arXiv:2303.00821v1 [cs.LG] 1 Mar 2023

学习高维因果效应

Aayush Agarwal*纽约大学库朗数学研究所 aka7919@nyu.edu

Saksham Bassi*纽约大学库朗数学研究所 sakshambassi@nyu.edu

摘要

高维因果推断数据集的稀缺限制了对复杂深度模型的探索。在这项工作中，我们提出了一种生成高维合成因果数据集的方法。该合成数据集使用 MNIST 数据集和伯努利处理值来模拟因果效应。这为研究各种因果效应估计模型提供了机会。我们在该数据集上使用 Dragonnet 架构（Shi 等人（2019 年））及修改后的架构进行实验。我们使用修改后的架构探索不同类型的初始神经网络层，并观察到修改后的架构在估计方面表现更好。我们发现，无需使用 Shi 等人（2019 年）提出的定向正则化，残差模型和 Transformer 模型就能非常精确地估计处理效应。†

1 引言

高维数据集使我们能够针对各种新任务测试最先进的模型。数据集维度较高时，我们便有机会尝试不同类型的深度学习模型。因此，解决方案是获取并构建图像或人类语言等高维数据，并以此合成用于因果推断的数据集。

深度学习模型无需预先了解底层信息，也能学习模式。Shalit 等人（2017年）的研究至关重要，因为该研究将治疗效应估计的学习能力应用于个体处理效应估计。个体处理效应指量化干预（处理）对单个输入的结果所产生的影响。其公式为$\tau(X) = E[Y|X, T = 1] - E[Y|X, T = 0]$，其中 X 是输入， Y是结果， T是处理。

除了个体处理效应，我们还有平均处理效应（ATE），其公式为 $E\Big[E[Y|X, T = 1] - E[Y|X, T = 0]\Big]$ 或 $E[\tau(x)],$，其中 τ是个体处理效应。在模拟数据集时，我们可以计算真实数据分布的平均处理效应。这样便可以将估计模型的平均处理效应与真实数据分布的平均处理效应进行比较，从而量化估计模型对处理效应的估计效果。这将作为我们实验的一项指标。对于常规回归或分类问题中常见的直接结果变量估计，也可以进行类似比较。当只需分析处理组的处理效应时，我们使用处理组平均处理效应（ATT），其公式为 $E\Big[E[Y_1 - Y_0|T = 1]\Big]$。这里， Y_1 是对输入实施处理时的结果， Y_0 是不对输入实施处理时的结果。 x的另一个重要函数是倾向得分，表示为 $g(X)$，即 $P(T = 1|X)$。倾向得分是输入接受处理的概率。Shalit 等人提出了一种深度学习模型，用于估计处理效应

*共同贡献

†本文的代码请见：https://github.com/sakshambassi/HD-Causal-Effect

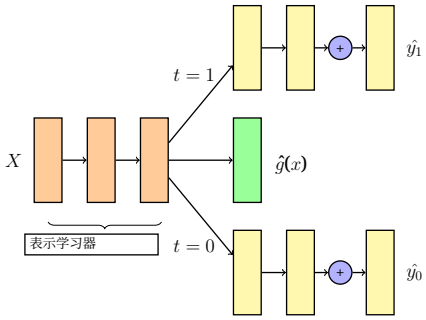

图1：Shi 等人（2019 年）提出的 Dragonnet 架构。$\hat{y}_1$ 和 $\hat{y}_0$ 是结果头。$\hat{g}(x)$ 是倾向得分的估计头。

该模型采用基于深度网络层的架构，包含两个结果头，分别输出Y_1 和 Y_0。Shi 等人提出了一种名为 Dragonnet 的架构，该架构受到 Shalit 等人（2017年）的启发，并在估计目标值 Y时加入了倾向得分估计。Dragonnet 模型提升了估计质量。

在这项工作中，我们研究以下内容：

- 首先，我们提出一种基于 MNIST 图像模拟高维因果效应数据集的方法，以便使用各种深度学习模型开展实验。
- 其次，我们在 Dragonnet 中试验不同的初始层（我们称之为表征学习器，见图1）。我们注意到，对于我们模拟的高维数据集，在不使用定向正则化目标（Shi 等人（2019 年）中介绍）的情况下，我们使用的学习器比 Dragonnet 架构中使用的学习器表现更好。

我们将在第3.2节中更详细地讨论设置，并在第4节中讨论结果。

2 相关工作

此前，在理解因果关系这一领域，人们曾将线性框架应用于非线性数据分布，以寻找因果方向，例如 Hoyer 等人（2008年）的研究。Shalit 等人通过引入一种名为 TARNET 的神经网络架构，在治疗效应估计方面取得了突破。TARNET 的架构与图1相似，但没有倾向评分头估计。Johansson 等人提出了两种方法，将表示学习应用于一组协变量，以进行反事实推断；其中一种方法是名为平衡神经网络的深度学习框架。Louizos 等人（2017年）提出了另一项研究，着重使用变分自编码器进行因果效应估计（个体和总体因果效应）。Curth 等人指出，理解数据生成过程十分重要，因此某些数据集可能比其他数据集更适合某些算法。该论文讨论了基于因果效应估计量的随机森林和神经网络在 IHDP 和 ACIC 数据集上的表现，以推断相同的因果效应。

Shi 等人在其研究中提出了两项内容：首先是一种修改后的神经网络架构（称为 Dragonnet；基于 Shalit 等人（2017年）提出的 TARNET），其次是一种作为目标函数的目标正则化技术（称为 t-reg）。Dragonnet 的修改之处在于，他们增加了另一个估计头，用于估计倾向得分 $g(x)$。他们发现，在治疗效应估计方面，Dragonnet 的表现更好，因为模型本身会在倾向得分误差与结果误差之间选择最佳权衡。Shi 等人还讨论了一种密集网络 Nednet，该网络以协变量 X 作为输入，以处理 t 作为输出。随后，移除 Nednet 的最后几层，并复用该密集网络的权重，以输出潜在结果 Y_0 和 Y_1。之后，论文比较了 Dragonnet 模型与 Nednet 模型，指出前者在处理倾向评分表示（$G(x)$）与预测准确度（Y）之间的权衡时表现更好。

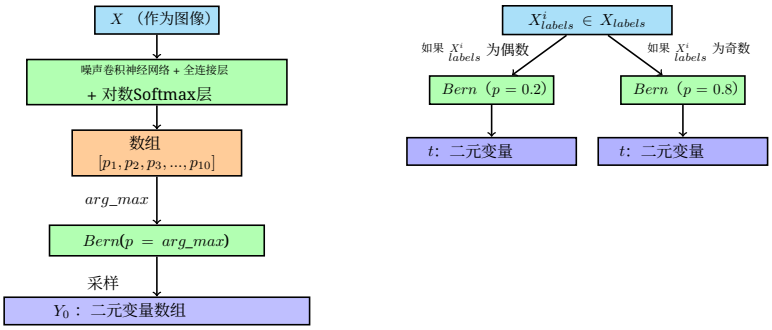

图 2：基于 MNIST 和若干伯努利分布生成合成数据集的模拟过程。正如图中所示，我们对 Y_0 进行了模拟，并通过略微调整模型参数（ Dropout 值和神经元数量），重复这一过程，使用卷积神经网络模拟 Y_1。

3 数据与方法

本节首先介绍 Shi 等人在其研究中使用的 IHDP 数据集。随后，详细阐述前文简要介绍的使用 MNIST 图像（LeCun 等人（2010 年））进行模拟的过程。本节最后一部分将说明我们在本研究中实验的修改后的架构。

3.1 使用 Dragonnet 对 IHDP 数据集进行模拟

Shi 等人（2019 年）使用 IHDP 数据集进行了 50 次重复实验（每次包含 747 个观测值）。每次重复实验包含处理 $t \in \{0, 1\}$、 μ_0 和 μ_1 （其中 $\mu_0 = E[Y_0|X]$ 和 $\mu_1 = E[Y_1|X]$）作为从不同分布模拟得到的输出，以及上述协变量，其中 19 个为二元值，6 个为连续值。根据这些 $treatment$ 和 μ 值，论文通过计算 $\mathbb{E}[\mu_1 - \mu_0]$ 得出平均处理效应的真实值。这有助于衡量估计的平均处理效应相对于该真实值的误差。

3.2 模拟 MNIST 数据集

本模块包含针对图像等高维属性模拟因果效应，并将其作为输入 X。我们选用 MNIST 手写数字数据集（LeCun 等人，2010 年）进行实验。与上述 IHDP 数据集的模拟过程类似，我们创建了诸如 Y_0、 Y_1、 X 和 t 这样的属性。我们通过在卷积神经网络（CNN）中采用不同的随机种子和 Dropout 组合，为 Y_0 和 Y_1 创建不同的分布。具体来说，我们使用带有对数 Softmax 输出的噪声卷积神经网络，找出输出中的最大值，并将其作为概率值传入伯努利分布，以采样单个输出值。我们采用这一过程生成一组 Y_0 和 Y_1。

处理 t 是通过根据 MNIST 数据集中输入图像标签为奇数还是偶数，从不同的伯努利分布中进行采样而生成的。如果标签为奇数，我们从伯努利分布中采样 t，其中取值为 1 的概率为 0.8。如果标签为偶数，我们从伯努利分布中采样 t，其中取值为 1 的概率为 0.2。根据给定样本的处理 t 是 0 还是 1，真实的 Y 取为 Y_0 或 Y_1。我们在图 2 中展示了模拟过程。

与上述模拟的 IHDP 数据集一样，现在我们可以计算真实的平均处理效应值，并将其作为参考，用于计算模型估计的平均处理效应的误差（绝对差值）。

3.3 表示学习器

如第1节引言中所述，高维输入使我们能够尝试不同类型的编码器（表征学习器，见图1），以表示图像嵌入。

在实验中，我们使用 Dragonnet 的稠密网络、ResNet50（He 等（2016））和 ViT（Dosovitskiy 等（2021））作为编码器。训练模型时，我们会隐藏模拟得到的潜在结果 Y_0和 Y_1，并向模型提供 X、 Y 和 t，因为目标函数只考虑依赖于 t 的结果头估计，并根据 Y 的真实值计算损失。相关结果见第4节。

Shi 等人（2019）提出的定向正则化针对的是以连续变量为结果的情形。然而，由于在我们的实验中，模拟的数据集最终结果是一个二元变量，因此我们将正则化修改为纳入扰动项的二元交叉熵，而不是扰动项的平方误差。

3.3.1 ResNet 作为表示学习器

增加输入数据的维度有助于将网络的前几层（编码器）设计为可插拔模块。我们实验使用的附加编码器之一是预训练的 ResNet50（He 等（2016））。随着神经网络深度增加，其参数数量也随之增加，因此模型拟合复杂分布的能力也会增强。然而，He 等（2016）的研究发现，18 层深的模型误差低于 34 层深的网络。这些有悖直觉的观察结果可能归因于损失向各层反向传播时梯度消失。ResNet 在卷积层之间引入了跳跃连接。这种跳跃连接的优点在于，在前向传播中，样本可以选择是否通过复杂卷积或更简单的线性变换进行转换。因此，当模型学习主干网络和跳跃连接的权重时，实际上也会将模型中的卷积层数量选作一个参数，从而使网络深度能够根据所建模的数据集灵活调整。

为了将其用于迁移学习，我们去除 ResNet 模型的分类头，并将图像的特征图传递给两个结果头和倾向评分估计头，如图1所示。

3.3.2 作为表示学习器的视觉Transformer

Transformer（Vaswani 等（2017））此前已用于自然语言处理任务，并带来了更高的准确率和更好的并行化能力。最近，Transformer（Salama（2018））在视觉任务中的应用也日益广泛。与自然语言处理任务类似，图像会被划分为图像块（这些图像块可以重叠），并作为 Transformer 的词元。将其传递给注意力头之前，会使用共享神经网络计算每个图像块的嵌入，然后将这些嵌入与表示该图像块相对于原始图像的位置编码的信息进行融合。

我们像在 ResNet 实验中一样去除末端的分类头，并将其用作表示学习器，如图1所示。对于视觉Transformer（ViT）网络，我们改编了 Salama（2018）中的代码。

4 实验

以下是表1中计算的指标公式。我们观察平均处理效应（称为 ATE_{AE}）的绝对误差（绝对差值）指标。正如上一节所述，真实数据分布的平均处理效应被用作 ATE_{AE} 的真实参考值。

$$ATE_{\text{AE}} = \left| \mathbb{E}\left[Y_1 - Y_0\right] - \mathbb{E}\left[\hat{Y}_1 - \hat{Y}_0\right] \right|$$

ˆ ˆ这里， Y_1、 Y_0 分别是 Y_1、 Y_0 的模型估计值。

我们的实验是在图 2 所示的模拟数据集上进行的。我们使用 Dragonnet 作为基线模型，并按照讨论所述，使用 ResNet50 和 ViT 作为图1中架构的表征学习器。图1中的神经网络唯一的变化是，在单独的实验中，我们还使用 ResNet 和 ViT 作为表征学习器。结果如表1所示。我们

模型	准确率		平均处理效应$_{AE}$	
	不使用 T-REG	T-REG	不使用 T-REG	T-REG
DRAGONNET	54.67	65.97	0.017	0.021
RESNET**50** + Dragon 头	**55.81**	63.50	**0.008**	**0.014**
VIT + Dragon 头	**55.09**	52.42	**0.001**	0.532

表1：模拟留出测试数据集上的结果指标。模型 ResNet50 + Dragon 头和 ViT + Dragon 头是采用 Dragon 头的表征学习器。

模型	治疗前偏倚	
	无 T-REG	T-REG
Dragonnet	−0.0078	−0.0007
ResNet**50** + Dragon 头	−0.0163	−0.0141
ViT + Dragon 头	−0.0230	−0.0113

表2：模拟测试数据上的治疗前偏倚结果。

知道（来自 Alves（2021年）），

$$\underbrace{\mathbb{E}\left[\mathbb{E}\left[Y \mid T=1\right]-\mathbb{E}\left[Y \mid T=0\right]\right]}_{\text{Association}}=\underbrace{\mathbb{E}\left[\mathbb{E}\left[Y_1-Y_0 \mid T=1\right]\right]}_{\text{ATT}}$$
$$+\underbrace{\mathbb{E}\left[\mathbb{E}\left[Y_0 \mid T=1\right]-\mathbb{E}\left[Y_0 \mid T=0\right]\right]}_{\text{Pre-treatment Bias}} \tag{1}$$

如表2所示，所得的治疗前偏倚接近于0。这一点很有帮助，因为该关联取决于我们已知的结果 Y，而不是取决于潜在结果 Y_0 和 Y_1，并且我们无法同时知晓二者。此外，公式（3）强调，处理组和对照组具有可比性，这是因果推断所期望具备的属性。

$$\underbrace{\mathbb{E}\left[\mathbb{E}\left[Y \mid T=1\right]-\mathbb{E}\left[Y \mid T=0\right]\right]}_{\text{Association}}=\underbrace{\mathbb{E}\left[\mathbb{E}\left[Y_1-Y_0 \mid T=1\right]\right]}_{\text{ATT}} \tag{2}$$
$$\therefore \mathbb{E}\left[Y_0 \mid T=1\right] \approx \mathbb{E}\left[Y_0 \mid T=0\right] \tag{3}$$

结论

我们的工作揭示了探索近期深度学习模型（如Transformer和残差网络）以估计处理效应的机会。我们还通过修改Dragonnet架构，研究了估计结果的变化。为了构建一个探索深度网络的平台，我们引入了一个用于创建因果效应数据集的模拟流程。在模拟数据集上的实验结果表明，ResNet50 + Dragon 头模型和ViT + Dragon 头模型在估计误差和准确率指标上表现良好。

参考文献

Alves，M. F.《因果关系导论》，2021年。
https://matheusfacure.github.io/python-causality-handbook/01-Introduction-To-Causality.html。

Curth, A., Svensson, D., Weatherall, J. 和 van der Schaar, M. 条件平均处理效应估计真的做得很好吗? 对处理效应估计中机器学习基准测试实践的批判性审视。载于第三十五届神经信息处理系统大会数据集与基准测试专题（第二轮），2021年。网址：https://openreview.net/forum?id=FQLzQqGEAH。

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby,N. 一张图像胜过 16×16

个词：大规模图像识别的 Transformer。载于《国际学习表征会议》，2021年。URL https://openreview.net/forum?id=YicbFdNTTy。He, K., Zhang, X., Ren, S., and Sun, J. 用于图像识别的深度残差学习。载于《2016年IEEE计算机视觉与模式识别会议》（计算机视觉与模式识别会议），第 770–778 页，06 2016年。doi: 10.1109/CVPR.2016.90。Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. 基于加性噪声模型的非线性因果发现。载于 Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L.（编），《神经信息处理系统进展》，第 21 卷。Curran Associates, Inc., 2008年。Johansson, F. D., Shalit, U., and Sontag, D. 用于反事实推断的表征学习。载于《第33届国际机器学习会议论文集》第 48 卷，ICML'16，第 3020–3029 页。JMLR.org，2016年。LeCun, Y., Cortes, C., and Burges, C. MNIST手写数字数据库。ATT实验室 *[在线]*。可用地址：http://yann.lecun.com/exdb/mnist，2，2010 年。Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., and Welling, M. 基于深度潜变量模型的因果效应推断。载于《第31届国际神经信息处理系统会议论文集》，NIPS'17，第 6449–6459 页，美国纽约州雷德胡克，2017年。Curran Associates Inc. ISBN 9781510860964。Salama, K. 使用视觉Transformer进行图像分类，2018年。URL https://keras.io/examples/vision/image_classification_with_vision_transformer

/.

Shalit, U., Johansson, F. D. 和 Sontag, D. 估计个体处理效应：泛化界与算法。载于第 34 届国际机器学习会议论文集——第 70 卷，ICML'17，第 3076–3085 页。JMLR.org，2017年。Shi, C., Blei, D. 和 Veitch, V. 适配神经网络以估计处理效应。载于 Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. 和 Garnett, R.（编），《神经信息处理系统进展》，第 32 卷。C urran Associates, Inc., 2019 年。Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. 和 Polosukhin, I. Attention Is All You Need。载于 Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. 和 Garnett, R.（编），《神经信息处理系统进展》，第 30 卷。Curran Associates, Inc., 2017年。