Learning Decision Trees Using the Area Under the ROC Curve

Cèsar Ferri [1], Peter Flach [2], José Hernández-Orallo [1]

[1] Dep. de Sist. Informàtics i Computació, Universitat Politècnica de València, Spain
[2] Department of Computer Science, University of Bristol, UK

The 19th International Conference on Machine Learning, Sydney, Australia, 8-12 July 2002

Evaluating classifiers

- Accuracy/error is not a good evaluation measure of the quality of classifiers when:
  - the proportion of examples of one class is much greater then the other class(es). A trivial classifier always predicting the majority class may become superior.
  - not every misclassification has the same consequences (cost matrices). The most accurate classifier may not be the one that minimises costs.

- Conclusion: accuracy is only a good measure if the class distribution on the evaluation dataset is meaningful *and* if the cost matrix is uniform.

Evaluating classifiers

- Problem. We usually don't know a priori:
  - the proportion of examples of each class in application time.
  - the cost matrix.

- ROC analysis can be applied in these situations. Provides tools to:
  - Distinguish classifiers that can be discarded under any circumstance (class distribution or cost matrix).
  - Select the optimal classifier once the cost matrix is known.

Evaluating classifiers. ROC Analysis

- Given a confusion matrix:

Real

Predicted	Yes	No
Yes	30	20
No	10	40

- We can normalise each column:

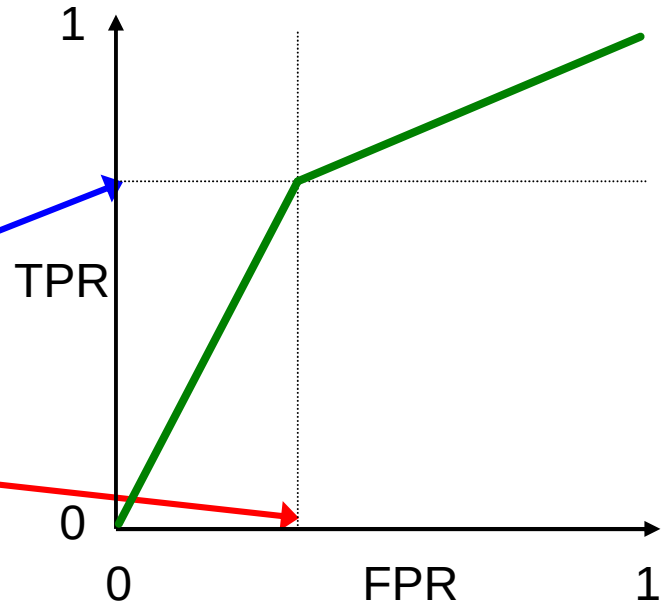
Real

Predicted	Yes	No
Yes	0.75	0.33
No	0.25	0.67

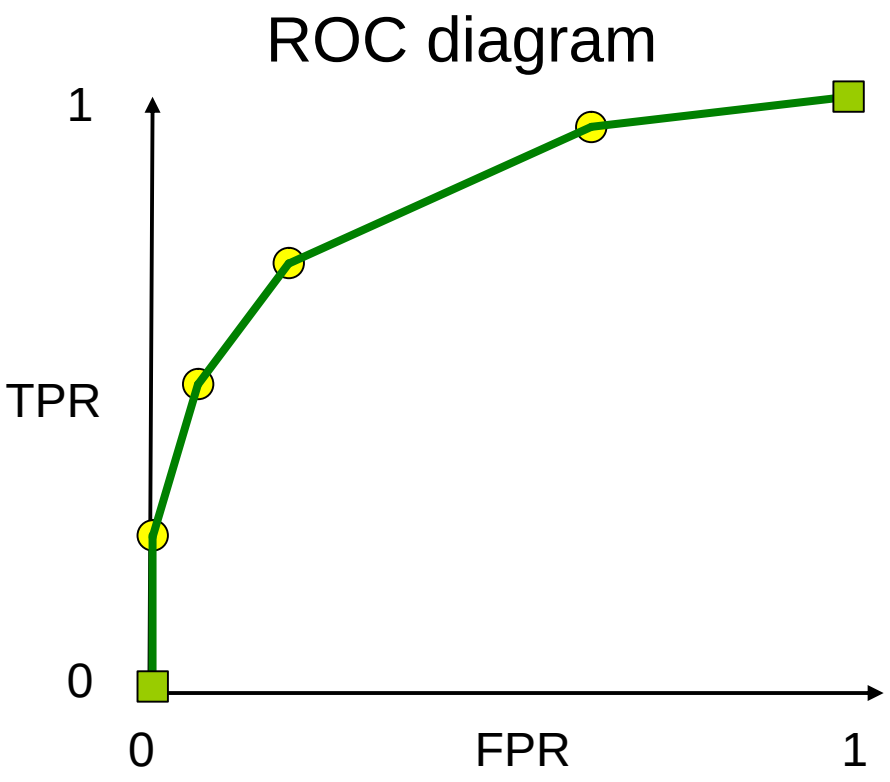
TPR

FPR

ROC diagram

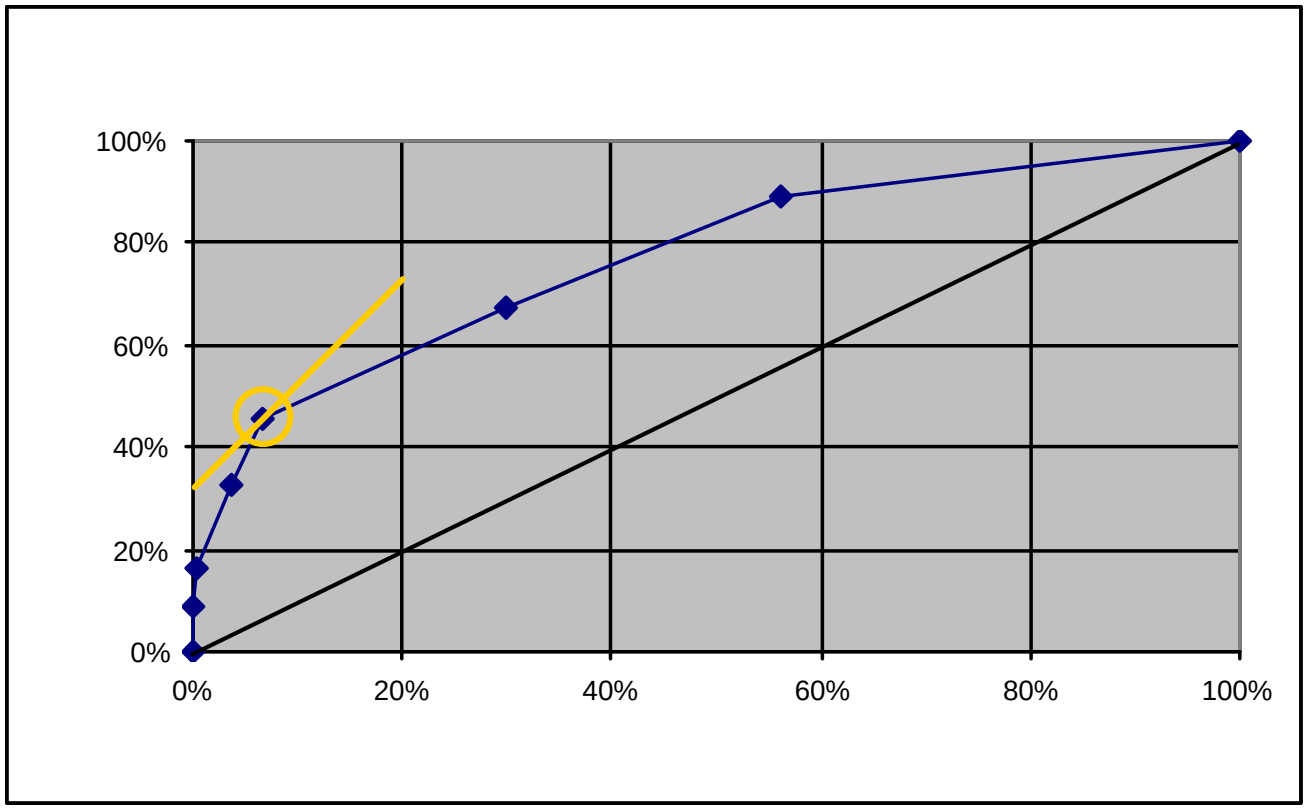


Evaluating classifiers. ROC Analysis



- Given several classifiers:
  - We can construct the convex hull of their points (FPR,TPR) and the trivial classifiers (0,0) and (1,1).
  - The classifiers falling under the ROC curve can be discarded.
  - The best classifier of the remaining classifiers can be chosen in application time...

Choosing a classifier. ROC Analysis

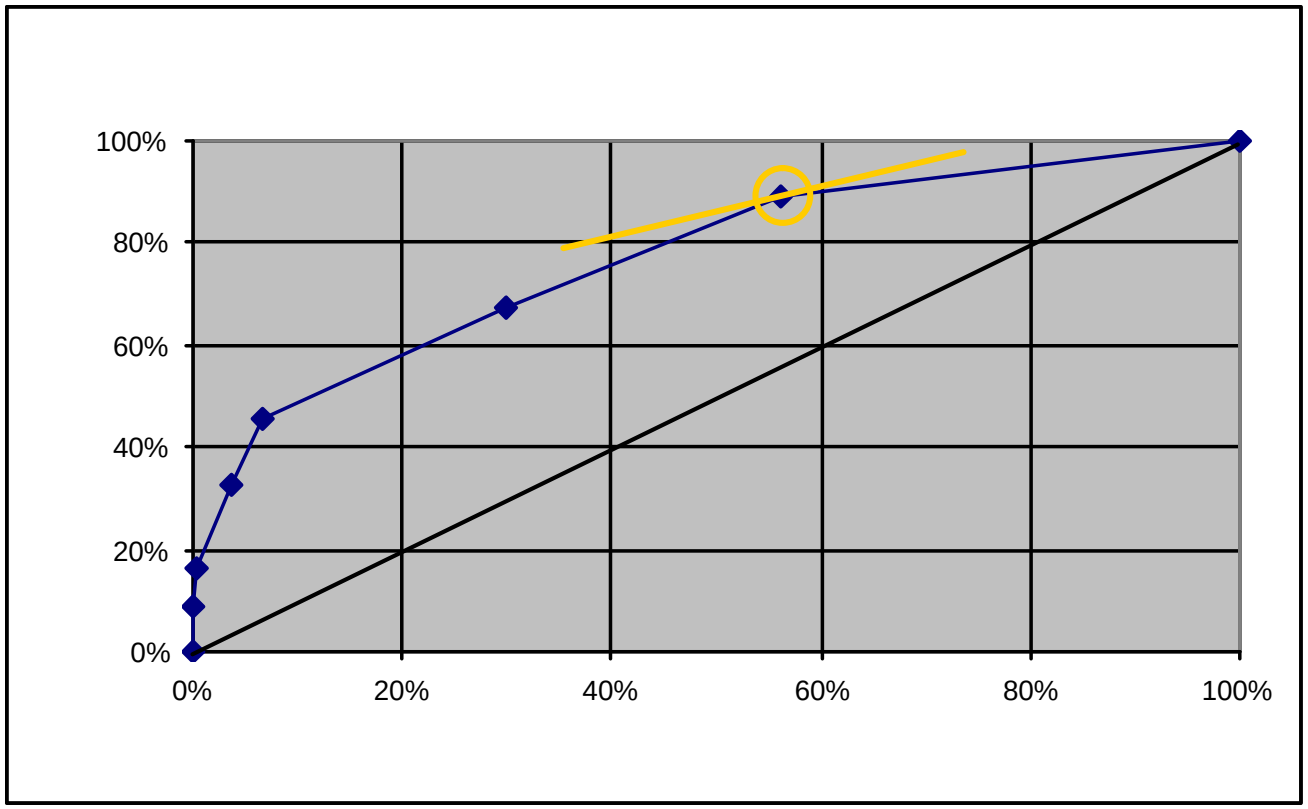


$$\frac{FPcost}{FNcost} = 1/2$$

$$\frac{Neg}{Pos} = 4$$

$$slope = 4/2 = 2$$

Choosing a classifier. ROC Analysis



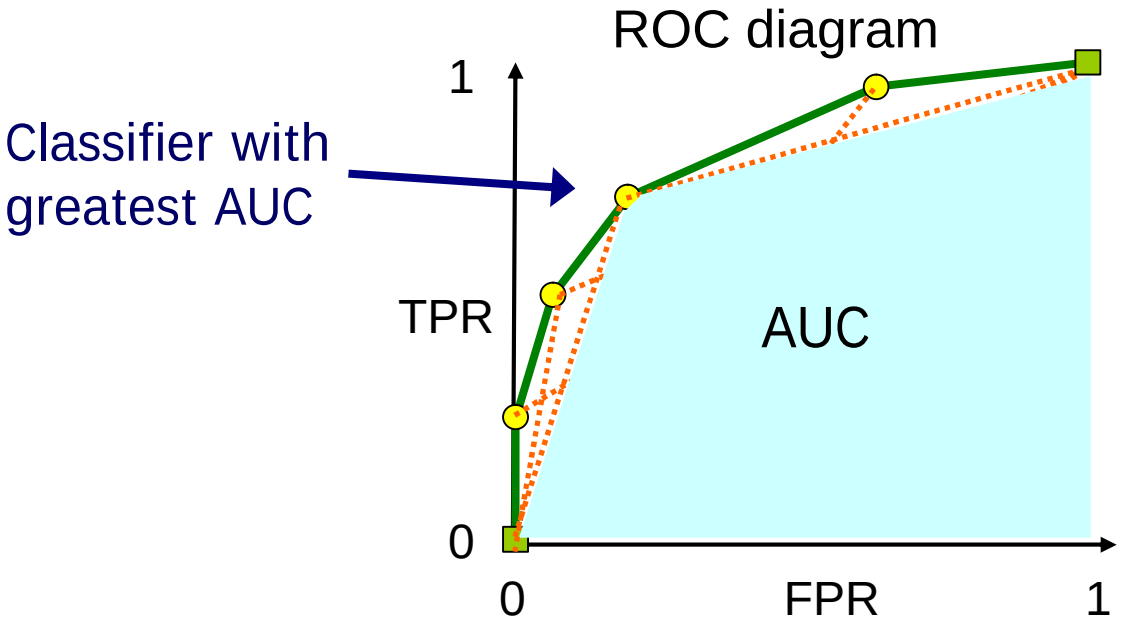
$$\frac{FPcost}{FNcost} = 1/8$$

$$\frac{Neg}{Pos} = 4$$

$$slope = 4/8 = .5$$

Choosing a classifier. ROC Analysis

- If we don't know the slope (expected class distribution)...



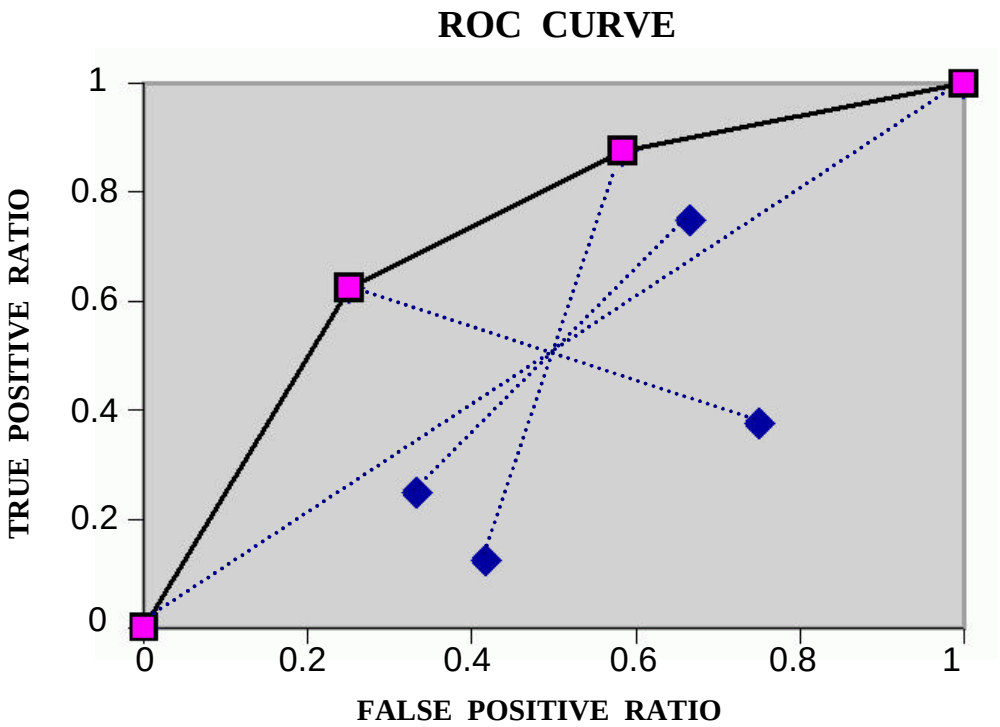
- The Area Under the Curve (AUC) can be used as a metric for comparing classifiers.

ROC Decision Trees

- A decision tree can be seen as an unlabelled decision tree (a clustering tree):
  - Given *n* leaves and 2 classes, there are 2^n possible labellings.
  - Clearly, each of the 2^n possible labellings of the *n* leaves of a given decision tree represents a classifier
  - We can use ROC analysis to discard some of them!

	Training Distribution		**Labellings**							
	T	F	1	2	3	4	5	6	7	8
Leaf 1	4	2	F	F	F	F	T	T	T	T
Leaf 2	5	1	F	F	T	T	F	F	T	T
Leaf 3	3	5	F	T	F	T	F	T	F	T

ROC Decision Trees



- Many labellings are under the convex hull.
- There is a special symmetry around (0.5,0.5).

- This set of classifiers has special properties which could allow a more direct computation of the optimal labellings.

ROC Decision Trees. Optimal Labellings

- Given a decision tree for a problem with 2 classes formed by n leaves $\{l_1, l_2, ..., l_n\}$ ***ordered*** by local positive accuracy, i.e, $r_1 \geq r_2, ..., r_{n-1} \geq r_n$, we define <u>the set of optimal labellings</u> $\Gamma = \{S_0, S_1, ... S_n\}$ where each labelling S_i, $0 \leq i \leq n$, is defined as: $S_i = \{A^1_i, A^2_i, ..., A^n_i\}$ where $A^j_i = (j,+)$ if $j \leq i$ and $A^j_i = (j,-)$ if $j > i$.

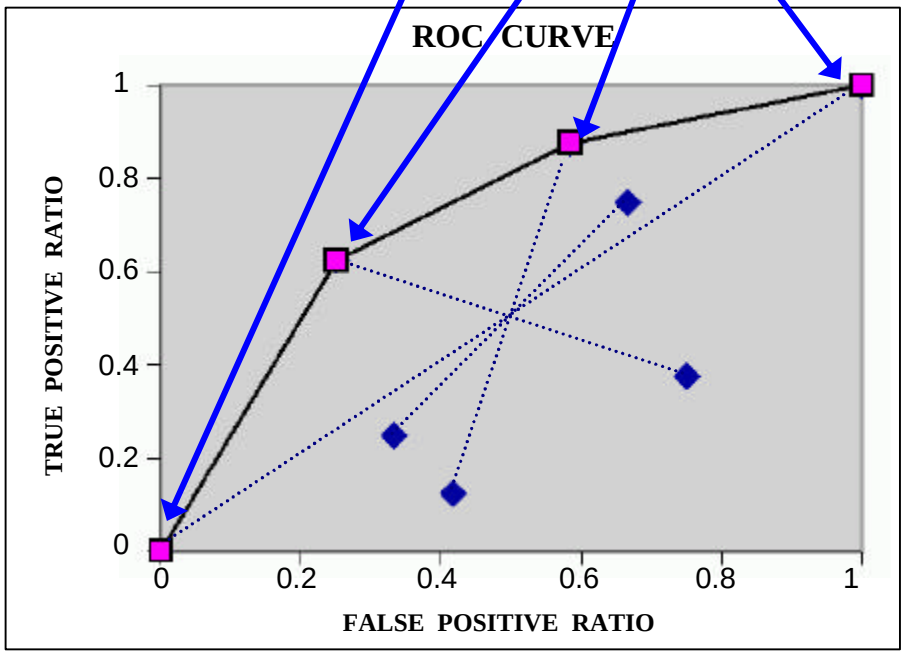
- **Theorem**: The convex hull corresponding to the 2^n possible labellings is formed **by and only by** all the ROC points corresponding to the set of optimal labellings Γ, removing repeated leaves with the same local positive accuracy.

Example

- We first order the leaves and then use only the optimal labellings:

	T	F				
Leaf 1	5	1	F	T	T	T
Leaf 2	4	2	F	F	T	T
Leaf 3	3	5	F	F	F	T

- That matches exactly with the convex hull:



ROC Decision Trees. Optimal Labellings

- Advantages:
  - Only n+1 labellings must be done (instead of 2^n).
  - The convex hull need not be computed.
  - The AUC is much easier to be computed: O(n log n).

- The AUC measure can be easily computed for unlabelled decision trees.
- Decision trees can be compared using it, instead of using accuracy.

Why don't we use this measure during decision tree learning?

AUC Splitting Criterion

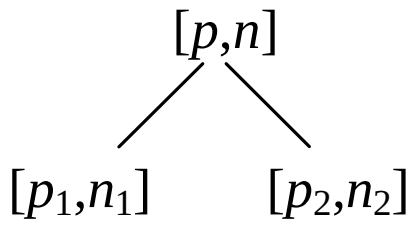
- **AUCSplit**:
  - Given a split *s* when growing the tree, we can compute the ordering of these leaves and calculate the corresponding ROC curve.
  - The area under this curve can be compared to the areas of other splits in order to select the best split.

AUC Splitting Criterion

- **AUCSplit vs. standard splitting criteria**:
  - Standard splitting criteria compare impurity of parent with weighted average impurity of children.

$$I(s) = \sum_{j=1..n_j} p_j \cdot f(p_j^+, p_j^-)$$

  - AUC is an alternative not based on impurity.
    - Example for 2 children:



$$AUCsplit = \frac{1}{2}\left(\frac{p_1}{p} - \frac{n_1}{n} + 1\right) = \frac{p_1 n + p n_2}{2pn}$$

Experiments

- Methodology:
  - 25 binary datasets UCI.
  - PEP Pruning.
  - 10-fold cross-validation.

- First we examine which is the best classical splitting criterion wrt. The AUC measure:

#	Gain Ratio	Gini	DKM	EErr
1	81.5 ± 14.0	79.8 ± 11.9	79.8 ± 11.9	82.2 ± 5.3
2	60.6 ± 10.4	57.7 ± 8.4	55.5 ± 7.9	69.8 ± 4.1
3	98.8 ± 1.6	98.7 ± 1.7	98.7 ± 1.7	95.4 ± 2.6
4	81.3 ± 8.0	80.6 ± 7.5	79.8 ± 8.1	76.4 ± 5.6
5	96.9 ± 2.5	96.9 ± 2.5	96.9 ± 2.5	96.9 ± 2.5
6	1 ± 0	99.9 ± 0.2	1 ± 0	1 ± 0.1
7	91.1 ± 6.6	90.9 ± 5.8	95.7 ± 5.3	93.6 ± 3.7
8	58.1 ± 24.4	66.4 ± 18.3	54.9 ± 18.6	51.2 ± 3.5
9	88.8 ± 10.2	56.1 ± 13.6	90.8 ± 5.0	59.0 ± 15.1
10	65.1 ± 6.7	63.4 ± 8.2	65.6 ± 8.4	59.9 ± 9.4
11	78.0 ± 5.2	27.8 ± 3.5	69.3 ± 25.7	30.5 ± 39.8
12	99.7 ± 0.4	99.3 ± 0.4	99.7 ± 0.3	98.3 ± 0.8
13	60.6 ± 10.2	69.7 ± 10.4	72.7 ± 6.8	68.1 ± 12.8
14	95.5 ± 2.5	95.2 ± 2.7	96.8 ± 2.1	94.8 ± 2.9
15	92.9 ± 12.4	65.4 ± 24.4	72.9 ± 26.3	65 ± 24.2
16	83.2 ± 16.5	48.6 ± 51.2	96.9 ± 5.7	34.8 ± 41.1
17	93.6 ± 3.2	49.7 ± 46.1	65.8 ± 45.5	3.7 ± 11.3
18	50.5 ± 25.9	48.9 ± 27.1	52.5 ± 24.5	21.5 ± 21.4
19	98.1 ± 0.7	98.2 ± 0.8	98.1 ± 0.8	97.8 ± 1.1
20	1 ± 0	1 ± 0	1 ± 0	1 ± 0
21	99.7 ± 0.6	98.2 ± 0.7	99.7 ± 0.3	96.3 ± 2.1
22	93.7 ± 3.7	81.7 ± 4.9	66.6 ± 21.6	50 ± 0
23	73.7 ± 3.1	66.6 ± 9.9	73.5 ± 4.3	51.0 ± 4.0
24	98.7 ± 1.0	95.9 ± 2.4	99.4 ± 0.5	85.7 ± 0.5
25	98.1 ± 2.3	95.9 ± 3.3	98.0 ± 2.6	96.0 ± 3.3
M	85.53	77.26	83.19	71.12

Experiments

- Methodology:
  - 25 binary datasets UCI.
  - PEP Pruning.
  - 10x10-fold cross-validation.
  - ✓ when differences are significant with t-test at 0.1.
- Next we compare the best classical splitting criterion with the AUCsplit:

	Gain Ratio		AUCsplit		Better?	
Set	Acc.	AUC	Acc.	AUC	Acc.	AUC
1	90.7±6.6	83.6±11.8	96.5±3.9	94.3±6.7	✓	✓
2	57.7±6.5	61.1±7.9	56.0±6.2	56.7±8.0	x	x
3	97.6±7.8	97.4±8.5	99.1±1.1	99.1±1.4	✓	✓
4	78.9±4.6	79.8±7.2	77.6±4.7	76.9±6.5	x	x
5	95.8±2.6	95.2±3.1	95.8±2.6	95.2±3.1		
6	1±0	1±0	1±0	1±0		
7	92.5±4.1	91.5±6.1	92.9±3.7	94.7±4.6		✓
8	72.1±10.2	61.3±16.9	69.5±10.6	59.3±16.2	x	
9	92.0±4.7	90.4±7.0	89.6±5.0	89.7±6.7	x	
10	62.6±8.8	64.2±10.6	64.0±9.0	65.8±10.1		
11	73.3±5.7	76.6±6.9	72.5±5.1	76.7±6.0		
12	99.1±2.3	99.5±1.6	99.2±0.6	99.5±0.6		
13	68.2±10.2	67.4±11.9	71.0±10.4	73.6±11.0	✓	✓
14	95.4±2.5	96.3±2.5	96.2±2.5	97.6±2.1	✓	✓
15	86.4±14.2	85.1±17.9	83.4±14.0	63.5±22.3		x
16	98.0±10.9	84.6±13.1	98.6±0.8	94.8±5.6	✓	✓
17	95.2±1.4	92.6±3.5	96.7±1.2	95.1±3.1	✓	✓
18	71.4±12.4	61.5±20.8	68.9±11.6	59.8±21.3		
19	95.0±1.8	98.2±0.9	94.8±1.9	98.1±1.0		
20	1±0	1±0	1±0	1±0		
21	99.6±0.3	99.6±0.5	99.6±0.2	99.4±0.6		
22	96.8±0.9	93.3±4.7	96.8±0.2	95.1±6.9		✓
23	70.4±3.9	72.2±4.9	71.1±3.6	73.3±4.0		✓
24	99.5±0.2	98.9±1.4	99.5±0.1	99.3±0.7	✓	✓
25	98.9±1.8	94.2±19.4	99.5±0.3	98.5±1.8	✓	✓
M.	87.49	85.78	87.55	86.24		

Experiments

- Methodology:
  - 6 of 25 binary datasets UCI with % of minority class < 15%.
  - PEP Pruning.
  - 10x10-fold cross-validation.

- Finally we compare the results when class distribution changes.

#	Original Dist.		50%-50%		Swapped Dist.		%min class
	GR	AUCs.	GR	AUCs.	GR	AUCs.	
16	98.0	98.6	88.3	93.5	78.6	88.3	6.06
17	95.2	96.7	88.6	92.6	81.9	88.4	11.83
21	99.6	99.6	99.0	98.7	98.4	97.8	10.4
22	96.8	96.8	89.8	89.7	82.9	82.7	10.23
24	99.5	99.5	96.0	96.6	92.5	93.6	3.95
25	98.9	99.5	95.8	98.4	92.7	97.3	9.86
M.	98.0	98.5	92.9	94.9	87.8	91.4	

Conclusions and Future Work

- **Labelling classifiers**:
  - One classifier can be many classifiers!
  - Optimal labelling set identified (order by local positive accuracy)
  - An efficient way to compute the AUC of a set of rules.
- **AUCsplit criterion**:
  - Better results for the AUC measure
- **Future work:**
  - Extension of the AUC measure and AUCsplit for $c>2$. ✓
  - Global AUC splitting criterion.
  - Pre-pruning and post-pruning methods based on AUC.