一种基于相互信息的公平和稳健的训练方法

**卢裕二**1 **李康虎**2 **义宗旺**1 **超昌浩**1

# 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.

# 项目简介

随着机器学习在软件中变得广泛使用

2.0时代（卡帕蒂，2017年），值得信赖的人工智能正变得越来越重要。除了简单地训练准确的模型，还迫切需要解决多种需求，包括公平性、健棒性、可解释性、透明度和问责制(IBM，2020)。特别是，我们关注公平性和鲁棒性，这是受相同训练数据影响的密切相关的问题。对于医疗保健、金融和自动驾驶汽车等敏感应用，训练有素的模型不能歧视汽车

1韩国大田，韩国高等科技学院(KAIST)，电气工程学院2美国威斯康辛州麦迪逊大学麦迪逊分校电气和计算机工程系。通讯：英宗华<[swhang@kaist.ac.kr](mailto:swhang@kaist.ac.kr)*>.*

*第97页的诉讼程序第*国际机器学习会议，在线，PMLR119,2020。版权所有2020，作者所有。

1<https://github.com/yuji-roh/fr-train>

基于包括年龄、性别或宗教在内的敏感属性。此外，由于应用程序通常依赖于外部数据集来进行训练数据，因此模型训练必须对有噪声的、主观的、甚至是对抗性的数据具有弹性。

传统上，模型公平研究（文卡塔苏布拉曼尼亚，2019；丘尔德科瓦和罗斯，2018；Verma和鲁宾，2018）专注于开发不同影响等指标（费尔德曼等。，均衡赔率(Hardt等人。，和机会平等(Hardt等人。，它捕捉到了歧视的各种概念。最近，缓解不公平的技术出现了激增。，它通过固定训练数据、训练过程或训练模型来提高模型的公平性。缓解不公平性通常涉及到模型的准确性和公平性之间的一些权衡。最近，生成式对抗网络(GANs)正在适应公平的环境(Zhang等。，2018a)。GANs的架构是合适的，因为准确性和公平性并不总是对齐的，而且同时训练两个模型是有意义的：一个使用输入特征预测标签的分类器，另一个使用分类器的预测标签预测敏感属性的对手。

稳健的模型训练也很重要，需要同时考虑。随着数据集的发布正成为主流，如Kaggle和谷歌数据集搜索等系统所证明的那样。，很容易发布有噪声、主观甚至敌对性的数据，我们下文称之为有毒数据。因此，也有大量的算法，使模型训练对数据中毒具有弹性。，2013年；Biggio等人。，2011年；Fre´nay&Verleysen，2014年)。然而，数据中毒攻击已经变得越来越复杂，而且防御所有这些攻击都是很困难的。, 2018).

在不解决数据中毒的情况下解决模型公平性可能会导致准确性和公平性之间更糟糕的权衡。例如，考虑一个银行系统，即有两个敏感的群体：男性和女性。假设我们使用了不同的影响。作为公平措施。如果模型的阳性预测率为M，女性为W，则不同的影响为最小*M , W*其中，如果一个值为1，则被认为是完美的公平的。图1显示了一个玩具示例五个月人和五个月女性：

*W*

*M*

*{ }*

那些需要贷款的人。每个人都与一个单维特征x相关联，只有那些有一个圆盒子的人才能偿还他们的贷款（即，他们的标签是正面的）。让我们训练一个阈值分类器，把人们分成两组，其中左边的人被拒绝贷款

**公平的分类器**

（Acc、DI）=（0.8、1）

**清洁**

**不公平的分类器**

（Acc、DI）=（1.0.5）



|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| F | M | F | F | M | F | M | M | F | M |





**X**

而右边的人也获得了贷款。关于清洁数据

**已中毒的**

**不公平的分类器**

**公平的分类器**

上面，一个分类器，不考虑公平的(不公平的

交流

波点

= 0.9

交流

波点

= 0.8

分类器，红色虚线)可以有完美的准确性，代价是产生0.5的不同影响，因为40%的女性获得贷款，而80%的男性获得贷款。另一方面，一个公平的分类器（蓝色实线）可以

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| F | M | F | F | M | F | M | M | F | M |

（计算干净的，di)=（0.9,0.67）



（计算干净的，di)=（0.6、1）





**X**

划分这样的人，使不同的影响是完美的，但精度只有0.8。现在假设我们在翻转5的标签*第*和7*第*人员（两者均为 男） 来自 阳性的 至 负数 作为 所显示的下面是一个例子。 当每个分类器对中毒的数据进行训练时，它的准确性是用干净的数据标签来测量的。对于不公平的分类器 经过培训 开启 这是 数据， 的 测试结果 是 混合的 其中： 精度从1下降到0.9，但不同的影响从0.5增加到0.67。然而，公平的分类器是严格的 更糟 测试结果 其中： 的 测量精度 减少的 来自 0.8到0.6，没有任何改变不同的影响。因此，在 公平 分类器的 准确度-公平性 权衡 是： 更糟 当数据中毒时。一个建议是在模型训练之前对数据进行消毒，但已知要消除中毒 没有的 任何 知识知识的 的 的 产品型号 是： 非常困难的(Koh等人。, 2018).

我们的主要贡献是一个名为FR-Train的集成解决方案，它训练准确的模型，对中毒也很公平和健壮。FR-Train扩展了一种最先进的仅公平的方法，即对抗性去偏置(AD)(Zhang等。，包括一个用于分类的生成器和一个区分预测和另一个敏感组，类似于GANs(古德费罗等。. ，2014)。鉴别器确保了预测yˆ独立于敏感属性的，我们首先使用相互信息提供了这种对抗性学习方法的解释。然后，我们将结果作为灵感，添加一个新的鲁棒性鉴别器，使用互信息来区分训练数据的（训练示例、预测)和一个单独和干净的验证集的(验证示例、验证标签）。该鉴别器确保训练数据上的模型预测与干净数据上的标签“一致”，其中干净验证集作为对培训的参考。此外，我们还利用鲁棒性鉴别器的结果，通过重新加权的例子来进一步提高公平性训练。在我们的实验中，我们证明了在模型训练中按顺序处理鲁棒性和公平性并不如在Fr-Train中同时处理它们那样有效。

另一个贡献是解决了构建一个干净的验证集并优雅地处理

*图1。*10个人需要贷款的小数据集（F：女性，M：男性）。圆角方框表示正标签。清洁数据（以上）通过翻转两个标签（以下）中毒。垂直线是不公平和公平阈值的决策界限分类器。DI是不同的影响，和Acc干净的（计算波点）是：清洁（有毒）数据的准确性。

它较小或不可用的情况。为此，我们演示了一种实用的众包方法，使用多数投票来构建一个干净的验证集，它比输入数据少。我们使用亚马逊机械土耳其人从真实的数据集中构建干净的验证集，并将它们作为社区资源发布。在最坏的情况下，当验证集不存在时，我们展示了如何调整自由度列车的参数，以保持保持合理的准确性和公平性。

在下面的部分中，我们展示了当前公平性方法的弱点，提出了FR-Train的实验，并提出了相关的工作。

# 公平性方法的脆弱性

我们进行的实验证明，最先进的公平方法确实是容易的，即使是简单的中毒攻击。我们生成一个合成数据集，如图2a所示（参见第4.1节中的生成细节）。那里

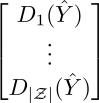
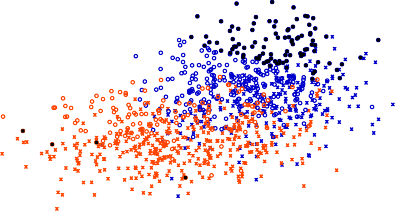
是两个非敏感的属性x1和x2，即：回复有的 在 的 x轴 以及 y轴， 分别为。的 实例

. 为了生成中毒数据，我们通过翻转属于特定z属性的例子的标签来毒害10%的训练数据(为此实验z

=1)，以最大限度地提高精度和性能下降。这种方法类似于现有的标签翻转方法。, 2018).为了创建一个验证集，我们随机选择相当于整个训练数据的10%的干净示例。

我们使用不同的影响作为公平衡量，并评估一种称为公平约束的公平方法(Zafar等。，它包含了一个正则化项，反映了基于凸边缘的分类器背景下的公平约束，如逻辑回归和支持向量机(S向量机)。 因为这种方法涉及到一个正则化的方法

5.0



软大

发电机 歧视者

（分类器） 为了公平性

歧视者

为了提高鲁棒性

|  |  |
| --- | --- |
| **z=0(pos)z=0(neg)z=1（pos)z=1(neg)**  **中毒的数据** |  |
|  |  |

0.0

**x2**



-5.0

0·8



**清理干净的数据，造成中毒的数据**

**完全不同的影响**

0·7

-10.0 -5.0 0.0 5.0 10.0

## x1

1. 带有标签翻转中毒的合成数据



*详见图3。* 外国列车的结构。

通过重新加权的例子进行公平训练。

## 公平性

我们用D来表示*tr* 培训数据集。 假设为D*tr* 具有

m个示例(i)，z(i)，y(i))}*m*



其中，x(i) 包含的

0·6

非敏感的属性，

*z*(i)

i=1

其中包含了这些敏感的属性，

0·5



0·4

0·725 0·750 0·775 0·800 0·825 0·850 0·875 0·900

**准确性**

1. 公平性约束下的准确性-公平性权衡曲线

*图2。* 上图显示了一个带有数据中毒的合成数据集。例子分为z=1（用圆标记）和z

=0（交叉），根据敏感属性z表示。 蓝色点表示正标签，而红点表示负标签。对于中毒，我们用z=1翻转了10%的例子的标签，以最大限度地提高精度和性能下降(Paudice等。，2018)。底部图显示，中毒显著恶化了公平约束方法的准确性-公平权衡（即曲线向左移）(Zafar等。, 2017).

平衡精度的因素λ和公平目标，我们可以通过调整其值得到一个权衡曲线。 图2b显示了用干净和有毒的合成数据集获得的两个精确-公平的权衡曲线。请注意，添加数据中毒会清楚地将曲线向左移动，这意味着精度会降低。这与我们的情况重合

和y(i) 是该标签。 敏感的属性和标签都可以 是 多类的， i.e.， 它们 可以 具有的 一个 的 多个 值。

为了符号上的简单起见，我们假设有一个敏感属性，它可以看作是字母表大小更大的多个敏感属性的合并结果。为了说明性的目的，我们关注不同的影响，留在补充我们的公式和实验结果，平等的机会和平等的机会。不同的影响目标是对每个敏感属性z具有相同的正预测比其中是可能的敏感属性值的集合。 我们使用以下定义来产生不同的影响：

*e z z*

**定义1。** *（完全不同的影响）*

*P(Y*ˆ =1|Z=z1)=P(Yˆ =1|Z=z2)，Vz1*，z*2 *e z.*

FR-Train中的第一个鉴别器可以区分预测

w.r.t. 一个敏感的群体来自于另一个群体。 不同的影响意味着，敏感的属性要独立于

该模型的预测，即，I（Z；Yˆ) = 0.

我们解释了FR-Train如何强制执行上述约束。 让P*Z*(z)是Z的分布，其中是zez。 让

直觉。 中毒混淆了模型，所以在那里

是更有偏见的例子来修复，这反过来使它吗

*反应过度，因此牺牲了更多的准确性。*  我们也是如此

*Y*ˆ*|Z=z~PY*ˆ *|z* (.)和Yˆ

*z∈Z PZ*(z)P*Y*ˆ *|z* (.).

*~PY*ˆ (.). 然后是P*Y*ˆ (.) =

在补充中留下了真实数据集上的公平性约束的准确性-公平性权衡曲线。结果表明，准确性和公平性都在降低

下面的定理断言互信息等价于以下函数优化，

最佳鉴别器D*\**(yˆ) = *PZ|Y*ˆ (z|yˆ)和

*z*

*z*

中毒的数据。在第4节中，我们将展示如何使用数据

中毒会影响其他公平的方法。

*z∈Z*

*D\**(yˆ)=1，Vyˆe y.

# FR-列车

**定理1。** *I(Z；Y*ˆ) =

最大值 *PZ*(z) *P*

*Dz* (yˆ)：

*Dz*(yˆ)=1，∀yˆ *z∈Z*

*Y z*

*z*

日志D*z*(Yˆ) +H(Z)。

我们现在描述快速列车列车（见图3）。 与传统的GANs不同，生成器是一个分类器，它接收一个示例xX并返回一个预测yˆ。 有两种鉴别器分别使用互信息来优化公平性和鲁棒性。此外，鲁棒性鉴别器的输出还可用于进一步改进

*e*

在将详细的证明推迟到补充材料的同时，我们提供了对证明的简要概述。由于RHS中的优化问题是凸的，我们通过求解KKT条件来找到最优鉴别器。然后证明了最优鉴别器得到的最大值等于相互信息

互信息的性质和广义延森-香农散度（Lin，1991）。

比显示上述等式更重要的是设计正确的优化问题。人们需要仔细地手工制作一个合理的优化问题，以便它的唯一解决方案匹配所需的数量。在这里，我们通过“猜测-检查”方法来设计优化问题，借助于互信息表达式中出现的KL差异的结构见解。

现在，我们将讨论如何实现上述表达式。

第4.1节）。在第3.3节中，我们还使用鲁棒性鉴别器，使用实例重加权来进一步提高公平性训练。

我们首先定义了X=VX+(1V)X圆形*，Z=V Z+(1*

*V)Z*圆形，和Y=VYˆ +(1V)Y圆形。在这里，请注意，V

*-*

*- -*

是一个指示器随机变量，表示是生成一个示例（V=1）还是来自验证集（V=0）。 然后，我们希望确保该分布

的（X、Z、Yˆ)匹配的 (X圆形*，Z*圆形*，Y*圆形). 这可以是

通过强制执行I(V；X，Z，Y)=0来完成，即，对训练数据的预测与标签无法区分

因为我们不知道P*Y*ˆ *z*

*|*

经验证版本如下：

(.)确切地说，我们计算出了

的验证集。因此，我们可以模拟干净的数据集

同时也期待着一个间接的消毒效果。

类似于公平鉴别器，我们证明了

最大值

*Dz* (yˆ)：

*PZ*(z)

1 日志D*z*(yˆ(i))+H(Z)。

相互信息等价于信息

*v*

降低函数优化，其中最优的降低

现在对于足够大的m，数字是m*z*

*z Dz* (yˆ)=1，∀yˆ *z∈Z*

*mz*

*i：z i* =z

的实例

犯罪犯D*\*(x，z，y)* = *P*

*V|X，Z，Y*

*(v|x、z、y)和*

带z(i) =z和P大致相同*Z*(z)m. -

*v*

*v∈V*

*D\*(x、z、y)=1、V(x、z、y)ex×z××*.

因此，上述表达式将变为：

证明类似于定理1。

**定理2。** *I(V；X，Z，Y)=*

最大值

1 日志D(yˆ(i))+H(Z)。

最大值

*Dz* (yˆ)： *z Dz* (yˆ)=1，∀yˆ *z∈Z i：z i* =z  *m*

*z*

*PV* (v) *P*

*Dv (x，z，y)：*  *v Dv (x，z，y)=1，∀(x，z，y)*

*v∈V*

*x，z，Y v*

日志D*v(X，Z，Y)*

+H(V)。

古德费罗等人。. 当=2时，我们还注意到，我们的公式不需要关于P的先验知识*Z*(z)。

*|z|*

有趣的是，这个公式和在

我们注意到对抗性偏偏差(AD)（Zhang等。还有一个额外的投影项，用来迫使分类器永远不会减少鉴别器的损失。 然而，我们在FR-Train中使用这个术语，因为它恶化了我们实验中的训练稳定性。

## 稳健性

鲁棒性鉴别器通过使用互信息来区分例子和前置-来确保鲁棒性训练

## 体系架构

Σ

Σ

我们在图3中描述了自由-列车体系结构。对于发电机的损失函数，我们采用交叉熵：

1 *m*

*- - - -*

*L*1 = *y*(i) 日志yˆ(i) (1 *y*(i))日志（1 *yˆ*(i)). *m*

i=1

我们设置了损失函数w.r.t。公平性鉴别器的表现如下：

最大值为L= 1 日志D(yˆ(i))+H(Z)

2

*D（·）*

*m*

*z*

*z∈Z i：z i* =z

其中，D（。。)：=(D1(.*), . . . ，D|Z|*(.)). 该条件

*z*

从一个干净的验证集。现在，让我们假设一下

*z∈Z*

*D\**(Yˆ) = 1可以通过添加一个软最大值来强制执行

存在这样的验证集（在第4.2节中，我们演示如何构建一个）。 然后，鉴别器用预测来区分训练数据(i)*, z*(i)*, yˆ*(i))}*m*  来自

i=1

一层到鉴别器。

最后，通过实现I（V；X、Z、Y），我们设置了损失函数w.r.t。鲁棒性鉴别器，如：

(i)

(i)

(i)

*m*圆形

验证集圆形 *，z*圆形 *，y*圆形 )}i=1 . 直观上说，如果

分类器被训练数据中的数据中毒所混淆，则其预测将与的标签不一致

3

*Dr* (·)

*m*

*圆形*

*圆形*

*圆形*

最大值为L=

1 日志D*r*(x(i) *，z*(i) *，y*(i) )+

*i：v i* =0

干净的数据和鉴别器将能够检测到这种差异。我们对验证集的使用是受到基于元学习的鲁棒训练算法的启发。，它还通过使用验证数据丢失作为一个元目标来防止中毒攻击。 然而，一个关键的区别是，我们采取了一种对抗性的学习方法，它引入了一个旋钮来控制鲁棒训练的重点。 我们发现这个旋钮使FR-Train对验证集的大小更健壮(请参见

1 日志（1） *Dr*(x(i)*，z*(i)*，yˆ*(i)))+H(V)。 *m*

*i：v i* =1

*-*

最终的目标函数是这些值函数的加权和：

最低L1 +λ1*L*2 +λ2*L*3*.*

*G（·）*

此处为λ1 以及 *λ*2 是 正在调整 自动控制器，旋钮 那个 播放 角色 至 强调公平和稳健的训练， 分别为。

## 公平培训的重例子

对于上述的体系结构，我们也利用了决策值D*r*(X， *Z， Y*ˆ以鲁棒性鉴别器为例，以进一步提高公平训练

图3，从鲁棒性鉴别器输出到分类器输入的箭头）。特别是，这两个损失是L1和L2现在是使用示例权重来计算。直觉是，通过更重视干净的例子，我们可以提高准确性-公平性的权衡。

问题是什么时候应用这些权重。如果我们过早地应用权重，那么D(X，Z，Yˆ)可能还不够准确，实际上会损害公平培训。直观上说，

*具体情况详见表1。*在合成测试数据集上的准确性和公平性性能。不同的影响(DI)。比较了两种方法：（1）公平性方法：FC(Zafar等。，2017)，LBC(Jiang&Nachum，2020)和AD(Zhang等人。其中“RML+”表示使用RML的消毒应用(Ren等。，之前；（2）不公平的方法：LR和RML。对于FR-Train和RML，验证集为的10%*tr*。中毒的量为中毒的10%*tr*。对于中毒数据的每个结果，我们与干净的数据结果进行比较，并显示百分比增加或 减少了。

处理方法 清理数据 中毒的数据

第 执行。 第 执行。

我们想使用鉴别器的结果时，当我们

要知道它的性能至少和分类器一样好。因此，对于一个更可靠的信号，我们使用分类器和鲁棒性鉴别器之间的相对性能来生成权重。给定分类器的损失L*c*和健壮性 鉴别器的 损失 *Ld*, 我们的 计算结果 的 最终的前

充足的重量，如W=R+D(X，Z，Yˆ）（1R），其中：

*× -*

*R=σ(Lc*C)是使用乙状函数σ和超参数C将损失比转换为概率。

*Ld*

*-*

我们注意到C作为损失比的一个阈值。

# 实验结果

我们提供了fr列车的实验结果。对于公平度，我们使用不同的影响，而在补充结果平等的赔率和机会的平等。我们分别评估所有的模型

清洁的测试集。在我们的实验中，我们使用了两个敏感的属性z1和z2，而不同的影响被测量为 比率 每分钟的{ *P* (Yˆ=1|Z=z ) *, P* (Yˆ=1|Z=z ) *}.*我们的 使用情况 Py-

*P* (Yˆ =1|Z=z )P(Yˆ =1|Z=z )

火炬(Paszke等人。所有的实验都是在使用Inteli7-6850CPU的服务器上进行的。更多的实施细节请在补充文件中进行详细说明。

## 合成数据的结果

对于合成数据，我们生成了2000个具有两个非敏感属性的例子x1和x2，敏感属性z和标签y，使用类似(Zafar等。，2017)。z和y都是双向的 无，和（x1，x2)对由两个正常的注意力组成

但是：（x1*，x*2)y=0（[2；2]、[10、1；1、3]）和

*| ~N--*

(x1*，x*2)y=1 ([2; 2], [5, 1; 1, 5]).z属性具有伯努利分布p(z=1)=p((x*j*1*，xj*2)y=1)/[p((x*j*1*，xj*2)y=0)+p((x*j*1*，xj*2)y=1)]其中(x*j*1*，xj*2)=(x1cos（π/4）x2罪（π/4），x1罪（π/4）+x2因为该公司（π/4))。最后，对于每个例子，都是x1和x2

*-*

*| |*

*|*

*| ~N*

值根据与y相关联的正态分布进行采样。对于数据中毒，我们用z=1翻转示例的标签，以最大限度地提高第2节所述的精度性能下降，中毒量为D的10%*tr*. 在补充文件中，我们也是

fc .822 .806 .831 (1.1%t)(5.7%l).760

lbc .819 .760 .827 (1.0%t)(5.9%l).715

广告广告的 .807 .811 .834 (3.4%t)(5.2%l).769

rml+的频率 .822 .806 .802 (2.4%l)(34.%l).529

rml+磅重数.819.760 .810 (1.1%l)(1.1%l).752

rml+广告 .807 .811 .808 (0.1%t) .756 (6.8%l)

lr .409 .885 .446 (9.1%t)(7.5%l).819

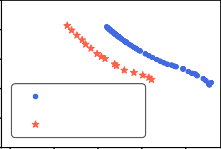
右毫升 .471 .876 .395 (16.% l) .869 (0.8%l)

## FR-列车 .818 .807 。827(1.1%t).814(0.9% *t)*

执行FR-列车，将中毒量从10%变化到40%。

**准确性和公平度，我们比较了FR-Train与各种基线。**首先，有公平的方法：公平约束(Zafar等。，2017）(FC)、标签偏差修正（Jiang&Nachum，2020）(LBC)和对抗性偏移（Zhang等。，2018年a）（广告）。如前几节所述，FC添加了一个惩罚术语，以捕获敏感群体之间的预测差异，而AD则利用对抗性学习来实现高公平性。LBC是一个例子的重加权算法，它假设存在真正的无偏但未知的标签。LBC提供了理论保证，对结果损失的训练对应于对真正的无偏标签的训练，从而产生了一个公平的模型。虽然还有其他的重新加权技术，包括(阿加瓦尔等人。，我们选择LBC是因为它在实验中表现最好(Jiang&Nachum，2020)。

据我们所知，由于FR-Train是第一个在模型训练中同时解决公平性和鲁棒性的方法，因此没有一种公平性方法可以使用干净的验证集来执行数据卫生处理。然而，(Ren等人。是一种最先进的基于元学习的健壮训练方法，使用一个干净的验证集，我们称之为RML。为了进行公平的比较，我们扩展了三种公平方法，首先执行RML，然后在公平训练中使用示例权重。此外，我们还与非公平的方法进行了比较：逻辑回归(LR)和RML。



* + 1. lbc



* + 1. 广告广告的

1·0

0·8

0·6

0·4

0·2

0·0

0·75 0·80 0·85

**交流**

* + 1. FR-列车

1·0

0·8



**清洁中毒**

0·6

0·4

0·2

0·0

0·75 0·80 0·85



**清洁中毒**

**交流**

* + 1. FR-列车

1.0

0.8

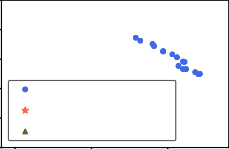
0.6

0.4

0.2

0.0

0.75 0.80 0.85



**干净的**

**中毒(λ2 =0.4)中毒(λ2 = 0.1)**

**交流**

* + 1. FR-列车

（变量。设置尺寸=10%)

（变量。设置尺寸=5%）

（变量。设置尺寸=0.1%)

*图4。*精确性-公平性的权衡曲线。图(a)和(b)显示，中毒恶化了LBC（Jiang&Nachum，2020）和AD（Zhang等）的准确性-公平性权衡。，2018a)。图(c)和(d)显示，FR-Train即使使用a5%的验证集。当验证集太小时（图(e)），FR-Train可以调整λ2 减少对不利影响 正在进行的培训。

表1将快速列车与基线进行了比较。我们使用的验证集是10%的*tr*. 我们的也应用适当的超参数，使不同的影响相似(约0。如果可能的话，跨所有方法的8)。执行时间：

*D*

正在设置λ1和λ2对于FR-Train，我们通常会修复λ2给一些

值，然后调整λ1使用一轮的交叉验证。逻辑回归没有超参数调整

以及基于元学习的鲁棒训练算法，因为它们没有调整公平性的知识。结果表明，对于公平性方法，数据中毒加剧了准确性-公平性的权衡。例如，FC的精度下降了5.7%，而它的不同影响仍然是一个相似的值。另一方面，变频列车的性能不会下降：不同的影响和精度分别增加了1.1%和0.9%。表1还显示，将公平性方法与RML（第4-6行）相结合并不总是能产生更好的准确性和公平性。事实上，使用消毒处理可能会降低精度或公平性(例如，RML+FC对中毒数据的精度为0.529，而FC则为0.760)。研究结果表明，消除中毒和消除偏见并不是那么有效的。

我们观察到准确性是如何在干净和有毒的数据集上与公平性进行权衡的。FC的计算结果如图2b所示。对于LBC，我们使用训练的数量作为一个旋钮来交换准确性与公平性，因为LBC通过在每次训练中反复更新示例权重来逐步提高公平性。如图4a所示，折衷曲线向左移动，这显示了一个明显的折衷退化。

对于AD，我们使用了α参数(张等人。，类似于λ1作为一个交易准确性和公平的旋钮。图4b显示的权衡曲线再次向左移动。

**验证集大小图4c到4e显示了验证集大小如何影响FR-Train的鲁棒性。**特别地，我们比较了FR-Train在干净数据和有毒数据上的准确性-公平性权衡

的验证集。当运行有毒数据时，我们修复了λ2 = 0.4和各种λ1. 我们的请参见此内容均匀5%的验证集（图4d）足以保持通过清洁数据获得的准确性和公平性。使用0.1%时（图示 4e)， 的 验证 设置 是： 也 小型的 以及 具有 安 不利

对培训的影响。然而，通过减少调谐旋钮λ2当下降到0.1时，我们可以弱化稳健训练，从而避免不利影响（图4e，绿色三角形）。这与RML形成了对比，后者在非常小的情况下遭受了不可忽略的性能下降

验证集。详见补充部分。

## 真实数据的结果

我们使用两个真实的数据集：Angwin等。和成人人口普查(Kohavi，1996)，分别有7214和45222个例子。我们使用与IBM的人工智能公平360(贝拉米等。并对两个数据集使用敏感属性SEX。对于数据中毒，我们使用了与在合成数据上使用的相同的方法：用z=1翻转标签，以最大限度地提高精度和性能下降。中毒量为D的10%*tr*.

虽然我们假设在以前的合成数据实验中有一个小而干净的验证集可用，但这样的假设在实践中并不成立。因此，对于真实数据实验，我们考虑一个场景，其中一个人首先构建一个小的(相当于5%*tr*)基于众包的验证集，然后用于FR-Train。我们在第4.5节中提供了关于如何构造这个验证集的详细信息。

*D*

表2和表3分别总结了各种训练算法在比较数据集和成人普查数据集上的公平性和精确性性能。在表1中，我们应用适当的超参数，以便在所有不同的方法中，对于干净和有毒的数据集都是相似的。结果与表1相似：由于数据中毒，三种公平性方法的影响和准确性更差；LR和RML的影响较差；FR-Train在公平性和准确性上都几乎没有下降。表2和表3还表明，将公平性方法与使用RML（第4-6行）的消毒处理相结合并不总是产生更好的准确性和公平性，甚至可能降低它们，这与合成数据的结果一致。人们可能会想知道，公平性基线是否会表现得更好

*具体详见表2。*比较测试数据的准确性和公平性性能。使用标签翻转攻击中毒训练数据的不同影响(DI)。比较了两种方法：

*具体详见表4。*中毒成瘾人口普查数据集上的混淆矩阵。不同的影响。其他设置与表相同 2.

处理方法 女性性的 男性的

（1）公平性方法：FC、LBC和AD，其中“RML+”表示

|  |  |  |
| --- | --- | --- |
|  | *yˆ=0* *yˆ=1* | *yˆ=0* *yˆ=1* |

预先使用RML进行消毒的应用；（2）非

公平性方法：LR和RML。对于FR-Train和RML，对于验证集为的5%*tr*。中毒的量为中毒的10%*tr*。对于中毒数据的每个结果，我们与

清理数据的结果，并显示百分比增加或 减少了。

处理方法 清理数据 中毒的数据

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 第 | 执行。 | 第 | 执行。 |  |
| fc lbc | .777  .866 | .682  .671 | 。794(2.2%t) | .612 (10.% l) |  |

.838 (2.8%l)（0.0%-）.671

.813 (6.1%l)(16.%l).570

广告广告的 .846 .680

rml+磅重数 .866 .671 .560 (28.% l)(5.4%l).645

rml+的频率 .777 .682

rml+广告 .846 .680 .869 (0.4%t)(3.7%l).646

.820 (3.1%l) .573 (16.% l)

rml+的频率

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| rml+广告 | *y=0*  *y=1* | 2,345  289 | 723  96 | 3,966  1,792 | 947  473 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| FR-列车 | *y=0*  *y=1* | 2,761  113 | 307  272 | 4,730  1,428 | 183  837 |

*具体详见表5。*关于比较试验数据的消融研究。使用标签翻转攻击中毒训练数据的不同影响(DI)。比较了四种方法：（1）FR-无R的列车(λ2 =0)，（2）FR-没有F的列车(λl= 0), (3)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| y=0 | 2,990 | 78 | 4,842 | 71 |
| *y=1* | 238 | 147 | 1,952 | 313 |

lr .465 .674 .454 (5.0%l)(6.4%l).631

右毫升 .493 .680 .575 (17.% t)(5.0%l).646

## FR-列车 .838 .676 .846 (1.0%t) .670 (0.9%l)

经过了清洁验证集的培训。在补充中，我们表明，性能实际上比表2和表3中的更差。这是因为干净的验证集太小，无法用作独立的列车数据。事实上，书中也有类似的观察结果。，2018b)。

*具体详见表3。*成人人口普查测试数据的准确性和公平性结果。不同的影响(DI)。其他设置与表2相同.

处理方法 清理数据 中毒的数据DI 执行。 第 执行。

fc .825 .826 .741 (10.% l)(3.0%l).801

lbc .825 .825 .760 (7.9%l)(4.0%l).792

|  |  |  |
| --- | --- | --- |
| 广告广告的 | .850 | .767 .755 (11.% l)(27.%l).563 |
| rml+的频率 | .825 | .826 .821 (0.5%l)(5.6%l).780 |
| rml+lbclml+广告  lr | .825  .850  .328 | .825 .762 (7.6%l)(4.5%l).788  .767 .834 (1.9% *l)* .647 (16.% *l)* |
| 右毫升 | .327 | .847 .189 (42.% l)(3.3%l).819  .846 .268 (18.% l)(0.7%l).840 |

## FR-列车 .828 .824 .847 (2.3%t) .809 (1.8%l)

表4显示了使用RML，使用有毒的成人普查数据集，对FR-Train、AD和FC之间的不同影响的混淆矩阵比较。当FR-Train、AD和FC实现（Acc、DI）=（0.809、0.847）、（0.647、0.834）和（0.780、0.821）时，报告结果，

分别为。FR-Train在所有方面都优于AD和FC，因为其鲁棒性鉴别器在消毒毒数据方面更有效。

FR列车没有示例重新加权（无RW）和（4）FR-列车。 对于 行数 2–4, 的 验证 设置 是： 5% 的 *tr* .

处理方法 清理数据 中毒的数据

第 执行。 第 执行。

无R型产品 .846 .678 .802 (5.2%l)(14.%l).580

无F型产品 .482 .681 .420 (13.% l)(7.2%l).632

无RW.832.677 .840 (1.0%t)(7.8%l).624

## FR-列车 .838 .676 .846 (1.0%t) .670 (0.9%l)

* 1. **消融术的研究**

在表5中，我们进行了消融研究，以研究FR列车的每个成分的影响。**（不带“R”）当λ时**2=0（即，没有健壮的训练），不同的影响很高，但是 测量精度 是： 低的 开启 的 中毒的 数据， 只是 类似 的 其他仅具有公平性的方法（表中：第2行、第1行至第3行）。(没有“F”)另一方面，当λ1=0（即无公平培训），准确性很高，但不同的影响是低的，就像其他不公平的方法一样（表中：2、第7行至第8行）。

（不需要重加权）最后，当不使用示例重加权时，精度和不同的影响都与FR-Rrain相似或更差。

综上所述，只有像FR-Train这样的整体框架才能实现优秀的模型公平性和训练鲁棒性。相比之下，其他只为其中一个目标定制的方法就失去了准确性、公平性或两者兼而有之。

## FR列车的误差范围

研究了自由列车的误差范围。对中毒数据的所有频率训练实验都用10个不同的随机种子重新进行，以产生平均的误差范围

(m)和标准偏差(s)值。在表6中以ms/2的形式报告了性能。在合成数据和成人人口普查数据集上，表现最低

*±*

（即ms/2）FR列车仍然分别优于表1和表3中的第二佳性能。对于比较数据集，FR-Train的最低性能比LBC相关算法的性能稍微差一些，

*-*

*具体详见表7。*众包标签的准确性比较（N：每个例子平均的答案数）和基于地面真相训练的逻辑回归模型的预测 标签。

数据集 众包业务 经过培训的模型

这可以解释为LBC算法

一开始就受到中毒的影响并不大。

*具体详见表6。*中毒数据集w.r.t上的错误范围。不同的影响(DI)。中毒的设置与以前的设置相同 实验结果。

数据集 中毒的数据

第 执行。

合成的0。795 ± 0.019 0.805 ± 0.008

N=1N=5N=11

对比线 0.609 0.656 0.667 0.659

成人人口普查0.6450.721 0.743 0.804

*详见表8。*在验证集使用众包标签时，FR-Train的准确性和公平性。的 培训数据中毒，详见表2和表3。

数据集 验证集为DIAcc。

众包业务 0.846 0.670

对比线 0.827 ± 0.027 0.653 ± 0.005

对比线

基本真相 0.899 0.674

成人人口普查 0.871 *±* 0.034 0.796 ±0.006

成人人口普查 众包业务 0.847 0.809

## 构建一个干净的验证集

我们现在演示如何使用众包构建一个干净的验证集。我们使用亚马逊机械土耳其人(AMT)为比较和成人人口普查数据集构建了验证集。虽然这些数据集都有标签，但我们假设它们不能作为干净的数据使用。我们还将数据集作为社区资源发布（参见描述和数据的补充部分），并相信我们的构造可以推广到其他数据集。虽然众包并不是构建一个干净的验证集的唯一方法，但它对我们的目的来说已经足够了。

我们通过要求一个工人对每个例子进行分类来为每个数据集设计AMT任务。对于成人人口普查数据集，工人会查看一个人的各种属性，并预测一个人的收入是否至少为5万$。答案不是是/否，答案必须是1到4的比例，这更准确地反映了工人的意见。比较数据集也有类似的设置，唯一的区别是，工人们需要预测一个罪犯在两年内是否会再次犯罪。每个任务显示大约30个问题，我们每个答案支付3美分。对于质量控制，每个任务还包含教育工人的小测验，还有一些问题被用来评估工人的表现。在收集到答案后，我们过滤掉表现不佳的人，平均每个问题最多有固定数量的N个答案，并与阈值2.5进行比较，以生成最终的标签。如果过滤掉不准确的工人，每个问题的答案数量可能会小于N。我们使用了美国、加拿大和英国所有人口统计数据的工人。虽然这种多数投票方法在我们的实验中已经很好，但我们还可以应用各种质量控制技术，如同行评审，这些技术可以进一步减少偏见(Karger等。, 2011).

重要的问题是众包标签的准确性，以及构建的验证集是否有结果

基本真相 0.864 0.809

在很高的准确性和公平性的免费培训。表7显示了当N从1增加到11时，众包标签的准确性。即使是最高的准确性，这些预测也不是完美的，因为工人们在没有任何其他上下文的情况下查看有限的信息（即，只有特征）。为了看看工人们是否能做得更好，我们还在地面真实标签上训练逻辑回归模型，并在测试数据上显示它们的准确性作为上界。因此，当比较和成人普查数据集的N=11时，准确性是可比的。因此，我们对所有的实验都使用这个设置。表8显示了我们所构建的验证集与使用由地面真实标签构成的相同大小的“完美”验证集相比的有用程度。对于这两个数据集，使用地面真实验证集会稍微高一些，但可比较的不同影响，同时获得几乎相同的精度，这证明了对FR-Train使用众包验证集是合理的。

# 相关的工作

歧视模型公平的概念有很多定义，通常来自于人们想要保证的某些社会目标。因此，已经提出了许多公平措施（Verma和鲁宾，2018年）。当我们关注的是群体公平，这确保了两个敏感群体之间的相似的统计数据，一个有趣的未来工作是考虑个人公平(Dwork等。，这保证了在附近的例子中也有类似的预测结果。最近，关于缓解不公平技术的研究也在激增。，2018b)。根据修复程序发生的位置，主要有三种方法：

（1）预处理技术(卡米兰和考尔德斯，2011年；duPinCalmon等。，2017年；Zemel等人。，2013年；费尔德曼等人。修复培训数据；（2）内处理技术(Zafar等。，2017年；江和纳丘姆出版社，2020年；张

等人。，2018年a；神岛等人。，2012年；科特等人。，2019年；2018年；阿加瓦尔等人。，在模型培训过程中解决了这个问题；以及（3）后处理技术(Hardt等。，2016年；Pleiss等人。，2017年；卡米兰等人。，2012年；Chzhen等人。在维护模型的同时操纵预测。在这三种技术中，内处理技术的优点是可以处理任何数据，并且对模型训练有更多的控制（文卡塔苏布拉曼尼亚，2019）。

虽然不是我们的直接关注点，但在公平研究方面还有其他值得注意的方向。基于因果性的公平性(基尔伯图斯等人。，2017年；Kusner等人。，2017年；张&巴林博伊姆，2018年；Nabi&Shpitser，2018；Khademi等。，2019年；)建议如何理解属性之间的因果关系，以克服非因果方法的局限性。正如非因果公平可以被相互信息捕获一样，我们怀疑因果公平和定向信息之间可能存在联系。另一种重要的方法(桥本等人。是基于分布鲁棒优化(DRO)(Sinha等。，关注敏感属性z未知时。基于dro的公平方法通过在不知道z的情况下均衡所有分布上的风险来确保公平的结果，但它并没有直接最小化公平指标，如不同的影响和公平的概率。相比之下，FRTrain假设了对z的完全了解，并利用它直接最小化公平指标。

正如我们在第2节中所演示的那样，现有的公平技术并不是为健壮的训练而定制的，所以它们很容易受到数据中毒的攻击。相比之下，FR-Train在同一模型训练过程中同时处理模型的公平性和鲁棒训练，因为它们密切相关，并受到相同训练数据的影响。

关于如何使模型训练对噪声甚至敌性数据的鲁棒性，有大量的文献。，2013年；Biggio等人。，2011年；Fre´nay&Verleysen，2014年；Kurakin等。, 2017).一个主要的挑战是，可能会有广泛的数据中毒攻击，并不断发展。虽然在模型训练之前消毒训练数据是一种选择，但防御所有可能的攻击似乎根本不可行，如(Koh等人所证明的那样。，2018)。最近的一个趋势是为在使用元学习的模型训练过程中的任何攻击开发通用的防御算法。，2017年；李等人。，2017年；肖等人。，2015年；亨德里克斯等人。，2018)。我们的FR-Train框架的灵感来自于使用元学习的鲁棒性训练(Ren等人。，但使用了一个基于GAN的模型来支持公平和健壮的培训，而不使用元学习。特别是，FR-Train的鲁棒性鉴别器的设计是基于基于相互信息的理论见解（第3.2节）。另一项研究是在测试期间防御对抗性攻击

gio等人。，2013年；古德费罗等人。，2015年；黄与科尔特出版社，2018年）。相比之下，我们的重点是防止训练数据上的数据中毒。

# 试验结论

我们提出了FR-Train，这是一个值得信赖的人工智能的整体框架，通过执行不公平的缓解和稳健的训练。我们的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，它具有结合两种方法的协同效应：（1）使用公平区分器来区分预测w.r.t。一个敏感组从其他组和（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，我们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。在我们的实验中，我们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对中毒具有稳健性，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。

# 确认情况

卢裕二和史蒂文E。Whang获得了谷歌人工智能重点研究奖和韩国国家研究基金会(NRF)资助的工程研究中心项目(NRF-2018R1A5A1059921)。该材料是基于美国空军科学研究办公室在奖项编号FA2386-19-1-1-4050下支持的工作。

# 参考文献

阿加瓦尔公司， A., 贝格尔齐默， A., Dud´ık，朗格福德和沃拉赫，公平分类的削减方法。ICML，第60-69页，2018年。

安格温，拉尔森，马图，S.，和基什内尔，L。机器偏见：全国各地都有一种软件被用来预测未来的罪犯。以及它对黑人的偏见。，2016年。

贝拉米，戴德，K.，金德，M.，霍夫曼，S.。 C., 霍德、坎南，洛伊亚、马蒂诺，梅塔，莫伊西洛维奇， A., 纳加尔，拉马默西，名词，理查兹，萨哈， D., 萨蒂吉里，辛格，瓦什尼，韩国和张。人工智能公平360：一个用于检测，理解、和减轻不想要的算法偏差的可扩展的工具包，2018年10月a。网址<https://arxiv.org/abs/1810.01943>.

贝拉米，戴德，K.，金德，M.，霍夫曼，S.。 C.,ᾏ

侯德、坎南、洛伊亚、马蒂诺、梅塔、莫伊西洛维奇，纳加尔，拉马穆西，K。名词，理查兹，

J.T.，萨哈，博士，萨蒂杰里，P.，辛格，M.，瓦什尼，K。R.，和Zhang，Y。人工智能公平性360：一个可扩展的工具包，用于检测、理解和减轻不需要的算法偏差。科，阿拉伯/1810.01943,2018b。

大乔， B., 尼尔逊先生， B., 和拉斯科夫，P支持对抗性标签噪声下的向量机。在ACML，第97-112页，2011年。

大乔， B., 科罗娜，我，迈奥卡岛， D., 尼尔逊先生， B., 斯兰迪奇，N.，拉斯科夫，P.，贾辛托，G.，和罗里，F。在测试时，对机器学习的规避攻击。在ECML PKDD中，pp。387–402,2013.

丘尔德科瓦， A. 还有罗斯， A. 机器学习中公平的前沿。CoRR，abs/1810.08810,2018年。

切珍，E.，丹尼斯，C.，赫比里，奥内托，L.，和庞蒂尔，

M.利用标记和未标记数据进行一致的公平二进制分类。在NeurIPS中，pp。12760–12770.2019.

科特，A.，古普塔，M.。R.，江，H.，斯雷布罗，N.，斯里德哈兰，K.，王，S.，伍德沃斯，B.E.，和你，S.培训充分推广公平度量和其他数据源约束的分类器。CoRR，abs/1807.00028,2018年。

有效的非凸约束优化。在ALT中，pp。300–332,2019.

杜平卡尔蒙，F，魏， D., 黄杉， B., 拉马默西，

K.和瓦什尼。R.优化了预防歧视的预处理程序。在NeurIPS中，pp。3995–4004,2017.

德沃克，哈特，先生，皮塔西，雷林戈尔德和泽梅尔，

R.的公平，通过意识来实现。在ITCS中，pp。214–226,2012.本文件编号：978-1-4503-1115-1。

费尔德曼，M，弗里德勒，S。 A., 莫勒，J.，谢德格， C., 和文，S。确认和消除不同的影响。在KDD中，第259-268页，2015年。

´不，´。还有弗莱森，M。标签噪声的分类：调查。*IEEE横向。神经网络。学习系统系统。*, 25(5):845–869, 2014.

古德费罗，阿巴迪，米尔扎，徐， B., 法利， D., 奥泽尔，南，考维尔， A.ᾏ C., 本加，

Y.生成的对抗性网。在NeurIPS中，pp。2672–2680,2014.

古德费罗、舍伦斯和斯泽格迪，解释和利用敌对的例子。2015年，在ICLR。

哈特，先生，普莱斯，E和斯雷布罗，N。有监督学习中的机会平等。*在NeurIPS中，pp。* 3315–3323, 2016.

桥本、斯里瓦斯塔瓦、南孔、h、梁，

P.公平没有人口统计在重复损失最小化。在ICML中，pp。1929–1938,2018.

风鸡， D., 马泽卡，M.，威尔逊， D., 和Gimpel，K.使用可信数据在被严重噪声损坏的标签上训练深度网络。在NeurIPS中，第10477-10486页，2018年。

ibm. 信任ai。 [https://www.research.](https://www.research.ibm.com/artificial-intelligence/trusted-ai/) [ibm.com/artificial-intelligence/](https://www.research.ibm.com/artificial-intelligence/trusted-ai/) [可信信](https://www.research.ibm.com/artificial-intelligence/trusted-ai/), 2020.

识别和校正机器学习中的标签偏差。*在AISTATS中，pp。* 702–712, 2020.

卡米兰，F.。和呼叫者，T。分类分类数据预处理技术。*知识知识。国际。系统节。*, 33(1):1–33, 2011.

卡米兰，F.，卡里姆，A.，和张，X。识别感知分类的决策理论。在ICDM中，pp。924–929, 2012.

神岛，T.，麻浩，S.，麻生，H.，和坂井，J.公平意识与消除偏见规则器的分类器。在ECML PKDD中，pp。35–50,2012.

卡格尔， D. R.，哦，S.，和沙阿， D. 为可靠的众包系统提供可靠的迭代学习。*在NIPS中，pp。* 1953–1961. 2011.

卡病，A。软件2.0。[https://medium.com/](https://medium.com/%40karpathy/software-2-0-a64152b37c35) [a64152b37c35](https://medium.com/%40karpathy/software-2-0-a64152b37c35), 2017.

卡德米市。还有霍纳瓦尔，五世。 G.ᾏ累犯预测中的算法偏差：一个因果视角（学生摘要）。在AAAI中，pp。13839–13840,2020.

卡德米， A., 李，S，福利， D., 和荣誉，五、算法决策中的公平：通过因果关系的探索。WWW，2907-2914页，2019。

基尔伯图斯，罗罗纳斯，罗贾斯-卡鲁拉，帕拉斯坎多罗，哈特，詹津， D., 还有Scho¨lkopf， B.ᾏ通过因果推理来避免歧视。在NeurIPS中，pp。656–666,2017.

亚当：一种随机优化的方法。*arXiv预印本：1412.6980,2014年。*

高高。W.，斯坦哈特，J.，和梁，P。更强大的数据中毒攻击，破坏了数据消毒防御。CoRR，abs/1811.00741,2018年。

提高天贝叶斯分类器的精度：一种决策树混合。在KDD中，pp。202–207,1996.

库拉金，古德费罗，和本吉奥，对抗性机器学习。2017年，在ICLR。

库斯纳先生。洛夫图斯，罗素，C.，和席尔瓦，R。反事实的公平性。在NeurIPS中，pp。4066–4076. 2017.

李，Y、杨、J、宋、Y、曹、L、罗、李，L。从蒸馏标签中学习。*在ICCV中，pp。* 1928– 1936, 2017.

林，J.。基于香农熵的发散测度。

《IEEE《信息论学报》，37（1）：145-151，

1991.

纳比，R.。还有什皮瑟，我。对结果的公平推断。AAAI出版社，第页。1931–1940,2018.

纳塔拉扬，德希隆，美国，拉维库马尔和特瓦里。用嘈杂的标签来学习。在NeurIPS中，pp。1196–1204,2013.

谷歌数据集搜索：在一个开放的网络生态系统中为数据集构建一个搜索引擎。*第28届网络会议(WebConf，2019年)，2019年。*

帕斯克先生， A., 格罗斯、钦塔拉、钱、昌、杨、e、德维托、林、德斯梅森， A., 安提加，L.和勒勒，

A.ᾏ在PyTorch中的自动分化。*在NIPS自动分配研讨会上，2017年。*

波迪斯，˜。-Gonza´，L.和卢普，E。 C.ᾏ标签消毒，防止标签翻转中毒攻击。在ECML PKDD中，pp。5–15,2018.

普里斯、拉格万、M.、吴、克莱因伯格和温伯格、K.Q。关于公平性和校准的问题。在NeurIPS中，pp。5684–5693,2017.

任，M.，曾，w.，杨，B.，和乌尔图森，R.学习重新计算例子为强大的深度学习。在ICML中，pp。4331–4340,2018.

辛哈市， A., 南孔，H.，和杜奇，J。 C. 通过原则对抗训练证明一些分布鲁棒性。ICLR，2017年。

否决权， A., 奥尔德林，N，切克，克拉辛，我，古普塔， A., 从嘈杂的大规模数据集中学习。在CVPR中，第6575-6583页，2017年。

算法公平性：度量、方法和表示。在PODS中，pp。481,2019.

维尔玛和鲁宾，J。公平性的定义被解释了一下。在

费尔韦尔@ICSE，页。1–7,2018.

黄，E.和科尔特，J。Z.可以通过凸外对抗性多面体来防御对抗性的例子。在ICML中，pp。5283–5292,2018.

肖、夏、T、杨、Y、黄、王，X。从大量的有噪声的标记数据中学习，以进行图像分类。在CVPR中，pp。2691–2699,2015.

扎法尔，M.B.，瓦莱拉，I.，戈麦斯-罗德里格斯，M.，和古马迪，K.P.公平约束：公平分类的机制。在AISTATS中，pp。962–970,2017.

泽梅尔，R.S.，吴，Y.，斯威尔斯基，K.，皮塔西，T.，和德沃克，

C. 学习公平的表现法。在ICML中，pp。325–333,2013.

张先生， B. H.、柠檬素， B., 还有米切尔，用对抗性学习来缓解不必要的偏见。在AIES，第335-340页，2018a。

张，J。还有巴林博伊姆。决策中的公平性-因果解释公式。2018年，在AAAI。

张，朱和赖特。J.培训集使用受信任的项目进行调试。在AAAI，2018年，b。