# ADMG 因果数据增强实用指南

**AudreyPoinsot**[1,2], **AlessandroLeite**[1]TAU、INRIA、LISN、巴黎萨克
雷大学[1]Ekimetrics[2]audrey.poinsot@inria.fr,
alessandro.leite@inria.fr

arXiv:2304.01237v1 [cs.LG] 3 Apr 2023

摘要

在小数据场景中应用机器学习（ML）时，数据增强至关重要。它能够生成符合观测数据分布的新样本，同时提高样本的多样性和变化性，帮助研究人员和从业者提升模型的鲁棒性，从而将模型部署到现实世界中。然而，表格数据中的数据增强仍有待改进，因为人们很少考虑底层数据机制的先验知识，这限制了生成数据的保真度和多样性。因果数据增强策略依靠因果图中编码的条件独立性，被认为是应对这些挑战的一种解决方案。在此背景下，本文通过实验分析了无环有向混合图（ADMG）因果数据增强方法在不同设置下的表现，以帮助研究人员和从业者了解先验知识在什么条件下有助于生成新的数据点，进而增强模型的鲁棒性。结果表明，所研究的方法：（a）不受底层模型机制的影响；（b）需要达到一定数量的观测值才能提升机器学习模型的准确率，而在小数据场景中，这一要求可能难以满足；（c）会将离群值传播到增强数据集中，从而降低模型性能；（d）对超参数取值敏感。

## 1 引言

机器学习（ML）模型需要高质量的数据，才能发现有用信息、在未见数据上表现良好，并能适应环境变化。尽管有些模型能够处理含噪声和高维数据集，但在高风险的小数据场景中，使用这些模型通常颇具挑战。在这种情况下，可以采用数据增强技术来弥补训练数据不足，通过人为增加样本数量和训练集的多样性，提升模型性能并减少过拟合（Van Dyk & Meng, 2001）。这类技术已成功应用于计算机视觉（CV）（Zhong etal.,2020; Hendryckset al.,2021）和自然语言处理（NLP）（Xie et al., 2020; Hao et al., 2023）任务中，在训练期间为模型提供正则化，从而帮助减少过拟合。不过，这些技术难以直接扩展到表格数据或时间序列数据（Talavera et al., 2022）。此外，这类技术通常侧重于增加样本的多样性或变异性（Wen et al., 2021），很少同时兼顾两者。

了解底层因果机制，或许能帮助数据增强技术利用因果图（CG）中编码的部分知识来应对这些问题。因此，一旦构建了因果图，我们就可以用它推断数据分布应满足的条件独立关系。由此，可以将干预分布中的数据与增强数据及观测数据相结合（Ilse et al., 2021），以提高数据集的多样性和可变性，并期望增强机器学习模型的鲁棒性。此类策略可以通过遵循因果提升流程（Little & Badawy, 2019），或探索因果图中编码的条件独立性先验知识（Teshima & Sugiyama, 2021）来实现。前者通过对来自干预分布的数据加权来生成新样本。相比之下，后者假设存在无环有向混合图（ADMG）（Richardson, 2003），并同时考虑从每个变量的条件经验分布出发的所有可能重采样路径，以生成新的数据样本。

在此背景下，本文通过实验[1] 评估 ADMG 数据增强方法（Teshima & Sugiyama, 2021）（第 2 节）的特性，并从保真度、多样性和泛化能力（Alaa et al., 2022）的角度，在小数据场景配置下考察不同的问题属性以及存在噪声数据时的表现（第3节）。本文旨在了解在何种条件下，该方法能够帮助从业者利用因果图中编码的先验知识扩充数据集，从而提高其机器学习模型的鲁棒性并应对过拟合。另一个目标是了解在何种条件下，不恰当的参数化设置会导致意外结果，即性能下降。

## 2 ADMG 数据增强

本节介绍 ADMG 因果数据增强方法（Teshima 和 Sugiyama，2021年）。从现在起，我们将此方法称为 CausalDA。

假设我们要使用损失函数 $L$ 在数据集 $D = (D_{train}, D_{test})$ 上训练模型 $f$，该数据集由训练集和测试集组成，且具有 $d$ 个维度。假设已知一个 ADMG 因果图 $\mathcal{G}$，它连接了按该图诱导的拓扑序排列的 $d$ 个变量 $\mathcal{G}$。

我们采用以下符号表示:

- $n = |D_{train}|$训练数据的数量
- $X_k$ 训练集中的 $k^{th}$ 数据点 $D_{train} = \{X_i\}_{i\in[1,n]}$
- $X_k^j$ $k^{th}$ 训练点中 $j^{th}$ 变量的取值， $X_k = \{X_k^1, ..., X_k^d\}$
- $X_k^J$ 对于 $J$ 一组变量， $k^{th}$ 训练点中 $J$ 这些变量的取值
- $D_{aug}$ 使用 $D_{train}$得到的增强数据集
- $Z_i$ 来自 $D_{aug}$的 $i^{th}$ 增强数据点
- $a$（$j$）因果图中该变量的祖先节点 $j$ $\mathcal{G}$

$D_{aug}$ 由训练集中所有观测变量的笛卡尔积构成:

$$D_{aug} = \{Z_i\}_{i\in[1,n^d]}, \quad Z_i = \{X_{i_1}^1, ..., X_{i_j}^j, ..., X_{i_d}^d\}$$

其中， $X_{i_j}$ 是用于复制其 $j^{th}$ 变量的值，以供增强数据点 $Z_i$使用的数据点。

每个 $Z_i$ 都关联一个权重 $w_i$，可将其解释为增强数据点 $Z_i$的存在概率。实际上， $w_i$ 衡量的是在给定变量的祖先值的条件下，增强数据点 $Z_i$ 中变量取值的概率。概率通过核函数估计，其中 $K^j$ 表示用于根据变量的祖先节点估计$j^{th}$ 变量概率的核函数。

$$w_i = \prod_{j=1}^{d} w_i^j = \prod_{j=1}^{d} \frac{K^j(Z_i^{a(j)} - X_{i_j}^{a(j)})}{\sum_{k=1}^{n} K^j(Z_i^{a(j)} - X_k^{a(j)})} \tag{1}$$

最后，使用加权损失在增强数据集上训练模型 $f$:

$$L_{aug}(f) = \sum_{i\in[1,n^d]} w_i L(f, Z_i) \tag{2}$$

实践中，权重通过算法1递归计算。为降低内存和计算成本，该方法允许我们选择一个概率阈值 $\theta \in [0, 1[$，当权重（及其对应的增强数据点）的当前值低于 $\theta$时，立即提前停止对其进行计算。

[1]代码可在以下地址获取github.com/audreypoinsot/admg_数据_增强

**算法1**CausalDA 算法

**输入：** $D_{train} = \{X_k\}_{k\in[1,n]}$, $\mathcal{G}$, $\theta$, $L$, $\{K^j\}_{j\in[1,d]}$ ▷ 假设训练集中的变量和核函数均按照图 $\mathcal{G}$的拓扑序排列

$W_{aug} \leftarrow \{\frac{1}{n}\}^n$
$Z_{aug} \leftarrow \{X_k^1\}_{k\in[1,n]}$
**对于** $j \in [2,d]$，**执行**
  $Z_{aug}^{new} \leftarrow \{\}$
  $W_{aug}^{new} \leftarrow \{\}$
  **对于** $Z_i,\ \ w_i \in Z_{aug},\ \ W_{aug}$，**执行**
    **对于** $i_j \in [1,n]$，**执行**
      $w_i^{new} \leftarrow w_i \cdot \frac{K^j(Z_i^{a(j)} - X_{i_j}^{a(j)})}{\sum_{k=1}^n K^j\ (Z_i^{a(j)} - X_k^{a(j)})}$ )
      $Z_i^{new} \leftarrow \{Z_i;\ \ X_{i_j}^j\}$
      **若** $w_i^{new} > \theta$**则**
        $Z_{aug}^{new} \leftarrow Z_{aug}^{new} \cup Z_i^{new}$
        $W_{aug}^{new} \leftarrow W_{aug}^{new} \cup w_i^{new}$
$Z_{aug} \leftarrow Z_{aug}^{new}$
$W_{aug} \leftarrow W_{aug}^{new}$

**输出：** $\hat{f} \in \underset{f}{\text{取最小值}} \sum_{(w_i,Z_i)i\in(W_{aug},\ Z_{aug})} w_i L(f,\ \ Z_i)$, $D_{aug} = (W_{aug},\ \ Z_{aug})$

## 3 实验设计

### 3.1 数据集

我们依靠合成数据完成了所有实验。这使我们能够全面控制数据所表示的问题。此外，所有模拟数据均从结构因果模型（SCM）中采样，因为 CausalDA 假设数据是通过因果模型生成的。关于结构因果模型的详细定义，请参见（Pearl，2009年）。

我们使用因果发现工具箱（Kalainathan 等，2020年；2022年）生成每个结构因果模型（SCM）。给定节点数量和期望度后，使用 Erd¨os-R´enyi 模型（Erd¨os 和 R´enyi，1959 年）生成每个结构因果模型的有向无环图（DAG）。每次为新增边采样后，我们都会检查是否会导致 DAG 中出现环。机制函数由一组参数化函数（例如线性函数或多项式函数）生成，其参数从给定的概率分布中随机抽取，详见附录 A.4。源变量（即因果图中没有父节点的顶点）使用具有四个成分和球形协方差的高斯混合模型（GMM）生成。最后，在因果机制中引入加性噪声变量。它们彼此独立同分布，并按照正态分布生成。构建好一个结构因果模型后，通过对源变量和噪声变量的实现值进行采样来生成数据。最后，机制函数按照因果图的拓扑序计算各变量的实现值。

### 3.2 评估方法

我们考察了不同的情景，以评估 CausalDA 的特性，并为从业者了解 CausalDA 对其问题可能具有的各种属性的响应提供一些见解。这些情景的默认参数详见附录 A.5，包括：

- **非线性数据生成设定**：通过改变机制所属的函数族，包括线性函数、多项式函数、Sigmoid函数、高斯过程和神经网络。
- **小数据场景**：通过将观测值数量从几个样本增加到一百个样本（即 $[30, 40, 60, 80, 100, 300, 500, 700]$）

- **高维场景**：通过将数据集中的变量数量从七个增加到二十五个（即 $[7, 8, 9, 10, 15, 20, 25]$）
- **高度依赖的输入变量设置**：通过改变 $[0, 1, 2, 3, 4, 5, 6, 7]$ 中因果图的期望度
- **高偶然不确定性设置**：通过改变 $[0.1, 0.2, 0.4, 0.6, 0.8, 1]$ 中的加性噪声幅度
- **带噪声的采集过程**（即离群值）：通过改变 $[0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15]$ 中的异常值比例
- **参数化不足场景**：通过改变 $\theta$ 第 2 节中定义的概率阈值。 $\theta \in [10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$

对于上述每种情形，我们通过测量 Kullback-Leibler 散度、Wasserstein距离以及变量的方差平均相对差异，比较了原始数据集与增强数据集的分布。

我们还将 CausalDA 与基线进行了比较。在这种情况下，我们按 70%，30% 的比例，将原始数据集（$\mathcal{D}$）划分为训练集和测试集。随后，我们分别在加权增强集（$D_{aug} := (Z_i,\ w_i)\quad i \in [1, n^d]$）和原始训练集上训练了两个极端梯度提升（XGBoost）模型，并将后者作为基线。我们在原始数据集的测试集上测量这两个模型的平均绝对百分比误差（MAPE）和 $R^2$ 分数，以预测问题中的每个变量。训练增强数据集上的每个 XGBoost 模型时，均考虑了数据权重；训练原始训练集上的模型时，则使用均匀权重。我们采用三折交叉验证，从 $n$ 个估计器 $\in [10, 50, 200]$ 和正则化系数 lambda $\in [1, 10, 100]$ 的组合中搜索最佳参数集。

对于离群值场景，我们还使用相同的指标比较经过改变的增强数据（即包含离群值）与常规增强数据（即不包含离群值）的分布。

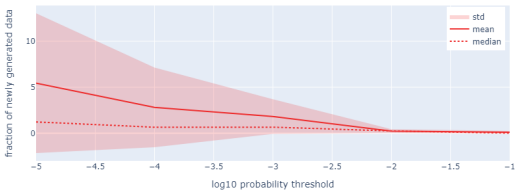
基于观测变量寻找因果图是一个 NP 难组合优化问题，这使得现有方法的可扩展性仅限于几十个变量（Chickering，1996；Chickering 等，2004）。因此，我们选择在本研究中开始探索这一局限性。不过，关于特征数量高于观测值数量的情况，我们将留待未来研究。

## 4 结果

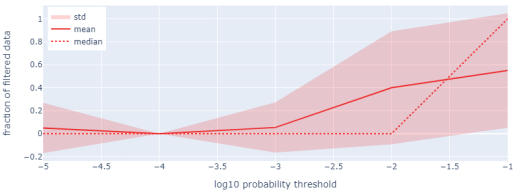
本节介绍第3节所述各场景的结果。附录A.1至A.3展示了补充结果，其中可以看到，当输入变量高度相关时，CausalDA 与基线的表现相近；此外，CausalDA 不受数据的偶然不确定性以及底层生成模型机制的影响。

**参数化不足**。 CausalDA 依赖其概率阈值参数 $\theta \in [0, 1]$（第 2 节）对增强数据进行剪枝，这会影响其分布。概率阈值接近 1 时，会强化观测数据中的相关性；阈值接近 0 时，则会根据因果图放宽这些相关性，从而生成更多数据点，如图 1a、1b 和 1d 所示。随着新生成数据占比增加，方差会降低（图 1c 与图 1a 对比），这表明 CausalDA 并不倾向于增加数据集的多样性，而是改变观测集密集区域中的分布。因此，若选择恰当的概率阈值，我们预计 CausalDA 能通过提供更精细的数据分布，提升机器学习模型在数据支撑集上的预测表现。图 1e 展示了这一发现。要提升 XGBoost 模型的性能，概率阈值参数似乎只有一个非常狭窄的取值范围。因此，从业者必须谨慎设定该参数。实际上，这一阈值可以解释为：给定某变量的父节点后，接受该变量新取值的最小概率。从公式（1）可以很容易看出 $w_i > \theta \implies w_i^j > \theta,\quad \forall j \in [1, d]$。因此，我们建议从业者分析给定各变量父节点时该变量的分布，以便有依据地选择所使用的概率阈值。

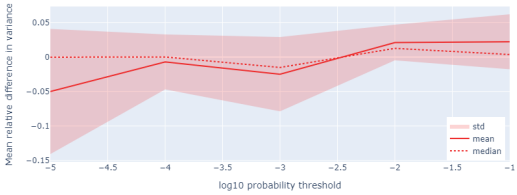
**小数据场景**。 图 2 显示，CausalDA 至少需要 300 个观测值才能提升 XGBoost 的性能。鉴于研究

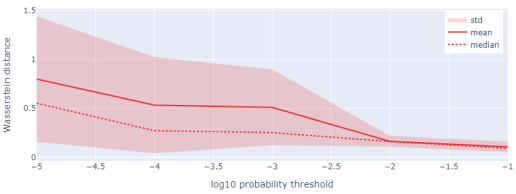



(a) 新数据比例

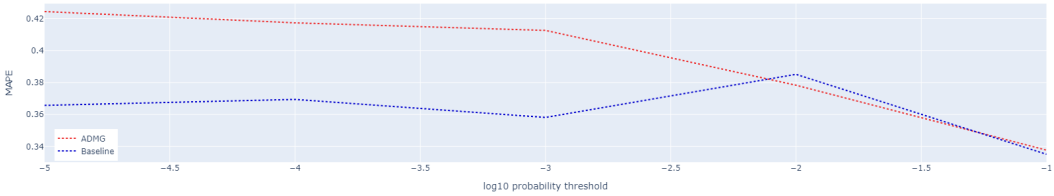



(b) 筛选数据比例



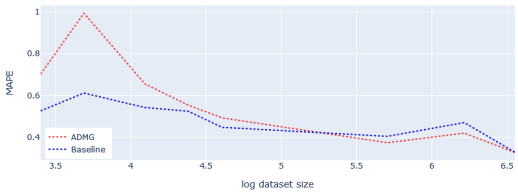




(c) 数据增强后的方差相对差异 (d) 原始集合与增强数据集之间的 Wasserstein距离

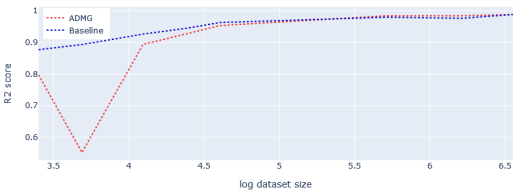



(e) XGBoost 的中位数 MAPE 分数

图 1：CausalDA 输出特征随概率阈值的变化




(a) MAPE 分数中位数

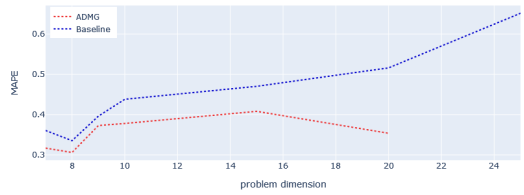



(b) R2 分数中位数

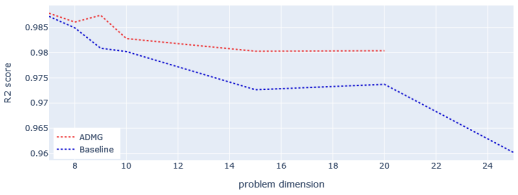
图 2：XGBoost 的中位数表现随数据集大小的变化

情境包括：(a) 无离群值，(b) 使用正确的因果图，(c) 分布内，(d) 数据由高斯混合模型生成，以及 (e) 使用没有明显不连续性或发散性的神经网络。因此，在小数据场景中，使用 CausalDA 改善某些机器学习模型预测可能颇具挑战。观察公式 (1) 可知，数据点数量很少时，核密度估计器会发生过拟合，这可以解释这一现象。同样，由于每个新数据点都是在其父节点取值的条件下生成的，CausalDA 需要有若干观测点具有相同的父节点实现值，才能生成新的数据点。因此，我们建议不要将 CausalDA 视为弥补数据不足的解决方案，而应将其视为一种通过加权数据增强来改进数据分布估计的方法。

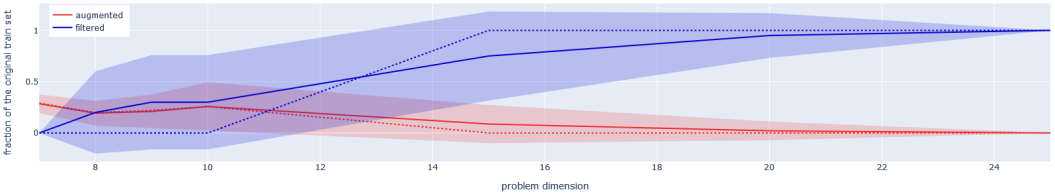
**高维场景**。 图 3 显示，维度增加有利于 CausalDA。事实上，CausalDA 利用因果图中编码的条件独立性先验知识，来提升机器学习模型的泛化能力。相关文献表明，在低维数据中，这类问题具有挑战性（Shah & Peters（2020年）），因为在已知因果图且预期度数确定的高维设置下，会产生更多的条件独立性。图 3 还强调，维度增加会提高整个数据集被筛选掉的概率。事实上，概率阈值可以解释如下：对于给定的 $Z_i$，在所有 $w_i^j$ 均等于 $c$ 的假设下，$\sqrt[d]{\theta}$ 是使 $Z_i$ 不被剪枝的 $c$ 最小值。由于 $\sqrt[d]{\theta}$ 随 $d$ 增加而增加，维度越高，权重不




(a) XGBoost 的 MAPE 分数中位数

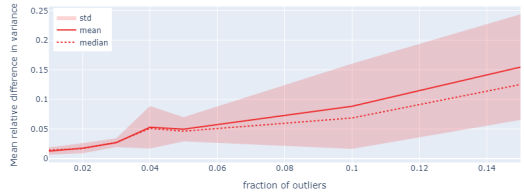



(b) XGBoost 的 R2 分数中位数

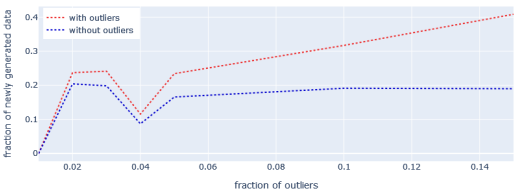



(c) 新增和筛选后数据点的比例，实线表示均值，虚线表示中位数

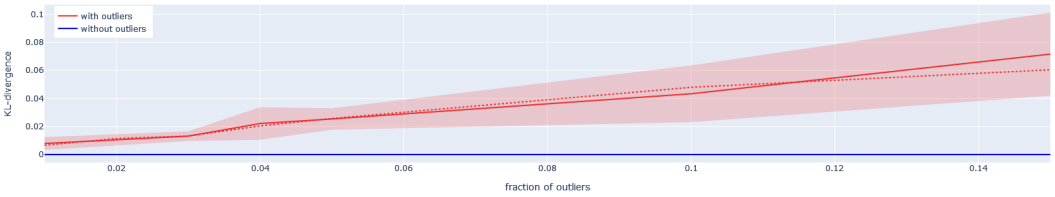
图 3：CausalDA 的性能随问题维度而变化



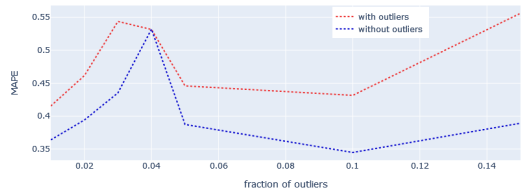

(a) 添加离群值后的方差相对差异



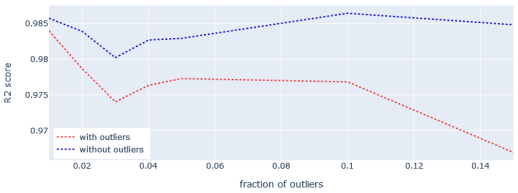

(b) 新数据比例的中位数




(c) 损坏集合与正常集合之间的 KL 散度，实线表示均值，虚线表示中位数




(d) XGBoost 的 MAPE 分数中位数




(e) XGBoost 的 $R^2$ 分数中位数

图 4：不同异常值噪声水平下 CausalDA 的表现

被剪枝的概率就越高（在概率阈值固定的情况下）。因此，从业者在选择适当的概率阈值时，也必须考虑问题的维度。

**带噪声的采集过程**。乍一看，我们可能会认为，得益于其剪枝策略，CausalDA 能够抵御离群值。事实上，当概率阈值足够高时，该方法理论上可以过滤掉概率较低的离群值。然而，即使在最理想的情况下，所有离群值都被过滤掉，它们仍会被核密度估计器纳入计算，从而使估计结果受到污染，并将离群值的影响扩散到增强数据集中。因此，CausalDA 对离群值的鲁棒性仍是一个悬而未决的问题。从图 4 可以看出，CausalDA 会将离群值传播到增强数据集中，导致 XGBoost 的性能下降。

**讨论**。 本节的结果帮助我们了解了 CausalDA 在哪些条件下能够帮助从业者改进其机器学习模型。首先，我们观察到，CausalDA 的性能与底层因果生成过程无关。不过，它会受到采集过程的影响，因为对离群值较为敏感。其次，根据我们的实验，该方法至少需要 300 个样本，因此不适用于小数据场景。相反，它可以利用因果图中编码的先验知识，提供更精细且忠实于观测数据的数据分布，同时不增加多样性，从而改善机器学习模型的泛化能力。第三，CausalDA 在很大程度上依赖概率阈值参数，而从业者可能难以选择该参数。事实上，迄今为止，还没有开发出能够引导用户更好地选择该参数的过程。我们认为，这一点是从业者在真实应用场景中使用该方法时面临的最关键问题。因此，我们希望将未来工作的重点放在自动选择概率阈值上。可能的解决方案包括使用待训练的机器学习模型自动调整该参数，或在计算权重时采用蒙特卡洛树搜索（MCTS）（Kocsis & Szepesv´ari，2006年），而不是剪枝过程。此外，我们还要指出，本文中的实验并不足以全面验证该方法。因此，我们希望进一步深入评估：首先，使用条件独立性检验，检查 CausalDA 是否确实增加了因果图中编码的条件独立关系；其次，在真实数据上评估该方法；第三，分析该方法对错误因果图的敏感性。事实上，构建因果图可能颇具挑战，据我们所知，目前还没有通用的方法来验证其真实性。

## 5 结论与后续工作

在医疗保健和金融等高风险领域应用机器学习时，数据稀缺是一个重大挑战。过去几年里，人们开发了各种方法，帮助研究人员和从业者人为扩充数据集，从而提升机器学习模型的鲁棒性和泛化能力。因果数据增强策略旨在利用因果图中编码的条件独立性来应对这一问题。

本文针对若干情境，对有向无环混合图数据增强方法（Teshima & Sugiyama, 2021）进行了实验分析。研究旨在帮助研究人员和从业者了解，在什么条件下，他们的先验知识有助于生成能够提升模型性能的新数据；同时考察在存在离群值、采用不同误差度量（即偶然不确定性）以及观测数据样本数最少的情况下，数据增强策略参数的影响。实验结果表明，样本量是应用该方法时的关键因素。此外，该方法会传播数据中存在的离群值。对于每个数据集，都必须谨慎设定其超参数。未来，我们计划首先开展更多实验，尤其是使用条件独立性检验和真实数据进行实验；其次，实现超参数优化过程的自动化。

## 致谢

本研究部分经费由欧盟委员会通过 HORIZON计划（TRUST-AI项目，合同编号：952060）提供。

## 参考文献

AhmedAlaa、BorisVanBreugel、Evgeny S Saveliev 和 Mihaela van derSchaar。您的合成数据有多忠实？用于评估和审计生成模型的样本级指标。载于国际机器学习会议，第 290–306 页，2022年。

David Maxwell Chickering。学习贝叶斯网络是 NP完全问题。从数据中学习：人工智能与统计学 *V*，第 121–130 页，1996 年。

Max Chickering、David Heckerman 和 Chris Meek。大样本贝叶斯网络学习是 NP困难问题。《机器学习研究杂志》，5:1287–1330，2004 年。

P. Erd¨os 和 A. R´enyi。《随机图 I》。德布勒森数学出版物，6:290–297，195 9.

Xiaoshuai Hao、Yi Zhu、Srikar Appalaraju、Aston Zhang、Wanqian Zhang、Bo Li 和 Mu Li。《MixGen：一种新的多模态数据增强》。发表于 *IEEE/CVF*冬季计算机视觉应用会议，第 379–389 页，2023年。

Dan Hendrycks、Steven Basart、Norman Mu、Saurav Kadavath、Frank Wang、Evan Dorundo、Rahul Desai、Tyler Zhu、Samyak Parajuli、Mike Guo等。《鲁棒性的多种面貌：对分布外泛化的批判性分析》。载于*IEEE/CVF*国际计算机视觉会议，第8340–8349页，2021年。

Maximilian Ilse、Jakub M Tomczak和Patrick Forr´e。《选择数据增强以模拟干预》。载于国际机器学习会议，第4555–4562页，2021年。

Diviyan Kalainathan、Olivier Goudet和Ritik Dutta。《因果发现工具箱：使用Python揭示因果关系》。《机器学习研究杂志》，21(1):1406–1410，2020年。

Diviyan Kalainathan、Olivier Goudet、Isabelle Guyon、DavidLopez-Paz 和 Mich`ele Sebag。结构无偏建模：因果图的对抗学习。《机器学习研究杂志》，23 62，2022 219 1( ): – 。

Levente Kocsis 和 Csaba Szepesv´ari。基于多臂老虎机的蒙特卡洛规划。载于第*17*届欧洲机器学习会议，第282–293页，2006年。

Max A Little 和 Reham Badawy。因果自助法。*arXiv:1910.09648*，2019年。

Judea Pearl。统计学中的因果推断：概述。《统计学综述》，3:96–146，2009年。

Thomas Richardson。无环有向混合图的马尔可夫性质。《斯堪的纳维亚统计学杂志》，30(1):145–157，2003年。

Rajen D. Shah 和 Jonas Peters。条件独立性检验的难度与广义协方差度量。《统计学年鉴》，48(3):1514 – 1538，2020年。

Edgar Talavera、Guillermo Iglesias、 ´Angel Gonz´alez-Prieto、Alberto Mozo 和 Sandra G´omez-Canaval。《时间序列领域的数据增强技术：综述与分类体系》，2022年。

Takeshi Teshima 和 Masashi Sugiyama。通过简单的数据增强，将因果图先验知识融入预测建模。在人工智能不确定性会议，第 86–96 页，2021年。

David A Van Dyk 和 Xiao-Li Meng。《数据增强的艺术》。《style id='1'>《计算与图形统计学杂志》，10(1):1–50，2001年。

Qingsong Wen、Liang Sun、Fan Yang、Xiaomin Song、Jingkun Gao、Xue Wang 和 Huan Xu。面向深度学习的时间序列数据增强：综述。在第三十届国际人工智能联合会议，2021年。

Qizhe Xie、Zihang Dai、Eduard Hovy、Thang Luong 和 Quoc Le。用于一致性训练的无监督数据增强。《style id='1'>神经信息处理系统进展，33:6256–6268，2020年。

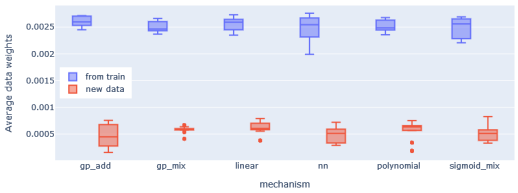
Zhun Zhong、Liang Zheng、Guoliang Kang、Shaozi Li 和 Yi Yang。随机擦除数据增强。*AAAI*人工智能会议，34(07):13001–13008，2020年。

# A 附录

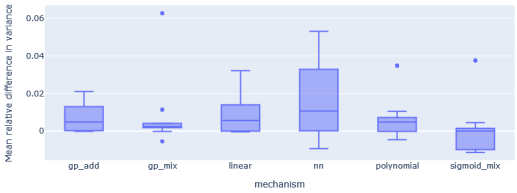
## A.1 非线性数据生成场景 – 结果

由于 CausalDA 只计算密度（见公式（1）），连接各变量的机制类型不应影响该方法的性能，图 5 展示了这一点。实际上，最重要的是变量属于连续变量还是离散变量，因为这会影响所选用的核。此外，每位实践者都可以根据给定其父节点时变量的分布，决定选择不同的核。对于连续变量，常见的做法是使用带 Silverman 带宽的高斯核；对于离散变量，则使用恒等核。

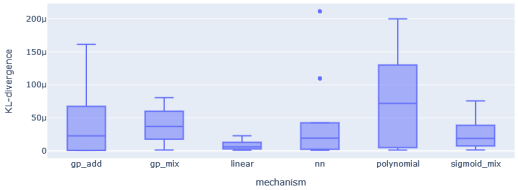
因此，CausalDA 可以在非线性设置下使用，无需特别处理。



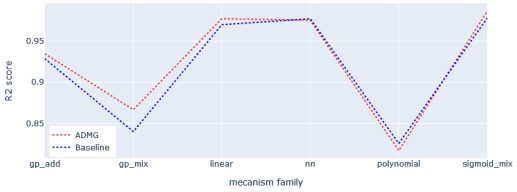

(a) 增强数据集的平均权重




(b) 增强后的方差相对差异




(c) 增强集合与原始集合的 KL 散度



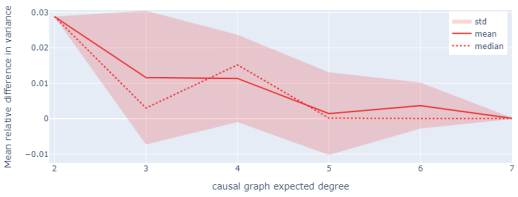

(d) XGBoost 的 $R^2$ 分数中位数

图 5：CausalDA 在不同底层机制下的表现

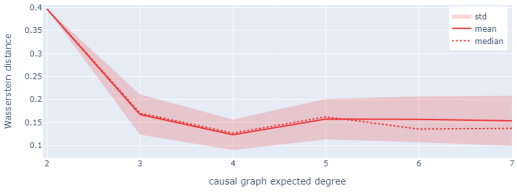
### A.2 高度相关输入变量情景——结果

从图 6 可以看出，输入变量之间的依赖性越强，CausalDA 就越接近基线。对照第 2 节中的方程，这一现象是合理的。实际上，因果图密度增加意味着平均而言父节点数量也会增加。因此，密度是在维度更高的支撑集上计算的，这使得密度值更小、更加稀释，不利于生成新样本。

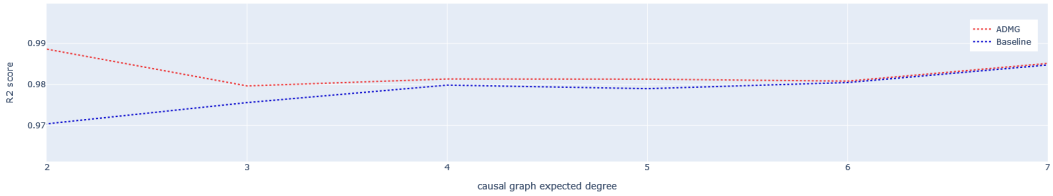
换言之，CausalDA 旨在利用因果图所隐含的条件独立性；边的数量增加时，条件独立性的数量会减少，因此该方法的效用也会降低，因为可供利用的先验知识更少。




(a) 增强后的方差相对差异




(b) 增强数据与原始数据之间的 Wasserstein 距离 sets

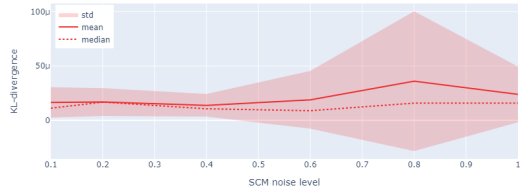



(c) XGBoost 的 $R^2$ 分数中位数

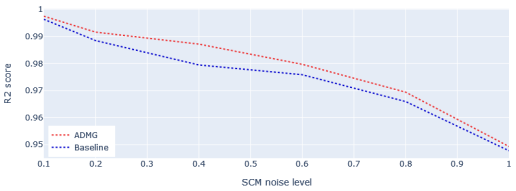
图 6：不同因果图连通程度下 CausalDA 的表现

### A.3 偶然不确定性–结果

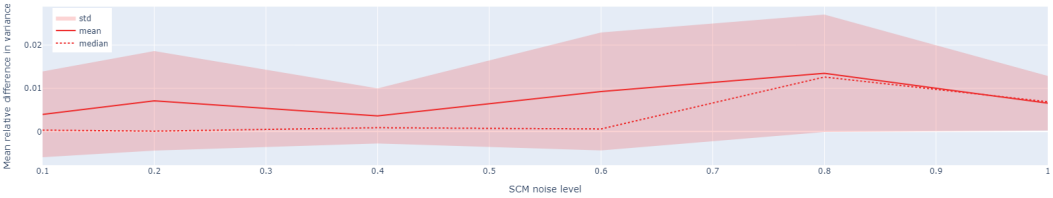
根据图 7 的结果可以断定，生成数据的结构因果模型中引入的噪声幅度对 CausalDA 的结果没有显著影响。这是合理的。事实上，改变结构因果模型噪声的幅度只会缩放密度分布，而这种影响很容易通过按照 Silverman 公式自动优化的高斯核带宽进行补偿。




（a）KL 散度：增强后的集合与原始集合




（b）XGBoost 的 $R^2$ 分数中位数




（c）数据增强后的方差平均相对差异

图 7：不同偶然不确定性水平下 CausalDA 的性能

### A.4 因果发现工具箱 –*AcyclicGraphGenerator* 模块说明

因果发现工具箱（Kalainathan 等，2020年；2022年）是一个用于因果推断的 Python 软件包。它主要专注于观测环境中的因果发现。然而，本节专门介绍我们在实验中使用的*AcyclicGraphGenerator* 模块。

*AcyclicGraphGenerator* 模块能够随机生成与数据集关联的结构因果模型。该生成器允许用户选择数据生成过程中所用的机制，以及噪声贡献的类型（加性和/或乘性）。

目前已实现的机制包括：

- **Linear**: $y = \mathbf{X}W + \times E$
  - $W \sim \mathcal{U}[0,1]^D$
- **Polynomial**: $y = \sum_{i=0}^{d} \mathbf{X}^i W_i + \times E$
  - $d$ the degree
  - $W_i \sim \mathcal{U}[0,1]^D \ \ \forall i$
- **Gaussian Process**: $y = \sum_{i=0}^{D} s_i + \times E$ with and
  - $D$ the number of causes
  - $s_i \sim \mathcal{N}(0, cov(\mathbf{X}_i)) \ \ \forall i$
- **Sigmoid**: $y = \sum_{i=1}^{D} (1+a) \cdot \frac{b \cdot (\mathbf{X}_i + c)}{1 + |b \cdot (\mathbf{X}_i + c)|} + \times E$
  - $D$ the number of causes
  - $a \sim Exp(4)$
  - $b \sim \tilde{b} \cdot \mathcal{U}[-2, -0.5] + (1 - \tilde{b}) \cdot \mathcal{U}[0.5, 2],\ \tilde{b} \sim Ber(0.5)$
  - $c \sim \mathcal{U}[-2, 2]$
- **Randomly initialized Neural network**: $y = \sigma((\mathbf{X}, E)W_{in})W_{out}$
  - $\sigma$ the hyperbolic-tangent activation function
  - $W_{in}$ and $W_{out}$ randomly initialized with the Glorot uniform

其中， $+\times$ 表示加法或乘法，$\mathbf{X}$ 表示维度为 $D$的原因向量， $E$ 表示用于涵盖所有未观测变量的噪声变量。如第3节所述，在我们的实验中， $E \sim \mathcal{N}(0,1)$。

要生成与数据集关联的随机结构因果模型，需要指定：

- 机制的函数族
- 生成过程中使用的噪声类型（“uniform”表示 $\mathcal{U}[-2,2]$ ，“gaussian”表示 $\mathcal{N}(0,1)$）
- 机制中的噪声比例

- 要生成的观测值数量
- 结构因果模型中的节点/变量数量
- 要生成的有向无环图类型（可以是从默认流程中采样的“default”，也可以是从 Erd¨os-R´enyi 模型（Erd¨os 和 R´enyi，1959）中采样的“erdos”；该模型会在新采样的边上添加条件，并检查其是否会导致环）– 如果是“default”：必须指定每个节点的最大父节点数– 如果是“erdos”：必须指定该有向无环图的期望度数

随后，按照以下步骤生成每个结构因果模型：

1. 生成有向无环图 2. 生成机制函数 3. 使用具有四个分量和球形协方差的高斯混合模型生成源变量（即因果图中没有父节点的顶点）。4. 在因果机制中引入噪声变量。它们都是独立同分布的。

构建好一个结构因果模型后，通过采样源变量和噪声变量的实现值来生成数据。接下来，机制函数按照有向无环图的拓扑序计算各变量的实现值。

### A.5 实验默认参数

表 1：未作为研究对象时的参数值

| 参数 | 值 |
|---|---|
| **网络架构** | 2 层全连接神经网络，采用双曲正切激活函数，并使用 Glorot 均匀初始化，包含 20 个神经元 |
| **变量数量** | 10 |
| **因果图期望度** | 3 |
| **加性噪声幅度** | 0.4 |
| **概率阈值** | $10^{-2}$ |
| **异常值比例** | 0 |
| **重复次数** | 20 |
| **核函数** | 采用 Silverman 带宽的高斯核 |