Complete decision tree induction functionality in scikit-learn

Ir. Sven Van Hove

Current status

- Reviewed literature
  - Result: overview of algorithm capabilities
- Implemented REP (classification) in scikit-learn
- Created automated test bench + plot generation
- Performed experimental comparison

Algorithm capabilities (1)

Capability	ID3	C4.5	Weka J48	CART	scikit DT	scikit RF	scikit GBT
Categorical attributes	Y	Y	Y	Y	N	N	N
Numerical attributes	N	Y	Y	Y	Y	Y	Y
Binary classification (y in [-1,1])	Y	Y	Y	Y	Y	Y	Y
Multiclass classification (y in [0, …, K-1])	N		Y		Y	Y	Y
Multilabel classification	N				Y	Y	N
Multioutput multiclass	N		N		Y	Y	N
Regression (y in R)	N	N	N	Y	Y	Y	Y
max_depth	N		N		Y	Y	Y
min_samples_leaf	N		Y		Y	Y	Y
min_samples_split	N		N		Y	Y	Y
max_leaf_nodes	N		N		Y	Y	Y
max_features	N		N		Y	Y	Y
predict_proba	N		N		Y	Y	Y
Reduced-error pruning (REP)	N	N	Y		N	N	N
Error based pruning (EBP, classification only)	N	Y	Y	N	N	N	N
Minimal cost complexity tree pruning (CCP)	N	N	N	Y	N	N	N
Pessimistic pruning	N	N	N		N	N	N
Rule-based post-pruning	N	Y	N	N	N	N	N
MDL-based pruning	N	N	N	N	N	N	N

Algorithm capabilities (2)

Capability	ID3	C4.5	Weka J48	CART	scikit DT	scikit RF	scikit GBT
Missing values	N	Y	Y	Y	N	N	N
Generate rulesets	N	Y	N	N	N	N	N
Binary splits on categorical values	N		Y	Y	Y	Y	Y
Non-binary splits on categorical values	Y	Y	Y	N	N	N	N
Class weights	N	N	Y		Y	Y	N
Purity split (classification)	N	N	N	N	N	N	N
Entropy split (classification)	Y		Y		N	N	N
Info gain split (classification)	Y	Y	Y	Y	Y	Y	N
Gain ratio split (classification)	N		Y		N	N	N
Gini split (classification)	N		N	Y	Y	Y	N
MSE split (regression)	N	N	N		Y	Y	Y
Friedman_MSE split (regression)	N	N	N		Y	N	Y
MAE split (regression)	N	N	N		Y	Y	Y
Chi-square stop criteria	Y	N	N	N	N	N	N
Hierarchical attributes	N	N	N	N	N	N	N
Learn oblique trees	N	N	N	N	N	N	N
Clustering (unsupervised)	N	N	N	N	N	N	N
Generate model tree	N	N	N	N	N	N	N
Online learning	N	N	N	N	N	N	N

Algorithm capabilities – key take-aways

- No nominal attribute support in scikit-learn
- No regression trees in weka
- No pruning in scikit-learn
  - Instead: pseudo-pruning
- EBP and REP in weka
- Only binary trees in CART, scikit-learn

Experimental setup

- Classifiers
  - PrunableDecisionTreeClassifier
  - J48
- Datasets
  - iris
  - wine
  - diabetes
  - ionosphere
  - wdbc
  - activity
- Pruning options
  - none
  - min_samples_leaf
  - REP [prune_percentage]
  - EBP [confidence_factor]
- Metrics
  - Number of nodes and leaves
  - Accuracy and F1 score
  - Fit and score duration
- 100 repeats, 10-fold cross-validation

J48 options

- Configured to behave similar to scikit-learn decision trees
  - Binary trees only
  - No tree collapsing
  - No subtree raising
  - No MDL correction
  - minNumObject=1 (default=2)
- TODO also test weka with default options (baseline)

Hypotheses – number of nodes and leaves

- Number of nodes ~ number of leaves
- Pruned trees have fewer nodes and leaves
- Pseudo-pruning (i.e., min_samples_leaf): even fewer nodes
- REP vs. EBP?

Plots – number of nodes and leaves

Hypotheses – Accuracy & F1 score

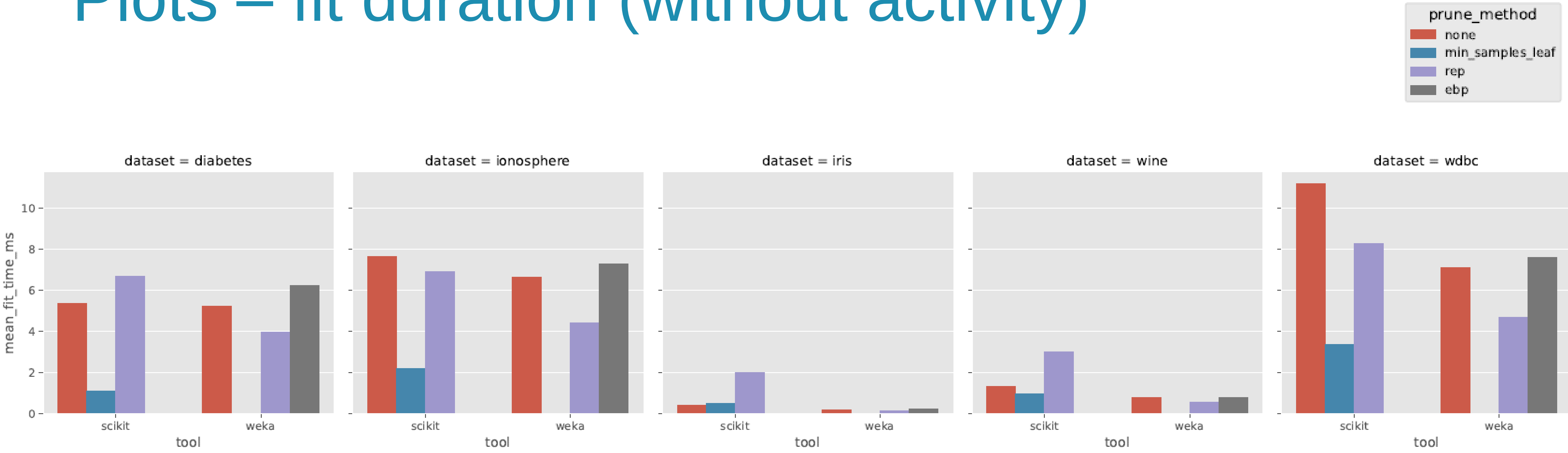
- Accuracy score ≈ F1 score (for balanced class distributions)
- Pruned trees have similar or better accuracy (less overfitting)
- Aggressive pruning (i.e., min_samples_leaf): lower accuracy
- Weka and scikit score similarly
    - Except for activity dataset?

Plots – Accuracy & F1 score

Hypotheses – fit and score duration

- Pruned trees fit more slowly
- Pseudo-pruned trees fit faster compared to unpruned trees
- Score time ~ tree size / max depth
  - Pruned trees score more quickly
- Notes
  - Weka (Java) and scikit (Python/Cython/C): apples and oranges
  - Using built-in timers of weka and scikit-learn
  - Measurement accuracy?

Plots – fit duration (without activity)



KU LEUVEN

Plots – fit and score duration

Next steps (MoSCoW)

Must have	Should have	Could/would have
REP for regression	Code documentation	Python 2.x compatibility
Thesis text (*)	Study effect on ensembles	Contribution-ready code
Other pruning algorithm(s)	Multi-output support	Missing values
Analyze score duration discrepancy	Improve memory usage	Nominal values
Reproduce accuracy problem	Speed up (Cython?)	Online learning

(*) Dutch thesis title?

Thank you