Data Mining with Weka

Simplicity first!

Ian H. Witten

Simplicity first!

Simple algorithms often work very well!

- There are many kinds of simple structure, eg:
  - *One attribute does all the work*
  - *Attributes contribute equally and independently*
  - *A decision tree that tests a few attributes*
  - *Calculate distance from training instances*
  - *Result depends on a linear combination of attributes*

- Success of method depends on the domain
  - *Data mining is an experimental science*

Simplicity first!

OneR: One attribute does all the work

- Learn a 1-level "decision tree"
  - *i.e., rules that all test one particular attribute*

- Basic version
  - *One branch for each value*
  - *Each branch assigns most frequent class*
  - *Error rate: proportion of instances that don't belong to the majority class of their corresponding branch*
  - *Choose attribute with smallest error rate*

Simplicity first!

```
For each attribute,
    For each value of the attribute,
    make a rule as follows:
        count how often each class appears
        find the most frequent class
        make the rule assign that class
          to this attribute-value
    Calculate the error rate of this attribute's rules
Choose the attribute with the smallest error rate
```

Simplicity first!

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Attribute	Rules	Errors	Total errors
Outlook	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temp	Hot → No*	2/4	5/14
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity	High → No	3/7	4/14
	Normal → Yes	1/7	
Wind	False → Yes	2/8	5/14
	True → No*	3/6	

* indicates a tie

Simplicity first!

Use OneR

- Open file weather.nominal.arff
- Choose OneR rule learner (rules>OneR)
- Look at the rule (*note: Weka runs OneR 11 times*)

Simplicity first!

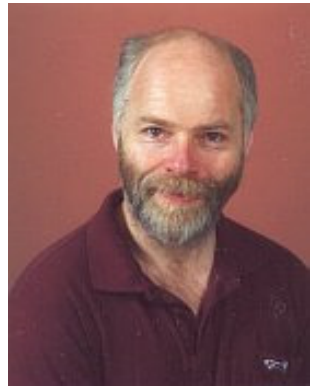
OneR: One attribute does all the work

- Incredibly simple method, described in 1993

  "*Very Simple Classification Rules Perform Well on Most Commonly Used Datasets*"

  - Experimental evaluation on 16 datasets
  - Used cross-validation
  - Simple rules often outperformed far more complex methods

- How can it work so well?
  - some datasets really are simple
  - some are so small/noisy/complex that nothing can be learned from them!



Rob Holte,
Alberta, Canada