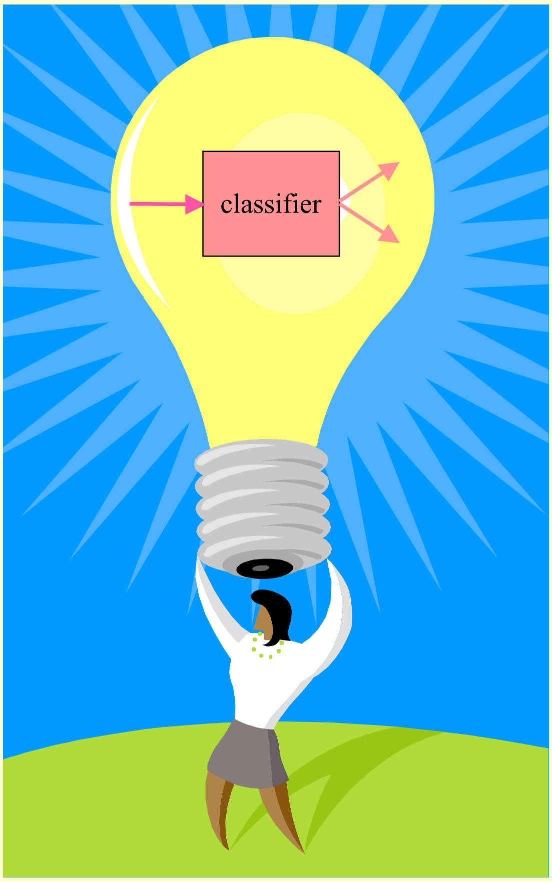
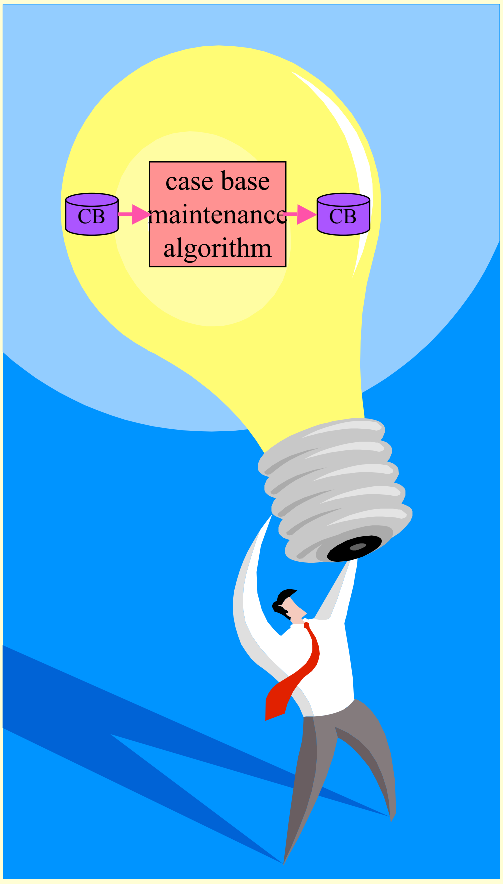
On the Role of Dataset Complexity in Case-Based Reasoning

Derek Bridge
UCC
Ireland

(based on work done with Lisa Cummins)



classifier



CB

case base maintenance algorithm

CB

Overview

- Dataset complexity measures
- Classification experiment
- Case base maintenance experiment
- Going forward

Overview

- Dataset complexity measures
- Classification experiment
- Case base maintenance experiment
- Going forward

Dataset Complexity Measures

- Measures of classification difficulty
  - *apparent* difficulty, since we measure a dataset which samples the problem space
- Little impact on CBR
  - Fornells et al., ICCBR 2009
  - Cummins & Bridge, ICCBR 2009
- (Little impact on ML in general!)

Dataset Complexity Measures

- Survey of 12 geometrical measures
  - Ho & Basu, 2002
- DCoL: open source C++ library of 13 measures
  - Orriols-Puig et al., 2009
- We have found 4 candidate measures in the CBR literature

Overlap of attribute values

F_1	Maximum Fisher’s Discriminant Ratio
F_2'	Volume of Overlap Region
F_3'	Maximum Attribute Efficiency
F_4'	Collective Attribute Efficiency

Separability of classes

N_1'	Fraction of Instances on a Boundary
N_2	Ratio of Average Intra/Inter Class Distance
N_3	Error Rate of a 1NN classifier
L_1	Minimized Sum of Error Distance of a Linear Classifier
L_2	Training Error of a Linear Classifier
C_1	Complexity Profile
C_2	Similarity-Weighted Complexity Profile
N_5	Separability Emphasis Measure

Manifold Topology & Density

L_3	Nonlinearity of a Linear Classifier
N_4	Nonlinearity of a 1NN Classifier
T_1	Fraction of Maximum Covering Spheres
T_2	Number of Instances per Attribute
T_3	Dataset Competence

Dataset Complexity Measures

- Desiderata
  - Predictive
  - Independent of what is being analyzed
  - Widely applicable across datasets
  - Cheap-to-compute
  - Incremental
  - Transparent/explainable

Overview

✓ Dataset complexity measures

➢ Classification experiment

➢ Case base maintenance experiment

➢ Going forward

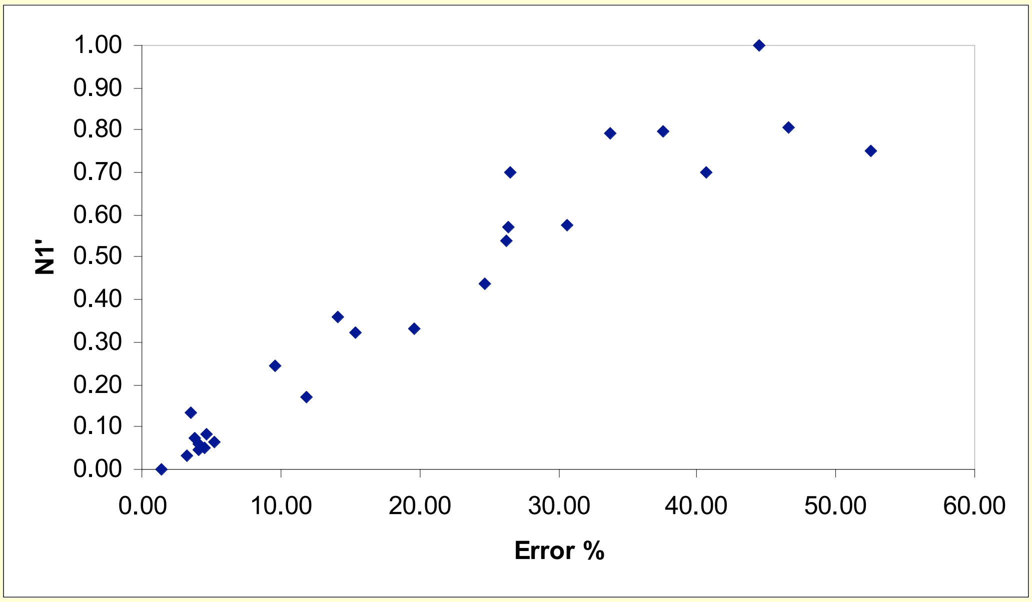
Classification experiment

- 25 datasets
  - 14 Boolean classification; 11 multi-class
  - 21 numeric-valued attributes only (12 Boolean classification; 9 multi-class)
- 4 Weka classifiers trained on 60% of dataset
  - Neural Net with 1 hidden layer
  - SVM with SMO
  - J48
  - IBk with k = 3
- Error measured on 20% of dataset
- Repeated 10 times

An example of the results

Dataset	NN	SVM	J48	IBk	Mean	N_1'
Iris	2.67	4.00	5.00	2.67	3.58	0.13
Lung Cancer	58.00	50.00	46.00	56.00	52.50	0.75

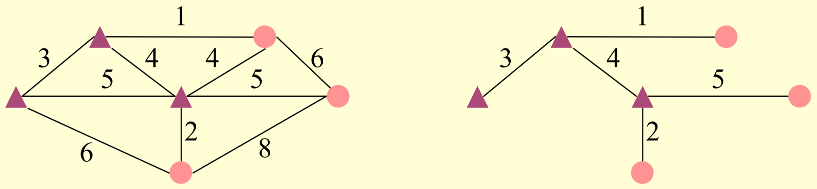
An example of the results



Correlation coefficient: 0.96

N_1' *Fraction of instances on a boundary*

- Build a minimum spanning tree



- Compute fraction of instances directly connected to instances of a different class
- Shuffle dataset, repeat, & average

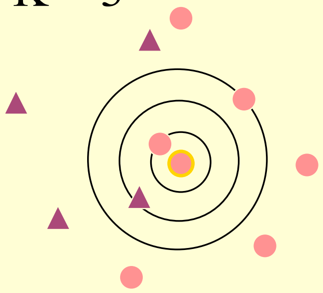
Other competitive measures

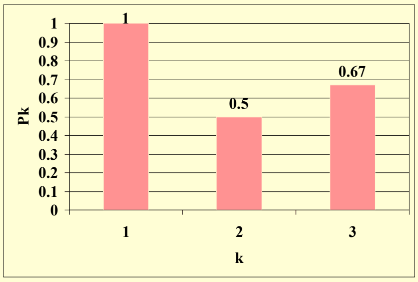
- N_3 Error Rate of a 1NN Classifier
  - leave-one-out error rate of 1NN on the dataset
- N_2 Ratio of Average Intra/Inter Class Distance
  - sum distances to nearest neighbour of same class
  - divide by sum of distances to nearest neighbour of different class
- L_2 Training Error of a Linear Classifier
  - build, e.g., SVM on dataset
  - compute error on original dataset
  - problems with multi-class; problems with symbolic values

C_1 *Complexity Profile*

- Computed for each instance, with parameter K [Massie et al. 2006]

For K = 3





- For a dataset measure, compute average complexity

Other measures from CBR

- C_2 Similarity-Weighted Complexity Profile
  - use similarity values when computing P_k
- N_5 Separability Emphasis Measure [Fornells et al. ’09]
  - $N_5 = N_1' \times N_2$
- T_3 Dataset Competence [Smyth & McKenna ’98]
  - competence groups based on overlapping coverage sets
  - group coverage based on size and similarity
  - dataset competence as sum of group coverages

Their predictivity

- C_1 Complexity Profile
  - Correlation coefficient: 0.98
- C_2 Similarity-Weighted Complexity Profile
  - Correlation coefficient: 0.97
- N_5 Separability Emphasis Measure
  - Between N_1' and N_2
- T_3 Dataset Competence
  - Correlation coefficient: near zero

Summary of experiment

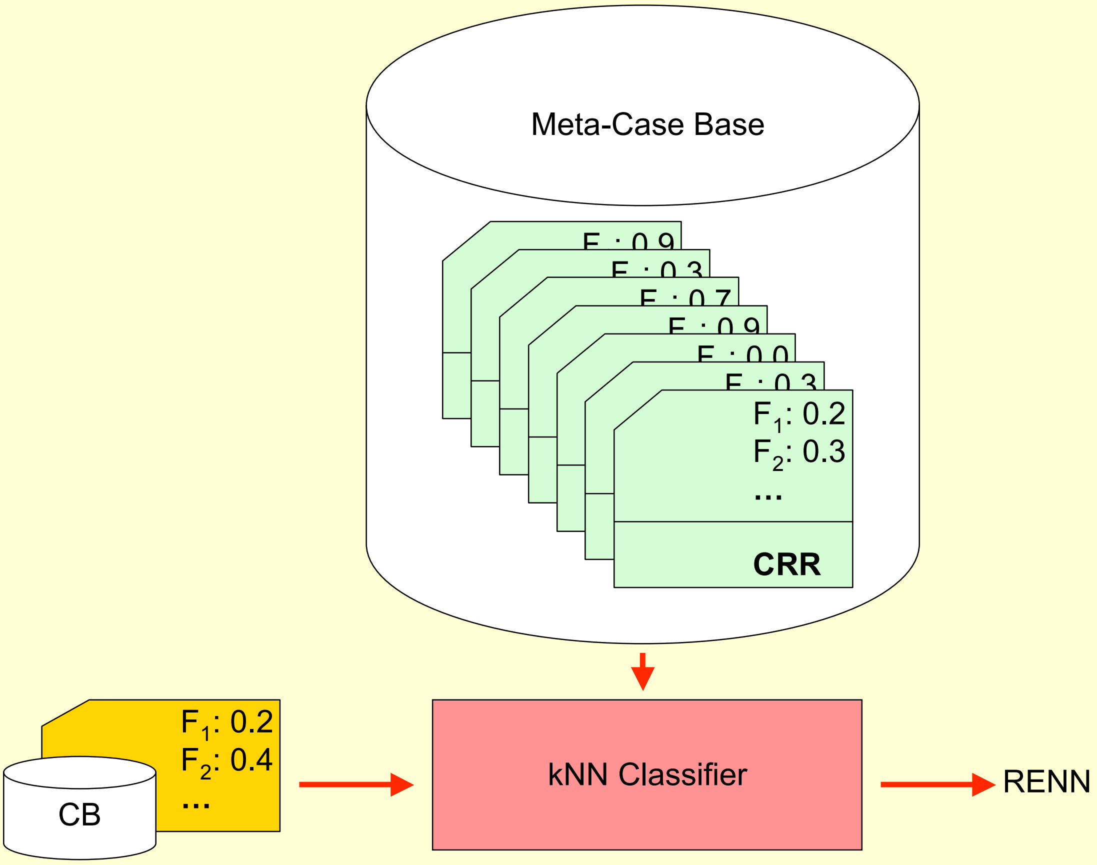
- Very predictive
  - C_1 Complexity Profile
  - N_3 Error Rate of 1NN Classifier
  - N_1' Fraction of Instances on a Boundary
- Predictive but problems with applicability
  - L_2 Training Error of a Linear Classifier
- Moderately predictive
  - N_2 Ratio of Average Intra/Inter Class Distance
- All are measures of *separability of classes*

Overview

- ✓ Dataset complexity measures
- ✓ Classification experiment
- ➢ Case base maintenance experiment
- ➢ Going forward

Meta-CBR for Maintenance

- Case base maintenance algorithms seek to:
  - delete noisy cases
  - delete redundant cases
- Different case bases require different maintenance algorithms
- The same case base may require different maintenance algorithms at different times in its life cycle
- We have been building classifiers to select maintenance algorithms



Meta-Case Base

F_1: 0.2
F_2: 0.3
...
CRR

CB

F_1: 0.2
F_2: 0.4
...

kNN Classifier

RENN

Case Base Maintenance Experiment

- Training (building the meta-case base)
  - From 60% of each dataset, create a case base
  - Create a meta-case to describe this case base
    - attributes are complexity measures
    - problem solution
      - run a small set of maintenance algorithms on each case base
      - record % deleted
      - record accuracy on the next 20% of each dataset
      - maintenance algorithm with highest harmonic mean of % deleted and accuracy becomes this meta-case's solution
- But, we use *feature selection* to choose a subset of the complexity measures
  - wrapper method, best-first search

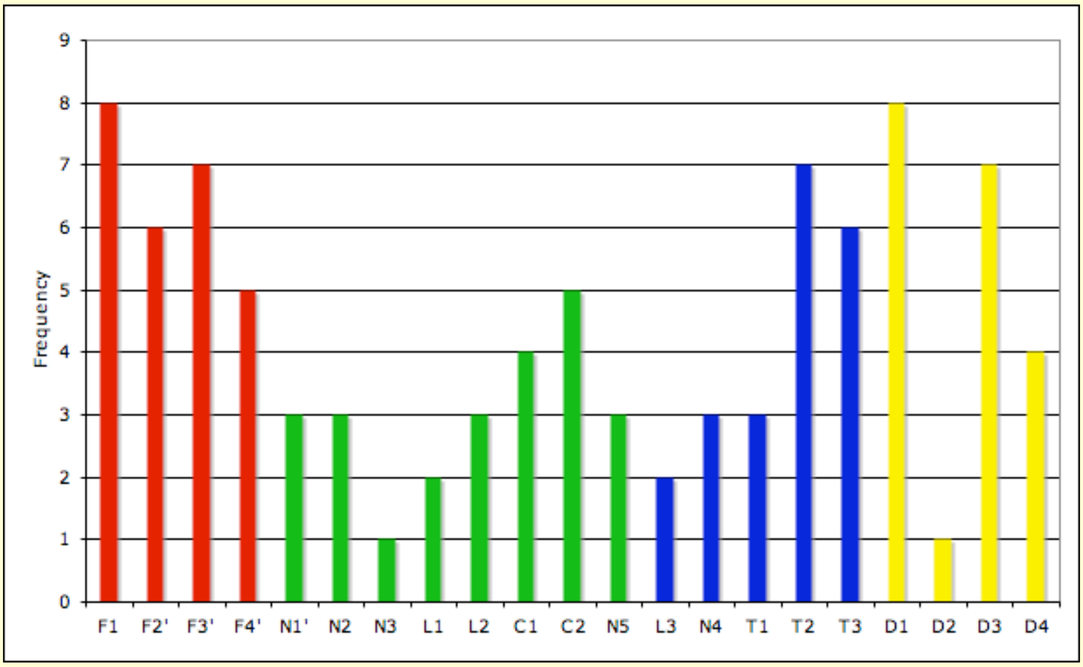
Case Base Maintenance Experiment

- Testing
  - Target problem is a case base built from remaining 20% of each dataset
    - attributes again are complexity measures
  - Ask the classifier to predict a maintenance algorithm
  - Run the algorithm, record % deleted, accuracy and their harmonic mean
- Compare meta-CBR with perfect classifier and ones that choose same algorithm each time

Example results

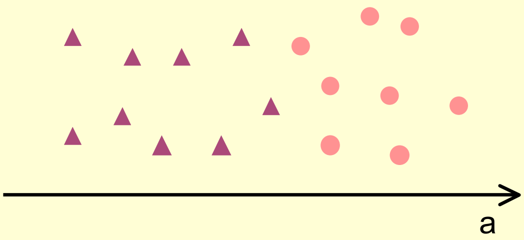
Classifier	Cases deleted (%)	Accuracy (%)	Harmonic mean
Choose-best	72.37	71.86	69.56
Meta-CBR	66.32	70.76	63.98
Choose ICF	64.54	69.63	62.29
Choose CBE	57.11	72.64	60.41

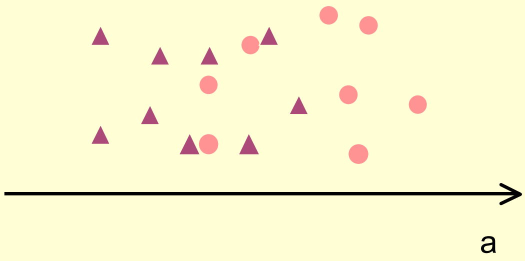
Which measures get selected?



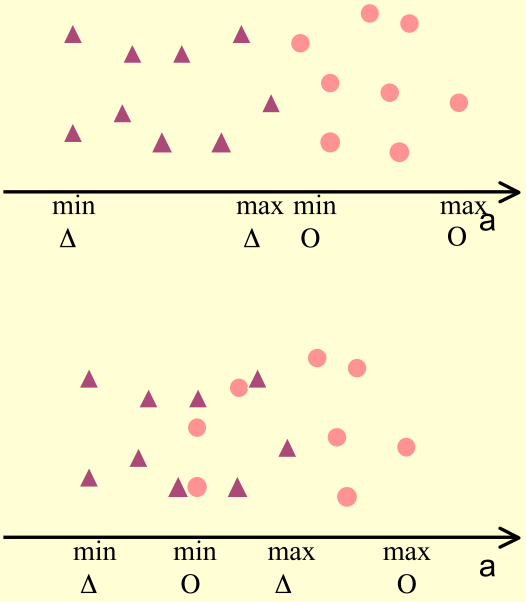
F_2' *Volume of Overlap Region*

- For a given attribute, a measure of how many values for that attribute appear in instances labelled with different classes

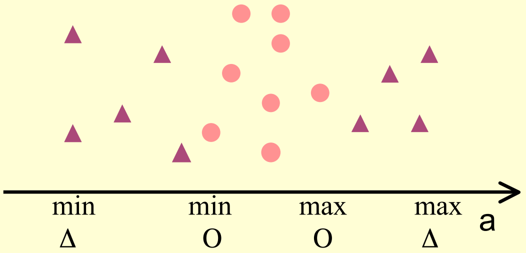




Quick computation of F_2



A problem for F_2



F_2' *Our version*

- o'(a) = count how many values are in the overlap
- r'(a) = count the number of values of a

$$F2' = \prod_{i=1}^{n} \frac{o'(a_i)}{r'(a_i)}$$

Summary of experiment

- Feature selection
  - chose between 2 and 18 attributes, average 9.2
  - chose range of measures, across Ho & Basu’s categories
  - always at least one measure of overlap of attribute values, e.g. F_2'
  - but measures of class separability only about 50% of the time
- But this is just one experiment

Overview

- ✓ Dataset complexity measures
- ✓ Classification experiment
- ✓ Case base maintenance experiment
- ➢ Going forward

Going forward

- Use of complexity measures in CBR (and ML)
- More research into complexity measures:
  - experiments with more datasets, different datasets, more classifiers,…
  - new measures, e.g. Information Gain
  - applicability of measures
    - missing values
    - loss functions
  - dimensionality reduction, e.g. PCA
  - the CBR similarity assumption and measures of case alignment [Lamontagne 2006, Hüllermeier 2007, Raghunandan et al. 2008]