132 | ICML 2018论文精读：聊一聊机器学习算法的“公平性”问题

2018-08-06 洪亮劼

AI技术内参

进入课程 >

讲述：初明明

时长 06:35　大小 3.02M

在上一次的分享里，我们介绍了今年 ICML 大会的一篇最佳论文，这是一篇非常优秀的机器学习和计算机安全相结合的论文。这篇论文剖析了目前在白盒攻击的场景下，攻击方如何绕过一种叫作“混淆梯度”的情况，来实现有效攻击的目的。

今天，我们来分享 ICML 2018 的另一篇最佳论文，题目 Delayed Impact of Fair Machine Learning。这篇论文主要探讨了“公平”（Fair）在机器学习中的应用。论文的五位作者都来自加州大学伯克利分校。

论文的背景

这篇论文所探讨的主题是机器学习的“公平性”问题。近些年，这个主题受到了学术界越来越多的关注，但是对于普通的人工智能工程师和数据科学家来说，这个议题依然显得比较陌

生和遥远。所以，我先来简单梳理一下这方面研究的核心思想。

机器学习有一个重要的应用，就是在各类**决策场景**中提供帮助，例如申请贷款、大学入学、警察执勤等。一个不可否认的特点是，这些决策很有可能会对社会或者个人产生重大的不可逆转的后果。其中一个重要的后果就是，针对不同的人群，有可能会产生意想不到的“不公平”的境况。比如，有一些普遍使用的算法，在帮助警察判断一个人是否可能是罪犯的时候，系统会认为美国黑人相对于白人更容易犯罪，这个判断显然存在一定的问题。

机器学习研究者已经注意到了这种算法中的“公平”问题，并且开始探讨没有任何限制条件的机器学习算法，是否会对少数族裔（Underrepresented Group）产生不公平的决策判断。基于这些探索，研究者们提出了一系列的算法，对现有的各种机器学习模型增加附带了公平相关的限制条件，希望通过这种方法来解决各种不公平定义下的决策问题。

论文的主要贡献

这篇论文从理论角度展开讨论，基于什么样假设和条件下的具有公平性质的机器学习算法，在决策场景中能够真正为少数族群带来长期的福祉。值得注意的是，这里所谓的少数族裔是一个抽象化的概念，指的是数目相对较少的，或者在某种特性下比较少的一组数据群体。这篇论文并不直接讨论社会学意义下的少数族群的定义。

作者们主要是比较两个人群 A 和 B，在不同的公平条件下，看这两组人群的某种“**效用**”（Utility）的差值会发生什么变化。这个差值可以是正的，没变化或者是负的。

论文的主要结论是，在不同的公平条件下，效用差值会有各种可能性。这其实是一个非常重要的结论。有一些公平条件，直觉上我们感觉会促进少数族群的效用，但这篇论文向我们展示了，即便出发点是好的，在某些情况下，效用差值也可能是负的。

除此以外，这篇论文还探讨了“**测量误差**”（Measurement Error）对效用差值的影响。作者们认为测量误差也应该被纳入整个体系中去思考公平的问题。

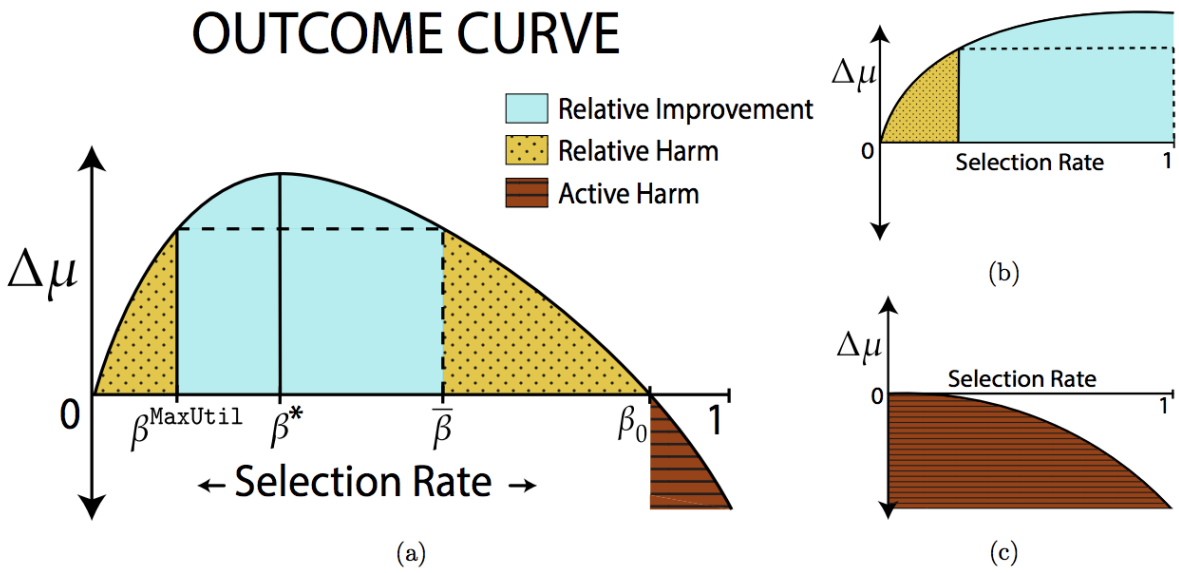
需要指出的是，论文的分析方法主要建立在时序关系的“**一步预测**”（One Time Epoch）基础上的。也就是说，我们利用当前的数据和模型对**下一步的决策判断**进行分析，并不包括对未来时间段所有的预测。从理论上说，如果在无限未来时间段的情况下，结论有可能发生变化。

论文的核心方法

这篇文章的核心思路是探讨针对人群 A 和 B 所采取的一种“策略”（Policy），是怎么样影响这两组人群的效用差别的。如果某种策略会导致某个群体的**效用差别为负**，那么我们就说这个策略对群体产生了“**绝对损坏**”（Active Harm）作用；如果**效用差别是零**，就说明这个策略对群体产生了“**停滞**”（Stagnation）作用；如果**效用差别是正的**，就说明这个策略对群体产生了“**推动**”（Improvement）作用。

除此以外，我们认为有一种不考虑人群 A 和 B 具体特征的期望最大化效用的策略，称之为“**最大化效用**”（MaxUtil）。这种策略其实就是在没有约束条件的情况下，利用一般的机器学习算法达到的效果。我们需要把新策略和这个策略进行比较，如果新的策略比这个策略好，就是产生了“**相对推动**”（Relative Improvement），反之我们说新的策略产生了“**相对损害**”（Relative Harm）。

为了进一步进行分析，作者们引入了一个叫“**结果曲线**”（Outcome Curve）的工具来视觉化策略和效用差值的关系。具体来说，曲线的横轴就是因为策略所导致的对某一个群体的选择概率，纵轴就是效用差值。当我们有了这个曲线之后，就能非常直观地看到效用差值的变化。



从这个曲线上我们可以看到，效用差值的确在一个区间内是“相对推动”的，而在另一个区间是“相对损害”的，在最右边的一个区间里是“绝对损害”的。这就打破了我们之前的看法，认为有一些选择策略会一致性地导致唯一结果。

在此基础上，我们专门来看这两种特殊的策略。第一种叫“**种族公平**”（Demographic Parity），思路是希望在两个人群中保持一样的选择概率。另一种策略叫“**公平机会**”（Equal Opportunity），思路是希望某个人群中成功的概率（例如申请到贷款、学校

录取等）和人群无关。这两种策略都是典型的试图利用限制条件来达到公平的方法。我们希望来比较的就是这两种策略以及之前说的最大化效用之间的一些关系，得出以下三个主要结论。

第一个比较出乎意料的结论是最大化效用这个策略并不会导致“绝对损害”。意思就是说，和人们之前的一些想法不同，最大化效用也有可能让少数族裔的效用得到提升或者不变。

第二个结论是，这两种公平策略都可能会造成“相对推动”。这也是推出这两种策略的初衷，希望能够在选择概率上进行调整，从而让少数族裔的效用得到提升。

第三个结论是，这两种公平策略都可能会造成“相对损害”。这是本篇论文的一个重要结论，正式地证明了公平策略在某个区间上其实并没有带来正向的“推动”反而是“损害”了少数族群。作者们进一步比较了“种族公平”和“公平机会”这两个策略，发现“公平机会”可以避免“绝对损害”而“种族公平”则无法做到。

小结

今天我为你讲了今年 ICML 的另一篇最佳论文。

一起来回顾下要点：第一，这篇论文讨论了计算机算法的公平性问题；第二，我们详细介绍了论文提出的两种策略以及得出的主要结论。

最后，给你留一个思考题，研究算法的公平性对我们日常的应用型工作有什么启发作用？

欢迎你给我留言，和我一起讨论。

上一篇　131 | ICML 2018论文精读：模型经得起对抗样本的攻击？这或许只是个错觉

下一篇　133 | ICML 2018论文精读：优化目标函数的时候，有可能放大了“不公平”？

精选留言

写留言

> 由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。