计算机科学
COMPUTER SCIENCE

基于深度强化学习的长期因果效应估计

柳家起, 汪玉杰, 相国督, 俞奎, 曹付元

引用本文

柳家起, 汪玉杰, 相国督, 俞奎, 曹付元. 基于深度强化学习的长期因果效应估计[J]. 计算机科学, 2026, 53(4): 235-244.

LIU Jiaqi, WANG Yujie, XIANG Guodu, YU Kui, CAO Fuyuan. Long-term Causal Effect Estimation Based on Deep Reinforcement Learning [J]. Computer Science, 2026, 53(4): 235-244.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250600043

基于深度强化学习的长期因果效应估计

柳家起[1,2]　汪玉杰[1,2]　相国督[1,2]　俞　奎[1,2]　曹付元[3]
1 合肥工业大学计算机与信息学院　合肥 230601
2 大数据知识工程教育部重点实验室(合肥工业大学)　合肥 230601
3 山西大学计算机与信息技术学院(大数据学院)　太原 030006
(liujiaqi@mail.hfut.edu.cn)

摘　要　因果效应估计旨在计算处理变量对结果变量的因果作用大小。现有主流因果效应估计方法主要适用于静态数据或时间序列中的单个时间点,无法有效估计处理变量在长期时间内对结果变量产生的累积影响。为解决这一问题,基于传统强化学习的长期因果效应估计方法通过线性基函数来拟合长期潜在结果,从而计算长期因果效应。然而,由于线性基函数在复杂场景下的表达能力有限,现有方法不能准确识别弱因果效应,同时在数据维度提高时会出现明显的性能退化问题。针对上述问题,提出了一种基于深度强化学习的长期因果效应估计方法。该方法采用对决网络估计长期潜在结果,能够有效估计处理变量对结果变量的影响,从而大幅提升算法对弱因果效应的识别能力;同时,所提方法避免了基函数选择不当而导致估计长期潜在结果时出现的偏差。实验结果表明,所提方法在统计学合成数据集和订单调度模拟数据集上优于现有算法。

关键词:长期因果效应估计;潜在结果模型;深度强化学习
中图分类号　TP181

Long-term Causal Effect Estimation Based on Deep Reinforcement Learning

LIU Jiaqi[1,2],WANG Yujie[1,2],XIANG Guodu[1,2],YU Kui[1,2] and CAO Fuyuan[3]
1 School of Computer Science and Information Engineering,Hefei University of Technology,Hefei 230601,China
2 Key Laboratory of Knowledge Engineering with Big Data(the Ministry of Education of China),Hefei University of Technology,Hefei 230601,China
3 School of Computer and Information Technology(School of Big Data),Shanxi University,Taiyuan 030006,China

Abstract　Causal effect estimation aims to calculate the magnitude of the causal effect of the treatment variable on the outcome variable. The existing prevalent causal effect estimation methods are mainly applicable to static data or a single time point in time series,and cannot effectively estimate the cumulative impact of the treatment variable on the outcome variable over a long period of time. To solve this problem,the long-term causal effect estimation method based on traditional reinforcement learning fits the long-term potential outcomes through linear basis functions,thereby calculating the long-term causal effect. However,due to the limited expressive power of linear basis functions in complex scenarios,existing methods cannot accurately identify weak causal effects,and at the same time,there will be significant performance degradation problems when the data dimension increases. In response to the above problems,this paper proposes a long-term causal effect estimation method based on deep reinforcement learning. This method uses the dueling network to estimate long-term potential outcomes,which can effectively estimate the impact of the treatment variable on the outcome variable,thereby greatly improving the algorithm's ability to identify weak causal effects. Meanwhile,the proposed method avoids the biases that occur when estimating long-term potential outcomes due to improper selection of basis functions. Experimental results show that the proposed method outperforms existing algorithms on statistical synthetic datasets and order scheduling simulation datasets.

Keywords　Long-term causal effect estimation,Potential outcome model,Deep reinforcement learning

因果效应估计用于计算处理变量对结果变量的影响强度,被广泛应用于临床医学[1]、广告推荐[2]等领域。例如,研究者通过计算药物对疾病的因果效应来判断药物的治疗效果,从而指导医生制定合适的医疗方案或研究[3]。随机对照试验被广泛认为是计算因果效应的"黄金标准"[4]。然而,在实际应用中,随机对照试验受到伦理、经济成本等方面的

到稿日期:2025-06-08　返修日期:2025-11-02
基金项目:国家科技重大专项(2021ZD0111801);国家自然科学基金(62376087)
This work was supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China (2021ZD0111801) and National Natural Science Foundation of China(62376087).
通信作者:俞奎(yukui@hfut.edu.cn)

约束，往往不可行[5]。研究者基于潜在结果模型提出了大量从观测数据中计算因果效应的方法，这些方法主要分为静态因果效应估计方法和时间序列因果效应估计方法两类。静态因果效应估计方法不关注处理变量对结果变量的影响随时间如何变化[6]。相比之下，时间序列因果效应估计方法在静态分析的基础上引入了时间维度[7]，旨在研究处理变量随时间推移对结果变量产生的动态影响。

这些方法尽管已经在多个领域被广泛应用，但是在策略评估[8]这类需要依据长期累积的结果进行决策的场景中仍然无法使用。因此，本文的工作更关注长期因果效应估计，即处理变量对结果变量产生的长期累积影响。

在实际应用中，长期累积影响可以依据长期利润、长期收入等累积的指标的变化程度来衡量。例如，在网约车平台（如Uber、滴滴）进行大规模网约车调度管理中，判断更新订单调度策略是否能帮助司机赚取更多利润时，更新订单调度策略带来的长期累积影响应该通过司机的长期累积收益的变化程度来衡量。在使用因果效应估计的方法进行策略评估时，将更新订单调度策略看作是接受处理，而维持原订单调度策略是不接受处理。在此过程中，会遇到两个乘客同时请求同一辆网约车的情况。如图1所示，司机使用不同的订单调度方式会接到不同的乘客，在未来到达对应的地点时会获得不同的利润。若使用时间序列因果效应估计进行策略评估，其会根据单个时间点的因果效应进行判断，如在t_0时刻由于策略$Policy_0$所产生的实时结果$Profit_0$更高，这种方法会得到更新订单调度策略不能帮助司机赚取更多利润的结论。相反，使用长期因果效应估计的方法进行策略评估，则会依据长期的累积结果来判断。如图2所示，该方法在t_0时估计接受处理和不接受处理后从t_0到t_n时刻的累积结果，从而可以判断更新订单调度策略有助于司机赚取更多的总利润。两种不同的因果效应估计方法得到了不同的结论，原因是$Policy_0$获得更大的单笔订单收益的同时会引导司机前往更远的地区，导致司机下次接单的位置在偏远地区（如郊区、工厂等）的概率更大，这种位置不利于司机获取订单，进而导致总收入降低。因此，使用时间序列的因果效应估计方法可能无法获得最优（长期利润最高）的订单调度策略。

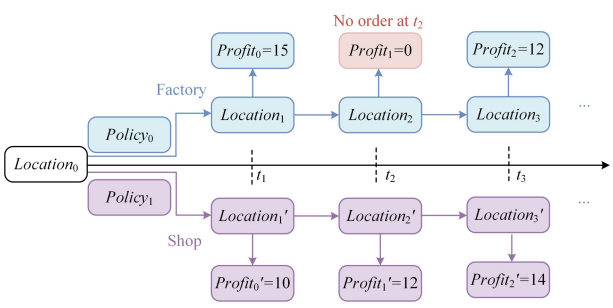


图1 在t_0时司机接受不同处理后的实时收益和位置信息

Fig. 1 Real-time earnings and positional information of drivers following their assignment to different treatment conditions at time t_0

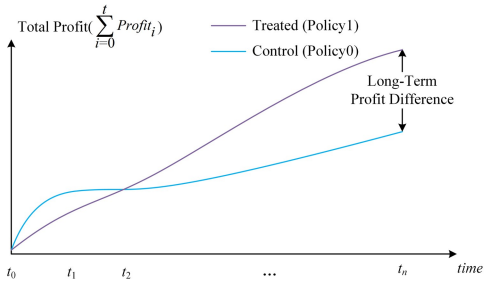


图2 t_0时刻司机接受不同处理后总收益的变化情况

Fig. 2 Variation in drivers' total earnings at time t_0 following their assignment to different treatment conditions

然而，在计算长期因果效应时，需要使用未来无法观测的数据。针对这一问题，研究者提出利用强化学习进行长期因果效应估计的方法[9]。该方法首先通过线性基函数拟合长期潜在结果，并推导出长期因果效应的分布，然后利用观测数据进行假设检验来判断处理变量是否对结果变量产生了长期累积影响。然而，这种方法存在以下缺点。首先，在基函数拟合的方法中，选择合适的基函数至关重要，但这通常需要对问题领域有广泛的先验知识，如果不能选择合适的基函数，会严重影响模型的性能[10]。其次，该方法拓展性较差。随着协变量维度的增加，所需的基函数快速增多，导致模型无法使用[10]。最后，该方法识别弱因果效应（即处理变量对结果变量影响较弱的情况）的准确率较低。在这种情况下，如果个体基数较大，那么弱因果效应识别不准确可能带来较严重的后果。

为了解决上述问题，本文提出了一种基于深度强化学习的长期因果效应估计方法（Long-term Causal Effect estimation based on deep reinforcement Learning，LCEL）。具体而言，在每个时间点，LCEL首先使用实时观测数据，依据奖励机制对所提出的长期潜在结果网络进行更新，从而可以不断地从观测数据中学习处理变量在各种条件下能产生的累积结果；然后，使用最新的长期潜在结果网络估计处理变量在该时变协变量下产生的累积结果，即长期潜在结果；最后，利用所得的长期潜在结果估计长期因果效应。本文的贡献总结为：

1）提出了一种基于深度强化学习的长期因果效应估计方法LCEL。相比于传统强化学习的方法，LCEL无须依赖大量的先验知识选取基函数，避免了基函数选取不当而导致的长期潜在结果估计不准确的问题；同时，LCEL有效缓解了现有方法在数据维度升高时不能有效计算的问题。

2）LCEL创新性地采用对决网络估计长期潜在结果，其独特的网络结构能够有效捕获不同处理所带来累积结果的差异，从而大幅提升算法对弱因果效应的识别能力，解决了现有长期因果效应估计识别弱因果效应准确率低的问题。

3）实验结果表明，LCEL算法在合成数据集和模拟订单调度的数据集上的表现均优于现有的长期因果效应估计方法。

1 相关工作

静态因果效应计算是指在静态场景中估计处理变量对结

果变量的因果影响。现有的工作可分为以下几类。

1)基于匹配的方法。这类方法将处理组和对照组中的个体进行匹配，来估计处理对结果的因果效应。例如，倾向得分匹配方法[11]通过构建倾向得分，并按照倾向得分将处理组和对照组进行匹配，进一步估计因果效应；最近邻匹配方法[12]将特征距离最近的处理组个体与对照组个体进行匹配，从而估计因果效应。

2)重加权方法。该方法为每个个体分配一个新的权重，构造一个处理组与对照组数据分布相似的伪总体。例如，逆概率加权方法[13]使用个体接受处理的概率的倒数作为权重，以此来平衡处理组和对照组之间的差异；双鲁棒估计方法[14]结合倾向得分模型和结果回归模型的估计结果(倾向得分是个体接受处理的概率，结果回归则是对结果变量和协变量之间关系的建模)，计算因果效应的估计值。

3)基于深度学习方法。这类方法建立深度神经网络模型来学习样本的特征表示或解耦去除混杂因素。例如，CFR[15]首先学习处理组和对照组的平衡表示，然后建立回归模型来估计反事实结果，最终估计出个体因果效应；CEVAE[16]结合变分自编码器，近似恢复潜在混杂因素与观测数据的联合分布，从而实现对个体和群体因果效应的估计。此外，基于深度学习的经典方法还有 SITE[17]，GANITE[18]，ESCFR[19]等。尽管研究者提出了多种静态因果效应的方法，但其无法处理随时间变化的因果关系。

与静态因果效应估计不同，时间序列因果效应估计关注因果关系随时间的动态变化。研究者最早在流行病学中引入了随时间变化的结果估计方法，并且广泛采用了简单的线性模型进行分析，例如边缘结构模型和结构嵌套模型[20]等。为了突破线性模型表达能力的局限性，研究者提出了几种贝叶斯非参数方法[21]和基于循环神经网络[22]的方法。然而，这些方法大多局限于单次治疗，导致其适用范围受限。近期，为应对时变混杂因素引起的偏差，研究者提出了 RMSN[23]，该方法通过建模时序边缘结构控制时变混杂，有效解决了连续多次处理场景下的因果效应估计问题；CRN[24]使用领域对抗训练来构建患者病史的平衡表示，并在每个时间步构建平衡不变的表示，消除个体与处理分配之间的关联；G-Net[25]的目标是同时预测结果和时变协变量，然后执行 G 计算以进行多步预测。这些方法都是建立在 LSTM 编解码器上，在捕获时变混杂因素之间的复杂依赖性方面的能力有限。为了弥补这一缺陷，CT[26]结合 Transformer 对时变协变量、历史信息以及处理结果交叉关注，得到平衡表示，在时间序列中估计多步的反事实结果。时间序列因果效应的主要目标是研究随着时间推移，结果变量随处理变量如何动态变化，大多数用于对反事实结果的预测。然而，其无法直接计算长期因果效应，无法从长远的视角评估处理变量对结果变量的累积影响。

本文主要关注长期因果效应估计，旨在评估一个处理变量在长时间范围内对结果变量的累积影响。最近，Shi 等[9]提出了使用强化学习估计长期因果效应的方法。该方法使用价值学习的原理，利用线性基函数拟合长期潜在结果，然后推导出平均因果效应的分布，最后通过假设检验判断是否存在长期因果效应。相比于之前的方法，此方法可以用于计算处理变量对结果变量的累积影响，为在潜在结果模型下计算长期因果效应提供了一种途径。但是，该方法是使用基函数拟合长期潜在结果，这导致在实际应用时基函数的选取比较复杂，同时在复杂数据中效果较差。为了解决这些问题，本文将引入深度强化学习计算长期因果效应的方法。

2 问题定义

假设 $a=1$ 表示接受处理，$a=0$ 表示不接受处理，$t=0,1,2,\cdots,T$ 表示时间点，$k=1,2,3,\cdots,N$ 表示个体。对于个体 k，t 时刻的时变协变量集合 $\boldsymbol{X}_t^{(k)}\in R^d$，$k$ 在 t 时刻接受的处理 $a_t^{(k)}\in\{0,1\}$，接受处理后的实时结果 $y_t^{(k)}\in R$。此外，使用 $Y_t(a)$ 表示在 t 时刻处理 a 的长期潜在结果，使用 $\boldsymbol{Y}_t=(Y_t(0),Y_t(1))^{\mathrm{T}}$ 表示在 t 时刻不接受处理和接受处理的长期潜在结果组成的向量。本文中长期因果效应涉及到的重要符号的定义如表 1 所列。

表 1 长期因果效应的重要符号定义

Table 1 Important symbolic definitions for long-term causal effects

符号	表示
t	时间点
k	个体
$\boldsymbol{X}_t$	个体在 t 时刻的协变量集
R^d	d 维实数向量集
a_t	t 时刻的处理
y_t	t 时刻的实时结果
$Y_t(a),\boldsymbol{Y}_{t,a}$	t 时刻处理 a 的长期潜在结果
$\boldsymbol{Y}_t$	t 刻所有处理的长期潜在结果的向量表示

本文将处理在未来一段时间内所带来的累积结果作为当前时刻该处理的长期潜在结果。如图 3 所示(其中黄色部分是在 t 时刻可以根据现有知识可得到的，蓝色部分是在 t 时刻未知的数据)，理论上 t 时刻接受处理或不接受处理的长期潜在结果表示为 $Y_t(a)=y_t+y_{t+1}+y_{t+2}+\cdots+y_{t+i}+\cdots$，然而在 t 时刻只能得到该式中的 y_t，而无法观测 y_{t+1} 及之后的数据，这导致未来的累积结果无法计算，因此传统方法在计算长期潜在结果时面临挑战。强化学习为解决这一问题提供了一种有效的工具。在强化学习中，价值函数被用于评估当前时刻的动作能带来的累积收获，其计算目标同样为累积值，与估计长期潜在结果的目标一致，因此本文将长期潜在结果的估计目标转化为价值函数的形式，如式(1)所示：

$$Y(a_t;\boldsymbol{X}_t)=\max\left\{\sum_{i=0}^{T}\gamma E[y_{t+i}\mid(a_t,\boldsymbol{X}_t)]\right\}\tag{1}$$

其中，γ 表示折扣因子；$a_t=1$ 表示 t 时刻接受处理，$a_t=0$ 表示 t 时刻不接受处理；$Y(1;\boldsymbol{X}_t)$ 和 $Y(0;\boldsymbol{X}_t)$ 分别表示个体在 t 时刻接受处理和不接受处理的长期潜在结果。

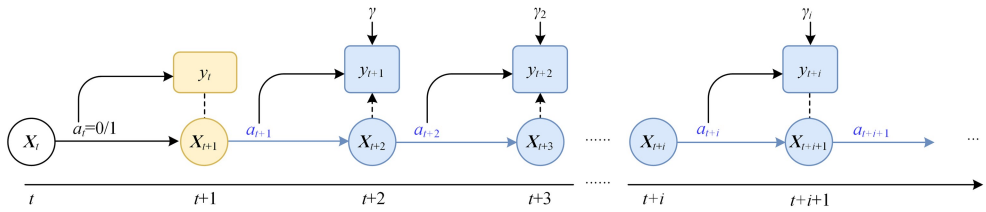


图 3 个体在 t 时刻接受处理或不接受处理后在未来每个时间点产生的实时结果以及时变协变量(电子版为彩图)

Fig. 3 Real-time results and time-varying covariates generated at each time point as time passes after an individual receives treatment or does not receive treatment at t time

为了保证上述长期潜在结果可预测,从而达成识别长期因果效应的目的,在满足因果效应计算问题中的基本假设即正值性假设、可忽略性假设等[27]的前提下,下面给出在长期因果效应估计问题中的假设条件。

假设 1 马尔可夫假设:存在一个马尔可夫转移过程 F,在任意的 $t\geqslant 0$ 时有 $\boldsymbol{X}_{t+1}=F(a_t,\boldsymbol{X}_t)$。该假设表明,下一个时刻的协变量 $\boldsymbol{X}_{t+1}$ 仅取决于上一时刻的协变量 $\boldsymbol{X}_t$ 和上一时刻的处理 a_t,而与更早的历史数据(如 $a_{t-1},\boldsymbol{X}_{t-1},\cdots,a_0,\boldsymbol{X}_0$)无关(见图 3)。

假设 2 条件平均独立性假设:存在一个函数 r,在任意的 $t\geqslant 0$ 时,都有 $E\{y_t\mid a_t,\boldsymbol{X}_t,a_{t-1},\boldsymbol{X}_{t-1},\cdots,a_0,\boldsymbol{X}_0\}=r(a_t,\boldsymbol{X}_t)$,在每个时刻的实时结果由该时刻的时变协变量和该时刻的处理决定。如图 3 所示,该假设表明,历史信息对当前结果的条件期望没有额外影响,即历史信息在给定 $(a_t,\boldsymbol{X}_t)$ 的条件下是冗余的,表示如下:

$$\begin{aligned}&E[y_t\mid a_t,\boldsymbol{X}_t,a_{t-1},\boldsymbol{X}_{t-1},\cdots,a_0,\boldsymbol{X}_0]\\&=E[y_t\mid a_t,\boldsymbol{X}_t,history]\\&=E[y_t\mid a_t,\boldsymbol{X}_t]\\&=r(a_t,\boldsymbol{X}_t)\end{aligned}\tag{2}$$

为了定量反映处理对个体或群体的长期因果效应,根据前面提出的长期潜在结果 $Y(a;\boldsymbol{X}_t)$ 以及因果可识别性条件定义了条件长期因果效应(Conditional Long-Term Treatment Effect, CLTE)、个体长期因果效应(Individual Long-Term Treatment Effect, ILTE)与平均长期因果效应(Average Long-Term Treatment Effect, ALTE)。

定义 1(条件长期因果效应) 个体 k 在 t 时刻的时变协变量下,接受处理的长期潜在结果 $(Y^{(k)}(1,\boldsymbol{X}_t))$ 与不接受处理的长期潜在结果 $(Y^{(k)}(0,\boldsymbol{X}_t))$ 的差异是此个体在此时变协变量下的长期因果效应,表达式为:

$$CLTE_k(\boldsymbol{X}_t)=Y^{(k)}(1,\boldsymbol{X}_t)-Y^{(k)}(0,\boldsymbol{X}_t)\tag{3}$$

定义 2(个体的长期因果效应) 单个个体 k 从 0 时刻到 T 时刻中在不同条件下条件长期因果效应的均值,表达式为:

$$ILTE_k=\frac{1}{T}\sum_{t=0}^{T-1}CLTE_k(\boldsymbol{X}_t)\tag{4}$$

定义 3(平均长期因果效应) 在整个群体中,所有个体(总数为 N)的长期因果效应的均值是该群体的平均长期因果效应,表达式为:

$$ALTE=\frac{1}{N}\sum_{k=1}^{N}ILTE_k=\frac{1}{N}\frac{1}{T}\sum_{k=1}^{N}\sum_{t=0}^{T-1}CLTE_k(\boldsymbol{X}_t)\tag{5}$$

3 LCEL 算法

基于第 2 章的假设和定义,本章提出了 LCEL 算法,其流程如图 4 所示。

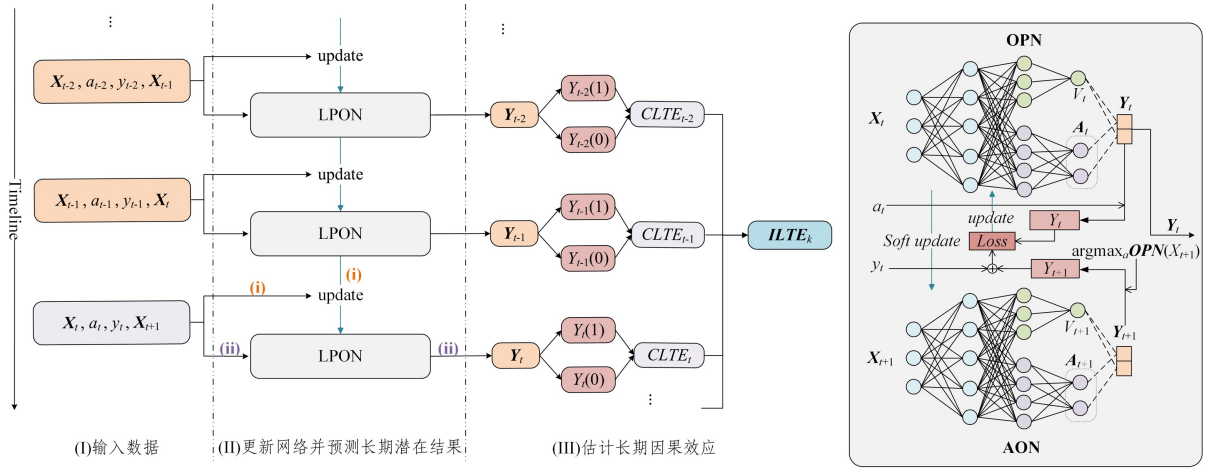


图 4 LCEL 算法流程图

Fig. 4 Flowchart of LCEL algorithm

如图 4(a)所示,LCEL 在每个时刻(以 t 时刻为例)的工作流程包括:1)输入 t 时刻的观测数据;2)使用输入数据更新 LCEL 的核心网络——长期潜在结果网络(Long-term Potential Outcome Network, LPON);3)通过 LPON 预测长期潜在结果;4)依据长期潜在结果估计该时刻的长期因果效应。在每个时刻重复上述过程(伪代码见算法 1)。

算法 1 LCEL 算法

输入：包含 N 个个体在 T 个时刻内的 N * T 条四元组数据$\{\mathbf{X}_t^{(k)}, a_t^{(k)}, y_t^{(k)}, \mathbf{X}_{t+1}^{(k)}\}$的数据集 D，经验回放数组，$\gamma$，$\tau$

输出：长期因果效应 ALTE，以及 N 个个体的 ILTE 数组

```
//计算 ALTE
for k=1:N do
    //计算 ILTE
    for t=0:T-1 do
    //计算 CLTE
    //优化 LPON
    计算该四元组的权重w_i，存入经验回放数组，并从中抽取 b 条四元组
    Step 1. Y_t(a)=f_OPN(X_t, a_t; θ)
    Step 2. a'_{t+1}=argmax_a f_OPN(X_{t+1}; θ)
    Y_{t+1}=Y_{t+1}(a'_{t+1})=f_AON(X_{t+1}, a'_{t+1}; θ')
      Step 3.
      δ(θ)_t=y_t+γ f_AON(X_{t+1}, a'_{t+1}; θ')−f_OPN(X_t, a_t; θ)
      Step 4. L(θ)=(1/b) Σ_{i=1}^{b} w_i · δ(θ)_t^2
      更新网络
      //计算该时变协变量下所有处理的长期潜在结果
      Y_t^{(k)}=f_OPN(X_t^{(k)})
      //计算该时刻的 CLT E_t^{(k)}
      CLTE(X_t)^{(k)}=Y(1; X_t^{(k)})−Y(0; X_t^{(k)})
      end for
    ILTE^{(k)}=(1/T) Σ_{t=0}^{T-1} CLTE(X_t)^{(k)}
    ILTE^{(k)} 存放到数组 ILTES 中
end for
ALTE=(1/N) Σ_{k=1}^{N} ILT E^{(k)}
return ALTE, ILTES
```

LCEL 在每个时间点（以 t 时刻为例）输入一条四元组数据$\{\boldsymbol{X}_t, a_t, y_t, \boldsymbol{X}_{t+1}\}$，其中，$\boldsymbol{X}_t$表示个体在 t 时刻的时变协变量集合，a_t表示个体在 t 时刻接受的处理，y_t表示 t 时刻的实时结果，$\boldsymbol{X}_{t+1}$表示接受处理a_t后该个体新的时变协变量集合。

LCEL 利用输入的数据，更新 LPON 并预测个体 k 在该时刻接受处理和不接受处理的长期潜在结果，进一步估计长期因果效应。

3.1 LPON

LPON 旨在使用输入数据预测个体在该时刻的时变协变量下接受处理与不接受处理的长期潜在结果。如图 4(b)所示，该网络分为两部分，分别是结果预测网络（Outcome Prediction Network，OPN）与辅助优化网络（Assisted Optimization Network，AON）。

3.1.1 OPN

OPN 完成 LPON 的核心任务：估计在时变协变量$\boldsymbol{X}_t$下接受处理和不接受处理的长期潜在结果。OPN 输入时变协变量集合$\boldsymbol{X}_t$，输出该时变协变量下接受处理与不接受处理的长期潜在结果$\boldsymbol{Y}_t$。

OPN 的网络结构采用强化学习中常用的对决网络[28]。这一设计有以下几个原因：首先，使用对决网络无须借助基函数拟合长期潜在结果；其次，对决网络能够适应复杂的时变协变量环境，数据维度上升时性能下降较少；最后，对决网络的特殊结构使其对弱因果效应具有高度敏感性，能够准确识别弱因果效应。如图 4(b)所示，OPN 有两条输出分支：第一条分支用于评估当前时变协变量$\boldsymbol{X}_t$产生的累积结果，称为 V 分支，其输出为V_t；第二条分支同时估计接受处理和不接受处理相对于当前时变协变量$\boldsymbol{X}_t$产生的累积结果，称为 A 分支，输出为$\boldsymbol{A}_t$。这种对处理和协变量产生的累积结果进行分解的方式，使得对决网络能够更细致地理解协变量和处理对未来结果的长期累积影响。因此，利用对决网络的优势，OPN 能够有效地评估处理变量带来的长期潜在结果[28]，从而真实地计算因果效应。V_t与$\boldsymbol{A}_t$的具体计算方式为：

$$\begin{cases} V_t = DNv(\boldsymbol{X}_t;\theta) \\ \boldsymbol{A}_t = DNa(\boldsymbol{X}_t;\theta) \end{cases} \tag{6}$$

其中，DNv表示对决网络的 V 分支，DNa表示对决网络的 A 分支，θ表示 OPN 的参数。

长期潜在结果由V_t和$\boldsymbol{A}_t$两部分组成。为了防止估计结果出现偏差，最终的长期潜在结果是V_t与$\boldsymbol{A}_t$的和减去接受处理和不接受处理产生的累积结果的均值。这种方式可以更精确地估计出长期因果效应，计算式如式(7)所示：

$$Y_t(a) = \boldsymbol{Y}_{t,a_t} = V_t + (\boldsymbol{A}_{t,a_t} - \bar{\boldsymbol{A}}_t) \tag{7}$$

然而，在实际应用中，研究者需要训练 OPN 并更新参数，以得到$\boldsymbol{Y}_t = f(\boldsymbol{X}_t;\theta)$（$f$ 即 OPN）的映射关系，但此时无法获取$\boldsymbol{Y}_t$的真实值，因此无法通过传统的监督学习方法对 OPN 进行训练和优化。为了解决这一问题，本文使用了时间差分（Time Difference，TD）算法[10]。TD 算法的思想是使用 TD 目标（即$y_t + \gamma Y_{t+1}$）作为Y_t的真实值来监督 OPN 更新。

根据最优贝尔曼方程[29]可知，长期潜在结果$Y_t = y_t + \max(\gamma y_{t+1} + \gamma^2 y_{t+2} + \cdots)$，在时间 t 只能观测到y_t，而$\max(\gamma y_{t+1} + \gamma^2 y_{t+2} + \cdots)$未知但是等价于$\gamma Y_{t+1}$。因此，本文使用辅助优化网络预测$\max(y_{t+1} + \gamma y_{t+2} + \cdots)$，即$Y_{t+1}$，由此可以得到 TD 目标，即$y_t + \gamma Y_{t+1}$。这种方法可以在没有真实值的情况下通过估计下一时刻的长期潜在结果来达成训练 OPN 的目的。

3.1.2 AON

AON 结构和工作原理与 OPN 相同，用于在 t 时刻根据$\boldsymbol{X}_{t+1}$通过正向传播来估计长期潜在结果Y_{t+1}，并结合观测到的实时结果y_t组合得到 TD 目标。

$$\begin{cases} V_{t+1} = DNv(\boldsymbol{X}_{t+1};\theta') \\ \boldsymbol{A}_{t+1} = DNa(\boldsymbol{X}_{t+1};\theta') \end{cases} \tag{8}$$

$$\begin{aligned} Y_{t+1}(a) &= \boldsymbol{Y}_{t+1,a_{t+1}} \\ &= V_{t+1} + (\boldsymbol{A}_{t+1,a_{t+1}} - \bar{\boldsymbol{A}}_{t+1}) \end{aligned} \tag{9}$$

其中，θ'是 AON 的参数。根据贝尔曼方程[29]，可以得出 OPN 的更新目标：$Y_t(a) = y_t + \gamma Y_{t+1}$。其中，$Y_t(a)$表示在 t 时刻处理 a 的长期潜在结果；y_t是已知的实时观测结果；γ是折扣因子；Y_{t+1}是在时变协变量$\boldsymbol{X}_{t+1}$下所有处理的长期潜在结果的最大值。

3.1.3 优化 LPON

根据前面对 OPN 和 AON 的介绍，LPON 每个时刻的迭代都需要依次计算出$Y_t(a)$和Y_{t+1}，然后结合输入的观测

数据，根据更新目标计算出 TD 误差和损失函数并对网络进行更新。更新的过程可以分为以下 4 个步骤。

Step 1 计算$\boldsymbol{Y}_t(a)$。

使用 OPN 估计 t 时刻时变协变量$\boldsymbol{X}_t$下处理a_t的长期潜在结果：

$$Y_t(a)=f_{\mathrm{OPN}}(\boldsymbol{X}_t,a_t;\theta) \tag{10}$$

Step 2 计算$\boldsymbol{Y}_{t+1}$。

由于此时的 OPN 是参数较新的网络，因此首先使用 OPN 估计出 $t+1$ 时刻所有处理的长期潜在结果，从而进一步判断出 $t+1$ 时刻的最优处理。然后用 AON 来估计在$t+1$时刻最优处理的长期潜在结果，从而得到Y_{t+1}。这个过程分为两步。

1）OPN 计算 $t+1$ 时刻的最优处理方式。OPN 网络根据 t 时刻输入的时变协变量$\boldsymbol{X}_{t+1}$，估计时变协变量$\boldsymbol{X}_{t+1}$下接受处理和不接受处理的长期潜在结果。通过该估计结果，得到 $t+1$时刻时变协变量$\boldsymbol{X}_{t+1}$下的最优处理方式a'_{t+1}。

$$a'_{t+1}=\underset{a}{\operatorname{argmax}}\, f_{\mathrm{OPN}}(\boldsymbol{X}_{t+1};\theta) \tag{11}$$

2）AON 估计 $t+1$ 时刻最优处理方式的长期潜在结果。AON 根据 OPN 确定的最优处理方式，进一步估计最优处理a'_{t+1}的长期潜在结果，并计算出Y_{t+1}：

$$Y_{t+1}=Y_{t+1}(a'_{t+1})=f_{\mathrm{AON}}(\boldsymbol{X}_{t+1},a'_{t+1};\theta') \tag{12}$$

其中，θ'表示辅助优化网络的参数。

Step 3 计算 TD 误差。

由贝尔曼方程[29]可知，理论上$Y_t(a)=y_t+\gamma Y_{t+1}$，由此可得结果预测网络的更新目标为：

$$f_{\mathrm{OPN}}(\boldsymbol{X}_t,a_t;\theta)=y_t+\gamma f_{\mathrm{AON}}(\boldsymbol{X}_{t+1},a'_{t+1};\theta') \tag{13}$$

根据更新目标计算出 TD 误差：

$$\delta(\theta)_t=y_t+\gamma f_{\mathrm{AON}}(\boldsymbol{X}_{t+1},a'_{t+1};\theta')-f_{\mathrm{OPN}}(\boldsymbol{X}_t,a_t;\theta) \tag{14}$$

Step 4 更新 LPON。

此时已经计算出一条四元组数据的 TD 误差，但每个时刻仅靠一条数据的误差更新 LPON，会导致 LPON 对环境的变化感知并不高效可靠，计算出的 CLTE 不能良好地反映真实的情况。因此，在每个时刻将观测到的四元组数据存入经验回放数组中，每次从经验回放数组中抽取多条四元组数据，计算多个 TD 误差，用于进一步计算损失函数并更新 OPN。

在经验回放数组中，根据 TD 误差δ_i给每条四元组数据赋予一个权重w_i，该权重用来衡量每条数据对 OPN 更新的重要性，表示为：

$$w_i=\frac{(M\cdot P(i))^{-\beta}}{\max_j(M\cdot P(j))^{-\beta}} \tag{15}$$

其中，β是重要性采样权重的指数修正参数，其在更新过程中逐渐增大；M 是数组中四元组样本的数量；优先级概率 $P(i)=p_i^\alpha/(\sum_j p_j^\alpha)$，在 $P(i)$中，α 是控制优先级影响的超参数，通常在 0 到 1 之间，$\alpha=1$时优先级影响最大；p_i是第 i 个数据的优先级，由 TD 误差决定，本文在训练时更加关注那些 TD 误差较大的经验，因此取$p_i=|\delta_i|+\epsilon$，ϵ是非常小的正数，防止第 i 条数据优先级为 0。在经验回放数组中，每个时间步都会有一条新的数据存入，并按照权重抽取 b 条数据进行批量训练。

计算 b 条数据的 TD 误差 δ 后对 OPN 进行更新，其损失函数表示为：

$$\mathcal{L}(\theta)=\frac{1}{b}\sum_{i=1}^{b}w_i\cdot\delta(\theta)_i^2 \tag{16}$$

因此可以更新 OPN 的参数，同时使用结果预测网络的参数对 AON 进行软更新，表达式如下：

$$\theta'\leftarrow\tau\theta+(1-\tau)\theta' \tag{17}$$

其中，$\tau\in(0,1)$是一个超参数，表示每次更新 AON 参数的比例，θ 和 θ'分别表示 OPN 和 AON 的参数。

3.2 估计长期因果效应

在 LPON 更新之后，使用其估计当前时变协变量下接受处理和不接受处理的长期潜在结果，其原理如式(6)和式(7)所示。

整体来看，利用 LPON 对该时变协变量信息($\boldsymbol{X}_t$)进行处理，估计出的长期潜在结果($\boldsymbol{Y}_t$)表示为：

$$\boldsymbol{Y}_t^{(k)}=f_{\mathrm{LPON}}(\boldsymbol{X}_t^{(k)}) \tag{18}$$

$$Y(0)_t^{(k)}=\boldsymbol{Y}_{t,0}^{(k)},Y(1)_t^{(k)}=\boldsymbol{Y}_{t,1}^{(k)} \tag{19}$$

在计算出 t 时刻接受处理和不接受处理的长期潜在结果之后，可以计算出个体在每个时变协变量下的条件平均因果效应、个体在所有时变协变量下的个体长期因果效应，以及在整个场景中所有个体的平均长期因果效应。

首先，根据计算得到的长期潜在结果，计算在该时变协变量下的条件长期因果效应(CLTE)。方式如下：

$$CLTE(\boldsymbol{X}_t)^{(k)}=\boldsymbol{Y}_{t,1}^{(k)}-\boldsymbol{Y}_{t,0}^{(k)} \tag{20}$$

其次，在可观察的时间内计算处理变量对个体 k 结果变量的个体长期因果效应，判断处理变量是否对该个体的结果变量产生累积影响。具体公式为：

$$\begin{aligned}ILTE^{(k)}&=\frac{1}{T}\sum_{t=0}^{T-1}CLTE(\boldsymbol{X}_t)^{(k)}\\&=\frac{1}{T}\sum_{t=0}^{T-1}[\boldsymbol{Y}_{t,1}^{(k)}-\boldsymbol{Y}_{t,0}^{(k)}]\end{aligned} \tag{21}$$

其中，k 为个体的序号，T 为所有观测到的总时间长度。

最后，在已经计算的个体长期因果效应基础上对整个场景中处理对结果的长期因果效应进行估计，用于评估整个群体中处理变量对结果变量的长期因果效应。

$$\begin{aligned}ALTE&=\frac{1}{N}\sum_{k=1}^{N}ILTE^{(k)}\\&=\frac{1}{N}\frac{1}{T}\sum_{k=1}^{N}\sum_{t=0}^{T-1}CLTE^{(k)}(\boldsymbol{X}_t)\end{aligned} \tag{22}$$

其中，N 表示观察的个体总数。

4 实验与分析

为验证 LCEL 方法的有效性，在合成数据集和模拟订单调度数据集上进行了实验。为了有效比较算法的性能，需要选取通过长期累积结果估计长期因果效应的方法进行对比，本文选取目前最优的 CausalRL[9]作为对比算法。CausalRL 方法通过线性基函数拟合长期潜在结果，并推导出长期因果效应及其分布，然后通过观察数据进行假设检验，判断处理变

量与结果变量之间是否存在长期因果关系。

4.1 合成数据集实验

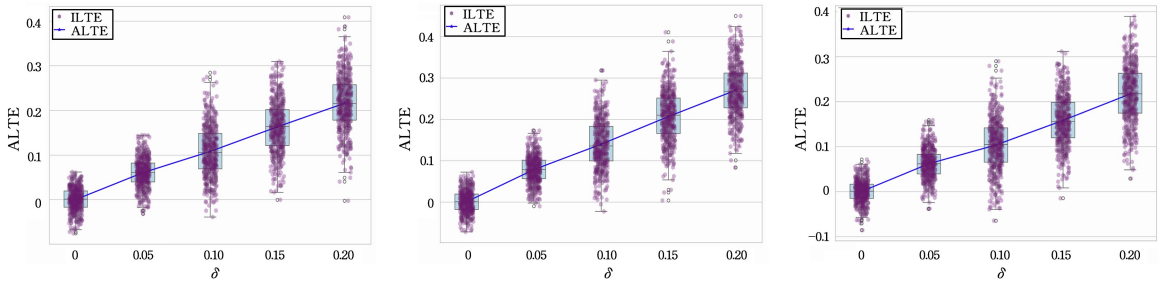
在合成数据集上的实验分为两部分：第一部分旨在验证算法对长期因果效应的敏感性；第二部分实验是在弱因果效应数据和较高维度的数据上将 LCEL 算法与对比算法在识别因果效应的准确性方面进行对比。

4.1.1 合成数据集

本小节采用的合成数据集取自文献[9]，此数据集中存在若干组数据，每组数据有 N 个个体在 T 个时间步内依据个体的时变协变量信息，使用不同的行为策略选取处理方式(0 或 1)获得实时结果，并转移到下一个时变协变量。在该数据集中，使用一个参数 δ 调控处理变量对结果变量真实的因果效应大小，参数 δ 与因果效应的具体关系表示为：$\delta=0$ 时，$ALTE=0$；$\delta>0$ 时，$ALTE>0$ 且 $ALTE$ 的大小随 δ 的增加而增加。在本次实验中，取 δ 分别为{0，0.05，0.1，0.15，0.2}的 5 种场景，令 $N=500$，$T=600$。

4.1.2 LCEL 对长期因果效应的敏感性实验与分析

使用文献[9]中提供的 3 种行为策略作为在每个时间步个体接受处理方式(0 或 1)的依据。使用 LCEL 算法依次计算出 CLTE，ILTE 和 ALTE。实验结果如图 5 所示。



(a)交替策略 (b)随机策略 (c)ε-贪婪策略

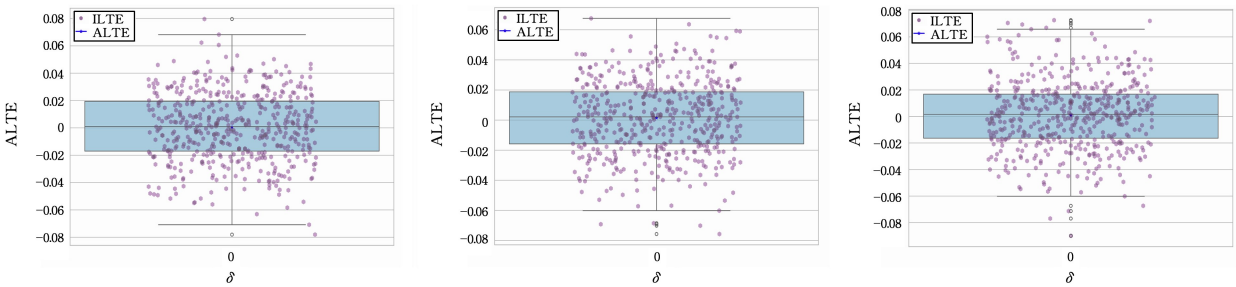
图 5 不同行为策略下的 ILTE 与 ALTE 结果(电子版为彩图)

Fig. 5 Results of ILTE and ALTE under different behavioral strategies

图 5 中的每个紫色点代表一个 ILTE 值，本文使用箱线图和散点图对其分布进行了可视化表示，并使用蓝色折线表示平均长期因果效应(ALTE)随参数 δ 变化的情况。实验结果表明，所提方法估计的平均因果效应的变化始终与真实的因果效应强弱程度(该强弱程度由数据集中的参数 δ 调控)保持一致。因此，LCEL 算法对长期因果效应具有较强的敏感性，可以有效地估计长期因果效应。

由于每个个体的时变协变量信息存在差异，因此计算出的个体长期因果效应在一个可接受范围内波动。在 $\delta=0$ 的场景下，处理变量对结果变量不存在长期因果效应，而在 $\delta=0.05$ 时，存在长期因果效应。但从图 5 所示实验结果的分布情况可以看出，在 $\delta=0$ 和 $\delta=0.05$ 这两种情况下，LCEL 算法计算的 ILTE 的数值范围出现了重叠。例如在 $\delta=0$ 时，LCEL 计算的 $ILTE\in(-0.06,0.06)$，在 $\delta=0.05$ 时，LCEL 计算的 $ILTE\in(-0.01,0.156)$，在这两个场景中计算出的 ILTE 存在(−0.01，0.06)的重叠，因此在此重叠区间的 ILTE 无法判定是否存在长期因果效应。为了解决该问题，本文通过实验确定无因果效应与有因果效应的边界。本文的目标是尽可能保证无因果效应的准确率大于 95%，从而在重叠部分的判断上更加严格。

为了找到这个边界，本文在 $\delta=0$ 时进行了多次实验。实验结果如图 6 所示，经过异常值剔除后，结果的波动在 0.05 以内。因此，本文选取 0.05 作为判断是否存在因果效应的边界。如果 $ILTE>0.05$，则认为处理变量对结果变量存在长期因果效应；否则，认为处理变量对结果变量不存在长期因果效应。



(a)交替策略 (b)随机策略 (c)ε-贪婪策略

图 6 不同的行为策略下在 $\delta=0$ 时的 ILTE 的分布

Fig. 6 Distribution of ILTE under different behavioral strategies when $\delta=0$

4.1.3 LCEL 与 CausalRL 的对比实验

1)弱因果效应识别准确率的对比

在 $\delta=0.05$ 和 $\delta=0.1$ 的数据中，处理变量对结果变量的累积影响较弱。为了证明 LCEL 方法在识别弱因果效应时优于现有方法，使用与 4.1.2 节相同的实验流程进行了对比实验。实验统计了识别个体长期因果效应的准确率，结果如表 2 所列。相比现有方法，LCEL 方法在保证识别无因果效应($\delta=0$)和较强因果效应($\delta=0.15$ 和 $\delta=0.2$)的准确率不下

降的前提下，能够在弱因果效应的场景（$\delta=0.05$和$\delta=0.1$）中取得更高的识别准确率。这表明，LCEL 在应对弱因果效应问题时优于现有方法。

表 2 在合成数据集上对因果效应识别的准确率

Table 2 Accuracy rate of causal effect identification on the synthetic dataset

Policy	algorithm	δ				
		0	0.05	0.1	0.15	0.2
Alternating	CausalRL	0.944	0.426	0.888	0.990	0.988
	LCEL	**0.980**	**0.670**	**0.938**	**1.000**	**1.000**
Random	CausalRL	**0.962**	0.486	0.902	**1.000**	**1.000**
	LCEL	**0.962**	**0.816**	**0.926**	0.994	**1.000**
ε-greed	CausalRL	0.950	0.392	0.860	**0.986**	**1.000**
	LCEL	**0.960**	**0.628**	**0.876**	0.976	**1.000**

2）更高维度数据上的实验

本文将数据集中的时变协变量从二维扩展到四维，数据集复杂度因此增加。在扩展后的数据集上进行对比实验，实验设置与之前的实验一致（结果如表 3 所列）。相比于低维数据，LCEL 方法在维度上升时的准确率几乎没有下降；而 CausalRL 在维度上升时性能严重下降，在多数存在因果效应的数据中会错误地判断为没有因果效应。

表 3 在四维时变协变量数据中的对比实验结果

Table 3 Results of comparative experiments in four-dimensional time-varying covariate data

Policy	algorithm	δ				
		0	0.05	0.1	0.15	0.2
Alternating	CausalRL	0.916	0.102	0.140	0.172	0.206
	LCEL	**0.992**	**0.654**	**0.868**	**0.988**	**0.996**
Random	CausalRL	0.966	0.068	0.086	0.090	0.104
	LCEL	**0.988**	**0.646**	**0.866**	**0.976**	**1.000**
ε-greed	CausalRL	0.978	0.064	0.092	0.152	0.228
	LCEL	**0.982**	**0.838**	**0.946**	**0.992**	**1.000**

4.2 订单调度模拟实验与分析

该数据集用于模拟网约车接单与送客的交互数据。在该实验中，借助模拟器可以恢复到t_0时刻的优势，在不同的场景进行了随机对照试验，对两种处理方式（1 是更新订单调度策略，0 是维持原订单调度策略）进行模拟，用于计算在对应场景下更新调度策略和不更新调度策略时司机的平均收入。

随机对照试验的结果为：$M=Income(1)-Income(0)$，其中 $Income(1)$表示 1000 个时刻之内 250 个司机接受处理的平均利润，$Income(0)$表示 1000 个时刻之内 250 个司机不接受处理的平均利润。

在计算长期因果效应时，将在每个时刻使用随机的方式进行处理 0 和 1 的分配，本次实验将每一时刻司机的位置作为其此刻的时变协变量信息，将 t 时刻订单调度的方式作为在 t 时刻的处理，t 时刻的收益作为实时奖励，下一时刻的位置作为新的时变协变量，按照这种方式延续下去。本文选择 250 个司机观察 1000 个时刻，其中前 500 时间步收集到的信息作为历史信息，用作对长期潜在结果网络的训练，后 500 个时间步在实时更新长期潜在结果网络的同时计算每个时变协变量下的长期因果效应。

该实验随机选取 6 个不同的场景（乘客分布不同），首先进行随机对照试验，结果如表 4 所列。在前 3 个场景中，随机对照试验的结果 M 都是负值，表明在前 3 个场景中司机更新订单调度策略会降低收益，所以更新订单调度对提升司机收益没有因果效应；相反，在后 3 个场景中随机对照试验的结果为正，则表明更新订单调度对提升司机收益有因果效应。

表 4 6 个场景中的随机对照试验结果以及算法的结果

Table 4 Results of randomized controlled trials and algorithmic outcomes in six scenarios

	1	2	3	4	5	6
随机对照试验结果 M	−313.3	−232.6	−123.5	37.3	140.2	470.0
是否存在长期因果效应	否	否	否	是	是	是
ALTE(Ours)	−0.107	−0.09	−0.045	0.02	0.05	0.254

从图 7 可以看出，所提方法计算出的因果效应值随着随机对照实验结果的增加单调增加，从而有效证明了所提方法对因果效应有敏感性。对比算法在第 4 个弱因果效应的场景中出现了错误，误识别为不存在因果效应，相比于对比算法，所提方法更加稳定。从图 8 展示的每个场景的 $ILTE>0$ 的比例结果中可以发现，LCEL 算法对个体因果效应的变化更敏感，能在复杂的场景中有效判断是否存在因果效应。

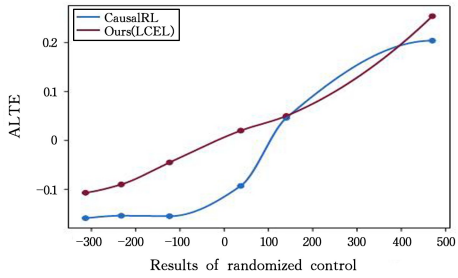


图 7 不同算法计算的 ALTE 随着真实收益的变化曲线

Fig. 7 Curve of ALTE computed by different algorithms as a function of the true reward

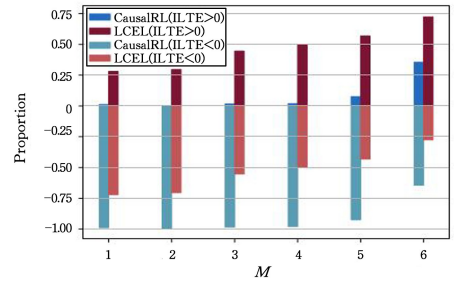


图 8 不同算法在不同场景中计算的 ILTE 的分布情况

Fig. 8 Distribution of ILTE computed by different algorithms across various scenarios

4.3 参数敏感性实验与分析

本节研究折扣因子 γ 以及参数 α 和 τ 的取值对算法性能的影响。在其他参数值保持不变且行为策略为随机策略，$N=500$，$T=600$ 时，在 $\delta=0$ 和 $\delta=0.05$ 的设定情况下的合成数据集上进行实验。

参数 $\gamma\in(0,1)$用于平衡实时结果与长期累积结果的权重，γ 接近 1 时模型更注重长期累积结果，本次实验中 γ 取值分别为{0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99}，结果如图 9 所示。γ 对算法的性能影响较大，且 $\gamma\in(0.8,0.95)$时算法性能

较优。γ 偏小时，模型可能会过度关注实时结果，导致模型偏离预期；γ 偏大时，模型会轻视实时结果，性能也会下降。

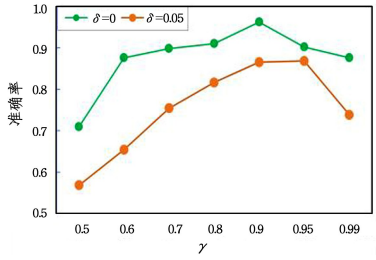


图 9 在合成数据集上 γ 取不同值时 LCEL 的准确率

Fig. 9 Accuracy of the LCEL with different γ values on a synthetic dataset

参数 α 是控制优先级影响的参数，本次实验中分别取 α 为[0.3，0.4，0.5，0.6，0.7，0.8]，结果如图 10 所示。$\alpha\in[0.5,0.6]$时，算法性能较优。α 过小，会导致算法对变化的场景不明显；过大，则会导致算法对环境变化敏感，从而导致性能下降。

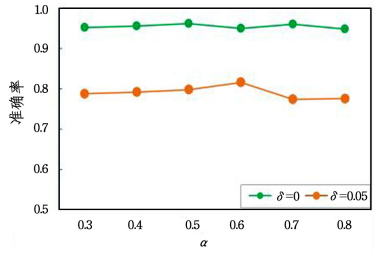


图 10 在合成数据集上 α 取不同值时 LCEL 的准确率

Fig. 10 Accuracy of the LCEL with different α values on a synthetic dataset

τ 表示每次软更新 AON 的参数比例，本次实验中 τ 的取值分别为[0.000 05，0.000 1，0.000 5，0.001，0.005]，结果如图 11 所示。$\tau\in[0.000\,5,0.001]$时效果较优。τ 取值过小，则 AON 更新过快，失去其稳定目标值估计的作用；过大，则导致 AON 更新过慢，不符合场景变化。

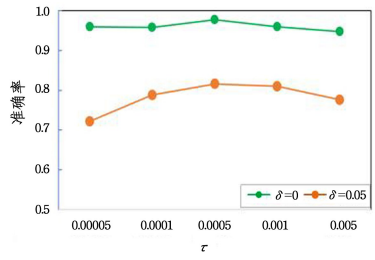


图 11 在合成数据集上 τ 取不同值时 LCEL 的准确率

Fig. 11 Accuracy of the LCEL with different τ values on a synthetic dataset

结束语 本文提出了使用深度强化学习估计长期因果效应的方法，使用价值学习的方法估计不同时变协变量下接受处理和不接受处理的长期潜在结果，并进一步计算长期因果效应。该方法弥补了现有方法的 3 个缺陷：1）现有方法依靠基函数拟合长期潜在结果，然而基函数选取不当会造成较大误差，本文使用的深度强化学习可以有效地解决基函数选取困难的问题，避免了因基函数选取不当造成的错误；2）现有方法在数据维度上升时性能严重下降，而本文提出的基于深度强化学习的方法在数据维度上升时也可以有效使用；3）现有方法在弱因果效应的识别上不准确，本文结合对决网络大幅提升了弱因果效应识别的准确率。在合成数据和模拟数据集上与对比算法进行的对比实验，验证了所提方法的有效性。接下来会探索如何使用状态价值函数描述长期潜在结果，同时结合策略学习提高算法学习各种条件下估计处理的长期潜在结果的能力，从而提高估计长期因果效应的准确度。

参 考 文 献

[1] KESSLER R C，BOSSARTE R M，LUEDTKE A，et al. Machine learning methods for developing precision treatment rules with observational data[J]. Behaviour Research and Therapy，2019，120：103412.

[2] ASSAEL H，ISHIHARA M，KIM B J. Accounting for causality when measuring sales lift from television advertising：Television campaigns are shown to be more effective for lighter brand users [J]. Journal of Advertising Research，2021，61(1)：3-11.

[3] SHALIT U. Can we learn individual-level treatment policies from clinical data? [J]. Biostatistics，2020，21(2)：359-362.

[4] RUBIN D B. Estimating causal effects of treatments in randomized and nonrandomized studies[J]. Journal of Educational Psychology，1974，66(5)：688.

[5] PEARL J，MACKENZIE D. The book of why：the new science of cause and effect[M]// Basic Books. 2018.

[6] WU A P，YUAN J K，KUANG K，et al. Learning decomposed representations for treatment effect estimation[J]. IEEE Transactions on Knowledge and Data Engineering，2022，35(5)：4989-5001.

[7] CAO D，ENOUEN J，WANG Y，et al. Estimating treatment effects from irregular time series observations with hidden confounders[C]// Proceedings of the 37th AAAI Conference on Artificial Intelligence. 2023：6897-6905.

[8] FOUGÈRE D，JACQUEMET N. Policy evaluation using causal inference methods[M]// Handbook of Research Methods and Applications in Empirical Microeconomics. Edward Elgar Publishing，2021：294-324.

[9] SHI C C，WANG X Y，LUO S K，et al. Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework[J]. Journal of the American Statistical Association，2023，118(543)：2059-2071.

[10] SUTTON R S，BARTO A G. Reinforcement learning：An introduction[M]// A Bradford Book. 2018.

[11] ROSENBAUM P R，RUBIN D B. The central role of the propensity score in observational studies for causal effects[J]. Biometrika，1983，70(1)：41-55.

[12] RUBIN D B. Matching to remove bias in observational studies [J]. Biometrics，1973，29(1)：159-183.

[13] HORVITZ D G，THOMPSON D J. A generalization of sampling without replacement from a finite universe[J]. Journal of the American Statistical Association，1952，47(260)：663-685.

[14] SCHARFSTEIN D O，ROTNITZKY A，ROBINS J M. Adjusting for nonignorable drop-out using semiparametric nonresponse models[J]. Journal of the American Statistical Associa-

tion, 1999, 94(448): 1096-1120.
[15] SHALIT U, JOHANSSON F D, SONTAG D. Estimating Individual Treatment Effect: Generalization Bounds and Algorithms [C]// Proceedings of the 34th International Conference on Machine Learning. 2017: 3076-3085.
[16] LOUIZOS C, SHALIT U, MOOIJ J M, et al. Causal Effect Inference with Deep Latent-Variable Models[C]// Proceedings of the 31st Conference on Neural Information Processing Systems. 2017: 6446-6456.
[17] YAO L Y, LI S, LI Y L, et al. Representation Learning for Treatment Effect Estimation from Observational Data[C]// Proceedings of the 32nd Conference on Neural Information Processing Systems. 2018: 2638-2648.
[18] YOON J, JORDON J, VANDERSCHAAR M. Ganite: Estimation of Individualized Treatment Effects Using Generative Adversarial Nets [C]// Proceedings of the 6th International Conference on Learning Representations. 2018: 50-60.
[19] WANG H, CHEN Z C, FAN J J, et al. Optimal transport for treatment effect estimation[C]// Proceedings of the 38th Conference on Neural Information Processing Systems. 2024: 1-21.
[20] ROBINS J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect[J]. Mathematical Modelling, 1986, 7(9/10/11/12): 1393-1512.
[21] XU Y, XU Y, SARIA S. A non-parametric bayesian approach for estimating treatment-response curves from sparse time series[C]// Proceedings of the 1st Machine Learning for Healthcare. 2016: 282-300.
[22] QIAN Z Z, ZHANG Y, BICA I, et al. Synctwin: Treatment effect estimation with longitudinal outcomes[C]// Proceedings of the 35th Conference on Neural Information Processing Systems. 2021: 3178-3190.
[23] LIM B, ALAA A, VAN DER SCHAAR M. Forecasting treatment responses over time using recurrent marginal structural networks[C]// Proceedings of the 32nd Conference on Neural Information Processing Systems. 2018: 7494-7504.
[24] BICA I, ALAA A M, JORDON J, et al. Estimating counterfactual treatment outcomes over time through adversarially balanced representations[J]. arXiv: 2002. 04083, 2020.
[25] LI R, HU S, LU M Y, et al. G-Net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime[C]// Proceedings of Machine Learning Research. 2021: 282-299.
[26] MELNYCHUK V, FRAUEN D, FEUERRIEGEL S. Causal transformer for estimating counterfactual outcomes[C]// Proceedings of the 39th International Conference on Machine Learning. 2022: 15293-15329.
[27] ROBINS J M. Optimal structural nested models for optimal sequential decisions[C]// Proceedings of the 2nd Seattle Symposium in Biostatistics: Analysis of Correlated Data. 2004: 189-326.
[28] WANG Z Y, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]// Proceedings of the 33th International Conference on Machine Learning. 2016: 1995-2003.
[29] BELLMANR. Dynamicprogramming[J]. Science, 1966, 153(3731): 34-37.

LIU Jiaqi, born in 2003, postgraduate. His main research interests include causal effect estimation and reinforcement learning.

YU Kui, born in 1977, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. 14259M). His main research interest is causal inference.

(责任编辑:何杨)