FR-Train: A Mutual Information-Based Approach to Fair and Robust Training

作者:

Yuji Roh　Kangwook Lee　Steven Euijong Whang　Changho Suh

摘要:

值得信赖的人工智能是机器学习中的一个关键问题，除了在训练精确的模型之外，还必须考虑在存在数据偏差和中毒的情况下进行公平和鲁棒的训练。然而，现有的模型公平技术错误地将有毒数据视为需要修复的额外偏差，导致严重的性能下降。为了解决这个问题，我们提出FR-Train，全面执行公平和稳健的模型训练我们提供了一个相互信息的解释现有的对抗训练基于公平的方法，并将这个想法应用于架构一个额外的鉴别器，可以识别有毒数据使用一个干净的验证集和减少其影响。在我们的实验中，FR-Train显示，在存在数据中毒时，通过减轻偏见和防止中毒，几乎使公平性和准确性没有下降。我们还演示了如何使用众包来构建干净的验证集，并发布新的基准数据集。

1.论文试图解决什么问题

本文提出了FR-Train方法，目的是当数据集存在偏差数据和中毒数据（有噪声、主观甚至敌对性的数据）时提高训练模型的公平性和鲁棒性。

该方法的输入是数据集x（包含偏差、中毒数据），使用FR-Train方法可以1）在训练过程中能更加准确，对中毒数据处理也很公平健壮。2）还可以进行数据清洗，构建一个新的干净的数据集x_1。

2.你的理解：论文中提到的解决方案之关键是什么?

- 本文对现有的基于公平性的对抗性训练方法提供一个互信息的解释，并将此想法应用于构建一个额外的判别器，该判别器可以使用干净的验证集识别中毒数据并减少其影响。
- 本文提出的方法还利用鲁棒性鉴别器的结果，通过重新加权来进一步提高公平性训练。
- 该方法不像传统的方法在模型训练中按顺序处理鲁棒性和公平性，而是同时处理它们。

3. 客观描述：论文中的实验是如何设计的?

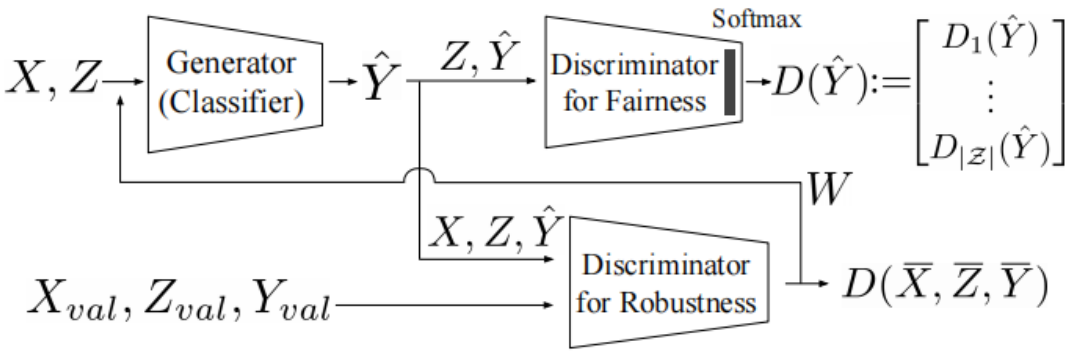

Figure 3. The architecture of FR-Train.

1. 分类器

先用一个分类器使得确保预测结果$\hat{Y}$，和前面的敏感属性无关

2. 鉴别器

- **鲁棒性鉴别器**：用于**区分新的训练数据和已经划分出的干净的验证集**，并且通过通过重新加权，返回到分类器输入口，来进一步提高公平性训练。
- **公平性鉴别器**：这一步就是传统的公平性训练方法，区别就在于输入已经做过了很多处理，并且该方法**鲁棒性和公平性的处理是同时进行**的。

4.用于定量评估的数据集是什么？代码有没有开源?

- 数据集 ProPublica COMPAS (Angwinet al., 2016) and AdultCensus (Kohavi, 1996), 分别有 7,214 and 45,222 个样例。
- 代码开源了，在附加文件可以找到。

5.论文中的实验及结果有没有很好地支持需要验证的科学假设?

有，该论文的主要假设或者观点是他们提出的训练模型在有中毒数据的时候有很高的公平性和鲁棒性。

-

Method	Clean data		Poisoned data	
	DI	Acc.	DI	Acc.
FC	.822	.806	.831 (1.1% ↑)	.760 (5.7% ↓)
LBC	.819	.760	.827 (1.0% ↑)	.715 (5.9% ↓)
AD	.807	.811	.834 (3.4% ↑)	.769 (5.2% ↓)
RML+FC	.822	.806	.802 (2.4% ↓)	.529 (34.% ↓)
RML+LBC	.819	.760	.810 (1.1% ↓)	.752 (1.1% ↓)
RML+AD	.807	.811	.808 (0.1% ↑)	.756 (6.8% ↓)
LR	.409	.885	.446 (9.1% ↑)	.819 (7.5% ↓)
RML	.471	.876	.395 (16.% ↓)	.869 (0.8% ↓)
FR-Train	**.818**	**.807**	**.827 (1.1% ↑)**	**.814 (0.9% ↑)**

- 上图比较了不同方法在干净数据和中毒数据集中的表现。发现存在数据中毒的时候，FR-Train方法的公平性和准确率均有所增加，说明了这个方法的鲁棒性和公平性

-

Dataset	Poisoned data	
	DI	Acc.
Synthetic	0.795 ± 0.019	0.805 ± 0.008
COMPAS	0.827 ± 0.027	0.653 ± 0.005
AdultCensus	0.871 ± 0.034	0.796 ± 0.006

- 上图是他们对每一个数据集做过数据清理后的新数据集的表现，发现即便是正确率最低的情况下，其正确率也比表一的正确率高。证明了该方法适用于构建一个新的干净的数据集。

6.这篇论文到底有什么贡献?

关键贡献是提供了一种使用相互信息的对抗性学习方法的解释，并提出了一种新的GAN架构（1）使用公平区分器来区分预测敏感属性和其他属性（2）使用鲁棒性鉴别器，用预测区分训练数据和一个干净的验证集，也用于通过示例重新加权进一步提高公平训练。

此外，该论文还演示了如何使用众包来构建一个干净的验证集，并发布了两个由亚马逊机械土耳其人构建的新数据集作为社区资源。证明了现有的公平方法容易遭受数据中毒，即使结合数据消毒。相比之下，FR-Train对中毒具有鲁棒性，即使验证集太小或不可用，也可以调整以保持合理的准确性和公平性。

7.下一步呢？有什么工作可以继续深入？

- 目前该方法研究的是群体公平，未来是否可以研究个人公平？
- 是否可以试着让机器能理解敏感属性和非敏感属性存在一些因果关系？这样该方法中的分类器是不是可以更好地工作？