软件所在机器学习公平性方面取得进展

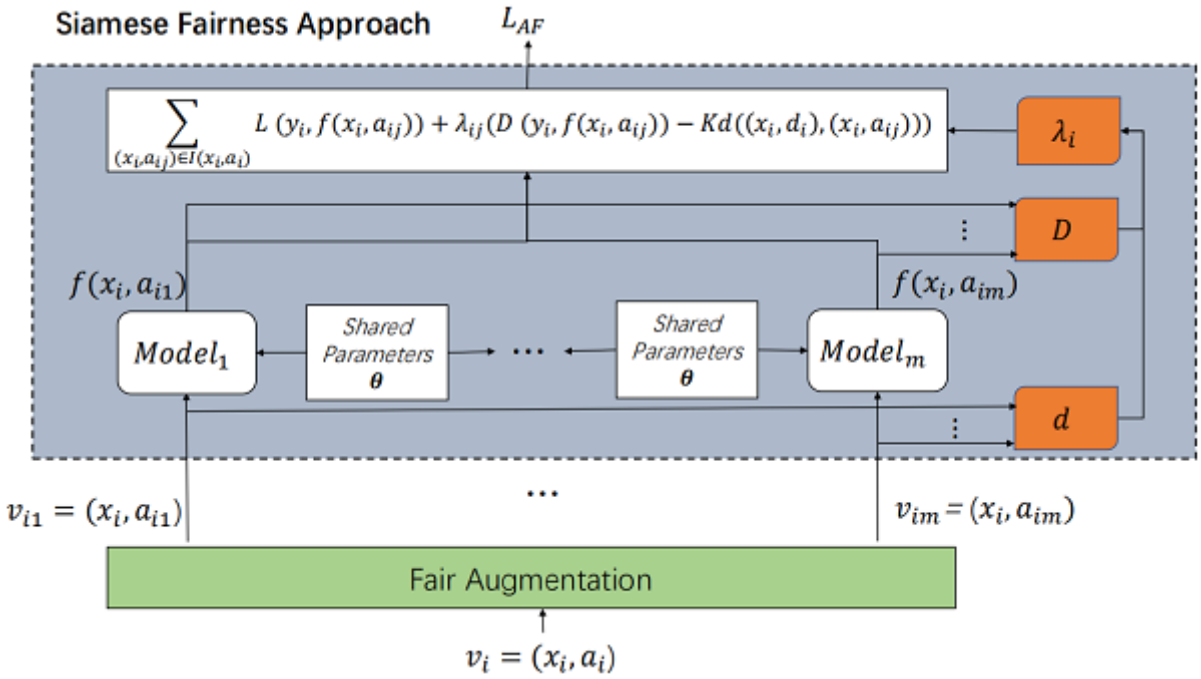
文章来源： | 发布时间：2023-03-14 | 【打印】 【关闭】

近日，软件所计算机科学国家重点实验室的研究论文“Accurate Fairness: Improving Individual Fairness without Trading Accuracy”被第37届AAAI人工智能大会接收，第一作者为硕士生李旭然，通讯作者为吴鹏副研究员。论文首次提出兼顾准确性的公平性评估标准——准确公平性（Accurate Fairness），以评估模型的预测结果是否既准确且公平。同时，首次提出孪生公平算法（Siamese Fairness Approach），实现了在提升模型准确公平性的同时，不损失其准确性与个体公平性，并应用于消除现实生活中的服务歧视问题。

随着机器学习的迅猛发展，如何保证机器学习系统的可靠性变得十分重要。可靠的机器学习系统不仅需要具有较高的性能，其预测结果还应满足社会、法律、伦理等公平性要求。但现有的机器学习评估标准无法同时兼顾准确性与公平性；现有机器学习算法在提升准确性或公平性二者之一时，往往以损害另一方为代价

为了从兼顾准确且公平的角度评估机器学习模型，论文首次将个体公平性标准的相似性条件与准确性标准中的真实标注信息相结合，提出新的公平性标准——准确公平性，从准确且公平的角度度量模型的可靠性，进一步提出了准确公平性度量标准，即公平查全率、公平查准率、公平-F1得分。论文首次将孪生算法应用于歧视消除，提出孪生公平算法。该算法可以在训练过程中同时接收多个相似输入，在提升模型公平性的同时不损失其准确性。

在公平性基准数据集Adult、COMPAS和German Credit上的实验结果表明，准确公平性标准可以发现准确性和个体公平性标准所忽略的准确但歧视、错误但公平的预测，同时孪生公平算法可以在提升模型准确公平性的同时，不损失其个体公平性与准确性。最后，论文将准确公平性标准以及孪生公平算法应用于检测并消除现实生活中的服务歧视问题，能够帮助酒店在不损失收益的情况下，消除服务歧视问题，为具有不同消费习惯的客户提供准确且无差别的房间服务。



算法流程图

论文地址：

https://arxiv.org/abs/2205.08704

模型地址：

https://github.com/Xuran-LI/AccurateFairnessCriterion