Data Mining

Introduction to RapidMiner
Outline

1. What is RapidMiner?
2. RapidMiner Terminology
3. RapidMiner User Interface
   1. The Design Perspective
   2. The Results Perspective
4. The RapidMiner Repository
5. Preprocessing Operators
6. Data Visualization
7. RapidMiner Resources
RapidMiner

- A very comprehensive open-source data mining tool
  - The data mining process is visually modeled as an operator chain
  - RapidMiner has over 400 built-in data mining operators
  - RapidMiner provides a broad collection of charts for visualizing data

- Project started in 2001 by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer at University of Dortmund, Germany

- Today: Maintained by a commercial company plus open-source developers

- RapidMiner Editions
  - Community Edition: Free (= Second Last Edition)
  - Enterprise Edition: Commercial (= Last Edition plus professional support)
### KDnuggets Poll 2014 (and 2013)

**What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [3285 voters]**

<table>
<thead>
<tr>
<th>Tool</th>
<th>2014 Users</th>
<th>2013 Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>RapidMiner (1453), 35.1% alone</td>
<td>44.2%</td>
<td>39.2%</td>
</tr>
<tr>
<td>R (1264), 2.1% alone</td>
<td>36.5%</td>
<td>37.4%</td>
</tr>
<tr>
<td>Excel (547), 0.1% alone</td>
<td>25.8%</td>
<td>28.0%</td>
</tr>
<tr>
<td>SQL (832), 0.1% alone</td>
<td>25.3%</td>
<td></td>
</tr>
<tr>
<td>Python (639), 0.9% alone</td>
<td>19.5%</td>
<td>13.3%</td>
</tr>
<tr>
<td>Weka (558), 0.4% alone</td>
<td>17.0%</td>
<td>14.3%</td>
</tr>
<tr>
<td>KNIME (492), 10.6% alone</td>
<td>15.0%</td>
<td></td>
</tr>
<tr>
<td>Hadoop (416), 0% alone</td>
<td>12.7%</td>
<td>9.3%</td>
</tr>
<tr>
<td>SAS base (357), 0% alone</td>
<td>10.9%</td>
<td>10.7%</td>
</tr>
<tr>
<td>Microsoft SQL Server (344), 0% alone</td>
<td>10.5%</td>
<td>7.0%</td>
</tr>
<tr>
<td>Revolution Analytics R (300), 13.3% alone</td>
<td>9.1%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Tableau (298), 1.3% alone</td>
<td>9.1%</td>
<td>6.3%</td>
</tr>
<tr>
<td>MATLAB (277), 0% alone</td>
<td>8.4%</td>
<td>9.0%</td>
</tr>
<tr>
<td>IBM SPSS Statistics (253), 0.4% alone</td>
<td>7.7%</td>
<td>8.7%</td>
</tr>
<tr>
<td>SAS Enterprise Miner (235), 1.3% alone</td>
<td>7.2%</td>
<td>5.9%</td>
</tr>
</tbody>
</table>

**Source:**
RapidMiner Terminology

- **Concept**: Class of examples, e.g. *person*, *flower*
- **Example**: Instance of a class, e.g. *Franz Müller*
- **Example Set**: The collection of all examples, that are mined.
- **Attributes**: Properties, e.g. *name*, *address*, *age*, *size*
  - Attributes have *Value Types* (see next slides)
  - Attributes have *Roles* (see next slides)
- **Process**: The complete data transformation and mining process.
- **Model**: The result of applying a process, e.g. *a classification model or set of rules*. 
## Value Types in RapidMiner

<table>
<thead>
<tr>
<th>Value Type</th>
<th>RapidMiner Name</th>
<th>Description</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>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>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>Date time</td>
<td>date_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>
<tr>
<td>Text</td>
<td>text</td>
<td>Random free text without structure</td>
</tr>
</tbody>
</table>
Attribute Roles in RapidMiner

- Special attributes
  - **Id**: A unique identifier of examples, e.g. *matriculation number*
  - **Cluster**: The assigned cluster of an example, e.g. *after clustering*
  - **Label**: The class attribute, e.g. *for classification tasks*
  - **Prediction**: The predicted label, e.g. *after applying a classification model*

- Regular Attributes
  - Regular attributes are used to describe characteristics of an example
The Design Perspective

Execute Process

Process View

List of Operators

Operators

Repository

Change Perspective

Parameter View

Help View
Specifying a Process by Chaining Operators

[Diagram showing the chaining of operators from Read Excel, Select Attributes, Set Role, Decision Tree, to Apply Model]
The Results Perspective

- Perspective for displaying and visualizing Example Sets and Models.

Data Statistics View

Raw Data View

Charts
The Repository

- Repository types
  - Local (files on your machine)
  - Remote (in the Cloud or own RapidMiner server)
- Location to store:
  - Datasets and its meta-data
  - Processes
- Data Import/Export Formats
  - CSV and XML Files
  - Excel Spreadsheets
  - Microsoft Access Tables
  - ARFF Files (Weka file format)
  - ...
(Some common) Preprocessing Operators

- **Type and Role Conversions**
  - Type: Provides for converting between different attribute types
  - Role: Provides for setting attribute roles (e.g. label, id)

- **Attribute Set Reduction and Transformation**
  - Select Attributes: Selects a sub set of the available attributes
  - Generate Attribute: Generates new attribute from existing attributes

- **Value Transformations**
  - Normalize: Normalizes attribute to certain range

- **Filtering**
  - Filter Examples: Filters the examples for a give criterion

- **Aggregation**
  - Performs a SQL-like aggregation (count, sum, ...)

Data Visualization

Visualization of data is one of the most powerful and appealing techniques for data exploration

• Humans have a well developed ability to analyze large amounts of information that is presented visually
• Can detect general patterns and trends
• Can detect outliers and unusual patterns

Visualization is the conversion of data into a visual format so that the characteristics of the data and the relationships among data items or attributes can be analyzed.
Visualization Techniques: Histogram

- Usually used to display the distribution of values of a **single attribute**
  - Divide the values into bins and show a bar plot of the number of objects in each bin
  - The height of each bar indicates the number of objects per bin
  - Shape of histogram depends on the number of bins

![RapidMiner Chart: Histogram](image)
Visualization Techniques: Scatter Charts

- Two-dimensional scatter charts are most commonly used.
- Often additional attributes/dimensions are displayed by using the size, shape, and color of the markers that represent the objects.
- It is useful to have arrays of scatter charts that can compactly summarize the relationships of several pairs of attributes.
- RapidMiner Scatter Charts
  - Scatter (single chart)
  - Scatter Multiple
  - Scatter Matrix
  - Scatter 3D
RapidMiner Chart: Scatter Matrix
RapidMiner Resources

- **RapidMiner 7:**
  - [https://my.rapidminer.com/nexus/account/index.html#downloads](https://my.rapidminer.com/nexus/account/index.html#downloads)
  - **License Key** can be found in ILIAS


- **Open Access Book covering RapidMiner**
  - Matthew North: Data Mining For The Masses: [http://dl.dropbox.com/u/31779972/DataMiningForTheMasses.pdf](http://dl.dropbox.com/u/31779972/DataMiningForTheMasses.pdf)

- **Operator Documentation:** [http://rapid-i.com/wiki/](http://rapid-i.com/wiki/)

- **RapidMiner Forum and Discussion Groups:** [http://forum.rapid-i.com/](http://forum.rapid-i.com/)

- **Video Tutorials**
  - by Rapid-I: [https://www.youtube.com/user/RapidIVideos](https://www.youtube.com/user/RapidIVideos)
  - by NDLR: [https://dspace.ndlr.ie/jspui/handle/10633/2353](https://dspace.ndlr.ie/jspui/handle/10633/2353)

- **MyExperiment:** process repository: [http://www.myexperiment.org/](http://www.myexperiment.org/)