Optimal selection of resampling methods in imbalanced data with high complexity

Annie Kim*, Inkyung Jung**

* Department of Biostatistics and Computing, Yonsei University Graduate School

** Division of Biostatistics, Department of Biomedical Systems Informatics, Yonsei University College of Medicine

Introduction

Background

The **overgeneralization problem** is a situation in which examples produced by the oversampling technique are introduced into the majority class domain. Some results have shown that oversampling worsens classification performance due to this problem. This study claim that this problem aggravates in complex data settings.

To mitigate the problem of overgeneralization in complex datasets, this study advises the use of two alternative approaches.

Purpose

The purpose of this study is to investigate the relationship between complexity and imbalance for classification. Through various scenarios of simulation and real data, an optimal resampling method for complex datasets is provided.

Method

To this day, new resampling methods are being developed, but they always fall into 3 categories: **oversampling**, **undersampling**, and **hybrid**.

Oversampling

Oversampling balances the number of samples between classes by adding an instance copy of an underrepresented class or generating synthetic data.

Undersampling

Undersampling is an efficient technique that does not need adding new data in imbalance dataset. It balances the number of samples between classes by deleting unnecessary instances.

Oversampling with filter(hybrid)

By cleaning the space resulting from oversampling, the overgeneralization problem can be resolved.

Oversampling	Undersampling	Oversampling with filter
• Random Over Sampling(ROS) • Synthetic minority oversampling technique (SMOTE) • Adaptive synthetic sampling technique (ADASYN) • Borderline SMOTE • SVM SMOTE • KMeans SMOTE	• Random Under Sampling (RUS) • Near Miss (NM) • Tomek Link (TL) • Condensed Nearest Neighbors (CNN) • Edited Nearest Neighbors (ENN) • Repeted Edited Nearest Neighbors (RENN) • ALL KNN • One Sided Selection (OSS) • Neighborhood Cleaning Rule (NCR) • Instance Hardness Threshold (IHT)	• SMOTE-Tomek Link • SMOTE-ENN • Dynamic SMOTE radial basis function (DSRBF) • TRIM-SMOTE • SMOTE-RSB* • NRSBoundary • NEATER • SMOTE-IPF • SMOTE-FRST-2T • NRAS

Simulation study

To investigate the relationship between data characteristics and selection of resampling methods, simulated data were generated with various combinations of concept of complexity, training set size, and degree of imbalance. Different kinds of resampling have been applied to the generated data.

Data generation

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \widetilde{x_4} \\ \widetilde{x_5} \\ \widetilde{x_6} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & \rho & 1 \end{bmatrix} \right), \quad where\ \rho = 0.3$$

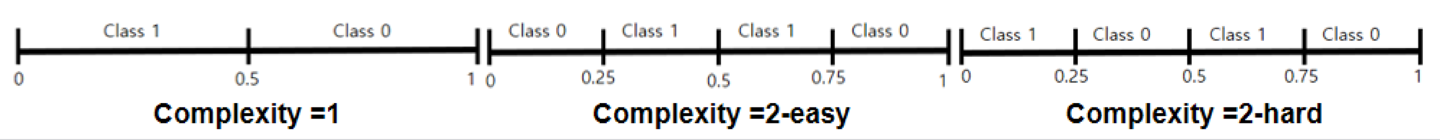
$$\mathrm{X}_4 = \begin{cases} 0\ if\ \widetilde{x_4} < \Phi^{-1}(0.3) \\ 1 \quad\quad o.w \end{cases}$$

$$\mathrm{X}_5 = \begin{cases} 0\ if\ \widetilde{x_5} < \Phi^{-1}(0.2) \\ 1 \quad\quad o.w \end{cases}$$

$$\mathrm{X}_6 = \begin{cases} 0\ if\ \widetilde{x_6} < \Phi^{-1}(0.15) \\ 1 \quad\quad o.w \end{cases}$$

$$\eta = {}^{1}\Big/_{\exp\left(-1 * \left(\frac{\mathrm{X}_1}{2} + \frac{\mathrm{X}_2}{4} + \mathrm{X}_4 - X_6 + \epsilon\right)\right) + 1} \qquad \epsilon \sim N(0,1)$$

Generated data were labeled using three different complexity level(c = 1,2-easy,2-hard). Three data set size and two level of class imbalance level were considered. By controlling complexity(c), imbalance(i) and size(s), we were able to generate 12 domains. Each domain was generated 50 times.

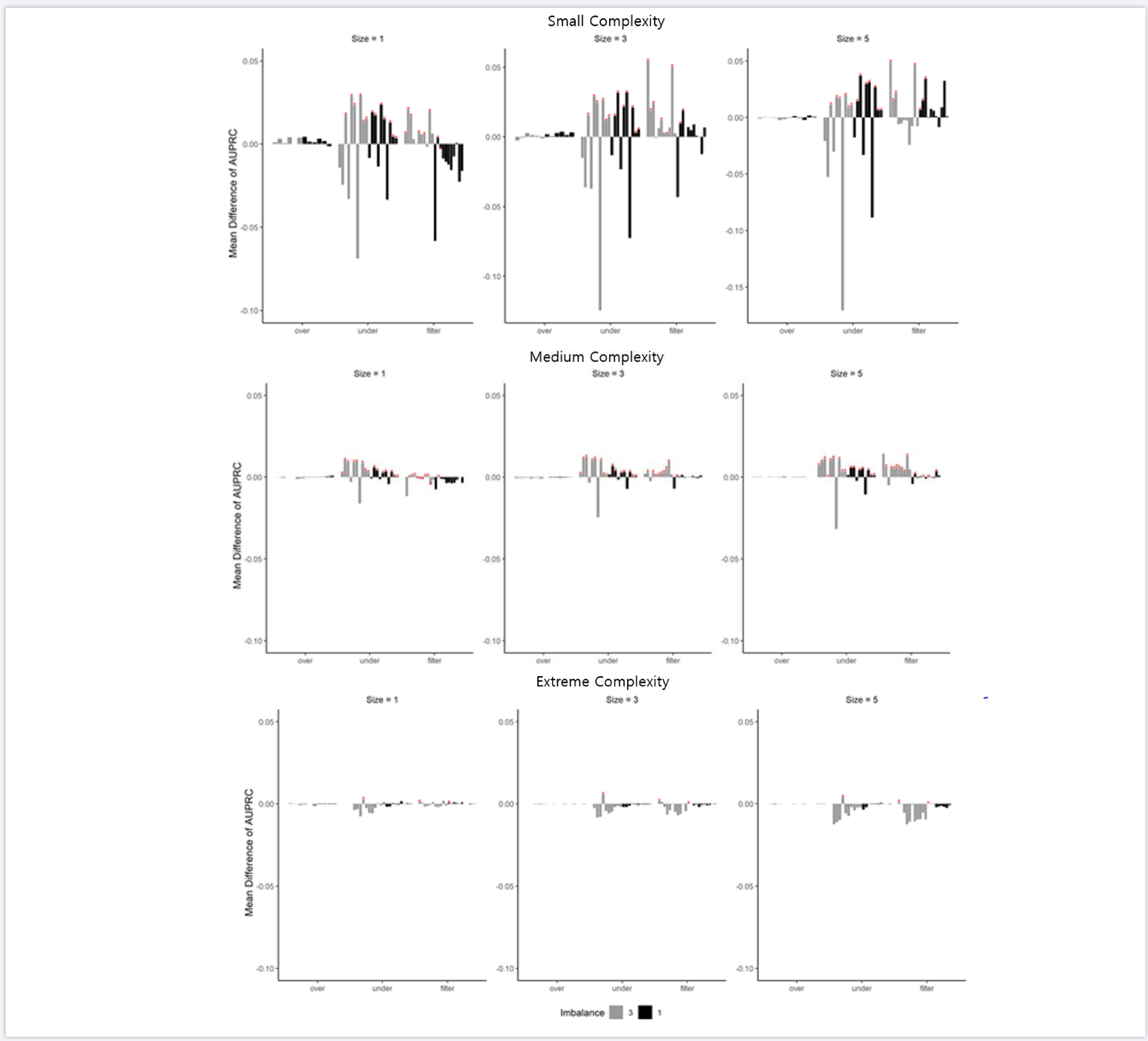


Decision tree classifier is used with parameter chosen by 3-fold cross validation. For performance evaluation AUPRC(area under precision recall curve) is used. The details of simulation framework are described in table below.

c	s	i	N	N+	IR
1,2 (easy, hard)	1	1	166	156	15.6
		3	195	156	4
	3	1	664	625	16
		3	781	625	4
	5	1	2656	2500	16
		3	3125	2500	4

Results

The result shows an increase or decrease in AUPRC when applying the resampling method at each imbalance, complexity, and sample size. Positive values indicate performance improvement; negative value indicate performance degradation.



- As the complexity of the data increases, it is more difficult to improve performance through resampling.
- Oversampling shows the least performance gain in all cases.
- In the case of undersampling, the degree of imbalance is less affected, but in the case of the oversampling with filter method, the performance difference varies depending on the degree of imbalance even at the same complexity.

Real data analysis

Real data were used to determine the relationship between data complexity and selection of resampling methods. Optimal resampling methods for each classification method using complexity measure were analyzed.

Data

109 labeled datasets are from UCI repository. For representation of the data characteristics ‘complexity measure’ is used.

complexity measure	
F1v	Directional-vector maximum Fisher’s discriminant ratio. complements F1 by searching for a vector able to separate two classes after the training examples have been projected into it
C2	Index computed for measuring class balance. Larger values of C2 are obtained for imbalanced problems.

Result

The result is divided into complex and non-complex data through the complexity measure. Decision Tree(DT), Random Forest(RF), Neural Network(NN), k-NN, and SVM were used as the classifier, and the tables show top 10 ranked results out of 130 combinations.

F1v top 25% datasets

rank	algorithm	resampling
1	RF	IPF
2	RF	SMOTETL
3	RF	NRSBoundary
4	RF	PSO
5	RF	FRST_2T
6	DT	NRSBoundary

F1v bottom 25% datasets

algorithm	resampling
RF	NRAS
RF	IPF
RF	DSRBF
RF	RSB
RF	SMOTE
RF	KMeansSMOTE

C2 top 25% datasets

rank	algorithm	resampling
1	RF	FRST_2T
2	RF	IPF
3	RF	NRSBoundary
4	RF	PSO
5	KNN	IPF
6	DT	FRST

C2 bottom 25% datasets

algorithm	resampling
RF	SMOTETL
RF	IPF
RF	DSRBF
RF	NRSBoundary
RF	NCR
RF	FRST_2T

oversampling
undersampling
oversampling with filter

- In the case of high F1v and C2, the combination of random forest and filtering method seems to be the best.
- When F1v is low(=low complexity), oversampling is one of the higher ranks.
- When C2 is low(=low imbalance), undersampling is one of the higher ranks.

Conclusion

This paper shows the optimal resampling method through various simulation scenarios and real data. It shows that when choosing a resampling method, data complexity and the imbalance ratio needs to take into account.