Data Mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?

Wouter G. Touw, Jumamurat R. Bayjanov, Lex Overmars, Lennart Backus, Jos Boekhorst, Michiel Wels, and Sacha A. F. T. van Hijum

Radboud University and NIZO food research, the Netherlands

Briefings in Bioinformatics 2013

Presented by

Nawanol Theera-Ampornpunt

PURDUE UNIVERSITY

Background

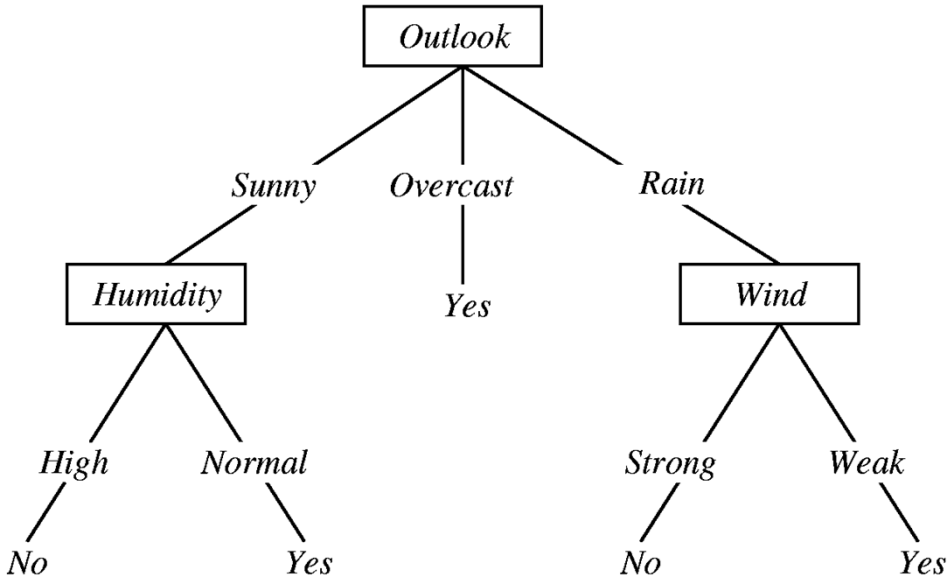
- Advancements in technology has allowed massive generation of 'Omics' data
  - genomics
  - proteomics
  - metabolomics
- Need tools to manage, visualize, and analyze these data
- Machine learning algorithms are *central* in knowledge extraction process
  - Typically as a classifier

Random Forest

- Random Forest (RF) has become popular
  - High prediction accuracy
  - Easy to interpret classifier
- Accuracy compares well to other algorithms
  - Support Vector Machine (SVM)
  - Artificial Neural Network (ANN)
  - Bayesian classifiers
  - Logistic Regression
- Life Science data sets have many more variables than samples
  - Curse of dimensionality

What is Random Forest?

- A collection of decision trees
  - Each decision tree created from a slightly modified version of the original training dataset
  - Final prediction = majority vote among trees

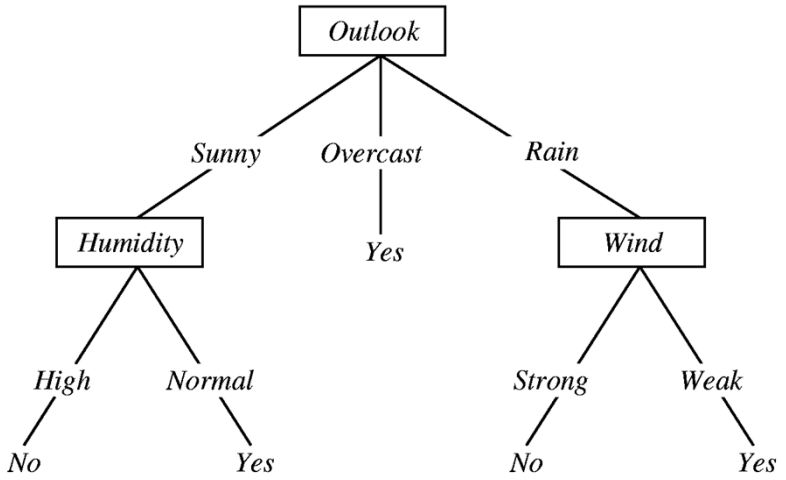


Generating Training Dataset

- Generate D' from D by
  - Sampling data points from D uniformly and **with replacement**
  - stop when |D'| = |D|
- This process is called Bootstrap Aggregating, also known as Bagging
  - Helps reduce variance
  - Avoid overfitting
- On average, D' will contain $1 - 1/e \approx 63.2\%$ of unique data points in D

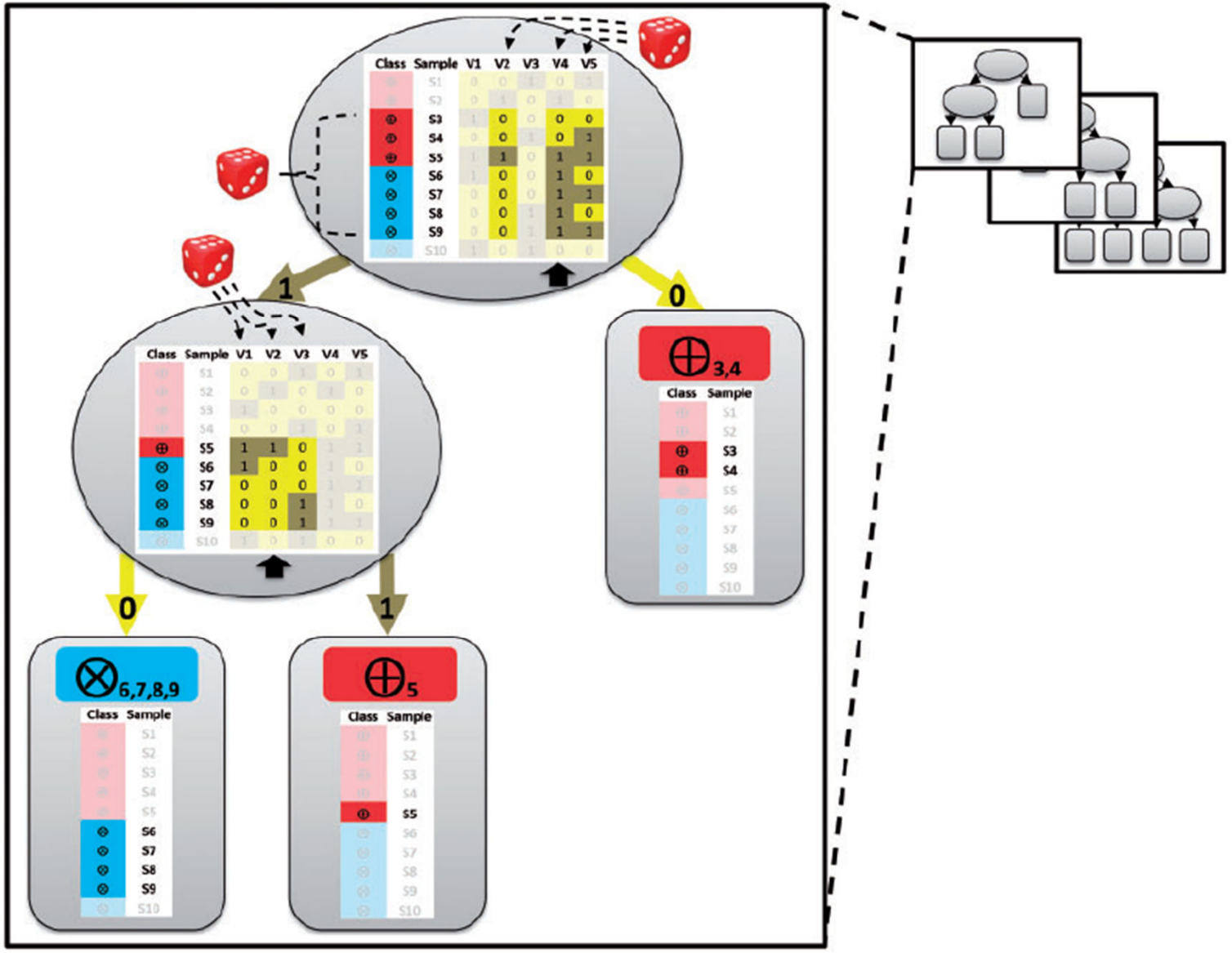
Decision Tree Training

- Start at root node
- Try each feature (and threshold) as a splitter
  - Compute score based on distributions of samples' label after the split
  - Use feature with highest score
  - Examples of scoring metric: entropy, Gini impurity
- Recursively build each subtree
  - until further split is not possible (full tree)
  - until gain / number of samples is below a threshold (pruning)



Random Forest Learning

- Differences from decision tree learning
  - Training dataset generated from bagging
  - At each split, only a random subset of features are considered
    - Typically $\sqrt{p}$ where p is number of features
    - Reduces correlation between decision trees
  - No pruning is used


Class Sample V1 V2 V3 V4 V5
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
1
0
⊕3,4
Class Sample
⊗6,7,8,9
⊕5

Benefits of Random Forest (1)

- Cross-validation is built-in
  - For each tree, out of bag samples can be used as test data
- Variable importance
  - Mean decrease in classification accuracy
    - Randomly permute values of the variable of test samples of each tree
    - Big decrease in accuracy = important variable
  - Gini impurity decrease
    - Sum of Gini impurity decrease across all nodes and trees where the variable is used for splitting

Benefits of Random Forest (2)

- Proximity score
  - Number of times two samples end up in the same leaf node of a tree
  - Outliers = samples with low proximity to all other samples from the *same* class
  - Subclasses (e.g., severe and mild subtypes of a disease) can be identified using proximity
- Conditional relationship between variables
  - If split on variable A is often followed by split on variable B, then A and B are conditionally dependent

RF Implementations

- 'randomForest' package in R
- Random Jungle framework
  - Fastest implementation of RF
  - Allows parallel computation
- Willows package
- WEKA workbench
  - Easy pre-processing
  - Easy comparison between algorithms

Conclusion

- RF is widely used in the Life Sciences
  - Can be used for both regression and classification
- Can be used as a black box
  - Feature selection and parameter tuning may improve accuracy
- Allow extraction of additional knowledge from data
  - Conditional relations between variables
  - Proximity of samples
  - Importance of variables
  - Individual trees can be analyzed