公平性与准确性之间是否存在权衡？使用错配假设检验的观点

**Sanghamitra Dutta** 1 2 **Dennis Wei** 1 **Hazar Yueksel** 1 **Pin-Yu Chen** 1 **Sijia Liu** 1 **Kush R. Varshney** 1

# 摘要

在现有的关于机器学习中的公平性的文献中，准确度和公平性之间的权衡是最被认为是一种必然的结果。然而，这并不是预设的，准确性应该随着公平性的提高而降低。新颖的是，我们通过*错配假设测试*的视角来研究公平分类：当给定两个有偏差的错配distribu- tributions时，试图找到一个能够区分两个理想dist- tributions的clas- siﬁer。使用信息理论中的工具Chernoff informa- tion，我们在理论上证明，与流行的看法相反，总是存在理想的分布，这样，操作的公平性和准确性（相对于理想分布）是同时实现的：不存在权衡。此外，同样的类别- siﬁer产生了缺乏与理想分布的权衡，同时产生了一个权衡，当精度测量与给定（可能有偏见）数据集。为了补充我们的主要结果，我们制定了一个优化来确定理想分布，并推导出基金的限制来解释为什么在给定的有偏数据集上存在权衡。我们还推导出在现实世界中，主动收集数据可以缓解公平性-准确性权衡的条件。我们的研究结果使我们认为，用有偏差的数据来衡量准确性是有问题的，相反，我们应该考虑的是理想的、无偏差的数据的准确性。

# 介紹

这项工作解决了算法公平性的一个基本问题（[Calmon等人](#_bookmark42)，[2017](#_bookmark42)；[Dwork等人。](#_bookmark32)

1IBM Research 2卡内基梅隆大学电气与计算机工程系。通讯作者：Sang- hamitra Dutta < >:Sang- hamitra Dutta *<* [sanghamd@andrew.cmu.edu>.](mailto:sanghamd@andrew.cmu.edu)

*Proceedings of the 97 th International Conference on Machine Learning*, Online, PMLR 119, 2020.Copyright 2020 by the au- thor(s).

[2012](#_bookmark32)；[Agarwal等人](#_bookmark29)，[2018](#_bookmark29)；[Hardt等人](#_bookmark46)，[2016](#_bookmark46)；[Ghassami等人](#_bookmark43)，[2018](#_bookmark43)；[Kusner等人](#_bookmark55)，[2017](#_bookmark55)；[Kilbertus等人](#_bookmark53)，[2017](#_bookmark53)；[Zemel等人](#_bookmark56)，[2013](#_bookmark56)）。)

*公平性和准确性之间是否有取舍？*

这种权衡的存在已经在sev- eral现有的作品中被指出（[Menon& Williamson，](#_bookmark59)2018[；](#_bookmark59)Chen [etal.，](#_bookmark27)2018[；](#_bookmark27)Zhao [& Gordon，2019），这些](#_bookmark57)作品也提出了不同的理论方法来描述它。然而，在公平性和准确性之间为什么要存在这样的权衡，并没有预设。例如，[Friedler等人](#_bookmark39)（[2016](#_bookmark39)）和[Yeom& Tschantz（](#_bookmark51)2018[）](#_bookmark51)提出，机器学习模型中的观察到的特征（如测试成绩）是一个可能是来自抽象构造空间（如真实能力）中的特征的噪声映射，在这个空间中不存在这种权衡。那么，为什么纠正偏差会恶化现实世界中的预测准确性？我们认为，退一步重新提出这个基本问题是有价值的。

在这项工作中，我们的主要论断是，现实世界中准确性和公平性（尤其是平等的对等性（[Hardt et al.](#_bookmark46) , [2016](#_bookmark46)））之间的权衡是由于机会、代表性等方面的历史性差异，导致无权群体的映射更加嘈杂（因此也有偏见），使得他们的正负标签"不太可分"。为了将这一观点具体化，我们采用了一种新颖的公平分类观点：错配假设检验的观点。在错配假设检验中，我们的目标是找到一个能区分两个"理想"分布的分类器，但相反，人们只能获得两个有偏差的错配分布。我们最重要的结果是在理论上表明，对于一个在给定的有偏数据分布上具有次优精度的公平分类器来说，总是存在着理想分布，这样，当对理想分布的精度进行测量时，公平性和精度是一致的。通过这个观点，公平性和准确性之间不存在权衡。

我们在这项工作中的贡献如下：

*分离性的概念来量化现实世界中的准确性-公平性权衡。*对于一个ob- served数据集中的一群人，我们用Chernoff信息将"可分离性"量化为正负类标签，这是一个信息理论上的最佳指数的近似值。

二进制分类的错误概率。我们证明（在定理[1](#_bookmark14)中），如果在观察到的数据集中，一个组的Chernoff信息低于另一个组的信息，那么使用组公平性标准修改最佳分类器会损害一个或两个组的错误指数（代表准确性），解释了准确性-公平性的权衡。这些工具不仅证明了权衡的存在（在一些存在- ing的作品中也有证明（[Menon & Williamson，](#_bookmark59)2018[；](#_bookmark59)Chen[等人，](#_bookmark27)2018[），](#_bookmark27)使用替代公式），而且它们还使我们能够近似地量化权衡，例如，我们可以使两个组的假阴性概率多么接近，以试图达到对准确性进行一定妥协的平等机会（见第4节的图3）。这种权衡的存在促使我们认为，分类器相对于现有的(可能有偏差的)数据集的准确性是一个有问题的性能衡量标准。相反，我们应该考虑相对于一个理想的数据集的准确性，这个数据集是一个无偏见的群体代表。

*理想的分布，公平性和准确性是在ac- cord。*新颖的是，我们通过错配假设检验的角度来研究公平分类的问题。我们表明(在定理[2](#_bookmark20)中)，存在理想的分布，使公平性(在现有和理想分布上机会均等的意义上)和准确性(相对于理想分布)是一致的。我们还制定了一个优化方案来说明如何在实践中去寻找这样的理想分布。理想分布提供了一个目标，以转移给定的偏态分布，并评估精度。它们的解释有两个方面。(i)从构造空间的"无偏"映射所产生的观测空间的可信分布；或(ii)构造空间本身的候选分布(在第[3.2](#_bookmark17)节进一步讨论)。

*标准来缓解现实世界中的准确性-公平性权衡。*接下来，我们还解决了另一个重要的问题，即，在现实世界中，我们什么时候可以通过额外的数据收集来减轻准确性和公平性的权衡，具体来说，我们必须在现实世界中工作。我们推导出一个信息理论标准（在定理[3](#_bookmark24)中），在这个标准下，收集更多的特征可以提高可分离性，从而提高现实世界中的准确度，缓解折衷。这也可以为我们选择理想分布提供参考。我们的分析可以作为主动公平性成功的技术解释（[Noriega-Campero等人](#_bookmark61)，[2019](#_bookmark61)；[Bakker等人](#_bookmark33)，[2019](#_bookmark33)；[Chen等人](#_bookmark27)，[2018](#_bookmark27)），使用额外的特征来提高公平性。

*数值实例。*我们通过一个例子(有分析性的封闭式)来展示分析如何进行。

**相关工作。**我们注意到，现有的几项工作，如[Garg等人](#_bookmark41)（[2019](#_bookmark41)）、[Menon & Williamson（](#_bookmark59)2018[）、](#_bookmark59)Chen[等人（](#_bookmark27)2018[）](#_bookmark27)和Zhao [& Gordon（2019），](#_bookmark57)也使用了in-。

形成理论或贝叶斯风险来描述准确性和公平性的权衡。然而，计算贝叶斯风险并不简单。事实上，即使对于高斯人，人们也会求助于Chernoff界限来近似Q函数。切尔诺夫信息是贝叶斯风险的近似值，它有一个易懂的几何解释（见图[2](#_bookmark16)）。这使我们能够在数值上计算出准确度-公平性的权衡（图[3](#_bookmark26)），也能理解通过数据收集可以提高"多少"准确度，超越了有一些改进的论断。据我们所知，现有的工作已经指出了基于贝叶斯风险的权衡的存在，但并没有提供一种ex- actly计算的方法，这促使我们引入额外的工具Chernoff信息来近似计算。Fur- thermore, this work goes beyond characterizing the trad-off imposed by the given dataset.我们的新颖之处在于采用了不匹配检测的观点，并证明存在理想的分布，从而在对理想分布进行精度测量时，公平性和准确性都是一致的。

[Wick等人](#_bookmark49)([2019](#_bookmark49))和[Sharma等人](#_bookmark64)([2020](#_bookmark64))最近的工作进一步阐明了定理[2](#_bookmark20)的意义，以及它如何提出了一个与"主流智慧"相矛盾的见解，即存在一个公平性和准确性一致的理想数据集。在某种意义上，我们的工作提供了一个理论基础，补充了[Wick等人](#_bookmark49)([2019](#_bookmark49))和[Sharma等人](#_bookmark64)([2020](#_bookmark64))的经验结果，明确了何时存在权衡，何时不存在。

也有几种现有的预处理数据的方法来生成公平的数据集（[Calmon等](#_bookmark44)，[2018](#_bookmark44)；[Feld- man等](#_bookmark36)，[2015](#_bookmark36)；[Zemel等](#_bookmark56)，[2013](#_bookmark56)）。在这里，我们的目标不是提出另一种通过预处理实现公平的竞争策略。相反，我们的重点是在理论上证明存在一个理想的数据集，这样一个公平的分类-ﬁer在准确性方面也是最优的，这在之前还没有正式证明。我们还关注机会均等而非统计均等（如[Calmon等人](#_bookmark44)（[2018](#_bookmark44)））。

我们的工具与[Varshney等人](#_bookmark65)([2018](#_bookmark65))(展示了可解释性如何改善Chernoff in- formation)以及一般的假设检验理论([Lee & Sung,](#_bookmark58) 2012[;](#_bookmark58) [Cover & Thomas, 2012)](#_bookmark30)有相似之处[。](#_bookmark30)我们的贡献在于在公平的机器学习- ing中使用这些工具，在那里它们还没有被用于我们所知的最好的边缘（例如，在[Menon＆Williamson（](#_bookmark59)2018[）](#_bookmark59)的先前分析中[；](#_bookmark59)[赵和戈登（2019）；](#_bookmark57)陈等[人（2018））。](#_bookmark27)

**备注1**（群体设置）**。***在这项工作中，我们在群体环境中操作（动机来自*[*Gretton等人*](#_bookmark45)*（*[*2007*](#_bookmark45)*）；*[*Ravikumar等人*](#_bookmark62)*（*[*2009*](#_bookmark62)*）；*[*Scott等人*](#_bookmark63)*（*[*2013*](#_bookmark63)*）），即随着样本数量的限制，允许使用数据的概率分布。这允许我们 量化*

*准确度和公平性权衡的基本限制。事实上，给定任何分类器，总是存在一个至少同样好的似然比检测器（见*[*Cover & Thomas（*](#_bookmark30)*2012）中的NP Lemma*[*）。*](#_bookmark30)

# 前言

**设置。**在这项工作中，我们专注于二元分类， 这出现在实践中常见的公平性文献， 例如，在决定是否应该接受或拒绝候选人 在应用程序，如招聘，贷款等。我们让Z表示受保护的属性，例如，性别、种族等。

在不失一般性的前提下，让Z=0为无特权组，Z=1为特权组。

受[Yeom& Tschantz（](#_bookmark51)2018[）](#_bookmark51)和[Friedler等人（](#_bookmark39)2016[）的](#_bookmark39)启发[，](#_bookmark39)我们假设存在一个抽象的建构空间，其中Xa是特征（如真实能力），Ya是真实的

标签（即取值0或1）。构造空间不是

我们可以直接访问。在现实世界中，我们反而有

访问一个观察到的空间，其中X表示特征向量，Y表示真实标签（即取值0或1）。为了简单起见，我们基于[Yeom& Tschantz（](#_bookmark51)2018[）](#_bookmark51)假设*Ya* = Y[。](#_bookmark51)[1](#_bookmark2) 观察到的特征是

从构造空间中的特征导出如下。X = *fY,Z*(*Xa*) 其中*fY,Z*( )是一个可能有噪声的映射，它可能取决于Y和Z。

*.*

让给定数据集中的特征在观测空间中具有如下分布。 X*|Y*=0*,Z*=0*~*P0(x)，并且

X*|Y*=1*,Z*=0*~*P1(x)。 同理，X*|Y*=0*,Z*=1*~*Q0(x)，且

X*|Y*=1*,Z*=1*~*Q1(x)。对于每个组Z=z，我们将分类器表示为*Tz*(x)*>τz*，即当Tz(x*)*>*τz*时，预测标签为1，否则为*0*。

**注2**（解耦分类器）**。***虽然这类模型可能会表现出不同的待遇（明确使用*Z*），但其目的是在决策中明确使用受保护的属性，以更好地缓解不同的影响（沿着公平的afﬁrmative行动的精神（*[*Dwork等人*](#_bookmark32)*，*[*2012*](#_bookmark32)*；*[*2018*](#_bookmark34)*））。 Fur-*

*热，如果Tz和*τ*z对两组都相同，那么一个不使用*Z的*分类器就成为我们分类器的特例。*

接下来，我们说明两个基本假设。(**A1**) Absolute Con-

tinuity。P(x)、P(x)、Q(x)和Q(x)均大于：

我们让PFP*,Ts*(τ*z*)为假阳性的概率(错误地接受负类标签；也称为

假阳性率(false positive rate，FPR))在组Z＝z上的概率，即PFP*,Ts(*τ*z*)＝Pr(*Tz*(X) τ*z* Y＝0，Z＝z)。同理，PFN*,Ts*(τ*z*)是假阴性(错误地拒绝阳性类标签的概率；也称为

*> |*

假阴性率(FNR))，给定。 PFN*,Ts* (τ*z*) = Pr (*Tz*(X) < τz Y = 1, Z = z)。一组的总体误差概率由以下公式给出。*Pe,Ts* (τ*z*) = π0PFP*,Ts (*τ*z*) + π1PFN*,Ts* (τ*z*), 其中π0和π1是给定Z = z的Y = 0和Y = 1的先验概率。为了简单起见，我们考虑给定Z = z的π0 = π1 = 1的情况，也考虑所有组Z = z的先验相等。

*|*

2

附录E中的不等前值。等前值也对应于平衡精度测量（[Brodersen等，2010](#_bookmark40)），它通常比普通精度更受青睐。

公平性的一个著名的定义是*均衡赔率*([Hardt et al.](#_bookmark46) , [2016](#_bookmark46))，它指出，如果一个算法对两组人的假阳性(错误地接受真阴性类标签)和假阴性(错误地接受真阳性类标签)的概率相等，那么这个算法就是公平的。

即Z=0和1.文献中广泛使用的这一衡量标准的一个宽松变体是*机会均等*，它强制执行了

只有两组的假阴性率相等（或等价的真阳性率相等）。在这项工作中，我们主要关注机会均等，尽管这些论点也可以扩展到其他公平性的衡量标准，例如，st- tistical parity（[Agarwal et al.](#_bookmark29) , [2018](#_bookmark29)）。

我们假设在构造空间中，准确度和机会均等之间不存在权衡，即群体的贝叶斯最优（[Cover & Thomas，](#_bookmark30)2012[）](#_bookmark30)分类器。

Z=0和Z=1也满足机会均等（假阴性概率相等）。在这项工作中，我们的目标是

是为了解释观察空间中的准确度和公平度的权衡，并试图确定不存在权衡的理想分布。现在，我们提供了一个关于分类器的误差指数的简要背景，以帮助跟随本文的其余部分。

**分类器的误差指数背景。**F PR和 FNR 的 误差 公式 给出

- 对数PFP*,Ts*（τ*z*）和*-*对数PFN*,Ts*（τ*z*）。 通常情况下，我们可以

0 1 0 1

这确保了似然比检测器，如对数P1(*x*)*>τ0*和Kullback-Leibler

P0(*x*)

(KL)之间的任何两个分布的分歧都是确定的。 (**A2**)不同的假设。 D(P0*||*P1)：

D(P1*||*P0)、D(Q0*||*Q1)和D(Q1*||*Q0)严格大于

不能得到准确的误差概率或其指数的闭式表达式，但指数可以用一个著名的下界*Chernoff界来逼近*（见Lemma [1](#_bookmark4)；附录A.1中的证明），众所周知，这个下界是非常紧密的（见Remark [3](#_bookmark5)，也见[Motwani & Raghavan（](#_bookmark60)1995[）；](#_bookmark60)[Berend & Kontorovich（2015））。](#_bookmark35)

比0，其中D( *.||.* )为KL离散度。

1这与[Yeom & Tschantz（](#_bookmark51)2018[）](#_bookmark51)中的"所见即所得"世界观是一致的，标签偏见可能是

**定义1.**P PFN*,Ts*(τ*z*)*的切诺夫指数被定义为：。*

FP*，Ts*

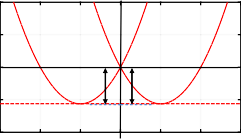
(τ*z*)*和*

忽视了我们所选择的公平衡量标准，即机会均等作为公平的衡量标准是合理的。

EFP*,Ts* (τ*z*) = sup(uτ*z)* Λ0(u))，*以及*

*u> 0*

*-*

2222



五FP

EFN

）

〇(u

⇤

)

l(u)

⇤



**对数生成函数**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  | 历 | FN | 历 | FP |  |
|  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  | E | FN | E | FP |  |
|  |  |  |  |  |  |

11

00

-1-1

-2u-2

11

00

EFNE

-1

FP

-1

-2-2

-1.5 - 11.5

-1.5 -1 -0.5

u 0 0.5 1 1.5

-1.5 -1-0.5

u 0 0.5 1 1.5

-1.5 -1 -0.5 u 0 0.5 1 1.5

*图1.*设P0(*x*)*～N*(1*，*1)，P1(*x*)*～N*(4*，*1)。对于似然比检测器*T*(*x*)=log P1(*x*)*>τ ，*我们可以计算出对数-。

P0(*x*)

生成函数如下。Λ0(*u*)=9 *u*(*u \_* 1)，Λ1(*u*)=9 *u*(*u* + 1)(在附录A.3中得出)。注意，Λ0(*u*)是严格的

2

2

凸，在*u*=0和*u*=1处为零，且Λ1(*u*)=Λ0(*u*+1)。我们得到EFP*,T* (*τ* )和EFN*,T* (*τ* )分别为切线Λ0(*u*)和Λ1(*u*)的y-截距的负值，斜率为*τ* 。当我们改变切线 (*τ* )的斜率时，EFP*,T* (*τ* )和EFN*,T* (*τ* )之间会有一个权衡，直到它们都在*τ* = 0时变得相等（左起第三个参数）。*τ* =0时的指数值(0斜率切线的y-截距的负值)被定义为Chernoff信息，由以下公式给出。C(P0*,*P1):=EFP*,T*(0)=EFN*,T*(0)*,* 在这个例子中等于9/8.

EFN*,Ts* (τ*z*) = sup(uτ*z)* Λ1（u））。

*-*

*u< 0*

*这里，*Λ0(u)*和*Λ1(u)*被称为对数生成函数，由*Λ0(u)=对数E[*euTs*(*X*)*|*Y=0，Z=z]*和*Λ1(u)=对数E[*euTs*(*X*) *|*Y=1，Z=z]*给出*。

*误差Pe,Ts* (τ*z*)*的能力被定义为：*

*Ee,Ts* (τ*z*) = min*{*EFP*,Ts* (τ*z*), EFN*,Ts* (τ*z*)*}*。

回想一下，在相等的前值下，我们有*Pe,Ts*(τ*z*)=。

1 PFP*,T* (τ*z*) + 1 PFN*,T* (τ*z*)。 *Pe,T*(τ*z*)的指数为

**Lemma 1**（Chernoff Bound）**。***指数满足：s 2s2s*

PFP*,T (*τ*z )**< e-EFP,*Ts *(τ*s *)*，*PFN*,T (*τ*z )*< e-EFN*,*Ts* (*τs )*。

的误差指数的最小值为主。

P

*s s* (τ )和P (τ )，而它又受制于

**Remark 3**（Chernoff约束的严密性）**。***对于Gaus-*

FP*,Ts z*

FN*,Ts z*

*的分布，尾部概率的特征为： 1.*

*的Q函数，它的上界和下界都是以Chernoff指数为单位的，其常量因子不会对指数产生明显影响（*[*Coˆte'等人*](#_bookmark28)*，*[*2012*](#_bookmark28)*）。Bhattacharya 边界（Chernoff 边界的特例）同时具有贝叶斯误差指数的上下限（*[*Berisha 等*](#_bookmark37)*，*[*2015*](#_bookmark37)*；*[*Bhattacharyya*](#_bookmark38)*，*[*1946*](#_bookmark38)*；*[*Kailath*](#_bookmark50)*，*[*1967*](#_bookmark50)*）。*

**切尔诺夫指数的几何解释。**切尔诺夫指数比精确误差指数更有洞察力，因为它们的几何解释，我们在此讨论（更多细节见附录A.2）。

为了便于理解，我们参考图[1](#_bookmark3)，在图中我们用一个数字例子来说明Chernoff指数的概念。在一般情况下，对数生成函数是凸的

并在u=0时变为0（见附录A.2）。此外，如果一个探测器是乖巧的[2](#_bookmark7)，即E[*Tz*(X) Y=1，Z=z]>0，E[Tz(X*)* Y=0，Z=z]<0，则Λ0(u)和Λ1(u)是严格的凸，并在两边达到最小值。

*|*

*|*

FPR和FNR的切尔诺夫指数的最小值

(定义[1](#_bookmark1))。较高的*Ee,Ts*(τ*z*)表示较高的精度，即较低的*Pe,Ts*(τ*z*)。为了理解这一点，首先考虑Z=T0(x)=log P1(*x*)形式的同类罩比检测器。

P0(*x*)

0. 当我们改变τ0时，PFP*,*T0(τ0)和PFN*,*T0(τ0)之间存在一个权衡，即当一个增加时，另一个减少。在它们的切尔诺夫指数中也观察到类似的权衡情况

(见图[1](#_bookmark3))。当τ0=0(对于等前值)且PFP*,*T0(0)=PFN*,*T0(0)时，*Pe,*T0(τ0)最小。 对于这个τ0=0的最优值，FPR和FNR的Chernoff指数也变得相等，即EFP,T0 (0)=EFN*,*T0 (0)，达到E*e,*T0 (τ0)=min EFP*,*T0 (τ0)，EFN*,*T0 (τ0)的最大-um值。这个指数称为Chernoff informa-。

*{}*

tion ([Cover& Thomas,](#_bookmark30) 2012[)。](#_bookmark30)为了完整起见，我们在附录A.4中加入了[Cover& Thomas(](#_bookmark30)2012[)](#_bookmark30)中关于Chernoff信息的一个著名结果的证明。

**Lemma 2.***对于*Y＝0*下的两个假设*P0(x)*和*Y＝1*下的两个假设*P1(x)*，贝叶斯最优分类器的误差概率的切尔诺夫指数由以下公式给出。*

的原点。Chernoff指数E

FP*，Ts*

(τ*z*)和

*切尔诺夫资料*[3](#_bookmark8)*。*

EFN*,Ts*(τ*z*)可以得到斜率为τ*z*的Λ0(u)和Λ1(u)切线的y-截距的负值(对于τ*e*(E[T(X)*|*Y=0，Z=z]，E[T(X)*|*Y=1，Z=z]))。

*z*

*z*

*z*

C(P0，P1) = min logP0(x)

- ul

*∈*(0*,*1)*x*

*1-u*

P1(x)

*u*、.(1)

**定义2。***总的proba-的Chernoff指数。*

2对于一个探测器*Tz*（*x*）*τz，*当*Y*=1时，我们会期望*Tz*（*X*）是高的，而当*Y*=0时，*Tz*（X）是低的，这证明了准则E[*Tz*（*X*）*Y*=1*，Z*=*z*]*>0*和 E[ Tz*（X）Y*=*0，Z=z]*<0。

*>*

*||*

因为 乖巧。 似然比检测器

**目标。**我们感兴趣的*准确度*指标是*Ee,*T0(τ0)和*Ee,*T1(τ1)，因为*Pe,Ts*(τ*z*)的Chernoff expo- nent值越高，意味着Z=0和Z=1这两个组的准确度越高。我们对*公平的*度量感兴趣。

*ness*是FNR的切尔诺夫指数之差，即。

*T*0(*x*)=log P1(*x*)*>τ0*在假设A2下表现良好，在

P0(*x*)

第[2](#_bookmark0)节因为我们有E[*Tz*(*X*)*|Y*=1*，Z*=*z*]=*D*(P1*||*P0)

而E[*Tz*(*X*)*|Y*=0*，Z*=*z*]=*\_ D*(P0*||*P1)。

3当P0(*x*)和P1(*x*)是连续分布时，用*x*上的积分代替求和(见附录A.3)。

EFN*,*T0 (τ0) EFN*,T1* (τ1)(从平等对立性中得到启发)。当EFN*,*T0（τ0）EFN*,*T1（τ1）=0时，一个模型是*公平的*，并且随着这个模型逐渐变得越来越不公平。

*|-|*

*|-|*

数量*｜*EFN*,*T0 (τ0) *-* EFN*,*T1 (τ1)*｜*增加。

我们的第一个目标是在现有的真实世界数据集上，即给定观察到的数据集上，用我们的指标来量化最佳准确性-公平性权衡的基本限制。

分布P0(x), P1(x), Q0(x), 和Q1(x).接下来，我们的目标是确定理想分布，其中公平性和accu-。

当用理想分布来衡量精度时，RACY是一致的。

# 主要结果

### 可分离性的概念。现实世界中对准确性-公平性权衡的基本限制。

考虑到第[2](#_bookmark0)节中的设置，我们表明，在观测空间中准确度和机会均等之间的权衡是由于非特权群体的噪声映射使得他们的正负标签不那么可分离。让

第一种情况是，从可分离性的角度来看，映射是无偏见的，在准确性和公平性之间没有权衡。第二种情况，在实践中比较常见，是由于数据集的固有局限性造成的歧视：来自构造空间的映射是有偏向的，并且没有足够的关于一个群体的分离性信息。

到另一个。在本文的其余部分，我们将重点讨论C(P0，P1)<C(Q0，Q1)的情况。在这种情况下，两组贝叶斯最优检测器的FNR的Chernoff指数为C(P0，P1)和C(Q0，Q1)，它们是

不平等，因此*不公平*。试图确保公平

因此，通过对任何一个组使用任何一个替代的似然比检测器，只会降低该组的准确度（错误概率的Chernoff指数），低于该组的Bayes最优（最佳）分类器，解释了准确度-公平性的权衡。我们在Lemma [3](#_bookmark10)中形式化了这一直觉（用于定理[1](#_bookmark14)的证明；见附录B）。

**Lemma 3.设C(P0，P1)< C(Q0，Q1)。***设*C(P0，P1)< C(Q0，Q1)。*假设有两个似然比检测器*T0（x）τ0*和*T1（x）τ1*。*

*> >*

*每组一个，使*EFN*,T*(τ0)=EFN*,T*(τ1)。

0 1

我们首先正式地定义一下我们的直观的可分离性概念。

**定义3.***对于在假设*Y=0*和*Y=1*下具有分布*P0(x)*和*P1(x)*的一群人，我们将其可分离性定义为他们的Chernoff信息*C(P0，P1)*。*

定义[3](#_bookmark11)是从Lemma [2](#_bookmark6)开始的，因为Chernoff信息本质上提供了一个信息理论上的近似，即在给定的数据集中，对一组人的最佳分类精度（在外在意义上）。接下来，我们从可分离性的角度来定义无偏映射。

*那么，下列说法中至少有一项是真的。*

*(一)Ee,*T0(τ0)<C(P0，P1)，*或(二)Ee,*T1*(*τ1)<C(Q0，Q1)。

接下来的两个结果显示了当前合理的公平分类方法是如何引起Lemma [3](#_bookmark10)中的两种情况的。考虑下面的优化问题，其目标是筛选出以下形式的分类器。

T0(x) τ0 和 T1(x) τ1，这两组的误差概率的 Chernoff 指数最大。

*> >*

的约束，即它们在给定数据集上是*公平的*。

**定义 4.***考虑第*[*2*](#_bookmark0)*节中的设置。映射*

*如果*C(P0，P1)=C(Q0，Q1)，则*从构造空间到观察**空间的*X=*fY,Z*(*Xa*)*说是无偏的*。

最大

T0*,τ*0*,T1,τ*1。

min *{*EFP*,*T0（τ0），EFN*,*T0（τ0）。

EFP*,T1*(τ1)，EFN*,*T1(τ1)*}。*

我们接下来的结果表明，公平性和准确性之间的权衡是由于从可分离性的角度来看映射的偏差引起的，即C(P0，P1) C(Q0，Q1)。Be

因为我们假设Z=0是无权组，所以我们

设C(P0，P1)等于，或小于C(Q0，Q1)。

**定理1**（解释权衡）**。***对于第*[*2*](#_bookmark0)*节中的设置，下列情况之一为真。*

1. *无偏映射，即*C(P0，P1)=C(Q0，Q1)*。该*

*贝叶斯最优检测器*T（x）*>τ*，*T（*x）>*τ，为*

令EFN*,*T0(τ0)=EFN*,*T1(τ1)。 (2)

这种优化符合现有工作的精神（[Zafar等人](#_bookmark54) ，[2017](#_bookmark54)；[Agarwal等人](#_bookmark29)，[2018](#_bookmark29)；[Donini等人](#_bookmark31)，[2018](#_bookmark31)；[Celis等人](#_bookmark47)，[2019](#_bookmark47)），在公平性约束下最大化精度。 从NP Lemma中，我们知道，给定任何一个

分类器，存在一个似然比检测器，它的准确度至少是一样好的。如果我们将T0(x)和T1(x)限制为对数P1(*x*)形式的似然比检测器。

P0(*x*)

和对数Q1(*x*)，则([2](#_bookmark13))有唯一解(τ *\**，τ *\**)。

*Q*(*x*)01

0

11

**Lemma 4.***设*C(P ，P )< C(Q ，Q )，且*T(*x)和

*误差概率*C(Q0，Q1)(=C(P0，P1))*的Chernoff指数的两组也达到了公平性。*

*即：|*EFN*,T*(τ0)*-*EFN*,T*(τ1)*|*=0*。*

0 1 0 1 0

T1(x)*被限制为似然比检测器。那么检测器*T0(x)*>τ0*\**，T1(*x)>*τ1\**，*则解出了*

0 1*优化*([2](#_bookmark13))*是贝叶斯最优检测器的未。*

1. *偏置映射，即*C(P0，P1)<C(Q0，Q1)。*贝叶斯最优检测器*T0(x)*>τ0*，*T1(*x)>*τ1。*

*特权组(*τ0*\**=0*)和一个次优的检测器来检测*

*Ee,*T*(τ*1*\**)*<*C(Q0，Q1)*的特权组(*τ1\*>0)。

1

*两组数据不公平，即|*E

FN*,*T0

(τ0) *-*

EFN*,T1* (τ1)*|* = 0*.此外，没有似然比值去掉*

*tector可以提高非特权组的错误概率的Chernoff指数，超过*C(P0，P1)*。*

作为证明简图，我们参考图[2](#_bookmark16)（左）。设τ0*\**=0，这就保证了EFN*,*T0（0）=EFP*,*T0（0）=C（P0，P1）。

44



↪So\_21E4↩=(u)*Z*丰u

↪So\_21E4↩0(u〉之二0。

-

𝐸!" ,$! 𝜏% =

↪Lu\_1D438↩!" ,$" (↪Ll\_1D70F↩& )0。

↪Lu\_1D438↩！ ' ,$" (↪Ll\_1D70F↩) 。

1

& )

1之*2*～．

ī（u）*之*=ó

⇤．已

⇤



⇤0(u〉*)*=1

⇤0(0）*2*=0

𝐸!" ,$! 𝜏% =

-1

↪Lu\_1D438↩!" ,$" (↪Ll\_1D70F↩&)

0

1

𝐸! ' ,$!(𝜏%)

u〉*之*=1

(Z)*2*=0

⇤1(

⇤1

2 2

0 0

-2-2

4-4

1

*图2.无特权组的分布*设非特权组(*Z*=0)的分布为P0(*x*) (1*，*1)和P1(*x*) (4*，*1)。另外，让特权组的分布为Q0(*x*) (0*，*1)和Q1(*x*) (4*，*1)。在这两个图中，红色和蓝色曲线分别表示*Z*=0和*Z*=1组的似然比检测器的对数生成函数（推导见附录A.3）。我们

*～N～N*

*～N～N*

有Λ0(*u*)*z*=1=*8u*(*u \_* 1)，Λ1(*u*)*z*=1=*8u*(*u* + 1)。 另外，Λ0(*u*)*z*=0=*9u*(*u \_* 1)，Λ1(*u*)*z*=0=*9u*(*u* + 1)。 请注意：

2

2

C（P0*，*P1）*< C（*Q0，*Q1）。*(**左**)该图对应于Lemma [4](#_bookmark15)的情况。组*Z*=0的检测器是贝叶斯最优检测器，*τ*0*\**=0，EFN*,*T0（*τ*0*\**）=EFP*,*T0（*τ*0*\**）=*C*（P0*，*P1）*。Z*=1组的检测器是一个次优检测器，因为为了满足机会均等，我们必须选择*τ*1*\**，使EFN*,T1*（*τ*1*\**）=EFN*,*T0（*τ*0*\**）=*C*（P0*，*P1），并且这严格小于*C*（Q0*，*Q1）。(**右**)该图对应于Lemma [5](#_bookmark19)的情况。*Z*=1组的检测器是贝叶斯最优检测器，*τ*1*\**=0，EFN*,T1*（*τ*1*\**）=EFP*,T1*（*τ*1*\**）=*C*（Q0*，*Q1）*。*为了满足机会均等，我们必须选择*τ*0*\**，使EFN*,*T0（*τ*0*\**）=EFN*,*T1（*τ*1*\**）=*C*(Q0*，*Q1)，该阈值严格大于*C*(P0*，*P1)。然而，这个阈值*τ*0*\**使得EFP*,*T0(*τ*0*\**)小于*C*(P0*，*P1)，导致*Z*=0组的次优检测器。

现在，只有斜率τ1*\**的值才会饱和。

组，且*Ee,*T0（τ0*\**）<C（P0，P1）。 的完整证明。

isfy EFN*,T1* (τ1*\**)=EFN*,*T0 (0)是一个τ1*\**>0 ，使得 EFN*,T1*(τ1*\**)=C(P0,P1)<C(Q0,Q1), 且?

因此EFP*,T1*(τ1*\**)> C(Q0，Q1)。 由此可得，min*{*EFP*,*T0(0)，EFN*,*T0(0)，EFP*,*T1(τ1*\*)*，EFN*,*T1(τ1*\**)*}*=C(P0，P1)。

对于τ0*\**=0，要么EFP*,*T0(τ0*\**)<C(P0，P1)<EFN,T0(*τ0*\*)。

或EFN,T0(τ0*\**)<C(P0，P1)<EFP,T0(*τ0*\*)，意味着，min{EFP,*T0*(*τ0*\*)，EFN,T0(τ*0*\*)，*EFP*,T1(*τ1*\*)，*EFN*,T1(*τ1*\*)}<*C*(*P0*，P1)。

这种降低特权组准确度的情况通常被解释为对特权组造成*主动伤害*。为了避免造成主动伤害，同时满足公平性标准，我们也可以考虑一种变体，即

我们不改变特权组的最优检测器（或精度）（即特权组的EFN*,T1*（τ1）=EFP*,T1*（τ1）=C(Q0，Q1)），而只改变特权的检测器。

非特权群体来实现公平。我们提出以下优化方案。

最大最小EFP*,T*(τ0)，EFN*,T*(τ0)

*{ }*0 0

T0*,τ*0

令EFN*,*T0(τ0)=C(Q0，Q1)。 (3)

同样，如果我们将T0(x)限制为似然比检测器，那么优化([3](#_bookmark18))存在一个唯一的解τ0*\**。

**Lemma 5.***设*T0(x)＝log P1(*x*)*，我们有*C(P0，P1)＜C(Q0，Q1)。 *求解优化*([3](#_bookmark18))*的检测器*T0(x) τ0*\*是无特权组的次优检测器，*Ee*,*T0 (τ0*\**) < C(P0，P1)。

*>*

P0(*x*)

作为一个证明草图，我们参考图[2](#_bookmark16)（右）。如果我们选择

τ0*\* /*= 0，我们得到了一个次优的无权检测器

定理[4](#_bookmark15)和[5](#_bookmark19)在附录B.3中提供。

**备注4**（Z上的相等前值）**。***按照平衡精度度量的思路，优化时假设*Z=0*和*Z=1*上的priors也相等。我们参考附录E.2*

*完善的优化，以考虑到不平等的问题。*

Z=0*和*Z=1*的前值。*

**注5**(推广到其他公平性措施)**。***虽然我们在这里关注的是机会均等，但这个想法也可以扩展到其他公平度量。例如，如果每组的最佳似然检测器，即*T0(x) 0*和*

*>*

*>*

T1(x) *不满足统计学上的奇偶性（*[*Agarwal etal*](#_bookmark29)*,*

[*2018年*](#_bookmark29)*），而还有其他的对检测器，对两个*

*的组，那么对于两个组中的至少一个组，正在使用一个次优的检测器。*

### 错配假设检验的观点。没有准确性-公平性权衡的理想分布。

在这里，我们将证明存在着理想的分布，使公平性和准确性相一致。由于在观察空间中，由于无特权群体的可分离性不足而产生的权衡，我们特别感兴趣的是寻找与特权群体可分离性相匹配的无特权群体的理想分布，而在Lemma [5](#_bookmark19)中以次优精度实现公平性的同一个de- tector现在实现了与理想分布相关的最优精度。我们证明了这种理想分布的存在，还提供了一个明确的构造。

**定理2**（理想分布的存在）**。***对于第*[*2*](#_bookmark0)*节中的设置，让*C(P0，P1) < C(Q0，Q1)。*让我们选择*

*的贝叶斯最优检测器*T1(x)=log Q1(*x*)*>0*，对于

Q0(*x*)

为小。基于这个观点，我们制定了以下优化方案，用于指定两个理想分布

*组*Z=1*。 那么，对于组*Z＝0*，存在*P←0(x) ← ←

*和*

*和*P←1(x)*的形式*P←0(x)=*。*

1

*z* P0(*x*)(1*\_o*)P1(*x*)*o*

P0(*x*)(1*\_w*)P1(*x*)*w*

*z* P0(*x*)(1*\_w*)P1(*x*)*w*

←P (x) =

*P*←0*,P*←1

P0和P1为无权组。

P0(*x*)(1*\_o*)P1(*x*)*o*

← ← *> 0*

←

~FN*,T*

*对于*w，v *e R，这样：*

min π0D(P←0*||*P0) + π1D(P←1*||*P1)

* *在给定数据上的公平性）贝叶斯最优检测器的。*

*-·*

P0(*x*)

0

*的理想分布，即*T(x)=log P1(*x*)0*等价于*

P0(*x*)

*与Lemma 5的检测器*T0(x) = log P1(*x*) *> τ* \**相关联。*

以致于，E (0)=C(Q0，Q1)，(4)

0

← ← *> 0*

~0

tor与理想分布和E

(0)是

其中T(x)=log P1(*x*)0为贝叶斯最优检测-。

*P*←0(*x*)

*在给定数据集上满足机会均等的，即：*

← ←

← *z z*

P0(*x*)(1*\_w*)P1(*x*)*w*

P0(*x*)(1*\_o*)P1(*x*)*o*

←

EFN*,*T0(τ0)=EFN*,T1*(0)=C(Q0，Q1)*。*

*(理想数据的准确性和公平性)切尔诺夫的前。*

*贝叶斯最优的误差概率的指数*

*探测器在理想分布上，即*C(P0，P1)=。

C(Q0，Q1)*，因此大于*C(P0，P1)*。*

证明在附录C中提供。第一条标准

证明我们总是可以找到理想的分布。

使得*公平*检测器对给定的区域----------------------。

但（见Lemma [5](#_bookmark19)）实际上是贝叶斯最优的去。

相对于理想分布的向量。请注意

存在多对(v，w)，使P0(x)=的(v，w)

P0(*x*)(1*\_w*)P1(*x*)*w*和P1(x)=P0(*x*)(1*\_o*)P1(*x*)*o*坐-。

证实了该定理的第一条标准。

第二个标准更进一步，证明在这样的理想分布对中，总能找到至少一个对，使它们和以下的理想分布一样可以分离。

← ←

的特权组（即C(P0，P1)=C(Q0，Q1)）。非特权组的贝叶斯最优检测器与重?

谱到理想分布，即T←0(x)=对数*P*←1(*x*)*>0。*

因此，不仅是*公平*

*P*←0(*x*)

在给定的数据集上，但也满足于

FN*,T*

假阴性概率的切尔诺夫指数

当对给定分布进行评估时，该检测器的

P0（x）和P1（x）。定理[2](#_bookmark20)已经表明，前述的。

所述优化是可行的。

本小节的结果可以扩展到优化。

([2](#_bookmark13))，或完全采用其他公平措施，例如：

统计均等，或其他类型的限制，如

最小的个体失真。

**与构造空间的关系。**理想分布

无权无势的群体，结合给定的。

特权群体的分布，有两种解释。

趋势。(一) 它们可以被看作是下列情况下的合理分布：

如果映射是无偏向的，那么观察到的空间就会有偏向性。

分离性的观点（回顾定义[4](#_bookmark12)）。(ii) 鉴于我们的

对建构空间的认识有限，他们也可以

被看作是构造空间中的候选分布。

如果Z=1组的映射是同位素，那么它本身就是同位素。

映射。 这可以被证明是合理的，因为我们没有

除了通过观察到的数据之外，对构造空间（甚至是它的二元性）有很多了解。假设他们会有一个可分离性，这并不是不可理解的。

至少为C(Q , Q )，也就是以下所展示的可分离性

理想数据上的机会均等，因为它的切尔诺夫前。

←←*||||*

FNR的成分也与特权群体的成分相等。

即C(Q0，Q1)。请注意，为了满足第二个

准则，我们限制自己选择v=1，即

导致w的适当值。

**备注6**（唯一性）**。***定理*[*2*](#_bookmark20)*提供了一个关于ex-的证明。*

*理想分布的重要性以及明确的构造。*

*tion。一般来说，可能存在其他对分布。*

*的形式，而不是《Theo-》中提到的特定形式。*

*rem* [*2*](#_bookmark20)*，但可能满足定理的两个条件。*

*因此，只给定*P0(x)*和*P1(x)*，理想的distribu-。*

*除非有进一步的假设，否则，这些产品不一定是唯一的。*

*对它们的理想特性提出了要求。*

为了在实践中找到这样的理想分布，我们必须要有一个理想的分布。

因此，我们提出了一个额外的理想属性。

的这样一个理想数据集。我们要求理想数据集是

一个给定数据集的有用代表。这就促使

约束π0D(P0 P0)+π1D(P1 P1)要尽可能的小

尽可能的，即理想的distribu-的KL分歧。

从各自给定的现实世界的分布中提取出来的数据

观察空间中的特权组。因此，定理[2](#_bookmark20)

也证明了构造空间是非空的。

0

1

**Remark 7** (Explicit Use of an Ideal Dataset)**.** *几个前*

*isting方法（*[*Calmon等人*](#_bookmark44)*，*[*2018*](#_bookmark44)*；*[*Feldman等人*](#_bookmark36)*，*[*2015*](#_bookmark36)*。*

[*Kamiran& Calders，*](#_bookmark52)*2012*[*）*](#_bookmark52)*提出对给定的*

*数据集，以生成一个满足CER-的替代数据集。*

*公平性和实用性（表征），在。*

*与优化*（[4](#_bookmark21)）的*精神相同，并在其上训练模型。*

*训练好的检测器可能是次优的。*

*给定的数据集，但被认为是公平的。这里面的结果*

*分节有助于解释为什么这些方法会导致的。*

*在给定的数据集上进行准确度和公平度的权衡，同时也是*

*证明准确性和公平性都能提高*

*当测量精度与*

*到备用/理想数据集。优化*([4](#_bookmark21))*也是remi-。*

*niscent of the formulation of* [*Jiang & Nachum (*](#_bookmark48)*2019*[*),*](#_bookmark48) *who*

*假设一个给定的偏标签函数最接近理想的*

*无偏标签函数在KL分歧方面。在该*

*然而，工作中，KL分歧被应用于条件的*

*与条件特征相对的标签分布pY eX。*

*分布pXeY 。此外，*[*江与纳丘姆（*](#_bookmark48)*2019*[*）。*](#_bookmark48)

*不分析取舍的特点。*

**Remark 8** (Implicit Use of an Ideal Dataset)**.***属于这一类的现有方法包括在损失函数中使用公平性正则化进行训练，或者对输出进行后处理以满足公平性标准。这些方法不是明确地生成一个理想的数据集，而是旨在寻找一个在给定数据集上满足公平性准则的类筛选器，并在给定数据集上以最小的精度妥协（召回优化*（[2](#_bookmark13)）*和*（[3）](#_bookmark18)）*。在这里，我们证明存在与这些公平检测器相对应的理想分布，使得给定数据集上的次优检测器可以相对于理想数据集是最优的。*

### 主动收集数据。通过改进知识来减轻现实世界的权衡。

在第[3.1](#_bookmark9)节中讨论的给定数据集中各组之间不一致的可分离性的固有限制，事实上是可以克服的，但需要付出相关的代价：主动收集数据。在本节中，我们将演示什么时候收集更多的特征可以帮助改善无特权组的Chernoff信息，而不影响特权组的信息。收集更多的特征可以帮助我们用初始数据集中没有的附加分离性信息对无特权组的成员进行更仔细的分类。事实上，这也是主动公平性背后的理念。

([Noriega-Campero et al.](#_bookmark61) , [2019](#_bookmark61); [Bakker et al.](#_bookmark33) , [2019](#_bookmark33); [Chen](#_bookmark27)

C（W0，W1）=C（P0，P1）。这与直觉相一致，如果X*/*完全由X决定，那么它所提高的可分离性不会超过单独使用X所能达到的水平。因此，对于C(W0，W1)>C(P0，P1)，我们要求X/贡献一些有助于更好地分离假设Y=0和Y=1的信息，这实质上导致了X/在X和Z=0的前提下*不*独立于Y。

如果新的数据改善了组Z=0的可分离性，其准确性与公平性的权衡就会得到缓解（见图[3](#_bookmark26)中。

第[4](#_bookmark22)节）。)新的理想分布也可以用第[3.2](#_bookmark17)节的技术找到，这些新的理想分布作为无偏映射下的候选观测-空间分布或构造-空间分布是比较可信的。如果新的数据改善了这两组分布的可分离性，新的理想分布也会有更好的可分离性。

# 数值示例

我们用一个简单的数字例子来说明我们的理论概念和结果是如何在实践中计算出来的。

**例1.** *设*Z =0*的考试分数为*

P0(x)*~N*(1，1)*和*P1(x)*~N*(4，1)*，并且对于*Z=1

*为*Q0(x)*~N*(0，1)*，*Q1(x)*~N*(4，1)*。*

让我们把自己限制在T0(x)=log P0(*x*)*>τ0*和T1(x)=log Q0(*x*)>*τ1*的形式的似然比检测器上。

P1(*x*)

Q1(*x*)

[等](#_bookmark27)，[2018](#_bookmark27)）。)我们下面的分析也可以作为主动公平成功的技术解释。

让X*/*表示额外的特征，所以(X，X*/*)现在被用于Z=0组的分类。 注意，X*/*也可以很容易地成为其他形式的额外信息

两组的对数生成函数。Z=1的对数生成函数

可以分析计算为。Λ0(u)*z*=1=8u(u 1)，Λ1(u)*z*=1=8u(u+1)(推导见附录A.3)，Chernoff信息可计算为C(Q0，Q1)=2。

*-*

现在，对于无权组Z=0，对数生成-。

Λ(u)=9 u(u *-* 1)。

包括额外的解释，以配合数据或。

和9

*0z*=02

决定，类似于[Varshney等人](#_bookmark65)（[2018](#_bookmark65)）。让（X，X*/*）有

以下分布。(X，X*/*)*|Y*=0*，Z*=0 *~* W0(x，x*/*)

和(X，X*/*)*|Y*=1*，Z*=0*～*W1(x，x*/)*，其中Y为真值。

标签。注意，P0(x)=

*x/* W0(x，x*/*)，P1(x)=。

Λ1(u)*z*=0=2 u(u+1)(同样见附录A.3)

推导）。)切尔诺夫信息为C(P0，P1)=9/8。

**准确度-公平性在现实世界中的权衡。**我们限制

特权组的检测器是贝叶斯最优的

*x/* W1(x，x*/*)。我们的目标是推导出条件不-。

探测器T1(x)= log Q1(*x*)*>0(*相当于x>2*)。*对于

其中，随着加入更多的材料，分离性会得到改善。

特征，即C(W ，W )>C(P ，P )。

这个探测器。

Q0(*x*)

EFP*,T1*（0）=EFN*,T1*（0）=C（Q0，Q1）=2.

0 1 0 1

**定理3**（提高可分离性）**。***在给定*X*/和*Y*不相互独立的情况下，如果且仅在*X*/和*Y*不相互独立的情况下，Chernoff信息*C(W0，W1)*严格大于*C(P0，P1)。

X*和*Z=0*，即条件互信息为*

I(X*/*；Y *|*X，Z=0)>0。

证明见附录D。请注意，一般情况下

C(W , W )*> C(P* , P )，因为可分离性只能是

0

1

0

1

现在，对于Z=0， 贝叶斯 最优 T0(x)=log P1(*x*)*>0*(或者，x> 1*.5)*将是不公平的，因为EFN,T0(0*)=*C(P0，P1)<EFN,T1(0)。 利用Chernoff信息的几何解释（回忆图[2](#_bookmark16)）。

我们可以计算FPR和FNR的切诺夫指数，即EFP*,*T0（τ0）和EFN*,*T0（τ0）的负值为

P0(*x*)

Λ0(u)*z*=0的切线的y截距和

Λ1(u)*z*=0，检测器T0(x)=log P1(*x*)*>τ0*。这使得

P0(*x*)

改善或保持不变（见附录D）。我们确定不等式严格的情况。

设x*/*是x的确定性函数，即f(x)。那么，如果x*/*=f(x)，W0(x，x*/*)=P0(x)，否则为0。同理，如果x*/*=f(x)，则W1(x，x*/*)=P1(x)，否则为0，由此可得

准确度之间的权衡。

(Ee*,*T0 (τ0)=min EFP*,*T0 (τ0),EFN*,*T0 (τ0) )和公平性( EFN*,*T0 (τ0) EFN*,*T1 (τ0) )通过改变τ0，如图[3](#_bookmark26)中蓝色曲线所示。

*|-|*

*{}*

需要注意的是，满足公平性的检测器(等额运算)。

1.4

𝜏% )

1.2



1

准确率(↪Lu\_1D438↩!,#!

0.8

)PwC，w1(

C(,0～ ,1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  | **文学士**  **不过** |
|  |  |  |  | **流亡** |
| **FAIR** | **但SUBOP** | **TIMAL on** | **现有** |  |
| **资料a** | **后的数据** | **集合** |  | **BAYES O** |
| **FAIR但**  **现存** | **SUBOPTIM**  **g数据** | **AL** |  | **不公平**  **之后** |

**是的 最佳方案 不公平**

**蜇伏数据**

0.4

0.2



0

0.4

0.2

z)

P0

(z)

P(z)

Q0(

**FAIR**

**BAYES**

Q1

P1

(z)

P(z)=P(z)

(z)

**BAYES**

0.6

0.4

**PTIMAL但**

0

0

02

1 1

4 6 8

0.2

0

**在现有数据的基础上进行收集**

*图4.*(顶部)对于例[1](#_bookmark23)中的分布，我们将其表示为

贝叶斯最优检测器对数Q1(*x*)*>0*(相当于x>*2)*，*对于*

0 0.2 0.4 0.6 0.8 1



𝜏(

特权组*Z*=1

Q0(*x*)

*Z*=0，则最佳的检测-

## |𝐸

- − 𝐸

|减少

P1(*x*)

*> 0*

.(底部)用于

& ' ,#!



𝜏%

&',#"

公平性

tor log *P*（*x*）0在给定的数据集上不满足机会均等，但一个次优的检测器却满足（注意到两组的假阴性率的等面积cor-反应）。然而，有

*图3.*用一个数字例子计算公平性和知识性之间的权衡。对于非特权组，我们让P0(*x*) (1*,* 1) 和P1(*x*) (4*,* 1)*.*我们将特权组的去向器限制为其贝叶斯最优检测器，*C*(Q0*，*Q1) = 2。蓝色曲线表示无特权组在现有数据集中准确性和公平性之间的权衡。现在假设我们能够为无特权组收集一个额外的fea- ture *X/，*使(*X，X/*) *Y* =0*,Z*=0。

*～N～N*

～～

((1*，*1)*，***1**)和(*X，X/*)*Y*=1*，Z*=0((4*，*2)*，***1**)，其中**1**是指

*N| ～N*

的2 2身份矩阵。绿色曲线显示了主动收集数据如何缓解公平性和准确性之间的权衡。

*x*

在给定的分布上，也可以在给定的分布上进行com-。

将其分析为对数P1(*x*) *> τ* \*，其中τ \*=*-3/2* (等价-)。

存在由P0=Q0和P1=P1=Q1给出的理想分布，这样该检测器在理想分布上是最优的，并且在现有分布和理想分布上也实现了公平性。

(X，X*/*)*Y*=0*，Z*=0*(*(1，1)，**I**)和(X，X*/*)*Y*=1*，Z*=0。

*|~*

((4，2)，**I**)，其中**I**为2 2身份矩阵。对数生成函数可推导为：。Λ0(u)=5u(u 1)，Λ1(u)=5u(u+1)。注意，Chernoff informa- tion(可分离性)C(W0，W1)=5/4，大于C(P0，P1)=9/8。因此，收集新特征提高了无权组的可分离性。

*-*

*N×*

现在，我们研究一下主动收集数据对?

通向

P0(*x*)0 0

在现实世界中的准确性和公平性的权衡。我们再次

x 2).这就导致FNR的指数相等，即EFN*,*T0 ( 3/2)=2=EFN*,*T1 (0)，但对于这个detec-。

*-*

*>*

tor EFP*,T* (τ *\**)=1/2，导致Chernoff前值减少

请参考图[3](#_bookmark26)（绿色曲线）。 考虑到可能性

Z=0的比值检测器，基于总集的fea-。

0

0 0tures，即T（x，x*/*）=log W0（*x，x/*）。

误差概率的比例（代表accu.com）。

0

W1(*x,x/*)

*> τ* .为了满足

racy），即*Ee,*T0（τ0*\**）=min*{*EFP*,*T0（τ0*\**），EFN*,*T0（τ0*\**）*}*=。

min*{*1/2，2*}*=1/2，小于C(P0，P1)=9/8。

我们的公平性约束，我们需要选择一个τ0*\**，从而使

EFN*,*T0(τ0*\**)=EFN*,*T1(0)=C(Q*′*0，Q1)=2。 经解-

ing，我们得到τ0*\**=5 *-*。40  *s -′*1.32. 为此

**理想分布。**我们参考图[4](#_bookmark25)。结果发现，定理[2](#_bookmark20)规定的一对理想分布是P0=Q0和P1=P1=Q1。贝叶斯最优检测器与

←←

*-s*

对于Z=0的理想分布，由以下公式给出。

log *P*←1(*x*)*>0*(相当于x>2*)。*请注意，这等于

←*P (x*) 0

lent to the detector log P1(*x*) *> τ* \*，其中τ \*= *-* 3*/2*，这满足了给定数据集上的机会均等。这个检测器

P0(*x*)

0

0

tor现在是贝叶斯最优的，相对于理想的distribu- tions P0和P1，并且具有整体的Chernoff指数。

←←

← ←

相对于理想分布而言，误差概率等于C(P0，P1)=2。因此，我们证明

以确保公平性（在现有数据集和理想数据集上机会均等的意义上）和准确性（相对于理想分布）是一致的。需要注意的是，人们也可以使用优化([4](#_bookmark21))或优化的任何变体，例如，使用统计奇偶性，来确定理想分布的备用对。

**主动收集数据。**现在假设我们能够收集到Z=0的额外特征X*/，*这样的话

τ0*\**的值，我们得到EFP*,*T0（τ0*\**）=7 40）0.68。该*公平*检测器的误差概率的切尔诺夫指数由min EFN*,*T0（τ0*\**），EFP*,*T0（τ0*\**）=给出。

min 2，0.68=0.68，大于0.5（谢尔-）。

*{}*

*{}*

公正检测器误差概率的off指数。

在收集附加特征X*/*）之前。)

# 结论

我们的结果提供了新颖的分析见解，解释和描述了真实数据集上的准确性-公平性权衡。我们基于Chernoff信息的分析可以帮助量化数据集的可分离性，甚至在任何分类算法被应用之前。我们相信，我们证明公平性和准确性与理想数据集是一致的，这将激励我们在算法公平性研究中使用相对于理想数据集的准确性作为性能指标（[Sharma等人](#_bookmark64)，[2020](#_bookmark64)；[Wick等人](#_bookmark49)，[2019](#_bookmark49)）。最后，我们的结果也为如何以及何时主动收集数据可以缓解现实世界中的权衡提供了信息。

# 鸣谢

作者要感谢Pulkit Grover、Shubham Sharma和匿名审稿人提出的宝贵建议。

# 参考文献

Agarwal, A., Beygelzimer, A., Dud´ık, M., Langford, J., and Wallach, H. A reductions approach to fair classiﬁca- tion.In *Proceedings of the International Conference on Machine Learning*，pp.60-69，2018.

Bakker, M. A., Noriega-Campero, A., Tu, D. P., Sattigeri, P., Varshney, K. R., and Pentland, A. S. On fairness in budget-constrained decision making.In *Proceedings of the KDD Workshop on Explainable Artiﬁcial Intelligence*, 2019.

Berend, D. and Kontorovich, A. A ﬁnite sample analysis of the naive Bayes classiﬁer.*Journal of Machine Learning Research*，16：1519-1545，2015.

Berisha, V., Wisler, A., Hero, A. O., and Spanias, A. Em- pirically estimable classiﬁcation bounds based on a non parametric divergence measure.*IEEE Transactions on Signal Processing*，64(3)：580-591，2015.

Bhattacharyya, A. On a measure of divergence between two multinomial populations.*Sankhya¯ : The Indian Journal of Statistics*, 7(4):401-406, 1946.

Boucheron, S., Lugosi, G. and Massart, P. *Concentration inequalities:A nonasymptotic theory of independence*.牛津大学出版社，2013年。牛津大学出版社，2013年。

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann,

J.M. The balanced accuracy and its posterior distribution.In *Proceedings of the International Conference on Pattern Recognition*，pp.3121-3124，2010.

Calmon, F. P., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized pre-processing for discrimination prevention.In *Advances in Neural Infor- mation Processing Systems*，pp.3992-4001，2017.

Calmon, F. P., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Data pre-processing for discrim- ination prevention:Information-theoretic optimization and analysis.*IEEE Journal of Selected Topics in Signal Processing*，12（5）：1106-1119，2018.

Celis, L. E., Huang, L., Keswani, V. and Vishnoi, N. K. Classiﬁcation with fairness constraints:A meta -algorithm with provable guarantees. - In Proceedings of ACM Conference on Fairness.In *Proceedings of the ACM Conference on Fairness, Accountability, and Trans- parency*, pp.319-328, 2019.

Chen, I. Y., Johansson, F. D., and Sontag, D. Why is my clas- siﬁer discriminatory?In *Advances in Neural Information Processing Systems*，pp.3539-3550，2018.

Coˆte´, F. D., Psaromiligkos, I. N., and Gross, W. J. A Chernoff-type lower bound for the Gaussian Q-function. *arXiv preprint arXiv:1202.6483*, 2012.

Cover，T. M. and Thomas，J. A. *Elements of Information Theory*.John Wiley & Sons，2012.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints.In *Advances in Neural Information Process- ing Systems*，pp.2791-2801，2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel,

1. 通过认识实现公平。In *Proceedings of the Innovations in Theoretical Computer Science Conference*，pp.214-226，2012.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classiﬁers for group-fair and efﬁcient machine learning.In *Proceedings of the Conference on Fairness, Accountability and Transparency*, pp.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and remove disparate impact.In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*，pp.259-268，2015.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian,

1. On the (im) possibility of fairness.*arXiv preprint arXiv:1609.07236*，2016.

Gallager, R. Detection, decisions, and hypothesis testing. [http://web.mit.edu/gallager/www/ papers/chap3.pdf](http://web.mit.edu/gallager/www/papers/chap3.pdf), 2012.

Garg, S., Kim, M. P., and Reingold, O. Tracking and improv- ing information in the service of fairness.In *Proceedings of the ACM Conference on Economics and Computation*，pp.809-824，2019.

Ghassami, A., Khodadadian, S., and Kiyavash, N. Fair- ness in supervised learning:A. Khodadadian, S. and Kiyavash, N. Fair- ness in supervised learning: An informationoretic approach.In *Proceedings of the IEEE International Sym- posium on Information Theory*，pp.176-180，2018.

Gretton, A., Borgwardt, K., Rasch, M., Scho¨ lkopf, B., and Smola, A. J. A kernel method for the two-sample- problem.In *Advances in Neural Information Processing Systems*，pp.513-520，2007.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning.In *Advances in Neural Informa- tion Processing Systems*，pp.3315-3323，2016.

Jiang, H. and Nachum, O. *arXiv preprint arXiv:1901.04966*, 2019.

Kailath，T. The divergence and Bhattacharyya distance measures in signal selection.*IEEE Transactions on Com- munication Technology*，15(1)：52-60，1967.

Kamiran, F. and Calders, T. Data preprocessing techniques for classiﬁcation without discrimination.*知识与信息系统*，33(1)：1-33，2012。

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Scho¨lkopf, B. Avoiding discrimination through causal reasoning.In *Advances in Neural Infor- mation Processing Systems*，pp.656-666，2017.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Coun- terfactual fairness.In *Advances in Neural Information Processing Systems*，pp.4066-4076，2017.

Lee, Y. and Sung, Y. Generalized Chernoff information for mismatched Bayesian detection and its application to energy detection.*IEEE Signal Processing Letters*，19（11）：753-756，2012。

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classiﬁcation.In *Proceedings of the Conference on Fairness, Accountability and Transparency*, pp.

Motwani, R. and Raghavan, P. *Randomized Algorithms*.剑桥大学出版社，1995年。

Noriega-Campero, A., Bakker, M. A., Garcia-Bulle, B., and Pentland, A. S. Active fairness in algorithmic decision making.In *Proceedings of the AAAI/ACM Conference on Artiﬁcial Intelligence, Ethics, and Society*, pp.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. Sparse additive models.*皇家统计学会杂志。B系列（统计方法学）*，71（5）：1009-1030，2009年。

Scott, C., Blanchard, G. and Handy, G. Classiﬁcation with asymmetric label noise:Consistency and maximal denoising.In *Proceedings of the Conference On Learning Theory*, pp.

Sharma, S., Zhang, Y., Aliaga, J. M. R., Bouneffouf, D., Muthusamy, V., and Varshney, K. R. Data augmentation for discrimination prevention and bias disambiguation.In *Proceedings of the AAAI/ACM Conference on Artiﬁcial Intelligence, Ethics, and Society*, pp.

Varshney, K. R., Khanduri, P., Sharma, P., Zhang, S. and Varshney, P. K. Why interpretability in machine learning?

An answer using distributed detection and data fusion theory.In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, pp.

Wick，M.，Panda，S.，和Tristan，J.-B.。Unlocking fairness: a trad-off revisited.In *Advances in Neural Information Processing Systems*，pp.8780-8789，2019.

Yeom, S. and Tschantz, M. C. Discriminative but not dis- criminatory:A comparison of fairness deﬁnitions under different worldviews. *ArXiv preprint arXiv:1808.08619*, 2018.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gum- madi, K. P. Fairness beyond disparate treatment & dis- parate impact:Learning classiﬁcation without disparate misreatment.In *Proceedings of the International Con- ference on World Wide Web*，pp.1171-1180，2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork,

C.Learning fair representations.In *Proceedings of the International Conference on Machine Learning*，pp. 325-333，2013.

Zhao, H. and Gordon, G. J. Inherent tradeoffs in learning fair representation.*arXiv preprint arXiv:1906.08386*, 2019.