Research Week 2022

Machine learning for cancer subtyping: sampling effects on predictive accuracy and feature selection

Brian Karlberg, Jordan Lee, Chris Wong, Josh Stuart, Kyle Ellrott
karlberb@ohsu.edu
Oregon Health and Science University, Department of Biomedical Engineering, Knight Cancer Institute

Keywords

Machine learning, feature engineering, cancer subtypes, power law curve fitting

Abstract

Accurate classification of cancer subtypes based on genomic data is a key component of precision oncology. Subtyping allows for better alignment of therapeutics with the specific biology of an individual patient's tumor. Training predictive models for classification requires collection, sequencing, and labeling of patient samples which incurs significant cost and thus motivates an estimation of the minimum number of samples required to attain a particular classification accuracy. In this work, a minimum sample size estimation strategy for machine learning cancer subtype prediction is developed by combining a sub-sampling method for learning curve generation with an inverse power law curve fitting method. A systematic search of statistical distributions fit to these predictions for 15 cancer cohorts in The Cancer Genome Atlas (TCGA) with sufficient sample sizes provides a means to predict, with quantified error, the minimum number of additional samples required for 11 under-sampled TCGA cancer cohorts. Additionally, the effect on cancer genomic feature selection can be interrogated via these sub-sampling methods. Minimizing the number of biomarkers comprising predictive clinical panels is one application of this. Here, computational feature selection is conducted within the minimal-sample framework to identify parsimonious feature sets of cancer-type specific biomarkers via mutual exclusivity analysis.