ 在 X_6 处形成碰撞点，因此在以 X_6 为条件时，Y 和 X_5 相关。这种关联是虚假的，因为 X_5 和 Y 实际上相互独立。因此，当使用 (X_3, X_6) 构建模型时，X_5 会在无意中被纳入模型。

我们通过一个示例来展示虚假关联对模型鲁棒性的影响。图 3(b) 给出了预测模型中包含虚假关联的示例。Y 是葡萄酒质量。因素是 Y 的父节点，Price 是 Y 的子节点，地点是 Y 的祖先节点（不是父节点），而 Market 是 Price 的父节点。地点和 Market 均未观测到。Price 与葡萄酒质量高度相关，因此看起来有助于预测葡萄酒质量。然而，Market 也会影响 Price。当使用 Price 预测葡萄酒质量时，Market 会因虚假关联而在无意中被引入模型。对于品质相同的葡萄酒，在两个市场中的价格可能相差很大。因此，在一个将 Price 作为特征的市场数据集上构建的模型，可能无法适用于另一个不同市场的数据集，因而不具备鲁棒性。

在模型构建过程中，无法使用正常的交叉验证来检验模型的鲁棒性。例如，如果用于构建模型的数据是在一个市场中收集的，那么训练数据集和测试数据集都来自同一个市场。这样一来，模型在训练数据集和测试数据集上的表现就会一致。然而，当模型用于不同的市场时，就会失效。

相反　因此，使用所有直接原因构建的模型不会受到未观测变量的影响，因此具有稳健性。　　ce

地点是因素的一个未观测原因。地点通过改变因素来影响葡萄酒质量，而地点所导致的变化会反映在因素中。在对因素进行条件化后，地点与葡萄酒质量相互独立，因此不会对预测 Y 起到作用。因此，使用因素构建的模型在不同地点都能给出一致的预测。

总之，我们得出以下结论。如果 $U \perp\!\!\!\perp Y$但 $U \not\perp\!\!\!\perp Y \mid \mathbf{X}'$，其中 $\mathbf{X}' \subseteq \mathbf{X}$，那么我们称未观测变量 U 参与模型$Y = f(\mathbf{X})$。如果某些 $U \in \mathbf{U}$ 参与模型，那么未知变量集 $\mathbf{U}$参与模型 $Y = f(\mathbf{X})$。

引理*4*（未观测变量在预测模型中的参与）。设特征集 X 包含与 Y相关的 $\{X_{\mathbf{A}}$、$X_{\mathbf{D}}$、$X_{\mathbf{I}}\}$。在数据集的底层因果图中，$(X_{\mathbf{A}}$、$X_{\mathbf{D}}$、$X_{\mathbf{I}})$ 分别由 Y 的祖先节点、 Y 的后代节点，以及 Y 的祖先节点的后代节点组成。 $\{X_{\mathbf{A}}$、$X_{\mathbf{D}}$、$X_{\mathbf{I}}\}$ 分别有未观测父节点 $\{U_{\mathbf{A}}$、$U_{\mathbf{D}}$、$U_{\mathbf{I}}\}$。设 $X_{\mathbf{A}}$ 包含 Y 的所有直接父节点。当 $PA(Y)$ 用于预测模型时，$X_{\mathbf{I}}$ 和 $(X_{\mathbf{A}} \setminus PA(Y))$ （ '\' 表示集合差）会被忽略。$U_{\mathbf{D}}$ 参与模型 $Y = f(PA(Y), X_{\mathbf{D}})$，但 $U_{\mathbf{A}}$ 、 $U_{\mathbf{D}}$ 或 $U_{\mathbf{I}}$ 不参与模型 $Y = f(PA(Y))$。

证明。根据马尔可夫条件（假设1），$\mathbf{X_I} \perp\!\!\!\perp Y|\text{PA}(Y)$，且（$\mathbf{X_A} \setminus \text{PA}(Y)$） $\perp\!\!\!\perp Y \mid \text{PA}(Y)$。因此，在构建包含 PA($Y$) 的预测模型时，会忽略 $\mathbf{X_I}$ 和（$\mathbf{X_A} \setminus \text{PA}(Y)$。

$f(\mathbf{X_A}, \mathbf{X_D})$ 使用条件概率 $P(y \mid (\mathbf{X_A}, \mathbf{X_D}))$ 进行预测。在数据中，$\mathbf{U_D} \perp\!\!\!\perp Y$。从 Y 到 $\mathbf{X_D}$ 的有向路径，以及从 $\mathbf{U_D}$ 到 $\mathbf{X_D}$ 的有向路径，在节点 $\mathbf{X_D}$ 处形成碰撞点。因此，$\mathbf{U_D} \not\perp\!\!\!\perp Y \mid \mathbf{X_D}$。因此，$\mathbf{U_D}$ 参与模型 $f(\mathbf{X_A}, \mathbf{X_D})$。

$f(\text{PA}(Y))$ 使用条件概率 $P(y \mid (\text{PA}(Y))$ 进行预测。根据马尔可夫条件（假设1），$\mathbf{U_A} \perp\!\!\!\perp Y \mid \text{PA}(Y)$ 且 $\mathbf{U_I} \perp\!\!\!\perp Y \mid \text{PA}(Y)$。因此，$\mathbf{U_A}$ 和 $\mathbf{U_I}$ 不参与模型 $f(\text{PA}(Y))$。此外，$\mathbf{U_D} \perp\!\!\!\perp Y \mid \text{PA}(Y)$，因此 $\mathbf{U_D}$ 不参与模型 $f(\mathbf{X_A})$。

引理得证。 □

引理4表明，仅使用 Y 的父节点构建预测模型，不会涉及系统中的其他未观测变量，因此模型能够抵御未观测变量的变化。相比之下，在模型中使用结果的后代变量，可能会使模型涉及未观测变量（虚假关联），并可能导致预测模型在新环境中的表现不佳。

在因果特征选择文献中， [47, 48], 马尔可夫毯被认为是分类的最优特征集。结果 Y 的马尔可夫毯包括 Y 的父节点（直接原因）、 Y 的子节点（直接受结果影响的变量），以及 Y 的每个子节点的父节点。研究表明，基于马尔可夫毯构建的预测模型能够抵御环境变化 [45]。当马尔可夫毯中的所有变量都可观测时，基于马尔可夫毯构建的模型既准确又稳健。我们考虑一种一般情形： Y 的某些子节点的部分父节点可能不可观测（例如图 3(b) 中的 Market 变量）。在这种情况下，使用受结果影响的变量会导致模型缺乏鲁棒性。需要注意的是，未观测变量在实践中十分常见，因此在构建模型时不使用受结果影响的变量，是一个合理的建议。

4 实验

实验将证明论文中的以下主要结论：1）当满足相关条件时，任何预测模型都能够估计任意特征对 Y 的控制直接效应（CDE）；2）基于所有直接原因（ Y 的父节点）构建的具有因果可解释性的预测模型，将比基于所有特征变量构建的预测模型更加稳健；3）具有因果可解释性的预测模型可能有助于个性化决策。

4.1 因果效应估计能力

由于大多数数据集中不存在真实因果效应，因此评估因果效应估计具有挑战性。我们使用两个合成数据集进行评估。以下将介绍数据生成的详细信息。

$\mathcal{G}_1$：$U \to X_1 \to Y, X \to X_2 \to Y, U \to X_3 \to Y, U \to X_4 \to Y, U \to X_5$和 $\mathcal{G}_2$：$U \to X_1 \to Y, X \to X_2 \to Y, U \to X_3 \to Y, U \to X_4 \to Y, U \to X_5, X_1 \to X_2 \to X_3, X_4 \to X_3$其中 U是一个未观测变量。

1	MLP	GBR	LR	DT	RF	SGD	线性深度度量学习	核深度度量学习	CF	线性深度强化学习	特征深度强化学习
2k	30.5 ± 10.7	34.2 ± 11.6	30.2 ± 12.1	35.5 ± 13.8	33.9 ± 11.7	32.2 ± 11.1	26.2 ± 8.1	38.3 ± 10.0	30.9 ± 12.4	30.0 ± 10.7	31.5 ± 11.2
	11.1 ± 9.9										
4k	32.3 ± 11.0	28.5 ± 7.6	28.6 ± 7.9	29.9 ± 7.3	30.0 ± 8.5	31.6 ± 9.7	30.1 ± 8.2	34.7 ± 6.5	32.3 ± 7.8	27.3 ± 9.1	29.9 ± 7.6
	9.6 ± 6.5	7.1 ± 5.0	7.4 ± 5.1	8.0 ± 6.2	8.8 ± 7.1	10.2 ± 6.5		8.1 ± 6.3	9.1 ± 5.3	10.2 ± 6.5	7.0 ± 4.5
6k	29.4 ± 6.2	27.6 ± 5.2	30.6 ± 8.3	29.4 ± 7.2	30.4 ± 8.0	29.4 ± 8.7	28.0 ± 7.7	33.3 ± 5.5	27.9 ± 6.7	28.1 ± 6.1	30.2 ± 9.2
	7.2 ± 6.0	5.5 ± 3.8	8.2 ± 6.6	7.0 ± 6.0	8.8 ± 6.5	6.8 ± 6.0	6.8 ± 5.9	7.3 ± 7.0	6.9 ± 5.0	6.7 ± 5.3	7.5 ± 6.1
8k	29.0 ± 5.9	29.9 ± 6.2	29.9 ± 6.9	30.2 ± 6.0	29.4 ± 4.6	31.1 ± 7.5	29.7 ± 5.8	32.8 ± 5.5	29.3 ± 5.7	31.7 ± 4.4	31.0 ± 5.3
	6.8 ± 5.8	5.7 ± 4.4	7.0 ± 4.7	5.5 ± 4.2	5.2 ± 3.3	8.8 ± 7.7	6.8 ± 4.7	6.4 ± 5.5	5.7 ± 3.9	4.7 ± 4.4	5.8 ± 5.2
10k	29.2 ± 5.4	30.6 ± 4.9	29.8 ± 7.1	29.6 ± 5.8	29.7 ± 5.1	30.5 ± 6.4	28.7 ± 6.4	31.5 ± 4.6	29.0 ± 6.6	30.0 ± 4.6	28.7 ± 5.6
	5.1 ± 4.4	4.4 ± 3.6	7.3 ± 4.8	4.7 ± 3.8	5.8 ± 4.1	5.6 ± 5.4	5.7 ± 3.9	5.7 ± 4.1	6.4 ± 4.9	4.1 ± 3.8	4.6 ± 3.9
20k	30.6 ± 4.0	28.7 ± 4.1	29.8 ± 4.2	28.7 ± 3.5	29.2 ± 3.1	28.7 ± 6.7	29.9 ± 3.9	31.5 ± 3.4	29.9 ± 3.3	30.0 ± 3.5	29.2 ± 3.6
	5.4 ± 3.4	4.0 ± 2.5	3.8 ± 3.5	3.1 ± 3.4	3.4 ± 2.3	5.4 ± 3.4	3.8 ± 2.5	4.4 ± 3.6	3.2 ± 2.6	4.0 ± 2.8	3.5 ± 2.5

2	MLP	GBR	LR	DT	RF	SGD	线性深度度量学习	核深度度量学习	CF	线性深度强化学习	特征深度强化学习
2k	27.9 ± 11.4	29.1 ± 12.7	27.9 ± 10.0	29.1 ± 11.0	28.0 ± 9.8	35.7 ± 13.0	29.8 ± 11.1	39.7 ± 10.4	27.2 ± 12.3	28.8 ± 11.7	32.4 ± 11.2
	14.1 ± 9.2			12 ± 8.2	11.7 ± 6.9						
4k	30.2 ± 10.7	29.2 ± 8.6	30.8 ± 8.9	30.2 ± 8.6	29.3 ± 8	32 ± 8.9	26.9 ± 8.0	35.9 ± 8.9	29.6 ± 8.9	27.6 ± 10.7	29.6 ± 9.6
	8.7 ± 7.6	8.9 ± 6.3	8.9 ± 6.5	8.5 ± 6.2	8.4 ± 5.5	9.8 ± 7.1	9.6 ± 7.5	12.8 ± 8.8	9.2 ± 5.1	10.9 ± 9.4	6.5 ± 5.1
6k	30.9 ± 6.8	28.8 ± 6.5	29.7 ± 5.4	27.7 ± 5.3	30.8 ± 6.1	30.0 ± 8.9	31.3 ± 5.9	36.2 ± 6.7	27.5 ± 7.2	26.2 ± 6.7	27.1 ± 8.1
	6.5 ± 6.9	5.8 ± 5.4	4.4 ± 4.5	5.4 ± 4.7	6.5 ± 5.3	7.6 ± 6.6	6.5 ± 4.8	10.3 ± 7.5	7.7 ± 7.6	7.4 ± 6.1	6.7 ± 5.2
8k	31.4 ± 7.7	29.2 ± 6.6	31.0 ± 6.5	30.0 ± 6.4	28.8 ± 6.1	32.8 ± 7.2	29.3 ± 6.6	32.4 ± 5.7	28.5 ± 6.6	28.9 ± 6.3	27.9 ± 5.8
	8.8 ± 5.6	6.4 ± 4.6	7.1 ± 5.9	6.6 ± 4.8	6.5 ± 4.5	7.6 ± 6.0	5.6 ± 4.0	7.4 ± 4.1	4.8 ± 3.6	6.3 ± 5.1	5.2 ± 3.5
10k	30.4 ± 6.2	30.9 ± 5.8	28.6 ± 5.4	28.0 ± 5.4	28.8 ± 4.6	28.1 ± 6.1	28.8 ± 5.9	32.3 ± 4.8	30.7 ± 4.5	29.0 ± 4.6	29.2 ± 5.7
	6.0 ± 4.7	5.4 ± 3.6	5.2 ± 4.2	4.6 ± 3.9	4.9 ± 3.6	6.0 ± 3.9	6.3 ± 4.8	5.7 ± 5.1	4.5 ± 3.0	5.6 ± 3.5	5.9 ± 4.5
20k	29.3 ± 4.5	29.2 ± 3.8	29.5 ± 3.9	30.0 ± 3.4	30.4 ± 2.8	29.8 ± 6.0	28.9 ± 3.9	33.4 ± 3.0	29.8 ± 3.1	30.1 ± 3.2	29.4 ± 4.1
	4.9 ± 3.6	3.5 ± 2.6	3.7 ± 2.9	3.6 ± 2.7	3.4 ± 2.5	5.7 ± 4.9	4.4 ± 2.9	5.3 ± 3.6	3.1 ± 2.2	3.2 ± 2.4	3.8 ± 3.3

表 1：两个表格中，两个合成数据集上的 CDE 和 CATE 估计偏差。含有／不含背景的行分别报告以父节点／所有变量作为输入时的结果。对所有方法而言，使用父节点比使用所有变量产生的偏差更低。左侧含背景的结果是 MODE 使用多种机器学习方法进行的 CDE 估计。右侧结果是使用基准方法进行的 CATE 估计。考虑置信区间时，MODE 与基准方法的偏差一致。

来自 $\mathcal{G}_1$ 的数据集的数据生成过程描述如下：

$$
\begin{aligned}
U \sim Normal(0,1);\ \ e \sim Normal(0,1);\\
X_1 \sim Binomial(1, \frac{1}{1+exp(-U)});\\
X_2 \sim Binomial(1, \frac{1}{1+exp(-U)});\\
X_3 \sim Binomial(1, \frac{1}{1+exp(-U)});\\
X_4 \sim Binomial(1, \frac{1}{1+exp(-U)});\\
X_5 \sim Binomial(1, \frac{1}{1+exp(-U)});\\
Y \sim 1 + 1.5X_1 + 1.5X_2 + 1.5X_3 + 1.5X_4 + e
\end{aligned}
$$

来自 $\mathcal{G}_2$ 的数据集的数据生成过程描述如下：

$$
\begin{aligned}
U \sim Normal(0,1);\ \ e \sim Normal(0,1);\\
X_1 \sim Binomial(1, \frac{1}{1+exp(-U)});\\
X_2 \sim Binomial(1, \frac{1}{1+exp(-U \times X_1)});\\
X_3 \sim Binomial(1, \frac{1}{1+exp(-U \times X_2 \times X_4)});\\
X_4 \sim Binomial(1, \frac{1}{1+exp(-U)});\\
X_5 \sim Binomial(1, \frac{1}{1+exp(-U)});\\
Y \sim 1 + 1.5X_1 + 1.5X_2 + 1.5X_3 + 1.5X_4 + e
\end{aligned}
$$

真实因果效应是使用*do* 演算规则 [29]在数据中计算得出的。

我们使用多种广为人知的预测《机器学习》模型实例化MODE：线性回归 (LR) [17],决策树 (DT) [32], 随机森林 (RF) [12], 随机梯度下降 (SGD) [11], 多层感知机 (MLP) [35] 和回归梯度提升(GBR) [15]。它们在 scikit-learn [31] 中实现。

我们将 MODE 与一些知名的CATE估计方法进行比较：线性双重机器学习（L-DML） [14], 核双重机器学习（K-DML） [27], 因果森林双重机器学习（CF） [43],线性双重稳健学习（L-DRL） [14], 以及森林双重稳健学习（F-DRL） [6]。它们在EconML[8] 中实现。当假设得到满足时，根据我们的结果，CDE 和 CATE 的结果是一致的，因此这种比较是恰当的。由于这些方法需要预先指定处理，我们在所有比较中都使用 X_1 作为处理。

贝叶斯模型的能力。估计的因果效应的标准误列于表1。每个估计值是30个数据集中的平均值。W e具有以下 以下结论。

首先，在满足条件的情况下，预测模型可以像其他专门设计的因果效应估计方法一样准确地估计因果效应，而其中大多数方法都是近年来设计的。在考虑置信区间时，两组方法的偏差是一致的。

其次，仅将 Y 的父节点作为输入变量，能显著改善所有方法的因果效应估计。

第三，所有方法都受益于数据集规模的扩大 ze.

结果表明，结合预测方法的 MODE 能够估计因果效应。

4.2 使用父节点的 ML 模型的鲁棒性

我们使用合成数据集测试将父节点作为特征的模型的鲁棒性。数据集基于图 3(a) 中的因果有向无环图生成，并进行了以下扩展：Factors 包含 3 个变量，且这些变量共享未观测变量 U_1。生成细节如下。

$$
\begin{aligned}
U_1 \sim Normal(0,1);\ \ U_2 \sim Binomial(0,0.5);\\
e_1 \sim Normal(0,1);\ \ e_2 \sim Normal(0,1);\\
X_1 \sim Binomial(1, \frac{1}{1+exp(-U_1)});\\
X_2 \sim Binomial(1, \frac{1}{1+exp(-U_1)});\\
X_3 \sim Binomial(1, \frac{1}{1+exp(-U_1)});\\
Y \sim 1 + 10X_1 + 10X_2 + 10X_3 + e_1;\\
P \sim 1 + 0.8 * Y + 2 * U_2 + e_2
\end{aligned}
$$

环境由 $U_2 = 0$ 和 $U_2 = 1$表示。“无环境变化”意味着训练和测试数据集均从一个环境的数据集中采样，而“有环境变化”意味着训练和测试数据集分别从两个环境的数据集中采样。

模型	相同环境		变化的环境	
	$0 \to 0$	$1 \to 1$	$0 \to 1$	$1 \to 0$
MLP	0.614 ± 0.024	0.614 ± 0.019	1.684 ± 0.123	1.551 ± 0.141
	0.999 ± 0.038	1		
GBR	0.816 ± 0.045	0.810 ± 0.042	3.817 ± 0.491	3.859 ± 0.591
	1.123 ± 0.053	1		
LR	0.612 ± 0.022	0.612 ± 0.027	1.566 ± 0.056	1.550 ± 0.078
	0.992 ± 0.037	0		
DT	0.885 ± 0.042	0.877 ± 0.057	4.478 ± 0.663	4.100 ± 0.443
	1.000 ± 0.040	1		
RF	0.756 ± 0.049	0.740 ± 0.037	3.957 ± 0.463	3.900 ± 0.614
	1.014 ± 0.033	1		
SGD	0.700 ± 0.062	0.701 ± 0.037	2.417 ± 0.526	2.531 ± 0.567
	1.003 ± 0.035	1		

表2：在环境发生变化（右侧）时预测模型的鲁棒性，以环境未发生变化的情况（左侧）为基准。含背景/不含背景的行分别表示以父节点/所有变量作为特征的模型的误差。使用父节点的模型的误差差异（右侧与左侧相比）显著小于使用所有变量的模型。

六个预测模型的误差见表2。在环境不变时，将所有变量用作特征所构建的模型，其误差低于将父节点用作特征所构建的模型，因为子节点 X_2 有助于预测 Y。然而，当环境发生变化时，使用所有变量的模型性能显著下降，而使用父节点的模型性能则保持在同一水平。

结果表明，仅使用父节点将产生不受环境（未观测变量）变化影响的鲁棒性模型。

然后，我们使用三个用于肺癌研究的真实世界数据集，测试将父节点作为特征的模型的鲁棒性。这些数据集来自哈佛 [10], 密歇根 [9], 和斯坦福 [16]的实验室。数据是在三种不同环境中，使用不同设备从不同患者身上收集的。数据包含数千个基因的基因表达值，以及表示癌症和正常的类别标签。

在三个数据集中匹配共同基因，并根据每个基因的中位数将其二值化。根据这些基因与类别标签的相关性，使用哈佛数据集选出 90 个基因。将密歇根和斯坦福数据集投影到同一组选定基因上，并用作测试数据集。

以下是数据集经二进制格式处理后的摘要。所有数据集均包含 90 个变量。哈佛、密歇根和斯坦福数据集的样本量分别为 156（17）、96（10）和 46（5），括号中的数字表示各数据集中的癌症病例数。

在哈佛数据集上构建模型，并在三个数据集上进行测试。训练数据集和测试数据集按 70 : 30 的比例划分。对于相同环境，使用来自哈佛的测试数据集；对于发生变化的环境，使用来自密歇根和斯坦福的测试数据集（也分别从每个数据集中抽取 30%）。重复次数为 30。使用因果发现工具箱 TETRAD [33]中实现的 PC 算法，在哈佛数据集中寻找父节点。在显著性水平$\alpha = 0.05$下，于哈佛数据集中找到三个父节点。

表 3 展示了几种知名机器学习模型在相同环境和变化环境下的误差。

通过比较每种方法在环境改变与未改变时的误差差异，使用父节点作为特征的模型，其误差差异显著小于使用所有基因作为特征的模型。这表明，使用父节点的模型比使用所有特征的模型更稳健。

4.3 个性化决策示例

为说明如何使用 MODE 进行个性化决策，我们假设德国信用数据集 [7] 满足这些假设。在德国信用数据集 [7], 中，分类属性中的每个值都被转换为二元变量，其中 1 表示该值存在。数值属性通过中位数进行二值化。数据集总共包含 61 个二元变量，包括类别变量。使用 PC-Select [13] , 并采用显著性水平 $\alpha = 0.05$, 在数据集中识别出七个父节点。在该数据集上构建了一个包含 500 棵树的随机森林。

用于说明 MODE 的案例 fo 个性化决策见表 4. 该人员被预测为

模型	Same	已更改	
	哈佛 → 哈佛 哈佛	→ 密歇根 哈佛	→ 斯坦福
MLP	0.0262 ± 0.0191	0.0631 ± 0.0435	0.0901 ± 0.0541
	0.0362 ± 0.0291	0.0447 ± 0.0384	0.0603 ± 0.0291
GBR	0.0184 ± 0.0165	0.0376 ± 0.0289	0.1050 ± 0.0340
	0.0447 ± 0.0252	0.0284 ± 0.0196	0.0716 ± 0.0368
LR	0.0262 ± 0.0221	0.0291 ± 0.0220	0.1043 ± 0.0281
	0.0170 ± 0.0162	0.0241 ± 0.0145	0.0199 ± 0.0167
DT	0.0362 ± 0.0275	0.1085 ± 0.0757	0.0858 ± 0.0456
	0.0461 ± 0.0231	0.0362 ± 0.0327	0.0525 ± 0.0374
RF	0.0220 ± 0.0153	0.0376 ± 0.0260	0.1199 ± 0.0540
	0.0291 ± 0.0240	0.0227 ± 0.0223	0.0652 ± 0.0353
SGD	0.0426 ± 0.0262	0.0844 ± 0.0643	0.1234 ± 0.0866
	0.0369 ± 0.0301	0.0418 ± 0.0328	0.0496 ± 0.0359

表 3：在来自不同实验室的三个真实世界数据集中，使用所有特征（无背景）/仅使用父节点（有背景）的预测模型的误差。环境变化与无环境变化时的误差差异反映了模型的鲁棒性。使用父节点的模型，其误差差异显著小于使用所有特征的模型。

“信用不良”，而 CDE 最高的三个特征让此人可以选择其中一个着手改进，以提升其未来的信用评级。

5 结论与讨论

我们提出了相关假设和理论，以支持估计任意特征对结果的 CDE。对于估计任意特征对结果的因果效应，我们的假设放宽了必须了解数据集背后的完整因果图这一要求。我们假设数据集包含结果的所有直接原因，且不包含受结果影响的变量。只要系统中不存在对 Y 的隐藏直接原因，我们就允许系统中存在隐藏变量。我们的假设为估计任意特征对结果的 CDE 提供了实用方法。我们还表明，在满足这些条件时，可以使用任何类型的预测模型来估计任意特征对结果的 CDE。

论文中的结果表明，由于可以估计每个特征在某种情境下对结果变化的贡献（即控制直接效应），预测模型满足我们所确定的条件，因而具有因果可解释性。此外，我们还表明，这种具有因果可解释性的模型在新环境中进行预测时具有稳健性，即使该环境不同于模型接受训练的环境。

所提出的方法的正确性依赖于一些假设，而验证这些假设需要领域知识。需要注意的是，所有因果效应估计方法都需要一些假设，而本文中的假设集合比假设已知数据背后的因果图更为宽松。

本文工作的意义有两方面。首先，用户很容易理解这组假设，从而知道能否利用预测模型估计数据中的因果效应。其次，在许多应用中，用户了解结果的大多数直接原因，并且可以接受估计的因果效应存在一定程度的偏差。例如，在产品推荐中，即使不知道因果效应的确切大小，了解某项行动是否会对销量产生正向影响也已经很重要。在这些应用中，本文提出的方法将帮助用户预测结果，并为每个个体找到合适的处理（行动）来改变结果。

总而言之，对预测模型进行因果解释需要领域知识。本文的结果提供了一种简便方法，用于整合最高层次的领域知识（包含结果的所有直接原因且不包含任何受结果影响的变量的数据集），而非详细的领域知识（完整的因果图）。这些结果支持在人工引导下开展《机器学习》，以构建具有因果可解释性和稳健性的模型。

参考文献

[1] J. Abrevaya、Y.-C. Hsu 和 R. P. Lieli。估计条件平均处理效应。*J*《商业与经济统计学杂志》，33(4):485–505，2015 年。

一位来自德国信用数据集的人	
原始结果	信用不良，概率为 95%
特征	储蓄账户.1 = 0 （储蓄账户不超过 100 德国马克）
CDE	0.93
干预	*do*(储蓄账户.1 = 1) (开设储蓄账户 >100 德国马克)
新结果	信用良好，概率为 98%
特征	住房.2 = 0 (不拥有房屋)
CDE	0.88
干预	*do*(住房.2 = 1) (拥有住房)
新结果	信用良好，概率为 93%
特征	Credit.history.5 = 0 (不存在其他积分)
CDE	0.85
干预	*do*(Credit.history.5 = 1)（是否有其他贷款）
新结果	概率为 90% 的信用良好

表4：使用 MODE 进行个性化决策的示例。第二行显示了对某人的结果预测。列出了控制直接效应最大的三个特征（及其原始输入值）。该人可以通过对任一特征进行干预 *do*()来改变信用评级。干预后的结果显示为新结果。

[2] A. Acharya、M. Blackwell 和 M. Sen。无偏解释因果发现：检测和评估直接效应。*A*merican Political Science Review，110(3):512–529，2016。[3] C. F. Aliferis、A. Statnikov、I. Tsamardinos、S. Mani 和 X. D. Koutsoukos。用于因果发现和分类特征选择的局部因果与马尔可夫毯诱导：第一部分：算法和实证评估。*J*ournal of Machine Learning Research，11:171–234，2010。[4] S. Athey。超越预测：使用大数据解决政策问题。*S*cience，355(6324):483–485，2017。[5] S. Athey 和 G. Imbens。异质性因果效应的递归分区。*P*roceedings of the National Academy of Sciences，113(27):7353–7360，2016。[6] S. Athey、J. Tibshirani 和 S. Wager。广义随机森林。*T*he Annals of Statistics，47(2):1148–1178，2019。[7] K. Bache 和 M. Lichman。UCI机器学习库，2013。[8] K. Battocchi、E. Dillon、M. Hei、G. Lewis、P. Oka、M. Oprescu 和 V. Syrgkanis。EconML：用于基于机器学习的异质性处理效应估计的 Python软件包。https://github.com/microsoft/EconML，2019。版本 0.x。[9] D. Beer 等。基因表达谱预测肺腺癌患者的生存情况。*N*ature Medicine，8(8):816–824，2002。[10] A. Bhattacharjee 等。通过 mRNA表达谱分析对人类肺癌进行分类，揭示了不同的腺癌亚类。第98卷，第13790–13795页，2001。[11] L. Bottou。使用随机梯度下降进行大规模机器学习。载于*P*roceedings of COMP- STAT'2010，第177–186页。施普普林格，2010。[12] L. Breiman。随机森林。*M*achine learning，45(1):5–32，2001。[13] P. B¨uhlmann、M. Kalisch 和 M. H. Maathuis。高维线性模型中的变量选择：部分忠实分布与 PC-Simple算法。*B*iometrika，97(2):261–278，2010。[14] V. Chernozhukov、D. Chetverikov、M. Demirer、E. Duflo、C. Hansen、W. Newey 和 J. Robins。用于处理和因果参数的双重/去偏机器学习。*a*rXiv 预印本 arXiv:1608.00060，2016。[15] J. H. Friedman。贪婪函数逼近：梯度提升机。*A*nnals of statistics，第1189–1232页，2001。

[16] M. Garber 等。肺腺癌中的基因表达多样性。《美国国家科学院院刊》，98(24)：13784–13789，2001。[17] A. Gelman 和 J. Hill。使用回归和多层次/层级模型进行数据分析。剑桥大学出版社，2006。[18] Y. Goyal、A. Feder、U. Shalit 和 B. Kim。使用因果概念效应（CaCE）解释分类器。arViv，2019。[19] M. A. Hern´an 和 J. M. Robins。因果推断，2010。[20] P. Hoyer、D. Janzing、J. M. Mooij、J. Peters 和 B. Sch¨olkopf。使用加性噪声模型进行非线性因果发现。《神经信息处理系统进展》，21，2008。[21] G. W. Imbens 和 D. B. Rubin。《统计、社会及生物医学科学中的因果推断》。剑桥大学出版社，2015。[22] B. Kim、M. Wattenberg、J. Gilmer、C. Cai、J. Wexler、F. Viegas 等。超越特征归因的可解释性：使用概念激活向量进行定量测试（TCAV）。载于《国际机器学习会议》，第2668–2677页。PMLR，2018。[23] M. C. Knaus、M. Lechner 和 A. Strittmatter。异质性因果效应的机器学习估计：经验蒙特卡洛证据。《计量经济学杂志》，24：134–161，2021。[24] S. R. K¨unzel、J. S. Sekhon、P. J. Bickel 和 B. Yu。使用机器学习估计异质性处理效应的元学习器。《美国国家科学院院刊》，116(10)：4156–4165，2019年2月。[25] C. Louizos、U. Shalit、J. Mooij、D. Sontag、R. Zemel 和 M. Welling。使用深度潜变量模型进行因果效应推断。载于《国际神经信息会议论文集》，第6449–6459页，2017。[26] M. Lu、S. Sadiq、D. J. Feaster 和 H. Ishwaran。使用随机森林方法估计观察性数据中的个体处理效应。《计算与图形统计学杂志》，27(1)：209–219，2018。[27] X. Nie 和 S. Wager。异质性处理效应的准预言式估计。《 Biometrika 》，108(2)：299–319，2021。[28] J. Pearl。直接效应和间接效应。载于《人工智能不确定性会议论文集》，第411?420页，美国加利福尼亚州旧金山，2001。Morgan Kaufmann Publishers Inc。[29] J. Pearl。《因果关系：模型、推理与推断》。剑桥大学出版社，第2版，2009。[30] J. Pearl。因果中介公式？路径与机制评估指南。《预防科学》，13(4)：426–436，2012。[31] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duchesnay。scikit-learn：Python 中的《机器学习》。《机器学习研究杂志》，12：2825–2830，2011。[32] J. R. Quinlan。《C4.5：机器学习程序》。Morgan Kaufmann，美国加利福尼亚州圣马特奥，1993。[33] J. D. Ramsey、K. Zhang、M. Glymour、R. S. Romero、B. Huang、I. Ebert-Uphoff、S. Samarasinghe、E. A. Barnes 和 C. Glymour。Tetrad？一个用于因果发现的工具箱。载于第8届国际气候信息学研讨会，2018。[34]J. M. Robins 和 S. Greenland。直接效应和间接效应的可识别性与可交换性。《流行病学》，第143–155页，1992。[35] D. E. Rumelhart、G. E. Hinton 和 R. J. Williams。通过反向传播误差学习表征。《自然》，323(6088)：533–536，1986。[36] U. Shalit、F. D. Johansson 和 D. Sontag。个体处理效应的估计：泛化界与算法。载于《国际机器学习会议论文集》，第3076–3085页，2017。

[37] S. Shimizu、P. O. Hoyer、A. Hyv¨arinen、A.Kerminen 和 M. Jordan。用于因果发现的线性非高斯无环模型。《机器学习研究期刊》，7(10)，2006 年。[38] P. Spirtes。因果推断简介。《机器学习研究期刊》，11:1643–1662，2010年。[39] P. Spirtes、C. Glymour 和R.Scheines。《因果关系、预测与搜索》。MIT 出版社，第2 版，2000 年。[40] P. Spirtes、C. C. Glymour 和R.Scheines。《因果关系、预测与搜索》。MIT 出版社，第 2版，2000年。[41] T.J.VanderWeele。受控直接效应和中介效应：定义、识别与界限。《斯堪的纳维亚统计学期刊》，38(3):551–563，2011 年。[42] T. J. VanderWeele。中介分析：实践者指南。《公共卫生年度评论》，37:17–32，2016 年。[43] S. Wager 和 S. Athey。使用随机森林估计和推断异质性处理效应。《美国统计学会会刊》，113(523):1228–1242，2018 年。[44] D.Westreich和 S.Greenland。表2谬误：呈现和解释混杂因素及修饰因素系数。《美国流行病学杂志》，177(4):292–298，2013 年。[45] S. Yang、K. Yu、F.Cao、L.Liu、H. Wang 和 J. Li。学习用于稳健领域自适应的因果表示。《IEEE 知识与数据工程汇刊》，35(3):2750–2764，2023 年。[46] L.Yao、Y. Li、S. Li、J.Liu、M.Huai、A.Zhang 和 J. Gao。从因果视角进行概念层面的模型解释。《IEEE 知识与数据工程汇刊》，2022 年。[47] K.Yu、X. Guo、L. Liu、J. Li、H. Wang、Z.Ling 和 X. Wu。基于因果关系的特征选择：方法与评估。《ACM 计算调查》，53(5):1–36，2020年。[48] K.Yu、L. Liu 和 J. Li。因果与非因果特征选择的统一视角。《ACM 知识发现与数据汇刊》，15(4):1–46，2021年。[49] W. Zhang、T.D.Le、L.Liu、Z.-H. Zhou 和 J. Li。挖掘异质性因果效应以实现个性化癌症治疗。《生物信息学》，33(15):2372–2378，2017 年。[50] W. Zhang、J.Li 和 L. Liu。处理效应异质性建模与提升建模的统一综述。《ACM 计算调查》，54(8):1–36，2022 年。[51] W. Zhang、L.Liu 和 J. Li。使用解耦潜在因素进行处理效应估计。载于：《AAAI 人工智能会议论文集，第 35 卷，10923–10930 页，2021 年。