



Big Data Analytics

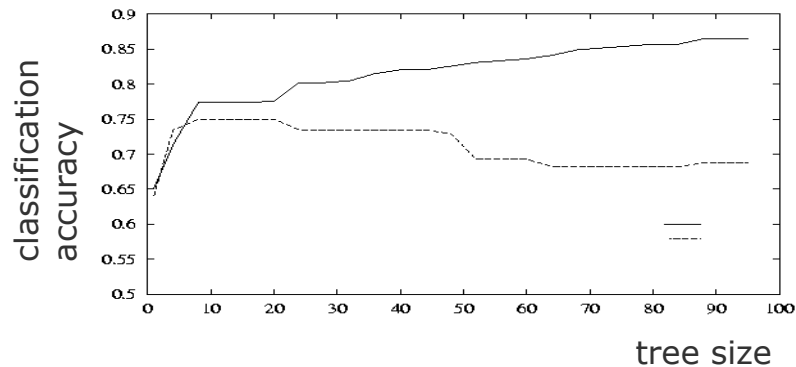
– Week 3: Decision Tree Pruning –

Knowledge Discovery and Data Mining
Prof. Dr. Emmanuel Müller

Avoid Overfitting in Classification

The generated tree may overfit the training data

- Too many branches, some may reflect anomalies due to noise or outliers
- Result has poor accuracy for unseen samples



Two approaches to avoid overfitting for decision trees:

1. **Post pruning** = pruning of overspecialized branches
 - Remove branches from a “fully grown” tree & get a sequence of progressively pruned trees
 - Use a set of data different from the training data to decide which is the “best pruned tree”

Overfitting: Avoidance

2. **Prepruning** = halt tree construction early, do not split a node if this would result in the goodness measure falling below a threshold

- Choice of an appropriate value for **minimum support**
 - *minimum support*: minimum number of data objects a leaf node contains
 - in general, *minimum support* $\gg 1$
- Choice of an appropriate value for **minimum confidence**
 - *minimum confidence*: minimum fraction of the majority class in a leaf node
 - typically, minimum confidence $\ll 100\%$
 - leaf nodes can absorb errors or noise in data records
- Discussion
 - With Prepruning it is difficult to choose appropriate thresholds
 - Prepruning has less information for the pruning decision than Postpruning. In general, it therefore can be expected to produce decision trees with lower classification quality.
 - Tradeoff: tree construction time $\leftrightarrow$ classification quality

Pruning of Decision Trees: Postpruning

Reduced-Error Pruning [Q87]

- Decompose classified data into training set and test set
- Creation of a decision tree E for the training set
- Pruning of E by using the test set T
 - determine the subtree S of E whose pruning reduces the classification error on T the most (i.e., replace S by a leaf assigning to it the most common classification of the training examples in S and compare the number of errors of the original and reduced tree)
 - remove that subtree
 - finished if no such subtree exists
- only applicable if a sufficient number of classified data is available

[Q87] J.R. Quinlan. *Rule induction with statistical data – a comparison with multiple regression*. In Journal of the Operational Research Society, 38, pages 347-352, 1987.

Pruning of Decision Trees: Postpruning

Minimal Cost Complexity Pruning [BFO+84]

- Does not require a separate test set
 - applicable to small training sets as well
- Pruning of the decision tree by using the training set
 - classification error is no appropriate quality measure
- New quality measure for decision trees:
 - trade-off of classification error and tree size
 - weighted sum of classification error and tree size
- General observation
 - the smaller decision trees yield the better generalization

[BFO+84] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth International Group, 1984.

Decision Tree Classifiers: Summary

Hierarchical linear classifier for data with attributes (categorical or numerical)

Pro

- Relatively fast learning speed (in comparison to other classification methods)
- Fast classification speed
- Convertible to simple and easy to understand classification rules
- Often comparable classification accuracy with other classification methods

Contra

- Not very stable, small changes of the data can lead to large changes of the tree