Proposal: A fair comparison of class imbalance correction methods

Alex Carriero (9028757)

1 | Introduction

The prevalence of prediction modelling employed in the field of medicine is rapidly increasing. In medicine, the goal of a prediction model is often to (accurately) predict a patient's risk of experiencing an event (e.g., stroke). Outcomes are therefore, binary (e.g., a patient either does, or does not, experience a stroke). Patients who experience the event are collectively referred to as the positive class, while those who do not, are referred to as the negative class. Typically, when developing clinical prediction models for binary outcomes, the information available to train the model is imbalanced (i.e., the number patients in one class dramatically outnumbers the other). This is referred to as class imbalance, and is often seen as a major problem in the field of machine learning.

Class imbalance in training data is thought to degrade the quality of prediction models. However, the quality of any prediction model is multi-faceted and is characterized by three specific criteria: discrimination, calibration and accuracy. Discrimination refers to a model's ability to yield higher risk estimates for patients in the positive class than for those in the negative class; it is quantified by area under the receiver operating curve (AUROC), also referred to as the concordance statistic. Calibration quantifies the reliability of the risk predictions themselves; for instance, a poorly calibrated model may produce risk predictions that consistently over- or under-estimate reality, or may produce risk estimates which are too extreme (too close of 0 or 1) or too modest. Finally, accuracy refers to the proportion of patients a model classifies correctly after a risk threshold is imposed. In other words, once a risk threshold is chosen, predictions are transformed back into a binary variable such that patients with a risk prediction greater than the threshold are predicted to belong to the positive class and those below it to the negative class, accuracy is then proportion of patients (in either the positive and negative class) that were classified correctly; model specificity and sensitivity may also often used as metrics to quantify accuracy. It is imperative to acknowledge that a model can have high accuracy and discrimination while the reliability of the risk estimates produced by the model (calibration) is poor (Van Calster et al. 2019). This is of particular relevance in the field of medicine, as the quality of the risk predictions themselves determine the clinical utility of a model. This is because it risk predictions tare used directly by clinicians to council patients and inform treatment decisions. Thus, if prediction models are not well calibrated, in a medical context, the consequences may be catastrophic.

This leads us to the dilemma which motivates our research question: class imbalance is thought to be a problem that needs solving. While an abundance of class imbalance correction methods exist, and are summarized well by Maheshwari, Jain, and Jadon (2018) and Lopez et al. (2013), comparisons of these various class imbalance corrections typically fail to assess model calibration. Recall that in medicine, data sets are often heavily imbalanced, due to the (thankfully) rare nature of many diseases, yet, it is unacceptable to apply corrections for class imbalance without consideration for their effects on model calibration, as calibration determines a model's clinical

utility. Furthermore, a recent study demonstrated that when predictive models are developed using logistic regression, applying particular class imbalance corrections dramatically deteriorated the calibration of the model (Goorbergh et al. 2022).

** RESEARCH QUESTION HERE... fair comparison of class corrections where calibration considered pls **

extra sentences:

- Three broad categories of class imbalance corrections emerge from the literature: pre-processing techniques, cost-sensitive learning and ensemble learning.

2| Analytic Strategy

2.1| Example

2.2| Simulation Study

study should be fair .. (Morris, White, and Crowther 2019)

2.3| Ethical Approval and other specifics

tech stuff gunna use r (R Core Team 2021).

References

Goorbergh, Ruben van den, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster. 2022. "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression." *Journal of the American Medical Informatics Association* 29 (9): 1525–34. https://doi.org/10.1093/jamia/ocac093.

Lopez, Victoria, Alberto Fernandez, Salvador Garcia, Vasile Palade, and Francisco Herrera. 2013. "An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics." *Information Sciences* 250: 113–41.

Maheshwari, Satyam, R. C. Jain, and R. S. Jadon. 2018. "An Insight into Rare Class Problem: Analysis and Potential Solutions." *Journal of Computer Science* 14 (6): 777–92.

Morris, Tim P., Ian R. White, and Michael J. Crowther. 2019. "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine* 38 (11): 2074–2102.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Van Calster, Ben, David J. McLernon, Maarten van Smeden, Laure Wynants, Ewout W. Steyerberg, Patrick Bossuyt, Gary S. Collins, et al. 2019. "Calibration: The Achilles Heel of Predictive Analytics." *BMC Medicine* 17 (1): 230. https://doi.org/10.1186/s12916-019-1466-7.