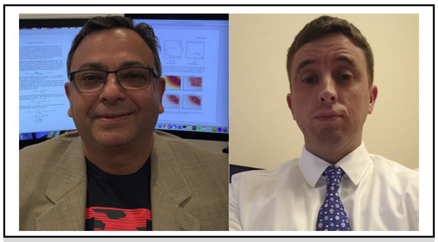
See Article page 1926.

Commentary: The problem of class imbalance in biomedical data

Hemant Ishwaran, PhD,[a] and Robert O'Brien, PhD[b]



Hemant Ishwaran, PhD, and Robert O'Brien, PhD

CENTRAL MESSAGE

Class-imbalanced data frequently occur in medical studies, which poses thorny issues for machine learning methods. Recent developments can provide a clear path forward in analyzing these data.

The main focus of the work by Bolourani and colleagues[1] is the development of a machine learning (ML) algorithm for predicting early readmission after esophagectomy. The authors provide a detailed multistep analysis that includes univariate and multivariate logistic regression, regularized lasso, random forests, and NearMiss. This is obviously a very complex analysis, and so at first glance readers might question why studying a simple binary outcome such as hospital readmission would entail so much effort. As the authors correctly identify, the difficulty here occurs due to the presence of "class-imbalanced data," which turns out to be a thorny scenario for ML procedures to overcome. Here we comment on and provide promising new developments for addressing this problem.

In class-imbalanced data, or simply imbalanced data, the outcome is binary (here, early readmission), such that the frequency of the observed classes is skewed to one realization—the majority class—vs the other possible realization—the minority class. In the analysis here, only 383 of the 2037 patients studied required early readmission: thus, the frequency of early readmission (minority class) to those not readmitted early (majority class) is 383 to 1654, an imbalance ratio (IR) of 4.3.

The problem is that many ML methods are "biased" toward the majority class in the presence of imbalanced data, especially when the IR is high. This is because ML classification is generally based on the Bayes decision rule, which classifies patients on the basis of their probabilities, with patients with probability ≥ 0.5 assigned to the minority class. Of course, the very nature of the imbalanced data makes this unlikely to occur, as the probability of being a minority class will almost certainly be less than 0.5 (except, perhaps, for a small subset), especially when IR is high. Hence ML classifiers tend to classify most of the data into the majority class in imbalanced data settings. Note that the same principle applies to standard procedures such as logistic regression if these use the Bayes decision rule for classification.

What is the answer? In ML, one approach has been to use what are called undersampling and oversampling techniques. As an example of oversampling, SMOTE[2] is a popular technique that creates artificial minority class examples in an effort to balance the data. Thus, for the data here, SMOTE would "manufacture" cases of early readmission, and the manufactured data would then be used in the analysis. NearMiss,[3] an example of undersampling, is the technique used in this work. NearMiss undersamples the majority class by removing patients not readmitted early in an effort to balance the data.

Unfortunately, although these types of methods have had reported success in the literature, as well as in this analysis, there is no theoretical justification for them that we are aware of. Most importantly, in subsampling the data by making use of clinical information, the resulting estimated values for probability will not be valid in general. Thus, the reported success of these methods is based primarily on their empirical performance in terms of classification (identifying which patients might be readmitted early), not on

From the [a]Division of Biostatistics, University of Miami, Miami, Fla; and [b]Department of Data Science, University of Mississippi Medical Center, Jackson, Miss.

Disclosures: The authors reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

Received for publication June 5, 2020; revisions received June 5, 2020; accepted for publication June 5, 2020; available ahead of print June 29, 2020.

Address for reprints: Hemant Ishwaran, PhD, University of Miami, Clinical Research Building Room 1058, 1120 NW 14th St, Miami, FL 33136 (E-mail: hemant.ishwaran@gmail.com).

J Thorac Cardiovasc Surg 2021;161:1940-1
0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2020.06.052

their ability to estimate probabilities (the probability a patient will be readmitted early). In our own experience with these methods, we have found that they can sometimes help improve classification; however, very delicate tuning and experience is required to do so.

There is another solution that provides a clearer path forward, however. This method is also based on subsampling the data but differs in a very important aspect, with the sampling done using only the value of the outcome and making no use of the associated clinical data. This type of sampling is called response sampling. In the ML literature, the most popular implementation is undersampling, in which the majority class is undersampled to match the frequency of the minority class. This is the technique used by balanced random forests (BRF), for example.[4] This method has been used quite widely and has been found to generally produce good results.

The theoretical explanation for why BRF and response-based undersampling works was provided in a recent paper by O'Brien and Ishwaran,[5] who showed that response-based undersampling is theoretically equivalent to replacing the Bayes rule with a different decision rule known as the quantile classification rule. Rather than classifying patients on whether their probability is >0.5, the rule adjusts the value 0.5 to match the underlying prevalence. Doing so yields a procedure with the optimal property of simultaneously maximizing sensitivity and specificity.

In fact, there is no need to subsample at all. O'Brien and Ishwaran[5] showed that we need only replace the Bayes rule with the new quantile rule to yield a procedure with theoretically justified properties. Furthermore, by forgoing sampling, the resulting estimated probabilities remain valid. Thus, we obtain not only a good classifier, but also one with valid probability estimates.

O'Brien and Ishwaran[5] have developed the quantile classifier for use with random forests, a method referred to as RFQ. We note that RFQ is available for general public use through the "imbalanced" function in the randomForestSRC R-package.[6] It can be used for classification, production of estimated probabilities, and calculation of variable importance values. The latter allow researchers to quickly determine which clinical variables are important and provide estimates of their effect size in terms of prediction error.

In conclusion, the authors have tackled an important medical issue for esophageal cancer patients. They provide a detailed analysis of class-imbalanced data, a setting common in medical studies but often misunderstood or overlooked. As the authors have found, such settings can be nuanced and difficult to analyze and require careful use of ML methods. Finally, we would like to thank the editors of the *Journal* for providing us with the opportunity to comment on this work.

References

1. Bolourani S, Tayebi MA, Diao L, Wang P, Patel V, Manetta F, et al. Using machine learning to predict early readmission following esophagectomy. *J Thorac Cardiovasc Surg*. 2021;161:1926-39.e8.
2. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-57.
3. Bao L, Juan C, Li J, Zhang Y. Boosted near-miss under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets. *Neurocomputing*. 2016;172:198-206.
4. Chen C, Liaw A, Breiman L. Using Random Forest to Learn Imbalanced Data: Technical Report. University of California Berkeley, Department of Statistics; 2004.
5. O'Brien R, Ishwaran H. A random forests quantile classifier for class imbalanced data. *Pattern Recognit*. 2019;90:232-49.
6. Ishwaran H, Kogalur UB. Fast unified random forests for survival, regression, and classification (RF-SRC). R package version 2.9.3; 2020. Available at: https://cran.r-project.org/package=randomForestSRC. Accessed January 21, 2020.