EXPLORE can achieve similar performance for prediction problems compared to other interpretable models on dataset with 50 covariates

Interpretable decision rules for patient-level prediction with EXPLORE

Background: EXPLORE (Exhaustive Procedure for LOgic–Rule Extraction) is an exhaustive search algorithm designed to find optimal decision rules. This algorithm has several features that make it attractive for patient-level prediction models. Prior work investigating the performance of EXPLORE on standard UCI datasets has shown promising results, but is limited as the studied prediction tasks are much simpler than real-world settings.

Result 1: On the full dataset LASSO, Random forest, and XGBoost using all candidate covariates had the best predictive performance across prediction tasks. However, models were very complex (221–5315 covariates).

Result 2: On the reduced dataset we observe there exists an interpretable model (3–10 covariates) with similar or even better performance than LASSO, Random forest, XGBoost (33–50 covariates) for 3/5 prediction tasks.

Methods

1. We investigated the performance (AUC) and complexity of prediction models developed using EXPLORE and other frequently used algorithms across five prediction tasks in the Dutch Integrated Primary Care Information (IPCI) database.

2. We created a reduced dataset with 50 covariates based on the highest (absolute) Pearson’s correlation coefficients.

3. Model complexity was measured as the total rule length (EXPLORE, RIPPER), number of nodes (Decision Tree), number of non-zero features (LASSO, Iterative Hard Thresholding), and number of features used in model (Random Forest, XGBoost).

Limitation: Univariate pre-variable selection was used to make EXPLORE computationally feasible. Future work needs to investigate into how to most effectively reduce data dimensionality for EXPLORE.