Application of Multi-Objective Metaheuristic Algorithms in Data Mining

Presented by:

Dr Beatriz de la Iglesia
University of East Anglia
Norwich, Norfolk, UK
email: bli@cmp.uea.ac.uk

UKKDD'07

Overview

- Why use Multi-Objective (MO) algorithms?
- An introduction to MO optimisation
- MO algorithms for classification
- Conclusions

Data Mining

Data Mining is *a step* in the KDD process.

It consists of the application of particular data mining algorithms to extract higher level information in the form of a model or a set of patterns from a large dataset.

Model selection

- Many models can fit the same data.
- Data mining is concerned with the improvement (optimisation) of the model to obtain the best prediction or description of the data, depending on the objectives of the KDD process.

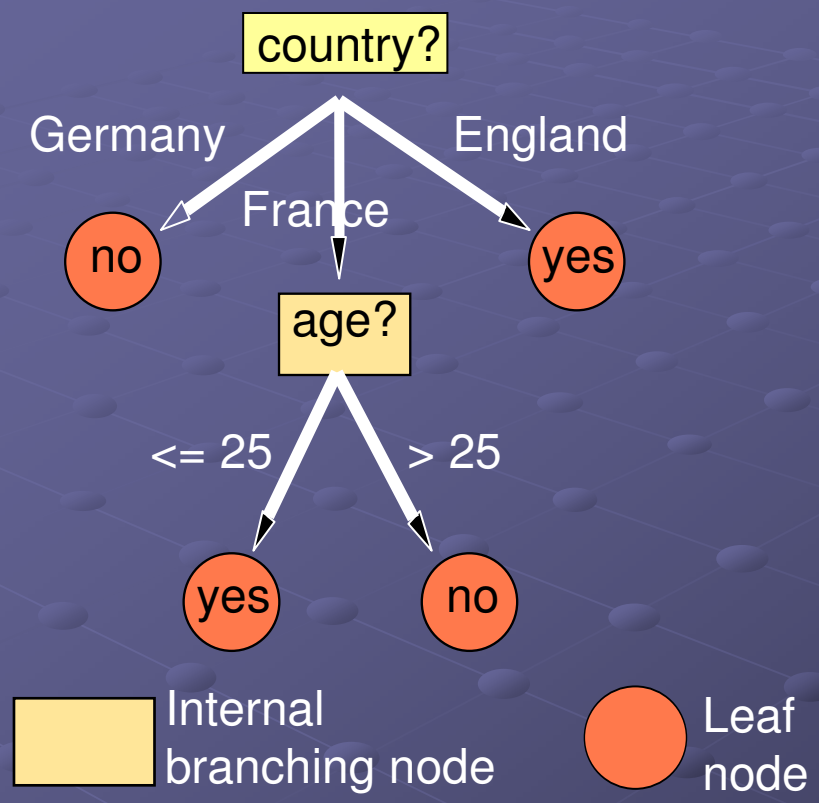
Marketing example

The goal is to predict whether a customer will buy a product given their sex, country and age (classification).

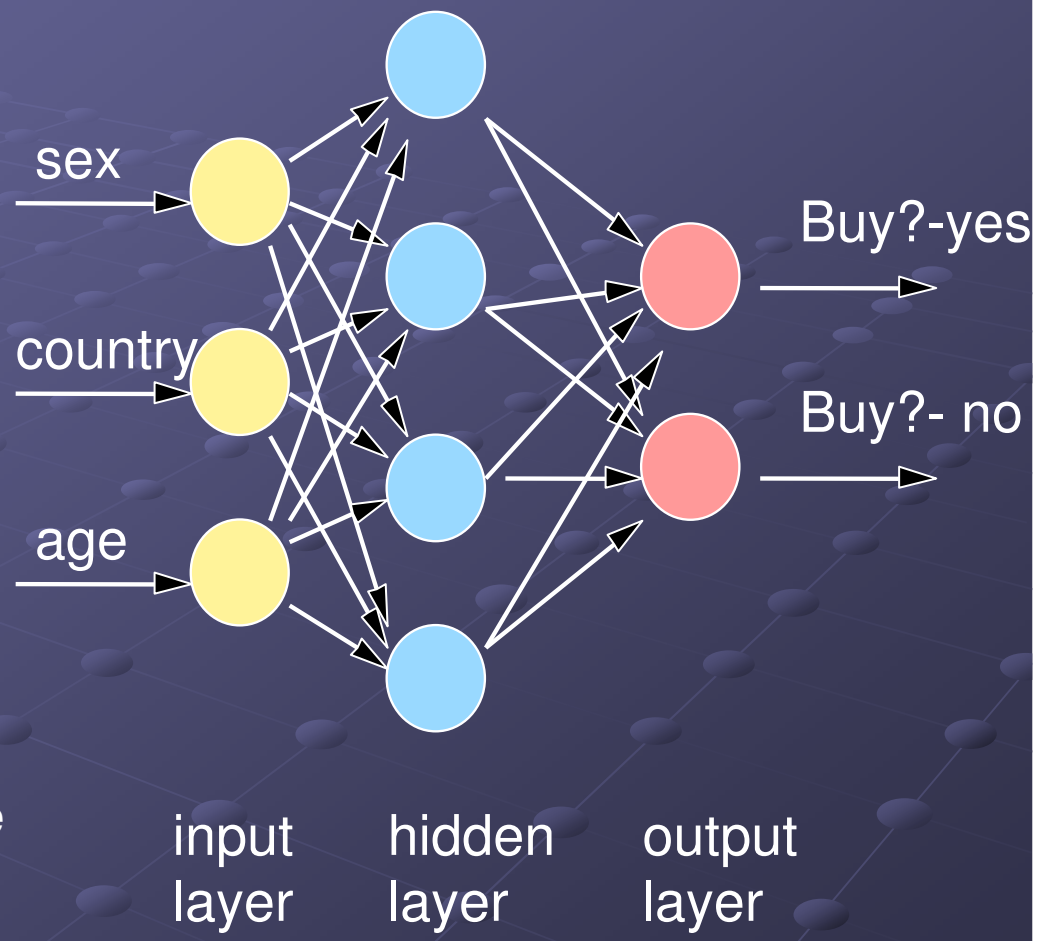
Sex	Country	Age	Buy? Goal/class
M	France	25	Yes
M	England	21	Yes
F	France	23	Yes
F	England	34	Yes
F	France	30	No
M	Germany	21	No
M	Germany	20	No
F	Germany	18	No
F	France	34	No
M	France	55	No

Freitas and Lavington (1998) Data Mining with EAs, CEC99.

Different models



Decision Tree



Neural Network

Optimising the model/patterns

- Data Mining is an optimisation process.
- We search for the best model or patterns according to some *evaluation* criteria.
- This normally requires adjusting parameters of the algorithm.

Generalisation

- Model should not only model the data used to build them (*train set*) but also the real-world process that is generating the data.
- Only then we may get a model that will generalise to other samples from the real-world process.
- We use an independent sample (*test set*) drawn from the real-world data to test the performance of the model on new data.
- Test set must not be compromised when building the model. A validation set should be used for any testing of the model in the intermediary stages.

Model selection criteria

- Most selections involve more than one criterion and there are often conflicts or trade-offs between different criteria.
- Eg. A couple are buying a house
  - She wants a very modern house with “wow” factor and gadgets
  - He wants a house with many rooms for family growth
  - The both want to find the house of their dreams as cheaply as possible

Multi-objective problem

Cost	Rooms	"wow" factor
£250,000	2	4
£300,000	4	5
£500,000	5	4
£500,000	4	3
£550,000	3	10
£900,000	4	3
£900,000	5	7
£1,000,000	10	5

Multi-Objective Data Mining

- In data mining there are also many conflicting criteria for model evaluation.
- Eg.
  - Decision trees and Neural Nets may be evaluated by their complexity and their generalisation error.
  - Association rules may be evaluated by their support and confidence.
  - Clustering solutions may evaluate entropy and purity or other measures of clustering quality.

Multi-Objective Optimisation

- Given two solutions with different objective values, it is not possible to state categorically than one solution is better than the other.
- Multi-objective algorithms must find the set of all such trade-off solutions.
- The user can then select a solution according to preference.

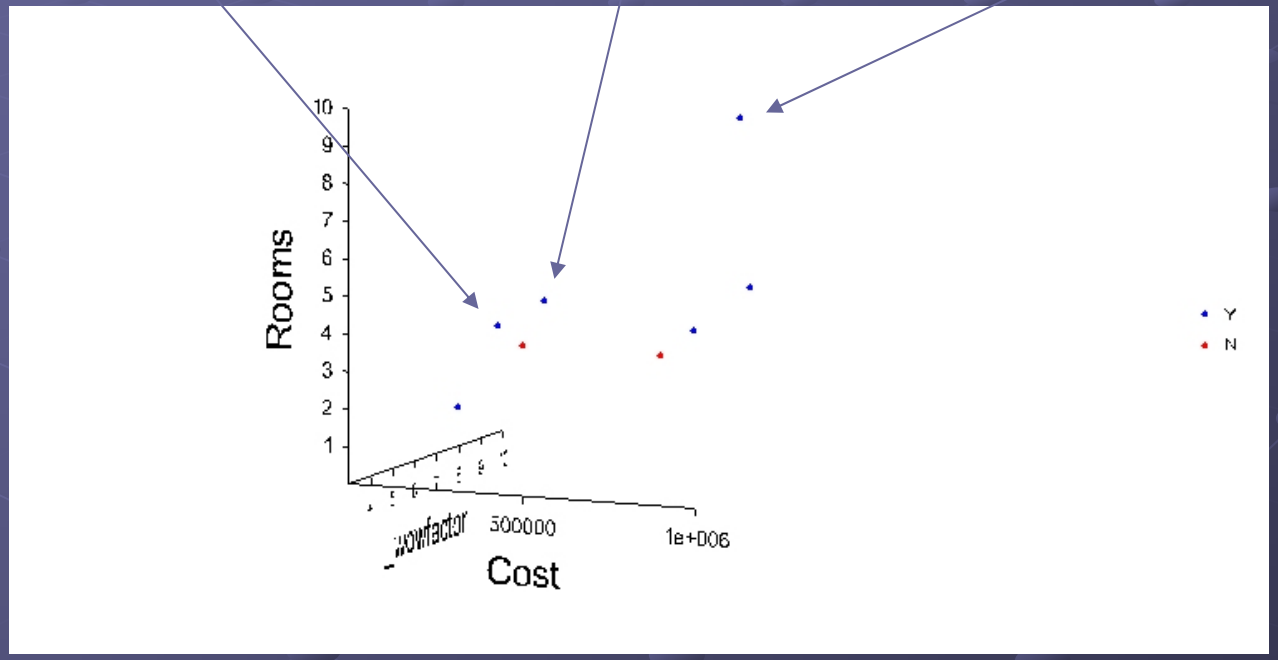
MO Opmimisation

- Given a problem with *n* objectives $f_1, \ldots, f_n$, each of which is going to be maximised, solution *a* *dominates* solution *b* if

$$f_i(a) \geq f_i(b) \forall i \in \{1, \ldots, n\} \text{and}$$
$$\exists \, \mathrm{j} \in \{1, \ldots, n\} \text{such that } \mathrm{f_j}(a) > f_j(b).$$

- Given a set of solutions, *S*, a solution $a \in S$ is *non-dominated* if there is no solution $s \in S$ that dominates a.

Pareto Front



MO algorithms

- Should approximate the Pareto-front.
- Should provide a good spread of solutions in the Pareto-front.
- Evolutionary Algorithms (EAs) are well suited to MO optimisation as they deal with a population of solutions.
- Many alternative EA approaches including
  - Aggregating functions
  - Lexicographical ranking
  - Pareto Dominance (e.g. PAES, NSGA II, SPEA 2)

Classification

- In classification a model is sought which can assign a class to each instance in the database. It relies on historical labelled data.
- Nugget discovery or partial classification seeks to find patterns that represent a “strong” description of a predefined class.
- Particularly relevant for “minority classes”.

A nugget

- Let *Q* be a finite set of *attributes,* each with an associated domain.
- A *record* specifies values for each attribute in *Q.*
- A tabular database*, D,* is a finite set of records.
- *Partial classification rules are of the form*

$$\text{antecedent} \Rightarrow \text{consequent}$$

where the antecedent and consequent are predicates used to define subset of records from the database *D* and the rule underlines an association between these subsets.

Attribute Tests

- In nugget discovery, antecedent is often constrained to be a conjunction of Attribute Tests (ATs).
- Numerical attributes are defined by a simple value, binary partition or range of values (e.g. age≥25)
- Categorical attributes are defined by a simple value, a subset of values or an inequality test (e.g. colour≠blue).
- The consequent represents the class assignment.

Strength of a rule

- If *A* represents the antecedent and *C* consequent the strength of the rule, *r*, can be expressed by

$$\mathrm{conf}(r) = \frac{|C|}{|A|} = \frac{c}{a} \qquad \mathrm{cov}(r) = \frac{|C|}{|B|} = \frac{c}{b}$$

A = {t ∈ D | antecedent (t)}

B = {t ∈ D | consequent (t)}

C = {t ∈ D | antecedent (t) ∧ consequent (t)}

- A strong rule is one that meets certain confidence and coverage thresholds.

Marketing example

Sex	Country	Age	Buy? Goal/class
M	France	25	Yes
M	England	21	Yes
F	France	23	Yes
F	England	34	Yes
F	France	30	No
M	Germany	21	No
M	Germany	20	No
F	Germany	18	No
F	France	34	No
M	France	55	No

If Country = England then "YES"
Conf (r) = 2/2 =1
Cov(r) 2/4 = 0.5

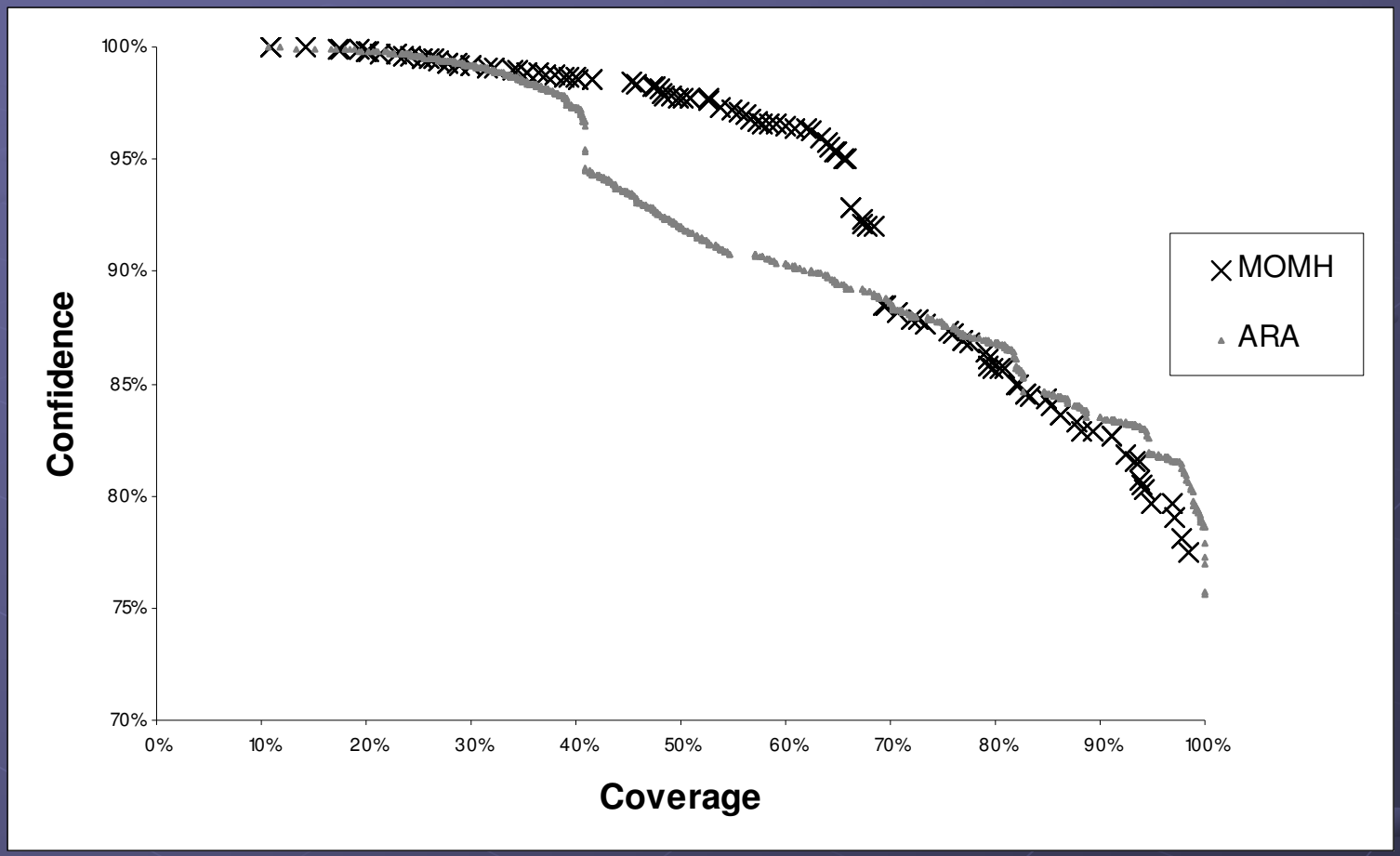
If Country = France then "YES"
Conf (r)= 2/5 = 0.4
Cov(r) = 2/4 = 0.5

If (Country = France AND Age≤ 25) OR (Country = England) then "YES"
Conf (r) = 4/4 = 1
Cov(r) = 4/4 = 1

Multi-Objective problem

- Each measure of the strength of the rule represents a different objective function to be optimised.
- Allows for the application of Multi-Objective metaheuristics (MOMH).
- Each Pareto optimal solution represents a different compromise between the objectives.

Partial Classification with MOMH



Rules for "<=50K" class of UCI Adult Dataset

Partial classification rules

IF CapitalGain <= 41310
THEN Salary <= 50K

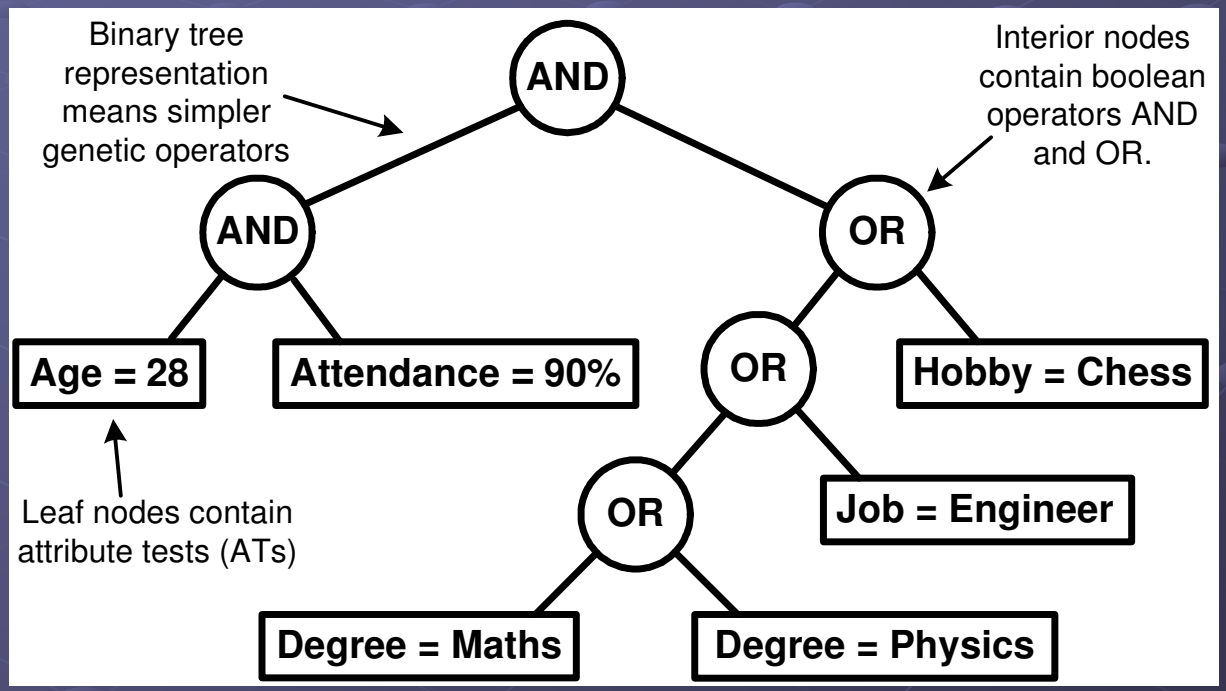
Confidence = 0.75
Coverage = 1

IF CapitalGain <= 6723 AND
Age <= 21 AND
HoursPerWeek <= 59 AND
MaritalStat != Married_civ_spouse
THEN Salary <= 50K
Confidence = 1
Coverage = 0.109341

IF CapitalLoss <= 2206 AND
CapitalGain <= 6849 AND
Relationship != Husband
THEN Salary <= 50K
Confidence = 0.92
Coverage = 0.69

Rule Trees for binary classification

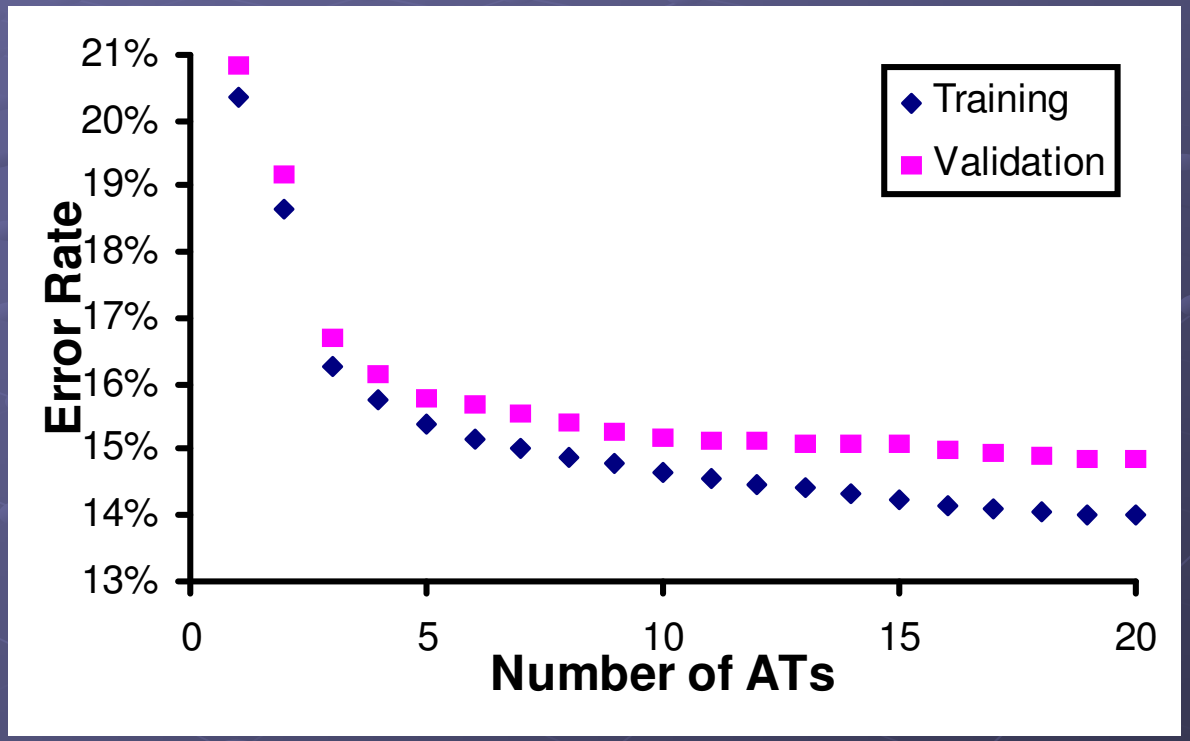
- Rule trees are more expressive than simple rules and more compact than sets of such rules.



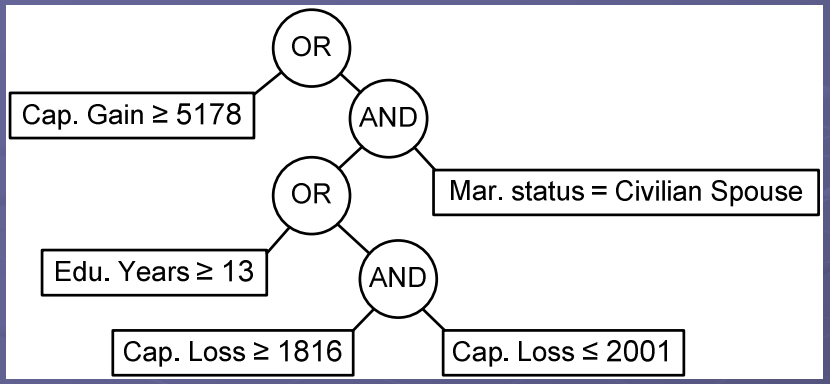
Objectives for MO approach

- **Minimizing misclassification costs:**
  - simple error rate;
  - balanced error rate;
  - any other measure of overall misclassification cost.
- **Minimizing rule complexity:**
  - to encourage the production of rules that are easily understood by the client;
  - to reduce the chance of overfitting the data.

Train/Validation/Test



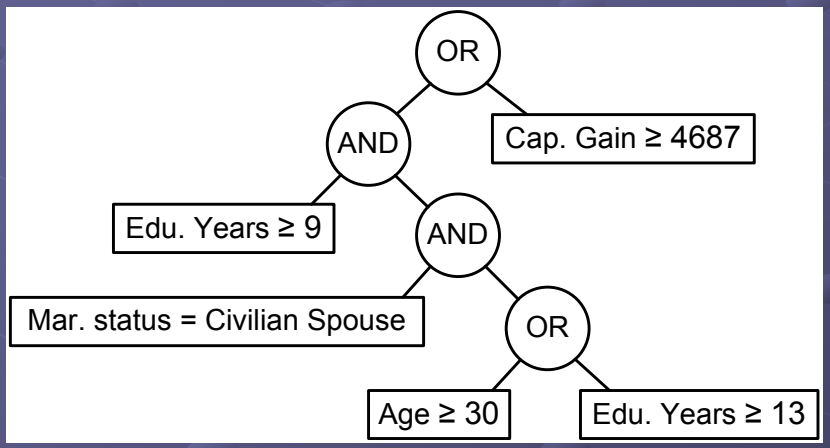
Rules for Adult “>50k”



Confusion matrix

		Predicted -	Predicted +
Actual	-	10732	628
	+	1710	1990

Simple error rate

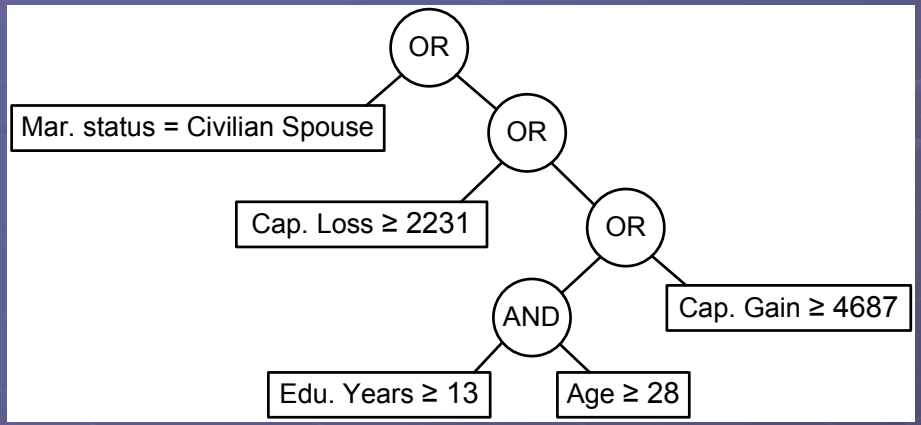


Confusion matrix

		Predicted -	Predicted +
Actual	-	8631	2729
	+	549	3151

Balanced error rate

Rules for Adult “>50k”



Confusion matrix

		Predicted	
		-	+
Actual	-	6524	4836
	+	140	3560

FN is 10 times the cost of a FP

Rule sets

If cap. gain ≥ 5178 then salary > 50 K

If mar. status = civilian spouse and cap. loss ≥ 1816 and cap. loss ≤ 2001 then salary > 50 K

If mar. status = civilian spouse and edu. years ≥ 13 then salary > 50 K

Otherwise salary ≤ 50 K

Result evaluation

- MOMHs to extract ETs produce complete, simple and understandable class descriptions.
- Competitive in terms in classification performance with other more sophisticated classifiers.
- Very flexible approach, providing user with a number of models.
- Can operate with different measures of complexity and different error measures.
- Can present knowledge in different formats: rule set, rule tree.

Conclusions

- Application of meta-heuristics to data mining has produced efficient and effective algorithms, scalable to large databases
- A MO approach has many advantages for both partial classification and binary classification.
- Algorithms for partial classification have compared well with All Rule Algorithms and algorithms for rule trees compare well with classification algorithms.
- The approach is very flexible: we can change rule representation, AT Types, objectives, etc.
- We are continuing research in this area.

Questions?