Part II. Descriptive Data Mining

- Subgroup discovery
- Relational data mining
- Association rule learning
Predictive vs. descriptive induction

- **Predictive induction**: Inducing classifiers for solving classification and prediction tasks,
  - Classification rule learning, Decision tree learning, ...
  - Bayesian classifier, ANN, SVM, ...
  - Data analysis through hypothesis generation and testing

- **Descriptive induction**: Discovering interesting regularities in the data, uncovering patterns, ... for solving KDD tasks
  - Association rule learning, Subgroup discovery, Symbolic clustering, ...
  - Exploratory data analysis
Descriptive DM

- Often used for preliminary explanatory data analysis
- User gets feel for the data and its structure
- Aims at deriving descriptions of characteristics of the data
- Visualization and descriptive statistical techniques can be used
Subgroup Discovery

- A task in which individual interpretable patterns in the form of rules are induced from data, labeled by a predefined property of interest.
- SD algorithms learn several independent rules that describe groups of target class examples
  - subgroups must be large and significant
Classification versus Subgroup Discovery

- **Classification** (predictive induction) - constructing sets of classification rules
  - aimed at learning a model for classification or prediction
  - rules are dependent

- **Subgroup discovery** (descriptive induction) – constructing individual subgroup describing rules
  - aimed at finding interesting patterns in target class examples
    - large subgroups (high target class coverage)
    - with significantly different distribution of target class examples (high TP/FP ratio, high significance, high WRAcc)
  - each rule (pattern) is an independent chunk of knowledge
Classification versus Subgroup discovery
Subgroup discovery in High CHD Risk Group Detection

**Input:** Patient records described by anamnestic, laboratory and ECG attributes

**Task:** Find and characterize population subgroups with high CHD risk (large enough, distributionally unusual)

From **best induced descriptions**, five were selected by the expert as **most actionable** for CHD risk screening (by GPs):

- high-CHD-risk $\leftarrow$ male & pos. fam. history & age > 46
- high-CHD-risk $\leftarrow$ female & bodymassIndex > 25 & age > 63
- high-CHD-risk $\leftarrow$ ...
- high-CHD-risk $\leftarrow$ ...
- high-CHD-risk $\leftarrow$ ...

(Gamberger & Lavrač, JAIR 2002)
Subgroup Discovery: Medical Use Case

- Find and characterize population subgroups with high risk for coronary heart disease (CHD) (Gamberger, Lavrač, Krstačić)

- **A1** for males: principal risk factors
  
  CHD $\leftarrow$ pos. fam. history & age $> 46$

- **A2** for females: principal risk factors
  
  CHD $\leftarrow$ bodyMassIndex $> 25$ & age $> 63$

- **A1, A2** (anamnestic info only), **B1, B2** (an. and physical examination), **C1** (an., phy. and ECG)

- **A1:** supporting factors (found by statistical analysis): psychosocial stress, as well as cigarette smoking, hypertension and overweight
Subgroup discovery in functional genomics

- Functional genomics is a typical scientific discovery domain, studying genes and their functions
- Very large number of attributes (genes)
- Interesting subgroup describing patterns discovered by SD algorithm

CancerType = Leukemia
IF KIAA0128 = DIFF. EXPRESSED
AND prostoglandin d2 synthase = NOT_ DIFF. EXPRESSED

- Interpretable by biologists

D. Gamberger, N. Lavrač, F. Železný, J. Tolar
Journal of Biomedical Informatics 37(5):269-284, 2004
Subgroups vs. classifiers

• Classifiers:
  – Classification rules aim at pure subgroups
  – A set of rules forms a domain model

• Subgroups:
  – Rules describing subgroups aim at significantly higher proportion of positives
  – Each rule is an independent chunk of knowledge

• Link
  – SD can be viewed as cost-sensitive classification
  – Instead of $FN_{cost}$ we aim at increased $TP_{profit}$
Classification Rule Learning for Subgroup Discovery: Deficiencies

• Only first few rules induced by the covering algorithm have sufficient support (coverage)

• Subsequent rules are induced from smaller and strongly biased example subsets (pos. examples not covered by previously induced rules), which hinders their ability to detect population subgroups

• ‘Ordered’ rules are induced and interpreted sequentially as a if-then-else decision list
CN2-SD: Adapting CN2 Rule Learning to Subgroup Discovery

- Weighted covering algorithm
- Weighted relative accuracy (WRAcc) search heuristics, with added example weights
- Probabilistic classification
- Evaluation with different interestingness measures
CN2-SD: CN2 Adaptations

- General-to-specific search (beam search) for best rules
- Rule quality measure:
  - CN2: Laplace: \( \text{Acc}(\text{Class} \leftarrow \text{Cond}) = \frac{\text{p(Class|Cond)}}{\text{WRAcc}(\text{Class} \leftarrow \text{Cond})} = \frac{(n_c+1)}{(n_{\text{rule}}+k)} \)
  - CN2-SD: Weighted Relative Accuracy
    \[ \text{WRAcc}(\text{Class} \leftarrow \text{Cond}) = \frac{\text{p(Cond)} (\text{p(Class|Cond)} - \text{p(Class))}}{\text{p(Cond)}} \]
- Weighted covering approach (example weights)
- Significance testing (likelihood ratio statistics)
- Output: Unordered rule sets (probabilistic classification)
CN2-SD: Weighted Covering

• Standard covering approach: covered examples are deleted from current training set

• Weighted covering approach:
  – weights assigned to examples
  – covered pos. examples are re-weighted:
    in all covering loop iterations, store count $i$ how many times (with how many rules induced so far) a pos. example has been covered: $w(e,i)$, $w(e,0)=1$

• Additive weights: $w(e,i) = 1/(i+1)$
  $w(e,i)$ – pos. example $e$ being covered $i$ times
Subgroup Discovery

Positive examples

Negative examples
Subgroup Discovery

Positive examples

Rule1: Cl=+ ⇐ Cond6 AND Cond2

Negative examples
Subgroup Discovery

Positive examples

Negative examples

Rule 2: Cl=+ ← Cond3 AND Cond4
Subgroup Discovery

Positive examples

Negative examples
CN2-SD: Weighted WRAcc Search Heuristic

• Weighted relative accuracy (WRAcc) search heuristics, with added example weights

\[
WRAcc(Cl \leftarrow Cond) = p(Cond) \cdot (p(Cl|Cond) - p(Cl))
\]

increased coverage, decreased # of rules, approx. equal accuracy (PKDD-2000)

• In WRAcc computation, probabilities are estimated with relative frequencies, adapt:

\[
WRAcc(Cl \leftarrow Cond) = p(Cond) \cdot (p(Cl|Cond) - p(Cl)) = \frac{n'(Cond)}{N'} \cdot \left( \frac{n'(Cl.Cond)}{n'(Cond)} - \frac{n'(Cl)}{N'} \right)
\]

- \( N' \) : sum of weights of examples
- \( n'(Cond) \) : sum of weights of all covered examples
- \( n'(Cl.Cond) \) : sum of weights of all correctly covered examples
SD algorithms in the Orange DM Platform

- **Orange** data mining toolkit
  - classification and subgroup discovery algorithms
  - data mining workflows
  - visualization

**SD Algorithms in Orange**

- SD (Gamberger & Lavrač, JAIR 2002)
- Apriori-SD (Kavšek & Lavrač, AAI 2006)
- CN2-SD (Lavrač et al., JMLR 2004): Adapting CN2 classification rule learner to Subgroup Discovery
Association Rule Learning

Rules: $X \Rightarrow Y$, if $X$ then $Y$

$X$ and $Y$ are itemsets (records, conjunction of items), where items/features are binary-valued attributes.

Given: Transactions

<table>
<thead>
<tr>
<th></th>
<th>i1</th>
<th>i2</th>
<th>...</th>
<th>i50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>1</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>$t_2$</td>
<td>0</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

Find: A set of association rules in the form $X \Rightarrow Y$

Example: Market basket analysis

beer & coke $\Rightarrow$ peanuts & chips $(0.05, 0.65)$

- Support: $\text{Sup}(X,Y) = \frac{\#XY}{\#D} = p(XY)$
- Confidence: $\text{Conf}(X,Y) = \frac{\#XY}{\#X} = \frac{\text{Sup}(X,Y)}{\text{Sup}(X)} = \frac{p(XY)}{p(X)} = p(Y|X)$
Association Rule Learning: Examples

• Market basket analysis
  – beer & coke ⇒ peanuts & chips (5%, 65%)
    (IF beer AND coke THEN peanuts AND chips)
  – Support 5%: 5% of all customers buy all four items
  – Confidence 65%: 65% of customers that buy beer and coke also buy peanuts and chips

• Insurance
  – mortgage & loans & savings ⇒ insurance (2%, 62%)
  – Support 2%: 2% of all customers have all four
  – Confidence 62%: 62% of all customers that have mortgage, loan and savings also have insurance
Association Rule Learning

**Given:** a set of transactions D

**Find:** all association rules that hold on the set of transactions that have

- user defined minimum support, i.e., support > \( \text{MinSup} \), and
- user defined minimum confidence, i.e., confidence > \( \text{MinConf} \)

It is a form of exploratory data analysis, rather than hypothesis verification
Searching for the associations

- Find all large itemsets
- Use the large itemsets to generate association rules
- If $XY$ is a large itemset, compute
  \[ r = \frac{\text{support}(XY)}{\text{support}(X)} \]
- If $r > \text{MinConf}$, then $X \Rightarrow Y$ holds
  (support $> \text{MinSup}$, as $XY$ is large)
Large itemsets

• Large itemsets are itemsets that appear in at least MinSup transaction
• All subsets of a large itemset are large itemsets (e.g., if A,B appears in at least MinSup transactions, so do A and B)
• This observation is the basis for very efficient algorithms for association rules discovery (linear in the number of transactions)
Association vs. Classification rules

- Exploration of dependencies
- Different combinations of dependent and independent attributes
- Complete search (all rules found)

rules

- Focused prediction
- Predict one attribute (class) from the others
- Heuristic search (subset of rules found)
MEDIANA – analysis of media research data

- Questionnaires about journal/magazine reading, watching of TV programs and listening of radio programs, since 1992, about 1200 questions. Yearly publication: frequency of reading/listening/watching, distribution w.r.t. Sex, Age, Education, Buying power,..

- Data for one year, about 8000 questionnaires, covering lifestyle, spare time activities, personal viewpoints, reading/listening/watching of media (yes/no/how much), interest for specific topics in media, social status

- good quality, “clean” data

- table of n-tuples (rows: individuals, columns: attributes, in classification tasks selected class)
MEDIANA – media research pilot study

• Patterns uncovering regularities concerning:
  – Which other journals/magazines are read by readers of a particular journal/magazine?
  – What are the properties of individuals that are consumers of a particular media offer?
  – Which properties are distinctive for readers of different journals?

• Induced models: description (association rules, clusters) and classification (decision trees, classification rules)
Simplified association rules

Finding profiles of readers of the Delo daily newspaper

reads_Marketing_magazine 116 $\Rightarrow$ reads_Delo 95 (0.82)
reads_Financial_News (Finance) 223 $\Rightarrow$ reads_Delo 180 (0.81)
reads_Money (Denar) 197 $\Rightarrow$ reads_Delo 150 (0.76)
reads_Vip 181 $\Rightarrow$ reads_Delo 134 (0.74)

Interpretation: Most readers of Marketing magazine, Financial News, Money and Vip read also Delo.
Simplified association rules

reads_Sara 332 ⇒ reads_Slovenske novice 211 (0.64)
reads_Ljubezenske zgodbe 283 ⇒
  reads_Slovenske novice 174 (0.61)
reads_Dolenjski list 520 ⇒
  reads_Slovenske novice 310 (0.6)
reads_Omama 154 ⇒ reads_Slovenske novice 90 (0.58)
reads_Delavska enotnost 177 ⇒
  reads_Slovenske novice 102 (0.58)

Most of the readers of Sara, Love stories, Dolenjska new, Omama in Workers new read also Slovenian news.
Simplified association rules

reads_Sportske novosti 303 \(\rightarrow\)
    reads_Slovenski delnicar 164 (0.54)
reads_Sportske novosti 303 \(\rightarrow\)
    reads_Salomonov oglasnik 155 (0.51)
reads_Sportske novosti 303 \(\rightarrow\)
    reads_Lady 152 (0.5)

More than half of readers of Sports news reads also Slovenian shareholders magazine, Solomon advertisements and Lady.
Given: a relational database, a set of tables, sets of logical facts, a graph, ...

Find: a classification model, a set of patterns
Relational Data Mining

- ILP, relational learning, relational data mining
- Learning from complex relational databases

<table>
<thead>
<tr>
<th>customer</th>
<th>ID</th>
<th>Zip</th>
<th>S gender</th>
<th>S gender</th>
<th>Income</th>
<th>Age</th>
<th>City</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>3478</td>
<td>34677</td>
<td>8587</td>
<td>female</td>
<td>male</td>
<td>60-70</td>
<td>32</td>
<td>small</td>
<td>suburban</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>order</th>
<th>Customer ID</th>
<th>Order ID</th>
<th>Store ID</th>
<th>Delivery Mode</th>
<th>Payment Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>3478</td>
<td>3446778</td>
<td>12</td>
<td>regular</td>
<td>cash</td>
<td></td>
</tr>
<tr>
<td>3478</td>
<td>47323861</td>
<td>17</td>
<td>express</td>
<td>check</td>
<td></td>
</tr>
<tr>
<td>3479</td>
<td>32334461</td>
<td>17</td>
<td>regular</td>
<td>check</td>
<td></td>
</tr>
<tr>
<td>3479</td>
<td>34758886</td>
<td>12</td>
<td>express</td>
<td>credit</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>store</th>
<th>Store ID</th>
<th>Size</th>
<th>Type</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>small</td>
<td>city</td>
</tr>
<tr>
<td>12</td>
<td>...</td>
<td>...</td>
<td>large</td>
<td>rural</td>
</tr>
</tbody>
</table>

Relational representation of customers, orders and stores.
Relational Data Mining (RDM)

- ILP, relational learning, relational data mining
  - Learning from complex relational databases
  - Learning from complex structured data, e.g. molecules and their biochemical properties
Relational and Semantic data mining

- ILP, relational learning, relational data mining
  - Learning from complex relational databases
  - Learning from complex structured data, e.g. molecules and their biochemical properties
  - Learning by using domain knowledge in the form of ontologies = semantic data mining
Propositionalization approach to RDM

Step 1

1. constructing relational features
2. constructing a propositional table
Propositionalization approach to RDM

Propositionalization

1. constructing relational features
2. constructing a propositional table

Step 2
Data Mining

model, patterns, …
Sample relational problem: East-West trains

1. **TRAINS GOING EAST**

1. 

2. 

3. 

4. 

5. 

2. **TRAINS GOING WEST**

1. 

2. 

3. 

4. 

5.
Relational data representation

<table>
<thead>
<tr>
<th>LOAD</th>
<th>CAR</th>
<th>OBJECT</th>
<th>NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>l1</td>
<td>c1</td>
<td>circle</td>
<td>1</td>
</tr>
<tr>
<td>l2</td>
<td>c2</td>
<td>hexagon</td>
<td>1</td>
</tr>
<tr>
<td>l3</td>
<td>c3</td>
<td>triangle</td>
<td>1</td>
</tr>
<tr>
<td>l4</td>
<td>c4</td>
<td>rectangle</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TRAIN</th>
<th>EASTBOUND</th>
</tr>
</thead>
<tbody>
<tr>
<td>t 1</td>
<td>TRUE</td>
</tr>
<tr>
<td>t 2</td>
<td>TRUE</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>t 6</td>
<td>FALSE</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CAR</th>
<th>TRAIN</th>
<th>SHAPE</th>
<th>LENGTH</th>
<th>ROOF</th>
<th>WHEELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>t1</td>
<td>rectangle</td>
<td>short</td>
<td>none</td>
<td>2</td>
</tr>
<tr>
<td>c2</td>
<td>t1</td>
<td>rectangle</td>
<td>long</td>
<td>none</td>
<td>3</td>
</tr>
<tr>
<td>c3</td>
<td>t1</td>
<td>rectangle</td>
<td>short</td>
<td>peaked</td>
<td>2</td>
</tr>
<tr>
<td>c4</td>
<td>t1</td>
<td>rectangle</td>
<td>long</td>
<td>none</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Diagram showing the relational data representation of a train, with tables for car shapes and train details.
Relational data representation
Propositionalization in a nutshell

Propositionalization task

**Transform** a multi-relational (**multiple-table**) representation to a propositional representation (**single table**)

Proposed in ILP systems
LINUS (Lavrac et al. 1991, 1994),
1BC (Flach and Lachiche 1999), RSD (Železny and Lavrac)
Propositionalization in a nutshell

Main propositionalization step: first-order feature construction

f₁(T):- hasCar(T,C), clength(C,short).

f₂(T):- hasCar(T,C), hasLoad(C,L), loadShape(L,circle)

f₃(T):- ....

Propositional learning:

t(T) ← f₁(T), f₄(T)

Relational interpretation:

eastbound(T) ← hasShortCar(T), hasClosedCar(T).
Relational Data Mining in Orange4WS

- service for propositionalization through efficient first-order feature construction (Železny and Lavrač, MLJ 2006)
  
  \[
  f_{121}(M) :- \text{hasAtom}(M,A), \text{atomType}(A,21) \\
  f_{235}(M) :- \text{lumo}(M,Lu), \text{lessThr}(Lu,1.21)
  \]

- subgroup discovery using CN2-SD
  
  mutagenic(M) ← feature121(M), feature235(M)
What is Semantic Data Mining

SDM task definition

Given:
- transaction data table, relational database, text documents, Web pages, ...
- one or more domain ontologies

Find: a classification model, a set of patterns
Using domain ontologies (e.g. Gene Ontology) as background knowledge for Data Mining

Gene Ontology

12093 biological process
1812 cellular components
7459 molecular functions

Joint work with Igor Trajkovski and Filip Zelezny
Using domain ontologies (e.g. Gene Ontology) as background knowledge for Data Mining

First-order features, describing gene properties and relations between genes, can be viewed as generalisations of individual genes.
First order feature construction

First order features with support > $\textit{min\_support}$

\begin{align*}
f(7, A) & :\text{-function}(A, 'GO:0046872'). \\
f(8, A) & :\text{-function}(A, 'GO:0004871'). \\
f(11, A) & :\text{-process}(A, 'GO:0007165'). \\
f(14, A) & :\text{-process}(A, 'GO:0044267'). \\
f(15, A) & :\text{-process}(A, 'GO:0050874'). \\
f(20, A) & :\text{-function}(A, 'GO:0004871'), \text{ process}(A, 'GO:0050874'). \\
f(26, A) & :\text{-component}(A, 'GO:0016021'). \\
f(29, A) & :\text{-function}(A, 'GO:0046872'), \text{ component}(A, 'GO:0016020'). \\
f(122, A) & :\text{-interaction}(A, B), \text{ function}(B, 'GO:0004872'). \\
f(223, A) & :\text{-interaction}(A, B), \text{ function}(B, 'GO:0004871'), \\
& \text{ process}(B, 'GO:0009613'). \\
f(224, A) & :\text{-interaction}(A, B), \text{ function}(B, 'GO:0016787'), \\
& \text{ component}(B, 'GO:0043231').
\end{align*}
Propositionalization

<table>
<thead>
<tr>
<th>diffexp g1 (gene64499)</th>
<th>random g1 (gene7443)</th>
</tr>
</thead>
<tbody>
<tr>
<td>diffexp g2 (gene2534)</td>
<td>random g2 (gene9221)</td>
</tr>
<tr>
<td>diffexp g3 (gene5199)</td>
<td>random g3 (gene2339)</td>
</tr>
<tr>
<td>diffexp g4 (gene1052)</td>
<td>random g4 (gene9657)</td>
</tr>
<tr>
<td>diffexp g5 (gene6036)</td>
<td>random g5 (gene19679)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
<th>f6</th>
<th>...</th>
<th>...</th>
<th>fn</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>g3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Propositional learning: subgroup discovery

<table>
<thead>
<tr>
<th></th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
<th>f6</th>
<th>...</th>
<th></th>
<th>...</th>
<th>fn</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>g3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>g4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>g5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Over-expressed

IF f2 and f3 [4,0]

diffexp(A) :- interaction(A,B) & function(B,'GO:0004871')
Semantic Data Mining in two steps

- **Step 1:** Construct relational logic features of genes such as:
  \[
  \text{interaction}(g, G) \land \text{function}(G, \text{protein}\_\text{binding})
  \]
  
  \((g \text{ interacts with another gene whose functions include protein binding})\)

  and propositional table construction with features as attributes

- **Step 2:** Using these features to discover and describe subgroups of genes that are differentially expressed (e.g., belong to class DIFF.EXP. of top 300 most differentially expressed genes) in contrast with RANDOM genes (randomly selected genes with low differential expression).

- Sample subgroup description:
  
  \[
  \text{diffexp}(A) \leftarrow \text{interaction}(A,B) \land \text{function}(B,'GO:0004871') \land \text{process}(B,'GO:0009613')
  \]
  
  \((A \text{ differentially expressed gene})\)
Semantic Data Mining

• Semantic subgroup discovery (Vavpetič et al., 2012)