Title: A Statistically Robust Approach to Machine Learning for Model Development and Validation Under the Constraint of Small Datasets

Authors: Cherub Kim*; Zhen Zhang*

*Johns Hopkins University, School of Medicine, Department of Pathology

Background

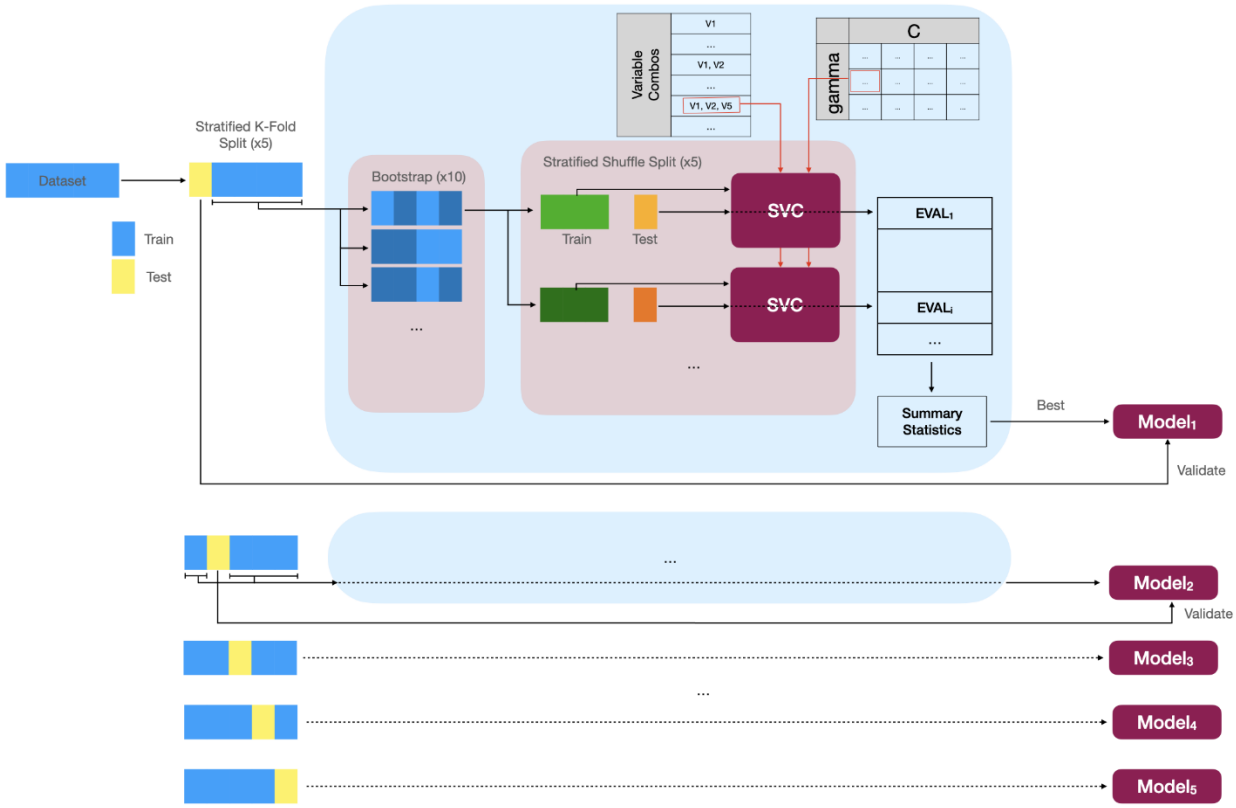
Despite the existence of a large amount of high-quality laboratory data of clinical analytes, most pathology datasets available for machine learning are annotated manually, which limits the sample size of available data for model development and validation. Machine learning models developed on small datasets may have high variance with poor generalizability in independent validation. We present a statistically principled workflow for machine learning model development for small datasets problems.

Technology

Grid-search using a combination of cross-validation, bootstrapping, and multiple rounds of stratified splitting were used in combination to select support vector machine parameters and analyte combinations. Scikit-learn v. 0.24.0, Pandas v. 1.2.0, Matplotlib v. 3.3.3, JupyterLab v. 3.0.0 libraries were used in Python3.8 to train and evaluate the models.

Methods

We perform a full grid-search over the range of model parameters and analyte combinations. At each grid point, the development set is bootstrapped ten times. Each bootstrap is stratified split five times into train and test sets. The support vector machine and analyte combination specified by each grid point is trained and tested with each of these datasets and the median and standard deviation are calculated as an overall score for each grid point. The top scoring parameter and variable combinations are used to train a five-member ensemble of support vector machines. These are evaluated with the initially left out validation set. This method was tested on an n = 81 prostate cancer marker dataset with 12 different cancer markers (including PHI) that is labeled as aggressive versus non-aggressive prostate cancer. The performance of our model was compared to that of PHI by area-under-curve, F1 and F2 score using stratified five-fold cross validation.



Results

Figure 1 depicts our parameter and variable selection process and result validation. Models developed using our method show good consistency (four of five splits) in outperforming PHI in F1 and F2 score.

Conclusions

We demonstrate a method for developing robust machine learning models given the small datasets that are common in pathology.