CTSI Module 8 Workshop
Biomedical Informatics, Data Standards & Practical Tools

Data Modeling and Evaluation with RapidMiner

Corey Arnold, PhD
Assistant Professor
Medical Imaging Informatics Group
Dept of Radiological Sciences
RapidMiner overview

- Open source software tool for data modeling and evaluation.
  - Logistic regression
  - Decision tree
  - Significance testing

- Tools for importing and transforming data from various sources.
  - Excel
  - Database
  - Delimited

- Implemented in Java with a user-friendly GUI.

- Available from http://rapid-i.com
Important Terms

• **Attribute.** The measurements/observations that describe the underlying process you are modeling (e.g., weight, ICD-9 code, SNPs, etc.).

• **Label.** A special attribute that identifies samples in relation to the research question (e.g., survival time, functional status, etc.).

• **Concept.** What is being represented by attributes (e.g., patients, doctors, nurses, etc.).

• **Example.** A specific instance of a concept in your dataset, which may be uniquely identified by an ID.
## Attribute Value Types

<table>
<thead>
<tr>
<th>Value type</th>
<th>RapidMiner name</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>nominal</td>
<td>Categorical non-numerical values, usually used for finite quantities of different characteristics</td>
</tr>
<tr>
<td>Numerical values</td>
<td>numeric</td>
<td>For numerical values in general</td>
</tr>
<tr>
<td>Integers</td>
<td>integer</td>
<td>Whole numbers, positive and negative</td>
</tr>
<tr>
<td>Real numbers</td>
<td>real</td>
<td>Real numbers, positive and negative</td>
</tr>
<tr>
<td>Text</td>
<td>text</td>
<td>Random free text without structure</td>
</tr>
<tr>
<td>2-value nominal</td>
<td>binominal</td>
<td>Special case of nominal, where only two different values are permitted</td>
</tr>
<tr>
<td>multi-value nominal</td>
<td>polynominal</td>
<td>Special case of nominal, where more than two different values are permitted</td>
</tr>
<tr>
<td>Date Time</td>
<td>data_time</td>
<td>Date as well as time</td>
</tr>
<tr>
<td>Date</td>
<td>date</td>
<td>Only date</td>
</tr>
<tr>
<td>Time</td>
<td>time</td>
<td>Only time</td>
</tr>
</tbody>
</table>
Analysis Overview

- **Process.** Defines the flow of data from its source, through transformation, modeling, and evaluation.
- **Operator.** Defines an action applied to data within a process. Has inputs and outputs.
Operators

• **Process Control.** Loops and conditional branches (e.g., reading all files in a directory).

• **Import/Export.** Reading/writing data and objects from/to external files and databases.

• **Data Transformation.** Data conversion, attribute selection, example filtering, imputation.

• **Modeling.** Data mining algorithms for classification, regression, clustering, association, correlation.

• **Evaluation.** Cross-validation, bootstrapping, significance testing.
RapidMiner Marketplace

• Repository for sharing RapidMiner code for different applications

• Extensions for:
  – Text analytics
  – High dimensional data
  – Parallel processing
  – Image processing
  – Web mining
  – Time series analysis