UNIVERSIDAD POLITECNICA DE VALENCIA
- Spain -

2nd Annual Conference of the ICT for EU-India Cross Cultural Dissemination

Valencia, Spain, November 14-15, 2005

Extending Decision Trees for Web Categorisation

Multiparadigm Inductive Programming group
(Extensions of Logic Programming group, ELP)
Universidad Politécnica de Valencia
José Hernández Orallo

Outline

- The MIP group
- Project Objectives
- Data Mining and Web Mining
- Data Mining for Web Categorisation
- A General-purpose Algorithm: DBDT
- DBDT for Web Classification
- Experimental Evaluation of DBDT
- Conclusions and Future work

The MIP group

- Began its research activities in 1997 inside the ELP group.
- Composed of
  - **3 PhD + 3 PhD students + 2 research collaborators**
- Research areas
  - Multiparadigm inductive programming (ILP, IFLP, …)
  - Multi-relational learning
  - Mainstream machine learning and data mining
  - Multi-classifier systems / ensemble methods
  - Cost sensitive learning and ROC analysis
  - Mimetic models
  - **Web mining and learning from complex data**
  - Other: Inductive debugging, theoretical foundations of machine learning, …

Project Objectives

- Two main objectives:
  - Effective knowledge extraction, handling and exchange, using “intelligent” software
  - Improve the accessibility of (cultural) information

- More and more inductive techniques are needed:
  - Knowledge discovery tools.
  - Knowledge transformation tools.
  - Software that learns and adapts.
  - Software that can handle non-specified situations.

Project Objectives

- Nowadays, the Web is the most important source for information

- Web information has special characteristics:
  - Heterogeneous.
  - Poorly structured.
  - Noisy.
  - Unpredictably volatile.
  - Huge.

- Specific tools are needed to help us handle such variety and quantity of information.

Data Mining and Web Mining

- Data mining (or more academically KDD) aims at discovering relevant knowledge from different sources of information.
- Web mining aims at discovering relevant knowledge from the Web.
- Web mining is classified into:
  - Content mining (text, title, keywords, …): classification, categorisation, summarisation, …
  - Structure mining (hyperlinks, website topology): finding hubs, authorities, …
  - Usage mining (log files, navigation trails): navigation patterns, user profiles, preferences, recommendations, …

Data Mining and Web Mining

- Web documents are especially difficult for classical DM techniques:
  - Non-structured.
  - Heterogeneous: textual, multimedia, hyperlinks, meta-labels, etc.

- Web mining adapts classical DM techniques or develops specific algorithms.
  - In general, lots of preprocessing is needed to convert the web data into simpler (flat and structured) data.

Data Mining for Web Categorisation

- Categorisation aims at finding one or more categories (from a set of categories) for a new document.
- When the number of possible categories is not very high, a feasible way of performing categorisation is trough several classifiers (one for each categoriy)
- Some simple approaches to Web document categorisation/classification take only the textual part into consideration.
  - Structure or usage information is not usually handled by the most common web mining tools.
    - But this information is also relevant!

Data Mining for Web Categorisation

Some techniques:

- Relational learning techniques
  - Special predicates: *has_word(), has_anchor_word(), link_to()*
- Bayesian techniques
  - Content information: *text+title (bags of words)*
- Support vector machines (upgraded)
  - Content information: *text + title*
  - Structure information: *anchor words*
- Decision trees (upgraded)
  - Content information: *keywords + some text (a few bags of words)*
  - Structure information: *hyperlinks*

Along with preprocessing (tags and natural language preprocessing)

A General-purpose Algorithm: *DBDT*

- Our proposal:
  - Use of structured (powerful) data types for representing each document feature (title, keywords, text, links, visits, …) as lists, trees, sets, etc.
  - Integration of web content, structure and usage in a unique framework, using a modification of decision tree learning in order to handle complex data

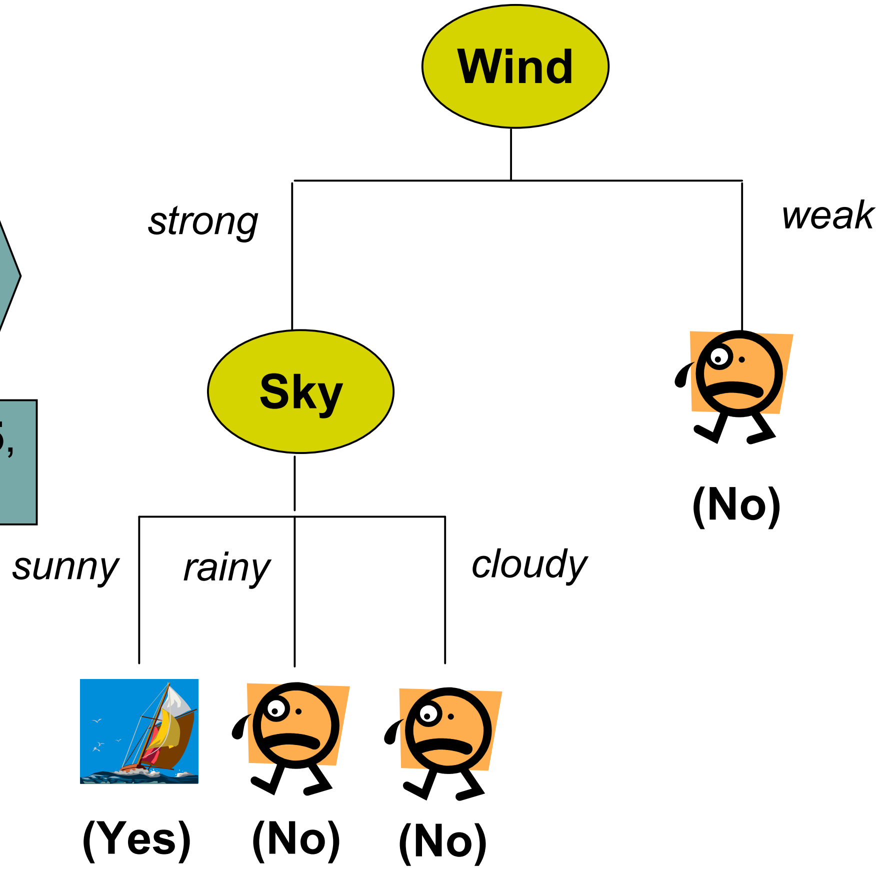
Distance-Based Decision Trees (DBDT)

A General-purpose Algorithm: *DBDT*

What is a Decision Tree?

Temp.	Wind	Sky	Sail?
Hot	Strong	Sunny	Yes
Hot	Weak	Sunny	No
Warm	Strong	Rainy	No
Cold	Strong	Rainy	No
Cold	Weak	Rainy	No
Hot	Strong	Cloudy	Yes

ID3,C4.5, …



A General-purpose Algorithm: *DBDT*

Decision Trees: partition rules?

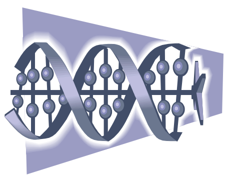
Nominal att. $X = a_1 \vee \ldots \vee X = a_n$

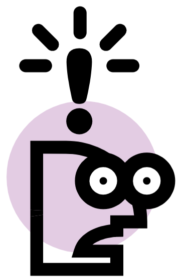
$X \in \{a_1, \ldots, a_n\}$ $X = a_i \vee (X = a_1, \ldots a_{i-1}, a_{i+1} \ldots, a_n)$

Numerical att. $X \leq h_1 \vee \ldots \vee h_i \leq X \leq h_{i+1} \vee \ldots \vee X \leq h_n$

$X \subseteq R$

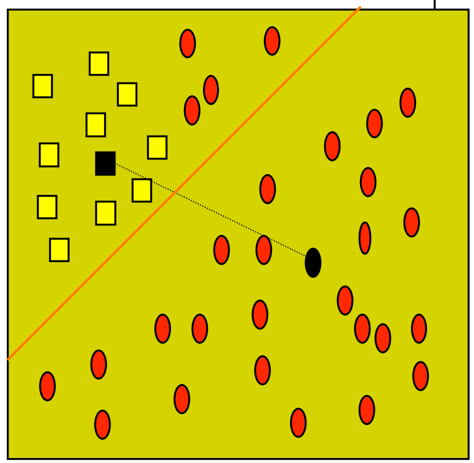
Structured att.

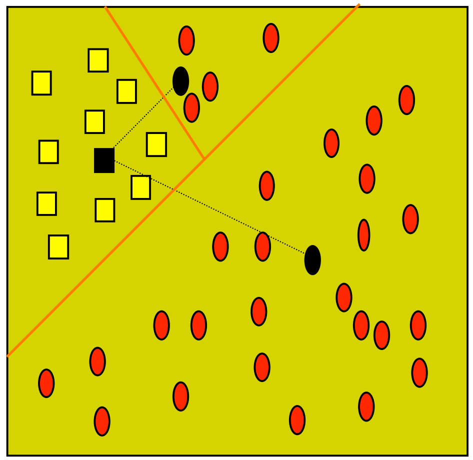




A General-purpose Algorithm: *DBDT*

- Centre splitting (Thornton 1995-2000)
  - Distance-based method for numerical attributes (linear discriminant).
  - One centre is calculated for each different class.
  - The space is divided according to these centres.
  - The process is iterated and stopped when all the regions are pure.
- Problems.
  - Requires a single distance between documents.
  - A simple distance loses information and doesn't provide too much knowledge.





A General-purpose Algorithm: *DBDT*

- Extension to convert this into a decision-tree technique
  - Apply the centre splitting technique for each attribute.
  - The centre must be a value of the dataset instead of computing the exact centre (which might not be a right element in the datatype).

- The extension:
  - Generates decision tree models (distance-based decision tree) in the form of rules.
  - Conditions are expressed in terms of distances to prototypes (proximity rules: “like {economy, politics}”), but can simplified in some cases.
  - Can handle nominal, numerical and **structured (complex)** attributes.
    - Defining a metric or a similarity function for each attribute.

A General-purpose Algorithm: *DBDT*

```
DBDT(input L_Nodes)
    For each atribute x:
        L_Proto ← Compute_Prototypes(x)
        If size(L_Proto)>1
            L_Splits ← Splitting(L_Proto,Data) // proximity, density
        EndIf
    EndFor

    Best ← Select_Best_Split(L_Splits)  // IG, GR, Accuracy, GINI
    L_Nodes ← ApplyBestSplit(Best)
    DBDT(L_Nodes) // recursively explore the new nodes
```

DBDT for Web classification

Id.	Daily conn.	Structure	Content	Sport news site?
1	10	{(Math,Topo,Analysis,Logic)↔(invariant,surfaces), (Math,Topo,Analysis,Logic)↔(Lie ope,tangent), (Math,Topo,Analysis,Logic)↔(Gödel,Fuzzy)}	{(Topo,3), (Analysis,5),(Logic,5)}	No
2	25	{(Linux,networking) ↔(shell,learners), (Linux,networking) ↔(TCP/IP,telnet,ftp)}	{(Linux,3),(php,6), (networking,8)}	No
3	30	{(economy,politics) ↔(Dow Jones,Yen), (economy,politics) ↔(interview,elections)}	{(economy,3),(politics,4), (law,10)}	No
4	38	{(soccer,championships, leagues) ↔(scorers,classif.), (scorers,classif.) ↔(best players,best referees)}	{(soccer,9),(league,8)}	No
5	41	{(soccer,champions league) ↔(scorers,classif.), (soccer,champions league) ↔(matches,semi-final)}	{(soccer,7), (league,5)}	Yes
6	32	{(soccer,champions league) ↔(scorers,classif.), (soccer,champions league) ↔(matches,referees)}	{(soccer,5), (league,5)}	Yes

DBDT for Web classification

Id.	Daily conn.	Structure		Sport news site?
1	10	{(Math,Topo,Analysis,Logic)↔(invariant,surfaces), (Math,Topo,Analysis,Logic)↔(Lie ope,tangent), (Math,Topo,Analysis,Logic)↔(Gödel,Fuzzy)}		No
2	25	{(Linux,networking)↔(shell,learners), P/IP,telnet,ftp)}	(networking,8)}	No
3	30	ow Jones,Yen), view,elections)}		No
4	38	) ↔(scorers,classif.), ers,best referees)}		No
5	41	→(scorers,classif.), matches,semi-final)}	{(soccer,7), (league,5)}	Yes
6	32	{(soccer,champions league) ↔(scorers,classif.), (soccer,champions league) ↔(matches,referees)}		Yes

<META
Invariant,surfaces
Name=keywords>
<BODY>
<\BODY>

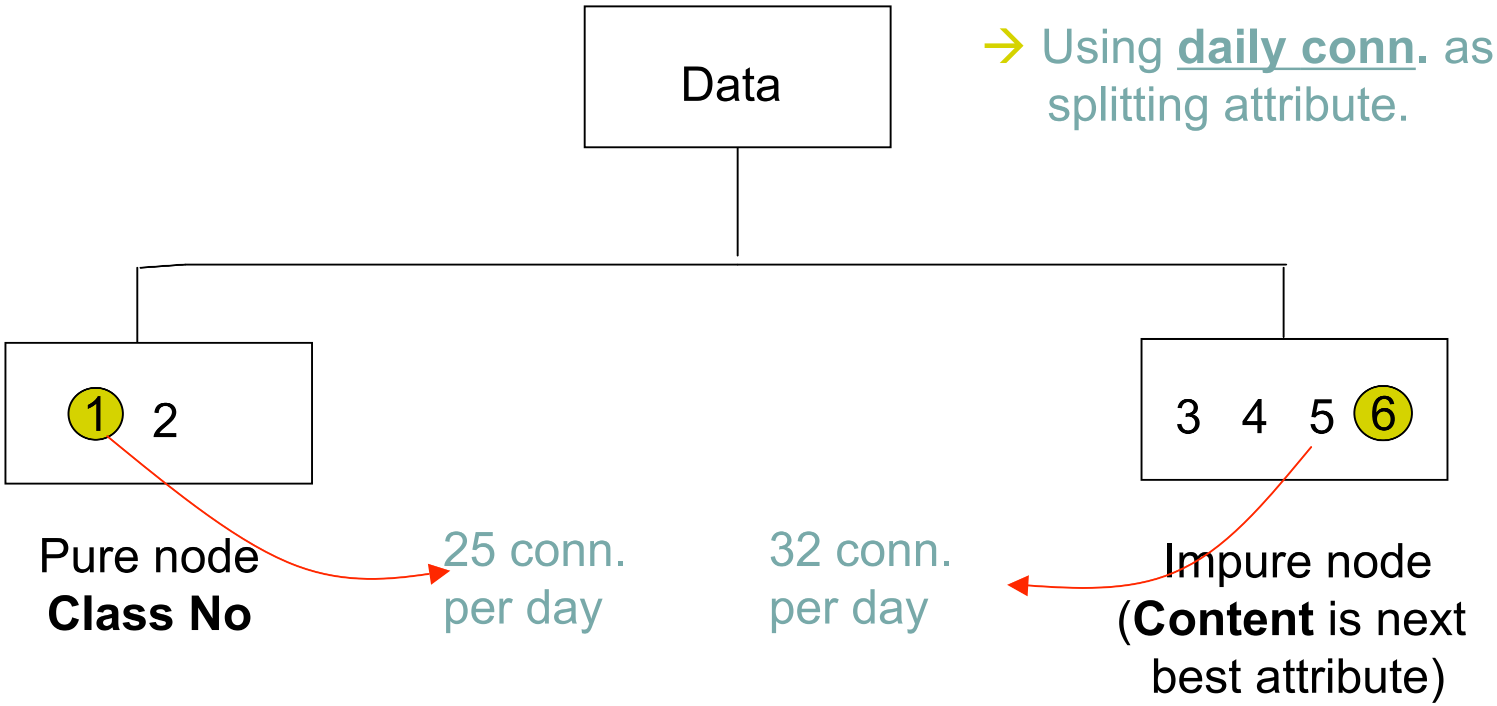
<META
Maths,Topo,Analysis,Logic
Name=keywords>
<BODY>
Topo
Logic …
<\BODY>

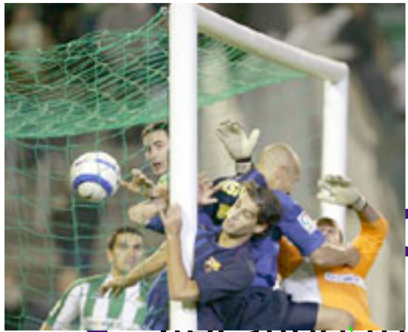
<META
Lie operator, tangent
Name=keywords>
<BODY>
<\BODY>

<META
Gödel, Fuzzy
Name=keywords>
<BODY>
<\BODY>

DBDT for Web classification

- After the 1st step… (heuristic: accuracy)

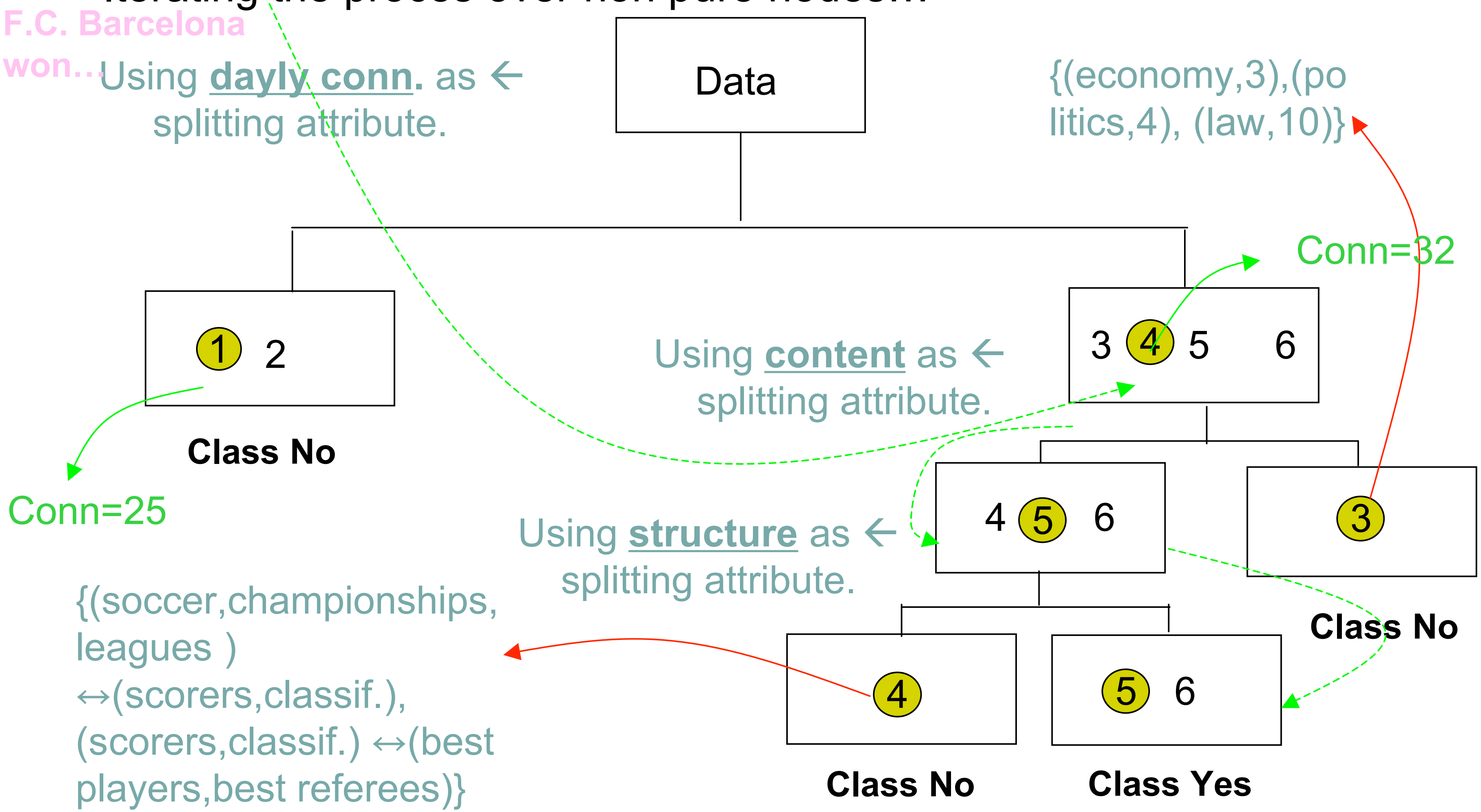




→ Daily conn. 40

...r Web classification

- ...iterating the proces over non pure nodes…



Experimental Evaluation of DBDT

- DBDT has been implemented in WEKA
- It includes several distance and (pseudo-)distance functions for nominal data, numerical data, lists and sets.
- In the experiments, lists and sets have been the structured data employed.
- Document representation
  - finite set of words (summary) from the title and the body selected according to its importance for classification.
  - the class label.

Experimental Evaluation of DBDT

- First experiment: classifying web sites by topic
  - 83 html documents downloaded from Internet
    - mathematics (biographies, technical pages, personal web sites, …)
    - sports (biographies, news, events, championships, …)

Num. of words	List (Acc. %)	Set (Acc. %)
50	100.0	93.6
75	98.1	91.5
100	95.9	91.4
125	98.4	94.8
150	97.6	92.5

Experimental Evaluation of DBDT

- Second experiment: Learning user profiles
  - *Syskill & Webert* data set (UCI repository)
    - several topics (clinical information, music events…)
    - documents ranked according to user preferences (hot, medium, cold)

Num. of words	Bands (Acc. %)	Biomedical (Acc. %)
50	74.5	71.5
75	79.7	81.3
100	77.9	83.0
125	82.0	84.0
150	81.7	79.6

Best Result (Bayesian method)

•*Biomedical*: 78.2%
•*Bands*: 74.6%

Best accurate result (DBDT)

Conclusions and Future Work

- *DBDT* is ...
  - A general-purpose algorithm
    - Like ID3, c4.5, CART, etc.
  - Able to handle structured attributes
    - Just necessary to define a similarity function for each attribute
  - Applicable for web mining
    - Web classification/categorisation problems
- Future work: How to transform the proximity rules into more “comprehensible” ones?
    - Instead of “close to {(economy,3),(politics,4), (law,10)}” we would prefer something like “having the words economy and politics”.
    - Defining a generalisation operator in metric spaces.

Conclusions and Future Work

- We plan to use the system for other web mining applications.
  - Recommender systems.
  - Personalisation.
  - Ontology categorisation (using metrics between ontologies).
  - …
- Other more general knowledge discovery areas (non-related to the project):
  - Extracting rules from incomprehensible models (black-box models).
  - Combination of data mining and simulation.
  - Applications in bioinformatics (complex data).
  - Ranking predictions and evaluating their quality.