Advances in Subgroup Discovery for Biomedical Research

Nada Lavrač

Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana
Slovenia

ML & CC, Ljubljana, 4 July 2013
Talk outline

• Knowledge Technologies at JSI

First Generation Data Mining
  – Background: Data Mining in a Nutshell

• Second Generation Data Mining
  – Selected Algorithms and Biomedical Applications

• Towards Third Generation Data Mining
  – A Service-oriented Knowledge Discovery platform
  – Semantic Data Mining: Using ontologies in DM

• Current work
Data Mining in a Nutshell

Data Mining

Given: transaction data table, relational database, text documents, Web pages

Find: a classification model, a set of interesting patterns

<table>
<thead>
<tr>
<th>Person</th>
<th>Age</th>
<th>Spect. presc.</th>
<th>Astigm.</th>
<th>Tear prod.</th>
<th>Lenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>17</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O2</td>
<td>23</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>SOFTWARE</td>
</tr>
<tr>
<td>O3</td>
<td>22</td>
<td>myope</td>
<td>yes</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O4</td>
<td>27</td>
<td>myope</td>
<td>yes</td>
<td>normal</td>
<td>HARD</td>
</tr>
<tr>
<td>O5</td>
<td>19</td>
<td>hypermetrope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O6-O13</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O14</td>
<td>35</td>
<td>hypermetrope</td>
<td>no</td>
<td>normal</td>
<td>SOFTWARE</td>
</tr>
<tr>
<td>O15</td>
<td>43</td>
<td>hypermetrope</td>
<td>yes</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O16</td>
<td>39</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>NONE</td>
</tr>
<tr>
<td>O17</td>
<td>54</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O18</td>
<td>62</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>NONE</td>
</tr>
<tr>
<td>O19-O23</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O24</td>
<td>56</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>NONE</td>
</tr>
</tbody>
</table>
Data Mining in a Nutshell

Given: transaction data table, relational database, text documents, Web pages
Find: a classification model, a set of interesting patterns

knowledge discovery from data

Data Mining

model, patterns, …

new unclassified instance
classified instance
black box classifier no explanation

symbolic model symbolic patterns explanation
Example: Learning a classification model from contact lens data

<table>
<thead>
<tr>
<th>Person</th>
<th>Age</th>
<th>Spect. presc.</th>
<th>Astigm.</th>
<th>Tear prod.</th>
<th>Lenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>17</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O2</td>
<td>23</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>SOFT</td>
</tr>
<tr>
<td>O3</td>
<td>22</td>
<td>myope</td>
<td>yes</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O4</td>
<td>27</td>
<td>myope</td>
<td>yes</td>
<td>normal</td>
<td>HARD</td>
</tr>
<tr>
<td>O5</td>
<td>19</td>
<td>hypermetrope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O6-O13</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O14</td>
<td>35</td>
<td>hypermetrope</td>
<td>no</td>
<td>normal</td>
<td>SOFT</td>
</tr>
<tr>
<td>O15</td>
<td>43</td>
<td>hypermetrope</td>
<td>yes</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O16</td>
<td>39</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>NONE</td>
</tr>
<tr>
<td>O17</td>
<td>54</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O18</td>
<td>62</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>NONE</td>
</tr>
<tr>
<td>O19-O23</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O24</td>
<td>56</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>NONE</td>
</tr>
</tbody>
</table>
Example: Learning a classification model from contact lens data

<table>
<thead>
<tr>
<th>Person</th>
<th>Age</th>
<th>Spect. presc.</th>
<th>Astigm.</th>
<th>Tear prod.</th>
<th>Lenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>17</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O2</td>
<td>23</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>SOFT</td>
</tr>
<tr>
<td>O3</td>
<td>22</td>
<td>myope</td>
<td>yes</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O4</td>
<td>27</td>
<td>myope</td>
<td>yes</td>
<td>normal</td>
<td>HARD</td>
</tr>
<tr>
<td>O5</td>
<td>19</td>
<td>hypermetrope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O6-O13</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O14</td>
<td>35</td>
<td>hypermetrope</td>
<td>no</td>
<td>normal</td>
<td>SOFT</td>
</tr>
<tr>
<td>O15</td>
<td>43</td>
<td>hypermetrope</td>
<td>yes</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O16</td>
<td>39</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>NONE</td>
</tr>
<tr>
<td>O17</td>
<td>54</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O18</td>
<td>62</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>NONE</td>
</tr>
<tr>
<td>O19-O23</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O24</td>
<td>56</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>NONE</td>
</tr>
</tbody>
</table>

Data Mining

- **tear prod.**: reduced, normal
- **astigmatism**: no, yes
- **spect. pre.**: myope, hypermetrope
- **Lenses**: NONE, SOFT, HARD, NONE
### Example: Learning a classification model from contact lens data

<table>
<thead>
<tr>
<th>Person</th>
<th>Age</th>
<th>Spect. presc.</th>
<th>Astigm.</th>
<th>Tear prod.</th>
<th>Lenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>17</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O2</td>
<td>23</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>SOFT</td>
</tr>
<tr>
<td>O3</td>
<td>22</td>
<td>myope</td>
<td>yes</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O4</td>
<td>27</td>
<td>myope</td>
<td>yes</td>
<td>normal</td>
<td>HARD</td>
</tr>
<tr>
<td>O5</td>
<td>19</td>
<td>hypermetrope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O6-O13</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O14</td>
<td>35</td>
<td>hypermetrope</td>
<td>no</td>
<td>normal</td>
<td>SOFT</td>
</tr>
<tr>
<td>O15</td>
<td>43</td>
<td>hypermetrope</td>
<td>yes</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O16</td>
<td>39</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>NONE</td>
</tr>
<tr>
<td>O17</td>
<td>54</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O18</td>
<td>62</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>NONE</td>
</tr>
<tr>
<td>O19-O23</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O24</td>
<td>56</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>NONE</td>
</tr>
</tbody>
</table>

- \text{lenses=NONE} \leftarrow \text{tear production=red}
- \text{lenses=NONE} \leftarrow \text{tear production=normal AND astigmatism=yes AND spect. pre.=hypermetrope}
- \text{lenses=SOFT} \leftarrow \text{tear production=normal AND astigmatism=no}
- \text{lenses=HARD} \leftarrow \text{tear production=normal AND astigmatism=yes AND spect. pre.=myope}
- \text{lenses=NONE} \leftarrow
First Generation Data Mining

• **First machine learning algorithms for**
  – Decision tree and rule learning in 1970s and early 1980s by Quinlan, Michalski et al., Breiman et al., …

• **Characterized by**
  – Learning from data stored in a single data table
  – Relatively small set of instances and attributes

• **Lots of ML research followed in 1980s**
  – Numerous conferences ICML, ECML, … and ML sessions at AI conferences IJCAI, ECAI, AAAI, …
  – Extended set of learning tasks and algorithms addressed
Second Generation Data Mining

- Developed since 1990s:
  - Focused on data mining tasks characterized by large datasets described by large numbers of attributes
  - New conferences on practical aspects of data mining and knowledge discovery: KDD, PKDD, ...
  - New learning tasks and efficient learning algorithms: Bayesian network learning, association rule learning, subgroup discovery, multi-relational data mining, statistical relational learning, SVMs, text and Web mining, social network analysis
# Task reformulation: Binary Class Values

Binary classes (positive vs. negative examples of Target class) - simplified single concept learning - “one vs. all” multi-class learning

<table>
<thead>
<tr>
<th>Person</th>
<th>Age</th>
<th>Spect. presc.</th>
<th>Astigm.</th>
<th>Tear prod.</th>
<th>Lenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>17</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>NO</td>
</tr>
<tr>
<td>O2</td>
<td>23</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>YES</td>
</tr>
<tr>
<td>O3</td>
<td>22</td>
<td>myope</td>
<td>yes</td>
<td>reduced</td>
<td>NO</td>
</tr>
<tr>
<td>O4</td>
<td>27</td>
<td>myope</td>
<td>yes</td>
<td>normal</td>
<td>YES</td>
</tr>
<tr>
<td>O5</td>
<td>19</td>
<td>hypermetrope</td>
<td>no</td>
<td>reduced</td>
<td>NO</td>
</tr>
<tr>
<td>O6-O13</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O14</td>
<td>35</td>
<td>hypermetrope</td>
<td>no</td>
<td>normal</td>
<td>YES</td>
</tr>
<tr>
<td>O15</td>
<td>43</td>
<td>hypermetrope</td>
<td>yes</td>
<td>reduced</td>
<td>NO</td>
</tr>
<tr>
<td>O16</td>
<td>39</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>NO</td>
</tr>
<tr>
<td>O17</td>
<td>54</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>NO</td>
</tr>
<tr>
<td>O18</td>
<td>62</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>NO</td>
</tr>
<tr>
<td>O19-O23</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O24</td>
<td>56</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>NO</td>
</tr>
</tbody>
</table>
Other tasks: Learning fromNumeric Class Data

<table>
<thead>
<tr>
<th>Person</th>
<th>Age</th>
<th>Spect. presc.</th>
<th>Astigm.</th>
<th>Tear prod.</th>
<th>LensPrice</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>17</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>0</td>
</tr>
<tr>
<td>O2</td>
<td>23</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>8</td>
</tr>
<tr>
<td>O3</td>
<td>22</td>
<td>myope</td>
<td>yes</td>
<td>reduced</td>
<td>0</td>
</tr>
<tr>
<td>O4</td>
<td>27</td>
<td>myope</td>
<td>yes</td>
<td>normal</td>
<td>5</td>
</tr>
<tr>
<td>O5</td>
<td>19</td>
<td>hypermetrope</td>
<td>no</td>
<td>reduced</td>
<td>0</td>
</tr>
<tr>
<td>O6-O13</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O14</td>
<td>35</td>
<td>hypermetrope</td>
<td>no</td>
<td>normal</td>
<td>5</td>
</tr>
<tr>
<td>O15</td>
<td>43</td>
<td>hypermetrope</td>
<td>yes</td>
<td>reduced</td>
<td>0</td>
</tr>
<tr>
<td>O16</td>
<td>39</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>0</td>
</tr>
<tr>
<td>O17</td>
<td>54</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>0</td>
</tr>
<tr>
<td>O18</td>
<td>62</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>0</td>
</tr>
<tr>
<td>O19-O23</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O24</td>
<td>56</td>
<td>hypermetrope</td>
<td>yes</td>
<td>normal</td>
<td>0</td>
</tr>
</tbody>
</table>
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
Subgroup discovery example: High CHD Risk Group Detection

**Input:** Patient records described by stage A (anamnestic), stage B (an. & lab.), and stage C (an., lab. & 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 $\leftarrow$ ...

Induced subgroups and their statistical characterization

Subgroup A2 for female patients:
high-CHD-risk ← female AND bodymassIndex > 25 AND age > 63

Supporting characteristics (computed using $\chi^2$ statistical significance test): positive family history and hypertension. Women in this risk group typically have slightly increased LDL cholesterol values and normal but decreased HDL cholesterol values.

(Gamberger & Lavrač, JAIR 2002)
Subgroup Discovery in DNA microarray data analysis

- **Functional genomics** is a typical scientific discovery domain, studying genes and their functions, characterized by
  - a very large number of attributes (genes) relative to the number of examples (observations).
  - typical values: 7,000-16,000 attributes, 50-150 examples

- **Sample problem:** Distinguishing between acute lymphoblastic leukemia (ALL, 27 samples) and acute myeloid leukemia (AML, 11 samples), with 34 samples in the test set. Every sample is described with gene expression values for 7,129 genes.

  [http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi](http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi)
Gene Expression Data: data mining format

<table>
<thead>
<tr>
<th>Patient #</th>
<th>Tumor Type</th>
<th>Gene #1</th>
<th>Gene #2</th>
<th>Gene #3</th>
<th>…</th>
<th>Gene #10,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>5.00</td>
<td>1.33</td>
<td>3.45</td>
<td>…</td>
<td>4.22</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>0.98</td>
<td>0.87</td>
<td>1.04</td>
<td>…</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>0.33</td>
<td>1.40</td>
<td>0.42</td>
<td>…</td>
<td>0.24</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>100</td>
<td>B</td>
<td>0.89</td>
<td>0.90</td>
<td>1.00</td>
<td>…</td>
<td>0.66</td>
</tr>
</tbody>
</table>
Standard Approach: Learning High-Dimensional Classifiers

- Neural Networks, Support Vector Machines, ...
- Usually good at predictive accuracy
  - Golub et al., Science 286:531-537 1999
  - Ramaswamy et al., PNAS 98:15149-54 2001
- Resistance to overfitting (mainly SVM, ensembles, ...)
- But black box models are hard to interpret

Apply subgroup discovery to learn symbolic rules instead, enabling explanation
Subgroup discovery in DNA microarray data analysis

- Interesting subgroup describing patterns discovered by SD algorithm

CancerType = AML

IF gene_20056 = DIFF. EXPRESSED
AND gene_23984 = NOT_DIFF. EXPRESSED
Subgroup discovery in DNA microarray data analysis

- Interesting subgroup describing patterns discovered by SD algorithm

CancerType = AML
IF gene_20056 = DIFF. EXPRESSED
AND gene_23984 = NOT_DIFF. EXPRESSED

CancerType = Leukemia
IF KIAA0128 = DIFF. EXPRESSED
AND prostoglandin d2 synthase = NOT_DIFF. EXPRESSED
Subgroup discovery in DNA microarray data analysis

• Interesting subgroup describing patterns discovered by SD algorithm

  CancerType = AML
  IF gene_20056 = DIFF.EXPR.
  AND gene_23984 = NOT_DIFF.EXPR.

• Interpretable by biologists
  – D. Gamberger, N. Lavrač, F. Železný, J. Tolar
    Jr Biomed Informatics 37(5):269-284 2004

• Note the accuracy-interpretability trade off
  – subgroup describing rules are less accurate but interpretable
SD algorithms in the Orange DM Platform

- **Orange** data mining toolkit
  - classification and subgroup discovery algorithms
  - data mining workflows
  - visualization
  - developed at FRI, Ljubljana

- **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
Freely Available Data Mining Platforms

• Second generation data mining platforms:
  – WEKA (Witten and Frank), Orange (Demšar and Zupan), KNIME (Berthold et al.), RapidMiner (Mierswa et al.)
  – include numerous data mining algorithms
  – enable data and model visualization
  – enable simple data mining workflow construction
Towards a Third Generation Data Mining Platform

- Prototype third generation data mining platform Orange4WS is available (Podpečan et al. 2009)
  - is service oriented (DM algorithms as Web services)
  - supports the entire Orange functionality, including user-friendly canvas for workflow construction; in addition, it includes
    - WEKA algorithms, implemented as Web services
    - relational data mining through propositionalization
    - semantic data mining using publicly available domain ontologies
  - enables simplified construction of Web services from available algorithms
  - includes a data mining ontology for planning DM workflows
Relational Data Mining in a nutshell

Given: a relational database, a set of tables, sets of logical facts, a graph, …
Find: a classification model, a set of interesting patterns
Relational data mining

- Relational data mining is characterized by using background knowledge (additional domain knowledge) in the data mining process
- Using background knowledge in data mining has been a topic of extensive research
  - Hierarchical attribute values (Michalski et al. 1986,…), hierarchy/taxonomy of attributes, …
  - Inductive logic programming - ILP (Muggleton, 1991; Lavrač and Džeroski 1994), …
  - Relational learning (Quinlan, 1993), …
  - Relational Data Mining (Džeroski and Lavrač, 2001), …
Relational data mining

- ILP, relational learning, relational data mining
  - Learning from complex multi-relational data
Relational data mining

- ILP, relational learning, relational data mining
  - Learning from complex multi-relational data
  - Learning from complex structured data: e.g., molecules and their properties in protein engineering, biochemistry, ...
Relational Data Mining through Propositionalization

Step 1

Propositionalization

Relational representation of customers, orders and stores.
Relational Data Mining through Propositionalization

**Step 1**
Propositionalization

**Step 2**
Data Mining

```
<table>
<thead>
<tr>
<th>ID</th>
<th>Zip</th>
<th>S</th>
<th>Sc</th>
<th>Income</th>
<th>Age</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
</tr>
</thead>
<tbody>
<tr>
<td>3478</td>
<td>34677</td>
<td>m</td>
<td>60-70</td>
<td>natures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3479</td>
<td>6366</td>
<td>f</td>
<td>80-90</td>
<td>nature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>Customer ID</th>
<th>Order ID</th>
<th>Store ID</th>
<th>Delivery Mode</th>
<th>PayMnt Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>3478</td>
<td>2146267</td>
<td>12</td>
<td>regular</td>
<td>cash</td>
</tr>
<tr>
<td>3478</td>
<td>3446778</td>
<td>12</td>
<td>express</td>
<td>check</td>
</tr>
<tr>
<td>3479</td>
<td>4728886</td>
<td>17</td>
<td>regular</td>
<td>credit</td>
</tr>
<tr>
<td>3479</td>
<td>3234444</td>
<td>17</td>
<td>express</td>
<td>credit</td>
</tr>
</tbody>
</table>

```

```
<table>
<thead>
<tr>
<th>Customer ID</th>
<th>Order ID</th>
<th>Store ID</th>
<th>Delivery Mode</th>
<th>PayMnt Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>3478</td>
<td>2146267</td>
<td>12</td>
<td>regular</td>
<td>cash</td>
</tr>
<tr>
<td>3478</td>
<td>3446778</td>
<td>12</td>
<td>express</td>
<td>check</td>
</tr>
<tr>
<td>3479</td>
<td>4728886</td>
<td>17</td>
<td>regular</td>
<td>credit</td>
</tr>
<tr>
<td>3479</td>
<td>3234444</td>
<td>17</td>
<td>express</td>
<td>credit</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>Store ID</th>
<th>Size</th>
<th>Type</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td></td>
<td>small</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>large</td>
<td>franchise</td>
</tr>
</tbody>
</table>
```

Relational representation of customers, orders and stores.

```
<table>
<thead>
<tr>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
<th>f6</th>
<th>fn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>g1</th>
<th>g2</th>
<th>g3</th>
<th>g4</th>
<th>g5</th>
<th>g6</th>
<th>fn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
<th>f6</th>
<th>fn</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
```

model, patterns, ...
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
  \[ \text{mutagenic}(M) \leftarrow \text{feature121}(M), \text{feature235}(M) \]
Relational data mining

- ILP, relational learning, relational data mining
  - Learning from complex multi-relational data
  - Learning from complex structured data: e.g., molecules and their properties in protein engineering, biochemistry, ...
  - Learning by using domain knowledge, e.g. by using ontologies as background knowledge for relational data mining
Using domain ontologies

Using domain ontologies as background knowledge

- E.g., the Gene Ontology (GO)
- GO is a database of terms, describing gene sets in terms of their
  - functions (12,093)
  - processes (1,812)
  - components (7,459)
- Genes are annotated to GO terms
- Terms are connected (is_a, part_of)
- Levels represent terms generality

.. figure:: image.png
   :alt: Gene Ontology (GO) hierarchy
   :align: center

   GO:0006520 amino acid metabolism
   GO:0009308 amine metabolism
   GO:0009309 amine biosynthesis
   GO:0006576 biogenic amine metabolism
   GO:00042401 biogenic amine synthesis
   GO:0008652 amino acid biosynthesis
Using domain ontologies

• Ontologies (Tim Berners-Lee), since 1989
  • accepted formalism for consensual knowledge representation for Semantic Web applications, a basic for the Semantic Web
  • Description logic, OWL, Protégé ontology editor
  – Using ontologies in data mining, since 2004

• Ontology-driven (semantic) data mining is an emerging research topic

• Semantic Data Mining (SDM) - a term denoting:
  – the new challenge of mining semantically annotated resources, with ontologies used as background knowledge in mining experimental data
Semantic Data Mining in Orange4WS

• Exploiting semantics in data mining
  – Using **domain ontologies** as background knowledge for data mining

• Semantic data mining technology: a two-step approach
  – Using propositionalization through first-order feature construction
  – Using subgroup discovery for rule learning
Semantic Data Mining

SDM task definition used in our work

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

Find: a classification model, a set of patterns
Example: Using Gene Ontology as Domain Knowledge in DNA Microarray Data Analysis

Gene Ontology

12,093 biological process
1,812 cellular components
7,459 molecular functions

Joint work with
Igor Trajkovski
Filip Zelezny and
Jakub Tolar
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{verbatim}
f(7,A):-function(A,'GO:0046872').
f(8,A):-function(A,'GO:0004871').
f(11,A):-process(A,'GO:0007165').
f(14,A):-process(A,'GO:0044267').
f(15,A):-process(A,'GO:0050874').
f(20,A):-function(A,'GO:0004871'), process(A,'GO:0050874').
f(26,A):-component(A,'GO:0016021').
f(29,A):- function(A,'GO:0046872'), component(A,'GO:0016020').
f(122,A):-interaction(A,B),function(B,'GO:0004872').
f(223,A):-interaction(A,B),function(B,'GO:0004871'),
          process(B,'GO:0009613').
f(224,A):-interaction(A,B),function(B,'GO:0016787'),
          component(B,'GO:0043231').
\end{verbatim}
## Propositionalization

<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>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>0</td>
</tr>
</tbody>
</table>

### Differential expression (diffexp)
- diffexp g1: gene64499
- diffexp g2: gene2534
- diffexp g3: gene51999
- diffexp g4: gene1052
- diffexp g5: gene6036

### Random selection (random)
- random g1: gene7443
- random g2: gene9221
- random g3: gene2339
- random g4: gene9657
- random g5: gene19679
Semantic Data Mining in two steps

• **Step 1:** Construct relational logic features of genes such as

\[ \text{interaction}(g, G) \text{ \& \ function}(G, \text{protein\_binding}) \]

(g 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:

\[ \text{diffexp}(A) \text{ :- interaction}(A,B) \text{ \& \ function}(B,'\text{GO:0004871}') \text{ \& \ process}(B,'\text{GO:0009613}') \]
Semantic Data Mining for DNA Microarray Data Analysis

• Semantic data mining integrates public gene annotation data through relational features
• It is implemented in the SEGS algorithm (Trajkovski, Železny, Lavrač and Tolar, JBI 2008), available in Orange4WS
• It can be combined with additional biomedical resources (BioMine), providing additional means for creative knowledge discovery from publicly available data sources
Biomine graph exploration
(Toivonnnen et al., University of Helsinki)

• **BioMine graph** contains information from public databases, including annotated sequences, proteins, orthology groups, genes and gene expressions, gene and protein interactions, PubMed articles, and different ontologies.
  
  – **nodes (~1 mio)** correspond to different concepts (such as gene, protein, domain, phenotype, biological process, tissue)
  
  – **semantically labeled edges (~7 mio)** connect related concepts

• **BioMine query engine** answers queries to potentially discover new links between entities by sophisticated graph exploration algorithms
The SEGS + BioMine Methodology

Microarray

gene1: ++
gene2: +
gene3: +
...
geneN: --

e.g. slow-vs-fast cell growth

Gene sets

Exploratory link discovery

Work by
Lavrač et al. 2009, 2010
Podpečan et al., BMC Bioinformatics 2011
Semantic Data Mining in Orange4WS: SEGS + BioMine workflow implementation
**SEGS + BioMine outputs**

**SEGS output:**

<table>
<thead>
<tr>
<th>#</th>
<th>Description</th>
<th>Set size</th>
<th>#DE_Genes</th>
<th>Fisher p-value (unadjusted p-value)</th>
<th>GSEA p-value (Enrichment score)</th>
<th>PAGE p-value (Z-score)</th>
<th>Aggregate p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Func(monovalent inorganic cation transporter activity), Proc(monovalent inorganic cation transport), Comp(integral to membrane)</td>
<td>26</td>
<td>10</td>
<td>0.000 (2.208e-07)</td>
<td>0.010</td>
<td>0.020 (3.787)</td>
<td>0.010</td>
</tr>
<tr>
<td>2</td>
<td>Func(monovalent inorganic cation transporter activity), Proc(monovalent inorganic cation transport), Comp(integral to membrane)</td>
<td>24</td>
<td>8</td>
<td>0.010 (4.228e-06)</td>
<td>0.010</td>
<td>0.020 (3.671)</td>
<td>0.013</td>
</tr>
<tr>
<td>3</td>
<td>Func(monovalent inorganic cation transporter activity), Proc(transport), Comp(integral to membrane)</td>
<td>26</td>
<td>8</td>
<td>0.010 (6.160e-06)</td>
<td>0.040</td>
<td>0.020 (3.831)</td>
<td>0.023</td>
</tr>
</tbody>
</table>

**BioMine query output:**

[Diagram showing network of nodes and edges]
Summary of SEGS + BioMine

• Semantic Data Mining algorithm SEGS discovers interesting gene group descriptions as conjunctions of concepts from three ontologies: GO, KEGG and Entrez

• Biomine finds cross-context links (paths) between concepts discovered by SEGS, using other ontologies, PubMed and other biomedical resources

• Initial results in stem cell microarray data analysis (EMBC 2009) indicate that the SEGS+Biomine methodology may lead to new insights – in vitro experiments are in progress at NIB to verify and validate the preliminary insights

• A general purpose Semantic Data Mining algorithm g-SEGS is also available in Orange4WS
Current work

- Semantic Subgroup Discovery (Vavpetič et al., 2012)
Summary and conclusion: Current work

- Semantic Subgroup Discovery workflows in Orange4WS and ClowdFlows (Vavpetič et al., 2012)
- ClowdFlows - browsed-based DM platform for data mining in the cloud and workflow sharing on the web (Kranjc et al. 2012)
Biomedical use cases

• Recent work: Subgroup discovery and their semantic explanation (submitted to JIIS, 2013)

• Workflow is available on the web for sharing in the ClowdFlows platform, http://clowdfloows.org/workflow/910/
Towards Third Generation Data Mining

- Orange4WS and ClowdFlows DM platforms
  - are service oriented (DM algorithms as web services)
  - user-friendly HCI: canvas for workflow construction
  - include functionality of standard data mining platforms
    - WEKA algorithms, implemented as Web services
- Include new functionality
  - relational data mining
  - semantic data mining
    - enable simplified construction of Web services from available algorithms
- ClowdFlows runs in a browser – enables data mining, workflow construction and sharing on the web
Summary and conclusion: Future work

• Current Semantic data mining scenario: Mining empirical data with ontologies as background knowledge
  • abundant empirical data, but
  • scarce background knowledge

• Future Semantic data mining scenario:
  – envisioning a growing amount of semantic data
  • abundance of ontologies and semantically annotated data collections
  • e.g. Linked Data
    – over 6 billion RDF triples
    – over 148 million links
Summary and conclusion: Future work

- We envision a paradigm shift from data mining (mining of empirical data) in standard data mining platforms to **knowledge mining on the web**
- The envisioned future Semantic data mining scenario in mining the Semantic Web:
  - mining knowledge encoded in domain ontologies,
  - constrained by annotated (empirical) data collections.
Acknowledgements

• Work on subgroup discovery was done jointly with D. Gamberger (RBI) and P. Kralj Novak (JSI)
• Work on relational data mining and semantic data mining was done jointly with I. Trajkovski (Skopje Uni.), F. Železny (CTU, Prague), J. Tolar (Univ. of Minnesota), I. Mozetič, A. Vavpetič, and colleagues (JSI)
• Work on new data mining platforms was done jointly with V. Podpečan and J. Kranjc (JSI)