Long COVID Prediction in Manitoba Using Clinical Notes Data: A Machine Learning Approach

Presented by: Surani Matharaarachchi

Joint work with:
Dr. Saman Muthukumarana, Dr. Mike Domaratzki, Dr. Alan Katz

November, 17 2023

University of Manitoba

Introduction

Long COVID Syndrome (LCS)

- A condition in which individuals experience symptoms for weeks or months after recovering from COVID-19.
- The need for consistent identification and treatment of Long COVID patients
  - 20-30% of COVID-19 survivors experience prolonged symptoms.
  - The condition can affect multiple organ systems.
  - Many are unaware of their condition.

Predictive Models for LCS

Challenges in Predicting LCS Patients at Risk

- Identifying ‘known LCS’ group for classification
  - Use Natural Language Processing (NLP) methodologies.
  - Conduct word extraction processes.
  - Perform manual refinement techniques.

- Class imbalance issue (Ratio: 0.96:0.04)
  - Used rebalancing techniques
  - Random Over-Sampling and Random Under-Sampling

Predicting Potential LCS Patients

- LCS Symptoms, Pre-COVID Symptoms, Sex, Sefi, Age Category
- Pre-COVID Symptom Scenario: within 90 days of the COVID index date
- Logistic Regression with Elastic Net Regularization
- Random Under-Sampling
- AUC - 0.94, Sensitivity - 0.95, Specificity - 0.81
- Identified LCS group in Risk: 1124 (24.7%) LCS patients from the set of 4556 COVID-19 cases

Class Imbalance Issue

- One or more classes are underrepresented.

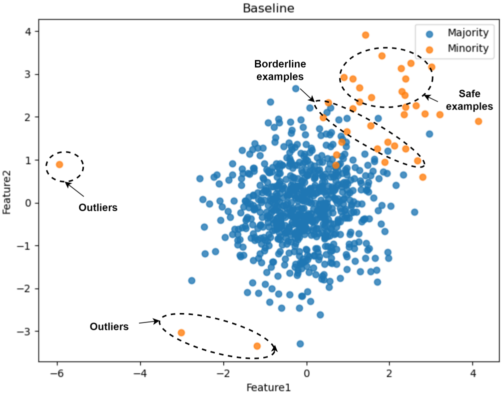


Figure: Outliers in minority class

Synthetic Minority Over-Sampling Technique (SMOTE)

- Create new samples for the minority class, helping to balance the dataset.
- Challenged by outliers within the minority class.

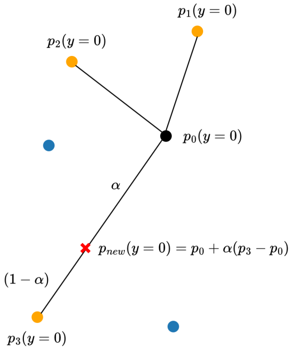


Figure: SMOTE data generation

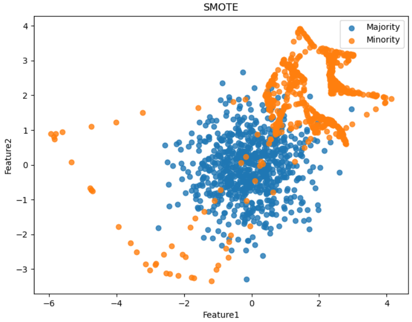


Figure: Re-sampled data with SMOTE

Novel Methods for Addressing Class Imbalance with Outliers

- Using a weighted average of neighbouring instances
- Improved robustness against outliers and noisy data

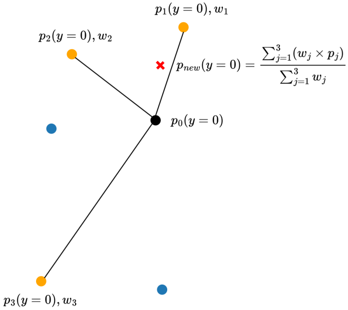


Figure: Proposed method data generation

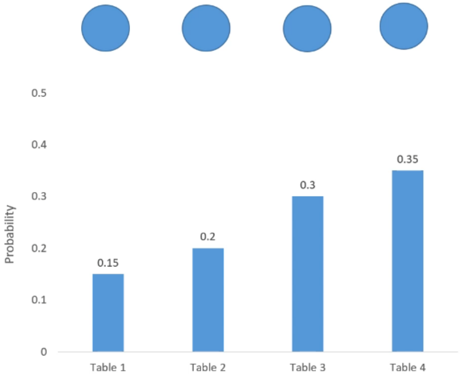
Novel Methods for Addressing Class Imbalance with Outliers

- Developing new SMOTE extensions
- Inverse distance between the median centroid of the minority class and the nearest neighbours

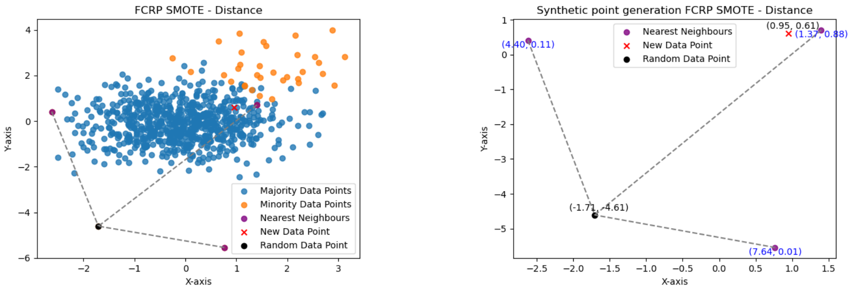
1. Distance extSMOTE
2. Dirichlet extSMOTE [1]
3. FCRP SMOTE (Finite Chinese Restaurant Process based SMOTE)

FCRP SMOTE

Showcasing the weight selection of FCRP SMOTE using Finite Chinese restaurant process with scaling parameter $\alpha = 0.1$



Synthetic Point Generation



(a) label 1

(b) label 1.1

Figure: One instance of generating a sample - FCRP SMOTE

Simulation Results

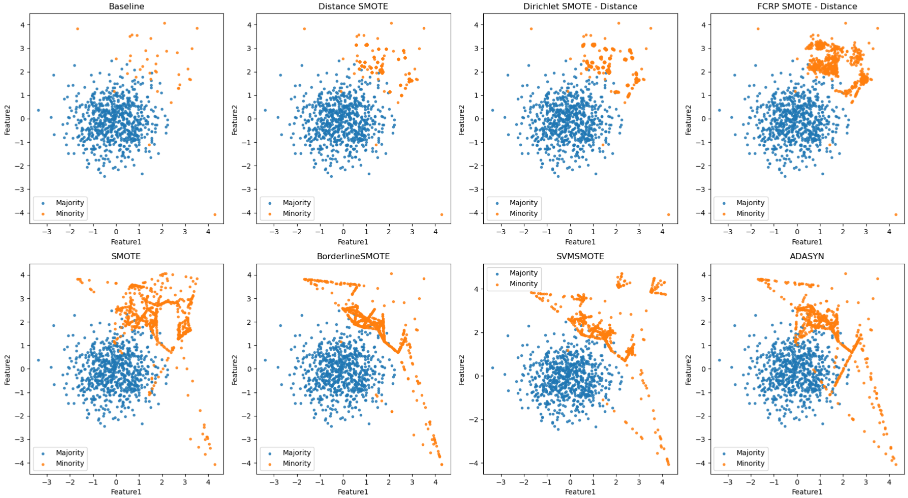


Figure: Comparison of resampled data

Simulation Results

- $X_{minority-outliers} \sim \mathcal{N}(\mu_1, \Sigma_1)$
- $X_{majority} \sim \mathcal{N}(\mu_2, \Sigma_2)$
- $X_{outliers} \sim Uniform(-10, 10)$

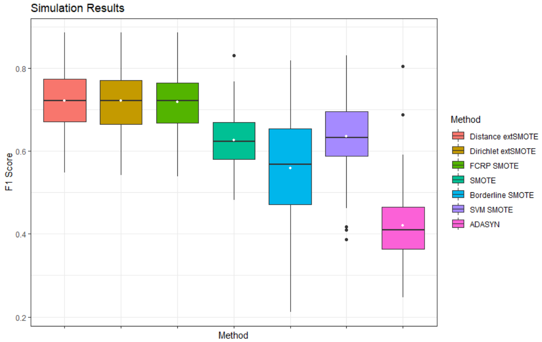


Figure: F1 Scores for 100 simulated datasets with 5-fold cross validation

Application Results

- 11 imbalanced datasets in UCI repository
- diabetes, mammographic_masses, ecoli, breast_cancer, abalone_19, isolet, car_eval_34, thyroid_sick, us_crime, oil, spectrometer

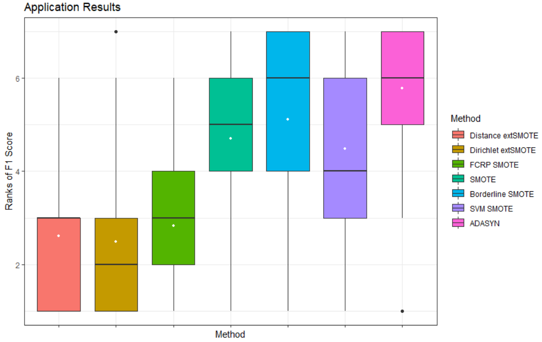


Figure: F1 Score Ranks for the datasets with 7 $\times$ 5-fold cross validation

Conclusion and Future Work

- Application of NLP in conjunction with machine learning techniques enables identifying established LCS patients at risk.
- Addressing class imbalance stands as a substantial challenge in classification tasks.
- Outliers within the minority class significantly affect SMOTE and related extensions.
- The proposed methodologies exhibit superior performance compared to existing techniques, showcasing efficacy in both simulated and application data, even in outlier-free scenarios.
- The proposed methods will be applied to predicting Long COVID patients in Manitoba.

References

[1] Bej, S., N. Davtyan, M. Wolfien, M. Nassar, and O. Wolkenhauer (2021). Loras: an oversampling approach for imbalanced datasets. *Machine learning 110*(2), 279–301.

Acknowledgment

I would like to express my special thanks of gratitude to

- my supervisors, Dr. Saman Muthukumarana and Dr. Mike Domaratzki, for their excellent guidance.
- Dr. Alan Katz for the constructive feedback and the Manitoba Centre for Health Policy (MCHP) for providing the data.
- the Department of Statistics and the staff for funding and resources.
- my family and friends for their continuous support.

Thank You!

Contact: matharas@myumanitoba.ca