Simulation Study Protocol

Alex Carriero

Department of Methodology and Statistics, UU
Julius Center, UMCU

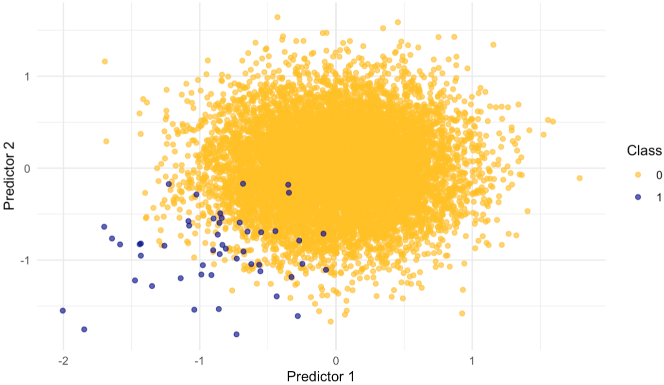
Overview

- Introduction
- Research Question
- Simulation Protocol
  - **A**ims
  - **D**ata generating mechanism
  - **E**stimands
  - **M**ethods
  - **P**erformance metrics
- Discussion / Questions

Introduction

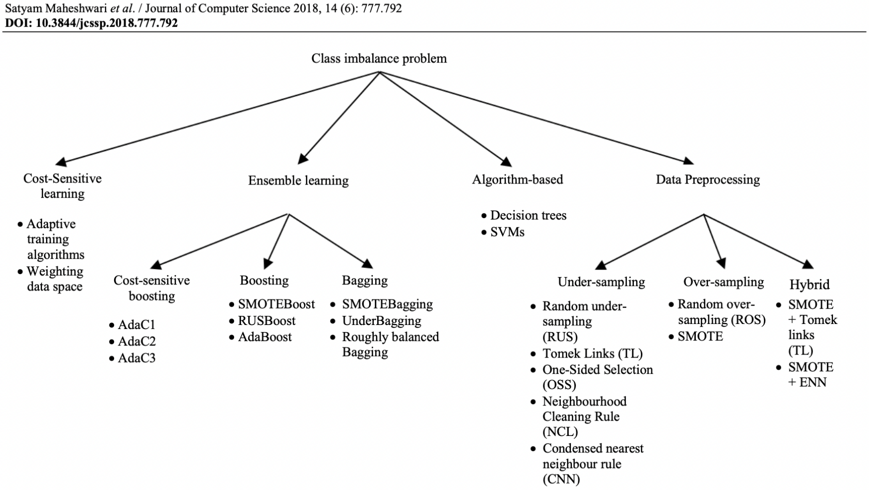
Context: Clinical prediction modelling for dichotomous risk prediction.

Problem: Class Imbalance



Introduction

Solution:



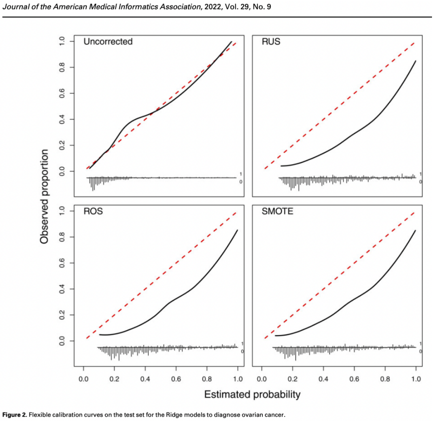
Satyam Maheshwari *et al.* / Journal of Computer Science 2018, 14 (6): 777.792
DOI: 10.3844/jcssp.2018.777.792

Class imbalance problem

- Cost-Sensitive learning
  - Adaptive training algorithms
  - Weighting data space
- Ensemble learning
  - Cost-sensitive boosting
    - AdaC1
    - AdaC2
    - AdaC3
  - Boosting
    - SMOTEBoost
    - RUSBoost
    - AdaBoost
  - Bagging
    - SMOTEBagging
    - UnderBagging
    - Roughly balanced Bagging
- Algorithm-based
  - Decision trees
  - SVMs
- Data Preprocessing
  - Under-sampling
    - Random under-sampling (RUS)
    - Tomek Links (TL)
    - One-Sided Selection (OSS)
    - Neighbourhood Cleaning Rule (NCL)
    - Condensed nearest neighbour rule (CNN)
  - Over-sampling
    - Random over-sampling (ROS)
    - SMOTE
  - Hybrid
    - SMOTE + Tomek links (TL)
    - SMOTE + ENN

Introduction

Goorbergh et al. (2022) results:



Journal of the American Medical Informatics Association, 2022, Vol. 29, No. 9

Figure 2. Flexible calibration curves on the test set for the Ridge models to diagnose ovarian cancer.

We need to make sure that the cure is not worse than the disease.

Research Question

Research Question: Can class imbalance corrections improve the performance of clinical prediction models, without compromising model calibration?

Simulation Protocol: Aim

Aim: Can class imbalance corrections improve the performance of clinical prediction models, without compromising model calibration?

Building on work of Goorbergh et al. (2022):

- consider variety of classification algorithms
- consider further imbalance corrections

Fair comparison: to determine which (if any) pair of imbalance correction and classification algorithm can outperform the classification algorithm alone.

Simulation Protocol: Data Generating Mechanism

- We consider 27 (3 x 3 x 3) scenarios,
  - Number of predictors: 8, 16, 32
  - Event fraction: 0.5, 0.2, 0.02
  - Sample Size: $\frac{1}{2}N$, N, $2N$

where N represents the minimum sample size for the prediction model calculated according to formulae in Riley et al. (2022).

- Data sets will have AUC = 0.85 (SD).
- For each scenario, generate 2000 data sets ⇒ 54000 data sets in total.
- Data sets partitioned into test and training data such that training set is 10x larger than the test set.

Simulation Protocol: Data Generating Mechanism

Data Generating Mechanism: data for each class is generated independently from two distinct *multivariate normal* distributions.

Class 0:

- choose n_0
- generate predictors $\sim \mathit{mvn}(\mu_0, \Sigma)$

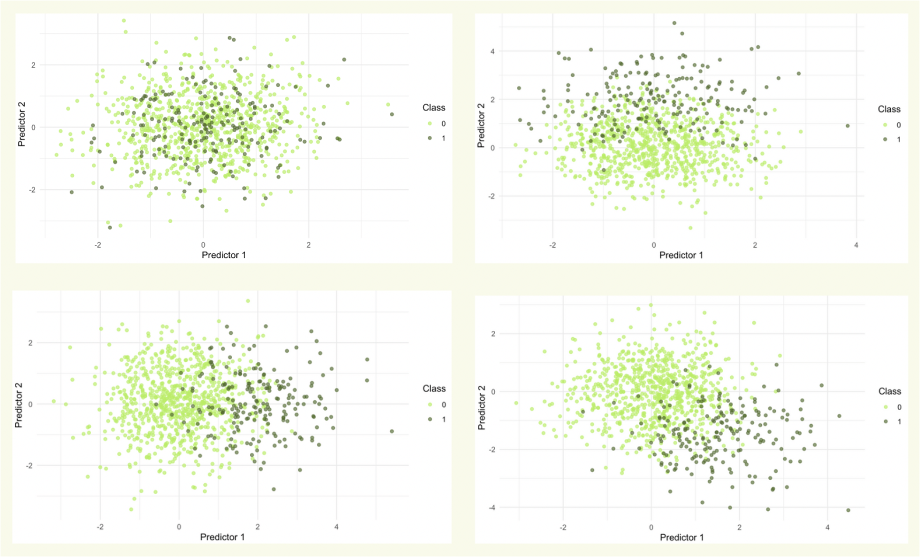
Class 1:

- choose n_1
- generate predictors $\sim \mathit{mvn}(\mu_1, \Sigma)$

Why?

- control event fraction
- control ΔC statistic (AUROC)
- not generated under a specific model (i.e., logistic regression)

Simulation Protocol: Data Generating Mechanism



Idea: vary parameters (μ and Σ) to control overlap of the distributions.

Simulation Protocol: Estimands

Estimands: out-of-sample predictive performance of clinical prediction models for dichotomous risk prediction.

Simulation Protocol: Methods

Full-Factorial Design: 5 imbalance corrections x 6 classification algorithms:

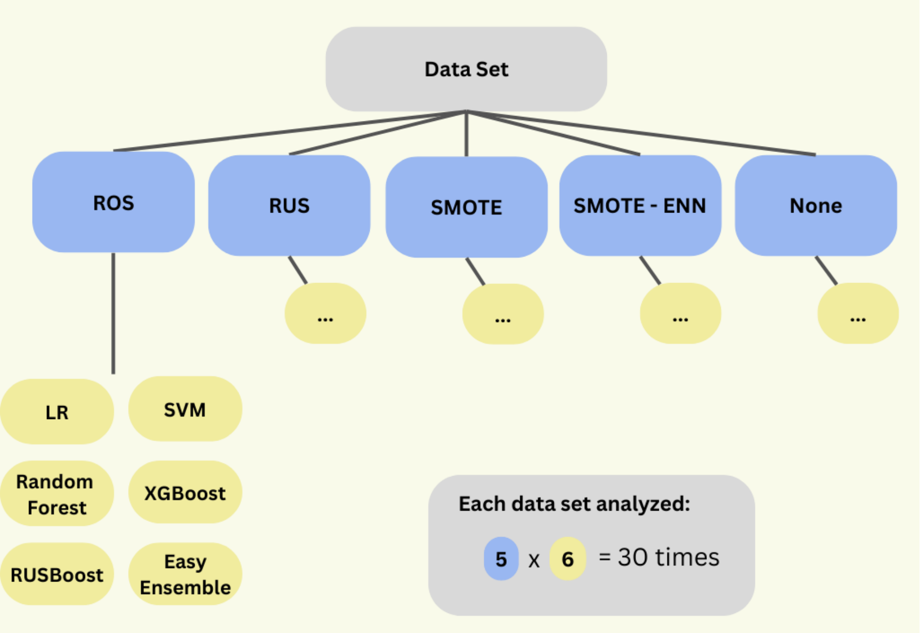
Imbalance Corrections: all pre-processing techniques

- Random Over Sampling (ROS)
- Random Under Sampling (RUS)
- Synthetic Minority Oversampling TEchnique (SMOTE)
- SMOTE-Edited Nearest Neighbours (SMOTE-ENN)
- None

Classification Algorithms:

- Logistic Regression
- Support Vector Machine
- Random Forest
- XG Boost
- RUSBoost
- EasyEnsemble

Simulation Protocol: Methods



Simulation Protocol: Performance Metrics

Performance Metrics: Out-of sample performance assessed in terms of accuracy, discrimination and calibration. Predictions generated using the test data set.

Accuracy:

- overall accuracy
- sensitivity
- specificity
- MCC

Discrimination:

- area under the receiver operator curve (ΔC Statistic)

Calibration:

- calibration intercept
- calibration slope

Discussion / Questions

- Under the data generating mechanism chosen, it is unclear how (if possible) to include interaction effects.

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.

Riley, Richard D., Gary S. Collins, Joie Ensor, Lucinda Archer, Sarah Booth, Sarwar I. Mozumder, Mark J. Rutherford, Maarten van Smeden, Paul C. Lambert, and Kym I. E. Snell. 2022. "Minimum Sample Size Calculations for External Validation of a Clinical Prediction Model with a Time-to-Event Outcome." *Statistics in Medicine* 41 (7): 1280–95.