计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(4):235−245 [doi: 10.15888/j.cnki.csa.009470]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

反事实增强的对抗学习序列推荐①

刘珈麟[1,3], 贺泽宇[2], 李　俊[1]

[1](中国科学院 计算机网络信息中心, 北京 100083)
[2](北京信息科技大学 计算机学院, 北京 100101)
[3](中国科学院大学, 北京 100049)
通信作者: 李　俊, E-mail: lijun@cnic.cn

摘　要: 最近, 强化学习技术在序列推荐系统取得成功, 它能从用户长期反馈信号中学习有效的推荐策略. 然而, 模型的激励函数设计面临区分度过低的难题. 这限制了模型学习不同用户反馈信号间的价值差异的能力, 并导致推荐策略总是次优的. 现有工作主要通过调节衰减因子来保证激励函数区分度, 但它依赖专家先验知识缺乏理论基础. 为了更合理地设计激励函数和提高其区分度, 本文依据因果论来分析推荐系统, 并提出一种基于反事实区分度增强的序列推荐算法 CAL4Rec. 首先, 所提出方法用结构因果图描述序列推荐过程, 并创造性地用因果图定义了因果可鉴别的价值激励区分度. 其次, 该方法用反事实生成对抗的自监督学习过程优化推荐策略网络, 以学习用户的真实倾向. 在一系列序列推荐基准数据集上, 对 CAL4Rec 开展了广泛对比和消融实验, 实验结果表明 CAL4Rec 的提升对多种网络实现结构有效 (平均 2.34%).

关键词: 反事实推理; 生成对抗学习; 结构因果模型; 序列推荐

引用格式: 刘珈麟,贺泽宇,李俊.反事实增强的对抗学习序列推荐.计算机系统应用,2024,33(4):235–245. http://www.c-s-a.org.cn/1003-3254/9470.html

Counterfactual Enhanced Adversarial Learning for Sequential Recommendation

LIU Jia-Lin[1,3], HE Ze-Yu[2], LI Jun[1]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China)
[2](Computer School, Beijing Information Science and Technology University, Beijing 100101, China)
[3](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Recently, reinforcement learning techniques have achieved success in sequence recommendation systems, as they can learn effective recommendation strategies from long-term user feedback signals. However, the design of the model's reward function faces the challenge of low discriminability. This limits the model's ability to learn the value differences between different user feedback signals, leading to suboptimal recommendation strategies. Existing studies mainly ensure discriminability of the reward function by adjusting decay factors, but this relies on expert prior knowledge and lacks a theoretical foundation. In order to more reasonably design the reward function and enhance its discriminability, this study analyzes the recommendation system based on counterfactual reasoning and proposes a sequence recommendation algorithm CAL4Rec based on counterfactual discriminability enhancement. Firstly, the proposed method uses structural causal graphs to describe the sequence recommendation process and creatively defines causally identifiable value reward discriminability using causal graphs. Secondly, this method uses a counterfactual generative adversarial self-supervised learning process to optimize the recommendation strategy network and learn the user's true preferences. Extensive comparative and ablation experiments were conducted on a series of sequence recommendation benchmark

① 基金项目: 国家自然科学基金 (61672490, 61602436); 中国科学院对外合作重点项目 (241711KYSB20180002); 国家重大研发计划子课题 (2022YFC3320900)
收稿时间: 2023-10-27; 修改时间: 2023-11-27; 采用时间: 2023-12-07; csa 在线出版时间: 2024-03-07
CNKI 网络首发时间: 2024-03-11

datasets for CAL4Rec, and the experimental results show that CAL4Rec's improvement is effective for various network implementation structures (average 2.34%).

Key words: counterfactual reasoning; generative adversarial learning; structural causal model; sequential recommendation

推荐系统致力于感知用户真实兴趣和解决信息爆炸问题, 近年来在流媒体内容分享[1,2], 电子商务[3]等在线服务中得到广泛应用, 具有重要的潜在商业价值. 在推荐问题中, 用户的兴趣倾向包含在与平台交互的历史行为记录中, 不同的行为反馈代表不同的推荐价值, 随着记录时间累积增长, 系统可以从大量的交互序列中挖掘用户的兴趣爱好, 从而实现个性化的推荐. 强化学习提供了一种建模上述序列推荐过程的工具, 其思路是将不同的用户反馈信号映射为不同的激励价值, 并通过最大化累积价值激励函数的优化过程, 同时挖掘用户的序列动态兴趣和长时平稳倾向.

尽管, 基于强化学习的推荐系统最近取得成功, 这得益于有效的价值激励函数设计, 但是启发式设计的价值激励函数往往面临区分度低的难题[4–6]. 价值激励函数区分度描述了不同用户行为反馈信号的量化价值之间的差异性[4–6]. 例如, 在电子商务场景中, 常见的用户行为包括点击反馈和购买反馈, 后者的价值直觉上远高于前者 (如可设置模型的点击价值为 0.2, 购买价值为 1.0), 则区分度可以定义为"购买-点击"价值比[4–6], 价值比越大反映不同反馈信号之间的区分性越高. 作为强化学习优化过程的指导信号, 价值激励函数的区分度直接影响推荐策略空间的区分度, 即最优策略和策略空间其他备选策略簇的差异性. 这意味着, 价值激励函数的区分度决定了模型学习最优推荐策略的能力. 此外, 不合理的价值激励函数及区分度设计会导致模型学习到错误的用户反馈信号. 固定价值比形式可能导致不同价值的反馈信号互相替换, 如用 5 次被连续点击的商品替代 1 个被真实购买的商品. 一些研究发现, 仅追求高价值比的区分度设计会降低"动作-状态"价值函数时序差分学习的稳定性.

目前, 基于强化学习的推荐系统主要分为两个研究分支: (1) 衰减因子加权, 如 DEAR[7], DEER[3], SQN[4], VPQ[6], SDAC[8], KERL[9], 其思路是调整超参数衰减因子强调高价值反馈的及时性, 降低可能等价替换的连续低价值推荐的累积激励函数, 间接惩罚低区分度策略. 这种方法需要依赖专家知识对特定领域任务进行大量调试才能得到性能满意的参数设置. (2) 逆激励函数学习, 如 IRecGAN[10]、InvRec[11]、PGPR[12], 其思路是设计单独的优化目标学习激励函数, 其区分度被隐式地包含到激励函数的归纳偏置中[13]. 逆激励函数学习方法建模的归纳偏置[14]无法通过假设检验验证. 因此, 上述方法对价值激励函数的设计都缺乏充分的理论基础, 并导致模型学到的推荐策略总是次优的.

为了更合理地设计具有高区分度的价值奖励函数, 本文将因果理论引入到基于强化学习的推荐系统中. 一方面, 从因果推理的角度看, 区分度本质上是 (反事实) 假想模型推荐系统与 (事实) 真实系统之间的用户行为差异. 也就是说, 区分度越高时, 价值函数指导下训练的推荐策略网络产生的推荐结果应趋近于事实系统并尽可能远离反事实系统. 另一方面, 因果推断有助于改善序列推荐模型表现. 最新的序列推荐研究工作[15–17]也尝试用时序因果模型来描述基于循环神经网络 RNN 序列推荐模型的马尔可夫决策过程. 时序因果模型的可鉴别性仍然有待探索和验证, 并没有得到充分的理论证明, 而静态的结构因果模型可鉴别性的相关理论结论也不能被直接运用. 综上, 本文将区分度可视为一个反事实因果量, 采用结果因果模型来描述强化学习推荐系统 (在实际应用中反事实因果模型一般未知[18]), 并用反事实推理来更好的学习推荐策略.

本文提供了一种新的基于反事实推理的序列推荐方法 (counterfactual adversarial learning for recommendation, CAL4Rec). 首先, 本文引入结构因果理论模型来分析序列推荐过程, 分析结果显示用户兴趣倾向和区分度共同决定了价值激励, 因此可通过因果干预操作, 避免价值函数的区分度过低问题. 其次, 所提出方法设计了一种新颖的反事实增强的高区分度激励函数, 以通过更合理的区分度设计让激励机制更好地指导模型学习. 此方法根据序列推荐因果模型和反事实推理操作重新定义了区分度, 即反事实区分度, 并将用户内在兴趣倾向显式构建为因果结构方程, 再通过 Gumbel-Max 神经网络学习该方程来获得因果关系可鉴别的价值激励函数. 不同于之前启发式的和依赖专

家知识的方法, 这种新设计具有更强的理论性, 因为模型所学习的用户兴趣归纳偏置是可检验的. 第三, CAL4Rec 模型使用反事实生成对抗训练框架, 来根据上述高区分度激励函数训练强化学习序列推荐模型. 具体的, 生成器网络用于模拟假想推荐系统和生成反事实价值, 鉴别器网络利用高区分度激励函数来区分事实观测数据分布和反事实数据分布. 通过鉴别分数来引导生成器网络逼近用户真实交互行为及兴趣. 最后, 多个基准数据集上的总体性能对比和消融实验分析证明了 CAL4Rec 设计的有效性.

1 研究基础及术语

本节针对 CAL4Rec 的工作基础进行阐述: 首先, 介绍了因果推理和强化学习相关的符号和定义; 然后, 详细介绍了序列推荐任务的研究工作进展.

1.1 术语

符号: 大写字母X表示随机变量, 小写字母x表示随机变量一次采样值. 粗体字母$\boldsymbol{X}$或$\boldsymbol{x}$表示随机变量或其取值的集合. 上角标表示t时刻的随机变量$X^{(t)}$, 下角标表示第k种干预变量x_k对随机变量Y的干预结果$Y_{k[x_k]}$: 测数据上的事实推荐($k=1$), 和推荐模型模拟随机干预的反事实推荐($k=2$).

强化学习将序列推荐问题定义为一个马尔可夫决策过程$(\mathcal{S},\mathcal{A},P,R,\rho_0,\gamma)$, 其中:

- $S\in\mathbb{R}^{d_s}$: 用户状态的随机变量, 由一个从交互序列学到的d_s维编码向量 (embedding) 表示.
- $A\in\mathbb{R}^{N}$: 推荐动作的离散随机变量, 系统的动作A是商品池中待推荐的选项, 由编码空间$E\in\mathbb{R}^{d_e\times N}$得到$A$的表征向量.
- $P:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}$用户状态分布的转移概率, 表示交互过程中用户收到推荐后的兴趣动态.
- $R:\mathcal{S}\times\mathcal{A}\to\mathbb{R}^{d_r}$表征用户反馈信号 (如点击, 购买等) 价值的离散随机变量. 系统将一个商品推荐给用户, 用户根据推荐质量给出反馈R, 系统使用价值R作为激励信号. 最大化累积激励函数是对用户长期兴趣倾向的挖掘, 弥补了深度序列模型作为序列推荐系统的目标函数的不足.
- $Y\in\mathbb{R}^{d_r}$: 用户内在倾向随机变量, 由当前状态S和系统推荐A共同影响的不可观测随机变量.
- ρ_0: 用户状态的初始分布.
- γ: 激励函数R的衰减因子, 奖励高价值A的及时性, 惩罚其可能的滞后性.

结构因果模型 (structural causal model, SCM) 是一组量化因果效应作用过程的方程[19], 定义为一个四元组$\langle\boldsymbol{U},\boldsymbol{V},\mathcal{F},P(\boldsymbol{U})\rangle$, $\boldsymbol{U}$表示影响 SCM 的环境因素 (外生变量), 其概率分布为$P(\boldsymbol{U})$. $\boldsymbol{V}=\{V_1,V_2,\cdots,V_n\}$是 SCM 描述的研究对象 (内生变量), 这些变量是由模型中的其他变量决定的. $\mathcal{F}=\left\{f_{V_1},f_{V_2},\cdots,f_{V_n}\right\}$是一组描述内生变量因果作用过程的函数, 第$i$个结构因果方程$f_{V_i}$是从因变量集$\boldsymbol{U}_{V_i}\cup\boldsymbol{Pa}_{V_i}$到果变量$V_i$的映射, 其中$\boldsymbol{U}_{V_i}\subseteq\boldsymbol{U},\boldsymbol{Pa}_{V_i}\subseteq\boldsymbol{V}\backslash V_i$, SCM 的结构因果方程簇$\mathcal{F}$构成了外生变量$\boldsymbol{U}$到内生变量$\boldsymbol{V}$的映射.

反事实联合概率分布 (counterfactual distribution) 描述了给定 SCM 中多种干预组合可能得到的潜在结果: 给定待学习 (未知参数为θ) 的结构因果模型$\mathcal{M}(\theta)$和干预集合$\boldsymbol{X}=\{\boldsymbol{X}_k:\boldsymbol{X}_k\subseteq\boldsymbol{V},k=1,\cdots,K\}$, 反事实联合概率分布$P^{\mathcal{M}(\theta)}\left(Y_{1[x_1]},\cdots,Y_{K[x_K]}\right)$可定义为:

$$\int_{\mathcal{D}_{\boldsymbol{\mu}}}1\left[\boldsymbol{Y}_{1[\boldsymbol{x}_1]}(\boldsymbol{\mu})=\boldsymbol{y}_1,\cdots,\boldsymbol{Y}_{K[\boldsymbol{x}_K]}(\boldsymbol{\mu})=\boldsymbol{y}_K\right]\mathrm{d}P(\boldsymbol{\mu}) \tag{1}$$

其中, $Y_{k[\boldsymbol{x}_k]}(\boldsymbol{\mu})$描述了第$k$个干预的反事实潜在结果:

$$\left\{f_{V_j}:V_j\in V\backslash X_k\right\}\cup\{f_X\leftarrow x:X\in X_k\} \tag{2}$$

1.2 研究基础

序列推荐作为推荐系统的重要研究分支受到研究人员的广泛关注[20,21], 其研究思路经历了深度学习时代前的协同滤波, 深度序列推荐模型和强化学习序列推荐 3 个阶段, 本节阐述不同阶段的研究工作基础.

1.2.1 序列推荐算法

传统推荐算法假设相似的用户具有相似的喜好倾向提出了基于矩阵分解的协同滤波算法[22–24]. BPR[22]提出了一种贝叶斯个性化排序推荐方法 (成对型排序损失函数), 使用一个有偏估计的分解矩阵作为推荐系统. 针对有偏估计矩阵分解的问题, NCF[24]首次提出使用深度神经网络估计“用户-商品”协同矩阵. FPMC[23]则针对矩阵分解的方法无法建模“用户-商品”交互过程的问题提出一种基于马尔可夫链的协同过滤模型, 将交互序列近似为一阶马尔可夫链, 并在序列化增强的成对型排序损失上优化. 上述方法无法建模高阶用户-商品交互过程.

传统推荐算法的缺点在于无法建模高阶用户-商品交互过程. 基于深度学习的推荐模型将“用户-商品”交互过程建模为时序序列, 模型的潜状态向量通过模

型学习可以挖掘用户的高阶动态兴趣倾向. GRURec[25]应用序列化神经网络预测下一时刻用户的兴趣倾向. 为了解决循环神经网络的梯度消散问题和计算效率问题, Caser[26]和 NextIt[27]使用卷积神经网络作为推荐骨干网络. SASRec[28]受到机器翻译等序列化生成任务的启发, 将 Trasformer[29]结构作为推荐骨干网络. 由于序列推荐系统中存在多种用户反馈信号, 不同类型的反馈信号对系统具有不同的价值, 基于深度学习的推荐模型的局限是没有考虑不同反馈信号的价值.

1.2.2 强化学习序列推荐算法

基于强化学习的序列推荐将用户交互序列建模为马尔可夫决策过程, 并通过最大化不同反馈信号的价值累积奖励函数来学习用户的长时兴趣, 弥补传统序列推荐方法受到梯度消散影响的不足. 其总体思路是将交互序列按照交互时刻的顺序划分为一系列连续的用户状态 (state) 转移, 该转移过程受到推荐动作 (action) 影响, 得到正反馈的推荐动作加入用户状态序列并通过编码映射学习用户状态向量, 并通过可调节的衰减因子最大化完整序列轨迹的累计价值奖励得到参数化的推荐策略π_θ:

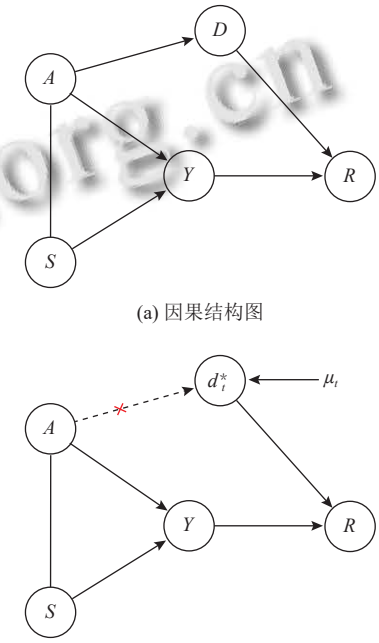
$$\max_{\pi_\theta}\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{|\tau|}\gamma^t r(s_t,a_t)\right] \tag{3}$$

其中, γ表示可调节的衰减因子, 用于调整不同时刻价值反馈$r(s_t,a_t)$的重要性. 已有工作按照采用的强化学习技术类别可以分为: (1) 基于梯度的方法, 考虑到推荐问题对实时用户交互的限制, off-policy REINFORCE[5]计算推荐策略网络的策略梯度以实现 YouTube 平台的视频推荐, 由于梯度计算是同策略估计需要对采样行为策略样本进行矫正以避免分布漂移引入干扰, 该方法提出一种基于倾向性分数的梯度重估方法. (2) 基于价值函数的方法, SQN[4]模型利用“动作-状态”价值函数时序差分[30]优化来学习累积价值激励最大化, 并通过联合优化交叉熵时序预测来学习用户的动态兴趣变化趋势, VPQ[6]在 SQN 的基础上利用重采样方法降低时序差分学习的方差. (3) 基于“动作-评价”结构的方法, SAC[4]利用“动作-状态”价值函数作为样本权重加权交叉熵时序预测, SDAC[8]针对策略网络动作空间的离散性提出了基于 Gumbel-Softmax 分布的策略估计模型. 基于强化学习的序列推荐系统需要依赖激励函数作为累积奖励最大化过程的优化信号, 虽然较低的购买点击比可以提高 Q 学习[30]的稳定性, 但同时会削弱区别性, 即给定当前交互历史, 一种连续推荐一系列低激励价值候选项的次优策略可能与另外一种直接返回一个高价值候选项的最优策略 (“命中”用户兴趣倾向) 具有相近的累积奖励, 已有研究工作通过分配时序敏感的衰减因子来增强不同价值用户反馈之间的区别性, 但设置合适的衰减因子也需要反复试错.

为了解决强化学习序列推荐算法存在的激励函数区分性不足的问题, 本文提出一种基于反事实生成对抗学习的序列推荐算法, 其核心思路是利用均匀分布模拟低区分度的推荐策略, 并通过生成对抗学习使得模型预测的反馈价值与事实观测值分布相近.

2 反事实生成对抗序列推荐模型

CAL4Rec 将结构因果图语言工具引入到序列推荐任务以提供可假设检验的和有理论依据保证的价值区分度模型. 本节首先介绍了序列推荐任务的定义, 在此基础上详述了 CAL4Rec 所提出的因果图模型和可学习的神经结构方程, 第 2.4 节阐述了用于学习图 1 所示结构因果模型的生成对抗算法, 最后讨论了 CAL4Rec 激励函数与已有工作的联系与区别.



(a) 因果结构图

(b) 模拟低区分度随机策略的后门干预

图 1 CAL4Rec 序列推荐的因果结构图

2.1 问题定义

序列推荐问题可以定义为给定用户M个浏览历史

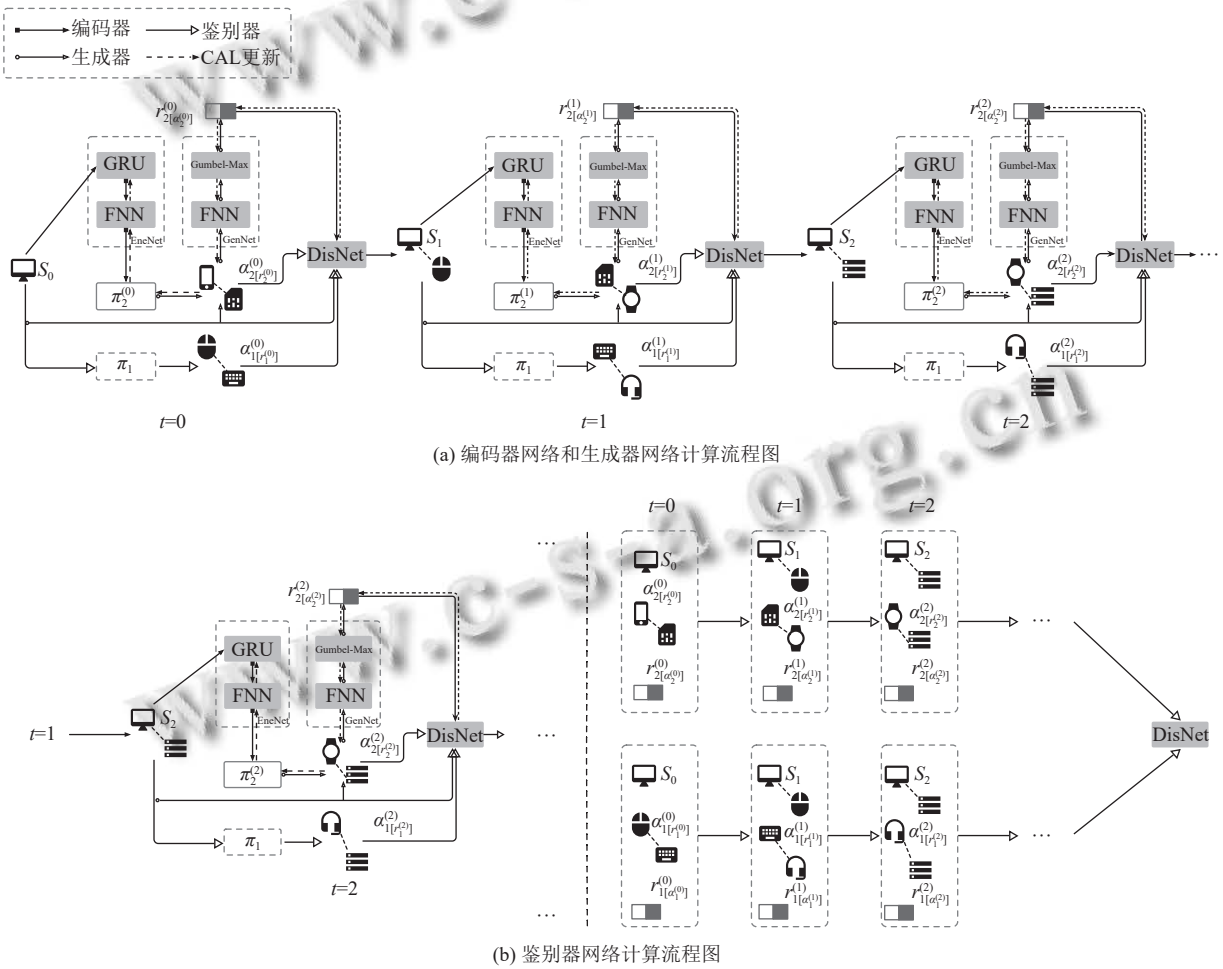
和推荐记录$\left\{\left[r_{1[a]}^{(0)},r_{1[a]}^{(1)},\cdots,r_{1[a]}^{(T_m)}\right]\right\}_{m=1}^{|A|}$, 系统学习预测下一时刻$t$用户的兴趣倾向$r_{2[a]}^{(t)}$和最优的推荐商品, 使得$r_{2[a]}^{(t)}\approx r_{1[a]}^{(t)}$. 其中, 不同的兴趣倾向$r$代表不同的两类用户行为, 因而系统对用户倾向预测的性能同时受到推荐策略区分度的影响.

2.2 框架概述

图 1 为 CAL4Rec 所提出的序列推荐因果结构与后门干预[31]. 图 1(a) 的因果图抽象了序列推荐过程各变量制约关系, 推荐系统根据用户当前的浏览状态S推荐选项A, 推荐项A和用户状态S影响用户的内在倾向Y, 同时不同的推荐选项A具有不同的推荐区分度D, 用户内在兴趣倾向Y和区分度D则决定了本轮推荐交互的价值激励R. 由图 1(a) 可知, $D\to R$存在后门路径$D\to A\to Y\to R$, $D\to A\to S\to Y\to R$且变量Y不可观测, 根据后门准则[31]可得干预区分度变量D可使用观测数据集代替随机控制实验[18]预测用户可能的反馈价值激励. 由于随机策略代表了一类低区分度的推荐策略 (不同价值的反馈信号均等概率采样), CAL4Rec 采用均匀分布作为后门干预模拟低区分度随机策略, 如图 1(b) 所示.

2.3 模型结构

图 1 所提出的因果图定性地描述了序列推荐问题的结构因果模型 SCM, 由于结构因果方程一般未知[18], CAL4Rec 采用了时序单向的神经网络结构 (如图 2 所示), 本节详细阐述该模型结构. 其中, $k=1$为观测数据集, 来自未知的采集策略π_1, 描述事实推荐策略; $k=2$为 CAL4Rec 生成的推荐序列, 来自推荐策略π_2 (式 (4)), 描述了反事实推荐生成策略.



(a) 编码器网络和生成器网络计算流程图

(b) 鉴别器网络计算流程图

图 2 CAL4Rec 总体架构

商品/用户的表征编码: 给定用户当前的浏览历史和推荐记录$\left[r_{1[a]}^{(0)},r_{1[a]}^{(1)},\cdots,r_{1[a]}^{(t-1)}\right]$, 推荐商品$a_{1[r_1]}^{(t-1)}$首先通过映射到编码空间$E\in\mathbb{R}^{d_e\times N}$得到一个唯一的低维编码向量$e^{(t-1)}\in\mathbb{R}^{d_e}$, 其中$N$是推荐池容量 (商品数量), d_e是

编码向量维度, 编码空间E由正态分布随机采样得到. 用户的状态编码向量$s^{(t)}\in\mathbb{R}^{d_s}$通过深度自回归模型$E$得到:

$$s^{(t)}=E\left(s^{(t-1)},e^{(t-1)};\theta_E\right) \tag{4}$$

其中, 初始状态$s^{(0)}=e^{(0)}$. 神经网络编码器E存在多种实现结构: 基于循环神经网络[25], 基于卷积神经网络[26,27]和基于注意力机制的前馈神经网络[28,32]. 实验中分别实现了这 3 类编码器并对进行了测试.

推荐策略网络 (因果图$s\to a$): 受 IRecGAN[10]模型启发, 文中采用了同样的神经后验概率估计思路. 具体来说, 根据t时刻用户的状态编码$s^{(t)}$, 系统推荐商品a的归一化概率表示为:

$$\pi\left(a\mid s^{(t)}\right)=\frac{\exp\left(W_a s^{(t)}+b_a\right)}{\sum_{j}^{N}\exp\left(W_j s^{(t)}+b_j\right)} \tag{5}$$

其中, $W\in\mathbb{R}^{N\times d_s}$, $b\in\mathbb{R}^{d_s}$为模型待优化网络参数, $\pi\left(a\mid s^{(t)}\right)$为系统的推荐策略.

用户内倾向学习 (因果图$(s,a)\to y$): 用户的不同反馈倾向$Y\in\mathbb{R}^{d_r}$受到用户当前状态和推荐商品共同影响:

$$Y=f_y\left(W_y\left[s^{(t)};a^{(t)}\right]+b_y\right) \tag{6}$$

其中, $W_y,b_y\in\mathbb{R}^{d_r}$是网络参数, $\left[s^{(t)};a^{(t)}\right]\in\mathbb{R}^{(d_s+d_e)}$是状态编码和推荐商品编码的拼接 (Concatenate). f_y表示非线性映射函数. 用户的内倾向映射学习结果等价于一个未归一化的神经后验概率$p_r(y\mid s,a)$.

反事实增强的高区分度激励函数 (因果图$(d,y)\to r$): 为了学习高区分度的用户反馈价值R, 受反事实异策略强化学习估计工作[33]的启发, 用户内倾向Y在当前策略上的反事实区分度可以定义为:

$$D_r=\log\frac{p_r(y\mid s,a)}{-\log\mu} \tag{7}$$

其中, $\mu\sim U(0,1)$均匀分布, 模拟完全随机的推荐倾向 (推荐区分度为 0). $D_r=0$表示推荐模型的区分度与完全随机推荐等价 (区分度为 0).

$$\begin{aligned}r_{2[a]}^{(t)}&=\arg\max_r\{D_r\}\\&=\arg\max\left\{\log p_r\left(y\mid s^{(t)},a^{(t)}\right)-\log\left(-\log\left(\mu^{(t)}\right)\right)\right\}\end{aligned} \tag{8}$$

其中, 由于推荐问题搜索空间庞大 ($|A|=N\gg 1$), 为了加速模型训练过程, 可以用 Softmax 算子近似 argmax 过程[8,34].

2.4 模型优化

为了学习第 2.3 节提出的模型结构, CAL4Rec 采用生成对抗学习 (算法 1) 编码器和生成器 (图 2(a)), 及鉴别器网络结构 (图 2(b)), 本节详细阐述该算法.

反事实生成对抗学习: 为了使价值激励函数 (式 (7)) 保持因果可鉴别的区分度的同时能反映用户真实倾向, CAL4Rec 使用鉴别器网络$D\left(r^{(t)}\mid s^{(t)},a^{(t)}\right)$鉴别价值激励函数的干预集合类别: 观测数据上的事实推荐$r_{1[a]}^{(t)}$, 模型的反事实推荐$r_{2[a]}^{(t)}$. 同时, 价值激励函数网络 (式 (7)) 的作用等价于生成器网络$G\left(\mu^{(t)}\mid s^{(t)},a^{(t)}\right)$. 鉴别器网络与生成器网络共同构成了生成对抗学习过程, 鉴别器网络的优化目标为:

$$\begin{aligned}\mathcal{L}_D^{(t)}=&\mathbb{E}_{r^{(t)}\sim p\left(R_{1[a_1]}\right)}\left[\log D\left(r_{1[a_1]}^{(t)}\mid s^{(t)},a^{(t)}\right)\right]\\&+\mathbb{E}_{\mu^{(t)}\sim P(U)}\left[\log\left(1-D\left(G\left(\mu^{(t)}\mid s^{(t)},a^{(t)}\right)\right)\right)\right]\end{aligned} \tag{9}$$

而生成器的优化目标为:

$$\mathcal{L}_G^{(t)}=\mathbb{E}_{\mu^{(t)}\sim P(U)}\left[\log\left(1-D\left(G\left(\mu^{(t)}\mid s^{(t)},a^{(t)}\right)\right)\right)\right] \tag{10}$$

由式 (9) 和式 (10) 可知, 生成器试图使得推荐系统预测的用户反馈$r_{2[a]}^{(t)}$观测数据的真实反馈$r_{1[a]}^{(t)}$相似, 鉴别器则进行相反过程, 当生成器网络与鉴别器网络优化过程达到动态平衡时有$r_{2[a]}^{(t)}=r_{1[a]}^{(t)}$[35].

自监督正则项: 由于鉴别器网络没有关于观测数据概率分布的先验知识, 因此其优化过程需要完全依赖生成器网络信息, 导致生成对抗学习过程难以挖掘用户兴趣倾向的序列动态性. 为了挖掘用户的序列动态特性以提升状态编码, CAL4Rec 在编码器学习中引入自监督交叉熵目标函数作为正则目标:

$$\mathcal{L}_{\text{ssl}}^{(t)}=-\sum_{a\in|A|}1\left[A_{1[r_1]}^{(t)}=a\right]\log\left(\pi\left(a\mid s^{(t)}\right)\right)$$

其中, 推荐结果a来自于观测数据集$\left\{\left[r_{1[a]}^{(0)},r_{1[a]}^{(1)},\cdots,r_{1[a]}^{(T_m)}\right]\right\}_{m=1}^{|A|}$.

算法 1. 反事实生成对抗 CAL 学习

输入: 观测数据$\left\{\left[r_{1[a]}^{(0)},r_{1[a]}^{(1)},\cdots,r_{1[a]}^{(T_m)}\right]\right\}_{m=1}^{|A|}$, 先验概率$U\sim Unif[0,1]$.

输出: 鉴别器, 生成器和编码器的网络参数$(\theta_D,\theta_G,\theta_E)$.

1) 初始化参数$\theta_D,\theta_G,\theta_E$
2) 令$i=1,\cdots,I$. 执行下述外循环操作:
3) 令$j=1,\cdots,K$. 执行下述内循环操作:
4) 批量采样B个外生变量U

5) 批量采样观测数据 $\left\{\left[r_{1[a]}^{(0)},\cdots,r_{1[a]}^{(T_b)}\right]\right\}_{b=1}^{B}$.
6) 计算梯度 $\nabla_{\theta_D}\frac{1}{B}\sum_{b=1}^{B}\sum_{t=1}^{T_b}\left[L_D^{(t)}\right]$ 更新 $D(r|s,a;\theta_D)$
7) 结束内循环
8) 批量采样 B 个外生变量 U
9) 计算梯度 $\nabla_{\theta_G}\frac{1}{B}\sum_{b=1}^{B}\sum_{t=1}^{T_b}\left[L_G^{(t)}\right]$ 更新 $G(\mu|s,a;\theta_G)$
10) 计算梯度 $\nabla_{\theta_E}\frac{1}{B}\sum_{b=1}^{B}\sum_{t=1}^{T_b}\left[L_{ssl}^{(t)}+L_G^{(t)}\right]$ 更新 $E(s,a;\theta_E)$
11) 结束外循环

值得指出的是, 算法 1 的生成对抗训练过程采用了经典的条件生成对抗网络 (CGAN) 的优化过程. 其相同之处在于, 当“生成-鉴别”过程趋于动态平衡时, 理论上生成器反事实预测分布与事实观测数据分布同簇 ($r_{2[a]}^{(t)}=r_{1[a]}^{(t)}$); 其区别之处在于, 反事实生成对抗训练中, 生成器依赖 Gumbel-Max 算子模块保证网络的因果鉴别性, 从因果推理角度看, 该模块将反事实区分度转换为鉴别性需求, 鉴别性理论上的保证使得反事实分布可以通过数据集分布进行观测学习, 该过程一般情况无法进行[31]; 从模型优化角度看, Gumbel-Max 模块起到缩减生成器网络参数空间的作用, 因为能够更有效地从离线数据集中学习.

2.5 相关讨论

图 1 提供了一种新的分类相关工作激励函数设计的因果视角. 深度序列推荐模型[25,26,28]和强化学习序列推荐模型[6,9,12,36,37]工作, 其编码器采用循环神经网络, 卷积神经网络和 Transformer[29]等自回归深度模型, 即 $S\to A$ 的结构因果方程相似. 区别是 $Y\to R$ 的结构因果方程的估计方法不同, 深度序列推荐模型使用 $R=Y$, 强化学习序列推荐模型设置 $R=\sum_t \gamma^t f_r(y)$, 其中 f_r 是一个启发式设计[36]:

$$f_r=\begin{cases}0.2, & \text{if } y=\text{点击}\\ 1, & \text{if } y=\text{购买}\end{cases} \tag{11}$$

上述方法或忽略了区分度 (深度序列推荐模型) 或区分度不可鉴别 (强化学习序列推荐模型, 由式 (11) 可知存在 5 个点击反馈等价于 1 个购买的可能). 而式 (7) 保证 CAL4Rec 的反馈激励函数 f_r 的区分度是因果可鉴别的.

3 实验结果及分析

为了验证 CAL4Rec 的有效性, 两个从真实电子商务场景采集的基准序列推荐数据集被用于实验分析. 为了证明提出方法有效地增加了激励函数的区分度, CAL4Rec 首先与具有代表性的强化学习推荐系统 SOTA 基线方法比较了总体 Top@k 推荐性能. 为了验证关键设计的有效性, 在 Retailrocket 数据集上进行的消融实验进一步证明了 Gumbel-Max 算子和生成对抗学习 (算法 1) 对推荐性能提升的帮助.

实验设置: Yoochoose 和 Retailrocket 是两个序列推荐的标准数据集, 分别包含两类交互正反馈 (点击和购买), 多类反馈的处理思路与两类反馈类似. 为了对比实验的公平一致性, 这里统一按照文献[4]的预处理过程, 删除了 Yoochoose 和 Retailrocket 中互动次数少于 3 次的交互序列, 得到的数据集统计结果如表 1 所示. 实验用于衡量推荐性能的两个指标是: 表征 Top@k 排序性能 ($k\in\{5,10,20\}$) 的归一化折损累计增益 (normalized discounted cumulative gain, $NG@k$)[4]; 反映召回性能的命中率 (hit ratio, $HR@k$), 以点击 (click) 为例:

$$HR=\frac{\#\text{点击命中数}}{\#\text{点击总数}} \tag{12}$$

表 1 数据统计

统计项	Retailrocket	Yoochoose
交互序列数	195 523	200 000
交互项目数	70 852	26 702
反馈点击数	1 176 680	1 110 965
反馈购买数	57 269	43 946

对比基线: 针对 3 种具有代表性的 SOTA 强化学习序列推荐算法被作为比较基线: 基于“状态-动作”价值函数的 SQN[4], VPQ[6], 和基于“动作-评论”架构的 SAC[4], 3 种方法联合优化时序差分学习[38] 和交叉熵自监督学习[28]. 为了证明 CAL4Rec 方法适用于不同编码器骨干网络, 实验考虑 3 种典型骨干网络结构: 循环神经网络 GRU[25], 卷积神经网络 NextIt[27]和 Transformer[29] 结构的 SASRec[28].

实现细节: 两个数据集采用的输入序列长度均为 10 个当前时刻的近期交互, 编码向量均采用 64 维, 干预变量 μ 均匀分布为 32 维, CAL4Rec 采用独热向量 (one-hot) 编码购买反馈和点击反馈, 批量输入 (batch size) 大小均为 256, 并通过 Adam 优化器同时更新鉴别器、生成器和编码器. 其中, 生成器和鉴别器的学习率为 0.005, 而编码器的学习率为 0.01. 由于 GPU 计算能力的限制, 编码器的骨干网络深度为 1 层, GRU 的隐状态维度 64, 膨胀卷积块 (dilated convolution block) 保持 6 层 64 通道卷积核大小 3 的设置[27], Transformer 骨干网络保持单头 1 层的自注意力块[28], 生成器网络

与鉴别器网络均为两层前馈神经网络 (隐藏层 64 维), 鉴别器输出非线性映射函数为 Sigmoid. SQN[4], SAC[4] 和 VPQ[6]均设购买价值为 1, 点击价值为 0.2, 骨干网络与 CAL4Rec 相同. 3 种骨干网络结构在训练过程中, 首先输入长度为 10 的最近交互序列, 经过编码器编码得到 64 维状态编码向量作为生成器的输入; 生成器首先采样 256 批量大小的随机均匀分布外生变量, 基于编码器结果拼接“状态-动作”编码向量对作为生成器输入, 经过 Gumbel-Max 算子前馈计算得到预测潜在的用户反馈; 鉴别器基于生成器预测结果, 拼接生成器“状体-动作”编码对作为条件输入项, 预测生成器预测反馈的数据来源 (事实观测数据分布或预测反事实生成分布).

3.1 总体性能比较

表 2 为 CAL4Rec 在 Retailrocket 和 Yoochoose 数据集上的性能比较. 进一步分析可以发现: (1) 对比不同骨干网络的实现结构, CAL4Rec 均有一致的提升, 证明了所提出方法适用于多种序列推荐骨干网络, 该提升性来自于 CAL4Rec 利用了用户反馈的不同激励价值作为生成对抗学习过程的信号, SQN[4], SAC[4]和 VPQ[6]同样利用到了反馈信号作为价值激励函数. (2) 对比同一个骨干网络的实现结构, CAL4Rec 较 SQN, SAC 和 VPQ 总体性能平均提升了 0.082 8, 证明了 CAL4Rec 的反馈奖励函数具备更大的区分度, 直观定性的分析, 这是因为生成对抗学习驱使编码器网络和生成器网络在均匀随机分布模拟低区分度推荐策略的干扰下, 拟合真实观测数据; 从因果角度定量的分析, 式 (6) 的鉴别性保证了用户反馈价值的区分度, 反馈价值的区分度进一步通过模型生成对抗训练过程保证了推荐策略式 (4) 的区分度.

表 2 整体性能比较

方法	Retailrocket				Yoochoose				Total
	Purchase		Click		Purchase		Click		
	HR@20	*NG*@20	*HR*@20	*NG*@20	*HR*@20	*NG*@20	*HR*@20	*NG*@20	
GRU-SQN	0.5258	0.3804	0.3345	0.2142	0.6339	0.3679	0.4771	0.2604	3.1942
GRU-SAC	0.5667	0.4211	0.3309	0.2144	0.6447	0.3721	0.4556	0.2472	3.2527
GRU-VPQ	0.5678	0.4223	0.3302	0.2145	0.6742	0.3950	0.4730	0.2599	3.3369
GRU-Our	**0.5783***	**0.4324***	**0.3408***	**0.2246***	**0.6847***	**0.4051***	**0.4835***	**0.2701***	**3.4195***
NextIt-SQN	0.6737	0.5108	0.3429	0.2278	0.5848	0.3353	0.4840	0.2660	3.4253
NextIt-SAC	0.6787	0.5321	0.3377	0.2176	0.5806	0.3377	0.4868	0.2595	3.4307
NextIt-VPQ	0.6858	0.5419	0.3449	0.d26	0.6061	0.3487	0.4789	0.2629	3.4918
NextIt-Our	**0.6963***	**0.5520***	**0.3554***	**0.2328***	**0.6167***	**0.3589***	**0.4895***	**0.2730***	**3.5746***
SASRec-SQN	0.6978	0.5320	0.3832	0.2401	0.6544	0.3708	0.5070	0.2797	3.6650
SASRec-SAC	0.6905	0.5556	0.3769	0.2390	0.6569	0.3868	0.4977	0.2728	3.6762
SASRec-VPQ	0.7156	0.5905	0.3763	0.2470	0.6879	0.4069	0.5099	0.2832	3.8173
SASRec-Our	**0.7261***	**0.6007***	**0.3868***	**0.2572***	**0.6985***	**0.4170***	**0.5205***	**0.2934***	**3.9002***

注: 最优结果用粗体表示, 次优结果下划线表示, “*”表示双侧t检验, p<0.05

综上所述, 表 2 的实验结果证明 CAL4Rec 适用于多种编码器骨干网络并取得一致的区分度提升.

3.2 消融实验分析

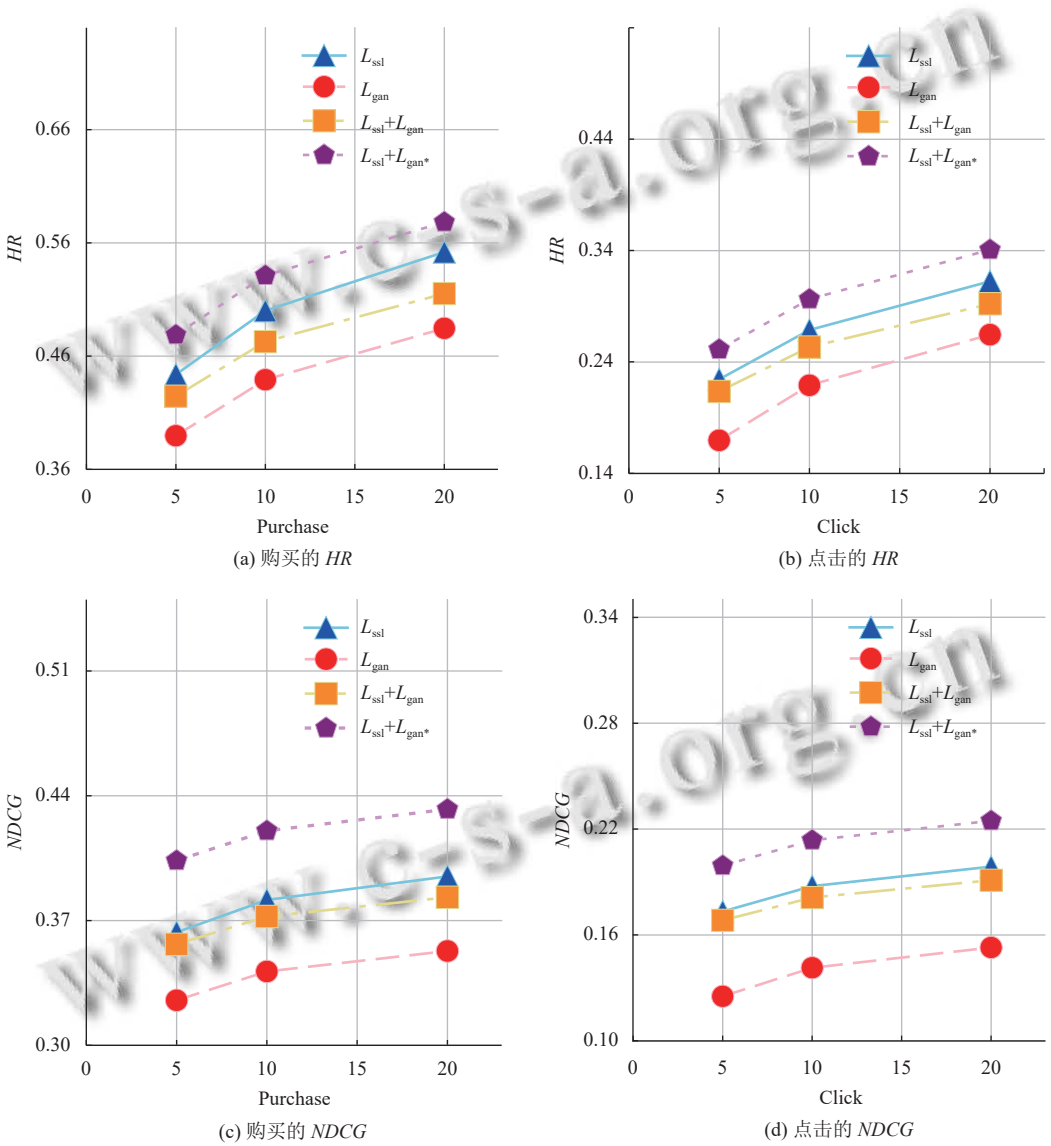
如图 3 所示, Retailrocket 数据集上的消融实验分析进一步研究了关键设计模块 Gumbel-Max 模块 (式 (7) 所示) 和生成对抗联合优化学习目标 (算法 1) 对 CAL4Rec 性能提升的作用, 实验中采用 GRU 作为编码器骨干网络, 不同的骨干网络均得到相似的变化趋势. 其中, L_{ssl} 表示仅依靠自监督目标函数优化策略推荐网络和编码器, 即经典序列推荐模型的优化过程. L_{gan} 表示仅依靠生成对抗学习无自监督正则优化 CAL4Rec; $L_{\text{ssl}}+L_{\text{gan}}$ 表示去掉反事实区分度增强的 Gumbel-Max 设计; $L_{\text{ssl}}+L_{\text{gan}*}$ 表示采用 Gumbel-Max 的联合优化也即 CAL4Rec 的完整模型/优化过程.

图 3(a)–(d) 的实验结果在两类反馈和两类评价指标上均得到了一致性的曲线趋势. 首先, 仅采用生成对抗学习过程 (虚线L_{gan}) 的推荐性能最差, 这是因为生成对抗学习本身不包括用户状态转移分布的信息, 因此需要自监督学习作为正则项进行规范 (虚线$L_{\text{ssl}}+L_{\text{gan}}$). 但自监督正则的生成对抗学习表现结果仍低于纯自监督学习 (实线L_{ssl}), 因为常规鉴别器网络没有关于观测数据收敛情况的先验知识. 最后, 价值区分度作为一类先验知识通过 Gumbel-Max 模块引入到 CAL4-Rec 设计的归纳偏置中, 提高了价值激励R的区分度进

而提升了整体推荐性能 (虚线$L_{ssl}+L_{gan*}$). 基于策略参数的鉴别器可以进一步提高性能, 这将作为未来改进方向.

4 结束语

针对强化学习序列推荐面临的价值激励函数区分度不足的问题, 提出了一种反事实区分度增强的生成对抗序列推荐方法 CAL4Rec. 它率先采用反事实推理定义区分度和价值激励函数, 其优点在于因果推理理论保证了该定义的鉴别性, 即 Gumbel-Max 的归纳偏置可假设检验. CAL4Rec 采用一种自监督规范的生成对抗学习过程学习结构因果方程. 在一系列基准数据集上进行的广泛实验证明所提方法的有效性. CAL4Rec 提供了一种利用因果图建模序列推荐过程并利用反事实鉴别性检验模型归纳偏置的思路, 为反事实缓解冷启动等关键问题提供了探索思路.



(a) 购买的 *HR* (b) 点击的 *HR*

(c) 购买的 *NDCG* (d) 点击的 *NDCG*

图 3 Retailrocket 上 Top@k 消融实验结果

参考文献

1 Hansen C, Hansen C, Maystre L, *et al*. Contextual and sequential user embeddings for large-scale music recommendation. Proceedings of the 14th ACM Conference on Recommender Systems. ACM, 2020. 53–62.

2 Covington P, Adams J, Sargin E. Deep neural networks for youtube recommendations. Proceedings of the 10th ACM Conference on Recommender Systems. Boston: ACM, 2016. 191–198.

3 Zhao XY, Zhang L, Ding ZY, *et al*. Recommendations with

negative feedback via pairwise deep reinforcement learning. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018. 1040–1048.

4 Xin X, Karatzoglou A, Arapakis I, *et al*. Self-supervised reinforcement learning for recommender systems. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2020. 931–940.

5 Chen MM, Beutel A, Covington P, *et al*. Top-k off-policy correction for a REINFORCE recommender system. Proceedings of the 12th ACM International Conference on Web Search and Data Mining. Melbourne: ACM, 2019. 456–464.

6 Gao CQ, Xu K, Zhou KQ, *et al*. Value penalized Q-learning for recommender systems. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid: ACM, 2022. 2008–2012.

7 Zhao XY, Gu CS, Zhang HSL, *et al*. DEAR: Deep reinforcement learning for online advertising impression in recommender systems. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI, 2021. 750–758.

8 Xiao T, Wang DL. A general offline reinforcement learning framework for interactive recommendation. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI, 2021. 4512–4520.

9 Wang PF, Fan Y, Xia L, *et al*. KERL: A knowledge-guided reinforcement learning model for sequential recommendation. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2020. 209–218.

10 Bai XY, Guan J, Wang HN. A model-based reinforcement learning with adversarial training for online recommendation. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 963.

11 Chen XC, Yao LN, Sun AX, *et al*. Generative inverse deep reinforcement learning for online recommendation. Proceedings of the 30th ACM International Conference on Information & Knowledge Management. ACM, 2021. 201–210.

12 Xian YK, Fu ZH, Muthukrishnan S, *et al*. Reinforcement knowledge graph reasoning for explainable recommendation. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris: ACM, 2019. 285–294.

13 Battaglia PW, Hamrick JB, Bapst V, *et al*. Relational inductive biases, deep learning, and graph networks. arXiv: 1806.01261, 2018.

14 Pearl J. Theoretical impediments to machine learning with seven sparks from the causal revolution. Proceedings of the 11th ACM International Conference on Web Search and Data Mining. Marina Del Rey: ACM, 2018. 3.

15 Tsirtsis S, De A, Gomez-Rodriguez M. Counterfactual explanations in sequential decision making under uncertainty. Proceedings of the 35th Conference on Neural Information Processing Systems. NeurIPS, 2021. 30127–30139.

16 Bongers S, Forré P, Peters J, *et al*. Foundations of structural causal models with cycles and latent variables. The Annals of Statistics, 2021, 49(5): 2885–2915.

17 Forré P, Mooij JM. Markov properties for graphical models with cycles and latent variables. arXiv:1710.08775, 2017.

18 Bareinboim E, Correa JD, Ibeling D, *et al*. On Pearl's hierarchy and the foundations of causal inference. In: Geffner H, Dechter R, Halpern JY, eds. Probabilistic and Causal Inference: The Works of Judea Pearl. New York: ACM, 2022. 507–556.

19 Peters J, Janzing D, Schölkopf B. Elements of Causal Inference: Foundations and Learning Algorithms. Cambridge: The MIT Press, 2017.

20 Zhang S, Yao LN, Sun AX, *et al*. Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys, 2019, 52(1): 5.

21 Fang H, Guo GB, Zhang DN, *et al*. Deep learning-based sequential recommender systems: Concepts, algorithms, and evaluations. Proceedings of the 19th International Conference on Web Engineering. Daejeon: Springer, 2019. 574–577.

22 Rendle S, Freudenthaler C, Gantner Z, *et al*. BPR: Bayesian personalized ranking from implicit feedback. Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Montreal: AUAI Press, 2009. 452–461.

23 Rendle S, Freudenthaler C, Schmidt-Thieme L. Factorizing personalized Markov chains for next-basket recommendation. Proceedings of the 19th International Conference on World Wide Web. Raleigh North: ACM, 2010. 811–820.

24 He XN, Liao LZ, Zhang HW, *et al*. Neural collaborative filtering. Proceedings of the 26th International Conference on World Wide Web. Perth: International World Wide Web

Conferences Steering Committee, 2017. 173–182.
25 Hidasi B, Karatzoglou A, Baltrunas L, *et al*. Session-based recommendations with recurrent neural networks. Proceedings of the 4th International Conference on Learning Representations. San Juan: ICLR, 2016.
26 Tang JX, Wang K. Personalized top-n sequential recommendation via convolutional sequence embedding. Proceedings of the 11th ACM International Conference on Web Search and Data Mining. Marina Del Rey: ACM, 2018. 565–573.
27 Yuan FJ, Karatzoglou A, Arapakis I, *et al*. A simple convolutional generative network for next item recommendation. Proceedings of the 12th ACM International Conference on Web Search and Data Mining. Melbourne: ACM, 2019. 582–590.
28 Kang WC, McAuley J. Self-attentive sequential recommendation. Proceedings of the 2018 IEEE International Conference on Data Mining. Singapore: IEEE, 2018. 197–206.
29 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
30 Mnih V, Kavukcuoglu K, Silver D, *et al*. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529–533. [doi: 10.1038/nature14236]
31 Pearl J. Causality. Cambridge: Cambridge University Press, 2009.
32 Sun F, Liu J, Wu J, *et al*. BERT4Rec: Sequential recommendation with bidirectional encoder representations from Transformer. Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing: ACM, 2019. 1441–1450.
33 Kumar A, Hong J, Singh A, *et al*. When should we prefer offline reinforcement learning over behavioral cloning? arXiv:2204.05618, 2022.
34 Meng ZQ, Liang SS, Fang JY, *et al*. Semi-supervisedly co-embedding attributed networks. Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver: NeurIPS, 2019. 32.
35 Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139–144. [doi: 10.1145/3422622]
36 Zha DC, Lai KH, Zhou KX, *et al*. Simplifying deep reinforcement learning via self-supervision. arXiv:2106.05526, 2021.
37 Zhou SJ, Dai XY, Chen HK, *et al*. Interactive recommender system via knowledge graph-enhanced reinforcement learning. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2020. 179–188.
38 Sutton RS, Barto AG. Reinforcement Learning: An Introduction. 2nd ed., Cambridge: MIT Press, 2018.

(校对责编: 孙君艳)