公平度量：一个用于群体公平性评估的 R 包

A PREPRINT

本杰明 · 史密斯
Department of Statistical Sciences
University of Toronto
Toronto, ON M5G 1X6
benyamindsmith@mail.utoronto.ca

高建辉
Department of Statistical Sciences
University of Toronto
Toronto, ON M5G 1X6
jianhui.gao@mail.utoronto.ca

詹 卡 格朗斯贝尔
Department of Statistical Sciences
University of Toronto
Toronto, ON M5G 1X6
j.gronsbell@utoronto.ca

2025 年 8 月 29 日

1 总结

公平性是机器学习（ML）中一个不断增长的领域，专注于确保模型不会为特定群体产生系统性的偏见结果，特别是那些由受保护属性定义的群体，如种族、性别或年龄。评估公平性是 ML 模型开发的关键方面，因为有偏见的模型可能会延续结构性不平等。{fairmetrics} R 包提供了一个用户友好的框架，用于严格评估众多基于群体的公平性标准，包括基于独立（例如统计平衡）、分离（例如等化机会）和充足性（例如预测平衡）的度量。基于群体的公平性标准评估模型是否在一组预定义的群体中具有相同的准确性或良好的校准，以便可以实施适当的偏见缓解策略。{fairmetrics} 通过一个方便的包装函数为多个指标提供点估计和区间估计，并包括从加强护理医学信息市场（MIMIC-II）数据库 (Goldberger et al., 2000; Raffa, 2016) 衍生出的示例数据集。

2 需求说明

机器学习模型越来越多地被整合到高风险领域中，以支持对个人和社会产生重大影响的决策过程，包括刑事司法、医疗保健、金融、就业和教育等领域 (Mehrabi et al., 2021)。越来越多的证据表明，这些模型在由受保护属性定义的不同群体之间往往表现出偏见。例如，在刑事司法领域，矫正罪犯管理剖析以供替代制裁使用的软件（COMPAS）——这是一种美国法院用来评估被告重新犯罪风险的工具——被发现错误地将黑人被告归类为高风险的概率几乎是白人被告的两倍 (Mattu, 2016)。这种偏见影响了黑人被告，可能导致更严厉的保释决定、更长的刑期和较少的假释机会与具有相似风险状况的白人被告相比。同样，在医疗保健领域，一种在美国部署用于识别复杂健康需求患者以供高风险护理管理项目使用的商业风险预测算法显示，对于黑人的校准显著低于对白人患者的校准 (Obermeyer et al., 2019)。这导致了患有相同健康状况的黑人患者与白人相比被推荐较少的基本护理服务。这些例子说明，在将机器学习模型部署到实际应用之前，从业者和研究人员必须确保这些模型支持公平决策的过程非常紧迫。

虽然现有的软件可以计算群体公平性标准，但它们仅提供点估计和/或可视化而不量化围绕这些标准的不确定性。这一限制使得用户无法确定观察到的不同群体之间的差异是否具有统计显著性，或者仅仅是由于有限样本量导致的随机变化的结果，这可能导致关于公平性违规的错误结论。{fairmetrics} R 包通过为基于差值和基于比率的群体公平性指标提供基于自助法的置信区间（CIs），解决了这一空白，使用户能够根据统计依据对其模型的公平性做出决策，在实践中这一点往往不一致。

3 范围

{公平指标}包旨在评估具有二元受保护属性的二元分类设置中的群体公平性。此限制反映了公平文献中的标准做法，并受到几个考虑因素的驱动。首先，二元分类在许多高风险应用中仍然很普遍，例如贷款审批、招聘决策和疾病筛查，其中结果通常被表述为接受/拒绝或正/负 (Mehrabi et al., 2021)。其次，群体公平性是二元分类任务中最广泛使用的框架 (Mehrabi et al., 2021)。第三，当受保护属性有多于两个类别时，如何评估群体公平性没有明确的共识 (Lum et al., 2022)。这种关注使{fairmetrics}能够为跨不同应用领域的二元分类任务中常用的群体公平度量提供统计基础的不确定性量化。

4 公平标准

群组公平标准主要分为三大类：独立性、分离性和充分性 (Barocas et al., 2023; Berk et al., 2018; Castelnovo et al., 2022; Gao et al., 2024)。独立性要求模型的分类在统计上与受保护属性无关，这意味着获得正面预测的可能性在整个受保护群体中相同。分离性要求分类和受保护属性在真实结果条件下相互独立，使得正面（或负面）结果类别内的受保护群体获得正面预测的概率相等。充分性要求结果和受保护属性在预测条件下相互独立，意味着一旦模型的预测已知，则受保护属性对真正结果不再提供额外信息。以下我们总结了 {fairmetrics} 包中可用的公平度量。

4.1 独立性

- **统计均等性：**比较各组之间正面预测的整体比率。

4.2 分离

- **平等机会：**比较组间假阴性率的差异，量化遗漏阳性结果的可能性的不同。
- **预测等式：**比较假阳性率（FPR）在各组之间的差异，量化将阴性结果错误标记为阳性的可能性的不同。
- **正类别的平衡：**比较真实结果为正的个体在各组中的预测概率平均值。
- **负类别的平衡：**比较真实结果为负的个体在不同组中的预测概率平均值。

4.3 充分性

- **阳性预测值公平性：**比较各组的阳性预测值，评估阳性预测的精度差异。
- **负预测率公平性：**比较各组的阴性预测值，评估阴性预测的精确度差异。

4.4 其他准则

- **布赖尔分数平等：**比较各组的布雷尔分数（即预测概率的均方误差），评估校准差异。
- **准确性一致性和**比较了预测模型在不同组中的总体准确性。

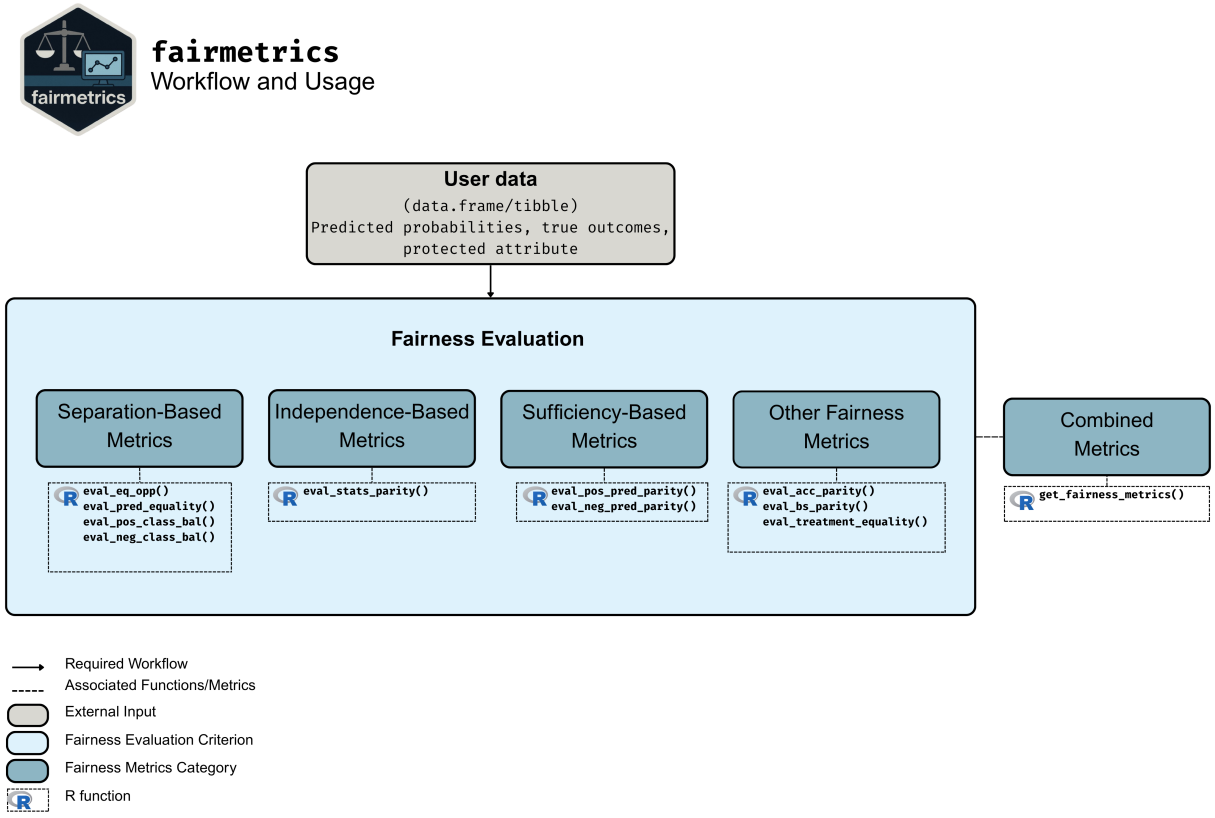

图 1: 使用{fairmetrics}评估模型公平性的流程跨多个标准。

- **处理平等性：**比较各组误漏报与误报的比率，评估正向结果未被检测出和负向结果错误警报之间的权衡是否平衡。

5 评估公平性标准

{fairmetrics}包的输入是一个数据框或 tibble，其中包含模型的预测概率、真实的结局以及感兴趣的保护属性。Figure 1展示了使用{fairmetrics}的工作流程。用户可以使用综合指标函数评估特定标准下的模型或多组公平性标准。

使用 {fairmetrics} 包的一个简单示例如下所示。该示例使用了模仿预处理数据集，这是 MIMIC-II 临床数据库 [1] 中的 Indwelling Arterial Catheter (IAC) 临床数据集的预处理版本 (Raffa, 2016; Raffa et al., 2016)。数据集由 1,776 名血流动力学稳定的呼吸衰竭患者组成，包括人口统计信息（患者年龄和性别）、生命体征、实验室结果、是否使用了 IAC，以及一个二元结果变量，表明患者在入院后 28 天内是否死亡。

尽管所选的公平性标准取决于具体情境，但我们通过获取公平性指标 () 函数展示了所有可用的标准以供说明。在这个示例中，我们评估了模型在受保护属性性别方面的公平性。对于条件统计均等性，我们对 60 岁以上的患者进行条件设定。该模型是在数据的一个子集上训练的，并且预测结果和评估是基于测试集进行的。

[1]该数据的原始版本由 PhysioNet(Goldberger et al., 2000) 提供，并可以通过加载模仿数据集在 {fairmetrics} 软件包中访问。

获取公平性指标 () 函数输出差异和比率为基础的指标及其相应的置信区间。当差异基础指标的置信区间不包括零或比率基础指标的区间不包括一时，表明在给定显著性水平上组间存在统计学上的显著差异。

```
library(fairmetrics)
library(dplyr)
library(magrittr)
library(randomForest)

# Load the example dataset
data("mimic_preprocessed")

# Split the data into training and test sets
train_data <- mimic_preprocessed %>%
  dplyr::filter(dplyr::row_number() <= 700)

test_data <- mimic_preprocessed %>%
  dplyr::mutate(gender = ifelse(gender_num == 1, "Male", "Female")) %>%
  dplyr::filter(dplyr::row_number() > 700)

# Train a random forest model
rf_model <- randomForest::randomForest(
  factor(day_28_flg) ~ .,
  data = train_data,
  ntree = 1000
  )

# Make predictions on the test set
test_data$pred <- predict(rf_model, newdata = test_data, type = "prob")

# Get fairness metrics
# Setting alpha=0.05 for 95% confidence intervals
get_fairness_metrics(
 data = test_data,
 outcome = "day_28_flg",
 group = "gender",
 group2 = "age",
 probs = "pred",
 cutoff = 0.41,
 alpha = 0.05
)
```

	Fairness Assesment	Metric
1	Statistical Parity	Positive Prediction Rate
2	Equal Opportunity	False Negative Rate
3	Predictive Equality	False Positive Rate

```
4  Balance for Positive Class           Avg. Predicted Positive Prob.
5  Balance for Negative Class           Avg. Predicted Negative Prob.
6  Positive Predictive Parity               Positive Predictive Value
7  Negative Predictive Parity               Negative Predictive Value
8          Brier Score Parity                                Brier Score
9     Overall Accuracy Parity                                   Accuracy
10         Treatment Equality (False Negative)/(False Positive) Ratio

   GroupFemale GroupMale Difference    95% Diff CI Ratio 95% Ratio CI
1         0.17      0.08       0.09   [0.05, 0.13]  2.12 [1.49, 3.04]
2         0.38      0.62      -0.24 [-0.39, -0.09]  0.61 [0.44, 0.86]
3         0.08      0.03       0.05   [0.02, 0.08]  2.67  [1.4, 5.08]
4         0.46      0.37       0.09   [0.04, 0.14]  1.24 [1.09, 1.42]
5         0.15      0.10       0.05   [0.03, 0.07]  1.50 [1.29, 1.74]
6         0.62      0.66      -0.04  [-0.21, 0.13]  0.94 [0.72, 1.22]
7         0.92      0.90       0.02  [-0.02, 0.06]  1.02 [0.98, 1.07]
8         0.09      0.08       0.01  [-0.01, 0.03]  1.12 [0.89, 1.43]
9         0.87      0.88      -0.01  [-0.05, 0.03]  0.99 [0.94, 1.04]
10        1.03      3.24      -2.21 [-4.38, -0.04]  0.32 [0.15, 0.68]
```

如用户希望计算个别标准，可以使用任何评估_*函数。例如，要计算平等机会，用户可以调用评估平等机会() 函数。

```
eval_eq_opp(
  data = test_data,
  outcome = "day_28_flg",
  group = "gender",
  probs = "pred",
  confint = TRUE,
  cutoff = 0.41,
  alpha = 0.05
)


There is evidence that model does not satisfy equal opportunity.
              Metric GroupFemale GroupMale Difference    95% Diff CI Ratio 95% Ratio CI
1 False Negative Rate       -0.42     -0.23      -0.19 [-0.33, -0.05]  1.83    [1.11, 3]
```

6 相关工作

其他类似于 {fairmetrics} 的 R 包包括 {fairness} (Kozodoi and V. Varga, 2021), {fairmodels} (Winiewski and Biecek, 2022) 和 {mlr3fairness} (Pfisterer et al., 2024)。{fairmetrics} 和这些其他包之间的差异有两点。主要的区别在于 {fairmetrics} 通过引导法计算群体公平标准及其对应公平度量置信区间之间的比率和差值，从而能够对公平标准进行更有意义的推断。此外，与{fairmodels}, {fairness} 和 {mlr3fairness} 包不同的是，{fairmetrics} 包没有外部依赖项，并且内存占用较低，结果是一个环境无关的工具，可以使用适度的硬件和旧系统运行。Table 1 显示了加载每个库时使用的内存和所需依赖项的比较。

包	内存（MB）	依赖关系
fairmodels	17.02	29
fairness	117.61	141
mlr3fairness	58.11	45
fairmetrics	0.05	0

表 1: 内存使用量（以 MB 为单位）以及公平指标与类似软件包的依赖关系。

对于 Python 用户，{fairlearn} 库 (Weerts et al., 2023) 提供了额外的公平性指标和算法。{fairmetrics} 包旨在与 R 工作流程无缝集成，使其成为基于 R 的机器学习应用的更便捷选择。

7 许可和可用性

{fairmetrics}软件包采用 MIT 许可证。它可以在 CRAN 上获取，并可以通过使用 `install.packages("公平度量")` 进行安装。一个更详细的教程可以访问：`https://jianhuig.github.io/fairmetrics/articles/fairmetrics.html`。所有代码都是开源的，托管在 GitHub 上。所有的错误和咨询都可以报告到 `https://github.com/jianhuig/fairmetrics/issues/`。

参考文献

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press, Cambridge, Massachusetts, 2023.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods Research*, 50(1):3–44, July 2018. doi:10.1177/0049124118782533. URL https://journals.sagepub.com/doi/10.1177/0049124118782533.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1), March 2022. doi:10.1038/s41598-022-07939-1. URL https://www.nature.com/articles/s41598-022-07939-1.

Jianhui Gao, Benson Chou, Zachary R. McCaw, Hilary Thurston, Paul Varghese, Chuan Hong, and Jessica Gronsbell. What is fair? defining fairness in machine learning for health, June 2024. URL https://arxiv.org/abs/2406.09307.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23):e215–e220, 2000. doi:10.1161/01.CIR.101.23.e215. URL https://doi.org/10.1161/01.CIR.101.23.e215.

Nikita Kozodoi and Tibor V. Varga. *fairness: Algorithmic Fairness Metrics*, 2021. URL https://CRAN.R-project.org/package=fairness. R package version 1.2.1.

Kristian Lum, Yunfeng Zhang, and Amanda Bower. De-biasing "bias" measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 379 – 389, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi:10.1145/3531146.3533105. URL https://doi.org/10.1145/3531146.3533105.

Lauren Kirchner Surya Julia Angwin Mattu, Jeff Larson. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July 2021. ISSN 0360-0300. doi:10.1145/3457607. URL https://doi.org/10.1145/3457607.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Florian Pfisterer, Wei Siyi, and Michel Lang. *mlr3fairness: Fairness Auditing and Debiasing for 'mlr3'*, 2024. URL https://mlr3fairness.mlr-org.com. R package version 0.3.2, https://github.com/mlr-org/mlr3fairness.

Jesse Raffa. Clinical data from the mimic-ii database for a case study on indwelling arterial catheters (version 1.0). https://doi.org/10.13026/C2NC7F, 2016. PhysioNet.

Jesse D. Raffa, Mohammad Ghassemi, Tristan Naumann, Mengling Feng, and Daniel J. Hsu. Data analysis. In *Secondary Analysis of Electronic Health Records*, pages 109–122. Springer, Cham, 2016. doi:10.1007/978-3-319-43742-2_9.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems, March 2023. URL https://arxiv.org/abs/2303.16626.

Jakub Winiewski and Przemysaw Biecek. fairmodels: a flexible tool for bias detection, visualization, and mitigation in binary classification models. *The R Journal*, 14(1):227–243, 2022. doi:10.32614/RJ-2022-019. URL https://rj.urbanek.nz/articles/RJ-2022-019/.