An Introduction to Mining Big and Complex Data

Sašo Džeroski
Jozef Stefan Institute, Ljubljana, Slovenia
Mining Big and Complex Data

• Introduction: Just what is big and complex data?
  • Volume & Velocity (Data Streams)
  • Variety (Structured Inputs and Structured Outputs)
  • Other complexity dimensions (Incompleteness, Context)

• The different tasks of structured output prediction

• Combination with other complexities
  • Semi-supervised
  • SOP on data streams

• Structured output prediction with predictive clustering
Data mining: Predictive modelling

• Predictive models focus on a target variable and predict its value from the values of input variables

• Classical problem: Medical diagnosis

• An example: Neurodegenerative diseases

• Target variable: Diagnosis; Possible values:
  • CN - Cognitively Normal (0)
  • SMC - Significant Memory Concern
  • EMCI - Early Mild Cognitive Impairment
  • LMCI - Late Mild Cognitive Impairment
  • AD - Alzheimer’s Disease (4)
Example task: Descriptive vars.; Biomarkers for Alzheimer’s

1. APOE4 – Genetic variations of APOE4 related gene
2. FDG – Positron emission tomography (PET) imaging results with $[^{18}\text{F}]$fluorodeoxyglucose
3. AV45 – Positron emission tomography (PET) imaging results with $[^{18}\text{F}]$-labeled amyloid imaging agent AV45
4. Ventricles
5. Hippocampus
6. WholeBrain
7. Entorhinal
8. Fusiform – Fusiform gyrus
9. MidTemp – Middle Temporal Gyrus
10. ICV – Intracerebral volume [Volumetric data 4-10]
Example: Decision tree for diagnosis

```
Example: Decision tree for diagnosis

PB.FDG_bl

<= 5.99515

PB.FDG_bl

<= 5.2823

AD (95.99/27.47)

> 5.2823

> 5.99515

EMCI (617.81/368.44)

PB.Entorhinal_bl

<= 3483

AD (101.51/59.88)

> 3483

EMCI (100.68/64.53)
```
Predictive modeling: Classification and regression

<table>
<thead>
<tr>
<th>Example 1</th>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>2</td>
<td>FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>1</td>
<td>FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>4</td>
<td>FALSE</td>
<td>0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 1</th>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>2</td>
<td>FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>1</td>
<td>FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>4</td>
<td>FALSE</td>
<td>0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
</tbody>
</table>
Big Data: Volume & Velocity

• Large number of columns (high dimensionality)
  • Need feature ranking/selection

• Large number of rows (massive data)
  • Need efficient data mining methods

• Streaming rows (data streams)
  • Need incrementality: Not all data available simultaneously
  • Data instances arrive at high velocities, in a specific order and their number is potentially arbitrarily large
  • The underlying concept (distribution) governing the data can change (concept drift)
  • We need fast processing (due to the high velocity)
  • The large and potentially infinite number of examples demands economical management of available memory
### Data streams: Regression

<table>
<thead>
<tr>
<th>Example</th>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Example n+5</td>
<td>1</td>
<td>TRUE</td>
</tr>
<tr>
<td>Example n+1</td>
<td>4</td>
<td>FALSE</td>
</tr>
<tr>
<td>Example n+2</td>
<td>6</td>
<td>FALSE</td>
</tr>
<tr>
<td>Example n+3</td>
<td>8</td>
<td>TRUE</td>
</tr>
<tr>
<td>Example n+4</td>
<td>6</td>
<td>TRUE</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
Big Data: Variety - Structured Input

Example:
Predicting biodegradability

<table>
<thead>
<tr>
<th>compound</th>
<th>activity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>0.37</td>
</tr>
</tbody>
</table>
Big Data: Variety - Structured Output

• Hierarchical classification
• Taxonomic classification of diatoms
• From microscopic images
• Taking into account the taxonomy of diatoms

<table>
<thead>
<tr>
<th>image</th>
<th>features/descriptors</th>
<th>taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Heuristic shape descriptors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>48 24 59 66 37 …</td>
<td>olivaceum</td>
</tr>
<tr>
<td></td>
<td>36 25 53 45 15 …</td>
<td>minutissimum</td>
</tr>
<tr>
<td></td>
<td>35 25 56 52 19</td>
<td>exigua</td>
</tr>
<tr>
<td>…</td>
<td>… … … … …</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>… … … … …</td>
<td>…</td>
</tr>
</tbody>
</table>
Structured-output prediction

• Multi-target prediction
  • Classification
  • Regression
  • Mixed

• Multi-label classification
  • Hierarchical multi-label classification

• Predicting (short) time series
# Multi-target prediction

## Classification

<table>
<thead>
<tr>
<th>Example</th>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>1 TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>Example 2</td>
<td>2 FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>Example 3</td>
<td>1 FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>Example 4</td>
<td>2 TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>Example 5</td>
<td>3 TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>Example 6</td>
<td>4 FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

## Regression

<table>
<thead>
<tr>
<th>Example</th>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>1 TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>Example 2</td>
<td>2 FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>Example 3</td>
<td>1 FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>Example 4</td>
<td>2 TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>Example 5</td>
<td>3 TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>Example 6</td>
<td>4 FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Example MTR task: Target vars.; Clinical scores for Alzheimer’s

1. CDRSB – Clinical Dementia Rating Sum of Boxes
2. ADAS13 – AD assessment scale
3. MMSE – Mini Mental State Examination
4. RAVLT (immediate, learning, forgetting, perc. forgetting) – Rey Auditory Verbal Learning Test (4 features)
5. FAQ – Functional Assessment Questionnaire
6. MOCA – Montreal Cognitive Assessment
7. EcogPt (Memory, Language, Visuospatial Abilities, Planning, Organization, Divided Attention, Total score) – Everyday cognition questionnaire – filled in by patient (7 features)
8. EcogSP (Memory, Language, Visuospatial Abilities, Planning, Organization, Divided Attention, Total score) – Everyday cognition questionnaire – filled in by study parter (7 features)
Example MTR model

- **Descr. attributes:** Biomarkers
- **Targets attributes:**
  - **diagnosis** (0-4)+
  - Clinical measurements/scores (23 of them)

- **Cluster 1**
  - N=154
  - FDG
  - AV45
  - ≤ 6623
  - > 6623

- **Cluster 2**
  - N=408
  - FDG
  - AV45
  - ≤ 1.23
  - > 1.23

- **Cluster 3**
  - N=108
  - FDG
  - AV45
  - ≤ 5.94
  - > 5.94

- **Cluster 4**
  - N=126
  - FDG
  - AV45
  - ≤ 3.38
  - > 3.38

- **DX**
- **CDRSB**
- **ADAS13**
- **MMSE**
- ...
Multi-Target Classification & Multi-Label Classification

• Learning models that simultaneously predict several nominal/binary target variables
• Input: A vector of descriptive variables
• Output: A vector of several nominal/binary targets

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Temperature</th>
<th>$K_2Cr_2O_7$</th>
<th>$NO_2$</th>
<th>$Cl$</th>
<th>$CO_2$</th>
<th>...</th>
<th>Cladophora sp.</th>
<th>Gongrosira incrustans</th>
<th>Oedogonium sp.</th>
<th>Stigeoclonium tenue</th>
<th>Melosira varians</th>
<th>Nitzschia palea</th>
<th>Audouinella chalybea</th>
<th>Erpobella octoculata</th>
<th>Gammarus fossarum</th>
<th>Baetis rhodani</th>
<th>Hydropsyche sp.</th>
<th>Rhyacophila sp.</th>
<th>Simulium sp.</th>
<th>Tubifex sp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID1</td>
<td>0.66</td>
<td>0.00</td>
<td>0.40</td>
<td>1.46</td>
<td>0.84</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ID2</td>
<td>2.03</td>
<td>0.16</td>
<td>0.35</td>
<td>1.74</td>
<td>0.71</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ID3</td>
<td>3.25</td>
<td>0.70</td>
<td>0.46</td>
<td>0.78</td>
<td>0.71</td>
<td></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>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Multi-Label Classification Example

- A decision tree for multi-label classification
## Hierarchical multi-label classification

<table>
<thead>
<tr>
<th>Example 1</th>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>TRUE</td>
</tr>
<tr>
<td>Example 2</td>
<td>2</td>
<td>FALSE</td>
</tr>
<tr>
<td>Example 3</td>
<td>1</td>
<td>FALSE</td>
</tr>
<tr>
<td>Example 4</td>
<td>2</td>
<td>TRUE</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Hierarchical multi-label classification: An example

- Gene function prediction
- Input: Tuple of primitives
- Output: A subhierarchy of a hierarchical catalog of gene functions, such as FunCat

![Descriptive attributes: [0.36, -0.49, -0.1, 0.21, -0.34, -0.27, -0.06, ...]

Target hierarchy subset:
- metabolism
- C-compound and carbohydrate metabolism
- regulation of C-compound and carbohydrate metabolism
- transcription
- RNA synthesis
- mRNA synthesis
- transcriptional control
- transcription activation

Fig. 1: An example task of HMLC: a single instance from the cellcycle dataset (Section 3) is shown, corresponding to one gene. The descriptive attributes are gene properties, the targets are gene functions from the FunCat hierarchy.
## Time-series prediction

<table>
<thead>
<tr>
<th>Example 1</th>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>Example 2</td>
<td>2</td>
<td>FALSE</td>
</tr>
<tr>
<td>Example 3</td>
<td>1</td>
<td>FALSE</td>
</tr>
<tr>
<td>Example 4</td>
<td>2</td>
<td>TRUE</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Predicting short time series

Table 2: An example task of predicting short time series. Three instances (genes) are shown: The descriptive attributes are gene functions, the target is a (short) time series of gene expression values in yeast responding to environmental stress (amino acid starvation in this case).

<table>
<thead>
<tr>
<th>Descriptive attributes</th>
<th>Target time series</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO 0000282</td>
<td>(a) [0.13, 0.48, 0.19, -0.23, -0.12]</td>
</tr>
<tr>
<td>GO 0000315</td>
<td>(b) [0.38, -0.57, 0.17, -0.04, 0.19]</td>
</tr>
<tr>
<td>GO 0000781</td>
<td>(c) [-2.25, -0.94, -0.09, 0.08, -0.15]</td>
</tr>
<tr>
<td>GO 0000785</td>
<td></td>
</tr>
<tr>
<td>GO 0000790</td>
<td></td>
</tr>
<tr>
<td>GO 0000819</td>
<td></td>
</tr>
<tr>
<td>GO 0080090</td>
<td></td>
</tr>
</tbody>
</table>

...
Even more complex SOs

- Mixed tuples (diff. arg. of tuple are of diff. types, e.g., a mix of discrete and real-valued targets)
- Besides tuples, sets & sequences of primitive values, also tuples, sets and sequences of structures (e.g., of the previously mentioned types of SOs
  - Tuples of hierarchies (The Gene Ontology has three hierarchies: BF, MP, Cellular Component)
  - Tuples of time series
  - Sets of tuples
  - Sequences of tuples
- ...

### Predicting tuples of time-series

<table>
<thead>
<tr>
<th>Example</th>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>1  TRUE 0.49 0.69</td>
<td><img src="chart1.png" alt="Chart 1" /> <img src="chart2.png" alt="Chart 2" /> <img src="chart3.png" alt="Chart 3" /></td>
</tr>
<tr>
<td>Example 2</td>
<td>2  FALSE 0.08 0.07</td>
<td><img src="chart1.png" alt="Chart 1" /> <img src="chart2.png" alt="Chart 2" /> <img src="chart3.png" alt="Chart 3" /></td>
</tr>
<tr>
<td>Example 3</td>
<td>1  FALSE 0.08 0.07</td>
<td><img src="chart1.png" alt="Chart 1" /> <img src="chart2.png" alt="Chart 2" /> <img src="chart3.png" alt="Chart 3" /></td>
</tr>
<tr>
<td>Example 4</td>
<td>2  TRUE 0.49 0.69</td>
<td><img src="chart1.png" alt="Chart 1" /> <img src="chart2.png" alt="Chart 2" /> <img src="chart3.png" alt="Chart 3" /></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
The other complexity aspects

- Incomplete annotations
- Network context
### Semi-supervised learning: Classification and regression

<table>
<thead>
<tr>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>1</td>
</tr>
<tr>
<td>Example 2</td>
<td>2</td>
</tr>
<tr>
<td>Example 3</td>
<td>1</td>
</tr>
<tr>
<td>Example 4</td>
<td>2</td>
</tr>
<tr>
<td>Example 5</td>
<td>3</td>
</tr>
<tr>
<td>Example 6</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>1</td>
</tr>
<tr>
<td>Example 2</td>
<td>2</td>
</tr>
<tr>
<td>Example 3</td>
<td>1</td>
</tr>
<tr>
<td>Example 4</td>
<td>2</td>
</tr>
<tr>
<td>Example 5</td>
<td>3</td>
</tr>
<tr>
<td>Example 6</td>
<td>4</td>
</tr>
</tbody>
</table>

...
Network regression

Node 1

Node 2

Node 3

Node 4

<table>
<thead>
<tr>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>FALSE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FALSE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>TRUE</td>
</tr>
</tbody>
</table>
Motivation for MAESTRA

Each of the individual complexity aspects above presents a major challenge to current ML/DM methods

Most approaches to

• Structured output prediction (e.g., multi-label learning)
• Mining data streams (e.g., VFDT)
• Semi-supervised learning (e.g., co-training)
• Learning in a network context (e.g. collective classification)

Consider each of the complexity dimensions individually
Simultaneous presence of several complexity aspects is a much harder challenge and is not addressed appropriately by current approaches

SOP [for different structured outputs] in all cases

- SOP + SSL (Semi-supervised structured-output prediction)
- SOP + Network Data
- SOP + Data Streams
- SOP + SSL + Data Streams
- ...
- SOP + SSL + Data Streams + (Dynamic) Network Data
SSL+SOP: Multi-target regression

<table>
<thead>
<tr>
<th>Example</th>
<th>Type</th>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Descriptive space</td>
<td>Target space</td>
</tr>
<tr>
<td>Example 1</td>
<td>1</td>
<td>TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>Example 2</td>
<td>2</td>
<td>FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>Example 3</td>
<td>1</td>
<td>FALSE</td>
<td>0.08</td>
</tr>
<tr>
<td>Example 4</td>
<td>2</td>
<td>TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>Example 5</td>
<td>3</td>
<td>TRUE</td>
<td>0.49</td>
</tr>
<tr>
<td>Example 6</td>
<td>4</td>
<td>FALSE</td>
<td>0.08</td>
</tr>
</tbody>
</table>

...
The MAESTRA project: Goals

Develop predictive modelling methods capable of simultaneously addressing several (and ultimately all) of the complexity aspects outlined above: Methods that can handle massive sets of network data incompletely annotated with structured outputs.

Develop the foundations (basic concepts/notions) and the methodology (design/implement algorithms) necessary.

Demonstrate the potential and utility of the developed approaches on showcase problems.
The MAESTRA pillars

Mining network data

Mining streaming data

Semi-supervised learning

Structured output prediction
MAESTRA Applications

• Life sciences / health
  • Fungal microbiology
  • Predicting gene function

• Sensor networks (smart grid, energy production)
• Social networks (e.g., sentiment analysis/Twitter)

• Multimedia
  • Image annotation
  • Image retrieval
A central approach in MAESTRA (but not the only one :-) )

- Learning tree and rule-based models in the context of predictive clustering, which unifies the tasks of **predictive modelling** and **clustering**

- Predictive clustering (PC) allows for
  - Handling different types of structured outputs
  - Efficient learning of trees, rules and ensembles thereof

- We are extending PC to consider semi-supervised learning; network context; and to learn from data streams

- We are considering different combinations of complexity aspects (e.g., SOP+data streams+SSL) in this context
Predictive Clustering for Predicting Structured Os
Predictive modeling

• Input: A table of data, a row is an object, single target

<table>
<thead>
<tr>
<th>Gender</th>
<th>Fusiform</th>
<th>Hippocampus</th>
<th>ICV</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>F</td>
<td>16471</td>
<td>6350</td>
<td>1445040,208</td>
</tr>
<tr>
<td>Example 2</td>
<td>M</td>
<td>20680</td>
<td>7440</td>
<td>1610298,246</td>
</tr>
<tr>
<td>Example 3</td>
<td>F</td>
<td>18751</td>
<td>6615</td>
<td>1257475,402</td>
</tr>
<tr>
<td>Example 4</td>
<td>M</td>
<td>22895</td>
<td>9311</td>
<td>1755672,837</td>
</tr>
<tr>
<td>Example 5</td>
<td>F</td>
<td>18446</td>
<td>6544</td>
<td>1527253,171</td>
</tr>
<tr>
<td>Example 6</td>
<td>F</td>
<td>16056</td>
<td>6869</td>
<td>1262875,649</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

• Output: A predictive model for the target

AD (95.99/27.47)
PB.FDG.bl
<= 5.99515
> 5.99515
EMCI (617.81/368.44)
<= 5.2823
> 5.2823
AD (101.51/59.88)
PB.Entorhinal.bl
<= 3463
> 3463
EMCI (100.68/64.53)
Clustering

Partition a set of objects into clusters of similar objects

- High similarity of objects within individual clusters, low similarity between objects from different clusters
- Minimize intra-cluster variance (ICV)
- Distance/similarity measure in the example space
Predictive clustering

- Combines prediction and clustering

- We can have hierarchical clustering (trees) and flat/overlapping clusterings (rules)

- With each cluster, predictive clustering provides
  - A description of the cluster
  - A prediction of the selected targets for that cluster

- The output of PC can be viewed both as a clustering and as a predictive model
Example predictive clustering tree

- **Descr. attributes**: Biomarkers
- **Targets attributes**: 
  - **diagnosis** (0-4+)
  - Clinical measurements/scores (23 of them)

- **FDG**
  - <= 5.94
  - > 5.94

- **Hippocampus**

- **AV45**
  - <= 6623
  - > 6623

- **Cluster 4**
  - N=126
  - 3.23
  - 3.38
  - 26.4
  - 24.8
  - ...

- **Cluster 3**
  - N=108
  - 2.31
  - 1.97
  - 18.2
  - 27.1
  - ...

- **Cluster 2**
  - N=408
  - 1.37
  - 0.70
  - 10.7
  - 28.8
  - ...

- **Cluster 1**
  - N=154
  - 2.00
  - 1.33
  - 14.9
  - 27.9
  - ...

- **DX**
- **CDRSB**
- **ADAS13**
- **MMSE**
- ...

---

*Example predictive clustering tree with attributes and clusters.*
Top-Down Induction of Decision Trees

To construct a tree $T$ from a training set $S$:

- If all the examples belong to the same class $C$, construct a leaf labeled $C$

- Otherwise:
  - Select the best attribute $A$ with values $v_1, \ldots, v_n$, which reduces the most the impurity of the target
  - Partition $S$ into $S_1, \ldots, S_n$ according to $A$
  - Recursively construct subtrees $T_1$ to $T_n$ for $S_1$ to $S_n$
  - Result: a tree with root $A$ and subtrees $T_1, \ldots, T_n$
To construct a tree $T$ from a training set $S$:

- If the examples in $S$ have low variance,
  construct a leaf labeled $target(prototype(S))$
- Otherwise:
  - Select the best attribute $A$ with values $v_1, \ldots, v_n$, which reduces the most the variance (measured according to a given distance function $d$)
  - Partition $S$ into $S_1, \ldots, S_n$ according to $A$
  - Recursively construct subtrees $T_1$ to $T_n$ for $S_1$ to $S_n$
- Result: a tree with root $A$ and subtrees $T_1, \ldots, T_n$
Learning PCTs

• Recursively partition data set into subsets (clusters) with low intra-cluster variance
  • Variance = avg. squared distance to prototype
    \[ ICV(S) = \sum_{y_j \in S} d(y_j, p(S))^2 \]

• For the variance, the distance is measured
  • In standard clustering, along all dimensions
  • In prediction, along a single target dimension
  • In predictive clustering, along a structured target, e.g., several target dimensions
**Predictive clustering:** A divides data into clusters 1 and 2 coherent along two dimensions
Distances/variances for SOP tasks

- The algorithm
- Variance for MT regression
  \[ \text{Var}(E) = \sum_{i=1}^{T} \text{Var}(Y_i). \]
- Variance for MT classification
  \[ \text{Var}(E) = \sum_{i=1}^{T} \text{Entropy}(E, Y_i) \]
- Variance for HMLC
  \[ \text{Var}(E) = \frac{1}{|E|} \cdot \sum_{E_i \in E} d(L_i, \bar{L})^2 \]
  \[ d(L_1, L_2) = \sqrt{\sum_{i=1}^{|L_i|} w(c_i) \cdot (L_{1,i} - L_{2,i})^2} \]

\[ \text{procedure} \text{ BestTest}(E) \]
1: \( (t^*, h^*, \mathcal{P}^*) = (\text{none}, 0, \emptyset) \)
2: \textbf{for each} possible test \( t \) \textbf{do}
3: \( \mathcal{P} = \text{partition induced by } t \text{ on } E \)
4: \( h = \text{Var}(E) - \sum_{E_i \in \mathcal{P}} \frac{|E_i|}{|E|} \text{Var}(E_i) \)
5: \textbf{if } (h > h^*) \text{ \&\& Acceptable}(t, \mathcal{P}) \textbf{ then}
6: \( (t^*, h^*, \mathcal{P}^*) = (t, h, \mathcal{P}) \)
7: \textbf{return } (t^*, h^*, \mathcal{P}^*) \]
Ensembles of PCTs

- Ensembles of PCTs use several methods for constructing base classifiers
  - Bagging & Random forests
  - Random subspaces & Bagged Random subspaces

- PCTs and Ensembles of PCTs implemented in SW package CLUS, jointly developed by JSI, Ljubljana and KULeuven, Belgium

- Written in Java

- Open source, available for download from http://sourceforge.net/projects/clus
Ensembles of PCTs: Bagging
**RandomForests & Feature Ranking**

```latex
procedure Induce_RF(E, k, f(x))
returns Forest, Importances
1: \( F = \emptyset \)
2: \( I = \emptyset \)
3: for \( i = 1 \) to \( k \) do
4: \( E_i = \text{Bootstrap}_\text{sample}(E) \)
5: \( Tree_i = PCT_{\text{rand}}(E_i, f(x)) \)
6: \( F = F \cup Tree_i \)
7: \( E_{OOB} = E \setminus E_i \)
8: \( \text{Update}_\text{Imp}(E_{OOB}, Tree, I) \)
9: \( I = \text{Average}(I, k) \)
10: return \( F, I \)
```
RandomForest Ranking

procedure Update_Imp\((E_{OOB}, \text{Tree}, I)\)
1: \(Err_{OOB} = \text{Evaluate}(\text{Tree}, E_{OOB})\)
2: for \(j = 1\) to \(D\) do
3: \(E_j = \text{Randomize}(E_{OOB}, j)\)
4: \(Err_j = \text{Evaluate}(\text{Tree}, E_j)\)
5: \(I_j = I_j + (Err_j - Err_{OOB})/Err_{OOB}\)
6: return

procedure Average\((I, k)\)
1: \(I^T = \emptyset\)
2: for \(l = 1\) to \(\text{size}(I)\) do
3: \(I_l^T = I_l/k\)
4: return \(I^T\)
RandomForest Ranking

$Importance(f_d) = \frac{1}{k} \cdot \sum_{i=1}^{k} \frac{Err_i(f_d) - Err(OOB_k)}{Err(OOB_k)}$

$k$ is the number of bootstrap replicates and $0 < d \leq D$
Random forest ranking for SOP

• This works for all types of outputs for which we can construct PCTs and ensembles thereof

• Multi-target classification

• Multi-label classification

• Hierarchical multi-label classification

• Multi-target regression
Combination of SOP with other complexity aspects

- RFs of PCTs & feature ranking therewith work with
  - Different types of SOP
  - With different degrees of supervision

- Unsupervised learning / clustering
- Semi-supervised learning
- Fully supervised learning for different types of SOP
  - Multi-target classification
  - Multi-label classification
  - Hierarchical multi-label classification
  - Multi-target regression
The MAESTRA foundation & pillars

• Incomplete annotations

• Massive/streaming data

• Network context
Coming up next:
Semi-supervised Learning of PCTs
Followed by:

Learning PCTs to Predict Structured Os from DS
Stay tuned for:

Learning in Networks

and

Applications of MBCD
Acknowledgements and announcement

We acknowledge European Commission support through the grants

• MAESTRA: Learning from Massive, Incompletely annotated, and Structured Data, grant 612944
• HBP SGA1: The Human Brain Project, grant 720270
• LANDMARK: LAND Management: Assessment, Research, Knowledge base, grant 635201

As well as the Slovenian Research Agency through

• P2-0103 Knowledge technologies
• L2-7509 Structured output prediction ...

And announce ...
ECML PKDD 2017
SKOPJE, MACEDONIA
18-22 September 2017