Data Mining with Weka

Pruning decision trees

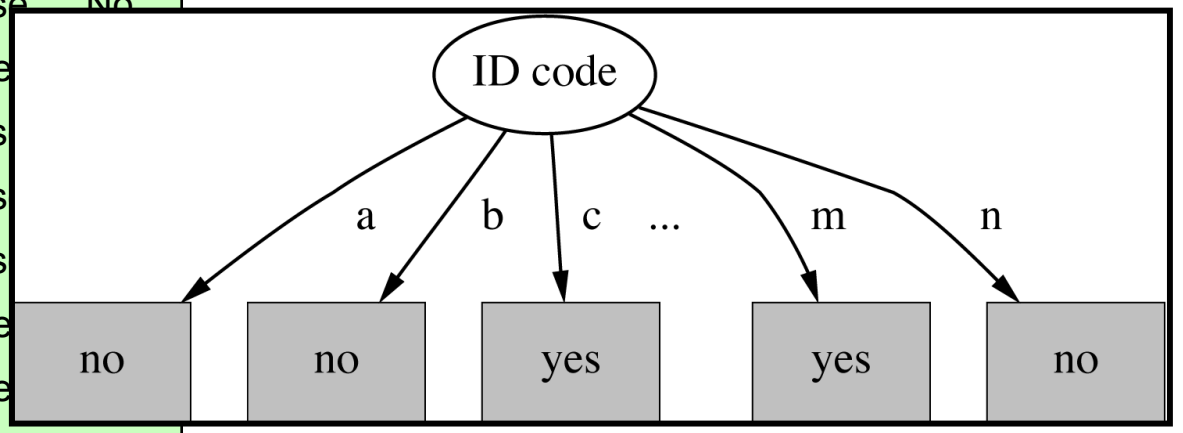
Ian H. Witten

Pruning decision trees

Pruning decision trees

Highly branching attributes — Extreme case: ID code

ID code	Outlook	Temp	Humidity	Wind	Play
a	Sunny	Hot	High	False	No
b	Sunny	Hot	High	True	
c	Overcast	Hot	High	Fals	
d	Rainy	Mild	High	Fals	
e	Rainy	Cool	Normal	Fals	
f	Rainy	Cool	Normal	True	
g	Overcast	Cool	Normal	True	
h	Sunny	Mild	High	False	No
i	Sunny	Cool	Normal	False	Yes
j	Rainy	Mild	Normal	False	Yes
k	Sunny	Mild	Normal	True	Yes
l	Overcast	Mild	High	True	Yes
m	Overcast	Hot	Normal	False	Yes
n	Rainy	Mild	High	True	No



Information gain is maximal (0.940 bits)

Pruning decision trees

How to prune?

- Don't continue splitting if the nodes get very small (J48 minNumObj parameter, default value 2)
- Build full tree and then work back from the leaves, applying a statistical test at each stage (confidenceFactor parameter, default value 0.25)
- Sometimes it's good to prune an interior node, raising the subtree beneath it up one level (subtreeRaising, default *true*)
- Messy … complicated … not particularly illuminating

Pruning decision trees

Over-fitting (again!)

Sometimes simplifying a decision tree gives better results

- Open file diabetes.arff
- Choose J48 decision tree learner (trees>J48)
- Prunes by default: 73.8% accuracy, tree has20 leaves, 39 nodes
- Turn off pruning: 72.7% 22 leaves, 43 nodes
- Extreme example: breast-cancer.arff
- Default (pruned): 75.5% accuracy, tree has 4 leaves, 6 nodes
- Unpruned: 69.6% 152 leaves, 179 nodes

Pruning decision trees

- C4.5/J48 is a popular early machine learning method
- Many different pruning methods
  - mainly change the size of the pruned tree
- Pruning is a general technique that can apply to structures other than trees (e.g. decision rules)
- Univariate vs. multivariate decision trees
  - Single vs. compound tests at the nodes
- From C4.5 to J48

Ross Quinlan,
Australian computer scientist