Some Light in a Dark Forest

A closer look at tree ensembles

Simon URBANEK

Department of Computer Oriented Statistics and Data Analysis
University of Augsburg
Germany

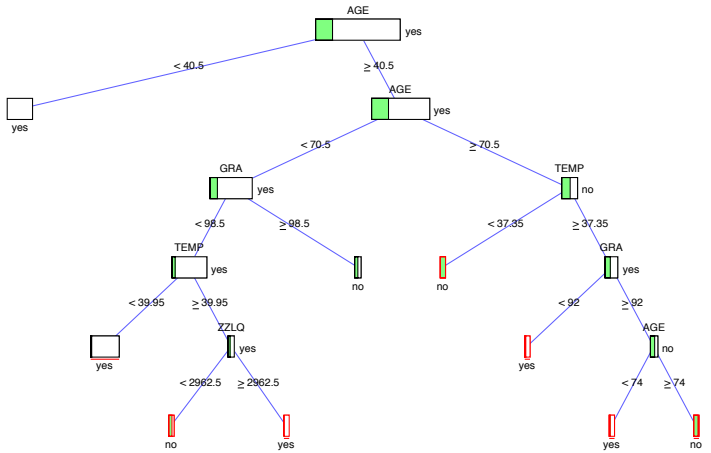
Overview

- **Tree model construction**
- **Ensembles and forests**
- **Comparing models**
- **Visualizing ensemble properties**
- **Future research**

Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Classification tree



Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Construction of tree models

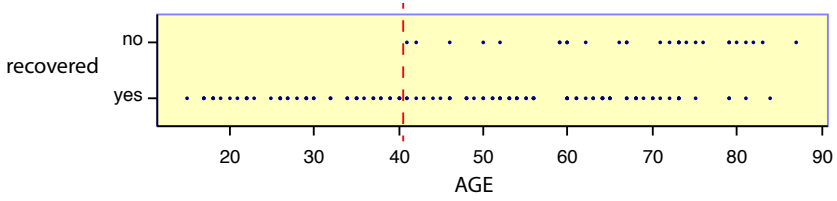
- **CART & derived algorithms**

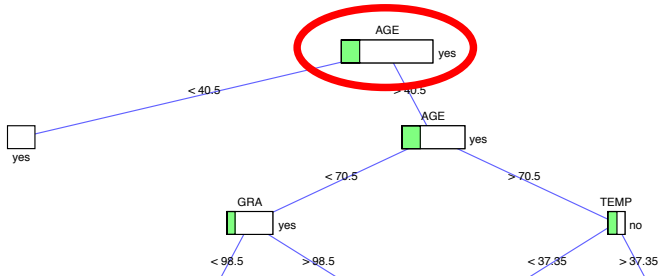
  local optimization in each node:

  - consider all variables
  - for each variable, consider all possible splits
  - select the "best" split according to an optimality criterion (Gini index, entropy, ...)

  repeat recursively for each child node

Split using a continuous variable

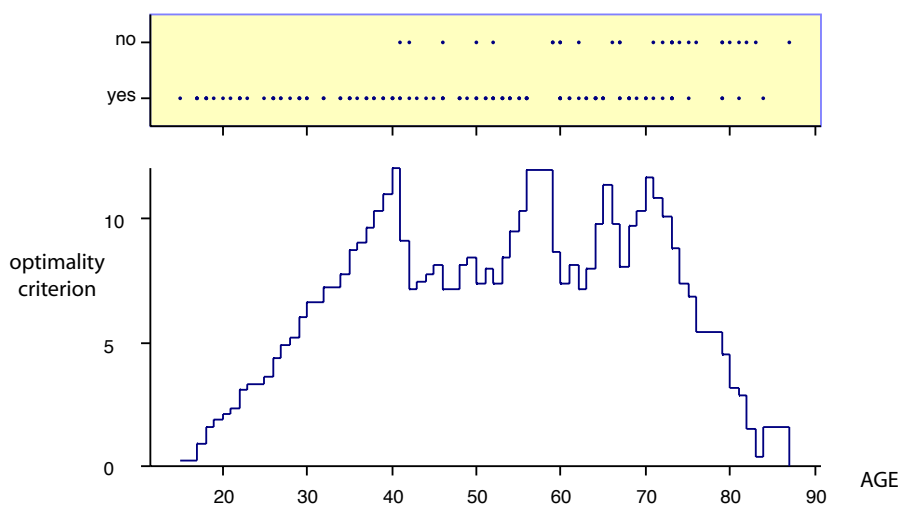




Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

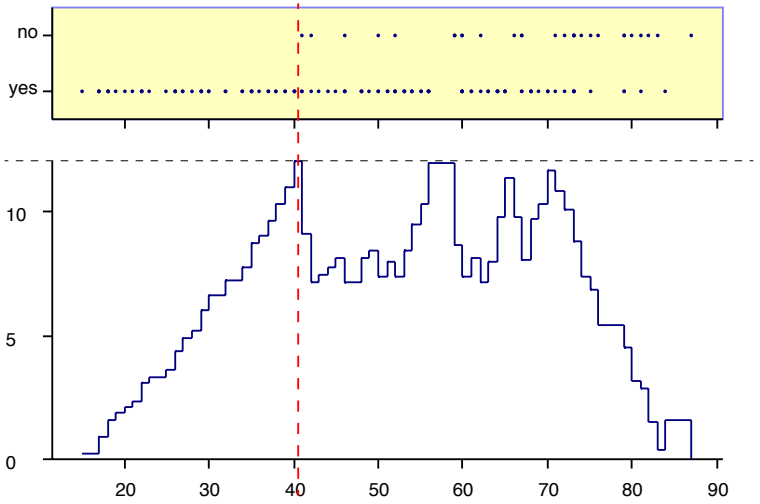
"Mountain plot"



Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

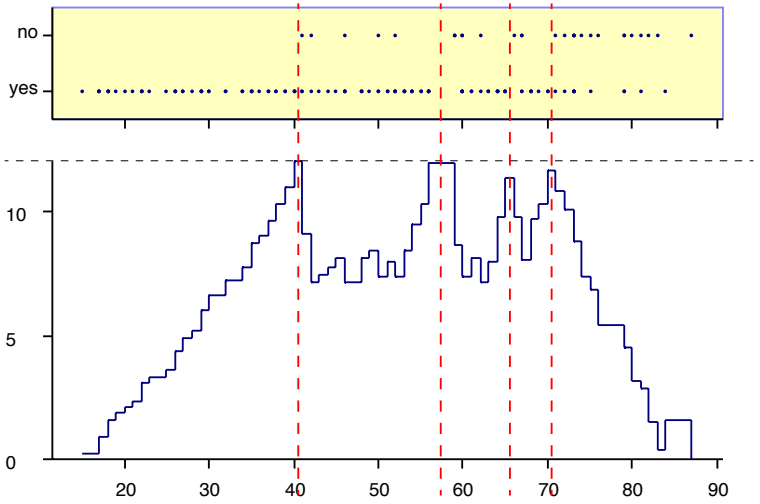
Optimal split



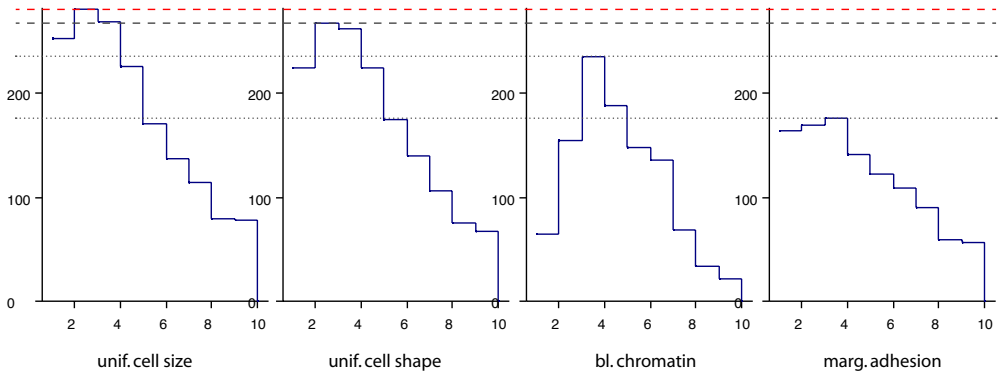
Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Competing splits using the same variable



Competing splits using different variables

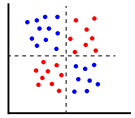


Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Individual trees

- **Advantage:**
  - easily interpretable
- **Possible issues:**
  - variable masking
  - weak splits
  - difficulties detecting locally inaccessible structures
  - credibility



Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Variations in model construction

- **Possible variations:**
  - different optimality criteria
  - use only subset of the data

    c.f. **bagging**

    changes in the data produce different models
  - use only subset of the covariates

    c.f. **random forests**

    effectively means to not insist on using the optimal split

Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Multiple models

- **Why multiple models?**
  - explore data structure
  - increase accuracy (aggregation or select the best)
- **Sources of multiple models**
  - interactive construction
  - bagging, boosting
  - random forests
- **Task: Compare models**

Conventional approach

- **Similarity measures**
  - matching coefficient
  - tree metric
  - partition metric
  - fit metrics
- **Difficulties**
  - resulting space is high-dimensional
  - MDS distance mapping of poor quality

Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Exploratory model analysis

- **Idea:**
  visualize different aspects of the models
  (variables and their contribution to the model,
  model structure, prediction, residuals, ...)

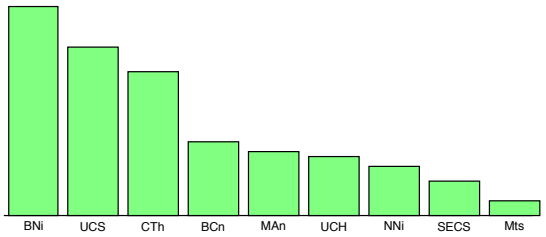
  learn more on the underlying data structure
  by looking at the behavior of the models

Workshop
Ensemble-Methods
Erlangen 2004

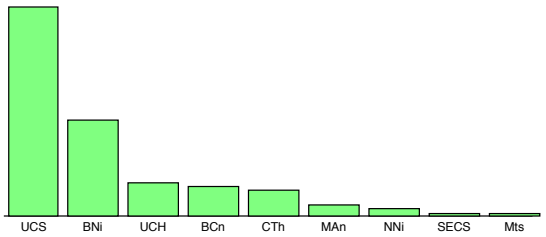
Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Overall variable importance

number of nodes using the variable for splitting



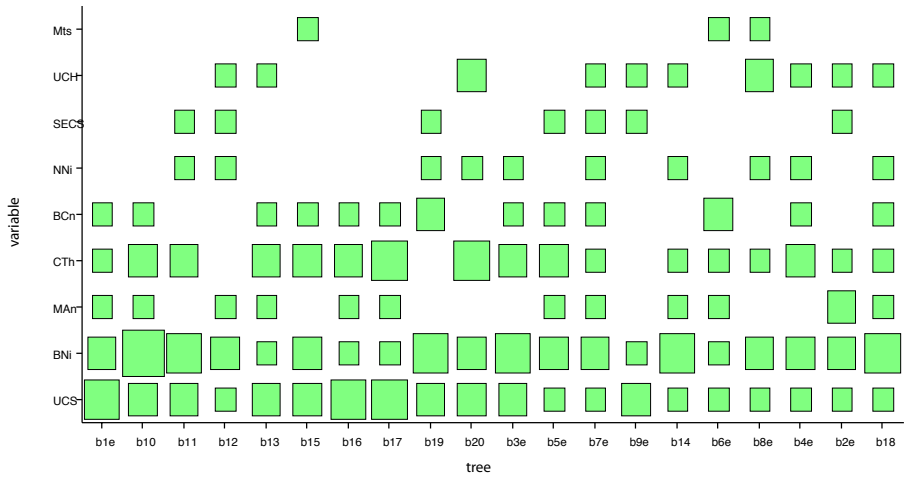
cummulated gain of each variable



Workshop
Ensemble-Methods
Erlangen 2004

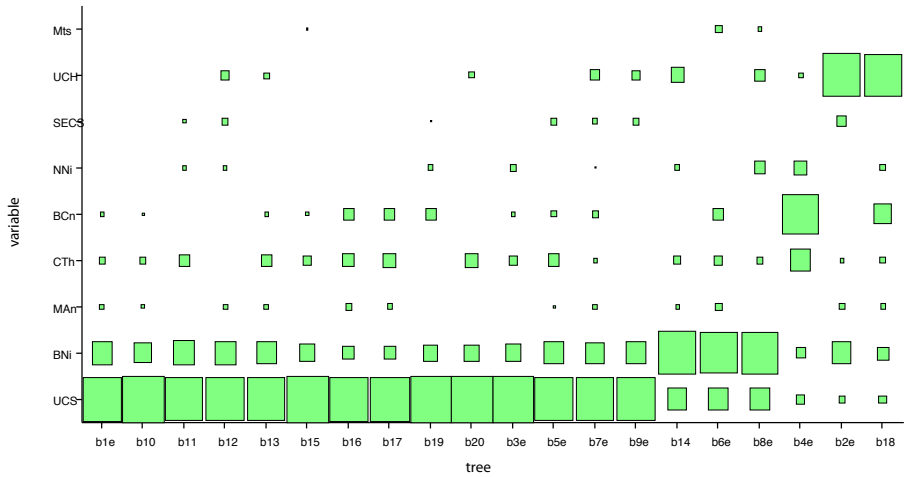
Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Frequency of used variables for each tree



Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

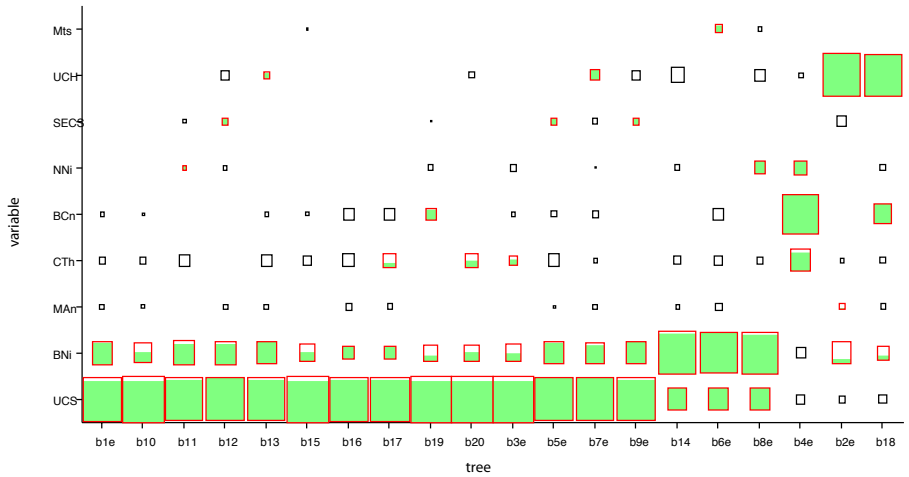
Variable importance for each tree



Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

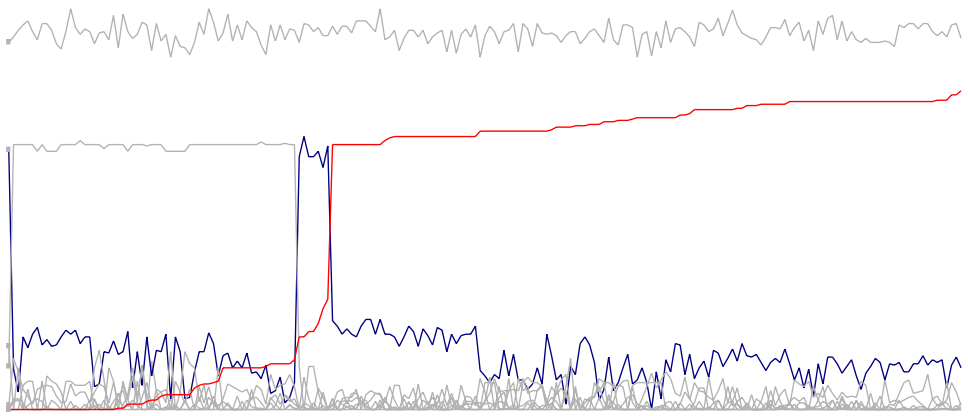
Linked highlighting



Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

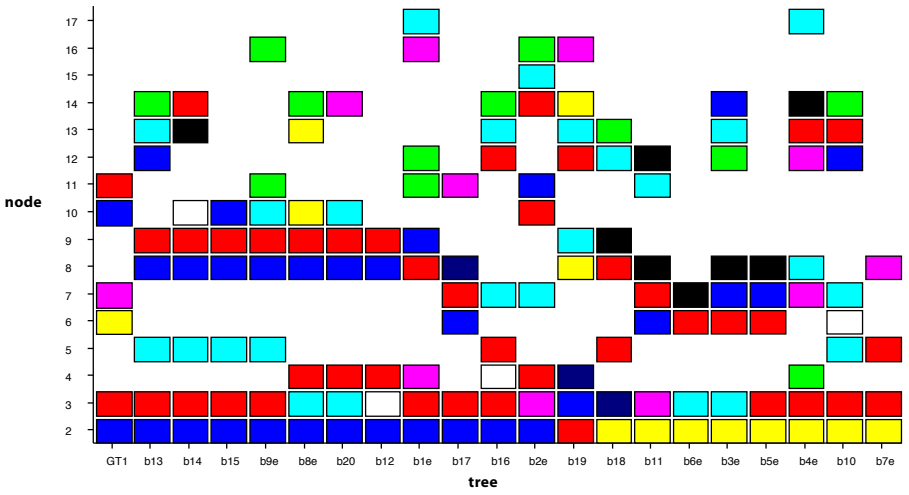
Larger number of models



Workshop
Ensemble-Methods
Erlangen 2004

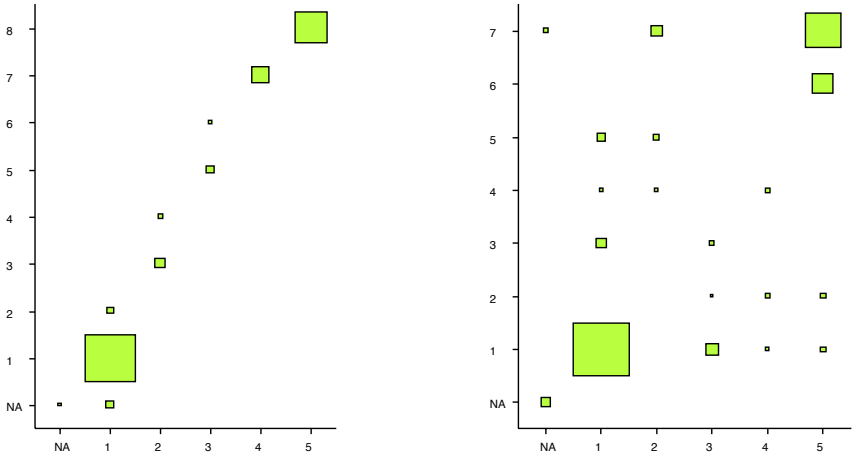
Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Spliting rules



Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

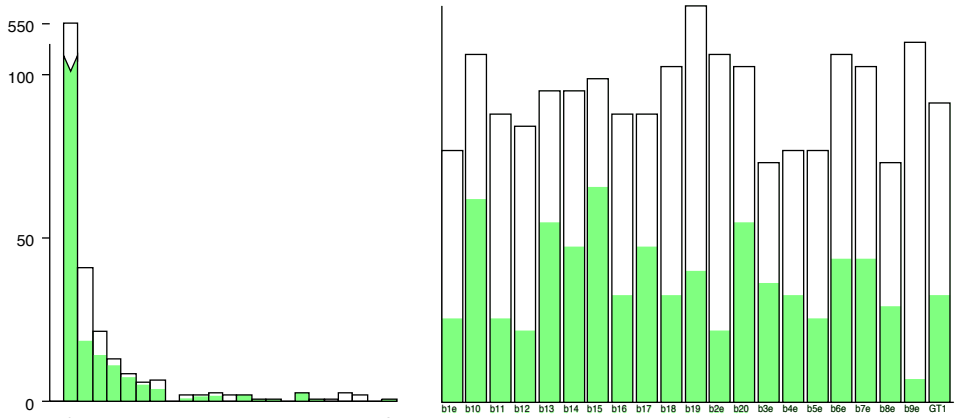
Pairwise comparison: terminal nodes



Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Prediction accuracy: Misclassification



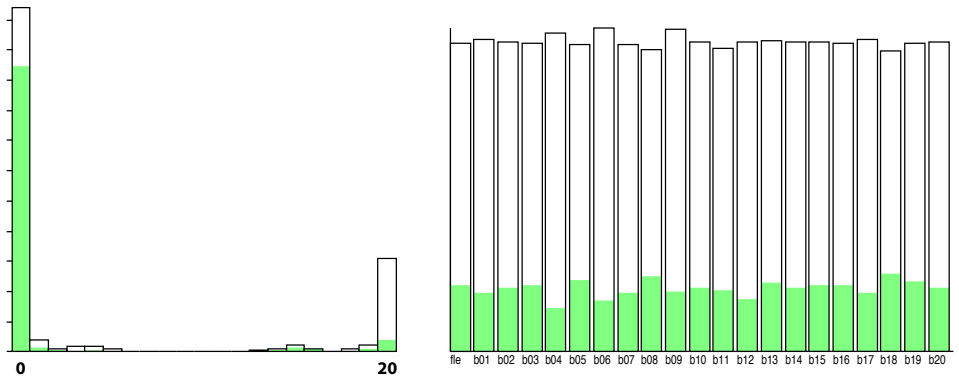
number of misclassifying models

misclassification plot: number of misclassified cases for each tree

Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Misclassification for stable trees



number of misclassifying models

misclassification plot: number of misclassified cases for each tree

Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Implementation of the presented methods

- **KLIMT**
  - interactive software for analysis of tree models and model ensembles
  - Java application, platform independent, free
  - uses R for construction of tree models
  - provides interactive tree construction and modification

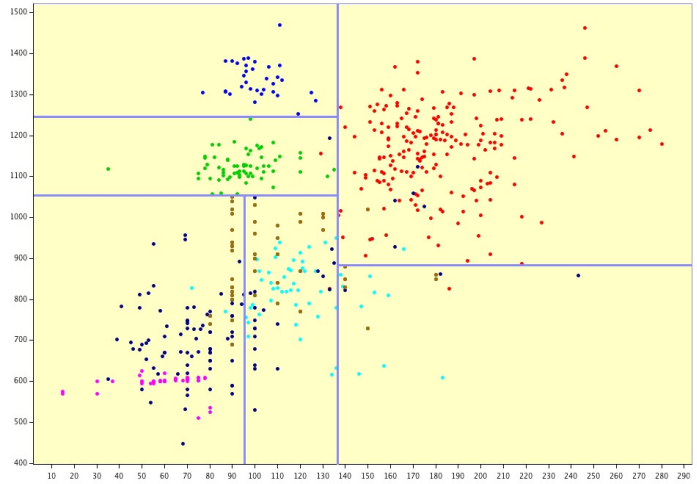
Future research

- **Take split positions into account**
- **Develop methods taking the hierarchical structure into account**
- **Look at more than one split level at a time when assessing individual splits**
- **Explore the entire model space**

Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

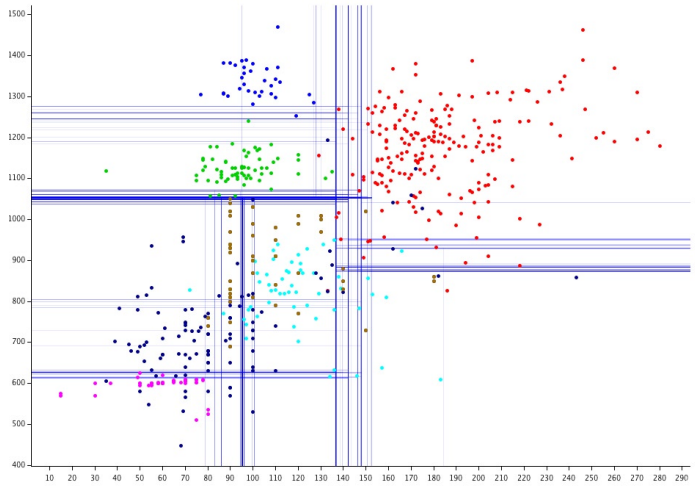
Split positions: sectioned scatterplots



Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

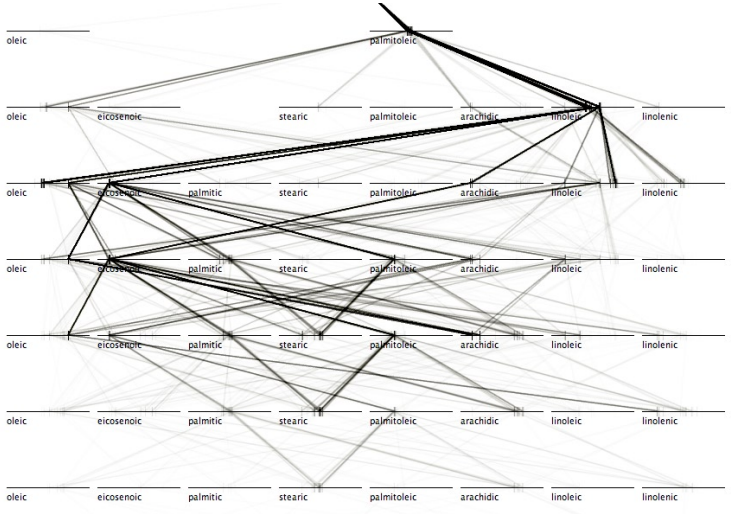
Split positions for 100 trees

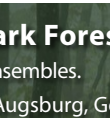


Workshop
Ensemble-Methods
Erlangen 2004

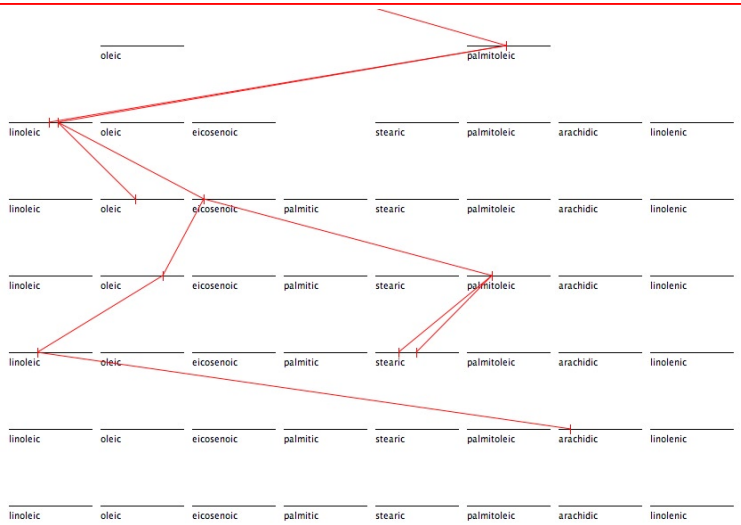
Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany

Split structure in 1000 trees





Split stucture for a single tree



Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany



Contact

Simon URBANEK
Department of computer oriented statistics and data analysis
University of Augsburg, Germany

e-mail: ***Simon.Urbanek@math.uni-augsburg.de***

Web: ***http://stats.math.uni-augsburg.de/~urbanek***

KLIMT: ***http://www.Klimt-Project.com/***

Workshop
Ensemble-Methods
Erlangen 2004

Some Light in a Dark Forest
A closer look at tree ensembles.
Simon Urbanek, University of Augsburg, Germany



Conclusions

- **Local optimality does not necessarily lead to best models**
- **Ensembles try to use wider model space**
- **Forests can be visualizated from different perspectives - recover information about the underlying data structure**
- **Implemented in research software KLIMT**