Introduction to classification and regression trees, random forests and model-based recursive partitioning in R

Marjolein Fokkema

Department of Methods & Statistics
Institute of Psychology
Leiden University

Ensemble methods (bagging, random forests)

have become a popular and widely used tool in many scientific fields, e.g., in genetics and bioinformatics, because they are applicable in high dimensional problems with complex interactions

(cf.,e.g., Furlanello et al., 2003, Gunther et al., 2003, Svetnik et al., 2003, Cummings and Myers, 2004, Cummings and Segal, 2004, Guha and Jurs, 2003, Lunetta et al., 2004, Segal et al., 2004, Arun and Langmead, 2006, Bureau et al., 2005, Huang et al., 2005, Shih, 2005, Diaz-Uriarte and de Andrés, 2006, Qi et al., 2006, Ward et al., 2006)

Problems of CART (and single trees in general)

II. instability to small changes in the training sample

Problems of CART (and single trees in general)

II. instability to small changes in the training sample

- small changes in the training sample $\Rightarrow$
  different splitting variable and value selected for a node $\Rightarrow$
  different tree from there on

Problems of CART (and single trees in general)

II. instability to small changes in the training sample

- small changes in the training sample $\Rightarrow$
  different splitting variable and value selected for a node $\Rightarrow$
  different tree from there on
  - different kinds of instability: two trees may look very different, but identify very similar subgroups and generate very similar predictions for new observations

Problems of CART (and single trees in general)

II. instability to small changes in the training sample

- small changes in the training sample $\Rightarrow$ different splitting variable and value selected for a node $\Rightarrow$ different tree from there on
  - different kinds of instability: two trees may look very different, but identify very similar subgroups and generate very similar predictions for new observations
  - extent of instability depends on characteristics of the data (e.g., signal/noise ratio, correlations between predictor variables)

Solution I: Evaluating stability

- Draw random samples from the training data and refit tree
- Assess stability of variables and values selected for splitting
- Philipp et al. (2016): Stability assessment of tree-based learners
- Implemented in function `stabletree` in package `stablelearner`

Solution II: Ensembling trees

main idea: utilize instability of individual trees

Solution II: Ensembling trees

main idea: utilize instability of individual trees

- grow a large number of trees on modified versions of the data (e.g., random samples, ...)

Solution II: Ensembling trees

main idea: utilize instability of individual trees

- grow a large number of trees on modified versions of the data (e.g., random samples, ...)
- generate predictions through averaging over or majority voting of the trees

Solution II: Ensembling trees

main idea: utilize instability of individual trees

- grow a large number of trees on modified versions of the data (e.g., random samples, ...)
- generate predictions through averaging over or majority voting of the trees
- E.g., bagging (Breiman, 1996a, 1998), random forests (Breiman, 2001), boosted tree ensembles (e.g., Breiman, 1997)

Ensemble learning

Motivation:

Can we improve the accuracy of a set of simple trees (weak learners) by combining them into an ensemble (a strong learner)?

Yes, we can!

- A weak learner is a method that does better than random guessing
- The predictive accuracy of the ensemble is better than any of its constituent members
- Can be applied to other learners than trees
- Works best for unstable methods

Decorrelating trees I: Bagging

take bootstrap samples from the original data

average over trees

bootstrap aggregating
(alternatively, we may use subsample aggregating)

Bootstrap sampling

from the original sample of size N draw a bootstrap sample of size N with replacement

$\Rightarrow$ some observations appear twice or more, some not at all

Bootstrap sampling

Bootstrap sampling

probability for one observation <u>not</u> to be drawn in one draw

$$1 - \frac{1}{n}$$

probability for one observation <u>not</u> to be drawn in any one of the n bootstrap draws

$$\lim_{n \to \infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} \approx 0.368 = 1 - 0.632$$

$\Rightarrow$ approx. 63.2% of all observations are in the bootstrap sample

$\Rightarrow$ approx. 36.8% of all observations are "out of bag"

Bagging

main idea: utilize instability of individual trees

Bagging

main idea: utilize instability of individual trees

- use large, unpruned trees (each tree has low bias but high variance)

Bagging

main idea: utilize instability of individual trees

- use large, unpruned trees (each tree has low bias but high variance)
- reduce variance by ensembling predictions

Bagging

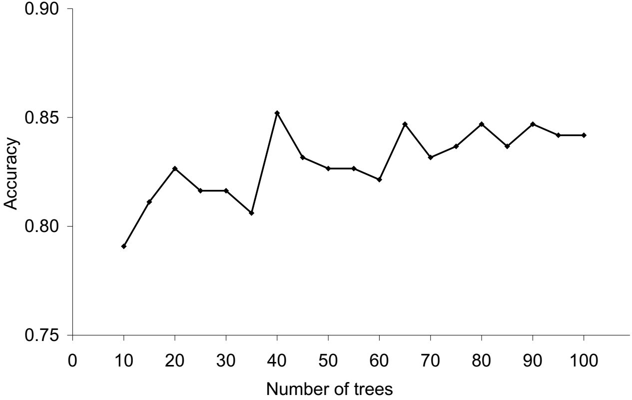
main idea: utilize instability of individual trees

- use large, unpruned trees (each tree has low bias but high variance)
- reduce variance by ensembling predictions

$\Rightarrow$ averaging increases prediction accuracy (Breiman, 1996a, 1998)

Bagging

prediction accuracy increases with the number of trees

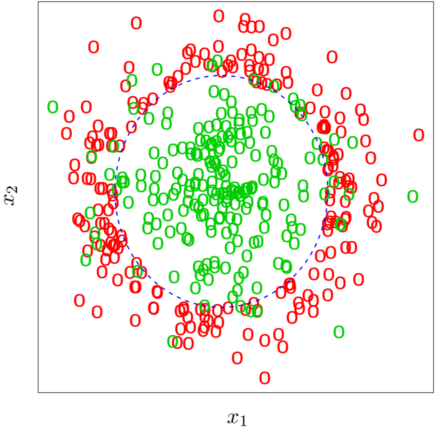


Bagging

decision boundaries are smoothed

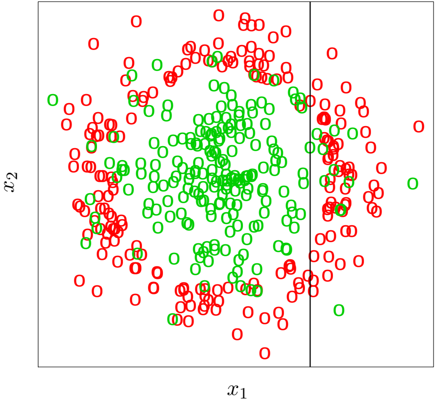
(thanks to Ji Zhu, University of Michigan, for the following graphical illustration)

Example: Nested Spheres

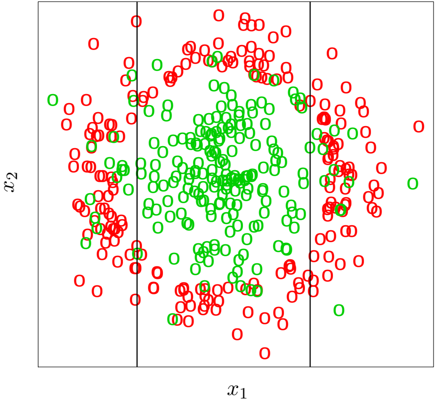


- Green class: two independent standard normal inputs X_1, X_2
- Red class: conditioned on $X_1^2 + X_2^2 \geq 4.6$

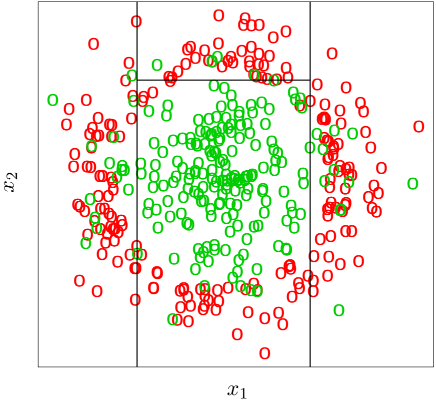
Classification Tree in Action: 1



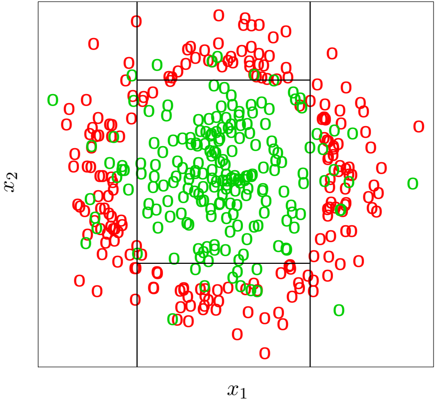
Classification Tree in Action: 2



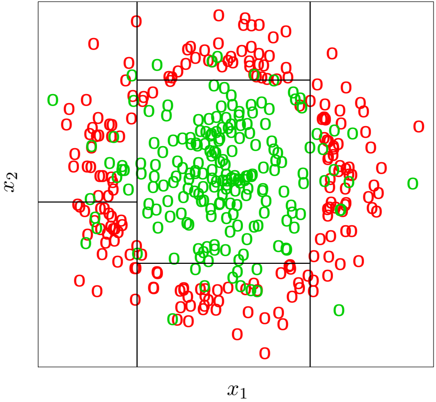
Classification Tree in Action: 3



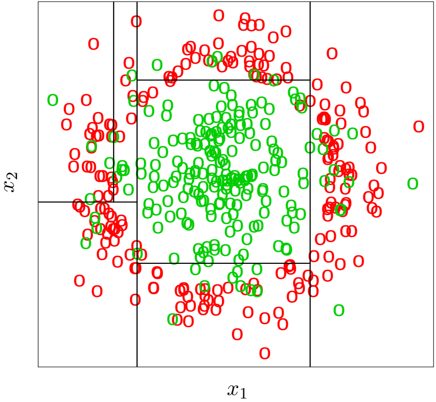
Classification Tree in Action: 4



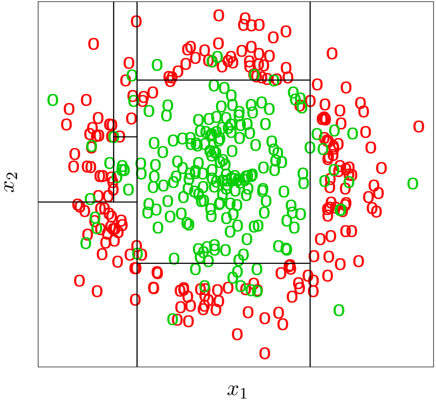
Classification Tree in Action: 5



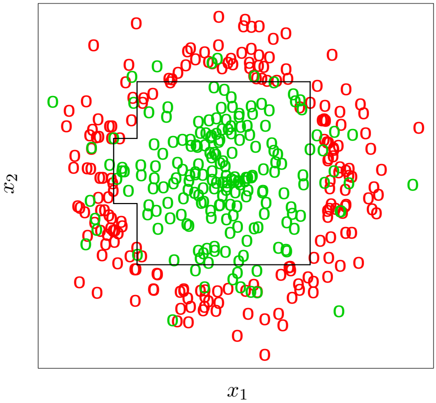
Classification Tree in Action: 6



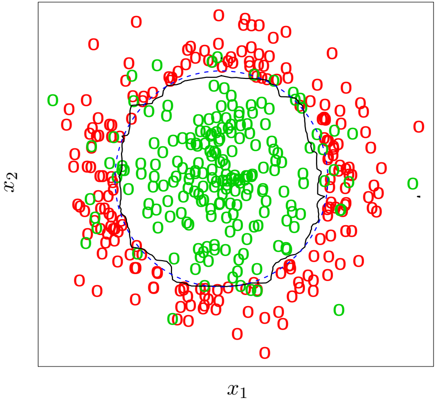
Classification Tree in Action: 7



Decision Boundary: Tree



Decision Boundary: Bagging



Bagging averages many trees, and produces more flexible decision boundaries.

Decorrelating trees II: Random forests

main idea: even more variation in individual trees

Decorrelating trees II: Random forests

main idea: even more variation in individual trees

- bootstrap sampling observations: vary "rows" of data set

Decorrelating trees II: Random forests

main idea: even more variation in individual trees

- bootstrap sampling observations: vary “rows” of data set
- add: sampling “columns” of data set (randomly draw variables)
  - a random subset of *mtry* variables is used for selecting each split in each tree

Decorrelating trees II: Random forests

main idea: even more variation in individual trees

- bootstrap sampling observations: vary “rows” of data set
- add: sampling “columns” of data set (randomly draw variables)
  - a random subset of *mtry* variables is used for selecting each split in each tree

$\Rightarrow$ individual trees look even more different

Decorrelating trees II: Random forests

main idea: even more variation in individual trees

- bootstrap sampling observations: vary "rows" of data set
- add: sampling "columns" of data set (randomly draw variables)
    - a random subset of *mtry* variables is used for selecting each split in each tree

$\Rightarrow$ individual trees look even more different

$\Rightarrow$ prediction accuracy is even higher (Breiman, 2001)

Advanced reading

- why bagging works: Bühlmann and Yu (2002)
  bagging smoothes hard decisions $\Rightarrow$ reduces variance
  (ugly asymptotics for trees as base learners)
- why random forests work: Lin and Jeon (2006)
  random forests can be viewed as adaptively weighted k-NN
  with terminal node size determining size of neighborhood

Bagging and random forests - tuning parameters

- number of trees
  argument `ntree` (default: 500)
  - more is better (does not negatively affect predictive accuracy) especially with many potential predictor variables
- number of randomly preselected predictors for each split
  argument `mtry` (usually $\sqrt{p}$ for classification, for smaller number of predictor variables sometimes $p/3$ is suggested, in `cforest` default: 5)
  - `mtry` $= p$ is bagging
  - different values for `mtry` can affect performance and estimates of variable importance

Random forests - tuning parameters

- tree depth
  - in bagging and random forests trees are usually grown large without pruning
  - only the minimum number of observations per node is fixed
  - results of Lin and Jeon (2006) indicate that the depth / number of observations per node do affect performance
- sampling size and method: for bootstrap sampling (default: N) and subsampling (default: $.632 \cdot N$)
  - subsampling gives better estimates of variable importance, see below

Decorrelating trees III: Boosting

main idea: fit trees on modified versions of the outcome variable Y

- can be used in addition to random sampling of rows and columns of X
- boosting is performed sequentially (in bagging and random forests, trees can be fit simultaneously)

Boosting

- At each stage $1 \leq m \leq M$ we have an imperfect model $F_(m-1)$
- we aim to improve the model with an estimator h_m, such that $F_m(X) = F_{m-1}(X) + h(X)$ provides a better model
  - How to find $h(X)$?
  - the perfect model would be $F_m(X) = F_{m-1}(X) + h(X) = y$
  - or, equivalently, $h(X) = y - F_{m-1}(X)$
- Therefore, we fit the model (the tree) at stage m to the residual $(y - F_{m-1}(X))$ instead of y
- Generally, a learning rate $0 < \nu < 1$ is applied in updating the ensemble:
  - $F_m(X) = F_{m-1}(X) + \nu \cdot h_m(X)$

Boosting

Requires more parameter tuning than bagging and random forests.

In noisy data, bagging and random forests are often more robust than boosting (Kotsiantis, 2011, e.g.,).

In R: function `gbm()` from package `gbm`

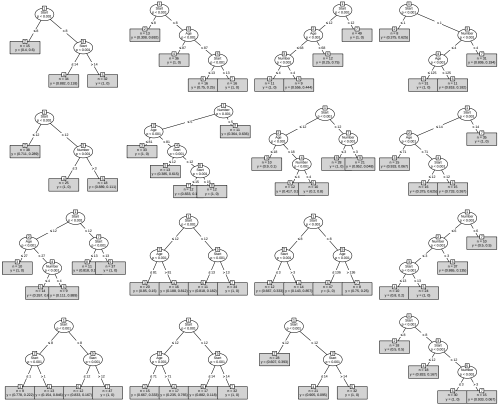
Boosting - tuning parameters

- `n.trees =`
  Increasing the number of trees may overfit the data with boosting. Optimal number of trees depends on learning rate and best determined by k-fold CV.
- `interaction.depth =`
  Typically, 4 through 8 terminal nodes work well, results are fairly insensitive to the exact choice (e.g., Hastie et al., 2009).
- `shrinkage =`
  a.k.a. learning rate ν. Typically, small values (e.g., $.001 \leq \nu < .01$) perform well (e.g., Efron et al., 2004, Bühlmann and Yu, 2003). Like number of trees, best determined by CV.

Note: list is non-exhaustive.

Interpretation

interpretation of predictor variables?



Measuring variable importance

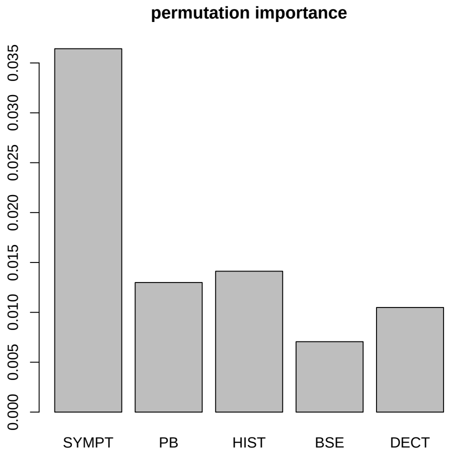
Measuring variable importance

- Gini importance
  mean Gini gain produced by X_j over all trees

Measuring variable importance

- Gini importance
  mean Gini gain produced by X_j over all trees
- permutation importance
  mean decrease in classification accuracy after permuting X_j over all trees
  - informative variables produce a systematic decrease in accuracy when permuted
  - uninformative variables produce a random decrease or increase in accuracy when permuted

Measuring variable importance



Measuring variable importance

Problems:

- Gini importance
  biased estimation of Gini gain in each tree $\Rightarrow$ Gini importance is biased in favor of continuous variables and variables with many categories

Strobl et al. (2007)

Measuring variable importance

Problems:

- Gini importance
  biased estimation of Gini gain in each tree $\Rightarrow$ Gini importance is biased in favor of continuous variables and variables with many categories
- permutation importance
  even if individual trees are unbiased, as in function `cforest`
  - bootstrap sampling affects variance of variable importance
  - variable importance of variables with many categories may be over/underestimated

  $\Rightarrow$ subsampling without replacement is used by default

Strobl et al. (2007)

Conditional permutation importance

spurious correlation between shoe size and reading skills in school-children

```
> mycf <- cforest(score ~ ., data = readingSkills,
+                     control = cforest_unbiased(mtry = 2))
> varimp(mycf)
nativeSpeaker           age      shoeSize
     12.62926      74.89542      20.01108
> varimp(mycf, conditional = TRUE)
nativeSpeaker           age      shoeSize
    11.808192     46.995336      2.092454
```

Strobl et al. (2008)

Measuring variable importance

Choice of hyperparameters:

- results are more stable when `ntree` is high
- results can vary for different `mtry`, especially in the case of correlated predictors
- subsampling size down to $.5 \cdot N$ (Buja and Stuetzle, 2006)

Random forests and bagging

pros	cons

Random forests and bagging

pros	cons
nonparametric approach	heuristic
easy to grasp	

Random forests and bagging

pros	cons
nonparametric approach	heuristic
easy to grasp	
applicable in	
small N large p problems	
with many irrelevant variables	

Random forests and bagging

pros	cons
nonparametric approach easy to grasp	heuristic
applicable in small N large p problems with many irrelevant variables	
applicable when predictors are correlated	

Random forests and bagging

pros	cons
nonparametric approach easy to grasp	heuristic
applicable in small N large p problems with many irrelevant variables	
applicable when predictors are correlated	
applicable to predictors of different types	selection can be biased

Random forests and bagging

pros	cons
nonparametric approach easy to grasp	heuristic
applicable in small N large p problems with many irrelevant variables	
applicable when predictors are correlated	
applicable to predictors of different types	

Random forests and bagging - continued

pros	cons
immune to outliers in predictors	

Random forests and bagging - continued

pros	cons
immune to outliers in predictors	
invariant under monotone transformations	

Random forests and bagging - continued

pros	cons
immune to outliers in predictors	
invariant under monotone transformations	
variable scales irrelevant	

Random forests and bagging - continued

pros	cons
immune to outliers in predictors	
invariant under monotone transformations	
variable scales irrelevant	
few hyperparameters work well "off-the-shelve"	

Random forests and bagging - continued

pros	cons
immune to outliers in predictors	
invariant under monotone transformations	
variable scales irrelevant	
few hyperparameters work well “off-the-shelve”	
variable importance	merely descriptive importance can be biased

Random forests and bagging - continued

pros	cons
immune to outliers in predictors	
invariant under monotone transformations	
variable scales irrelevant	
few hyperparameters work well "off-the-shelve"	
variable importance	merely descriptive

Random forests and bagging - continued

pros	cons
immune to outliers in predictors	
invariant under monotone transformations	
variable scales irrelevant	
few hyperparameters work well "off-the-shelve"	
variable importance	merely descriptive
out of bag error estimates	

Random forests and bagging

each tree is grown on a training (bootstrap or sub-) sample

each tree brings its own test (out of bag; OOB) sample

$\Rightarrow$ OOB error estimates are not overly optimistic (e.g., Breiman, 1996b)

Highly recommended

- Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning.* Second edition. Springer, Berlin.
  - Especially chapters 1, 2, 9, 10, 15
  - Yay, free! http://statweb.stanford.edu/~tibs/ElemStatLearn/download.html
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.
  - More introductory version of Friedman et al. (2009)
  - Yay, free! http://www-bcf.usc.edu/~gareth/ISL/getbook.html
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323.
  - Note of first author: please do not interpret negative importances as a measure of statistical significance

References I

K. Arun and C. J. Langmead. Structure based chemical shift prediction using random forests non-linear regression. In T. Jiang, U.-C. Yang, Y.-P. P. Chen, and L. Wong, editors, *Proceedings of the Fourth Asia-Pacific Bioinformatics Conference, Taipei, Taiwan*, pages 317–326, 2006.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996a.

L. Breiman. Out-of-bag estimation. *Unpublished Technical Report, Statistics Department, University of California at Berkeley, CA, USA*, 1996b.

L. Breiman. Arcing the edge. Technical report, Technical Report 486, Statistics Department, University of California at Berkeley, 1997.

L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–849, 1998.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

P. Bühlmann and B. Yu. Analyzing bagging. *The Annals of Statistics*, 30:927–961, 2002.

P. Bühlmann and B. Yu. Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–339, 2003.

A. Buja and W. Stuetzle. Observations on bagging. *Statistica Sinica*, 16:323–351, 2006.

A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh. Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, 28(2):171–182, 2005.

M. P. Cummings and D. S. Myers. Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. *BMC Bioinformatics*, 5:132, 2004.

M. P. Cummings and M. R. Segal. Few amino acid positions in rpoB are associated with most of the rifampin resistance in Mycobacterium tuberculosis. *BMC Bioinformatics*, 5:137, 2004.

R. Diaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3, 2006.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2): 407–499, 2004.

References II

C. Furlanello, M. Neteler, S. Merler, S. Menegon, S. Fontanari, D. Donini, A. Rizzoli, and C. Chemini. GIS and the random forest predictor: Integration in R for Tick-Borne desease risk assessment. In F. Leisch K. Hornik and A. Zeileis, editors, *Proceedings of the 3rd international workshop on distributed statistical computing DSC*, 2003.

R. Guha and P. C. Jurs. Development of linear, ensemble, and nonlinear models for the prediction and interpretation of the biological activity of a set of PDGFR inhibitors. *Journal of Chemical Information and Computer Sciences*, 44(6):2179–2189, 2003.

E. C. Gunther, D. J. Stone, R. W. Gerwien, P. Bento, and M. P. Heyes. Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles *in vitro*. *Proceedings of the National Academy of Sciences*, 100(16):9608–9613, 2003.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2nd edition, 2009.

X. Huang, W. Pan, S. Grindle, X. Han, Y. Chen, S. J. Park, L. W. Miller, and J. Hall. A comparative study of discriminating human heart failure etiology using gene expression profiles. *BMC Bioinformatics*, 6:205, 2005.

S. Kotsiantis. Combining bagging, boosting, rotation forest and random subspace methods. *Artificial Intelligence Review*, 35(3):223–240, 2011.

Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

K. L. Lunetta, L. B. Hayward, J. Segal, and P. Van Eerdewegh. Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genetics*, 5:32, 2004.

M. Philipp, A. Zeileis, and C Strobl. A toolkit for stability assessment of tree-based learners. Technical report, Working Papers in Economics and Statistics, 2016.

Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63(3):490–500, 2006.

M. R. Segal, J. D. Barbour, and R. M. Grant. Relating HIV-1 sequence variation to replication capacity via trees and forests. *Statistical Applications in Genetics and Molecular Biology*, 3(1):2, 2004.

Y.-S. Shih. Tumor classification by tissue microarray profiling: Random forest clustering applied to renal cell carcinoma. *Modern Pathology*, 18:547–557, 2005.

References III

C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25, 2007.

C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307, 2008.

V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random Forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 2003.

M. M. Ward, S. Pajevic, J. Dreyfuss, and J. D. Malley. Short-term prediction of mortality in patients with systemic lupus erythematosus: Classification of outcomes using random forests. *Arthritis and Rheumatism*, 55(1):74–80, 2006.