Proposal: assessing the impact of class imbalance corrections on model calibration.

Alex Carriero (9028757)

1 | Introduction

Prediction modelling in medicine is gaining increasing attention; clinicians are often interested in predicting a patient's risk of disease. Due to the (thankfully) rare nature of many diseases, the data available to train clinical prediction models are often heavily imbalanced (i.e., the number of patients in one class dramatically outnumbers the other) [**summary_m**]. This is referred to as class imbalance. Class imbalance is seen major problem in the field of machine learning as it is known to degrade model performance [**yu**]. Consequently, imbalance correction methodologies are proposed as a solution [**yu**].

An ideal imbalance correction will improve all aspects of model performance. These criteria include: classification 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 [**achilles**].

Class imbalance is not unique to medical data sets and literature introducing imbalance correction methods arises from many disciplines. An abundance of imbalance corrections exist and are summarized by [**summary_b**, **summary_lp**, **summary_m**, **summary_h**, **summary_k**]. Information regarding the effect of these corrections on model calibration is sparse. In medicine, it is essential that model calibration is assessed. This is because in practice, risk predictions from the model are given directly to a clinician who will use the information to council patients and inform treatment decisions. If a model is poorly calibrated, the personal costs to the patient may be enormous. It is entirely possible for a model to exhibit excellent classification accuracy and discrimination while calibration is poor [**achilles**]. Therefore, assessing only discrimination and accuracy is insufficient.

Only one study has assessed the impact of imbalance corrections on model calibration. **rubenempty citation** demonstrated that implementing imbalance corrections lead to dramatically deteriorated model calibration, to the extent that no correction was recommended [**ruben**]. In this study, models were developed using logistic regression and penalized (ridge) logistic regression [**ruben**]. Motivated by the work of **rubenempty citation**, we must ensure that the "cure" is not worse than the disease. In our research, we aim to assess the impact of imbalance corrections on model calibration for prediction models trained with a wider variety of classification algorithms including: linear classifiers, ensemble learning algorithms and algorithms specifically designed to handle class imbalance. Furthermore, we aim to answer the question: can imbalance corrections improve overall model performance without comprising model calibration?

2| Analytic Strategy

We will evaluate the performance of several imbalance corrections in a simulation study. We will adhere to the ADEMP guidelines for the design and reporting of our simulation [**tim_morris**].

2.1| Simulation Study

The aim of the simulation study is to determine which pair(s) of imbalance correction and classification algorithm can outperform the classification algorithms without imbalance corrections.

Imbalanced data will be simulated to reflect 27 scenarios. The following criteria will be varied: number of predictors, event fraction and sample size. The number of predictors will vary through the set {8,16,32} and event fraction, through the set {0.5, 0.2, 0.02}. The minimum sample size for the prediction model (N) will be computed according to formulae from **rileyempty citation**. Sample size will then vary through the set $\{\frac{1}{2}N, N$ and $2N\}$.

Under each scenario, 2000 data sets will be generated. More specifically, test and training data will be generated such that the training set is 10x larger than the test set. Each simulated data set will be analysed by 30 methods = 6 (classification algorithms) x 5 (imbalance corrections). The classification algorithms and imbalance corrections we will include in our simulation are detailed in Table 1.

Table 1: Classification algorithms and imbalance corrections to be evaluated.

Index	Classification Algorithms	Imbalance Corrections
1	Logistic Regression	None
2	Support Vector Machine	RUS (random under sampling)
3	Random Forest	ROS (random over sampling)
4	XG Boost	SMOTE (synthetic majority over sampling)
5	RUSBoost	SMOTE - ENN (SMOTE - edited nearest neighbours)
6	Easy Ensemble	

Finally, performance criteria will include measures of model discrimination, accuracy and calibration. Discrimination will be measured by area under the receiver operator curve (AUROC). Classification accuracy by Matthew's correlation coefficient (MCC), overall accuracy, sensitivity and specificity. Finally calibration will be measured in terms of calibration intercept and calibration slope.

2.2| Software

All analyses will be conducted using the open source statistical software R [**base**]. Additionally, our simulation study is expected to be quite computationally intensive. Therefore, we intend to run the simulation using the high performance computers at the UMC.