Learning under constraints with EXPLORE

PRESENTER: **Aniek Markus**

INTRO:

The rule induction algorithm EXPLORE has several features that make it attractive for patient-level prediction models:

- The resulting model is a (short) decision rule and can be considered interpretable
- It is possible to specify additional constraints, e.g. minimum specificity or mandatory features that should be included

METHODS:

1. We investigated the performance of EXPLORE in comparison to LASSO logistic regression and RandomForest.
2. As default setting for EXPLORE we use a maximum rule length of 3 and maximize accuracy.
3. We also investigate learning with EXPLORE under two types of constraints: minimum specificity 0.9 and selecting the best predictor in LASSO logistic regression as mandatory feature.

RESULTS:

1. The methods perform roughly similar on the same prediction problems, even though some prediction problems seem to be more difficult than others (see Figure 1).
2. RandomForest uses most features, followed by LASSO logistic regression, and EXPLORE with the lowest number for 11/12 prediction problems.
3. Imposing a minimum specificity of 0.9 leads to a slightly lower AUC.
4. Adding one mandatory feature led to a different model in 4/13 prediction problems. In these cases, the AUC decreased only marginally (0.005-0.016) and increased once (0.01).

Results on standard UCI datasets show that the rule induction algorithm EXPLORE can achieve similar performance as LASSO logistic regression and RandomForest, with substantially smaller models.

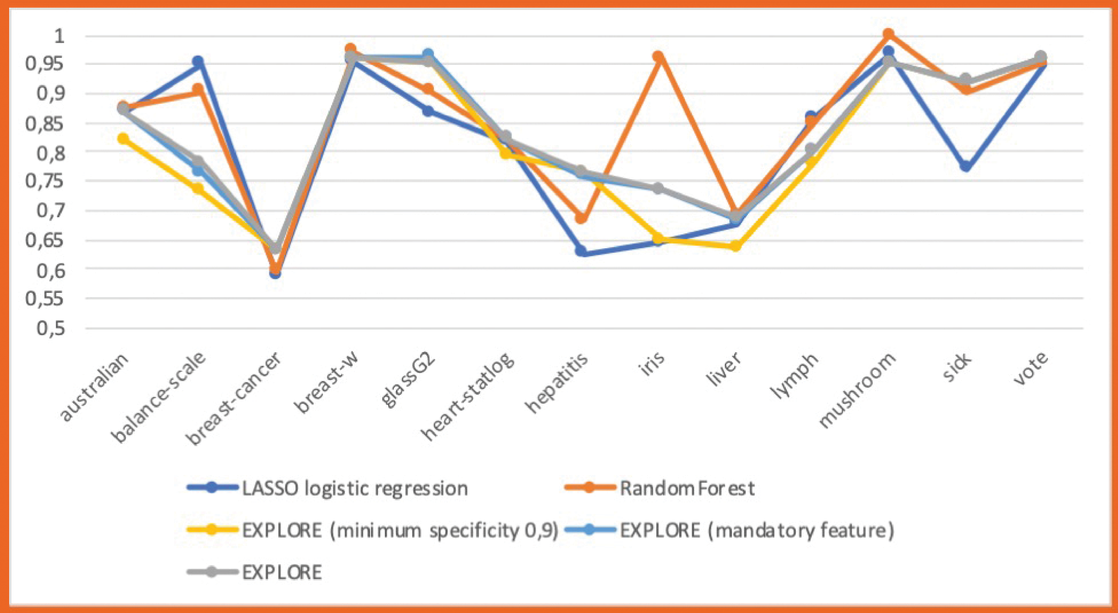

Figure 1: Comparison performance of AUC on standard UCI datasets between LASSO logistic regression, RandomForest, and EXPLORE (maximum rule length 3).

R package to run EXPLORE can be downloaded from GitHub

About EXPLORE:

1. EXPLORE (**Exhaustive Procedure for LOgic-Rule Extraction**) is an exhaustive search algorithm designed to find optimal decision rules.

2. EXPLORE generates decision rules of pre-specified length in **disjunctive normal form** (DNF). A formula in DNF is a disjunction of terms (OR), where the terms are conjunctions (AND) of literals, and the literals are feature-operator-value triples (A > a). An example of a DNF formula is (A > a AND B = b) OR C >= c, the resulting decision rule has the form: if (DNF formula) then class = 1 else class = 0.

3. To find the best decision rule EXPLORE performs an exhaustive search of all possible rules of pre-specified length using smart techniques to reduce the search space. For example, by reducing the number of values for each feature that need to be checked (**subsumption pruning**) and disregarding subspaces that cannot contain the optimal decision rule (**branch-and-bound**).

Aniek F. Markus, MSc,
Jan A. Kors, PhD,
Peter R. Rijnbeek, PhD

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.