CWI - Information Access Group

Extended Methods to Handle Classification Biases

Emma Beauxis-Aussalet, Lynda Hardman emma@cwi.nl

Why?

Correcting Classification Biases

Classifiers can systematically confuse one class for another, and they can **over- or under-estimate the class sizes** (counts of items per class). Without correcting such biases, no scientific conclusion can be drawn from classification data.

Detailing the Error Decomposition

The biases also concern the errors between specific classes. **Within the items classified as Class X how many truly belong to Class Y?** Detailing the error decomposition between each class characterizes the quality of classification data, and enables uncertainty-aware data analyses.

How?

Reclassification Method

Also called Double Sampling [1,3], it uses error rates based on output class sizes (e.g., Precision).

$$e_{xy} = \frac{n_{xy}}{n_{.y}}$$ Error Rate (Precision-like); Output Class Size

$$\widehat{n'_{xy}} = e_{xy}\, n'_{.y}$$ Error Decomposition

$$\widehat{n'_{x.}} = \sum e_{xy}\, n'_{.y}$$ Corrected Class Size

Misclassification Method

Also called Matrix Inversion method [2-4], it uses error rates based on true class sizes (e.g., Recall).

$$\theta_{xy} = \frac{n_{xy}}{n_{x.}}$$ Error Rate (Recall-like); True Class Size

$$\widehat{n'_{xy}} = \theta_{xy}\, \widehat{n'_{x.}}$$ Error Decomposition

$$\begin{pmatrix} \widehat{n'_{1.}} \\ \widehat{n'_{2.}} \\ \dots \\ \widehat{n'_{x.}} \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{21} & \dots & \theta_{x1} \\ \theta_{12} & \theta_{22} & \dots & \theta_{x2} \\ \dots & \dots & \dots & \dots \\ \theta_{1x} & \theta_{2x} & \dots & \theta_{xx} \end{pmatrix}^{-1} \begin{pmatrix} n'_{.1} \\ n'_{.2} \\ \dots \\ n'_{.x} \end{pmatrix}$$ Corrected Class Size

New: Ratio-to-TP Method

Its atypical error ratio are based on True Positives. It gives the exact same results as the Misclassification method. It has properties of interest to ensure matrix invertibility and to predict results variance.

$$r_{xy} = \frac{n_{xy}}{n_{xx}}$$ Error Ratio; True Positives

$$\widehat{n'_{xy}} = r_{xy}\, \widehat{n'_{xx}}$$ Error Decomposition

$$\begin{pmatrix} \widehat{n'_{11}} \\ \widehat{n'_{22}} \\ \dots \\ \widehat{n'_{xx}} \end{pmatrix} = \begin{pmatrix} 1 & r_{21} & \dots & r_{x1} \\ r_{12} & 1 & \dots & r_{x2} \\ \dots & \dots & \dots & \dots \\ r_{1x} & r_{2x} & \dots & 1 \end{pmatrix}^{-1} \begin{pmatrix} n'_{.1} \\ n'_{.2} \\ \dots \\ n'_{.x} \end{pmatrix}$$ True Positives in End-Results

		True Class c_1	c_2	...	c_i	Output Count
Output Class	c_1	n_{11}	n_{21}	...	n_{i1}	$n_{.1}$
	c_2	n_{12}	n_{22}	...	n_{i2}	$n_{.2}$
	...	...	...	...	...	...
	c_i	n_{1i}	n_{2i}	...	n_{ii}	$n_{.i}$
True Count		$n_{1.}$	$n_{2.}$	...	$n_{i.}$	$n_{..}$

*Confusion matrix and notation (**n** for test set, **n'** for target set)*

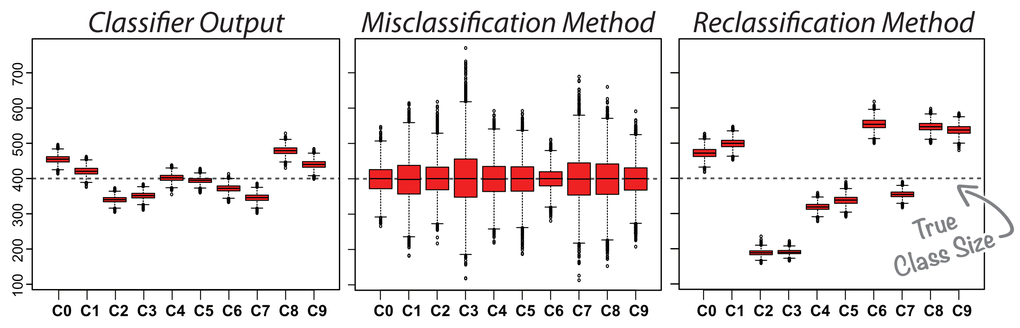
But...

Varying Class Proportions

The Reclassification method is biased if class proportions vary. **Misclassification and Ratio-to-TP methods are unaffectted.**

Varying Error Rates

Random error rate variations yield **important variance with the Misclassification and Ratio-to-TP methods.** These are not recommended for small data. Variance estimation methods depend on how the test set is sampled. We provide a novel method to addresses disjoint test and target sets.



New: Sample-to-Sample Method

Combined with Fieller's theorem, it estimates the **variance of the Misclassification method for disjoint test and target sets.**

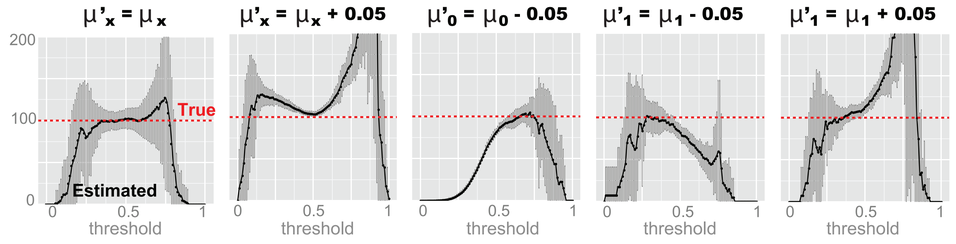
$$V\left(\widehat{\theta'_{xy}}\right) = \frac{\theta_{xy}\,(1-\theta_{xy})}{n_{x.}} + \frac{\theta_{xy}\,(1-\theta_{xy})}{n'_{x.}}$$ Error Rate Variance; Variance w.r.t. Test Set; Variance w.r.t. Target Set

New: Maximum Determinant Method

We observed that **the larger the determinant of the error rate matrix, the smaller the variance** of the Misclassification method. It can help to predict the variance, and to choose a classifier. But theory must be established (e.g., impact of number of classes, Misclassification or Ratio-to-TP error rate matrix, sample sizes).

Varying Feature Distributions

If target sets systematically differ from the test set, it can bias the error rates and thus the Misclassification & Reclassification methods. Linear models can infer error rates from feature values [5] but this approach is complex with the Misclassification method.



So...

Directions for Future Research

- **Address varying feature distributions.**
- **Identify the misclassified items individually**, given the error decomposition and class probabilities for each item.
- **Investigate Maximum Determinant method.**

[1] Tenenbein A.: A double sampling scheme for estimating misclassified multinomial data with application to sampling inspection (1972)
[2] Shieh M.S.: Correction Methods, approximate biases, and inference for misclassified data (2009)
[3] Buonaccorsi J.P.: Measurement error: models, methods and applications (2010)
[4] Beauxis-Aussalet E. et al.: Multifactorial uncertainty assessment for monitoring population dynamics (2015)
[5] Boom B.J., Beauxis-Aussalet E. et al.: Uncertainty-aware estimation of population abundance using machine learning (2016)