From Ensemble Methods to Comprehensible Models

Cèsar Ferri, José Hernández-Orallo, M.José Ramírez-Quintana
{cferri, jorallo, mramirez}@dsic.upv.es
Dep. de Sistemes Informàtics i Computació,
Universitat Politècnica de València,
Valencia, Spain

The 5th International Conference on Discovery Science
Lübeck, 24-26 November 2002

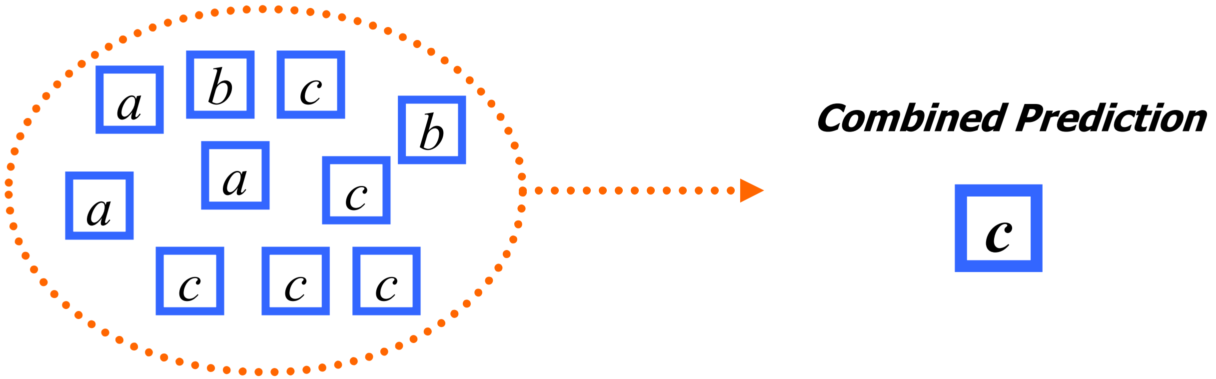
Introduction

- Machine Learning techniques that construct a model/hypothesis (e.g. ANN, DT, SVM, ...):
  - usually devoted to obtain **one** single model:
    - As accurate as possible (close to the "target" model).
    - Other (presumably less accurate) models are discarded.
  - An old alternative has recently been popularised:
    - "Every consistent hypothesis should be taken into account"

But... How?

Ensemble Methods (1/3)

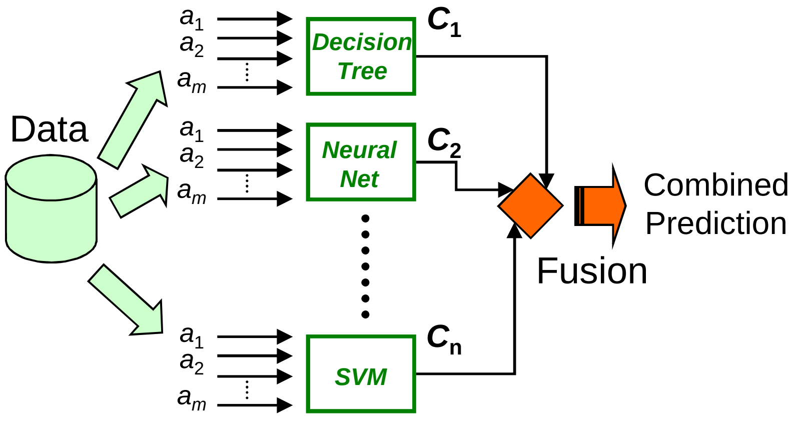
- Ensemble Methods (Multi-classifiers):
  - Generate multiple (and possibly) heterogeneous models and then combine them through *voting* or other fusion methods.



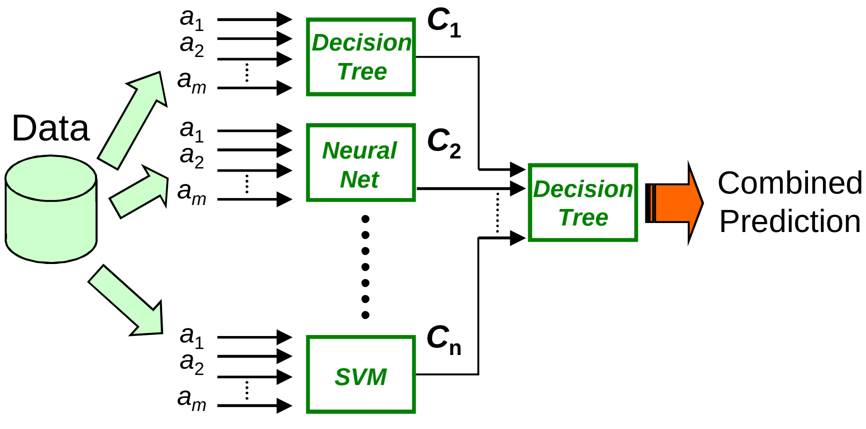
  - Much better results (in terms of accuracy) than single models when the number and variety of classifiers is high.

Ensemble Methods (2/3)

- Ensemble Methods (Multi-classifiers):
  - Different topologies: simple, stacking, cascading, …



Simple Combination



Stacking

  - Different generation policies: *boosting, bagging, randomisation, …*
  - Different fusion methods: majority voting, average, maximum, …

Ensemble Methods (3/3)

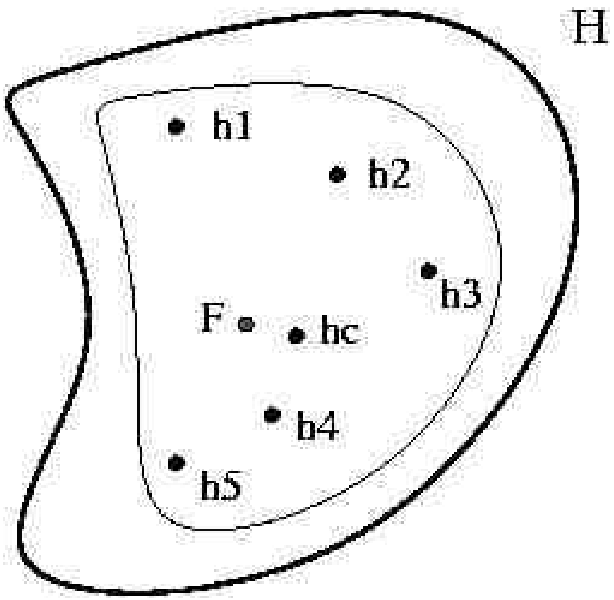
- Main drawbacks:
  - **Computational costs**: huge amounts of memory and time are required to obtain and store the set of hypotheses (ensemble).
  - **Throughput**: the application of the combined model is slow.
  - **Comprehensibility**: the combined model behaves like a black box.

The solution of these drawbacks would boost the applicability of ensemble methods in machine learning applications.

Archetype (1/2)

- The question is to reduce to one hypothesis from the combination of *m* hypotheses without losing too much accuracy.
- One possibility is to select one hypothesis according to the semantic similarity to the combined hypothesis

Archetype (2/2)



- The intuitive idea is to select the component of the ensemble closest to the the combined hypothesis

Hypotheses similarity

- Measures of similarity of hypothesis must be considered:
- Given two classifiers, an unlabelled dataset of *n* examples, with *C* classes, we can construct a *C x C* confusion or contingency matrix $M_{i,j}$

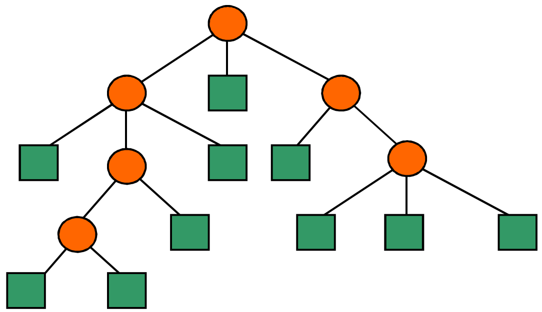
$$\theta = \sum_{i=1}^{C} \frac{M_{i,i}}{n} \qquad \kappa = \frac{\theta - \theta_2}{\theta - 1} \qquad Q = \frac{\prod_{i=1}^{C} M_{i,i} - \prod_{i=1, j=1, i \neq f}^{C} M_{i,j}}{\prod_{i=1}^{C} M_{i,i} + \prod_{i=1, j=1, i \neq f}^{C} M_{i,j}}$$

Random Invented Dataset

- We require a dataset to establish the similarity between the hypotheses
- We could employ a subset of the training dataset as *validation dataset*
- A better possibility is the generation randomly of an unlabelled *invented dataset*

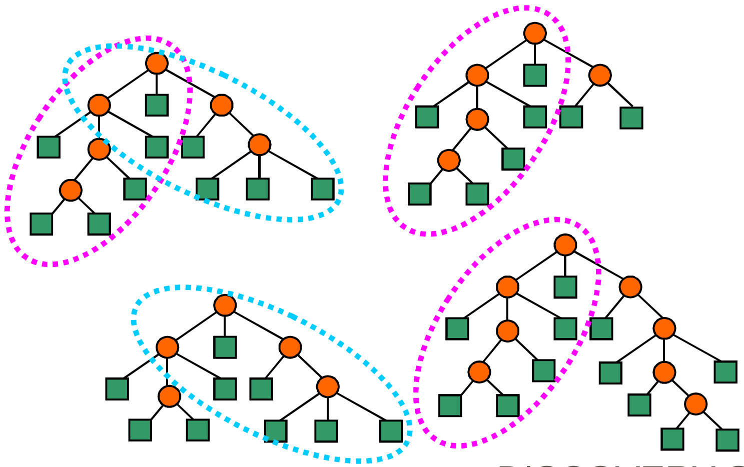
Ensembles of Decision Trees

- Decision Tree:



  - Each internal node represents a condition.
  - Each leaf assigns a class to the examples that fall under that leaf.

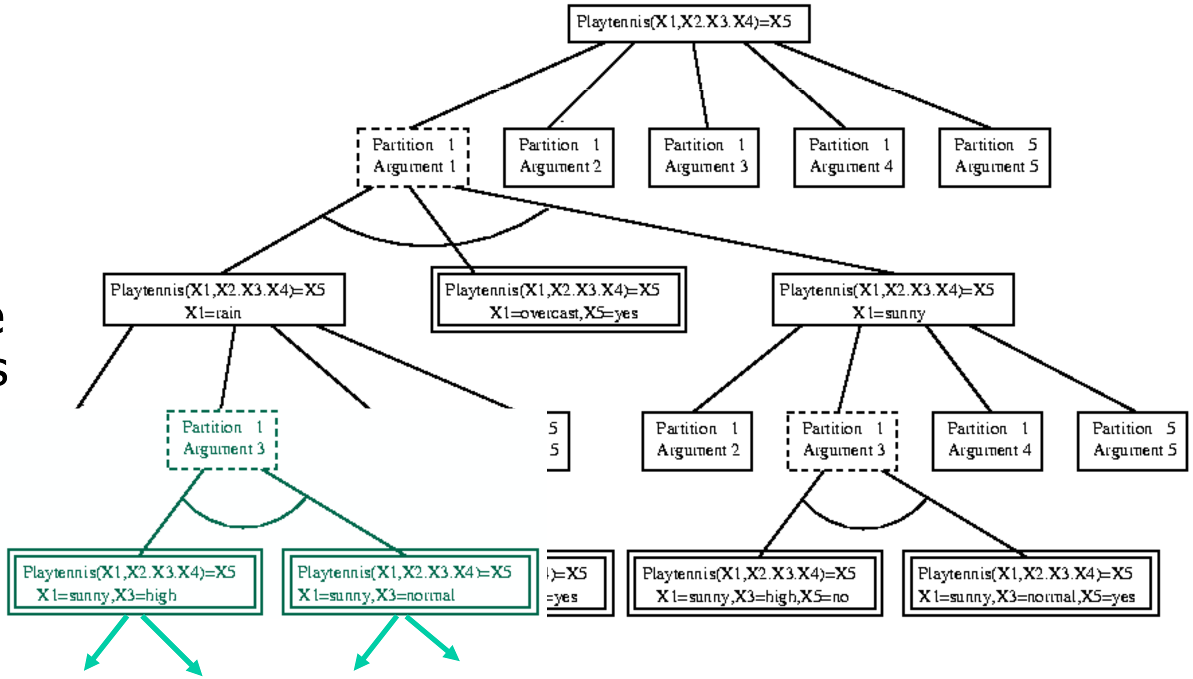
- Forest: several decision trees can be constructed.



  - Many trees have common parts.
  - Traditional ensemble methods repeat those parts:
    - memory and time ↑↑↑

Decision Tree *Shared* Ensembles

- Shared ensemble:
  - Common parts are shared in an AND/OR tree structure.



- Construction space and time resources are highly reduced
- Throughput is also improved by this technique.

Decision Tree *Shared* Ensembles

- Previous work:
  - Multiple Decision Trees (Kwok & Carter 1990)
  - Option Decision Trees (Buntine 1992)
    - The AND/OR tree structure is populated (partially) breadth-first.
  - Combination has been performed:
    - Using weighted combination (Buntine 1992).
    - Using majority voting combination (Kohavi & Kunz 1997).
  - Different conclusions on where alternatives are especially beneficial:
    - At the bottom of the tree (Buntine).
      - Trees are quite similar → Accuracy improvement is low.
    - At the top of the tree (Kohavi & Kunz).
      - Trees share few parts → Space resources are exhausted as in other non-shared ensembles (boosting, bagging, ...).

Multi-tree Construction

- New Way of Populating the AND/OR Tree:
  - The first tree is constructed in the classical eager way.
  - Discarded alternative splits are stored in a list.
    - Repeat *n* times:
      - Once a tree is finished, the best alternative split (according to a “wakening” criterion) is chosen.
      - The branch is finished using the classical eager way.
  - This amounts to a ‘beam’ search → Anytime algorithm.
    - Extensions and alternatives can happen at any part of the tree (top, bottom).
    - The populating strategy can be easily changed.
    - The fusion strategy can also be flexibly modified.
    - The desired size of the AND/OR tree can be specified quite precisely.

Fusion Methods

- Combination on the Multi-tree:
  - The number of trees grows exponentially w.r.t. the number of alternative OR-nodes explored:
    - Advantages: ensembles are now much bigger with a constant increase of resources. Presumably, the combination will be more accurate.
    - Disadvantages: the combination at the top is unfeasible.
  - Global fusion techniques would be prohibitive.

Local Fusion

- First Stage. Classical top-down:
  - Each example to be predicted is distributed top-down into many alternative leaves.
  - The example is labelled in each leaf (*class vector*).
- Second Stage. The fusion goes bottom-up:
  - Whenever an OR-node is found. The (possibly) inconsistent predictions are combined through a *local fusion method*:

- Fusion of millions or billions of trees can be performed efficiently.

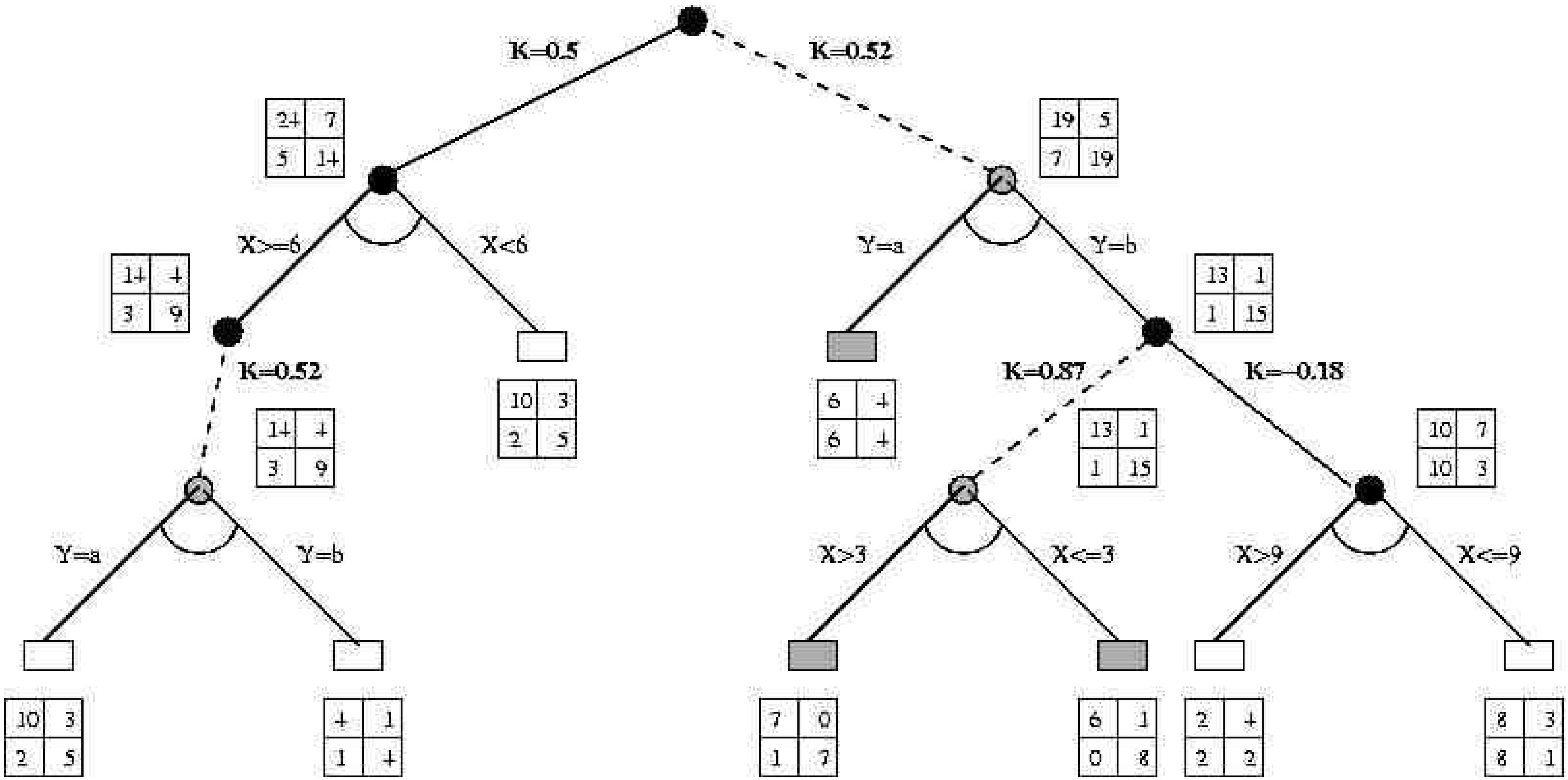
Selection of an Archetype (1/3)

- Due to huge amount of hypotheses is could be not feasible to compute the similarity of each hypothesis with respect to the combined hypothesis
- We could make compute the similarity for each internal node, and then extract the most similar solution

Selection of an Archetype (2/3)

1. We label the invented dataset w.r.t. the combined hypothesis
2. We fill a contingency matrix M in each leaf of the multi-tree according with the labeled invented dataset
3. We propagate upwards the contingency matrix:
   - For the AND-nodes we accumulate the contingency matrix of their children nodes: $M = M_1 + M_2 + .. + M_i$
   - For the OR-nodes, we compute a similarity measure of their children, and the M of the node with highest similarity is copied in the AND node. The selected node is marked

Selection of an Archetype (3/3)



Archetype Technique

1. **Multi-tree generation**: The first step consists in the generation of a multi-tree from a training dataset.
2. **Invented dataset**: In this phase, an unlabelled invented dataset is created, by a random dataset
3. **Multi-tree combination**: The invented dataset is labelled by the combination of the shared ensemble
4. **Calculation and propagation of contingency matrices**: A contingency matrix is assigned to each node of the multi-tree, using the labelled invented dataset and a similarity metric.
5. **Selection of a solution**: An archetype hypothesis is extracted from the multi-tree by descending the multi-tree through the marked nodes.

Experiments (1/4)

- Experimental setting:
  - 13 datasets from the UCI repository.
  - Multi-tree implemented in the **SMILES** system.
  - Splitting criterion: GainRatio (C4.5).
  - Second node selection criterion (wakening criterion): random.

Experiments (2/4)

- Evaluating Similarity Metrics

#	1st	Comb	Arc. κ	Arc. θ	Arc. Q
1	92.3	100	100	100	100
2	74.8	77.4	76.1	76.2	75.8
3	97.5	97.5	97.6	97.6	97.6
4	78.2	82.7	78.2	78.3	78.5
5	93.6	96.0	94.4	93.9	94.2
6	60.9	66.3	63.8	64.3	61.9
7	76.8	83.1	80.1	80.1	79.8
8	97.3	96.5	96.5	91.0	47.0
9	89.8	93.6	90.6	89.9	74.3
10	89.0	91.0	89.6	89.6	89.3
11	62.9	64.5	61.9	62.9	49.8
12	92.6	92.6	92.8	92.9	91.4
13	77.5	79.9	79.4	78.9	76.7
gmeans	82.41	85.45	83.78	83.45	76.24

Experiments (3/4)

- Influence of the Size of the Invented Dataset:

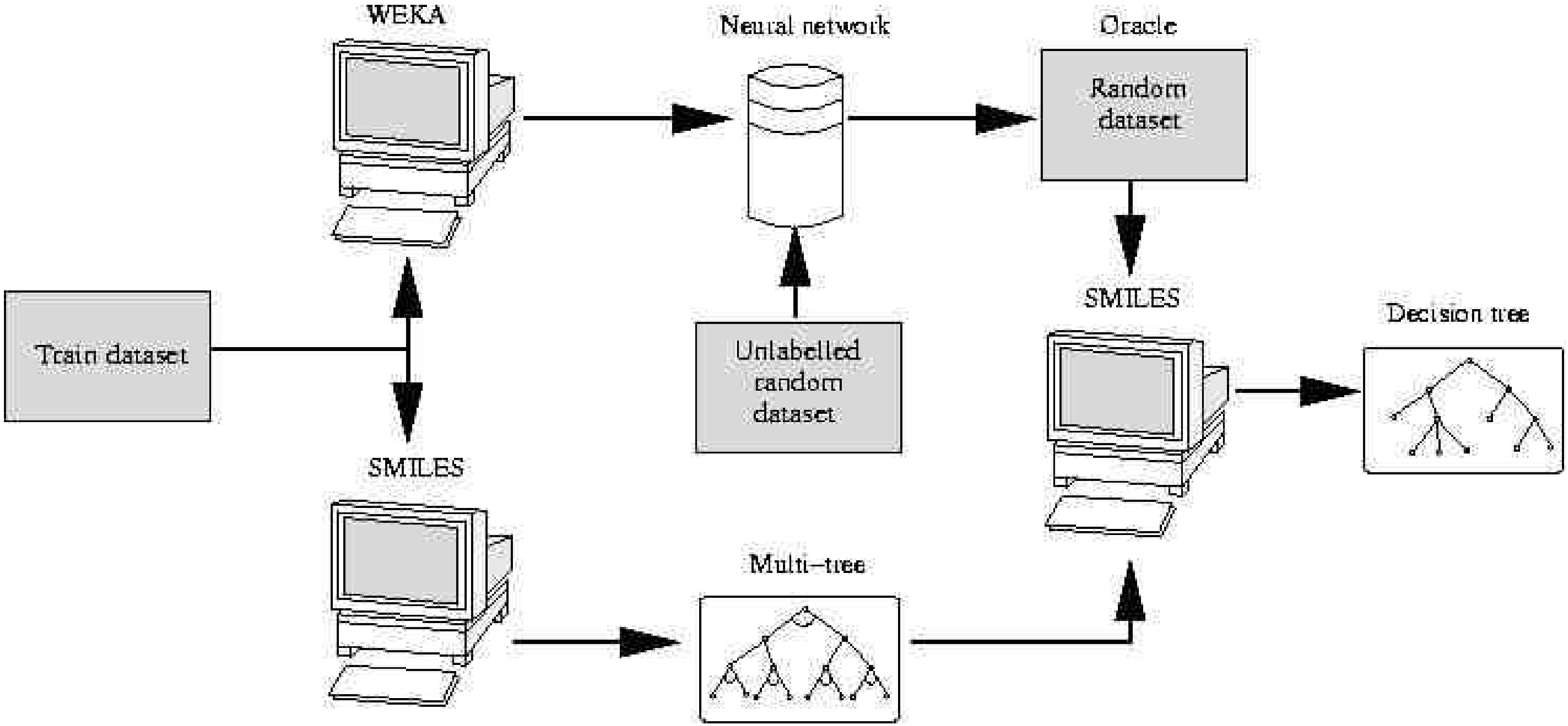
		10	100	1000	10000	100000
#	Comb	Arc	Arc	Arc	Arc	Arc
1	99.8	72.3	93.3	99.8	100	99.9
2	77.3	64.6	61.0	75.2	76.1	76.2
3	97.6	82.9	94.5	97.6	97.6	97.6
4	82.9	65.9	70.3	78.0	78.2	78.6
5	95.8	73.7	92.4	94.4	94.4	93.8
6	67.5	69.1	63.6	63.9	63.8	63.5
7	83.0	62.5	75.4	79.4	80.1	79.9
8	95.0	68.8	93.3	95.0	96.5	96.5
9	93.6	45.6	84.7	90.5	90.6	89.9
10	91.0	71.0	75.4	88.1	89.6	89.8
11	63.7	44.3	54.3	59.1	61.9	61.2
12	92.5	73.8	89.3	91.3	92.8	92.6
13	80.0	46.8	73.9	77.9	79.4	79.0
gmeans	85.36	63.57	77.40	82.88	83.78	83.57

Experiments (4/4)

- Combination Resources compared to other Ensemble Methods:

	1	10				100				1000			
#	1st	Comb	Arc	Occ	#Sol	Comb	Arc	Occ	#Sol	Comb	Arc	Occ	#Sol
1	92.3	96.1	96.0	96.5	107	100	100	100	8.7×10^{8}	100	100	100	1.6×10^{19}
2	74.8	74.9	74.3	74.3	148	77.4	76.1	72.5	2.6×10^{10}	82.3	82.1	70.4	3.2×10^{20}
3	97.5	97.7	97.7	97.6	46	97.5	97.6	97.5	80×10^{4}	97.7	97.7	97.6	7.1×10^{14}
4	78.2	79.0	78.1	78.3	257	82.7	78.2	78.6	2.7×10^{12}	84.6	79.8	79.5	3.1×10^{38}
5	93.6	94.9	94.2	93.9	63	96.0	94.4	93.6	26×10^{5}	95.7	94.1	93.9	5.6×10^{11}
6	60.9	63.8	61.8	60.0	55	66.3	63.8	62.3	59674	68.5	65.9	62.1	2.1×10^{9}
7	76.8	77.9	77.2	76.8	131	83.1	80.1	76.7	3.4×10^{8}	88.0	83.5	76.8	1.2×10^{18}
8	97.3	97.0	98.0	97.5	23	96.5	96.5	96.8	38737	95.0	93.3	96.3	1.8×10^{18}
9	89.8	91.3	90.6	90.1	92	93.6	90.6	90.2	3.3×10^{7}	93.8	91.1	90.8	1.2×10^{10}
10	89.0	89.6	89.1	89.0	151	91.0	89.6	89.1	1.7×10^{9}	91.6	90.0	89.1	2.8×10^{24}
11	62.9	62.5	62.3	61.9	97	64.5	61.9	62.1	1.5×10^{6}	64.5	60.9	61.1	4.6×10^{14}
12	92.6	93.2	92.6	92.6	26	92.6	92.8	93.0	3392	90.7	92.6	93.7	6.1×10^{7}
13	77.5	79.1	77.6	77.8	57	79.9	79.4	78.4	1134750	80.3	78.2	77.0	3.8×10^{8}
gm.	82.41	83.49	82.85	82.55	78.31	85.45	83.78	82.91	4.3×10^{7}	86.44	84.49	82.65	6.2×10^{14}

Archetype as a Hybrid Method



Conclusions

- Archetyping as an method to obtain comprehensible solutions from an ensemble method:
  - The use of multi-trees permits the extraction of a hypothesis from an exponential number of hypotheses
  - An invented dataset avoids the loss of part of the training evidence as validation datasets
- The Archetype solution can also be considered as an explanation of the combined hypothesis

Conclusions

- Some further improvements:
  - The experimental study of archetype as an hybrid method.
  - The study of methods that could select analytically the archetype solution, without the necessity of employing an invented dataset

- **SMILES** is freely available at:
  - **`http://www.dsic.upv.es/~flip/smiles/`**