Advances in Data Mining for Biomedical Research

Nada Lavrač

Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana
Slovenia

MedInfo 2010 invited lecture, September 2010
Slovenia – Ljubljana (capital)

Ljubljana, Slovenia
Jožef Stefan Institute, Ljubljana, Slovenia

• Jožef Stefan Institute (JSI, founded in 1949)
  – named after a distinguished physicist
    Jožef Stefan (1835-1893)
  – leading national research organization in natural sciences and technology (~700 researchers and students)

• Jožef Stefan International Postgraduate School (founded in 2004)
  – Offers three MSc and PhD programs (in English):
    ICT, nanotechnologies and ecotechnologies
Department of Knowledge Technologies

- **Head:** Nada Lavrač, **Staff:** 40 researchers, 15 students
- **Knowledge Technologies**
  - Making AI techniques operational for practical problems
Department of Knowledge Technologies

- **Head:** Nada Lavrač, **Staff:** 40 researchers, 15 students
- **Knowledge Technologies**
  - Making AI techniques operational for practical problems
- **Main research areas**
  - Data Mining and Machine Learning
  - Text Mining and Human Language Technologies
  - Web Services and Semantic Web
  - Ontologies and Knowledge Management
  - Decision Support Systems
- **Applications**
  - Medicine, Bioinformatics, Public Health
  - Ecology, Finance, …
Department of Knowledge Technologies

- My research preferences
  - Data Mining
  - Text Mining
  - Web Services and Workflows
  - Knowledge Management

- Applications
  - Medicine, Bioinformatics
  - Public Health
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
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
Data Mining in a Nutshell

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

Find: a classification model, a set of interesting patterns

Data Mining

model, patterns, ...

knowledge discovery from data

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 in a Nutshell

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

- **Given**
  - Data Mining
  - Knowledge discovery from data
  - Data mining data
  - Given: transaction data table, relational database, text documents, Web pages
  - Find: a classification model, a set of interesting patterns

**Table:**

<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>hypermetro</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>hypermetro</td>
<td>no</td>
<td>normal</td>
<td>SOFT</td>
</tr>
<tr>
<td>O15</td>
<td>43</td>
<td>hypermetro</td>
<td>yes</td>
<td>reduced</td>
<td>NONE</td>
</tr>
<tr>
<td>O16</td>
<td>39</td>
<td>hypermetro</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>hypermetro</td>
<td>yes</td>
<td>normal</td>
<td>NONE</td>
</tr>
</tbody>
</table>

**New unclassified instance**

- New unclassified instance
- Classified instance
- Black box classifier
- No explanation

**Symbolic Model**

- 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
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>

lenses=NONE ← tear production=red
lenses=NONE ← tear production=normal AND astigmatism=yes AND spect. pre.=hypermetrope
lenses=SOFT ← tear production=normal AND astigmatism=no
lenses=HARD ← tear production=normal AND astigmatism=yes AND spect. pre.=myope
lenses=NONE ←
Task reformulation: Binary Class Values

<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>

Binary classes (positive vs. negative examples of Target class)
- simplified single concept learning
- “one vs. all” multi-class learning
## Other tasks: Learning from Numeric 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>

**Numeric class values – regression analysis**
Other tasks: Learning from Unlabeled Data

Unlabeled data - clustering: grouping of similar instances
(similar instances – many common values)
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
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
Text Mining Example: Clustering of PubMed Articles

Slide adapted from D. Mladenić, JSI
OntoGen Applied to Clustering of PubMed Articles on Autistic Spectrum Disorders

Work by Petrič et al. 2009

www.ontogen.si
Fortuna, Mladenič, Grobelnik 2006
Finding linking terms (idea based on early work of Swanson 1990) – using OntoGen?

Literature A (magnesium)

Literature C (migraine)

- Linking term B1
- Linking term B2
- Linking term B3
Finding linking terms (idea based on early work of Swanson 1990) – using OntoGen?

Literature A (calcineurin)

Literature C (autism)

Work by Petrič et al. 2009

- Linking term B1
- Linking term B2
- Linking term B3
Outlier analysis from two document sets for cross-context link discovery

2-dimensional projection of documents (about autism (red) and calcineurin (blue). Outlier documents are bolded for the user to easily spot them.

Our research has shown that most linking terms appear in outlier documents. (Lavrač, Sluban, Grčar, Juršič 2010)
Using OntoGen for Outlier Document Identification

Slide adapted from D. Mladenić, JSI
Method 2 results on autism-calcineurin: Outlier calcineurin document CN423

Work by Petrič et al. 2010
Selected Text Mining Tools Developed at JSI

Document-Atlas

Content-Land

OntoGen

Semantic-Graphs

Contexter

M. Grobelnik, D. Mladenić, et al.
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
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: 7000-16000 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 7129 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 Classification Models

• 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 interprete
Subgroup discovery in DNA microarray data analysis

• Interesting subgroup describing patterns discovered by SD algorithm

<table>
<thead>
<tr>
<th>CancerType = AML</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF gene_20056 = DIFF.EXPR.</td>
</tr>
<tr>
<td>AND gene_23984 = NOT_DIFF.EXPR.</td>
</tr>
</tbody>
</table>

• **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.), …
  – 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
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
  – Relational Mining Supported in a Service-oriented Knowledge Discovery platform
  – Semantic Data Mining: Using ontologies in DM
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 (ILP)

- Learning from multiple tables
  - patient records connected with other patient and demographic information
- Complex relational problems:
  - temporal data: time series in medicine, ...
  - structured data: representation of 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

Relational representation of customers, orders and stores.
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
  \[ \text{mutagenic}(M) \leftarrow \text{feature}_{121}(M), \text{feature}_{235}(M) \]
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
  – Relational Mining Supported in a Service-oriented Knowledge Discovery platform
  – Semantic Data Mining: Using ontologies in DM
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
Example: Using Gene Ontology as Domain Knowledge in DNA Microarray Data Analysis

Gene Ontology

12093 biological process
1812 cellular components
7459 molecular functions

Joint work with
Igor Trajkovski
Filip Zelezny and
Jakub Tolar
Example: Using Gene Ontology as Domain Knowledge in DNA Microarray Data Analysis

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:004871'), 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

diffexp g1 (gene64499)  random g1 (gene7443)
diffexp g2 (gene2534)   random g2 (gene9221)
diffexp g3 (gene5199)   random g3 (gene2339)
diffexp g4 (gene1052)   random g4 (gene9657)
diffexp g5 (gene6036)   random g5 (gene19679)

<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>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>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g4</td>
<td>1</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>0</td>
</tr>
<tr>
<td>g5</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>0</td>
<td>0</td>
</tr>
<tr>
<td>g1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</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>1</td>
<td>1</td>
<td>1</td>
<td>0</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>0</td>
<td>1</td>
</tr>
<tr>
<td>g5</td>
<td>0</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>
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:

\[
\text{diffexp}(A) \leftarrow \text{interaction}(A,B) \text{ AND } \text{function}(B,'GO:0004871') \text{ AND } \text{process}(B,'GO:0009613')
\]
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
(Toivonnen 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

SEGS

Biomine

Work by
Lavrač et al. 2009, 2010
Podpečan et al. 2010
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</th>
<th>GSEA p-value</th>
<th>PAGE p-value</th>
<th>Aggregate p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Func (monovalent inorganic ion transporter activity), Pro (monovalent inorganic ion transporter), Comp (integral to membrane)</td>
<td>26</td>
<td>10</td>
<td>0.000 (9.28e-07)</td>
<td>0.010 (0.362)</td>
<td>0.020 (3.767)</td>
<td>0.010</td>
</tr>
<tr>
<td>2</td>
<td>Func (monovalent inorganic ion transporter activity), Pro (monovalent inorganic ion transporter), Comp (integral to membrane)</td>
<td>24</td>
<td>8</td>
<td>0.010 (4.23e-06)</td>
<td>0.010 (0.352)</td>
<td>0.030 (3.671)</td>
<td>0.013</td>
</tr>
<tr>
<td>3</td>
<td>Func (monovalent inorganic ion transporter activity), Pro (monovalent inorganic ion transporter), Comp (integral to membrane)</td>
<td>26</td>
<td>8</td>
<td>0.010 (9.16e-06)</td>
<td>0.040 (0.323)</td>
<td>0.020 (3.931)</td>
<td>0.023</td>
</tr>
</tbody>
</table>

#### BioMine query output:

![BioMine query output](image)

**Project:** [Image]

**Enriched genesets for class A**

- found by Combining p-values
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
Overall Summary

• 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
  – Relational Mining Supported in a Service-oriented Knowledge Discovery platform
  – Semantic Data Mining: Using ontologies in DM
  – En passant - Creative Knowledge Discovery from Biomedical Literature, explored in ongoing FP7 EU project BISON (2008-11)
Acknowledgements

• The work on relational data mining and semantic data mining was done in collaboration with I. Trajkovski (Skopje Uni.), F. Železny (CTU, Prague) and J. Tolar (Univ. of Minnesota)

• Third generation platform development is joint work with V. Podpečan (JSI)

• BISON is EU funded project, coordinated by M. Berthold (Konstanz University)

• Thanks to Riccardo Belazzi, Johanna Westbook and the MedInfo 2010 organizing team for the opportunity to present this talk