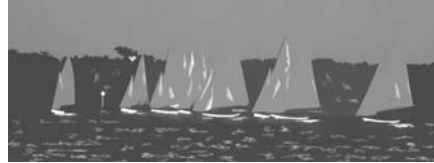


# Meta-heuristics in data mining

---



**Dr George D Smith**  
**School of Computing Sciences**  
University of East Anglia, Norwich  
NR4 7TJ, England

---

## Contents

---

- ❖ What is data mining?
- ❖ Rule Induction using heuristic search
- ❖ Feature construction using genetic programming

## What is data mining?

---

- ❖ “DM is the search for valuable information in large volumes of data” - Weiss & Indurkha (1998)
- ❖ “DM helps end users extract useful business information from large databases” - Berson & Smith (1998)



**Companies have large amounts of data from which useful knowledge must be extracted**

## **DM - definition**

---

- ❖ Data mining is the search for implicit relationships and patterns in data held in large databases.
- ❖ These relationships represent valuable knowledge about the database and of the real world represented by the database.

## **Example scenario**

---

- ❖ Insurance companies have massive amounts of data on customer details, premiums, claims and hence costs.
- ❖ A pattern in this data may represent a niche market.
- ❖ If exploited, it may give them an edge in the market, but only briefly.

## Induction v/s deduction

---

- ❖ DM is an inductive (not deductive) process
  - Deduction - extract information that is a logical consequence of the data in the database. Supported by most DBMSs and OLAP.
  - Induction - infer knowledge that is generalised from the data in the database. Generally not supported by DBMSs or OLAP.

## Rationale

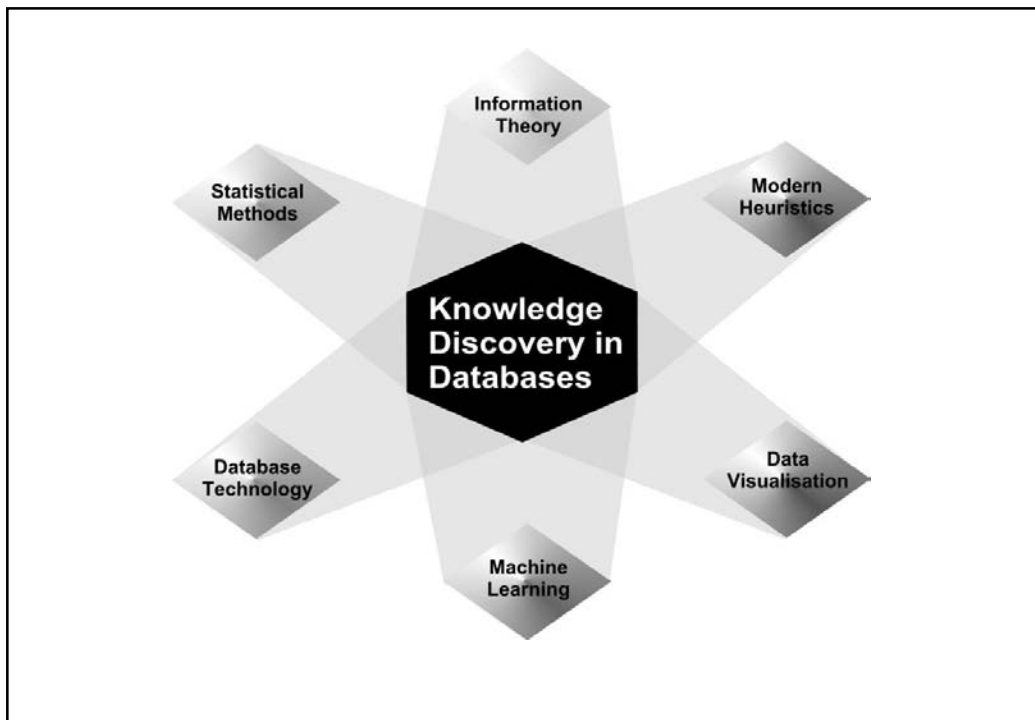
---

- ❖ With databases of enormous size, the user needs help to analyse the data more effectively than just simply reporting and querying.
- ❖ Semi-automatic methods to extract useful, unknown (higher-level) knowledge in a concise format will help the user make more sense of their data.

## Roots of Data Mining

---

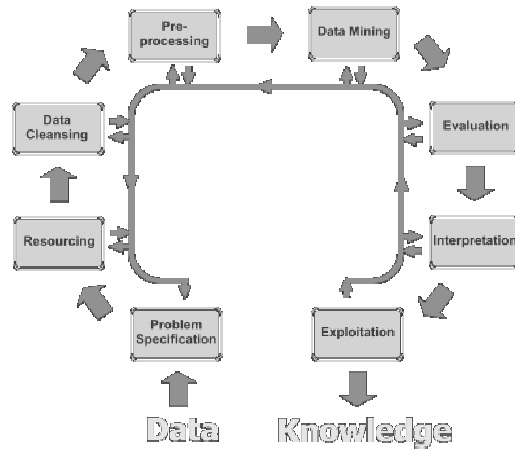
- ❖ Data mining draws on the concepts of three major paradigms:
  - Database technologies
  - Statistical methods
  - Machine learning
- ❖ as well as secondary technologies, such as heuristics and visualisation, and also domain knowledge.



## The kdd process

Data mining is one stage of the knowledge discovery in databases (kdd) process.

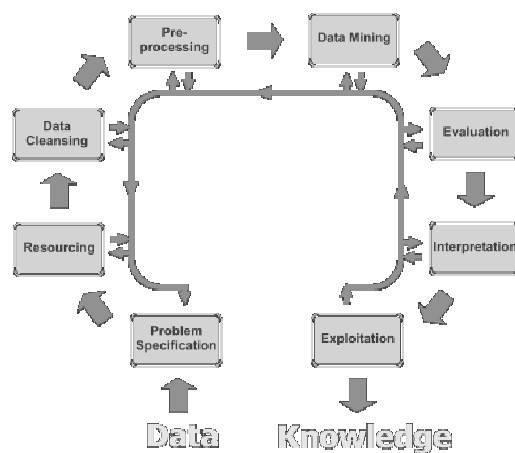
The other stages are geared to preparing the data for the mining process and interpreting and implementing the results.



## The kdd process

A data mining project should:

- follow a specification
- be fully resourced
- be flexible, with key decision points
- leave an audit trail
- be managed and controlled accordingly



## DM tasks

---

- ❖ There are two recognised categories of goals (or high level tasks) in DM:
  - Prediction (models are constructed using historical cases to predict outcomes for all new cases)
  - Description (models are constructed to describe particular patterns or relationships in the data - nugget discovery)

## Predictive DM

---

- ❖ Consider fraud detection in the use of credit cards.
- ❖ Here a model is constructed using historical data (with % of known fraud cases) to predict fraud for future instances. The model is designed to handle all possible future transactions and is thus a global model.

## Descriptive DM

---

- ❖ These models are concerned with identifying patterns in subsets of the database, usually with the objective of discovering nuggets of knowledge, which in turn represent business opportunities.
  - For example, in insurance data, finding that young male drivers with fast cars, but car is old, do not claim very often, therefore low risk.

## DM tasks or activities

---

- Classification (direct marketing, retention)
- Clustering (segmentation, www marketing)
- Regression (credit scoring)
- Association techniques (market basket)
- Visualisation
- Time series (forecasting, trends analysis)
- event stream analysis (sequencing)

## **Rule Induction using Heuristic Search**

---

- ❖ We will now see how to induce rules for classification using heuristic search techniques.
- ❖ Specifically, we use simulated annealing.

## **Classification**

---

- ❖ Classification is used to classify each database record into one of a number of pre-defined classes based on values of other (predicting) attributes for the record.
- ❖ The class is one of the attributes of the database, and each record of the historical DB has already been assigned to one of these classes.

## Marketing example

---

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

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

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

## Classification Techniques

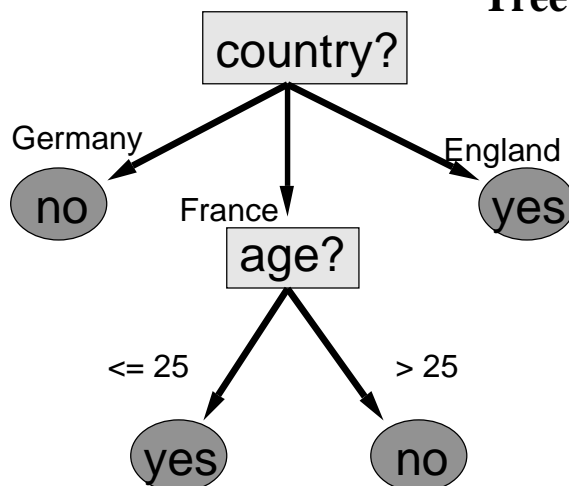
---

- ❖ There are many techniques used for classification, including:
  - Tree induction (decision trees)
  - Rule induction
  - ANNs
  - Statistical models

## Tree Induction


- ❖ Decision trees are mainly predictive models that classify the data by asking a classification question at each branch.
- ❖ The internal nodes of the tree relate to the predicting attributes, while the leaf nodes are the predicted class.
- ❖ ID3, C4.5, C5, CART, CHAID/XAID

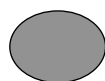
### Tree Induction



This is a decision tree induced by the Marketing example data, using C4.5.

The first branch is called the root of the tree.

 Internal branching node

 Leaf node

## Tree induction algorithms

---

- ❖ vary in their splitting criterion;
- ❖ are usually greedy and iterative;
- ❖ can build very large trees for moderate data sets;
- ❖ can generate rules derived from trees (not the same as rule induction).

## Rule induction (RI) algorithms

---

- ❖ As their name suggests, RI algorithms generate rules outright, not via decision trees.
- ❖ Rules take the form: If (condition) then (class), where the condition has to be True for class to be assigned to a record.
- ❖ For example, in marketing DB,  
If (country=Germany) then (buy?=No)

## RI - terminology

---

- ❖ When we refer to a rule we are really talking about a classification rule:

If (set of conditions) then (class),

or

$\alpha$  implies  $\beta$

- ❖  $\alpha$  is referred to as the antecedent of the rule.
- ❖  $\beta$  is the consequent of the rule.

## RI - notation

---

- ❖ Assume we have a database  $D$  of  $d$  records.
- ❖ Each record has  $n$  attributes (fields),  $A_i$ .
- ❖ Each  $A_i$  is defined over domain  $Dom_i$ .
- ❖ We are asked to find a rule

$$\alpha \Rightarrow \beta$$

that appears to hold for some records in the database.

- ❖ Let  $r$  denote a particular record in  $D$ .

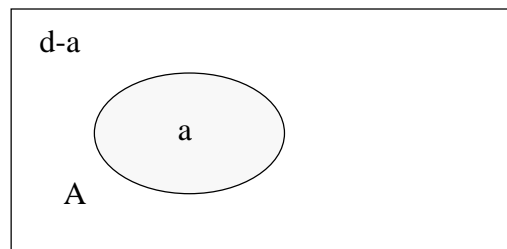
## RI - performance measures

---

- ❖ Associated with any rule  $\alpha \Rightarrow \beta$  are three sets of records:

$A = \{r|\alpha(r)\}$  - the set of records matching condition & hence classified as belonging to a particular class.

Let  $a = |A|$ .

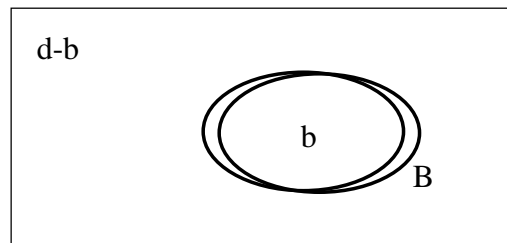


## RI - performance measures

---

- ❖  $B = \{r|\beta(r)\}$  - the set of records that actually belong in this class - fixed for each class/data set.

❖ Let  $b = |B|$

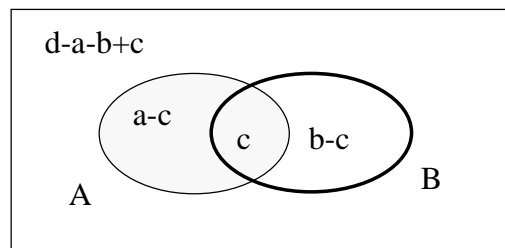


## RI - performance measures

---

❖  $C = \{r \mid \alpha(r) \text{ and } \beta(r)\} = A \cap B$ . Thus  $C$  is the set of records that are accurately classified by the rule.

❖ Let  $c = |C|$ .



## RI - performance measures

---

❖ Likewise, there are a number of measures we use to determine how good the rule is:

- Coverage is the number of records satisfying the condition part of rule, i.e. how often the rule applies. Coverage =  $a$ .
- It can also be expressed as a proportion of the whole database. Coverage =  $a/d$ .
- Sometimes referred to as applicability (support).

## RI - performance measures

---

- ❖ Accuracy is a measure of how often the rule is correct.
- ❖ Accuracy is the number of instances it predicts correctly, expressed as a proportion of all instances matched by rule.
- ❖ Accuracy =  $c/a$
- ❖ Sometimes referred to as confidence.

## For example:

---

If (Country = France)  
then (Buy = Yes)

Coverage =  $5/10 = 50\%$

Accuracy =  $2/5 = 40\%$

Support = Acc \* Cov  
=  $2/10 = 20\%$

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

## Interestingness

---

- ❖ is another measure used.
  - It should increase/decrease with accuracy for a fixed coverage.
  - It should increase/decrease with coverage for a fixed accuracy.
- ❖ Various combinations of coverage and accuracy used.

## Domain knowledge is also useful ...

---

- ❖ Consider the following rule (patient DB):
- ❖ If (patient = pregnant) then (gender = female)
  - Accuracy is 100%, we hope!
  - Coverage may also be significant.
  
- ❖ Rule is completely useless!

## RI with heuristics

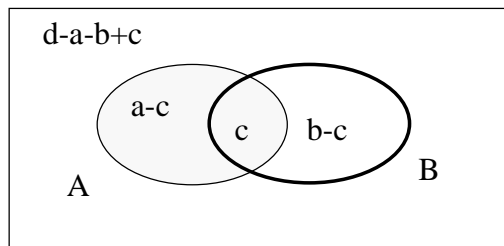
❖ We have two competing objectives:

- Maximise  $c$
- Minimise  $a-c$  (error)

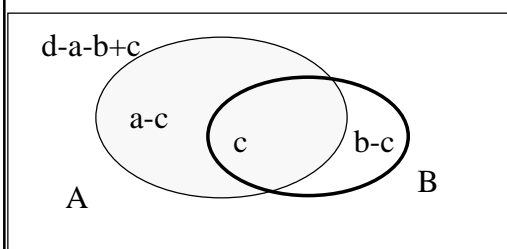
❖ Consider

- maximise  $(\lambda c - a)$

$\lambda$  is the balance between accuracy and coverage

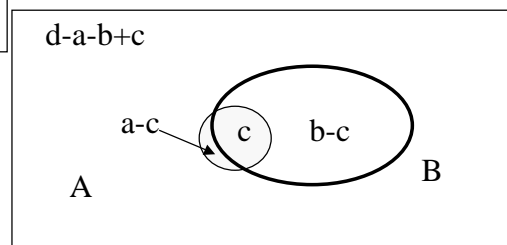


## Why competing?



Increasing  $c$  :  
matching more records  
but  $a-c$  may also increase

Decreasing  $a-c$  :  
matching fewer records  
but  $c$  may also decrease



## RI with heuristics

---

- ❖ With the objective function  $\max(\lambda c - a)$ , rule induction becomes a 'search' problem ( $\lambda$  is fixed during search).
- ❖ Given a set of 'operators' that can modify an initial (random) rule, we use simulated annealing to search for the best rule that can maximise our measure.
- ❖ Our methodology is to start with a random rule of the form  $\alpha_1 \wedge \alpha_2 \wedge \dots \wedge \alpha_s$  - for example  $\{(age < 16) \wedge (sex = M)\}$  then modify the rule by making a small change (attribute/ value/operator/ condition). This is equivalent to a simple neighbourhood move. This is evaluated on training data.

## Simulated annealing

---

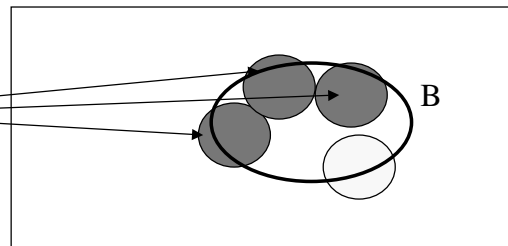
- ❖ Local search samples neighbour solutions, moving only if the neighbour solution is better.
- ❖ Simulated annealing relaxes this condition in local search, allowing a move to a worse neighbour solution with a probability that decreases during the search.
- ❖ Thus we can escape from local minima.

## Covering rules

---

- ❖ To 'cover' the target class, the methodology is to aim for a rule with high accuracy, strip out all records which match condition (whether accurate or not) then repeat the process, each time relaxing  $\lambda$ .

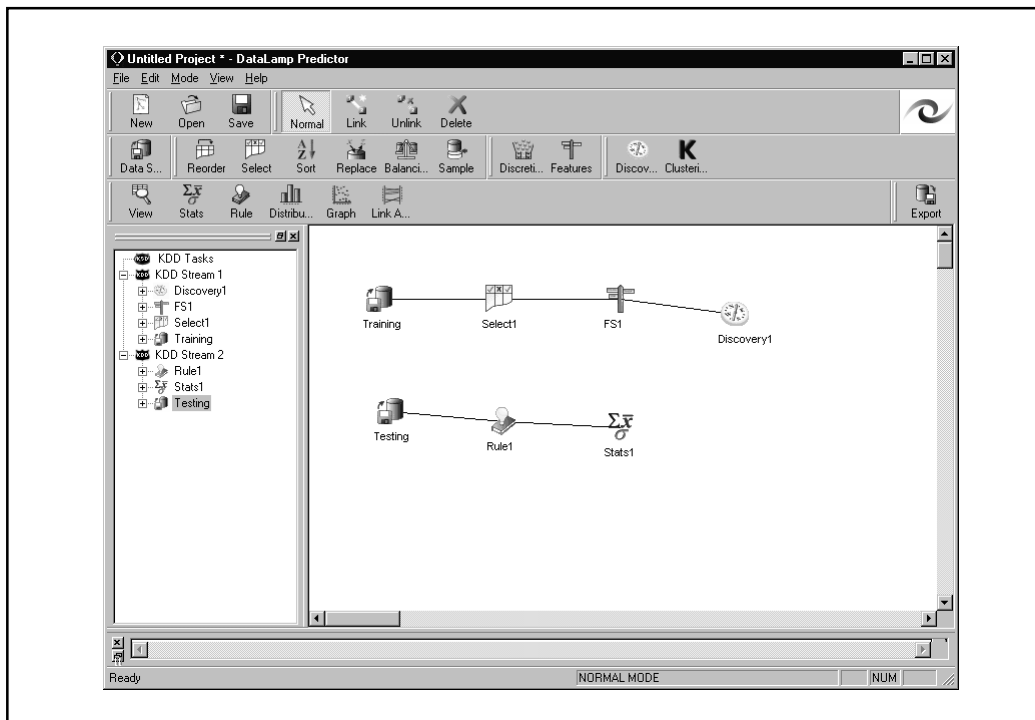
Records matched previous rule and removed from the training set



## Data mining software

---

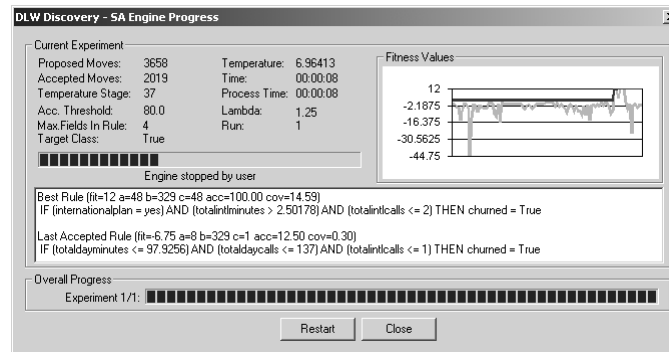
- ❖ As part of a TCS project with Lanner Group, UEA has designed and developed a data mining toolkit, DataLamp, which is aimed at capturing the market for comprehensive data pre-processing and data mining software.
- ❖ The product, DataLamp, is in release 2 and is available on Windows. (Witness Miner)



## DataLamp Discovery dialogue box



## DataLamp - rule induction using SA



## Fraud in mobile phone networks

- ❖ A large telecommunications company supplied us with a customer information database.
- ❖ Each record represented a snapshot of the customer, taken every 2 hours, and represented the difference between current and historical behaviour.
- ❖ Our task was to derive patterns from this DB that could be used to identify customers who were likely to be using the services fraudulently.

## Fraud in mobile phone networks

---

- ❖ There are 34,104 records in the DB.
- ❖ Each record contains 32 fields, one of which was a class field, fraud. Approximately 5.8% of the records were classified as fraud.
  - training set - 17052 records
    - fraud - 1006 (5.9%)
    - non-fraud - 16046
  - test set - 17052 records
    - fraud - 966 (5.7%)
    - non-fraud - 16086

## Fraud in mobile phone networks

---

- ❖ Tree induction- requires the data set to be balanced
  - Clementine (C4.5)
  - KnowledgeSeeker (XAID/CHAID)
- ❖ Rule induction
  - DataLamp (Witness Miner, using simulated annealing)
- ❖ ANNs
  - MLP - two different topologies

## Fraud in mobile phone networks

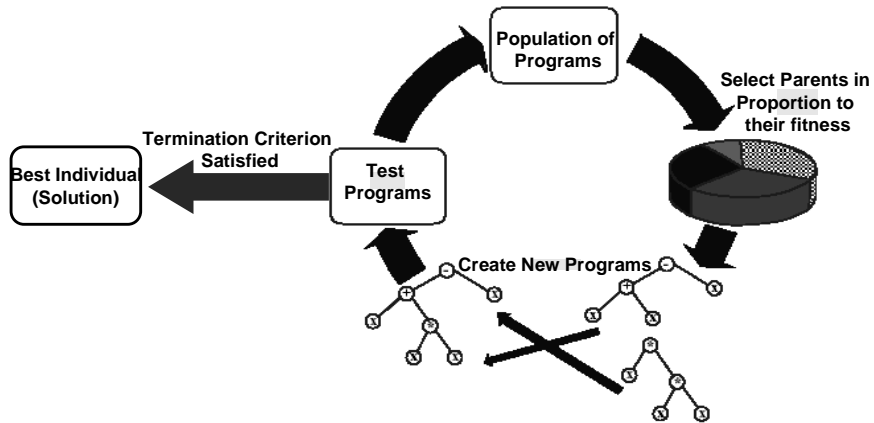
Actual	N-F	N-F	F	F			
Predicted	N-F	F	N-F	F	correct	no. of	no. of
Baseline	16086	0	966	0	94.3%	rules	cond.
C4.5- all rules	15743	343	203	763	96.8%	32	3 to 5
ANN default	16012	74	567	399	96.2%		
ANN optimised	16019	67	589	377	96.2%		
KS-whole tree	15856	230	305	661	96.9%	66	6
SA - rule 10	15848	238	295	671	96.9%	1	5

Note the overall accuracy that SA achieves with 1 rule, compared to the tree induction algorithms, C4.5 and KS. Subsequent covering improves this first rule.

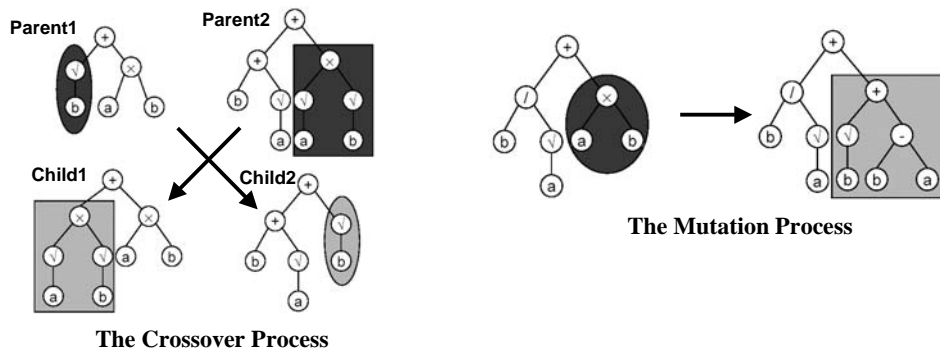
## Feature construction using GP

- ❖ Feature construction is the process of constructing new attributes which are linear or non-linear combinations of original attributes.
- ❖ Tree & rule induction algorithms typically construct each condition on a field-by-field basis. Any combination of attributes which presents a much stronger prediction will therefore be missed.
- ❖ Oblique classification (OC1) - considers linear combinations of attributes at each split.

# Genetic programming

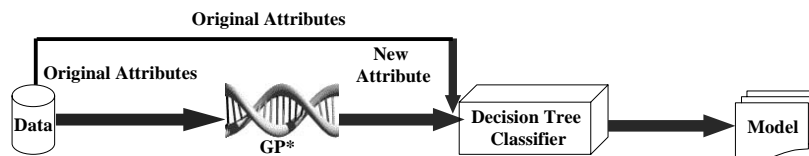


# Creation of new programs



## Methodology

---



We apply a GP to construct an attribute which is a function of the original attribute set. We add this new attribute to the attribute set and compare the performances of the classifiers (three tree induction algorithms C5, CHAID, CART and also an ANN) with and without this constructed attribute.

## GP Parameters

---

- ❖ Terminal set - {original attribute set, 1}
- ❖ Function set - {+, -, x, / }
- ❖ Population size 600
- ❖ Tournament selection (7)
- ❖ Crossover rate 70%, mutation rate 50%
- ❖ Fitness - Information Gain (Gini Index). This is the value an attribute would be assigned at a splitting node in the tree induction algorithm C5 (CART). The higher the value, the more predictive power of the (constructed) attribute.

## Experimental data sets

<b>Abalone</b>	<b>4177</b>	<b>28</b>	<b>8</b>
<b>Balance-Scale</b>	<b>625</b>	<b>3</b>	<b>4</b>
<b>Bupa Liver Disorder</b>	<b>345</b>	<b>2</b>	<b>6</b>
<b>Waveform</b>	<b>300</b>	<b>3</b>	<b>21</b>
<b>Wine</b>	<b>178</b>	<b>3</b>	<b>13</b>

Data sets from [www.ics.uci.edu/~mlearn/MLRepository.html](http://www.ics.uci.edu/~mlearn/MLRepository.html)  
10 fold cross-validation used.

## Error rates - average on test sets (10xCV)

Abalone	Original	GP - IG	GP - GI	Balance	Original	GP - IG	GP - GI
C5	79.26	79.09	80.07	C5	22.42	0.00	0.00
CHAID	76.45	74.77	75.40	CHAID	28.39	5.65	5.49
CART	74.69	73.02	73.86	CART	22.74	0.00	0.00
ANN	72.45	72.49	73.14	ANN	10.00	9.36	9.19

BUPA	Original	GP - IG	GP - GI
C5	36.47	32.35	32.94
CHAID	40.00	30.00	32.06
CART	32.35	30.29	31.18
ANN	37.65	35.29	31.47

Waveform	Original	GP - IG	GP - GI	Wine	Original	GP - IG	GP - GI
C5	22.94	19.54	19.64	C5	6.47	5.29	4.12
CHAID	28.36	24.92	24.64	CHAID	17.06	17.65	15.29
CART	24.58	20.02	19.54	CART	10.59	5.88	3.53
ANN	17.15	15.20	15.77	ANN	3.53	2.94	1.76

## Tree Sizes

---

- ❖ Finally, apart from Abalone data set, the use of the evolved attribute leads to a significant reduction in the size of tree induced by C5, CART & CHAID.
- ❖ For Balance data set, for example:
  - Original set: average 80.8 nodes (depth 9.7)
  - GP - IG set: always 5 nodes (depth 3)
  - GP - GI set: always 5 nodes (depth 3)
  - In GP sets, evolved attribute is always  $(f_1 * f_2) / (f_3 * f_4)$ , ie 7 nodes, and this appears twice in the new induced trees (C5 & CART), giving a total size of 17 nodes (depth 5).

## References

---

- ❖ Predictive Data Mining, Sholom Weiss & Nitin Indurkha, Morgan Kaufmann (1998), ISBN 1-55860-403-0
- ❖ Data Mining, Ian Witten & Eibe Frank, Morgan Kaufmann (1999), ISBN 1-55860-552-5
- ❖ de la Iglesia, B. and Debuse, J.C.W. and Rayward-Smith, V.J., "Discovering Knowledge in Commercial Databases Using Modern Heuristic Techniques", *In 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 44-49, 1996 Sponsored by IEEE Computer Society
- ❖ Muharram & Smith, The effect of evolved attributes on classification algorithms, AI2003, 16th Australian Conf. AI, Lecture Notes in AI, no 2903, Springer-Verlag
- ❖ Muharram & Smith, Evolutionary feature construction using information gain and gini index, To appear in EuroGP2004, Coimbra, Portugal.

# Thank you

---

