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”？

2018-08-08 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 06:21 大小 2.91M

今天我们要分享的是 ICML 2018 的一篇最佳论文提名，题目是 Fairness Without Demographics in Repeated Loss Minimization。

这篇论文讨论了这样一个话题，在优化目标函数的时候，如何能够做到针对不同的子群体，准确率是相当的，从而避免优化的过程中过分重视多数群体。这篇论文的作者都来自斯坦福大学。

论文的主要贡献

这篇论文其实也是希望讨论算法带来的“公平性”问题，但是出发的角度和我们上一篇讨论公平性的论文非常不一样。这篇论文的核心思想，是希望通过机器学习目标函数优化的原

理，来讨论机器学习和公平性的关系。

作者们发现，基于“平均损失”（Average Loss）优化的机器学习算法，常常会给某一些少数群体带来巨大的不准确性。这其实并不是模型本身的问题，而是优化的目标函数的问题。在这样的情况下，目标函数主要是关注有较多数据的群体，保证这些群体的损失最小化，而可能忽略了在数量上不占优势的少数群体。

在此基础上，还带来了另外一个用户“**留存度**”（Retention）的问题。因为少数群体忍受了比较大的优化损失，因此这些群体有可能离开或者被这个系统剔除。所以，长期下去，少数群体的数目就可能逐渐变少。这也许是目标函数的设计者们无从想到的一个平均损失函数的副产品。作者们还把这个现象命名为“**不公平的放大**”（Disparity Amplification）。

这篇论文的一个重要贡献是发现**ERM（Empirical Risk Minimization，经验风险最小化）其实就存在这种不公平的放大性**。ERM 包含了很多经典的机器学习模型的目标函数，例如支持向量机（Support Vector Machines）、对数回归模型（Logistic Regression）以及线性回归等。作者们还发现，ERM 可以让即便在最初看上去公平的模型，在迭代的过程中逐渐倾向于不公平。

为了解决 ERM 的问题，作者们开发了一种新的算法框架，DRO（Distributionally Robust Optimization，分布式健壮优化）。这种框架是为了最小化“最差场景”（Worst-Case）的风险，而不是平均风险。作者们在真实的数据中展示了 DRO 相比于 ERM 更能够解决小众群体的不公平性问题。

论文的核心方法

为了解决在 ERM 下的对不同群体的不公平性问题，作者们首先对数据做了一种**新的假设**。

作者们假设数据中有隐含的 K 个群体。每一个数据点，都有一定的概率属于这 K 个群体。我们当然并不知道这 K 个群体本身的数据分布，也不知道每个数据点对于这 K 个群体的归属概率，这些都是我们的模型需要去估计的隐含变量。

对于每一个数据点而言，在当前模型下，我们都可以估计一个“**期望损失**”（Expected Loss）。在新的假设框架下，因为每个数据点可能属于不同的 K 个群体，而每个群体有不同的数据分布，因此会导致在当前群体下的期望损失不一样，也就是会出现 K 个不一样的期望损失。我们的目的，是要控制这 K 个损失中的**最差的损失**，或者叫**最差场景**。如果我

们可以让最差的损失都要小于某一个值，那么平均值肯定就要好于这种情况。这也就从直观上解决了不公平放大的问题。

那么，如果我们直接在这样的设置上运用 ERM，会有什么效果呢？这里，有一个数值是我们比较关注的，那就是在整个框架假设下，每个群体的**期望人数**。这个数值等于在期望损失的情况下，当前群体剩余的人数加上新加入的人数。作者们在论文中建立了对这个期望人数的理论界定。

这个结论的直观解释是，如果在当前更新的过程中，期望人数的数值估计能够达到一个稳定的数值状态，那么就有可能稳定到这里，不公平放大的情况就不会发生；而如果没有达到这个稳定的数值状态，那么不公平放大的情况就一定会发生。也就是说，在 ERM 优化的情况下，群体的大小有可能会发生改变，从而导致人群的流失。

在这个理论结果的基础上，作者们提出了 DRO。DRO 的核心想法就是**要改变在优化过程中，可能因为数据分配不均衡，而没有对当前小群体进行足够的采样**。

具体来说，**DRO 对当前群体中损失高的人群以更高的权重，也就是说更加重视当前目标函数表现不佳的区域**。对于每一个数据点而言，损失高的群体所对应的群体概率会被放大，从而强调这个群体当前的损失状态。换句话说，DRO 优先考虑那些在当前情况下损失比较大的小群体。这样的设置就能够**实现对最差情况的优化从而避免不公平放大**。

作者们在文章中展示了 DRO 所对应的目标函数可以在递归下降的框架下进行优化，也就是说任何当前利用 ERM 的算法，都有机会更改为 DRO 的优化流程，从而避免不公平放大的情况。

论文的实验结果

作者们在一个模拟的和一个真实的数据集上进行了实验。我们这里简单讲一讲真实数据的实验情况。

作者们研究了一个“自动完成”（Auto Completion）的任务。这个任务是给定当前的词，来预测下一个词出现的可能性。而数据则来自两个不同人群，美国白人和黑人所产生的推特信息。在这个实验中，作者们就是想模拟这两个人群的留存度和模型损失。这里面的隐含假设是，美国白人和黑人的英语词汇和表达方式是不太一样的。如果把两个人群混合在一起进行优化，很有可能无法照顾到黑人的用户体验从而留不住黑人用户。

在实验之后，DRO 相比于 ERM 更能让黑人用户满意，并且黑人用户的留存度也相对比较高。从这个实验中，DRO 得到了验证，的确能够起到照顾少数人群的作用。

小结

今天我为你讲了今年 ICML 的最佳论文提名。

一起来回顾下要点：第一，这篇论文也讨论了算法带来的“公平性”问题，是从机器学习目标函数优化的角度来考虑这个问题的；第二，这篇论文的一个重要贡献是发现 ERM 确实存在不公平的放大性，基于此，作者们开发了一种新的算法框架 DRO；第三，文章的实验结果验证了 DRO 的思路，确实能够解决小众群体的不公平性问题。

最后，给你留一个思考题，这两期内容我们从不同的角度讨论了算法的公平性问题，你是否有自己的角度来思考这个问题？

欢迎你给我留言，和我一起讨论。

上一篇　132 | ICML 2018论文精读：聊一聊机器学习算法的“公平性”问题

下一篇 134 | ACL 2018论文精读：问答系统场景下，如何提出好问题？

精选留言 (2)

写留言

幻大米
2018-08-13

没看原始论文前会有些疑问：关照了少数人群，多数人群会不会有损失呢？损失大于少数人群的提升吗？

刘洋
2018-08-09

通过对少数群体补充训练样本的方式，采用erm方式来进行优化，应该也可以吧？