De-biasing "Bias" Measurement

Kristian Lum, Yunfeng Zhang, Amanda Bower

META (Machine Learning Ethics, Transparency, and Accountability), Twitter Inc.

Github: [https://github.com/twitter-research/double-corrected-variance-estimator](https://github.com/twitter-research/double-corrected-variance-estimator)
Kristian Lum
Pronouns (She/Her)
@KLdivergence

Yunfeng Zhang
Pronouns (He/Him)
@zywind

Amanda Bower
Pronouns (She/Her)
@amandarg__
Which model has the lowest group-wise performance disparities?
Existing Bias Metrics are Inadequate

- Visualizations are inadequate for a human to make judgements about the degree of disparities when the number of groups involved is large.
- Canonical bias metrics such as statistical parity difference and predictive parity difference only measure differences between two groups.

Statistical Parity Difference:
\[ P(\hat{Y}=1|G=a) - P(\hat{Y}=1|G=b) \]

False Positive Rate Difference:
\[ P(\hat{Y}=1|Y=0, G=a) - P(\hat{Y}=1|Y=0, G=b) \]

Which model has the lowest group-wise performance disparities?
# Meta-Metrics

| Meta-Metric Name                      | Formula                                                                 | Type       | Used by   |
|--------------------------------------|-------------------------------------------------------------------------|------------|-----------|
| max-min difference                    | $M_{\text{mm-diff}}(Y) = \max_k Y_k - \min_k Y_k$                     | Extremum   | [10]      |
| max-min ratio                         | $M_{\text{mm-ratio}}(Y) = \frac{\max_k Y_k}{\min_k Y_k}$              | Extremum   | [10, 22]  |
| max absolute difference               | $M_{\text{max-abs-diff}}(Y) = \max_k |Y_k - \bar{Y}|$             | Extremum   | [2, 32]   |
| mean absolute deviation               | $M_{\text{mad}}(Y) = \frac{1}{K} \sum_k |Y_k - \bar{Y}|$           | Variability| [31]      |
| variance                              | $M_{\text{var}}(Y) = \frac{1}{K-1} \sum_k (Y_k - \bar{Y})^2$          | Variability|           |
| generalized entropy index ($\alpha \neq 0, 1$) | $\frac{1}{K\alpha(\alpha-1)} \sum_{k=1}^K \left[ \left( \frac{Y_k}{\bar{Y}} \right)^\alpha - 1 \right]$ | Variability| [9, 44, 46]|
Meta-Metrics Have Upward Statistical Bias

Simulation setup
- 5000 individuals divided equally into K groups
- $K \in [5, 150]$
- Generate performance metric from the following model
  \[ Z_k \sim \text{Binomial}(n_k, \mu_k) \quad Y_k = \frac{Z_k}{n_k} \]
- $\mu_k$ equally spaced on $[l, .9]$, where $l$ is the lower bound.
- Results are averaged over 1000 simulations

Simulations show upward statistical bias in every meta-metric.

$K$: Number of groups
$Y$: Vector of observed model performance $[Y_1, ..., Y_K]$
$\mu$: Vector of true model performance $[\mu_1, ..., \mu_K]$
Meta-Metrics Have Upward Statistical Bias

We can also show mathematically that $M(Y)$ is an upward statistically biased estimate of $M(\mu)$ for several meta-metrics: mean absolute deviation, max absolute deviation, and variance.

\[
\mathbb{E}[M_{\text{mad}}(Y)] = \mathbb{E} \left[ \frac{1}{K} \sum_{k=1}^{K} |Y_k - \bar{Y}| \right] \\
= \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} |Y_k - \bar{Y}| \\
> \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [Y_k - \bar{Y}] \\
= \frac{1}{K} \sum_{k=1}^{K} |\mu_k - \bar{\mu}| \\
= M_{\text{mad}}(\mu),
\]

$K$: Number of groups  
$Y_k$: Observed model performance for group k  
$\mu_k$: True model performance $E[Y_k]$
Intuition Behind the Statistical Bias

True model performance

Sample model performance

Statistical bias is induced by sampling
Statistically Biased Meta-Metrics Are Problematic

1. The overestimation of the true group-wise performance disparity may cause unnecessary model adjustments. Because of fairness-accuracy trade-off, such adjustments could lead to less performant models for all individuals.

2. Statistically Biased meta-metrics cannot be compared fairly across different grouping methods. Different grouping could lead to different amount of bias.
Correcting for Statistical Bias in The Variance Meta-Metric

\[ \hat{M}_{\text{var}}(\mu) = M_{\text{var}}(Y) - \frac{1}{K} \sum_k \hat{\sigma}_k^2 \]

Truncated:
\[ \hat{M}_{\text{var}}(\mu) = \max(0, M_{\text{var}}(Y) - \frac{1}{K} \sum_k \hat{\sigma}_k^2) \]

- Introduced by Cochran (1954) and popularized by Hedge and Olkin (1985) for meta-analysis of between study variability.
- Larger \( \hat{\sigma}_k^2 \) (e.g. caused by small sample size) leads to larger statistical bias.
- Untruncated version is statistically unbiased when statistically unbiased estimates of the standard errors are available. We follow the convention to use the truncated version to preclude negative variance estimates.

\( \sigma_k \): Standard error of the observed model performance.

\( M_{\text{var}} \): The between-group variance meta-metric.
Simulation shows that the correction works

\[ n=5000, \; K=100 \]

Equal group size, \( n_k=50 \)
Unequal group size, \( n_k=10 \) to 90

Equal performance, \( \mu_k=0.8, \; M_{\text{var}}(\mu)=0 \)
Unequal performance, \( \mu_k=0.1 \) to 0.9, \( M_{\text{var}}(\mu)=0.055 \)

Black vertical lines show the true variance.

Simulation:
- Generate samples for each group using the parameters on the left.
- Calculate sampling variance of each group as:
  \[ \hat{\sigma}_k^2 = \frac{Y(1-Y)}{n_k} \]
- Apply the \( M_{\text{var}}(\mu) \) equation.
- Repeat 1000 times.

The distributions of corrected variance estimates are centered at the truth.
Uncertainty Quantification for MetaMetrics

Uncertainty intervals inform decision makers about the confidence of the statistical estimate.

Point estimate

Uncertainty interval

Between-group performance variance

0
Naïve application of the corrected variance estimator to each bootstrap sample results in upwardly shifted uncertainty intervals that do not even cover the true value.

\[
X_k \sim \text{Binomial}(n_k, \mu_k) \quad Y_k = \frac{X_k}{n_k}
\]

\[
X_k^* \sim \text{Binomial}(n_k, Y_k) \quad Y_k^* = \frac{X_k^*}{n_k}
\]
Correcting for Statistical Bias in Uncertainty Intervals

Bootstrapping induces variance by itself:

\[ \text{Var}(Y_k^*) = \mathbb{E}(\text{Var}(Y_k^* | Y_k)) + \text{Var}(\mathbb{E}(Y_k^* | Y_K)) \]

Sampling variance of bootstrapped \( Y_k^* \) is:

\[ \hat{\sigma}_{Y_k}^2 = \frac{2Y_k^*(1-Y_k^*)}{n_k} - \frac{Y_k^*(1-Y_k^*)}{n_k^2} \]
Corrected Uncertainty Quantification

variable:
- uncorrected_var
- corrected_var
- double_corrected_var
Corrected Uncertainty Quantification

|                        | uncorrected_var | corrected_var | double_corrected_var |
|------------------------|-----------------|---------------|----------------------|
| Equal Size; Equal Perf | 0.0             | 0.0           | 99.7                |
| Unequal Size; Equal Perf | 0.0           | 0.0           | 99.3                |
| Equal Size; Unequal Perf | 15.4           | 67.6          | 94.9                |
| Unequal Size; Unequal Perf | 10.4           | 60.4          | 93.0                |

Table 3: Empirical coverage of the 95% bootstrap intervals over 1000 replicates.
Application on The Adult Income Dataset

- ~48K individuals from the 1994 census database.
- 14 features
- Label: whether an individual’s annual income was above $50K
- Split the data into 70% train and 30% test
- Trained a gradient-boosted trees classifier.
- 87% accuracy.

| #  | age | workclass | education | marital-status | occupation         | relationship | race |
|----|-----|-----------|-----------|----------------|--------------------|--------------|------|
| 25 | 25  | Private   | 11th      | Never-married  | Machine-op-imput   | Own-child    | Black|
| 38 | 38  | Private   | HS-grad   | Married-civ-spouse | Farming-fishing  | Husband      | White|
| 28 | 28  | Local-gov | Assoc-acdm | Married-civ-spouse | Protective-serv   | Husband      | White|
| 44 | 44  | Private   | Some-college | Married-civ-spouse | Machine-op-imput | Husband      | Black|
| 18 | 18  | ?         | Some-college | Never-married  | ?                  | Own-child    | White|
| 34 | 34  | Private   | 10th      | Never-married  | Other-service      | Not-in-family| White|
| 29 | 29  | ?         | HS-grad   | Never-married  | ?                  | Unmarried    | Black|
| 63 | 63  | Self-emp-not-inc | Prof-school | Married-civ-spouse | Prof-specialty   | Husband      | White|
| 24 | 24  | Private   | Some-college | Never-married  | Other-service      | Unmarried    | White|
Application on The Adult Income Dataset

Age: 8 groups. 15-95 at 10 year interval.
Race: 5 groups. White, Black, American Indian and Eskimo, Asian and Pacific Islander, and other.
Performance metrics: False positive rate, selection rate, and true positive rate.
Application on The Adult Income Dataset

The double corrected uncertainty intervals show that in some cases, had we used standard methods, we could have *erroneously* concluded that there are large disparities with statistical confidence.
Contributions and Conclusion

1. We identified meta-metrics for measuring group-wise model performance disparities, particularly in consideration of large numbers of groups.
2. We showed that these meta-metrics are statistically biased measurements.
3. We developed an unbiased estimator for between-group variance based on prior work.
4. We also developed a double-corrected estimator for quantifying the uncertainty of between-group variance.

Future work

- Examine other methods for measuring between-group variability, particularly those from the meta-analysis literature.
- Investigate corrections other metametrics such as max-min difference.

Caveat: Meta-metrics cannot capture the entirety of the impact of ML systems. Small measured disparities should not be taken as a guarantee that the system is fair or free from adverse impacts.