« Causal Enhanced Uplift Model » 复现报告

复现论文作者：He X, Xu G, Yin C, et al

摘要

提升模型研究的是因果推断问题，以对个体做出干预手段为原因，以收获个体带来的潜在价值为结果。其本质是为了从人群中区分出易受干预影响人群，作为干预手段实施的目标人群，来最大化潜在价值。但是与传统的监督学习不同的是，没有直接的监督标签，因为在模型训练中不能直接得到提升量，只能得到实施干预或不实施干预的响应结果。为了能够利用提升信息来得到更好的预测模型，此次复现的论文提出了因果增强提升模型，通过是否进行干预和提升效果来划分人群，构建偏序对，利用偏序对的提升差异反向传播指导模型训练，进而保证模型的正确性，此外还利用批损失函数保证个体和群体治疗效果的一致性。本次工作复现了因果增强模型，并且为了进一步得到更准确的模型，在偏序对的提升差异做了正负样本数量收缩。

关键词：因果增强；提升模型；深度学习

1 引言

提升建模（Uplift model）旨在预测一个激励手段（即治疗）对于一个用户能够带来的响应增量价值，例如向一个用户发放优惠券与不发放相比，能增加多大的购买概率。提升建模已经被广泛应用于电子商务[1]，精准医疗[2]和公共政策[3]等诸多方面，在构建提升模型时，最主要的问题是不能直接测量得到提升量，也被称为因果基本问题，在现实生活中，我们只能得到一个用户被干预或者没有被被干预的响应结果，因此提升量（即两者的差）没有监督信号可以使用。为了解决这个问题，本次复现的论文第一次将监督信号应用于神经网络中，通过构造偏序对，使用了隐含的提升信息，得到因果增强模型，进一步增强提升模型的效果。

2 相关工作

现有的提升建模方法主要分为基于元学习者[4]、基于树[5]和基于边界[6]的方法。基于元学习者的方法通常分为两个步骤：首先对治疗照和控制组的响应分别建立一个预测器，然后通过两个预测器的差值进行排序决策。这些预测器被独立地训练，专注于提高各自响应的准确性，但是提升建模更关注的是响应的变化。此外，直接使用两个预测器的差值进行排序可能会遭受耦合误差；基于树的方法通过末端节点测量的提升直接预测单个提升，但是在特征规模较大的场景中训练时间和性能都会遭受挑战；基于边界的方法专注于优化特定评估指标的边界。但是这些指标都与提升没有直接关系。

3 本文方法

3.1 本文方法概述

为了解决现有方法存在的限制，提出了一种因果增强提升模型（CEUM），通过在正负样本上构造偏序对来利用隐含的提升信息，最终提升模型的目的是为了在批中找到平均提升效应，来接近个体提升，在训练过程中利用损失函数保证个体平均提升和群体治疗的一致性。

3.2 构造具有偏序关系的样本对；

本质上，我们希望模型将人群划分为以下四类人, 如图 1(a) 所示：干预敏感人群：只要增加干预手段就会带来潜在价值。同时不加干预手段，就不会带来潜在价值; 自然转化人群：不管有没有干预都会带来潜在价值的人群; 无动于衷人群：不管有没有干预会带来潜在价值人群; 反作用人群：没有干预会带来潜在价值，有了干预反而不会带来潜在价值. 然而，在现实世界中，完全区分这四类人是不可能的，所以论文中对人群进行重新划分，根据是否干预和是否带来潜在价值将人群划分为四类，如图 1(b) 所示：g_1: 代表所有在干预组并且会带来潜在价值的人群。那么这其中就包括了干预敏感人群和自然转化人群; g_2: 代表所有在对照组并且会带来潜在价值的人群。其中包括了自然转化人群和反作用人群; g_3: 代表所有在治疗组但是不会带来潜在价值的人群。其中包括了无动于衷人群和反作用人群; g_4: 代表所有在对照组也不会带来潜在价值的人群。其中包括了干预敏感人群和无动于衷人群。

从图 1(c) 中可知，g_1 的提升要高于 g_2,g_3, 同时 g_4 提升高于 g_2,g_3, 将 g_1 和 g_4 作为一个集合，g_2 和 g_3 作为一个集合，通过这俩个集合构造偏序对 <x_1,x_2>,x_1,x_2 可以分别属于 g_1 和 g_4 组成的集合，也可以属于 g_2 和 g_3 组成的集合，但是 x_1,x_2 不能同时属于一个集合，同时引入 y_{diff} 这个变量，当 x_1 属于 g_1 和 g_4 组成的集合时，y_{diff} 值为 1，当 x_2 属于 g_2 和 g_3 组成的集合时，y_{diff} 值为-1，

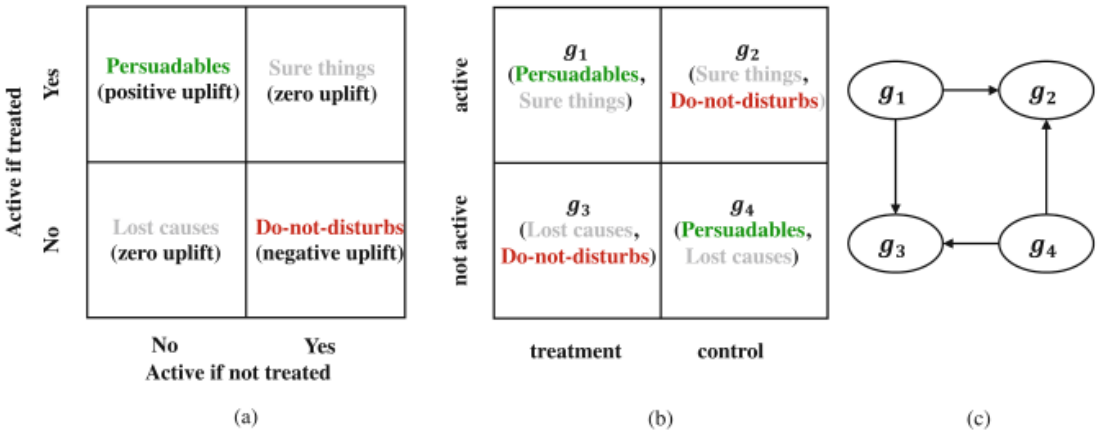

图 1: 偏序对构造图 (来自原论文)

3.3 提升计算

将 x 表示为用户的特征，将 t 表示为潜在的干预。用户 x 在处理条件 t 下对应的干预结果记为 Y_t(x). 对于二元治疗集 T=0,1, 其中 0 表示“对照组”，1 表示“干预组”，分别对应结果变量 Y_0(x) 和 Y_1(x)。在这个意义下，个体治疗效果 ITE，被定义为 ITE=Y_1(x)-Y_0(x)。对于提升模型，要推断的目标量是条件平均处理效果 CATE[7]来逼近个体治疗效果,CATE 被定义为 CATE=E[Y_1(x) - Y_0(x)|X=x].

有了上面的符号，我们可以简单描述模型。给定数据集 (x_i,t_i,Y_1), 其中 Y_i(x)=$t_i Y_1(x_i)$+(1-t_i)$Y_0(x_i)$, 训练分类模型 f:X x T->Y, 来拟合 $\hat{y}_i$=f(x_i,t_i), 那么未观察到的反事实结果为 $\hat{y}_i^C$ 可以通过 f(x_i,1-t_i) 被计算, 用户的提升值为 $\hat{u}_i = (-1)^{1-t_i}\left(\hat{y}_i - \hat{y}_i^C\right)$。

那么对于研究论文的数据集是由偏序对构成,<x_1,x_2>, 当 x_1 取自 g_1 和 g_4 时，D_1=(t_1,x_1,x_2,y),y_{diff}=1, 当 x_2 取自 g_2 和 g_3 时，D_2=(t_1,x_1,x_2,y),y_{diff}=-1, 总的数据集为 D=D_1UD_2.

3.4 损失函数

由于提升模型的主要目的是将用户进行分类得到营销敏感人群，所以损失函数使用交叉熵损失：

$$L_{CE}(y,\hat{y}) = -\sum_{i=1}^{n}\left(y_i \log \hat{y}_i + (1-y_i)\log\left(1-\hat{y}_i\right)\right) \tag{1}$$

每一对偏序对都包含了提升差的隐含信息，利用提升差反向传播增强提升效果，使用合页损失函数来优化：

$$L_{\text{diff}} = \sum_{i=1}^{n}\left[\text{margin}_1 - y_{\text{diff}}\, diff_{\text{uplift}}\right]_+ \tag{2}$$

其中， $\text{diff}\, f_{\text{uplift}} = \text{uplift}_{x_1} - \text{uplift}_{x_2}$,$margin_1$ 是量化辨别阈值的超参数

为了保证提升模型的准确性和个体和群体的一致性的损失函数,

$$L_{ATE} = \sum_{i=1}^{B}\left[-\text{margin}_2 + \|\hat{\tau}_i - \tilde{\tau}_i\|^2\right]_+ \tag{3}$$

其中,$\hat{\tau}_i = \sum_{j\in b_i}\frac{1}{n_{i1}} t_j y_j - \frac{1}{n_{i0}}\left(1-t_j\right) y_j$,$\tilde{\tau}_i = \frac{1}{n_i}\sum_{j\in b_i}\text{uplift}_j$

所以总的损失函数是

$$L = L_{CE} + \lambda L_{diff} + \beta L_{ATE} \tag{4}$$

3.5 评估指标

提升曲线下面积（AUUC）[8]，n_T 和 n_C 分别表示数据集 D 中的处理子集和控制子集，n_T 和 n_C 表示对应的数据集大小。$f\left(D_T, \frac{p}{100}n_T\right)$ 和 $f\left(D_T, \frac{p}{100}n_C\right)$ 分别为 n_T 和 n_C 的前 p 个百分比，且按照模型 f 的预测排序。

$$\widehat{AUUC}\left(f, D_T, D_C\right) = \int_0^1 V(f,x)dx \approx \sum_{p=1}^{100} V\left(f, \frac{p}{100}\right) \tag{5}$$

其中，

$$V\left(f, \frac{p}{100}\right) = \frac{1}{n_T}\sum_{i\in f\left(D_T, \frac{p}{100}n_T\right)} y_i - \frac{1}{n_C}\sum_{j\in f\left(D_C, \frac{p}{100}n_C\right)} y_j \tag{6}$$

4 模型介绍与改进

4.1 模型设计

ceum 模型是一个多任务的模型，主任务是对给定特征向量 x 的用户进行 t 干预（此处是二值干预）预测用户带来提升量，还包括俩个辅助任务，一个是利用正负样本构建的具有偏序关系的对之间的提升量差值进行反向传播，对模型进行指导训练。另一个是利用损失函数维持在个体层面和群体层面的一致性，维持模型正确性。

4.2 模型实现

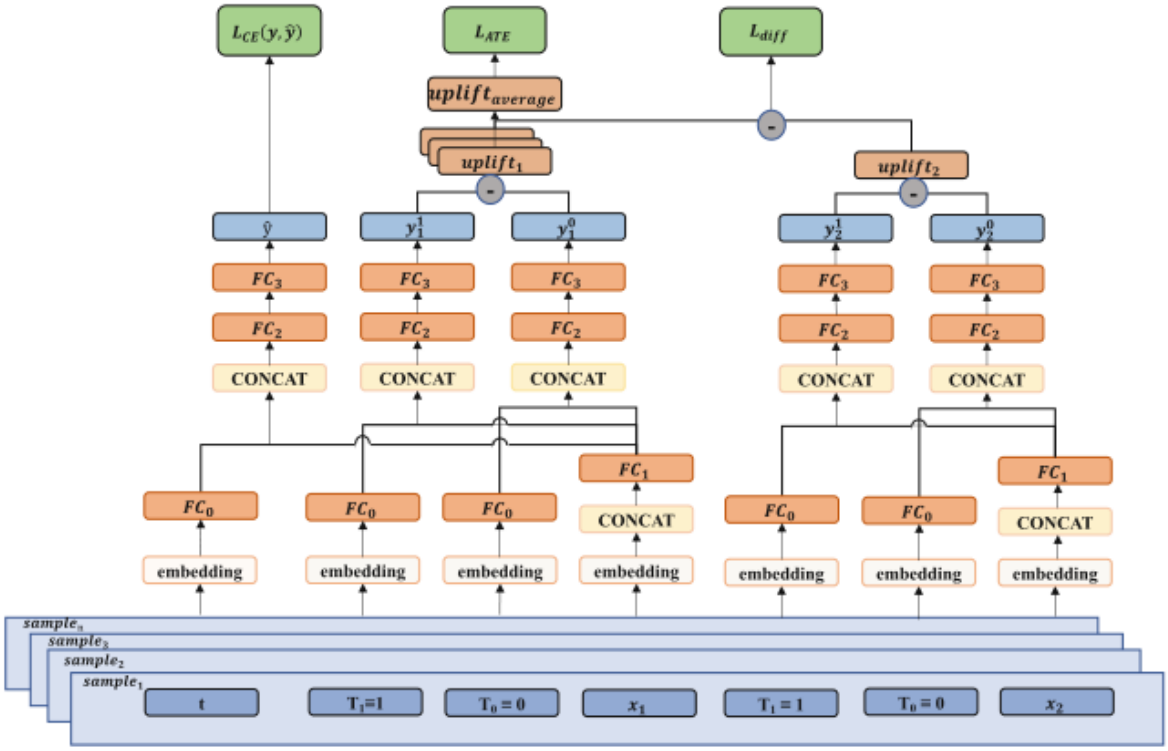

图 2: 因果增强模型结构图 (来自原论文)

治疗表示层: 将所有的干预 (即 t, T_1≡1 and T_0≡0) 通过 embedding 转换为向量，然后通过线性层 fc_0 得到它们的表示。

用户表示层: 每个 x 包含 n 个特征值，x = x_1，…，x_i，…，x_n，其中 x_i 是个体的第 i 个特征的索引。通过 embedding 层将它们分别嵌入到不同的向量 v = v_1，…，v_i，…v_n,n 个向量的拼接，v_1;…;v_i;…;v_1 被馈送到一个全连接层 (fc_1)，ReLu 作为激活函数。输出的是特征 x 的编码表示。

我们连接 r_t 和 rx_1 俩个向量，并以 ReLu 和 sigmoid 作为激活函数，将其传入两个连续的全连通层 (即 fc_2 和 fc_3)。最终得到各种响应结果，利用偏序对的提升量之间的差值，以及预测响应和实际响应差值，数据批内的平均提升量和预测的平均处理效果差值，构建损失函数反向指导模型训练。

4.3 与已有开源代码对比

没有开源代码

算法伪代码：

Procedure 1 Causal Enhanced Uplift Model

Input: feature X_1, treatment t_1,table y_1,feature X_2, treatment t_2,table y_2
Output: metrics $u\_at\_k, qini\_coef, uplift\_auc, wau$
$Initialize\ model$
for *epoch* ***in*** *num_epcho* **do**
  for *b* ***in*** *batch_size* **do**
    $uplift_{x1}$, $uplift_{x2}$ *and* $\hat{y_1}$ *are obtained by neural network training.*
    $Calculate\ the\ loss\ L_{CE}(y_1, \hat{y_1}) = -\sum_{i=1}^{n}\left(y_1 \log \hat{y_1} + (1-y_1)\log\left(1-\hat{y_1}\right)\right)$
    $Calculate\ the\ loss_{\text{diff}} = \sum_{i=1}^{n}\left[\text{margin}_1 - y_{\text{diff}}\, diff_{\text{uplift}}\right]_+$
    $where\ \ \text{diff}\, f_{\text{uplift}} = \text{uplift}_{x_1} - \text{uplift}_{x_2}$
    $Calculate\ the\ loss\ L_{ATE} = \sum_{i=1}^{B}\left[-\text{margin}_2 + \|\hat{\tau}_i - \tilde{\tau}_i\|^2\right]_+$
    $where, \hat{\tau}_i = \sum_{j\in b_i}\frac{1}{n_{i1}} t_j y_j - \frac{1}{n_{i0}}\left(1-t_j\right) y_j, \tilde{\tau}_i = \frac{1}{n_i}\sum_{j\in b_i}\text{uplift}_j$
    $loss = L_{CE} + \lambda L_{diff} + \beta L_{ATE}$
    $backward\ propagation,\ calculate\ the\ gradients,\ update\ the\ paramters$
  end
  $calculate\ u_at_k,\ qini_coef,\ uplift_auc,\ wau$
end

4.4 创新点

<x_1,x_2>为偏序对，如果 x_1 和 x_2 分别在样本采集时，能确定的数量关系有，num($x1_{g1+g4}$)=num($x2_{g2+g3}$), num($x2_{g1+g4}$)=num($x1_{g2+g3}$), 但是在这其中 g_1,g_2,g_3,g_4 单独的比例是不确定的，那么在 $uplift_x1$ 与 $uplift_x2$ 作差时，由于 g_1,g_2,g_3,g_4 在实验中起的作用或者贡献是不同的，比例不同，造成提升结果差异，将影响实验的正确性。为了减小最上面提到的不同比例，贡献不同，造成误差，可以做缩放。

5 实验结果分析

5.1 数据集介绍

Hillstrom[9]是电子邮件广告随机试验后收集到的数据集。在这个数据集中顾客被随机分为两个实验组和一个对照组。将接收“女性商品电子邮件”作为干预，响应结果是客户的访问状态。Criteo[10]提升模型数据集 (Criteo-v2) 是从数字广告业务的随机对照实验中收集。干预手段是顾客收到促销广告，响应结果是顾客的访问状态。有关数据集的详细信息请参见表 1。

Metric	Hillstrom	Criteo-v2
Total size	42,693	13,979,592
Treatment/control ratio	1.0038	5.6667
Features variables	8	12
Treatment positive rate	0.1514	0.04854
Control positive rate	0.1062	0.03820
Group positive rate	0.1288	0.04699

表 1: 数据集信息

5.2 实验结果

由于数据集数据量太大，所以在数据集中进行部分采样，训练集与测试集数据量比例为 7:3，在训练集中构建偏序对并对模型进行训练，最终得到实验结果如表 2 所示。

从实验结果分析，利用偏序对构建的因果增强提升模型比传统的 S_Learner 模型提升效果好, 对偏序对进行收缩得到的改进后的因果增强提升效果更好。

	Hillstorm	**Criteo-v2**
Models	AUUC	AUUC
S-Learner	0.01955	0.00630
ceum	**0.02320**	**0.00765**
ceum$_{imp}$	**0.02593**	**0.00902**

表 2: 实验结果

6 总结与展望

复现论文是一个多任务提升模型框架，包括主任务和俩个辅助任务。该模型缓解了没有监督标签问题。利用数据集中治疗和转化的关系构建偏序对，实际上是增加了辅助量，利用具有偏序关系的对之间的提升差异，反向传播矫正 ATE, 减少预测偏差。本次工作复现了因果增强模型还在偏序对的构建上可以进一步改进，获得更好的提升模型。

参考文献

[1] LI S, VLASSIS N, KAWALE J, et al. Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns[J]. IJCAI'16 2016: 3768-3774.

[2] JASKOWSKI M, JAROSZEWICZ S. Uplift modeling for clinical trial data[J]., 2017.

[3] GRIMMER J, MESSING S, WESTWOOD S J. Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods[J]. Political Analysis, 2017, 25(4): 413-434. DOI: 10.1017/pan.2017.15.

[4] KÜNZEL S R, SEKHON J S, BICKEL P J, et al. Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning[J]., 2017.

[5] NICHOLAS J, PATRICK D, SURRY R. Real-World Uplift Modelling with Significance-Based Uplift Trees[J]., 2012.

[6] BETLEI A, DIEMERT E, AMINI M R. Uplift Modeling with Generalization Guarantees[J]., 2021.

[7] ZHANG W, LI J, LIU L. A Unified Survey of Treatment Effect Heterogeneity Modelling and Uplift Modelling[J]. ACM Comput. Surv., 2021, 54(8).

[8] DEVRIENDT F, GUNS T, VERBEKE W. Learning to rank for uplift modeling[J]., 2020.

[9] HILLSTROM K. The minethatdata e-mail analytics and data mining challenge[J]., 2018.

[10] EUSTACHE D, ARTEM B, RENAUDIN C, et al. A large scale benchmark for uplift modeling.[J]., 2018.