RDM APPLICATIONS: AN OVERVIEW

Sašo Džeroski

Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, SLOVENIA

Outline

- RDM applications overview by areas
- Some more details on selected RDM applications

Types of applications

- Proof-of-the-principle applications
- Real-world applications

Proof-of-the-principle applications

- Related to important practical problems
- Formulate a generic task as a data mining problem
- Solve a benchmark/illustrative instance of the generic task
  This instance is (much) simpler than typical real-world problems

Real-world applications

- Real-world data available
- Domain expert interested in solving a specific practical problem
- Expert evaluation of the obtained results

Evaluation criteria

- Theories learned evaluated by a domain expert
- Theories learned used to solve new problem instances
  (testing on unseen cases)

Application areas

- Bioinformatics applications
  - Drug design
  - Predicting mutagenicity and carcinogenicity
  - Predicting protein structure and function
- Medical applications
- Environmental applications
- Traffic engineering applications

Application areas (continued)

- Mechanical engineering applications
- Text mining/Web mining/Natural language processing
- Business data analysis
- Various other applications
  - Software engineering
  - Music
  - Dynamic systems
    (control, design, diagnosis and modelling)
  - Adaptive system management

Bioinformatics applications

Greatest success of ILP/RDM

Three major classes of bioinformatics applications

- Drug design
- Predicting mutagenicity and carcinogenicity
- Predicting protein structure and function

A number of applications exists in each class

Drug design

- QSAR for pyrimidines and triazines (King et al. 1992)
- QSAR for Alzheimers disease drugs / tacrine analogues (King et al. 1995)
- SARs for modulating transmembrane calcium movement (Srinivasan and King 1996)
  - PROGOL, used to analyze a class of calcium-channel activators (molecules), generated a boolean feature that signifi cantly increases the correlation of two existing features with the activity variable. The fi nal result is comparable with much more complex models derived using computational chemistry.
- Diterpene structure elucidation (Džeroski et al. 1996, 1998)
- Pharmacophore discovery for ACE inhibition (Muggleton et al. 1996)
- Characterizing successful enantioseparations (Bryant 1997)

PHARMACOPHORE DISCOVERY FOR ACE INHIBITION

Identify the structure (pharmacophore) responsible for the activity of
ACE (Angiotensin-converting enzyme) inhibitors
from 28 molecules that display the activity of ACE inhibition

Molecules described by 3D-atom and bond information

Background knowledge about atom groups
and distances between pairs of groups

P-PROGOL discovered a four-piece pharmacophore
with one zinc-binder and three hydrogen acceptors
that is present in all molecules and
is equivalent to the generally accepted pharmacophore for ACE inhibition
(according to expert opinion)

DITERPENE STRUCTURE ELUCIDATION

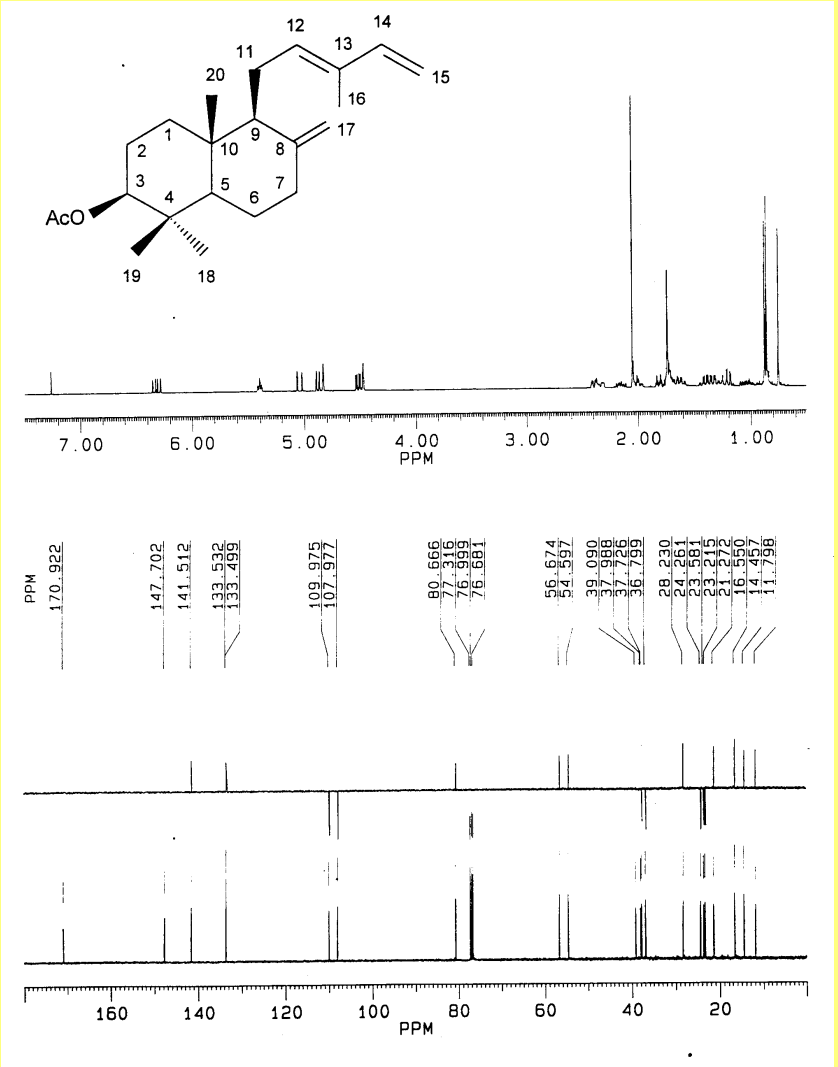
Diterpenes

- Natural organic compounds of significant commercial interest
- Used as lead compounds in the search for new drugs
- Have low molecular weight and a skeleton of 20 carbon atoms

The task

- Given is the ^{13}C NMR spectrum of the diterpene
  including the frequencies and multiplicities for each carbon atom
- The task is to classify the diterpene (spectrum)
  into the appropriate skeleton group

Diterpene structure elucidation (ctd.)



A diterpene belonging to the skeleton type labdan, with acetyl as a residue

At the bottom is the ^{13}C NMR spectrum, at the top the ^{1}H NMR spectrum

The additional measurements above the ^{13}C NMR spectrum show the multiplicity of carbon atoms

The multiplicity of a carbon atom is the number of hydrogen atoms directly connected to it (**s**ingulet, **d**oublet, **t**riplet, **q**uadruplet)

Diterpene structure elucidation (ctd.)

```
red(v1,t,39.00).
red(v1,t,19.30).
red(v1,t,42.20).
red(v1,s,33.30).
red(v1,d,55.60).
red(v1,t,24.30).
red(v1,t,38.30).
red(v1,d,148.60).
red(v1,d,57.20).
red(v1,s,39.70).
red(v1,t,17.60).
red(v1,t,41.40).
red(v1,d,73.50).
red(v1,t,145.10).
red(v1,q,111.40).
red(v1,q,27.90).
red(v1,q,106.30).
red(v1,q,33.60).
red(v1,q,21.60).
red(v1,q,14.40).
labdan(v1).
```

1503 examples (diterpene molecules with known skeleton groups and their spectra)

23 skeleton groups ($labdan$, $clerodan$, ...)

NMR spectra represented by frequencies of (peaks belonging to) carbon atoms and their multiplicities

Preprocessing corrects for multiplicity changes caused by residues (reduced multiplicities)

Matching problem.
Propositional version NOT straightforward.

Diterpene structure elucidation (ctd.)

ILP problem formulation

- Background predicate `red(MoleculeID,M,F)`
- 23 target predicates `labdan(MoleculeID)`, `clerodan(...`

Representation engineering

- For each of the four multiplicities, add the number of atoms with that multiplicity as a feature
- Background predicate `prop(NoOfSingulets,NoOfDublets,NoOfTriplets,NoOfQuadruplets)`

Diterpene structure elucidation - results

Example rule (TILDE), covers 347 of the 357 Labdans

```
c52 :- prop(A,B,C,D), A /= 4, A /= 3,
red(q, I), 20.97 <= I <= 59.5, B <= 4,
red(d, J), 49.12 <= J <= 93.8, D = 6.
```

Accuracy on unseen cases (10-fold CV)

Problem/System	FOIL	RIBL	TILDE	ICL	C4.5
red	46.5%	86.5%	81.6%	65.3%	N/A
prop	70.1%	79.0%	78.5%	79.1%	**78.5%**
red+prop	78.3%	**91.2%**	**90.4%**	86.0%	N/A

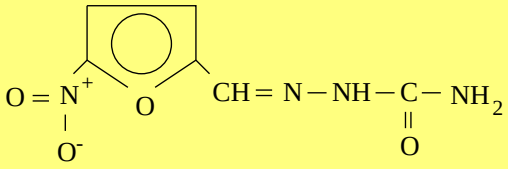
This is close to expert accuracies (which can only be estimated, as experts often use sources other than ^{13}C-NMR, e.g., ^{1}H-NMR)

Predicting mutagenicity and carcinogenicity

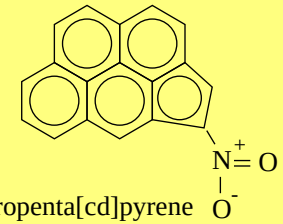
- Predicting carcinogenicity (King and Srinivasan 1996)
  - Data originates from US National Toxicology Program
  - Carcinogenicity assesed using rodent bioassays
  - PROGOL applied to discover structural alerts.
    Best result (of either human or machine origin) using only the data given.
    Comparable to methods that use extra information from biological rodent tests.
  - New compounds tested, challenge posed at IJCAI'97
  - ILP approaches turn out to perform surprisingly well in terms of predictive accuracy
  - Followed up by ECML/PKDD-2001 Predictive toxicology challenge
- Predicting mutagenicity of nitro-aromatic compounds
  (Srinivasan et al. 1994)

PREDICTING MUTAGENICITY

Nitro-aromatic compounds

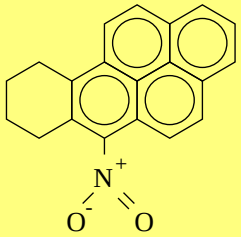


nitrofurazone

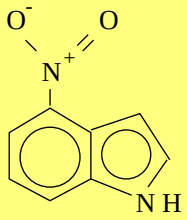


4-nitropenta[cd]pyrene

High mutagenicity



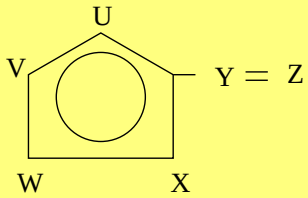
6-nitro-7,8,9,10-tetrahydrobenzo[a]pyrene



4-nitroindole

Low mutagenicity

New structural alert



Examples: 188 regression-friendly and 42 unfriendly compounds

Background knowledge: atoms, bonds, atom groups and properties

ILP system applied: PROGOL

A new structural alert for high mutagenicity discovered.

Better than regression on unfriendly set (88% vs 69%), comparable on regression-friendly set (88% vs 89%)

Predicting protein structure and function

- Predicting protein secondary structure
  (Muggleton et al. 1992; Mozetič and Hodošček 1997)
- Classifying protein domains into three-dimensional folds
  (Turcotte et al. 1998; 2001)
- Recognising neuropeptide precursor proteins (Muggleton et al. 2000)
- Genome scale prediction of protein functional class (King et al. 2000)
- mRNA signal structure detection (Horvath et al. 2001)

Medical applications

- Learning rules for early diagnosis of rheumatic diseases (Lavrač et al. 1993)
- Identifying glaucomatous eyes from ocular fundus images (Mizoguchi et al. 1997)
- Classifying X-ray angiograms of a patient’s cerebral vasculature (Sammut and Zrimec 1998)
- Modeling the therapeutic effects of drugs on patients in intensive care (Moriket al. 1999)
- Analysis of epidemiological data from San Francisco tuberculosis clinic (Getoor et al. 2001)

Applications in ecology

- Biological classification of river water quality (Džeroski et al. 1994)
- Modelling algal growth in the Lake of Bled (Kompare et al. 1997)
  - FORS applied to predict growth of maximal biomass quantity in the metalimnion of the east basin of the lake, using background knowledge about seasonal changes, multiplication and linear regression
  - An excellent model produced (expert's comment) that agrees with expert knowledge on algal growth in the lake and produces good predictions.
- Predicting biodegradation rates of chemical compounds (Džeroski et al. 1999)
- Predicting physical and chemical parameters of water quality in Slovenian rivers from bioindicator data (Blockeel et al. 1999)

BIOLOGICAL CLASSIFICATION OF BRITISH RIVERS

Given a list of biological indicators present at a sampling site (families of macro-benthic invertebrates) and their abundance levels
Classify the sample into one of five classes B1a, B1b, B2, B3, B4

Examples: 300 samples from the upper Trent catchment (British Midlands), classified by expert river ecologist

Background knowledge: relations between abundance levels

ILP systems applied: GOLEM, CLAUDIEN

Interesting rules discovered (according to expert evaluation)

$$b1b(X) \leftarrow ancylidae(X, A), gammaridae(X, B), hydropsychidae(X, C), \\ rhyacophilidae(A, D), greater\_than(B, A), greater\_than(B, D)$$

PREDICTING BIODEGRADATION RATES

- QSAR problem: predict the rate of aquatic biodegradation of chemicals from their structure represented by constituent atoms and bonds
- Database of 328 structurally diverse compounds
- Structure of compounds available in SMILES notation
- Background knowledge on functional groups
- Several ILP systems used (TILDE, SCART, ICL, FFOIL) as well as propositionalization + several propositional systems
- Better results obtained (correlation r=0.7) than state-of-the-art biodegradability prediction system based on linear regression (r=0.6)

Predicting biodegradation rates (ctd.)

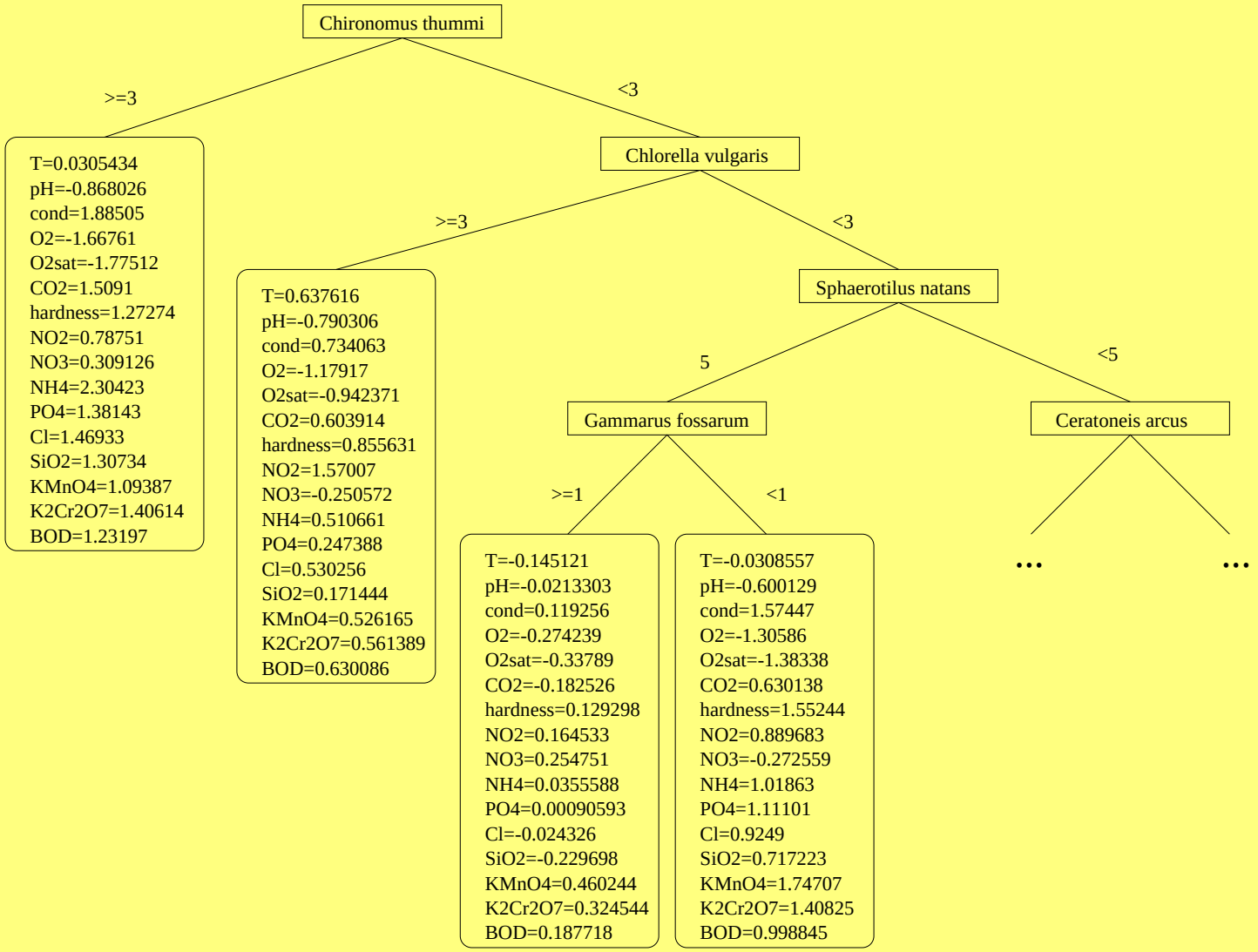
Example rule induced

A compound M degrades fast IF
M *contains* an atom A1 and
atom A1 is a nitrogen atom and
atom A1 *is connected to* atom A2 with bond B and
bond B is an aromatic bond and
the molecular weight of M is less than 110 units and
the logP value (hydrophobicity) of M is greater than zero

PREDICTING PHYSICAL/CHEMICAL PARAMETERS OF RIVER WATER QUALITY FROM BIOINDICATOR DATA

- Predictive clustering trees (TILDE) were used
- A single tree is built
- Simultaneously predicts the values of 16 physical and chemical parameters of river water quality from bioindicator data
- A biological sample is a set of organisms (taxa) present in the water at a given time (varying number of taxa)
- Better prediction accuracy achieved with single tree as compared to 16 trees (one for each target parameter)

Predicting physical/chemical parameters of river water quality from bioindicator data (ctd.)



Example tree

Applications in traffic engineering

- Identifying traffic problems in Spain (Džeroski et al. 1998)
- Classify road accidents in UK as ’young male driver to blame’ or otherwise (Roberts et al. 1998)
  - Goal: find out if YMDs are overrepresented in a particular type of accident or just have more accidents in general
  - Factual details: accident record, casualty record(s), vehicle record(s) Include road surface conditions, weather, vehicle maneuvering
  - Contributory factors report resulting from investigations More subjective information, e.g., who caused the accident, causes
  - Conclusion: YMDs have more accidents due to inexperience without mitigating circumstances

IDENTIFYING TRAFFIC PROBLEMS IN SPAIN

- Use sensor data (velocity, occupation, flow / saturation) and Road network structure as background knowledge
- Accident and congestion situations as examples
- Realistic simulator data on traffic around Barcelona used
- ILP systems used: ICL, PROGOL, TILDE; Compared to C4.5
- ILP systems perform much better, showing it is essential to take the (variable) road geometry into account
- Example rules:

  There is an accident at section A if

  the saturation at A is low and the occupation at the section preceding A is high.

  There is an accident at section A if

  the saturation is low and the occupation is high at A, whereas the velocity at the next section is high.

Applications in mechanical engineering

- Finite element mesh design (Dolšak et al. 1991-96)
- Electrical discharge machining (Karalič 1995)
  - FORS applied to model behavior of human operator controlling two variables: gap and flow. The rules derived for each variable combined in a model that
    a) is comprehensible to the domain expert and
    b) directs parameter changes towards the optimal process working region.
- Steel grinding (Karalič 1995)
  - FORS applied to learn rules for predicting workpiece roughness from properties of sound produced during grinding.
  - Induced rules involve inequalities between the frequencies of the maximum area peak and the spectrum area central point, as well as linear regression, and reveal domain properties not found by earlier statistical analysis and propositional learning methods.

FINITE ELEMENT MESH DESIGN

Given a geometric **structure** and **loadings/boundary conditions**
Find an **appropriate resolution** for a finite element mesh

Examples: ten structures with appropriate meshes (cca. 650 edges)

Background knowledge

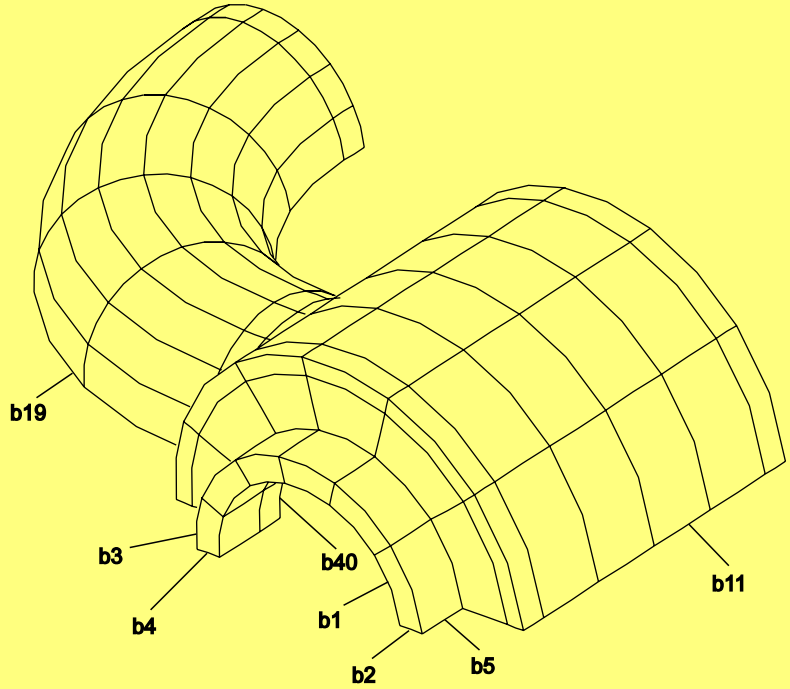
- Properties of edges (short, loaded, two-side-fixed, ...)
- Relations between edges (neighbor, opposite, equal)

ILP systems applied: GOLEM, CLAUDIEN

Many interesting rules discovered (according to expert evaluation)

Finite element mesh design (ctd.)

Example structure with an appropriate mesh



Example rules

$$mesh(Edge, 7) \leftarrow usual\_length(Edge),$$
$$\quad neighbour\_xy(Edge, EdgeY), two\_side\_fixed(EdgeY),$$
$$\quad neighbour\_zx(EdgeZ, Edge), not\_loaded(EdgeZ)$$
$$mesh(Edge, N) \leftarrow equal(Edge, Edge2), mesh(Edge2, N)$$

Web mining / Text mining

- Extracting knowledge bases from the web (Craven and Slattery 2001)
  - Classifying web pages
  - Learning prototypical patterns of hyperlink connectivity among pages
- Information extraction from text
  - Protein-related DB entries from MEDLINE (Craven 1999)
    For localization, associated diseases, drug interactions
  - Entries for a jobs DB from newsgroup postings
    (Califf and Mooney 1999)
- Text categorization (Cohen 1995)
- Information retrieval via text categorization
  (Dimec et al. 1999; Loggie 2000)

Applications in natural language processing

- Automated construction of natural language parsers (Zelle and Mooney 1995)
- Parsing natural language database queries (Zelle and Mooney 1996)
- Semantic lexicon acquisition for learning parsers (Thompson and Mooney 1997)
- Learning transfer rules (Bostrom and Zemke 1997)
  - Learn rules that transform a quasi logical form (QLF) of a sentence in English to a QLF of an equivalent sentence in Swedish, from example QLF pairs.

Applications in natural language processing (ctd.)

- Morphology
  - Generating diminutives of Dutch nouns (Dehaspe et al. 1995)
  - Generating the past tense of English verbs (Mooney and Califf 1995)
  - Nominal paradigms of Slovene nouns (Džeroski and Erjavec 1997)
  - Multilingual morphology (Manandhar et al. 1998)
- Learning grammars
  (Zelle and Mooney 1993, Dehaspe et al. 1995, Muggleton et al. 1996)
- Part-of-speech tagging (Cussens 1997; ...)
- For most recent NLP applications see (Cussens and Džeroski 2000)

LEARNING MORPHOLOGY - **Past tense of English verbs**

- Examples in orthographic form: `past([s,l,e,e,p],[s,l,e,p,t])`
- Background knowledge contains the predicate `split(Word,Prefix,Suffix)`, which works on nonempty lists
- An example decision list induced from 250 examples:

```
past([g,o], [w,e,n,t]) :- !.
past(A,B) :- split(A,C,[e,p]), split(B,C,[p,t]),!.
...
past(A,B) :- split(B,A,[d]), split(A,C,[e]),!.
past(A,B) :- split(B,A,[e,d]).
```

- FOIDL (Mooney and Califf 1995) achieves much higher accuracy on unseen cases as compared to a variety of propositional approaches

LEARNING MORPHOLOGY

Inflections of Slovene nouns (and adjectives)

- The Slovene language has a rich system of inflections
- Nouns in Slovene are lexically marked for **gender** (masculine, feminine or neuter) and inflect for **number** (singular, plural or dual) and **case** (nominative, genitive, dative, accusative, locative, instrumental)
- Learn analysis and synthesis rules
  (Synthesis: *base form* $\Rightarrow$ *oblique forms*,
  Analysis: *oblique forms* $\Rightarrow$ *base form*)
- Orwell's **1984** Corpus from COPERNICUS Project MULTEXT-East split into a training and testing set
- An overall accuracy of 94% achieved (97% analysis, 92% synthesis)
- Also done for other MULTEXT-East languages: CS, EN, ET, RO, SL

Business data analysis

PKDD-2000 discovery challenge: Analyze data from a relational database describing the operation of a Czech bank

Database contains eight tables

- Client and account information (*account*, *client*, *disposition*)
- Usage of products (*permanent order*, *transaction*, *loan*, *credit card*)
- Demographic information about 77 Czech districts (*demographic data*)

Data mining task: determine loan quality of account

(Knobbe et al. 2001) use propositionalization by aggregation and C5.0 and achieve very high predictive accuracy

Various other applications

- Software engineering
- Music
- Dynamic systems
  (control, design, diagnosis and modelling)
- Adaptive system management

Applications in software engineering

- Program construction (Grobelnik 1992, Bergadano et al. 1993-96)
- Inducing invariants for program verification (Bratko and Grobelnik 1993)
- Inductive test case generation (Bergadano et al. 1993)
- Recovering an abstract specification of a large software system (Cohen 1994)
- Software fault prediction (Cohen and Devanbu 1997)
  - FOIL and FLIPPER applied to predict whether a fault occurs in a C++ class, given background knowledge on coupling relationships to other C++ classes

Applications in music

- Analysis of Rahmaninoff's piano performance (Dovey 1995)
- Analysis and prediction of piano performances (Van Baelen and De Raedt 1996)
  - CLAUDIEN used to induce theories for predicting MIDI files from the musical analysis of Mendelssohn's Lied Ohne Worte.
- Composing the two-voice counterpoint (Pompe et al. 1996)
  - SFOIL applied to learn rules that help generate the counterpoint melody from the cantus firmus melody.
  - The learned rules together with some rules from musical theory produced counterpoints that professional musicians and nonmusicians judge to be at the level of an acceptable junior composer.

Applications in dynamic systems (control, design, diagnosis and modelling)

- Learning diagnostic rules for a satelite power supply (Feng 1992)
- Learning qualitative models from example behaviors (Bratko et al. 1992, Džeroski 1992)
- Design from first principles (Bratko 1993)
- Learning flight stage transition rules (Camacho 1994)
- Learning pole balancing (Džeroski et al. 1995)
- Learning relational concepts from sensor data of a mobile robot (Klingspor et al. 1996). The ILP system GRDT is applied to learn abstract operational concepts, e.g., 'the robot moves along a wall', from sequences of sonar sensor measurements and robot actions.

ADAPTIVE SYSTEMS MANAGEMENT
(Knobbe et al. 2000)

- Identifying causes of low performance in complex computer networks
- Upholding service level agreements (SLAs), which prescribe, e.g., maximium database access times
- Data from performance monitors on individual network components
- Structure of the network (connections between components) as background knowledge
- Typical rule:
  Low performance if any NSF server has a monitor with high value
  Propositional rules involve particular servers and particular monitors, e.g., high-usage of disk 14 on server 11

Summary

- Decreasing number of proof-of-the principle,
  increasing number of real applications
- A steady overall increase in the number of RDM applications
- Several of these have generated useful new knowledge
- In several domains RDM clearly outperforms propositional learners
- Most successes in the domain of bioinformatics
- Slow start on business data analysis
- Increased interest for RDM applications in text and web mining