齐智通讯 第 123 期

推断什么对于对齐的推荐系统是有价值的

Inferring what is valuable in order to align recommender systems

齐智通讯是每周出版物，其中包含与全球人工智能对齐有关的最新内容。在此处找到所有齐智通讯资源。特别是，你可以浏览此电子表格，查看其中的所有摘要。此处的音频版本（可能尚未启用）。请注意，尽管我在 DeepMind 工作，此齐智通讯仅代表我的个人观点，而不是我雇主的观点。

强调

从优化参与度到衡量价值（Smitha Milli 等人）（由 Rohin 总结）：本文试图为现有推荐系统创建一个比参与度更好的目标，这种方式可以应用于 Twitter 等现有公司。基本方

法是将要优化的变量（用户值）视为潜在变量，并使用概率推断来推断特定建议有价值的可能性。

通常，使用这种方法的主要挑战是指定观察模型：潜变量如何引起观察数据。对于 Twitter，这将要求你回答以下问题："如果用户不重视推文，那么用户点击"赞"按钮的可能性有多大？这是一个很难回答的问题，因为也许用户喜欢使用推文来停止对话，或者因为他们目前正在上瘾但实际上并不有价值，等等。

一种简单的启发式方法是获取两个数据集，我们知道一个数据集比另一个数据集具有更有价值的建议。然后可以将这些数据集之间的用户行为差异视为与价值的相关性。作者提供了一种从此类数据集推断观察模型的定量方法，在此不再赘述，因为它主要是启发式基线。一个明显的问题是，如果"最佳"数据集是通过优化（例如）点击产生的，则点击的增加可能是由于价值增加以外的原因，但是这种启发式方法会将整个增加归因于价值增加。

我们如何做得更好？本文的主要见解是，如果你拥有大量历史数据，则可以通过识别锚点来获得很多收益：一种反馈，当给出该反馈时可以明确地表明其潜在价值。在 Twitter 上，这被视为"很少见"（SLO）按钮：如果单击此按钮，则我们可以肯定地知道这是没有用的，无论用户采取了任何其他措施。然后，可以通过查看那些行为与锚点之间的联系来推断价值与其他行为（如推文）之间的联系，我们可以根据历史数据进行估算。

形式上，作者假定访问描述各种可能行为之间的关系的图（几乎所有行为都有潜在值 V 作为父项）。其中之一被标识为锚点节点A，对于该节点，假定 P(V = 1 | A = 1)是已知的，并且独立于所有其他变量。但是，P(V = 1 | A = 0)并不独立于其他变量：直观地，如果未单击 SLO 按钮，则我们需要回头看一下其他变量以估算值。

作者然后表明，在锚变量的一些合理假设下，如果你有一个历史数据集来估计 P(A, B)（其中 B 包含所有其他跟踪的行为），则无需指定观察模型 P(B|V) 的所有行为，你只需要为A的父母指定观察模型，即 P(parents(A)|V)。其他所有事物都是唯一确定的，使我们能够计算最终目标P(V|A, B)。（有关如何有效执行此操作的算法详细信息；有关详细信息，请参见论文。）在这种情况下，他们使用上面概述的启发式方法来估计 P(parents(A)|V)。

不幸的是，他们没有一个很好的方法来评估他们的方法：他们显然无法通过查看它是否带来更高的点击次数来评估它，因为整个目的是要放弃点击次数作为优化目标。（我认为在 Twitter 上进行用户研究是不可行的。）评估的主要形式是运行模型并报告学习到的概率，并证明它们似乎合理，而 Naive Bayes 模型输出的结果则不可行。

Rohin 的观点：我真的很喜欢这篇论文：似乎在尝试调整当今的推荐系统时确实遇到了麻烦，并且可能取得了实质性进展。

对于使用因果图形模型，我有些怀疑：如果你创建的模型具有一些条件独立性关系，而该条件独立性关系在实践中不成立，那么你的模型可能会有一些非常糟糕的推断。实际上发生了这种情况：当他们仅基于Twitter的UI元素对关系建模时，特别是没有对SLO对点击的依赖关系进行建模时，他们的结果就很糟糕，点击某个帖子被解释为该帖子没有价值的证据。。

如本文所述，我们可以放弃对贝叶斯网络的因果解释。这使我们能够在B中更多节点之间绘制边缘，以使我们的模型更具表现力，并部分防止这种类型的错误指定，同时还使我们能够表达更复杂的关系。例如，我认为（但不确定）在他们当前的图表中，喜欢和转发一条帖子将被视为独立的证据来源。如果我们在它们之间占据优势，你将获得模型学

习用户喜欢和喜欢的能力。转推一个帖子，该帖子的总和超过了喜欢和转发的贡献的总和。请注意，你仍然不能将所有内容都连接到锚点A，因为它们要求A不能有子代，并且如果你将父代添加到A，则必须通过上述启发式方法来估计它们，这可能不是很好。因此，你仍然需要为A建立条件独立性模型，并且当B变得更加复杂时，这可能会变得更加困难。

考虑到动机，这也是有道理的：由于想法是获取有关“流经锚点”到B中变量的值的信息，因此你似乎不必过多担心B中变量之间的关系，并且最好在它们之间建立任意复杂的关系模型。除了删除因果关系解释并在B中添加大量边缘之外，另一种选择是在图形中添加更多功能：例如，也许你还希望包括推文获得的全部赞次数，或者是否包括当前周末。你需要确保你的模型不会表现得那么强，以至于现在无法适合数据集，例如“ 9月24日发布的推文获得了38756个赞，这对爱丽丝来说就不值钱了”。但是，他们使用的数据集可能很大，

本文是否与本期齐智通讯的主题超级智能人工智能系统的对准有关？我不认为这是相关的，因为主要技术似乎涉及我们将（锚变量的）知识有效地硬编码到目标中，这种方式对推荐系统有意义，但可能不适用于更一般的系统系统。无论如何，我一直在强调它，因为我认为它特别新颖且有趣，它似乎可以解决现有人工智能系统的对准问题，并且它是研究谱系的一部分，我认为它与对准有关：如何为人工智能系统创建良好的目标。即使你只在乎超级智能系统的对准，似乎仍应遵循当今使用的技术以及其应用中出现的问题，

技术性人工智能对齐

学习人类意图

从隐式人类反馈中学习任务的EMPATHIC框架（崔玉uch，张启平等人）（由 Rohin 总结）：从人类反馈中学习的一个问题是，收集人类反馈的成本非常高。相反，我们可以从人类自动做出的面部表情中学习吗？本文表明答案是肯定的：他们首先在观察自治智能体的同时记录人类的反应，然后使用它来训练预测给定人类反应的奖励的模型。然后，他们将此模型转移到新任务。

人类也要学习：使用优化的人类输入来更好地进行人与人工智能交互（Johannes Schneider）（由 Rohin 总结）：人与人工智能交互中的大多数工作都集中在优化人工智能系统以使其与人的性能良好。但是，我们也可以教人如何与人工智能系统一起工作。本文是在一个简单的绘画游戏的背景下研究这个想法的，其中人类必须在一分钟内画出某个单词的草图，然后人工智能系统必须猜测该单词是什么。

作者开发了一种系统，对人类绘制的图像进行细微修改，以使其更易于识别-与对抗示例的设置非常相似。在用户研究中，向人们展示了一张图片，并要求重绘该图片。与更改后的图像一起显示时，与原始图像一起显示后，重新绘制的图像被更正确地分类，并且绘制时间更少。

阅读更多：人工智能安全需要社会科学家（AN＃47）

奖励学习理论

反对大脑中的经济价值的案例（Benjamin Y. Hayden等人）（由 Rohin 总结）（H / T Xuan Tan）：在神经经济学文献中，通常假设（基于过去的研究）假设大脑明确地计算一些价值观念，以便做出选择。本文认为这是错误的：大脑实际上没有明确地计算出值，而是直接学习产生作用的策略是合理的。

Rohin 的观点：如果你以前对逆向强化学习和类似技术持乐观态度，因为你认为它们可以推断出大脑计算出的相同价值观念，那么这似乎是一个重要的论点。但是，应该指出的是，作者并没有在争论大脑没有在优化某些特定的价值观念：只是没有明确地计算出这种价值观念。（类似地，强化学习的学习策略可以优化奖励功能，即使它们不需要显式计算每个状态的奖励来选择一个动作也是如此。）因此，你仍然可以希望大脑正在优化一些价值观念，而不是显式计算，然后使用逆强化学习恢复值的概念。

防止不良行为

安全意识强化学习（SARL）（Santiago Miret 等人）（由 Rohin 总结）：许多安全方法都依赖于从可信任的监督者（通常是人类）那里学习，包括迭代扩增（AN#40），辩论（AN#5）），养育（AN#53），委托强化学习（AN#57）和量化（AN#48）。本文将这种想法应用于避免在SafeLife环境中产生副作用（AN#91）。他们训练安全性智能体以最小化副作用分数，以用作可信任监督的智能体，然后训练常规强化学习智能体以优化奖励，同时惩罚偏离安全性智能体政策的情况。他们发现安全性智能体可以零缺陷地转移到新环境中，并且还有助于减少这些环境中的副作用。

处理智能体群组

多智能体社会强化学习可提高泛化能力（Kamal Ndousse 等人）（由 Rohin 总结）：我们之前已经看到，在探索困难的稀疏奖励环境中，进行专家演示以避免进行所有探索非常有用自己(1 (AN #14), 2 (AN #65), 3 (AN #9))

。但是，这假设演示程序在环境“外部”，而实际上我们希望将其建模为环境的一部分，例如在辅助博弈 assistance games (AN #69)。然后看起来像社交学习，智能体通过查看环境中其他智能体的提示来学习如何执行任务。

但是我们如何在高维环境中做到这一点？本文研究了一种方法：添加辅助损失，智能体必须在其中辅助预测环境的下一个状态。由于环境本身包含从事有用工作的专家，因此智能体暗中必须了解这些专家在做什么以及他们的行为产生了什么影响。

他们发现，这样的智能体学会了遵循专家的提示，因此相对于单独训练的智能体而言，奖励得到了显着提高。实际上，这些智能体可以转移到新颖的环境中，在那里他们继续遵循专家的提示以获取高回报。但是，这意味着在专家不在场时他们不会学习如何操作，因此在独奏设置中会失败。这可以通过在单独的设置和与专家在场的设置混合的情况下进行训练来解决。

Rohin 的观点：我非常喜欢将人类建模为环境的一部分，因为我们最终将使人工智能系统与人类合作并与人类互动-它们不会像现在那样“在人工智能宇宙的外部”目前通常建模。

杂项（对齐）

人工智能接管的日期不是人工智能接管的日期（Daniel Kokotajlo）（由 Rohin 总结）：这篇文章指出，当根据AGI时间轴做出决策时，相关日期不是人工智能实际接管世界的日期，而是我们可以对此做任何事情的最后一点。

人工智能战略与政策

未来指数：群体预测如何 为大局提供参考（Michael Page 等人）（由 Rohin 总结）：本文解释了 CSET 最近的预测项目 Foretell 背后的方法。我们想知道未来 3-7 年内可能发生的几种潜在地缘政治情况中的哪一种。通过向相关专家征求意见，我们可以对此有所了解，但是通常很多专家会不同意，因此很难得出结论。

我们想通过利用群体的智慧来减轻这种情况。不幸的是，这将要求我们对方案进行清晰，准确的操作；我们感兴趣的场景很少适合这种操作。取而代之的是，我们可以找到许多可以针对特定情况辩护的预测变量，并确定一个或多个指标，这些指标本身是清晰准确的，并且可以为我们提供有关某些预测变量的信息。我们可以利用群体的智慧来获得这些指标的预测。然后，我们可以计算群体预测和历史数据的简单趋势外推之间的偏差，并使用观察到的趋势方向作为支持或反对特定方案的参数。

本文在涉及美国，中国和人工智能的潜在场景中说明了这一点。一个重要预测指标的例子是“中美紧张局势”。相关指标包括中美贸易额，中国 O 签证数量等。在这种情况下，群体预测表明指标中存在趋势偏差，这表明美中紧张关系加剧。