eResearch 2014
Data Mining Workshop

Scientific workflow management with ADAMS
Outline

● Basics
  • Bit different, eh?
  • Features, User interface

● Data Mining
  • Feature generation, evaluation, visualization
  • Generate and use model

● Scripting
  • R, Groovy, Jython
Basics

• Canvas
Basics (2)

- Tree
Basics (3)

• How it works
  • No explicit connections
  • Actors snap into place in tree
  • Color coded with name (+ annotation)
    - standalone (no I/O)
    - source (only O)
    - transformer (I/O)
    - sink (only I)
    - control actor (data flow)
Basics (4)

• Advantages
  • compact layout
  • scales to 1000s of actors
  • context-aware adding of actors
  • interactive components
  • modular framework (Maven)
  • easy to add actors: 1 Java class, 1 icon
Basics (5)

- Limitations
  - only 1-to-n connections with tree layout
  - only single input/output

- Countermeasures
  - callable actors: n-to-1
  - containers: multiple outputs
  - variables: change options at runtime
  - internal storage: reuse data in multiple locations
# Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine learning/data mining</td>
<td>WEKA, WEKA webservice, MOA, MEKA, parameter optimization, experiment generation on-the-fly, setup generators, time series</td>
</tr>
<tr>
<td>Data processing</td>
<td>WEKA, R-Project, XML, XSLT, XPath, HTML, JSON</td>
</tr>
<tr>
<td>Streaming</td>
<td>MOA, Twitter (record/replay)</td>
</tr>
<tr>
<td>Spreadsheets</td>
<td>MS Excel (r/w), ODF (r/w), CSV (r/w), Gnumeric (r/w)</td>
</tr>
<tr>
<td>Imaging</td>
<td>ImageJ, JAI, ImageMagick, Gnuplot, OCR (tesseract)</td>
</tr>
<tr>
<td>Graphics output</td>
<td>BMP, JPG, PNG, TIF, PDF</td>
</tr>
<tr>
<td>Visualization</td>
<td>Scatter and line plots, Images, GIS (OpenStreetMap)</td>
</tr>
<tr>
<td>Scripting</td>
<td>Groovy, Jython</td>
</tr>
<tr>
<td>Documentation</td>
<td>DocBook, HTML</td>
</tr>
<tr>
<td>Web</td>
<td>HTTP, FTP, SFTP, SSH, Email, Webservices</td>
</tr>
<tr>
<td>Other</td>
<td>de/-compression (tar, zip, bzip2, gzip, lzma), Java code generation</td>
</tr>
</tbody>
</table>
User interface

- Main interfaces available from WEKA, MOA, MEKA, ImageJ
- Visualization
  - Preview browser, Time series explorer, Openstreetmap
- Tools
  - Flow editor/runner, Text editor/diff, PDF Viewer, Spreadsheet file viewer
- Misc preference and configuration panels
Data Mining

- **Lesson:** Image processing
  - Caltech 101 dataset
  - Pictures of objects belonging to 101 categories
  - Examples

- **Lesson downloads**
Data Mining (2)

- Get a feel for the data
Data Mining (3)

• Analyze categories

Flow
- SelectDirectory directory: "$HOME/documents/conferences/eResearchNZ2014/workshop/data/101_ObjectCategories" [stops flow if canceled]
- SetVariable $dir [REPLACE]

list categories
- FileSystemSearch DirectorySearch $dir [SORT_BY_NAME]
- SetVariable $sub_dir [REPLACE]

count
- FileSystemSearch FileSearch $sub_dir [SORT_BY_NAME]
- ArrayLength $num
- SetVariable $num [REPLACE]

determine category
- BaseName $category
- SetVariable $category [REPLACE]

output info
- CombineVariables $category, $num
Data Mining (4)

- “Fancy” display of categories

- Flow
  - SelectDirectory directory: ${HOME}/documents/conferences/eResearchNZ2014/workshop/data\101_ObjectCategories [stops flow if canceled]
  - SetVariable @\{dir\} [REPLACE]
  - init spreadsheet
    - NowSpreadSheet name: Category, Count, cols: Category, Count, row type: DENSE, sheet: Spreadsheet
    - SetStorageValue sheet
  - list categories
    - FileSystemSearch DirectorySearch (@\{dir\} [SORT_BY_NAME])
    - SetVariable @\{sub_dir\} [REPLACE]
    - FileSystemSearch FileSearch (@\{sub_dir\} [SORT_BY_NAME])
    - ArrayLength
    - SetVariable @\{num\} [REPLACE]
  - determine category
    - BaseName
    - SetVariable @\{category\} [REPLACE]
  - add row
    - StorageValue sheet
    - SpreadsheetInsertRow after: last, insert: '?' [no copy]
    - set category row: last/col: 1, value: @\{category\}, no copy
    - set count row: last/col: 2, value: @\{num\}, no copy
    - setStorageValue sheet
  - display data
    - StorageValue sheet
    - SpreadsheetDisplay X: left, Y: top, W: 640, H: 480, font: Monospaced-PLAIN-12, decimals: 3, optimal
Generate features/training data

- Flow
- CallableActors
  - progress
  - SetVariable: @{dir} [REPLACE]
- create output filename
  - AppendName: train.arff
  - SetVariable: @{outfile} [REPLACE]
- list categories
  - FileSystemSearch: DirectorySearch(@{dir} [SORT_BY_NAME])
  - SetVariable: @{sub_dir} [REPLACE]
- determine category
  - BaseName
  - SetVariable: @{category} [REPLACE]
  - CallableSink: progress
- traverse images
  - FileSystemSearch: FileSearch(@sub_dir) [SORT_BY_NAME]
  - JAIReader: Histogram
  - JAIFlattener: Histogram
  - WekaFilter: Add-1 STR -N category -C last
  - WekaSetInstanceValue: last -> @{category}
  - WekaInstanceDumper: ARFF: @{outfile}
Data Mining (6)

- Test classifier setups
Data Mining (7)

- Evaluate classifiers
Data Mining (8)

• Build and use model

```
Flow
  SetVariable @{model} = {TMP}/101.model
  Start

  determine # of classes
  Variable @{model}:
  WekaModelReader
  ContainerValuePicker Header [outputs switched]

  Tee
  WekaInstancesInfo NUM_CLASS_LABELS
  SetVariable @{num_labels} [REPLACE]
  WekaInstancesInfo CLASS_LABELS
  SequenceToArray Length @{num_labels}, Class: -from 1st element-
  StringJoin glue ,
  SetVariable @{labels} [REPLACE]

  make predictions
  SelectFile directory: {CWD} [stops flow if canceled]
  JAIReader
  JAIFlattener Histogram
  UpdateProperties props: filter.nominalLabels, vars: labels
  WekaFilter Add -T NOM -N category 1. Faces, airplane...
  WekaClassSelector last, name:*
  WekaClassifying @{model}
  ContainerValuePicker Classification label [outputs switched]
```
Scripting

• **Lesson:** Scripting with R, Groovy, Jython
  • Near Infrared (NIR) data in proprietary Bruker SPC format
  • hyperSpec R package reads certain SPC files
    http://cran.r-project.org/web/packages/hyperSpec/
  • Install R packages
    > install.packages("Rserve")
    > install.packages("hyperSpec")
  • Configure Jython $HOME/.jython
    python.security.respectJavaAccessibility=false
Scripting (2)

- Use rkward to inspect data
Scripting (3)

• Connect to R via Rserve and run script to extract double matrix

- **Flow**
  - Rserve host: localhost port: 6311
  - SelectFile directory: ${CWD}
  - SetVariable @{file} [REPLACE]
  - RTransformer script: library(hyperSpec); nspc = hyperSpec::rea..., return: DoubleMatrix
  - Convert DoubleMatrixToSpreadSheet
  - Convert-1 SpreadsheetAddRowID (start: 1)
  - Convert-2 TransposeSpreadSheet
  - Branch parallel, threads: #cores
  - SpreadsheetDisplay X: left, Y: top, W: 640, H: 480, font: Monospaced-PLAIN-12, decimals
  - Sequence
    - SpreadsheetPlotGenerator generator: SimplePlotGen
    - SequencePlotter X: left, Y: top, W: 800, H: 350
Scripting (4)

- Use Groovy to compute 1\textsuperscript{st} derivative
Scripting (5)

- Use Jython to computer 1st derivative
Discussion

Questions?