Evaluating Study Replicability in Supervised Machine Learning with Epi-transcriptomic Data: Impact of Data Harmonization and Classifier Validation

In recent decades, study replicability has emerged as a critical concern in biomedical research, especially in the development of sample classifiers using molecular data like (epi-)transcriptomics. This concern stems partly from the well-recognized issue of systematic yet irreproducible artifacts pervasive in molecular data and the less-discussed tendency to aureate data harmonization (including data normalization and batch-effect correction) in addressing these artifacts. However, data harmonization may inadequately remove artifacts or excessively smooth data variability. These challenges are further compounded by the widespread use of cross-validation and split-sample validation to assess classifier accuracy, potentially leading to data leakage during random data splits and resulting in over-optimistic error estimation.

Emerging supervised machine learning techniques hold promise for deriving more accurate classifiers from molecular data. However, they are not immune to replicability concerns due to the same issues of pervasive artifacts, inadequate harmonization, and inappropriate validation. In this study, we conducted simulated experiments to investigate the impact of data harmonization methods and classifier validation approaches on the replicability and accuracy of machine learning using microRNA expression data collected with microarrays or sequencing.

Our simulations employed resampling strategies, utilizing paired microRNA datasets from the same set of tumor samples representing two tumor types, previously gathered at Memorial Sloan Kettering Cancer Center. Each pair consisted of one dataset collected with uniform handling and balanced design, and another without. In silico datasets were generated with varying levels of biological signals and magnitudes of data artifacts to train classifiers for tumor types. Specifically, they were harmonized using popular normalization methods or batch-effect correction methods, trained using machine learning techniques such as K-nearest neighbors, Random Forest, Support Vector Machine, and XGBoost, and then assessed for accuracy using cross-validation and split-sample validation. These accuracy estimates were compared against a gold standard derived from in silico datasets generated at similar signal levels and free of artifacts.

In this poster, we will present the results from the simulations regarding the sensitivity of machine learning to data harmonization and classifier validation. Furthermore, we will discuss the preferred choices of harmonization method, learning technique, and validation approach, based on considerations of both accuracy and replicability.