ICML 2019 | 强化学习用于推荐系统，蚂蚁金服提出生成对抗用户模型

[机器之心](javascript:void(0);) 今天

**论文：Generative Adversarial User Model for Reinforcement Learning Based Recommendation System**

**作者：Xinshi Chen、Shuang Li、Hui Li、Shaohua Jiang、Yuan Qi、Le Song**

**机器之心编译**

**参与：李诗萌、shooting**



论文地址：https://arxiv.org/pdf/1812.10613.pdf

将强化学习用于推荐系统，能更好地考虑用户的长期效益，从而保持用户在平台中的长期满意度、活跃度。但是，强化学习需要大量训练样本，例如，AlphaGoZero [1] 进行了 490 百万局模拟围棋训练，Atari game 的强化学习在电脑中高速运行了超过 50 个小时 [2]。而在推荐系统的场景中，在线用户是训练环境，系统需要与用户进行大量的交互，利用用户的在线反馈来训练推荐策略。该过程将消耗大量交互成本、影响用户体验。在蚂蚁金服被 ICML 2019 接收的这篇论文中，作者们提出:

用生成对抗用户模型作为强化学习的模拟环境，先在此模拟环境中进行线下训练，再根据线上用户反馈进行即时策略更新，以此大大减少线上训练样本需求。

此外，作者提出以集合（set）为单位而非单个物品（item）为单位进行推荐，

并利用 Cascading-DQN 的神经网络结构解决组合推荐策略搜索空间过大的问题。

在推荐系统中应用强化学习（RL）有很大的研究价值，但也面临诸多挑战。在这样的配置中，在线用户是环境（environment），但是并没有明确定义奖励函数（reward）和环境动态（transition），这些都对 RL 的应用造成了挑战。

本文提出利用生成对抗网络同时学习用户行为模型（transition）以及奖励函数（reward）。将该用户模型作为强化学习的模拟环境，研究者开发了全新的 Cascading-DQN 算法，从而得到了可以高效处理大量候选物品的组合推荐策略。

本文用真实数据进行了实验，发现和其它相似的模型相比，这一生成对抗用户模型可以更好地解释用户行为，而基于该模型的 RL 策略可以给用户带来更好的长期收益，并给系统提供更高的点击率。

**推荐系统的 RL 挑战**

几乎对所有的在线服务平台来说，推荐系统都是很关键的一部分。系统和用户之间的交互一般是这样的：系统推荐一个页面给用户，用户提供反馈，然后系统再推荐一个新的页面。

构建推荐系统的常用方式是根据损失函数评估可以使模型预测结果和即时用户响应之间差异最小化的模型。换句话说，这些模型没有明确考虑用户的长期兴趣。但用户的兴趣会根据他看到的内容随着时间而变化，而推荐者的行为可能会显著影响这样的变化。

从某种意义上讲，推荐行为其实是通过凸显特定物品并隐藏其他物品来引导用户兴趣的。因此，设计推荐策略会更好一点，比如基于强化学习（RL）的推荐策略——它可以考虑用户的长期兴趣。但由于环境是与已经登陆的在线用户相对应的，因此 RL 框架在推荐系统设置中也遇到了一些挑战。

首先，驱动用户行为的兴趣点（奖励函数）一般是未知的，但它对于 RL 算法的使用来说至关重要。在用于推荐系统的现有 RL 算法中，奖励函数一般是手动设计的（例如用 ±1 表示点击或不点击），这可能无法反映出用户对不同项目的偏好如何 (Zhao et al., 2018a; Zheng et al., 2018)。

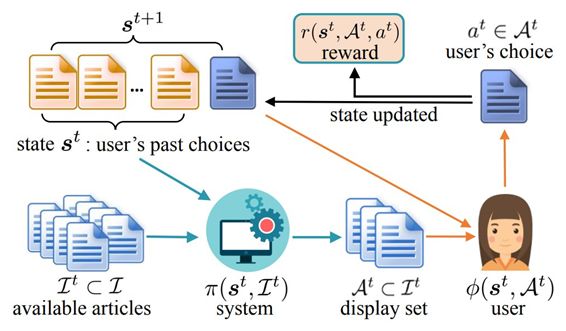
其次，无模型 RL 一般都需要和环境（在线用户）进行大量的交互才能学到良好的策略。但这在推荐系统设置中是不切实际的。如果推荐看起来比较随机或者推荐结果不符合在线用户兴趣，她会很快放弃这一服务。

因此，为了解决无模型方法样本复杂度大的问题，基于模型的 RL 方法更为可取。近期有一些研究在相关但不相同的环境设置中训练机器人策略，结果表明基于模型的 RL 采样效率更高 (Nagabandi et al., 2017; Deisenroth et al., 2015; Clavera et al., 2018)。

基于模型的方法的优势在于可以池化大量的离策略（off-policy）数据，而且可以用这些数据学习良好的环境动态模型，而无模型方法只能用昂贵的在策略（on-policy）数据学习。但之前基于模型的方法一般都是根据物理或高斯过程设计的，而不是根据用户行为的复杂序列定制的。

**解决方案**

为了解决上述问题，本文提出了一种新的基于模型的 RL 框架来用于推荐系统，该框架用统一的极小化极大（minimax）框架学习用户行为模型和相关的奖励函数，然后再用这个模型学习 RL 策略。



*图 1：用户与推荐系统之间的交互。绿色箭头表示推荐者信息流，橙色箭头表示用户信息流。*

本文的主要技术贡献在于：

* 开发了生成对抗学习（GAN）方法来模拟用户行为动态并学习其奖励函数。可以通过联合极小化极大优化算法同时评估这两个组件。该方法的优势在于：

（i）可以得到更准确的用户模型，而且可以用与用户模型一致的方法学习奖励函数；

（ii）相较于手动设计的简单奖励函数，从用户行为中学习到的奖励函数更有利于后面的强化学习；

（iii）学习到的用户模型使研究者能够为新用户执行基于模型的 RL 和在线适应从而实现更好的结果。

* 用这一模型作为模拟环境，研究者还开发了级联 DQN 算法来获得组合推荐策略。

动作-价值函数的级联设计允许其在大量候选物品中找到要显示的物品的最佳子集，其时间复杂度和候选物品的数量呈线性关系，大大减少了计算难度。

用真实数据进行实验得到的结果表明，从保留似然性和点击预测的角度来说，这种生成对抗模型可以更好地拟合用户行为。根据学习到的用户模型和奖励，研究者发现评估推荐策略可以给用户带来更好的长期累积奖励。此外，在模型不匹配的情况下，基于模型的策略也能够很快地适应新动态（和无模型方法相比，和用户交互的次数要少得多）。

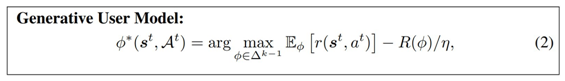
**生成对抗用户**

本文提出了一个模拟用户顺序选择的模型，并讨论了该模型的参数化和评估值。用户模型的建立受到了模仿学习的启发，模仿学习是根据专家演示来学习顺序决策策略的强大工具。研究者还制订了统一的极小化极大优化算法，可以根据样本轨迹同时学习用户行为模型和奖励函数。

**将用户行为作为奖励最大化**

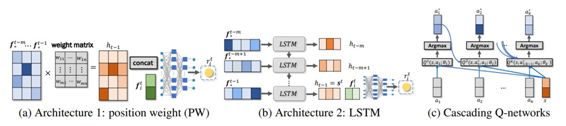
研究者还根据两个现实的假设模拟了用户行为：（i）用户不是被动的。相反，当给用户展示 k 个物品的集合时，她会做出选择，从而最大化自己的奖励。奖励 r 度量了她对一个物品的兴趣有多大或满意程度。另外，用户可以选择不点击任何物品。然后她得到的奖励就是没在无聊的物品上浪费时间。（ii）奖励不仅取决于所选物品，还取决于用户的历史。

例如，一个用户可能一开始对 Taylor Swift 的歌没什么兴趣，但有一次她碰巧听到了她的歌，她可能喜欢上了这首歌，于是开始对她的其他歌感兴趣。此外，用户在反复听 Taylor Swift 的歌之后可能会觉得无聊。换句话说，用户对物品的评价可能会随着她的个人经历而产生变化。模型公式为：



**模型参数化**

图 2 说明了模型的整体参数化。简单起见，研究者将奖励函数中所有参数表示为 θ，将用户模型中的所有参数集表示为 α，因此分别用符号 γ\_θ 和 φ\_α 表示。

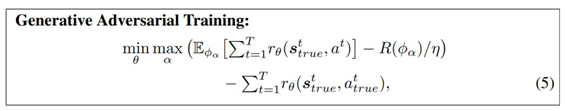


*图 2：由 (a) 位置权重 (PW) 或 (b) LSTM 参数化的模型架构。(c) 级联 Q 网络。*

**生成对抗训练**

在实践中，用户奖励函数 γ(s^t, a^t) 和行为模型 φ(s^t，A^t) 都是未知的，需要评估数据得到。行为模型 φ试图模仿真实用户的行为序列，该用户采取行为以最大化奖励函数 γ。与生成对抗网络相似：(i) φ 作为生成器，会根据用户的历史来生成她的下一个行为；(ii) γ 作为判别器，试图将行为模型 φ 生成的行为与用户的实际行为区分开来。因此，受 GAN 框架的启发，研究者通过极小化极大方法同时评估了 φ 和 γ。

更确切地说，给定某个用户的 T 个观测到的行为的轨迹 {a^1\_true, a^2\_true, . . . , a^T\_true} 及相应的所点击物品的特征 {f^1\_∗ , f^2\_∗ , . . . , f^T\_∗ }，研究者通过求解下面的极小化极大优化方法共同学习到用户的行为模型和奖励函数：

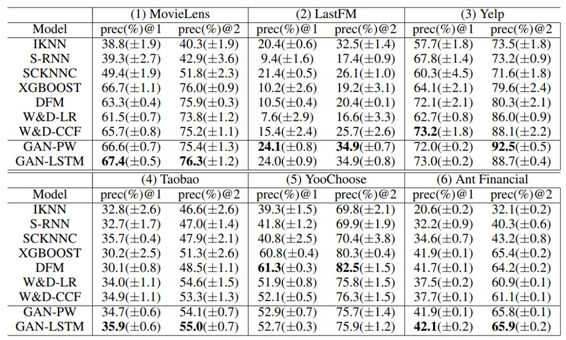


研究者用 s^t\_true 强调这是在数据中观测到的值。

**实验**

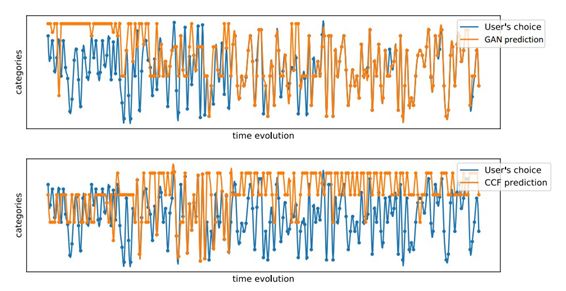
研究者用三组实验来评估其生成对抗用户模型（GAN 用户模型）和由此产生的 RL 推荐策略。该实验旨在解决下列问题：（1）GAN 用户模型可以更好地预测用户行为吗？（2）GAN 用户模型可以带来更高的用户奖励和点击率吗？（3）GAN 用户模型是否有助于降低强化学习的样本复杂度？

下面展示的是具有位置权重（GAN-PW）和 LSTM（GAN-LSTM）的 GAN 用户模型的预测准确率，表 1 结果表明 GAN 模型的性能显著优于基线。此外，GAN-PW 的性能几乎和 GAN-LSTM 一样，但训练效率更高。因此后续实验使用的是 GAN-PW（后面统称 GAN）。



*表 1：预测性能的比较，研究者在 GAN-PW 和 GAN-LSTM 中使用的是香农熵。*

另一个在 Movielens 上得到的结果很有趣，如图 3 所示。蓝色曲线表示用户随时间推移的实际选择。橙色曲线则是 GAN 和 W&D-CCF 预测的行为轨迹。



*图 3：对比用户选择的真实轨迹（蓝色曲线）、GAN 模型预测得到的模拟轨迹（上部分图中的橙色曲线）和 W&D-CFF 预测得到的模拟轨迹（下图中的橙色曲线）。Y 轴表示 80 个电影类别。https://mmbiz.qpic.cn/mmbiz_png/KmXPKA19gW8Zfpicd40EribGuaFicDBCRH6IOu1Rnc4T3W3J1wE0j6kQ6GorRSgicib0fmNrj3yzlokup2jia9Z0YVeA/640?wx_fmt=png&wxfrom=5&wx_lazy=1&wx_co=1*