The C4.5 Project

Overview of algorithm with results of experimentation

Summary

- Terminology
- C4.5 vs. ID3
- Datasets
- C4.5 results on datasets

Terminology

- Training cases
- Test cases
- Unseen cases

Gain vs. Gain Ratio

- ID3 creates complex trees using gain
- C4.5 uses a different measure
  - Gain ratio considers what ID3 does not
  - Minimum number of instances per leaf node
- Meaning: C4.5 creates more useful models

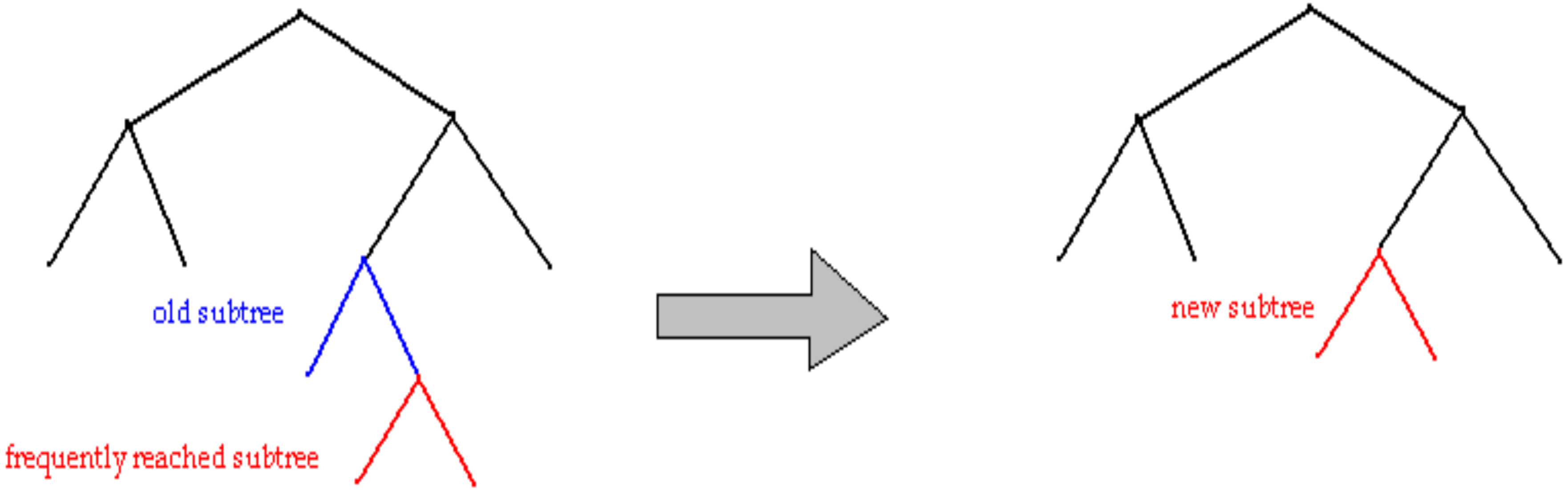
Missing Data

- ID3 does not make allowances
- C4.5 adjusts the gain ratio to favor attributes with existing values
- Classifying training and unseen cases
  - C4.5 uses probabilistic weights

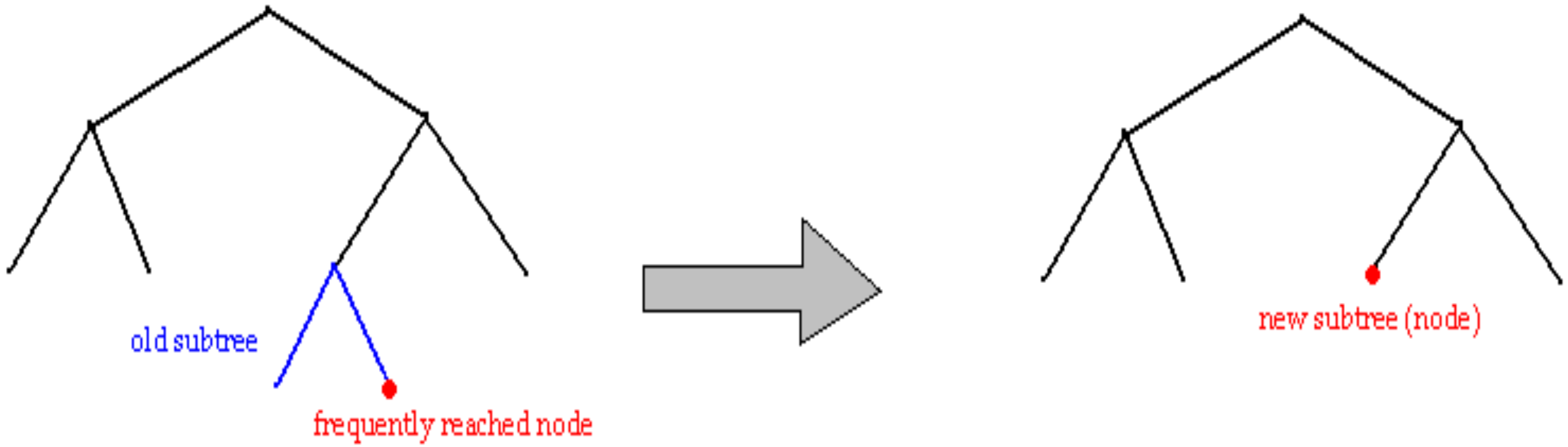
Pruning

- ID3 produces complex trees
- C4.5 prunes trees
  - Pessimistic error prediction
  - Subtree raising
  - Subtree replacement

Subtree Raising



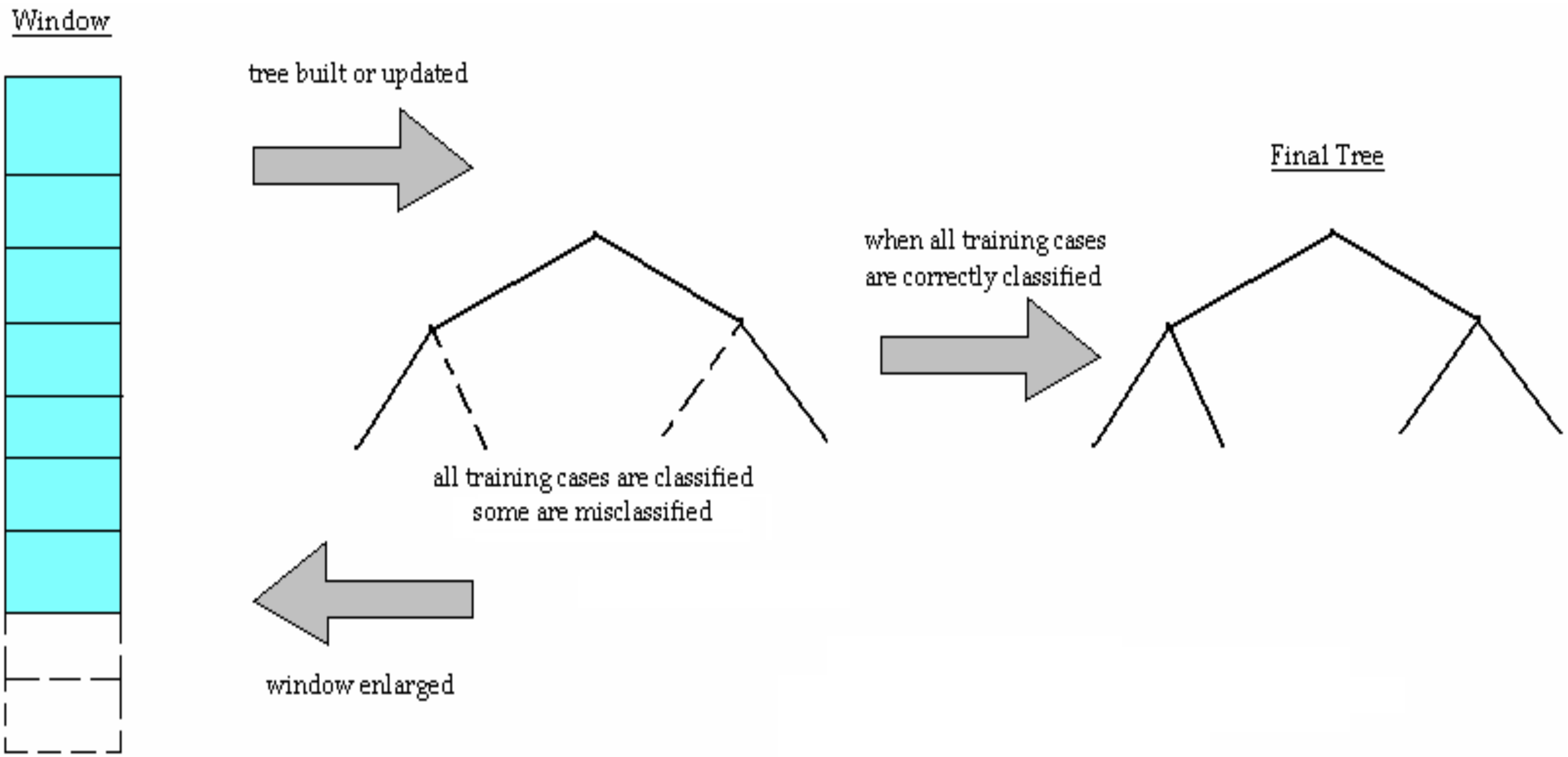
Subtree Replacement



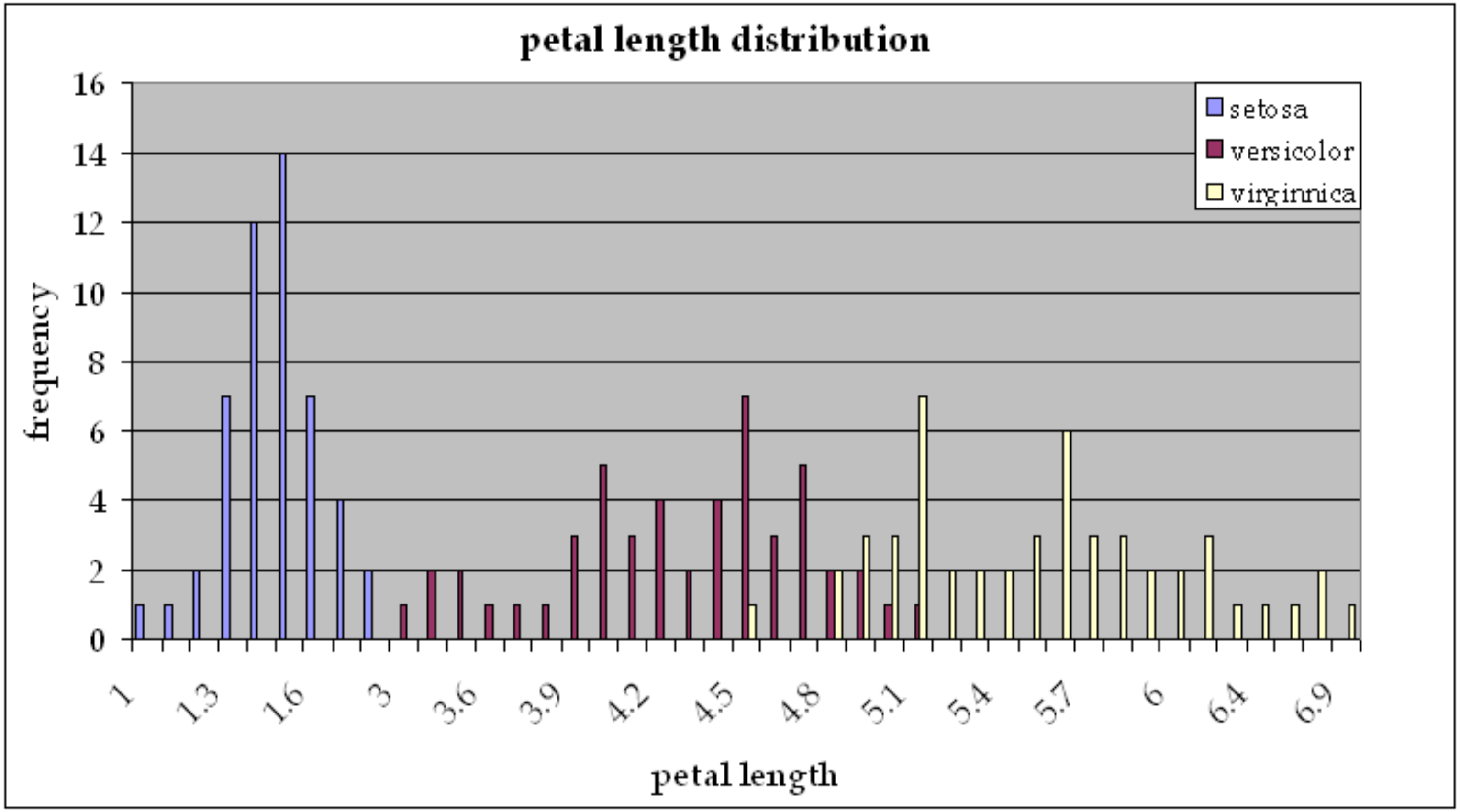
Features of C4.5

- Rules
- Consulter
- Categorical data
- Windowing

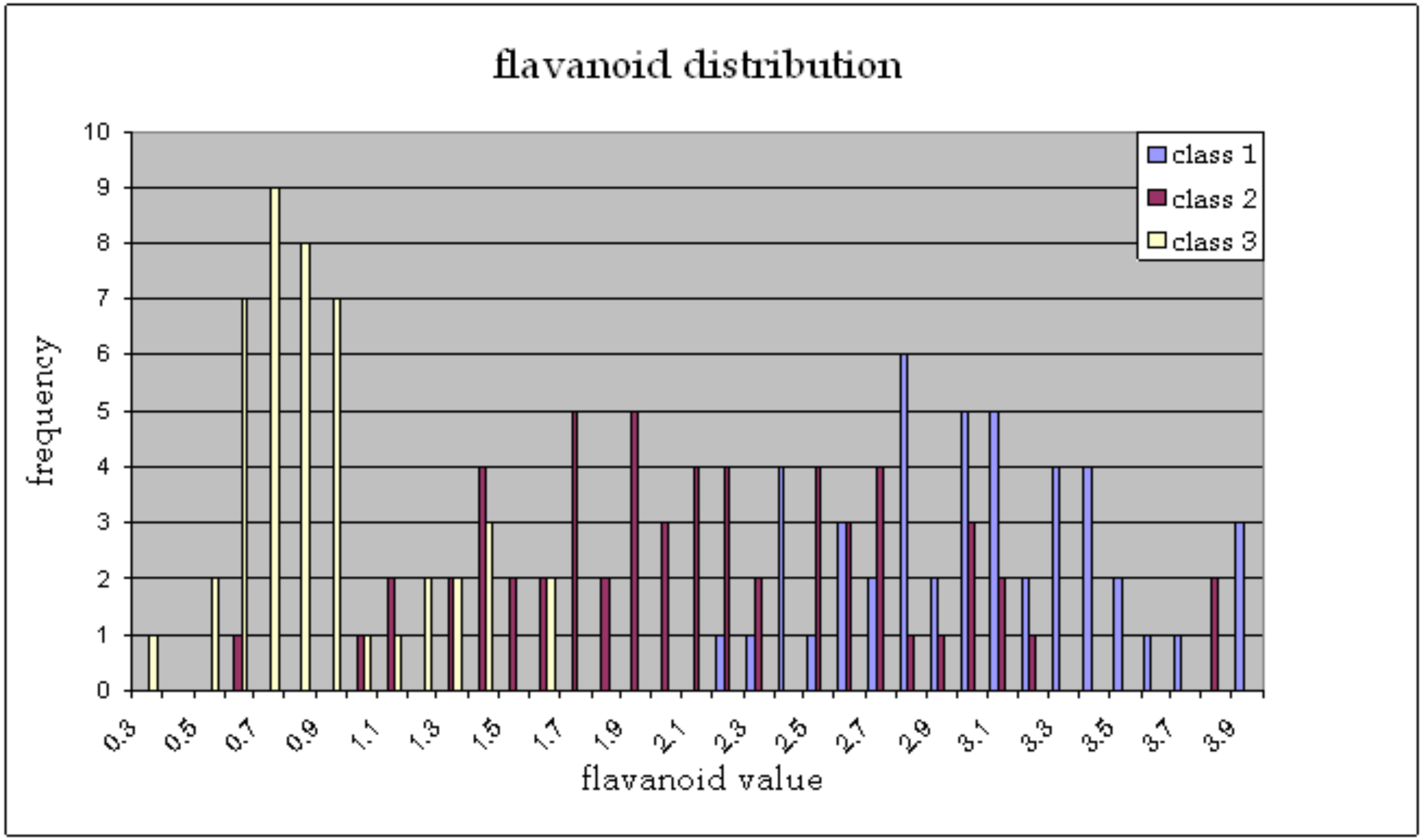
Windowing


Window
tree built or updated
Final Tree
when all training cases
are correctly classified
all training cases are classified
some are misclassified
window enlarged

Iris Dataset

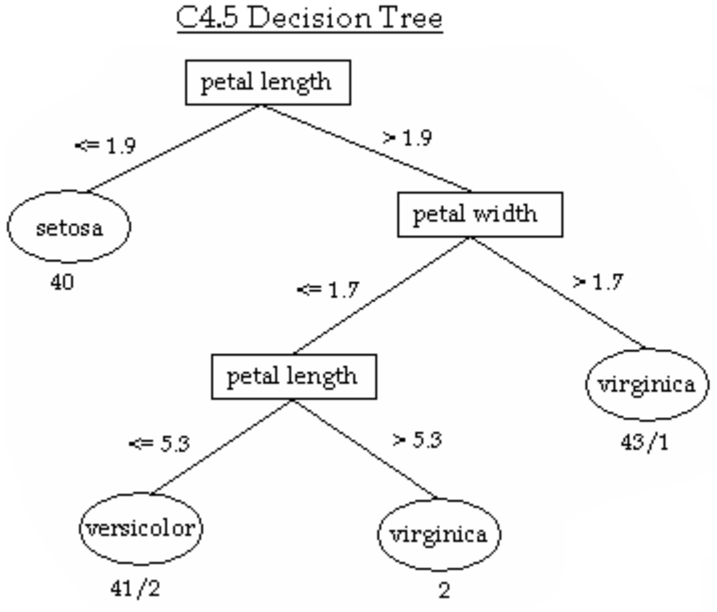


Wine Dataset



Results of C4.5 on Datasets

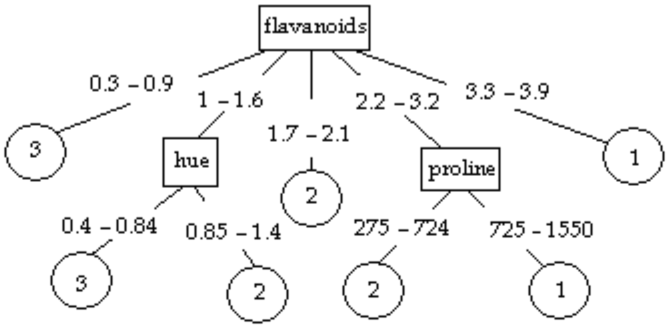
- Iris dataset: similar results



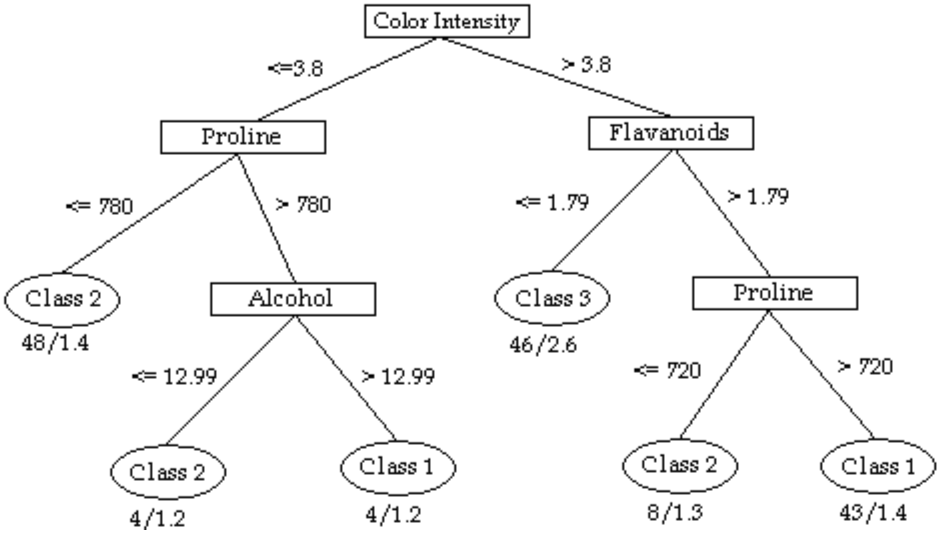
Results of C4.5 on Datasets

- Wine dataset: different results
  - Possible reasons for differences

ID3 Decision Tree



C4.5 Decision Tree



Closing Summary

- C4.5 vs. ID3
  - Gain vs. gain ratio
  - Missing data
  - Pruning
  - Features of C4.5
- C4.5 Results
  - Iris – similar results
  - Wine – different results

The End

reasonable questions welcomed.