计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(10):53−60 [doi: 10.15888/j.cnki.csa.007547]

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

# 基于反事实学习及混淆因子建模的文章个性化推荐①

杨梦月[1,2], 何洪波[1], 王闰强[1]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
通讯作者: 何洪波, E-mail: hhb@cnic.cn

**摘　要**: 如今, 互联网推荐系统已经成为了一个热门话题, 自动化推荐极大程度上方便了人们的生活, 帮助人们从海量的信息当中寻找到最感兴趣的关键信息. 互联网上每时每刻都在产生新的文章信息, 已有的信息是一个非常庞大的数据集合, 这些被记录的大量数据能够帮助统计出用户偏好以及文章内容的受欢迎程度. 目前互联网上有许多种类的推荐系统, 他们综合考虑了用户特征, 文章特征. 基于互联网各大社交媒体上的数据, 现有的用户个性化推荐系统通过构建特定的模型对用户进行精准推荐. 目前, 推荐算法主要通过监督学习与在线学习的方法进行构建, 但这些方法进行个性化推荐的时候往往忽略了一个问题: 历史记录当中的推荐策略往往是部分观测数据, 具有分布不平衡的劣势, 通过现有的历史记录不能保证算法能够得到无偏的推荐结果, 也不能适应线上的环境以及推荐策略变化. 本文提出了一种基于反事实学习并考虑系统当中混淆因子的文章个性化推荐. 这种方法有更强的理论保证, 并且在实验结果当中也显示了比现有方法更加好的算法表现.

**关键词**: 推荐系统; 反事实学习; 因果推理; 混淆因子; 个性化推荐

引用格式: 杨梦月,何洪波,王闰强.基于反事实学习及混淆因子建模的文章个性化推荐.计算机系统应用,2020,29(10):53–60. http://www.c-s-a.org.cn/1003-3254/7547.html

## Counterfactual Learning in Article Recommendation with Confounder

YANG Meng-Yue[1,2], HE Hong-Bo[1], WANG Run-Qiang[1]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

**Abstract**: Nowadays, the Internet recommendation system has become a hot topic. Automatic recommendation has greatly facilitated people's life and helped people find the most interesting key information from the massive information. Now news information is generated every moment on the Internet, and the existing information is a very large data set, which can help to count the user preferences and popularity of news content. At present, there are many kinds of recommendation systems on the Internet. They comprehensively consider the characteristics of users and articles to be recommended. Based on the data on various social media on the Internet, they build models and can use these models for accurate personalized user recommendation. The existing recommendation system is usually a supervised learning system which takes a lot of user characteristics into account. These methods often ignore the following issue: the recommendation strategy in the history is often imbalance. Through the existing historical records, we cannot guarantee an unbiased result. So in this study, we propose a kind of personalized recommendation based on counterfactual learning. This method has stronger theoretical guarantee and also shows better algorithm performance than existing methods in the experimental

① 基金项目: 中国科学院"十三五"信息化建设专项 (XXH13504-04)
Foundation item: CAS Special Fund for Informatization Construction in 13th Five-Year Plan (XXH13504-04)
收稿时间: 2019-12-31; 修改时间: 2020-02-08; 采用时间: 2020-03-11; csa 在线出版时间: 2020-09-30

results.

**Key words**: recommendation system; counterfactual learning; causal inference; confounding factor; personalized recommendation

互联网文章推荐系统是一个帮助用户, 为用户提供文章阅读建议的平台, 推荐系统的存在已经成为了减轻用户信息负载, 实现文章个性化获取的一个重要方式. 互联网推荐系统近年来在各种商业领域得到了广泛的应用, 不仅仅用在了文章领域, 在社交网站, 电商网站等等网站都有大量的应用[1,2]. 可以说, 人们的日常网上生活已经离不开互联网推荐系统. 文章信息相较于其他信息来说, 具有更强的主题性, 其具有生命周期较短、访问记录稀疏、文本表示复杂的特点, 所以在其基础上的推荐系统构建相对来说更加复杂. 现有的推荐算法按照训练方法分类一般情况下主要分为两大种类, 第一类是使用在线学习 (online learning)[3–6] 的方法通过跟环境不断交互的方式进行各推荐单元的期望回报预估, 通过这种方法能够适应线上动态变化的环境, 并且能够执行更好的推荐探索策略. 这种方法具有很大的优势, 是因为他简单易行, 并且非常适合线上环境变化. 但其劣势也非常明显, 单一的特征处理结构不能满足其面对的更加复杂的异构数据, 并且增加模型的复杂度目前还没有合适的数学方法能够给出该算法所要使用的置信区间, 另外线上环境执行具有不确定性, 并且线下训练采样率非常低. 第二类是使用历史数据进行监督学习[7–10], 对当前的历史记录进行拟合并且计算所要预测请求的回报. 这类方法的优势是能够改变特征提取和组合的方式, 使其自动提取出来的特征对分类结果更有帮助, 常见的例如基于机器学习以及深度学习的推荐系统.

第一种方法需要在线环境进行交互迭代训练, 但是通常情况下, 使用真实环境进行训练的代价十分高昂, 并且迭代速度缓慢. 在真实的工业场景中, 一般采用构建离线模拟器的方式. 但是构建模拟器对于文章预测并不可行, 因为文章预测没有一定的规律性. 第二种方法不需要线上环境的交互过程, 监督学习的方法往往通过大量的历史数据进行学习, 但在从历史数据中学习会产生一定的问题. 那就是历史数据往往不能包含所有的决策情况, 所以, 使用历史数据预测往往会产生分布偏移的情况, 使得决策策略更加倾向于历史数据当中存在的事件. 模拟线上环境的实现对于文章推荐场景来说较为困难, 所以本文将关注基于历史数据信息的学习. 本文需要解决现有的从历史数据中直接进行监督学习会产生策略分布偏差的问题, 并且解决存在在无法直接观测到的混淆因子的情况下, 模型产生的偏差.

针对以上问题, 本文引入了因果推断的思想, 通过反事实学习对环境中未出现的情况进行推理, 进行分布建模, 从而能够避免样本量涵盖范围不足产生的分布偏差. 本文同时考虑推荐系统中含有大量的对结果产生影响的隐性因素 (在这里把它称为混淆因子 (confounder)), 提出了混淆因子存在情况下的策略学习方法.

本文主要贡献如下:

(1) 本文针对现有的各类深度学习推荐系统方法的分析结构和目标特点, 构建出深度反事实学习框架, 降低推荐系统依赖于历史策略产生的偏差.

(2) 本文提出了使用了变分自编码机对推荐系统中混淆因子进行推断, 使得模型能够在这些混淆因子的影响下构建出一个无偏倾向分数, 降低模型偏差.

## 1 文章个性化推荐方法

### 1.1 基于监督学习的个性化推荐

监督学习的推荐系统一般会收集大量的历史推荐数据, 其中包括用户被观测到的特征, 以及大量的文章特征. 其中用户需求一般被表示为一个向量, 这个向量中包含一些用户特征, 通常被表示为稀疏特征. 同时一般用户会有一个随机推荐的项目范围, 这个范围通常情况下在针对用户请求的召回阶段是非常巨大的, 通常这时会使用一些简易的相似度匹配方法从超大样本空间中选取出较为相似的小样本, 针对小样本空间进行精确推荐. 而我们的模型就工作在这一精确排序的层面. 如下给出一些符号表示:

以$X \in \mathbb{R}^{b*d}$作为用户样本集合, 其中 $b$ 为样本数量,

$d$ 为样本空间维数大小. 同时用户候选推荐项目集合作为$A = \{a_1, a_2, \cdots, a_k\}$, 其中$a_k \in \mathbb{R}^s$, $s$ 表示文章向量的向量空间大小. 与此同时, $Y \in \mathbb{R}^b$则代表数据集中一个请求和推荐之后得到的用户反馈, 在文章推荐系统当中, 主要研究文章事件的点击率 CTR. 一般情况下, 一个事件可以被看作是用户请求和推荐商品以及用户反馈的三元组 $(x,a,y)$.

现有的监督学习模型主要分为两种类别, 其一是通过设计单一模型的最大化提取当前有效特征嵌入层, 通常情况下根据推荐项目的不同有利用组织单一的多层感知机 (MLP), 卷积神经网络 (CNN), 循环神经网络 RNN 神经网络, 或与条件随机场 CRF 的模型构建的神经网络, 这种网络的设计需要根据数据的表现和结果的偏差进行微调, 大多数情况下一种模型并不具有推广性.

其二是使用多模型融合的方法进行联合特征提取, 比如经典的 wide & deep 模型, 这种方法结合了传统机器学习和深度学习的特征提取优势设计嵌入层融合, 从而能够得到更具有记忆性和泛性的隐藏层特征, 得到偏差更小的分类模型. 常用的融合方式有 CNN 与 RNN 网络结构融合, CNN 与自动编码器 (Autoencoder) 的融合等[11−13], 这类方法在各类问题上都展现出其巨大的优势, 主要是因为其能够利用各大网络结构的设计特点针对原始数据进行不同程度上的特征提取, 从而提升模型的无偏性. 虽然一个特定的融合模型不能推广到各类场景, 但是该类方法的思想具有很强的应用价值.

如果说模型设计可以被看作一种人工的先验知识, 这个先验知识限制了假设空间的大小, 那么针对各类场景问题的特定模型设计就格外重要.

监督学习中使用的损失函数一般根据任务不同稍有偏差, 在文章推荐中由于一般将 CTR 作为评价指标, 并且一般情况下使用较差熵损失函数:

$$L = \frac{1}{N}\sum_{n} y \log \hat{y}$$

其中, $N$ 为采样样本空间大小, $y$ 为样本标签真值, $\hat{y}$为当前模型预测值.

这种方法的优势是其可用性强, 简单, 容易操作. 但存在一些缺陷: 历史数据的推荐在实际线上系统应用中存在偏差, 因为针对一个在特定时间出现的用户请求, 历史数据当中往往只包含其中一种推荐结果, 换句话说, 在其他情况下的推荐结果是未知的, 另外由于本身收集到的数据并不能保证其不会偏向于某种特定的非最优策略分布, 所以使用监督学习往往会产生训练偏差, 但这些偏差很多时候都很容易被忽略.

### 1.2 基于在线学习的个性化推荐

在线学习是在当推荐系统需要一定的探索性时的解决方案, 算法通过直接与环境交互推荐得到的反馈结果进行更新.

在线学习的主要方法是基于置信区间上界的方法, 该方法提供了一种探索策略, 即在算法早期由于训练样本较少, 所以此时针对每一种推荐的参数在早期置信度较低. 该置信度可以使用霍夫丁不等式来具体衡量. 早期的在线学习仅仅考虑伯努利实验, 而不考虑用户上下文信息, 在 2010 年, Li 等[3] 提出了一种方法可以将在线学习算法扩展到上下文形式, 主要使用了线性模型对上下文进行了回馈函数建模, 并且使用了阿祖玛不等式对其估计上界进行限制, 从而达到一定的探索效果, 该算法取得了巨大的成功, 并且被工业界广泛使用.

在线学习主要分为以下步骤:

1) 首先对系统中的每个可推荐项目, 都设置同样的初始化参数模型, 并进行随机探索推荐一个推荐项目.

2) 得到该推荐项目针对该用户的推荐回报结果, 并用推荐结果更新算法中参数.

3) 算法计算每一个可推荐项目的期望回报结果, 并且预估其置信上界, 选择当前置信上界值最高的可推荐项目进行推荐.

4) 得到推荐结果之后, 返回第 2) 步.

从以上步骤可以看出, 在线学习本质上是非监督化的训练, 在没有监督信息的情况下, 对回报函数的预估和拟合会出现一定的偏差, 会出现过估计的情况, 即其预估值会越来越高, 导致推荐结果出现偏差, 其次是该方法必须需要在线执行, 离线采样率非常低, 但在线训练的代价通常很高, 因为线上系统面向真正的用户, 所以在算法早期会有大量的不确定推荐因素. 造成用户损失以及公司损失.

### 1.3 基于反事实学习的个性化推荐

监督学习的方法虽然已经有不错的效果, 但是其需要大量的训练数据涵盖出各种不同的推荐情况. 实

际情况下, 往往线上数据保留下来的记录只是遵循一种或几种推荐策略, 并不能涵盖出所有的推荐结果, 所以如果在这种样本上训练的话, 监督学习得到的推荐策略就有一定的倾向性. 比如, 当前策略下, 正负样本数量差距较大, 那么期预测的正负结果数量差距可能会更大, 原因是算法更倾向于对不确定的上下文执行更小可能出错的推荐, 从而算法更容易被历史策略先验影响.

贝叶斯网络之父 Pearl 在 1986 年提出使用因果实现真正的机器智能, 其思想核心就是发现数据之间的因果关系[14–16]. 在推荐系统领域, 传统机器学习只能根据数据之间的相关性发现其中的相关关系, 但是学习出相关关系之后, 并不能给出一个准确推荐结果. 举个例子, 阳光照射在物体上时, 会在地面投射出影子, 是因为阳光的位置和物体的形态位置决定了影子的形状和大小. 基于相关关系的学习仅仅只能发现三种事物相关性, 但是在探索次数有限的情况下. 假如想得到一种特殊形态的影子, 使用监督学习很难在历史记录当中学习出来该如何摆放物体. 而基于因果的反事实学习是在假设已经存在一个因果关系结构, 通过控制变量的方法得到每次的影子投射结果, 就可以对历史中不存在的情况进行分布建模, 从而能得到无偏估计, Swaminathan 等首先定义了在历史记录中进行反事实学习的机器学习框架[17], 并且针对其模型结构在深度学习上进行推广[18] 以及进一步归一化[19,20]. 此外, 反事实学习也被扩展在表示学习以及日志学习领域[21,22].

## 2 基于因果反事实推理的个性化推荐

针对监督学习训练模式在数据策略缺失较多的情况下, 本文引入反事实学习作为训练策略. 反事实策略能够帮助算法通过重要性采样的方式发现完善在历史情况下没有被观测到的策略结果分布, 即反事实分布.

参考推荐系统训练的一般过程, 本段将给出一些符号表示: 使用$Y(a):a\in A$表示不能直接观测到的潜在环境回报方程, 表示当前选择文章 $a$ 时环境给出的 CTR 预估结果. 使用$D_t=\{D_{t1},D_{t2},\cdots,D_{tk}\}$,其中$D_{ta}$表示观测到第 $t$ 轮时, 历史策略 $a$ 的推荐指示, 既如果 $D_{ta}=1$, 则表示在时间步 $t$ 时 $a$ 被选择. 另外, 本文将历史数据的策略看作一个固定的策略, 这个策略分布可以使用$p_t=\{p_{t1},p_{t2},\cdots,p_{ta}\}$表示, 训练策略主要被表示为如下步骤:

1) 假设数据集当中收集的事件三元组在每一个训练时间步中满足独立同分布 (i.i.d.) 采样, 并且其概率服从一个特定的概率分布.

2) 在每步训练中使用神经网络估计得到当前步的历史策略倾向分数$\hat{p}_t$, 通过该分布进行反事实重要性采样, 从而达到反事实学习的目的.

$$\hat{R}=\frac{1}{T}\sum_t\sum_A y_t D_{ta}\frac{\pi(a|x_t)}{\hat{p}_t} \tag{1}$$

3) Swaminathan 等在 2015 年提出[19], 该方法存在一定的倾向性过拟合, 因为在进行该算法时, 需要估计历史策略分布值, 当对该值不加限制的情况下, 作为分母, 如果该值过小, 可能会导致严重的训练不稳定性. 为防止倾向性过拟合, 他提出对当前计算结果进行自归一化.

$$\hat{R}_N=\frac{\hat{R}}{\frac{1}{T}\sum_t\sum_A D_{ta}\frac{\pi(a|x_t)}{\hat{p}_t}}$$

同样, 该方法仍然也会有一定的不准确性, 在系统不知道$\hat{p}_t$先验值的情况下, 必须首先估计历史记录里面的策略分布, 而对该分布的估计必须要基于人工设置的算法模型, 相当于该算法模型本身就带有一定的先验性, 导致结果出现偏差. Narita 等[23] 使用机器学习模型对该策略分布进行预估, 但是仍然不能达到较好的效果, 因为添加先验进入该环境系统中, 会导致结果发生倾向偏差.

### 2.1 使用因果关系对历史策略建模方法

为了防止在推荐系统中出现偏差, 本文将反事实引入用于构建无偏的历史策略分布. 通过反事实学习, 虽然已经可以从历史数据中学习到无偏策略. 但此时仍然需要估计历史策略的倾向性分数, 即需要估计$\hat{p}_t$的值, 如果此时采用机器学习监督算法去直接估计计算该值, 则仍然会在结果中产生偏差, 其中的原因是, 系统中观测到的仅仅是当前用户的信息, 推荐项目的信息, 以及最终的结果, 这种观测未考虑到混淆因子存在的情况, 并且每个不同的模型都有其预测特点, 不能保证其无偏性.

混淆因子是指在系统中无法观测到的影响因素, 混淆因子最早源于辛普森悖论, 是指当对具有较大的肾结石患者执行疗程 A 时, 成功治愈的可能性较高, 对具有较小肾结石患者执行 B 时, 同样也有较高的治愈概率. 但某些情况下, 疗程 A 对于较小肾结石者比疗

程 B 更加有效. 之所以产生该悖论的原因是系统中存在不能直接观测的混杂因素, 该混淆因素直接影响到状态的观察, 决策, 以及结果的反馈.

前人提出的估计历史策略倾向分数的方法, 大多为使用某中机器学习模型, 但该做法具有较强的先验性. 会导致策略学习产生倾向一种模型策略结果的偏移, 并且在使用该方法的过程中, 由于未考虑环境中混淆因子的存在, 及其有可能产生较大的偏差, 所以为了防止该偏移的产生, 本文将混淆因子进行整体建模.

文章推荐系统中的因果关系可以被构建成两种形式.

第一种是在不考虑混淆因子的情况下, 在模型中, 当前观测到的状态 $X$, 策略 $A$ 与结果 $Y$ 的因果关系如图 1 所示, 可以观察到当前观测回报结果仅仅由当前所选择的推荐项目 $A$ 影响. 不考虑混淆因子的存在. 也就是之前大部分研究工作所采用的因果关系图.

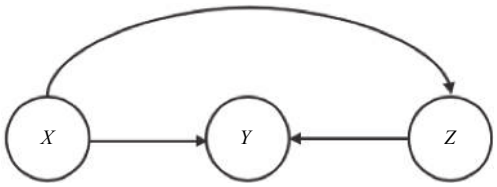


图 1 假设在不存在混淆因子情况下的因果关系图

第二种考虑了系统中混淆因子 $U$ 的存在下的因果关系, 如图 2 所示, 可以看出系统中的观测 $X$, 推荐项目 $A$, 以及最终结果 $Y$ 都受到环境当中的混淆因子影响. 而本文主要研究的就是在这种情况下的建模问题.

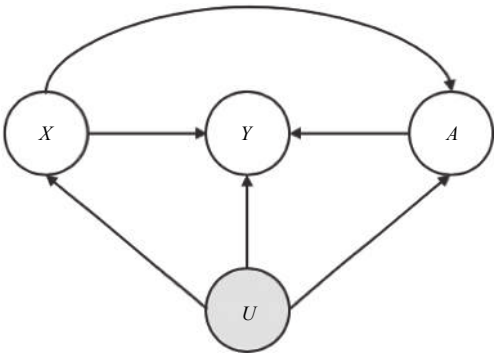


图 2 假设系统中存在未知的混淆因子变量

### 2.2 基于混淆因子建模的倾向分数建模

在因果推断的框架下, 混淆因子被看作环境中的隐变量, 由于隐变量属于系统中无法观测的变量, 所以可以根据其余可观测变量对其进行推断. 本文提出的模型基于生成模型中的变分自编码器框架对隐变量执行推断.

首先, 需要考虑潜在回馈方程, 即$Y(x) = h(x,a)$, 如果潜在的混淆因子可以使用, 则可以采用如下关系式对回馈方程进行建模:

$$Y(x) = h(x,a) = \mathbb{E}[y_a|x] = \mathbb{E}[y|a,u]$$

式中存在未知隐变量 $u$, 在系统中表现为无法直接观测以及施加影响的混淆因子. 所以需要对该隐变量进行推断, 即推断模型$q_\phi(u|x,y,a)$, 在本文中, 基于变分自编码机 (VAE) 的思想, 将混淆因子 $u$ 作为隐变量, 采用变分自编码机的编码器进行推断.

变分自编码机由 Kingma 在 2013 年提出[24], 其假设系统的观测值是由少数的隐变量控制, 其假设隐变量满足一个各向同性的标准高斯分布[25,26]. 模型主要分为编码器和解码器, 编码器用于将生成分布拟合成为标准高斯分布, 以达到解耦目的. 解码器根据从标准高斯分布中采样出的隐变量获得新的观测数据. 一般认为, 变分自编码器相较于其他的生成模型具有更高的泛化性能.

本文采用生成模型联合训练的方式, 隐变量推断过程参考[27], 同时对隐变量和回馈函数进行推断和学习. 该学习过程参考图 2 因果性所示, 对于一个观测到的推荐三元组记录 $(x,a,y)$ 来说, 其对数似然如下所示:

$$\log p_\theta(x_i,y_i,a_i) = \int \log p_\theta(x_i,y_i,a_i|u_i)p(u_i)du \quad (2)$$

在该方法中采用了基于隐变量推断的生成模型. 该模型的主要方法是采用变分自编码器的训练方法, 该模型的训练目标即为最大化似然函数 (式 (2)). 对任意变量有如下生成模型方程, 在本文中, 假设 $x$, $a$, $y$ 的生成都能由非线性函数控制, 即:

$$\begin{cases} x_i p_\theta(x|u_i) \\ a_i p_\theta(a|u_i,x_i) \\ y_i p_\theta(y|u_i,a_i,x_i) \end{cases}$$

其中, $\theta$可以为任何非线性函数的参数, 而在本文中该函数参数为神经网络中的参数, 由神经网络优化拟合.

假设混淆因子 $u$ 的先验分布 $p(u)$ 满足一个各向同性的多元高斯分布即$u \sim N(0,I)$, 此时, 根据 VAE 假设, 在本文中, 生成的 $u$ 的在观察数据三元组 $(x, y, a)$ 条件下的条件概率满足一个多元高斯分布:

$$p_\phi(u|x,y,a) = N(\mu,\sigma)$$

公式中$\phi$的表示变分编码器参数. 根据之前的 $x$, $y$, $z$ 的生成模型, 从混淆因子中解码得到的三元组可以被表示为如下表达式:

$$\log p_\theta(x_i,y_i,a_i|u_i) = \log p_\theta(y_i|u_i,a_i) + \log p_\theta(a_i|u_i) + \log p_\theta(x_i|u_i)$$

由于在该生成模型中, $u$ 的后验概率由历史记录当中的所有观察值三元组共同决定, 所以一定程度上避免了只考虑一项因素所产生的混淆偏差.

在生成模型变分推断自编码器 (VAE) 中, 变分自编码器主要采用两层结构——编码器和解码器.

编码器的作用主要是通过训练的三元组数据得到隐变量 $u$ 的后验方程. 本文假设 $u$ 的先验概率为一个各向同性的标准高斯分布, 所以编码其采用 $u$ 的后验概率与 $u$ 的先验产生的偏差作为误差, 该误差使用分布距离 KL 散度衡量.

解码器的作用是通过从一个标准高斯分布中采样出隐变量 $u$, 并且根据该隐变量 $u$ 的值还原出观察值, 其误差采用经验误差, 即最小化由观测值通过编码器产生的隐变量经由解码器生成的新观测值以及训练观察值之间的经验误差.

采用监督经验误差作为解码器误差, 优化目标为置信下界 (*ELBO*). 该置信下界可以被写作如下形式:

$$ELBO = \sum_{(x,a,y)\in obs} E_{q_\phi(u|x,a,y)}[\log p_\theta(x,a,y|u)] - KL[q_\phi(u|x,a,y)||p(u)]$$

注意到, 观测到的因果关系数据, 即历史记录, 仅仅记录了一部分的推荐结果, 为了得到无偏倾向分数 $\hat{p}(a|x)$. 在变分自编码机当中推断后验概率的同时, 需要在同时估计历史策略分布$a_i \sim q_\phi(a|x_i)$. 所以本文提出将其加入损失函数当中, 该损失函数是由 Louizos [28] 首次形式化成如下形式:

$$L = ELBO + \sum_{(x,a)\in obs} \log q_\phi(a|x) \tag{3}$$

在该式中, $q_\phi(a|x_i)$为一个分类模型的似然函数, 通过高斯分布拟合, 通过似然函数的最大化更新参数, 预测结果为各分类的分数. 其中$(x_i,y_i,a_i)$为数据集当中的三元组, 使用最小化该损失函数中的可训练参数$\theta$优化生成函数的结果, 即解码器. 并且通过最小化该方程中的可训练参数$\phi$用于优化倾向分数, 最后可以将其用于反事实推断.

## 3 实验及结果分析

### 3.1 实验设置

在推荐系统中对于在线方法直接验证是一件比较困难的事, 因为没有直接的标签可以用于参考. 所以在这里直接使用推荐系统中的累积回报对最终反事实学习进行结果验证.

本文实验基于真实的文章推荐场景, 数据选用 Yahoo！Today’s Module 公开新闻推荐数据集. 数据采集于 2009 年五月的 Yahoo! front page, 其中包括了用户信息, 新闻推荐可选择范围, 文章推荐结果作为一个事件. 数据一共包含了 468 万个事件 ($x$,$a$,$y$), 其中 $x$ 包含用户的 6 维特征, 是由原始特征提取出的主题特征, 每一维特征都在范围 [0, 1], 文章特征同样采用 6 维主题特征范围同样在 [0, 1]. 反馈值 $y$ 代表了当前推荐项目的点击情况, 其中 1 代表当前用户点击了该新闻, 0 代表用户没有点击该新闻数据. 实验中随机筛选了其中推荐最多的 6 种文章, 并对数据进行了充分混合, 以防选择顺序对结果产生影响. 最终实验选择一共 128 万个事件的集合. 其中前 100 万个事件作为训练集, 剩余的 28 万个事件当作测试集. 实验采用的对照组分为 3 类.

1) 使用监督学习估计的方法

将用户特征 $x$ 与推荐新闻的新闻特征 $a$ 进行拼接当作训练数据, 将回报 (CTR) 当作训练标签 $y$, 对其进行监督学习. 实验中一共尝试了四种监督学习方法, 其中包括: 对率回归 (Logistic regression), 梯度提升机 (Gradient boosting machine), 随机森林 (random forest).

2) 使用监督学习方法估计倾向性分数, 使用反事实学习的方法

使用对率回归, 梯度提升机和随机森林得到的策略作为倾向性得分, 使用反事实学习的方法得到最终策略, 实验结果通过测试集拒绝采样进行验证.

3) 使用因果混淆因子建模的反事实训练

使用对混淆因子建模的方法, 计算出倾向性得分, 通过该倾向性得分进行反事实学习, 最终得到结果.

### 3.2 反事实学习神经网络实现细节

本文假设隐变量符合标准高斯分布$N(0,1)$, 其中隐变量维度设置为 8 维. 编码器中的$采用单层全联接作为隐层模型, 隐层维数为 8. 解码器中的 3 类三元组估

计同样采用单层全联接作为隐层模型, 其中$\log p_\theta(y_i|u_i,a_i)$隐层维数为 16, $\log p_\theta(a_i|u_i,x_i)$、$\log p_\theta(x_i|u_i)$分别为 16, 8. 由于模型需要同时估计策略模型$q_\phi(a_i|x_i)$似然函数中的高斯分布参数, 隐层维数为 16. 为了更好地提升 GPU 运算效率, 隐藏层维数设置为 2 的倍数, 另外为避免过拟合现象产生, 隐层维数与解码器均与输入数据维数接近, 均设置为 8 与 16.

### 3.3 实验结果分析

实验结果图 3 所示, 可以看出对率回归表现结果欠佳, 是因为线性模型的假设空间的复杂度不够高, 从而不容易捕捉更加复杂的模式, 所以拟合度会有一些欠缺, 尤其是在数据量过大的情况下, 需要的假设空间会更大. 所以其表现相对于梯度提升机和随机森林较差.

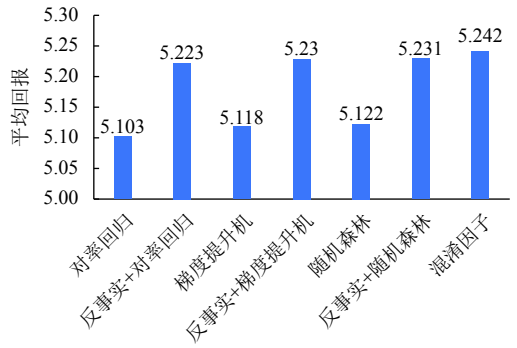


图 3 各对照组实验的平均回报

梯度提升机和随机森林都是基于树结构的方法, 在该实验中, 对与两种方法试验均采用 50 个估计器以及 15 层树深度以达到最佳的精准率并防止过拟合. 得到的结果相差并不多, 可见这两种方法的复杂度要比线性模型高出许多, 所以表现也更加优秀.

可以从图 3 中看出, 采用反事实学习的策略, 得到的效果比直接使用监督学习高许多, 由于在反事实学习中倾向性分数作为分母, 为避免过拟合, 在实验过程中对该倾向性分数做了一定的限制, 要求其不为 0 且与当前策略的比值不低于 0.05, 不高于 0.95. 但也可以看出, 虽然采用反事实学习虽然能提高性能, 但是也受到倾向性分数的影响, 倾向性分数较高的, 通过反事实学习也能得到更高的分数.

实验同样也测试了本文提出的算法, 该算法得到了最佳的结果. 可以看出本文提出的基于混淆因子的反事实学习算法效果最好. 这是因为使用了因果建模之后, 估计出来的历史策略包含更大的范围, 所以能够拥有更高的无偏性, 降低了系统出现的偏差.

## 4 结论与展望

通过实验发现了当对倾向性分数使用混淆因子建模, 有利于反事实学习得到最佳结果, 实验结果证明, 本文提出的基于因果混淆因子的反是学习具有更低的偏差. 本文首先对推荐系统中存在的因果关系进行建模, 此外, 本文使用了隐变量推断的方法对混淆因子的分布进行推断, 并同时训练出生成模型, 从而得到了反事实学习中的倾向性分数. 该倾向性分数作为反事实学习重要性采样的分母进行训练. 结合因果关系中混淆的方法, 是一种创新的鲁棒性的方法. 未来的研究将采用一些更加复杂的主题数据, 并且在该方法解决过拟合上做更多的探索.

## 参考文献

1 Wang GX, Liu HP. Survey of personalized recommendation system. Computer Engineering and Applications, 2012, 48(7): 66–76.
2 刘辉, 郭梦梦, 潘伟强. 个性化推荐系统综述. 常州大学学报 (自然科学版), 2017, 29(3): 51–59.
3 Li LH, Chu W, Langford J, *et al*. A contextual-bandit approach to personalized news article recommendation. Proceedings of the 19th International Conference on World Wide Web. New York, NY, USA. 2010. 661–670.
4 Li LH, Chu W, Langford J, *et al*. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. New York, NY, USA. 2011. 297–306.
5 Abbasi-Yadkori Y, Pál D, Szepesvári C. Improved algorithms for linear stochastic bandits. Proceedings of the 24th International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2011. 2312–2320.
6 王高智, 肖菁. 基于内容和最近邻算法的多臂老虎机推荐算法. 华南师范大学学报 (自然科学版), 2019, 51(1): 120–127.
7 赵鹏, 耿焕同, 王清毅, 等. 基于聚类和分类的个性化文章自动推荐系统的研究. 南京大学学报 (自然科学), 2006, 42(5): 512–518.
8 辛菊琴, 蒋艳, 舒少龙. 综合用户偏好模型和 BP 神经网络的个性化推荐. 计算机工程与应用, 2013, 49(2): 57–60, 96. [doi: 10.3778/j.issn.1002-8331.1109-0547]
9 杨永健. 基于模糊认知图和人工神经网络的个性化推荐算

法研究. 天津职业院校联合学报, 2009, 11(5): 54–56.
10 马华, 王清, 韩忠东, 等. 决策树分类算法在个性化图书推荐中的应用. 软件, 2012, 33(8): 100–101, 104. [doi: 10.3969/j.issn.1003-6970.2012.08.027]
11 周朴雄, 张兵荣, 赵龙文. 基于 BP 神经网络的情境化信息推荐服务研究. 情报科学, 2016, 34(3): 71–75.
12 邹润. 基于模型组合算法的用户个性化推荐研究 [ 硕士学位论文 ]. 南京: 南京大学, 2014.
13 祝婷, 秦春秀, 李祖海. 基于用户分类的协同过滤个性化推荐方法研究. 现代图书情报技术, 2015, 31(6): 13–19.
14 Didelez V, Pigeot I. Judea Pearl: Causality: Models, reasoning, and inference. Politische Vierteljahresschrift, 2001, 42(2): 313–315. [doi: 10.1007/s11615-001-0048-3]
15 Louizos C, Shalit U, Mooij J, *et al*. Causal effect inference with deep latent-variable models. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2017. 6449–6459.
16 Van Den Broek P. Causal reasoning and inference making in judging the importance of story statements. Child Development, 1989, 60(2): 286–297. [doi: 10.1111/j.1467-8624.1989.tb02715.x]
17 Swaminathan A, Joachims T. Counterfactual risk minimization: Learning from logged bandit feedback. Proceedings of the 32nd International Conference on Machine Learning. Lille, France. 2015. 814–823.
18 Joachims T, Swaminathan A, De Rijke M. Deep learning with logged bandit feedback. International Conference on Learning Representations. 2018. 1–12. https://openreview.net/pdf?id=SJaP_-xAb
19 Swaminathan A, Joachims T. The self-normalized estimator for counterfactual learning. Advances in Neural Information Processing Systems. 2015. 3231–3239. http://papers.nips.cc/paper/5748-the-self-normalized-estimator-for-counterfactual-learning.pdf
20 Swaminathan A, Joachims T. Batch learning from logged bandit feedback through counterfactual risk minimization. Journal of Machine Learning Research, 2015, 16(52): 1731–1755.
21 Johansson F, Shalit U, Sontag D. Learning representations for counterfactual inference. Proceedings of the 33rd International Conference on Machine Learning. New York, NY, USA. 2016. 3020–3029.
22 Agarwal A, Basu S, Schnabel T, *et al*. Effective evaluation using logged bandit feedback from multiple loggers. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA. 2017. 687–696.
23 Narita Y, Yasui S, Yata K. Efficient counterfactual learning from bandit feedback. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 4634–4641.
24 Kingma D P, Welling M. Auto-encoding variational Bayes. arXiv preprint arXiv: 1312.6114, 2013.
25 Huszár F. Variational inference using implicit distributions. arXiv preprint arXiv: 1702.08235, 2017.
26 Higgins I, Matthey L, Pal A, *et al*. Beta-VAE: Learning basic visual concepts with a constrained variational framework. ICLR, 2017, 2(5): 6.
27 Zhu FJ, Lin AD, Zhang GQ, *et al*. Counterfactual inference with hidden confounders using implicit generative models. Proceedings of the 31st Australasian Joint Conference on Artificial Intelligence. Wellington, New Zealand. 2018. 519–530.
28 Louizos C, Shalit U, Mooij J, *et al*. Causal effect inference with deep latent-variable models. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2017. 6446–6456.