Title: Ensemble learning in class imbalanced data

Emmanuel Ogundimu

General description

In predictions and classification problems, class imbalance occurs when the class of interest (positive or minority class) is relatively rare compared to the other classes (negative or majority classes). As a result, classifiers can be heavily biased toward the majority class. Consider, for example, a credit risk data set with 10 defaulters (minority class) and 990 non-defaulters (majority class). A classifier that correctly classifies all the 990 non-defaulters while misclassifying the 10-defaulters will have 99% accuracy. This classifier is not useful.

A key strategy for handling class imbalanced problem is data balancing techniques. The strategy includes over-sampling (duplication of minority class examples), under-sampling (discarding majority class examples) and hybrids of both methods. The most popular hybrid method is called SMOTE or the Synthetic Minority Over-sampling Technique. Its basic idea is to artificially synthesize new minority class instances and add them to original data. For a sample $\boldsymbol{X}$ in minority class, Euclidean distance from samples in minority class is calculated to obtain its k-nearest neighbours. Then, a sample $\boldsymbol{X}^*$ from its k-nearest neighbours is randomly selected. Finally, new minority sample $\boldsymbol{X_{new}}$ is generated as:

$$\boldsymbol{X_{new}} = \boldsymbol{X} + u.(\boldsymbol{X}^* - \boldsymbol{X}); \quad u \sim U(0,1).$$

Ensemble methods have been developed to handle class-imbalanced data. These are techniques for combining two or more algorithms of similar or dissimilar types called based learners. SMOTE sampling strategies have also been integrated with ensemble learning techniques in form of “Bagging” (SMOTEBagging) and “Boosting” (SMOTEBoost) by the data science community.

This project will explore SMOTE and Ensemble methods for class imbalanced data. The topic area is wide and there are novel research areas that can be explored. Some possibilities include, but are not limited to:

- Evaluation of the impact of the degree of class imbalance on ensemble methods
- Development of new ensemble methods based on variants of SMOTE techniques
- Bagging and Boosting based on statistical models for class imbalanced data (e.g. Firth penalized logistic regression)
- Evaluation of techniques for quantifying overly optimistic predictions in imbalanced data settings (e.g. k-fold cross-validation, and Bootstrap and its variants).

References

(1) Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). SMOTE: Synthetic Minority OverSampling Technique. J. Artif. Intell. Res. 16, 321–357.

(2) Fernandez, A., S. Garcia, F. Herrera, and N. V. Chawla (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. J Artif Intell Res 61, 863–905.

(3) Firth, D. (1993) Bias reduction of maximum likelihood estimates. Biometrika, 80, 27–38.