Delegating Classifiers

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

1 *Dpt. de Sistemes Informàtics i Computació,*
Universitat Politècnica de València, Spain.

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

2 *Department of Computer Science,*
University of Bristol, UK.

University of BRISTOL

{ cferri@dsic.upv.es, Peter.Flach@bristol.ac.uk, jorallo@dsic.upv.es }

ICML-2004, Banff, July 5, 2004.

Outline

- Introduction.
- Delegation as Separate-and-Conquer
- Establishing the Thresholds
- Scenarios
- Experiments
- Discussion
- Conclusions and Future Work

Introduction

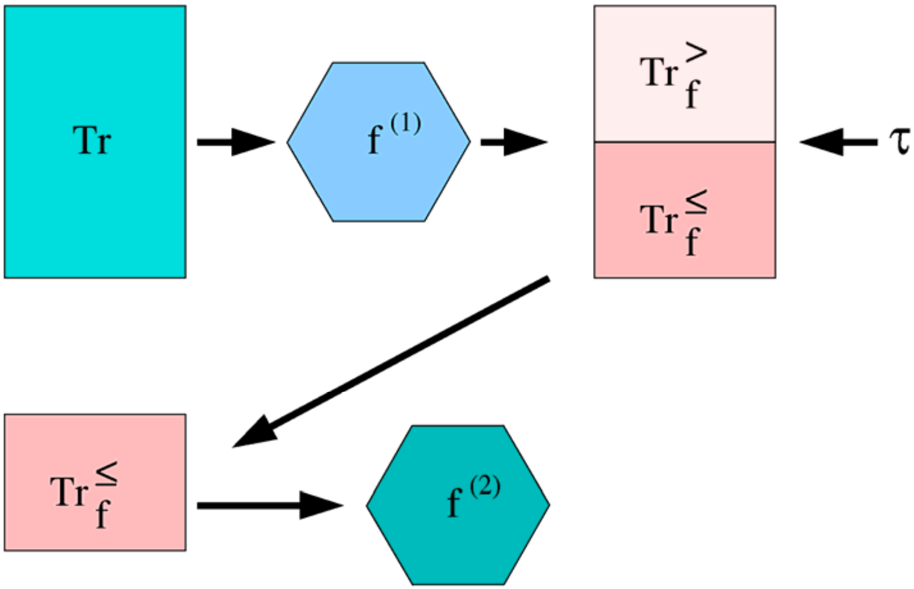
- Many collaborative views of learning:
  - combination (ensembles, stacking/cascading),
  - co-learning, ...
- Generally composed of “total”, non-specialised classifiers, usually under-utilised.
- Learner specialisation:

But who determines the areas of specialisation?
And how?

  - Pre-refereeing: meta-learning, analysis of separability, ...
  - Post-refereeing: stacking, cascading, arbitrating, grading, ...
  - Self-refereeing: ???

Introduction

- Delegation:



- Training set *Tr* is split using a threshold τ.
- The threshold τ defines the degree of self-confidence.

$$Tr_f^{>} = \{ e \in Tr : f_{CONF}(e) > \tau \}$$

$$Tr_f^{\leq} = Tr - Tr_f^{>}$$

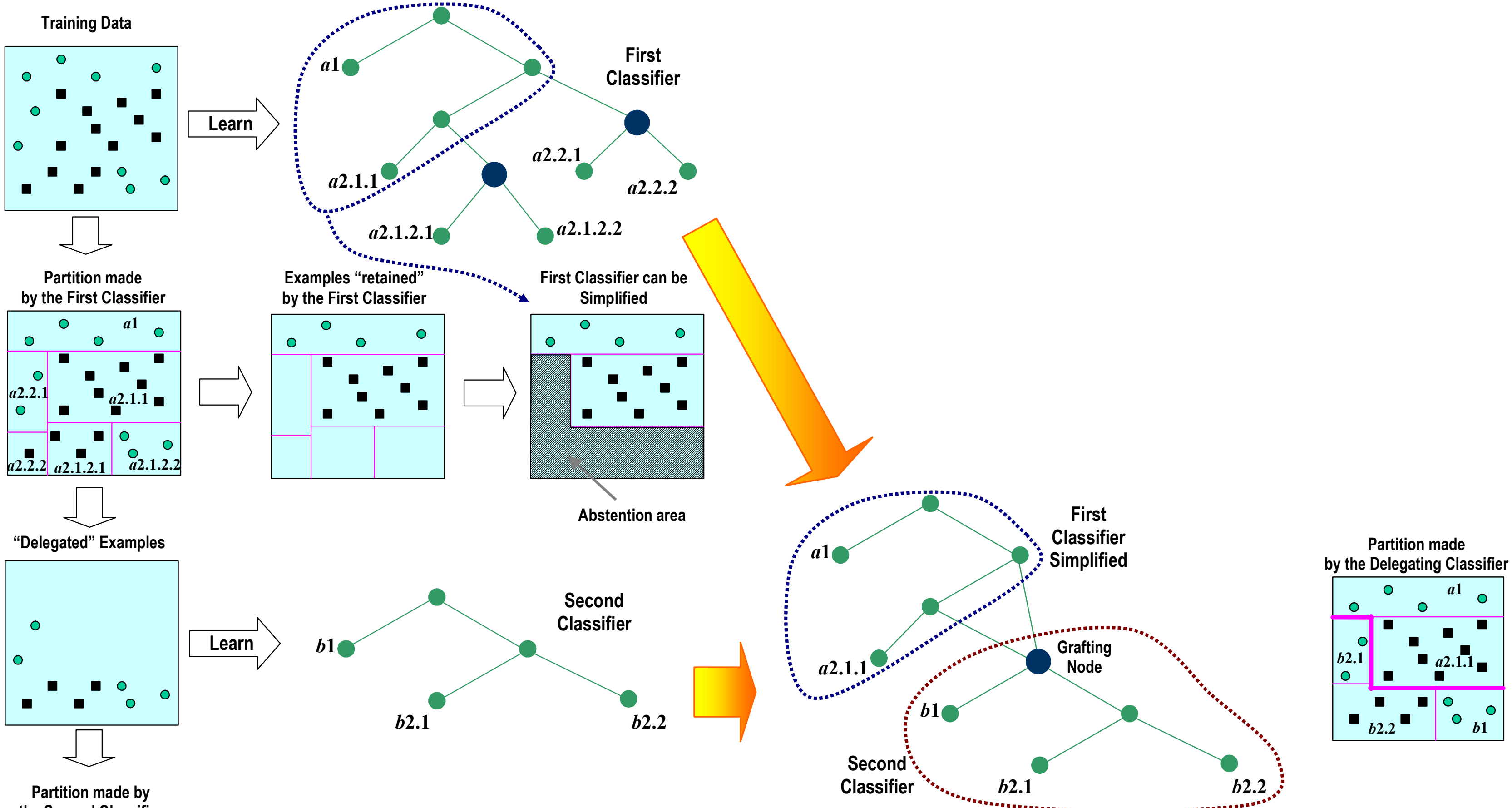
- The overall classifier is used as follows:

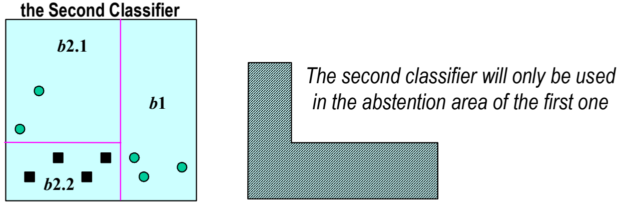
Decision Rule for a Delegating Classifier with threshold τ:
IF $f_{CONF}^{(1)}(e) > \tau$ THEN PREDICT $f_{CLASS}^{(1)}(e)$
ELSE PREDICT $f_{CLASS}^{(2)}(e)$

Introduction

- If the task is classification, it is a multi-classifier method.
  - *Self-refereeing*: each classifier self-assigns its area of expertise.
  - *Serial*: not parallel or hierarchical.
  - *Transferring*: each prediction is made by only one classifier (no combination).
  - *Specialised*: based on partial classifiers.
  - *Attribute-preserving*: no new attributes are generated.

Delegation as Separate-and-Conquer





- The resulting model can be comprehensible.

Establishing the Threshold

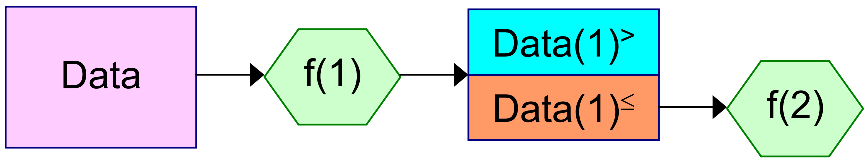
- We use the *same* threshold for prediction (and for the test set).

How do we determine this threshold?

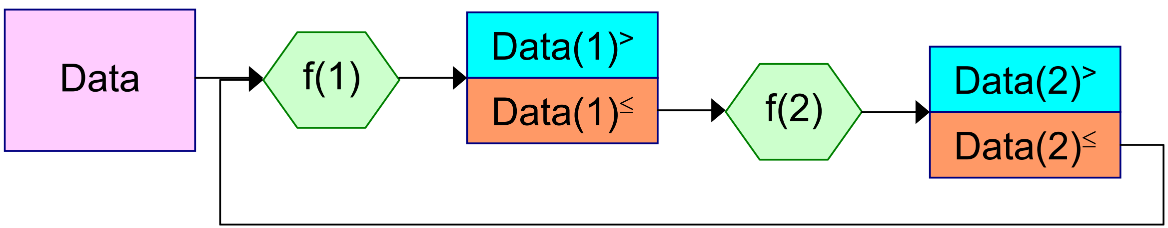
- This threshold is very dependent on the problem.
- Instead, we define a percentage of retention.
  - Two different ways:
    - Global Absolute Percentage (GAP): retain a fraction of the ρ best ranked examples.
    - Stratified Absolute Percentage (SAP): retain a fraction of the ρ best ranked examples *per class*.

Scenarios

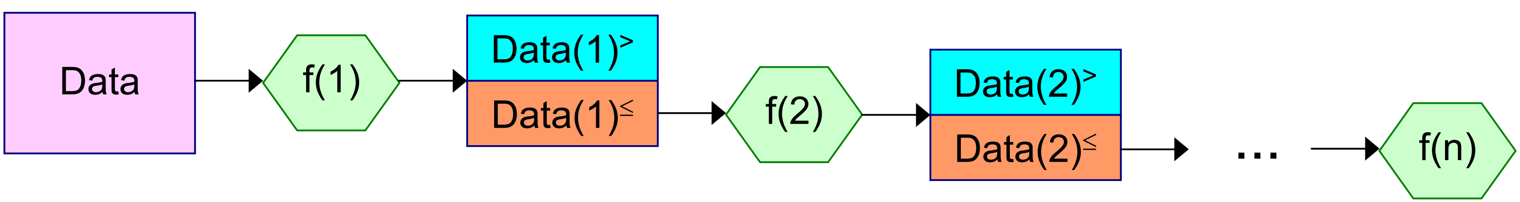
- We investigated three different scenarios:
  - Two stages: a master classifier and a slave classifier.



  - Two stages with "round rebound".



  - Iterative: several chained stages.



Experiments

- Experimental methodology:
  - 22 datasets from UCI repository
  - Trained PETs (Probability Estimation Trees):
    - Smiles and Weka J4.8 variants of C4.5.
    - Pruning disabled.
    - Probability smoothing.
  - Evaluation:
    - 20x5-fold cross-validation.
    - Accuracy and AUC used as metrics.

Experiments

- Importance of a good probability estimation.
  - Four methods of PETs:
    - With pruning and no smoothing (Pr NoSmooth)
    - No pruning and no smoothing (NoPr NoSmooth)
    - No pruning and Laplace smoothing (Pr Laplace)
    - No pruning and Mbranch smoothing (Pr Mbranch) (ECML'03)
  - Two-stage scenario. GAP (ρ=0.5).
    - Averaged results for the 22 datasets.

	Pr NoSmooth	NoPr NoSmooth	NoPr Laplace	NoPr Mbranch
Single Acc	83.88	83.72	83.72	83.72
Single AUC	86.46	87.16	90.18	90.78
Del50% Acc	84.01	83.81	84.77	84.73
Del50% AUC	85.93	87.16	90.89	91.31

- The way in which the master classifier is able to estimate its reliability is key to the success of the delegating method.

Experiments

- The proportion ρ and Global/Stratified.
  - Two-stage scenario. GAP and SAP. Varying proportions ρ.
    - Averaged results for the 22 datasets.

	None	20%	33%	45%	50%	55%	67%	80%
GAP Acc	83.72	84.23	84.13	84.67	84.73	84.72	84.61	84.60
GAP AUC	90.78	91.02	91.14	91.27	91.31	91.24	91.04	90.92
SAP Acc	83.73	84.29	84.42	84.37	84.34	84.32	84.48	84.32
SAP AUC	91.31	90.79	90.79	90.65	90.61	90.47	90.26	89.85

  - Proportions around 0.5 are optimal.
  - The improvement is obtained with just around a 50% overhead.
  - Stratified thresholds do not improve the results of global thresholds in general.

Experiments

- Iterative Scenario:
  - The greater the # of iterations the better the results.

	50%	33%	20%	10%	5%	2%	1%
GLOBAL ACC	84.73	85.20	85.33	85.64	85.82	85.85	85.93
GLOBAL AUC	91.31	91.40	91.43	91.61	91.75	91.82	91.82
GLOBAL #IT	2.00	3.16	4.68	7.74	12,50	21.64	31.31
STRAT. ACC	84.34	84.75	85.09	85.30	85.42	85.53	85.58
STRAT. AUC	90.61	90.70	90.74	91.02	91.25	91.44	91.51
STRAT. #IT	2.00	3.06	4.33	6.70	9.83	15.30	18.24

  - Once again, the “Global Absolute” variant is the best one.
  - Execution times for 2% and 1% are around 8 and 10 times higher, respectively, than a single classifier.
  - With similar times, delegation is close to bagging and not far behind boosting.

Discussion

- Factors that affect “Delegation”:
  - Reliability estimation (confidence) crucial.
  - Cannot be justified as a reduction of variance.
  - Patterns removed iteratively, as in Sep&Conq.
  - Class distribution is modified (better specialisation?).
  - Overfitting not so crucial as expected.

- Some of these factors may also explain a better improvement for accuracy than for AUC.

Conclusions and Future Work

- Delegation is a key idea in machine learning.
  - This work has used it systematically, using learners as building blocks for different scenarios.
  - As long as classifiers perform better probability estimation, they are more reliable for self-refereeing, crucial in delegation.
  - The method is simple, general and efficient.
  - In some configurations, it can preserve the comprehensibility of the base models by pruning and grafting them.
    - Only the useful parts are maintained.

Conclusions and Future Work

- Future work:
  - Use very efficient classifiers for the first stage and then more data-intensive ones for the subsequent stages.
  - Investigate the “combination” of the predictions.
  - Apply to regression and clustering.
  - Investigate other methods to determine the threshold (e.g. AUC-based, validation dataset, …).