THE C4.5 PROJECT

By
Fatine Bourkadi

Outlines

- *Introduction to C4.5*
- Training Set
- Test set
- Data Sets
- results

Introduction to C4.5

- C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan..
- C4.5 is an extension of Quinlan ID3 algorithm
- C4.5 builds decision trees from a set of training data using the concept of information entropy.

Training Set

- Entropy
  - $H(\mathrm{R},\mathrm{A})=\sum\nolimits_{i} p(class_i / R) \times \log[p(class_i / R)]$

Training Set

- The training data is a set of $S=s\downarrow 1 , s\downarrow 2 ,...$ of already classified examples. Each sample $S\downarrow i ,= x\downarrow 1 , x\downarrow 2 ,...$ is a vector where $x\downarrow 1 , x\downarrow 2 ,$ represent attributes or features of the sample. The training data is augmented with vector $C=c\downarrow 1 , c\downarrow 2 ,...$ represent the class to each sample belongs.
- Name file
  - Provides names for classes, attributes, and attribute values.
- Data file
  - Describe the training cases from which decision trees are to be constructed.

Test Set

- Test file
  - Test set to evaluate the classifier that C4.5 have produced.

Iris Flower Datasets

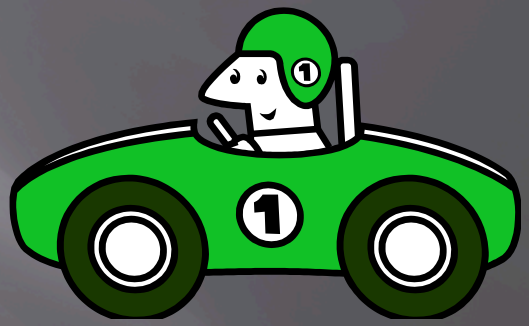
- Iris flower
  - 150 instances
  - three classes:
    - Iris-setosa
    - Iris-versicolor
    - Iris-virginica
  - Four Attributes in cm:
    - Sepal width
    - Sepal length
    - Petal width
    - Petal length

Wine Dataset

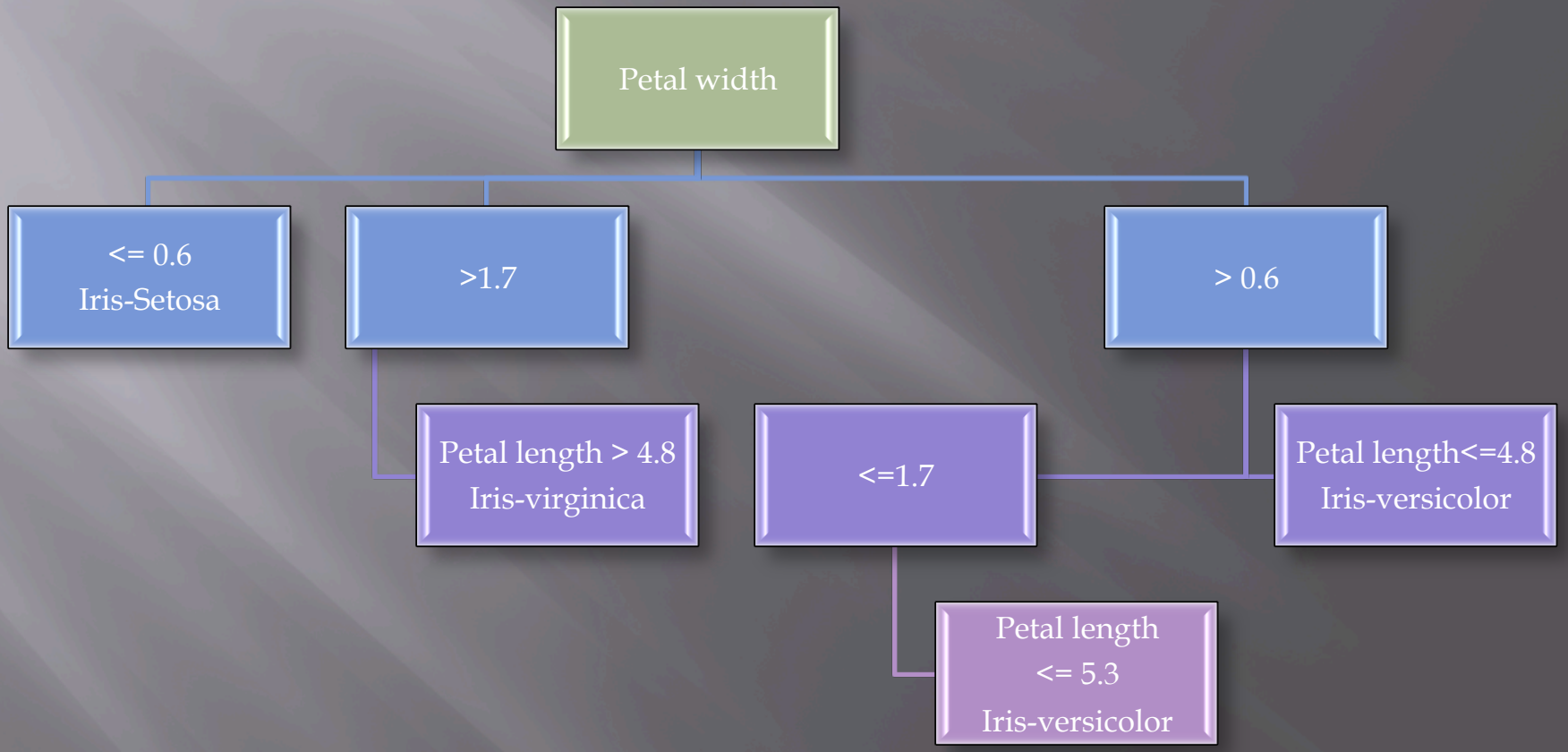
- Wine
  - 153 instances
  - Three classes:
    - Class 1
    - Class 2
    - Class 3
  - 13 attributes:
    - Alcohol, Malic acid, Ash, Alkalinity of ash, Flavonoids, Magnesium, Nonflavanoid Phenols, Proanthocyanins, color intensity, Hue, OD280/OD315 of diluted wines, Praline.

Car Evaluation Dataset

- Car Evaluation
  - 1728 instances
  - Four classes:
    - Unacceptable
    - Acceptable
    - Good
    - Very good
  - Six attributes:
    - Buying
    - Maintenance
    - Doors
    - Person
    - Lug-boot
    - Safety



Iris Dataset Decision Tree



Wine Dataset

- Rules for wine dataset:
  - Rules examines the original decision tree produced by the C45. program and derives from it a set of production rules of this form.

```
Rule 5:
        Color intensity > 3.4
        OD280/OD315 of diluted wines > 2.11
        Praline > 714
        ->  class 1  [96.1%]

Rule 4:
        Color intensity <= 3.4
        ->  class 2  [96.2%]

Rule 3:
        OD280/OD315 of diluted wines > 2.11
        Praline <= 714
        ->  class 2  [93.6%]

Rule 1:
        Hue <= 0.96
        OD280/OD315 of diluted wines <= 2.11
        ->  class 3  [92.5%]

Default class: 2
```

Wine dataset

- The statistics for the first rule

Evaluation on training data (115 items):

Rule	Size	Error	Used	Wrong	Advantage	
5	3	3.9%	35	0 (0.0%)	35 (35|0)	1

Car Evaluation dataset

- Following the report on each rule there is a summary and a confusion matrix showing where the misclassifications of the training cases occur.

Tested 1127, errors 61 (5.4%) <<

(a)	(b)	(c)	(d)	<-classified as
124			1	(a): class acc
36	7		2	(b): class good
13		895		(c): class unacc
9			40	(d): class vgood

Result Review

- Iris Dataset:
  - Has 100% accurate data
- Wine Dataset:
  - Has 89.7% accurate data
- Car Evaluation Dataset:
  - Has 98.7% accurate data

References

- Aleshunas, J. (n.d.). *http://mercury.webster.edu/aleshunas/Data%20Sets/Supplemental%20Excel%20Data%20Sets.htm*.
- Dunham, M. H. (2003). *Data Mining Introductory and Advanced Topics*. Saddle River, New Jersey, USA: pearson Education.inc.
- Khoa D, D. (2006). *Comparing Classification using C4.5, Naive Bayes, K-nearest Neighbors, and Backpropagation Neural Network Algorithms*.
- Kumar, X. W. (2009). *The Top Ten Algorithms in Data Mining*. Boca Raton: Taylor and Francis Group, LLC.
- Pang-Ning Tan, M. S. (2006). *Introduction to Data Mining*. Boston, MA, USA: Pearson Education, Inc.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Meteo: Morgan Kaufmann .