Proposal: A fair comparison of class imbalance correction methods.

Alex Carriero (9028757)

Date: 13/10/2022

Word Count: XXX

Program: *Methodology and Statistics for Behvarioual, Biomedical, and Social Sciences*

Supervisors: dr. Maarten van Smeden, Utrecht University Medical Center and dr. Ben van Calster, Leuven University and Leiden University Medical Center.

Host Institution: Julius Center for Health Science and Primary Care, UMC.

Candidate Journal: Statistics in Medicine

FETC-approved: 22-XXXX

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 a particular event (e.g., a stroke). Patient outcomes are therefore, binary (event or no event). When developing clinical prediction models for binary outcomes, the information available to train the model is typically imbalanced (i.e., the number patients in one class dramatically outnumbers the other) (Maheshwari, Jain, and Jadon 2018). This is referred to as class imbalance, and is often seen as a major problem in the field of machine learning (Maheshwari, Jain, and Jadon 2018). Furthermore, due to the (thankfully) rare nature of many diseases, class imbalance is especially common in medical data sets (Maheshwari, Jain, and Jadon 2018).

Class imbalance is thought to diminish the quality of prediction models (Yu and Zhou 2021). "Quality" in this context is too general a term as the quality of a prediction model is multifaceted. It is characterized by three criteria: accuracy, discrimination, and calibration. Accuracy refers to the proportion of patients that a model classifies correctly (after a risk threshold is imposed). 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. Finally, calibration refers to 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 produce risk estimates which are too extreme (too close of 0 or 1) or too modest.

Calibration is the metric which is most interesting when developing clinical prediction models. This is because in practice, the risk predictions from the model are given directly to a clinician who will use the information to council patients and inform treatment decisions. Thus, it is essential that these predictions are accurate (i.e., calibration is good) otherwise, the personal costs to the patient may be enormous. It is entirely possible that a model has great accuracy and discrimination while the reliability of the risk predictions produced by the model (calibration) is poor (Van Calster et al. 2019). Thus, all three criteria should be considered when discussing the consequences of class imbalance on the quality of clinical prediction models. This is rarely the case and unfortunately, it is often calibration that is forgotten (Van Calster et al. 2019).

Class imbalance is not unique to medical data sets thus, literature focusing on imbalance correction methods arises from many disciplines. An abundance of imbalance corrections exist and are well summarized by Haixiang et al. (2017). Yet, information regarding the effect of theses corrections on model calibration is sparse. One recent study demonstrated that implementing common imbalance corrections lead to dramatically deteriorated model calibration, to the extent that no correction was recommended (Goorbergh et al. 2022). In this study, models were developed using logistic regression and penalized (ridge) logistic regression (Goorbergh et al. 2022).

Motivated by the work of Goorbergh et al. (2022), we aim to assess the impact of imbalance corrections on model calibration when prediction models are trained with a wider variety of classification algorithms: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), XG Boost (XG), RUSBoost (RB) and Easy Ensemble (EE). More specifically, we aim to determine if any combination of class imbalance correction and classification algorithm can out preform the classification algorithm alone. Special attention will be paid to model calibration while accuracy and discrimination will also be considered.

2| Analytic Strategy

To provide a fair comparison of imbalance corrections a simulation study will be designed and implemented. Subsequently, we will demonstrate the implementation of all methods considered in the simulation through a case study of empirical data.

2.1 Simulation Study

Classification Algorithms: LR, SVM, RF, XG, RB, EE

Imbalance Corrections: none, RUS, ROS, SMOTE, SMOTE-ENN

6 x 5 = 30 combinations.

We will adhere to the ADEMPT guidelines for the design and reporting of our simulation study (Morris, White, and Crowther 2019).

2.2 Case Study

– discuss details wednesday ? –

2.3 Technical Considerations

All analyses will be conducted using the open source statistical soft R (R Core Team 2021). Furthermore, the simulation study planned is quite computationally intensive, therefore, we intend to run the simulation using the super computer at the UMC.

References

Branco, Paula, Luis Torgo, and Rita P. Ribeiro. 2016. "A Survey of Predictive Modeling on Imbalanced Domains." *ACM Comput. Surv.* 49 (2). https://doi.org/10.1145/2907070.

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.

Haixiang, Guo, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. "Learning from Class-Imbalanced Data: Review of Methods and Applications." *Expert Systems with Applications* 73: 220–39.

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.

Yu, Lian, and Nengfeng Zhou. 2021. "Survey of Imbalanced Data Methodologies." arXiv. https://doi.org/10.48550/ARXIV.2104.02240.