Review of: "Chronic disease treatment default prediction with random sampling optimization."

Sally McClean[1]

[1] Ulster University

Potential competing interests: No potential competing interests to declare.

Class imbalance is often a major problem for classification algorithms, due the majority class overwhelming the minority one, which is often the class of most interest. Previously this problem has been addressed by random over-sampling of the minority class, thus evenly balancing the samples.

The current study investigates strategies for random sampling optimization using five classification-based algorithms, namely: extreme gradient boosting, gradient boosting, random forest, support vector machines and logistic regression, to evaluate predictive performance for a real-world healthcare dataset of patients suffering from hypertension with comorbidities.

The problem of class imbalance has been well known for over 20 years and various solutions have been proposed, such as so-called “bagging” (typically involving sampling as discussed), and “boosting” (supplementing the minority class with typical additional data). In both cases the aim is to balance the classes thus improving prediction accuracy, model generalization and model behavior, as the authors say. However, I don’t think it is correct to say that this problem has received “little attention”. I suppose it would be useful to discuss some early work on the topic, citing authors such as Brieiman or Efron and Hastie who provide more recent summaries.

The results, as expected, demonstrate the need to balance the data in order to get a useful result and the importance of choosing meaningful and useful performance criteria.

Chen, C., Liaw, A. and Breiman, L., 2004. Using random forest to learn imbalanced data.*University of California, Berkeley*, *110*(1-12), p.24.

Efron, B. and Hastie, T., 2021. *Computer age statistical inference, student edition: algorithms, evidence, and data science* (Vol. 6). Cambridge University Press.