Random Forests

Presentation

Patric Weber

Agenda

1. Definition
2. Analogy
3. Flowchart
4. Model tuning
5. Advantages & Disadvantages
6. Conclusions
7. Bibliography

Definition

Random Forests

- **Random Forests** builds lots of trees so that the correlation between trees gets smaller and the majority of the predicted classes wins.
- Ensemble method uses multiple learning models to extract more accurate results.
- For supervised and unsupervised learning
- **Panacea**
- **Bagging =** average noisy and unbiased model → low variance → ***combine learners!***

Panacea lat. = a remedy for all ills or difficulties

Definition

Random Forests Analogy

- **Analogy “Ask the audience”**
  - **Majority of vote is correct**
  - **Why**?
  - Individuals have different experience, different “**data**” to decide
  - Individuals have different knowledge , different “**variables**” to decide

Random Forests is a **hybrid** of the **Bagging** algorithm and the random subspace method **(RSM).**

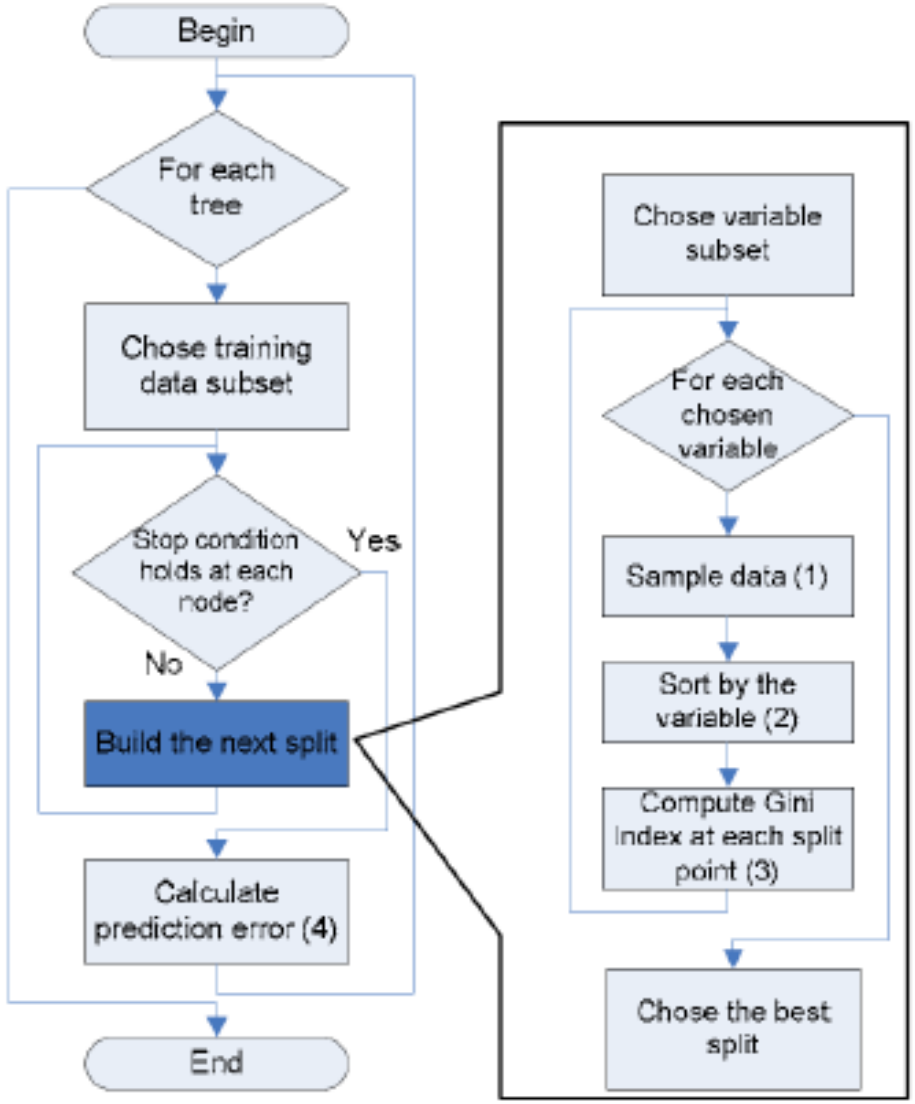
Experience = Bagging

Knowledge = Random Subspace Method

Random Forests

Flowchart

Forest



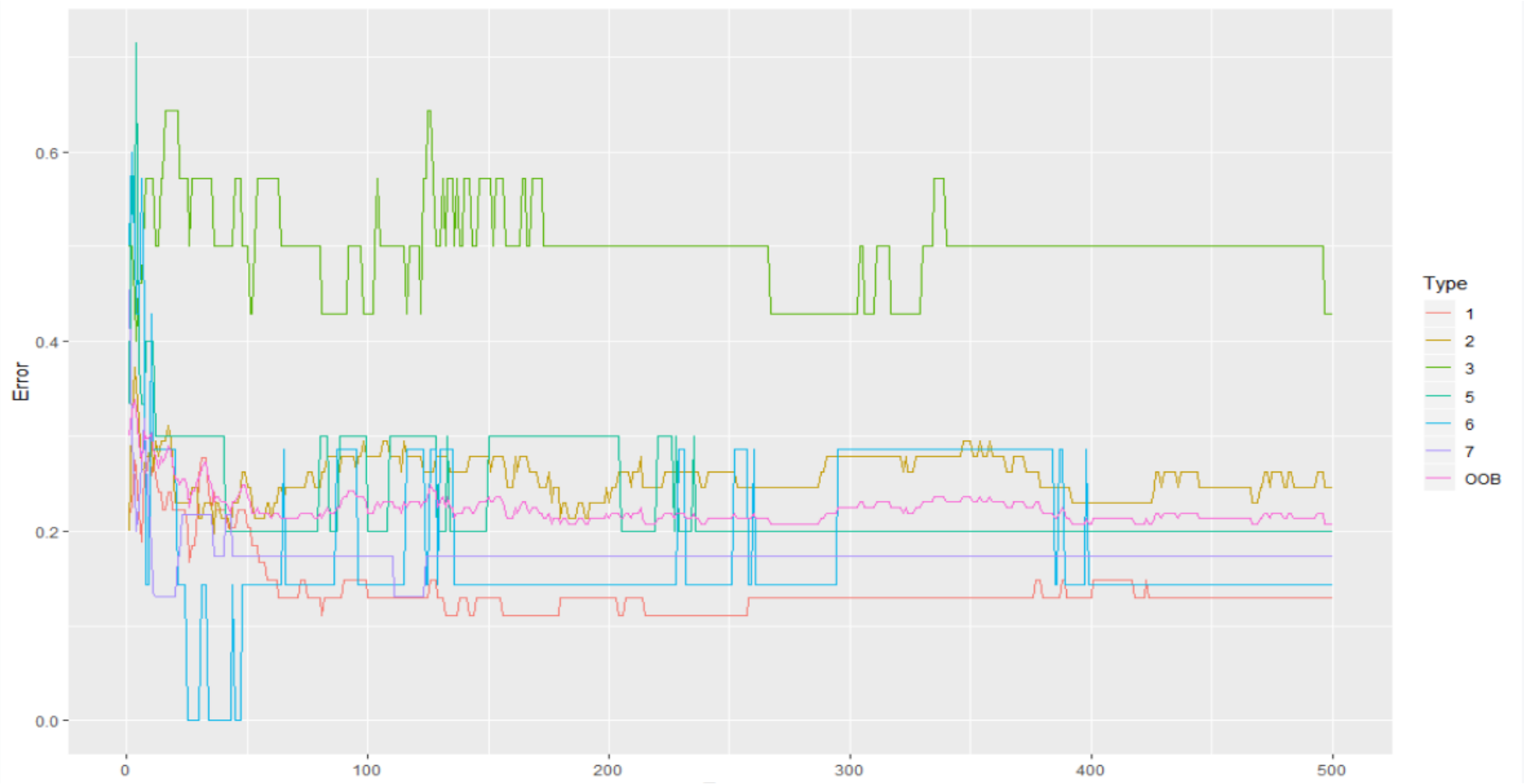
Random

Random Forests

Model Tuning

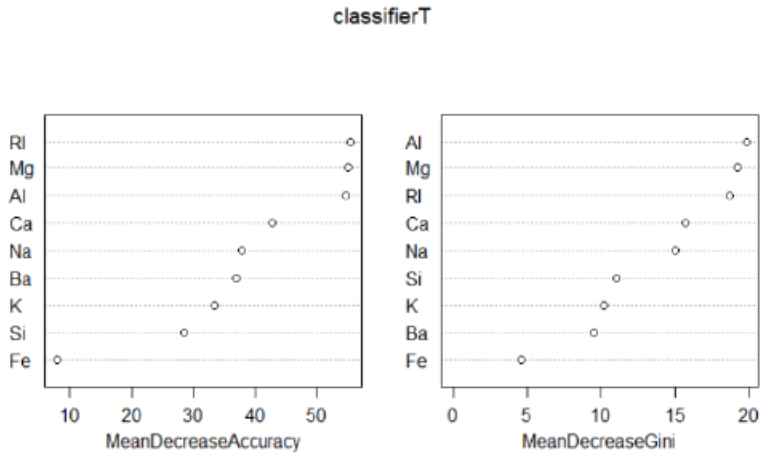
- **Trial and error approach**
- Splits are chosen according to a **purity measure**:
  - i.e. regression → squared error and Gini index or deviance for classification
- How to select ntrees?
  - Grow trees until the no longer out-of-bag-error decreases
- How to select mtry? *(variables for each split)*
  - Take default parameters, half of them and twice of them and pick the best!
  - Loop through to find out where the error is low.

Plot(ClassifierT)



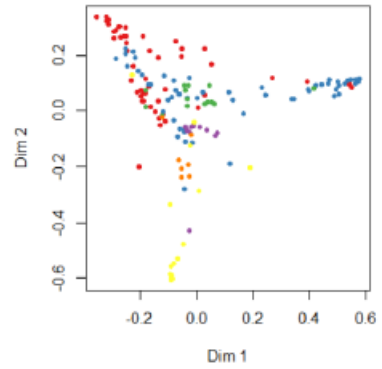
Advantages and Disadvantages

Random Forests



+ **Advantages**
  - Great out-of-the-box learner
  - Works on both regression and classification
  - Variable Importance

- **Disadvantages**
  - Less to control
  - Slow on large datasets
  - Quite accurate but not comparable to advanced boosting algorithms (AdaBoost)



Conclusions

Random Forests

- Easy to implement
- Robust against outliers, can handle data without preprocessing
- Best first hands-on approach algorithm
- Great data exploration features including Multidimensional scaling (MDS) plots

- *Algorithm: Random Forests → Breiman et. Al.*
- *Implementation R: "randomForest" → Andy Liaw from* MERCK

Bibliography

Great Resources

Jain Kunal, Ray Sunil, and Singh Simran. A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python), April 2016. URL https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/.

Saulo Pires de Oliveira. A very basic introduction to Random Forests using R | Oxford Protein Informatics Group, April 2017. URL https://www.blopig.com/blog/2017/04/a-very-basic-introduction-to-random-forests-using-r/.

Kailash Awati. A gentle introduction to random forests using R, September 2016. URL https://eight2late.wordpress.com/2016/09/20/a-gentle-introduction-to-random-forests-using-r/.

Raul Eulogio. Introduction to Random Forests, August 2017. URL https://www.datascience.com/resources/notebooks/random-forest-intro.

Peter Harrington. *Machine learning in action*. Manning Publications Co, Shelter Island, N.Y, 2012. ISBN 978-1-61729-018-3. OCLC: ocn746834657.

Anish Singh Walia. Random Forests in R, May 2018. URL https://datascienceplus.com/random-forests-in-r/.

Nina Zumel and John Mount. *Practical data science with R*. Manning Publications Co, Shelter Island, NY, 2014. ISBN 978-1-61729-156-2. OCLC: ocn862790245.

Claude Sammut and Geoffrey I. Webb, editors. *Encyclopedia of machine learning*. Springer, New York ; London, 2010. ISBN 978-0-387-30768-8 978-0-387-34558-1 978-0-387-30164-8. OCLC: ocn651073009.

Leo Breiman and Adele Cutler. Random Forests. URL https://www.stat.berkeley.edu/~breiman/RandomForests/.

Leo Breiman and Adele Cutler. Random Forests. URL https://www.stat.berkeley.edu/~breiman/RandomForests/.

Random Forests · UC Business Analytics R Programming Guide. URL https://uc-r.github.io/random_forests#idea.

Fred Hamprecht. Random Forest Feature Importance, 2012. URL https://www.youtube.com/watch?v=WE67TSz-a7s.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9 (1):307, July 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-307. URL https://doi.org/10.1186/1471-2105-9-307.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition, 2009. ISBN 978-0-387-84857-0. URL //www.springer.com/de/book/9780387848570.