Stat 98: Literature Review

Masaoud Haidar

April 2022

1 Citation

Blagus, Rok, and Lusa, Lara. "SMOTE for High-Dimensional Class-Imbalanced Data," *BMC Bioinformatics*, no. 14 (2013).

2 Review

This paper aims to find the properties of SMOTE (Synthetic Minority Oversampling Technique) and how it affects classification models, mainly in a high dimensional settings. To do so, it takes three approaches. First, it derives theoretical properties of SMOTE-augmented datasets and discusses how they would affect classification models. Second, it runs tests on simulated data to compare the performance of SMOTE on different settings and on different classification models. The classification models discussed are CART, KNN, DLDA, DQDA, Random Forest, SVM, PAM, and penalized logistical regression (Lasso and Ridge). Finally, the paper tests SMOTE on three real-world datasets to confirm the results of the simulation and provide real world applications.

The paper finds that SMOTE doesn't change the in-class mean and it reduces the in-class variance of the data. This translates to almost no effect of SMOTE on DLDA and DQDA and results in underestimated p-values for other models. The paper derives that SMOTE reduces the Euclidean distance between the test data and the augmented minority samples. This translates to a high effect of SMOTE when used with KNN or PAM and when used with variable selection. Finally, for all the other models, the paper finds that SMOTE has positive effect in a low dimension setting, but a limited to no-effect effect in a high dimensional setting.

The paper does a great job deriving theoretical properties and using them to explain the results of the simulation. What it lacks is an explanation of why these properties and results are different between low and high dimensional data. It seemed that the theoretical properties should still apply either way, but the experimental results are very different between low and high dimension settings.

The theoretical properties in the paper offered a point of reference for our project in terms of what to test for. But instead of looking at in-class mean and variance, our project looked at general mean and variance and found that they both significantly change because of SMOTE. And instead of looking at p-values of coefficients, we looked at the bias of the coefficients and found a significant bias in them, which means that the coefficients of the models trained on SMOTE-augmented data shouldn't be used for interpretation.

The experimental results of the paper weren't as relevant to our project because they focus on comparing low and high dimensional settings, where our projects only studies a low dimensional setting (with two predictors).