RAPIDMINER 5.0

Ricardo Aler Mur

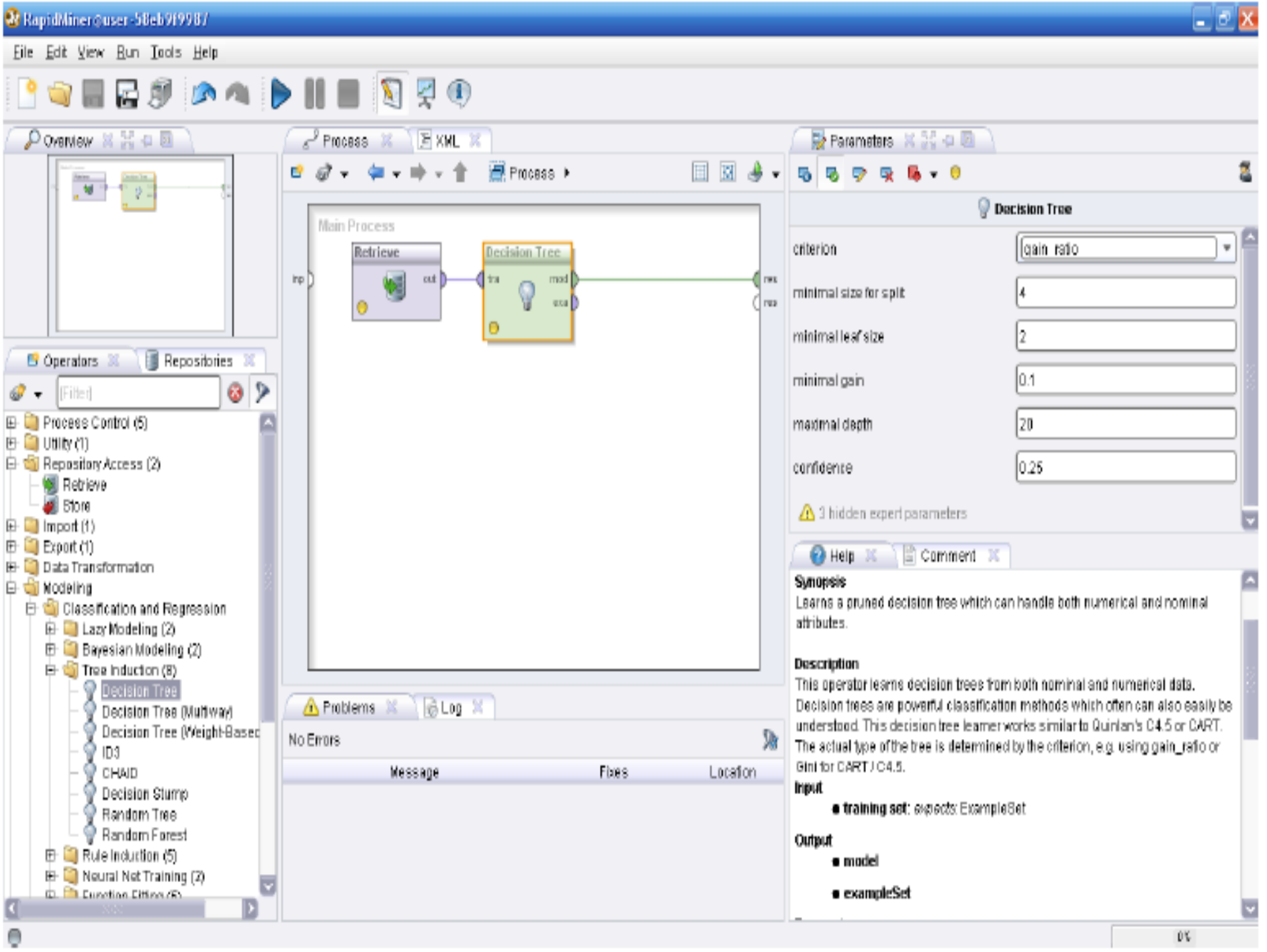
ÍNDICE

1. Árbol de decisión
2. Reglas
3. Árbol de decisión con validación cruzada
4. Árbol de decisión vs. knn con validación cruzada
5. Árbol decisión vs. knn con validación cruzada evaluando el coste (con matriz de costes)
6. Árbol de decisión vs. Metacost con árboles de decisión y matriz de costes
7. Comparación de curvas ROC

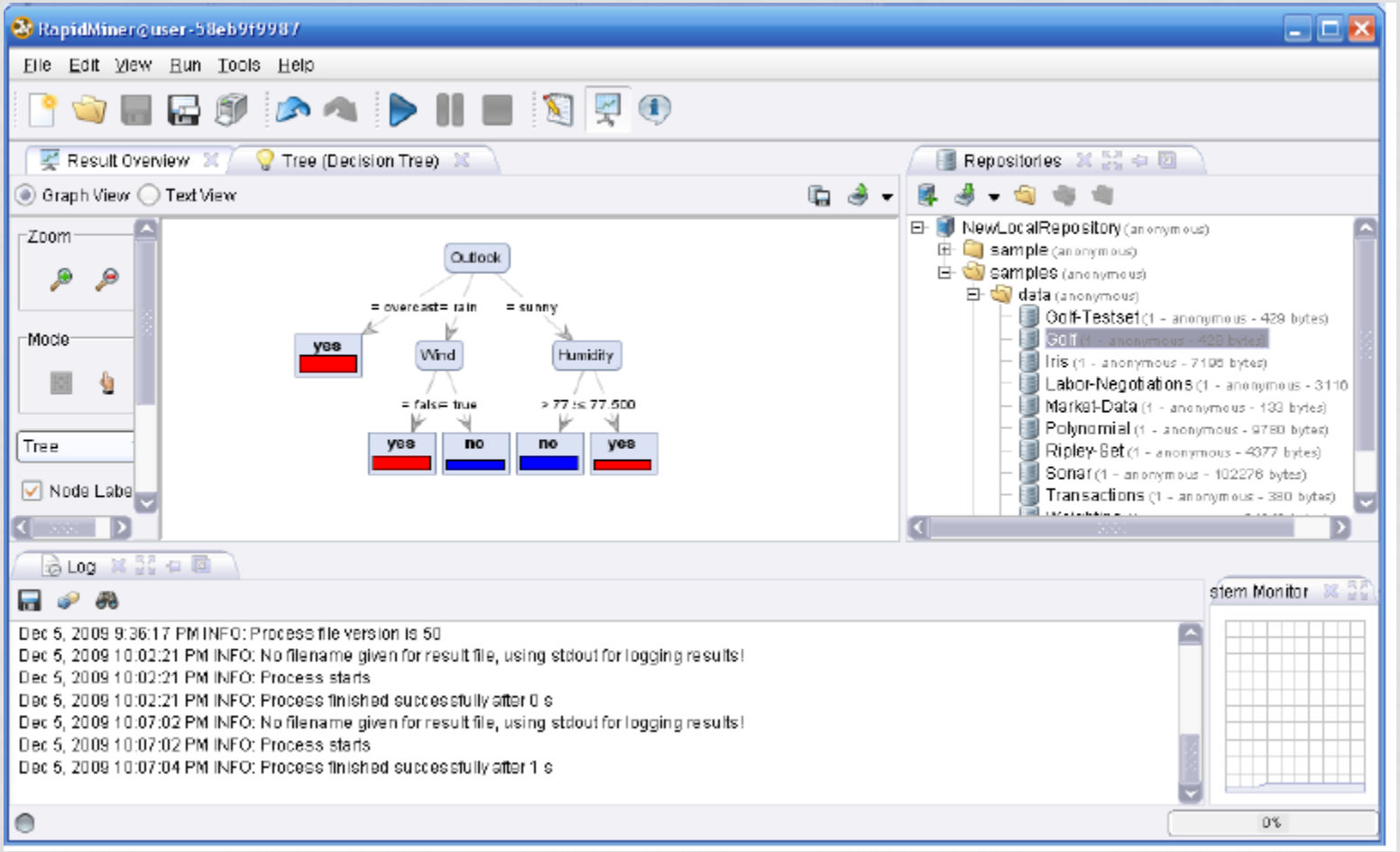
1. ÁRBOL DE DECISIÓN

- Ejemplo 1 del tutorial:

- http://www.dataprix.com/files/RapidMiner_Tutorial_online_Operadores.pdf


RapidMiner@user-58eb9f9987
File Edit View Run Tools Help
Overview
Process
XML
Process
Main Process
Retrieve
Decision Tree
Parameters
Decision Tree
criterion
gain_ratio
minimal size for split
4
minimal leaf size
2
minimal gain
0.1
maximal depth
20
confidence
0.25
3 hidden expert parameters
Operators
Repositories
Process Control (5)
Utility (1)
Repository Access (2)
Retrieve
Store
Import (1)
Export (1)
Data Transformation
Modeling
Classification and Regression
Lazy Modeling (2)
Bayesian Modeling (2)
Tree Induction (8)
Decision Tree
Decision Tree (Multiway)
Decision Tree (Weight-Based)
ID3
CHAID
Decision Stump
Random Tree
Random Forest
Rule Induction (5)
Neural Net Training (2)
Problems
Log
No Errors
Message
Fixes
Location
Help
Comment
Synopsis
Learns a pruned decision tree which can handle both numerical and nominal attributes.
Description
This operator learns decision trees from both nominal and numerical data. Decision trees are powerful classification methods which often can also easily be understood. This decision tree learner works similar to Quinlan's C4.5 or CART. The actual type of the tree is determined by the criterion, e.g. using gain_ratio or Gini for CART / C4.5.
Input
training set: expects: ExampleSet
Output
model
exampleSet

ÁRBOL DE DECISIÓN



TECLAS IMPORTANTES

- F8 o View -> Perspectives -> Design
  - Vista de procesos
- F9 o View -> Perspectives -> Results
  - Vista de resultados

2. REGLA DE DECISIÓN

- Pinchar con el botón derecho del ratón sobre la caja del árbol de decisión y reemplazar el modelo para usar reglas (Rule induction)

8. Reemplazar el aprendiz por otro esquema de aprendizaje para tareas de clasificación. Hacer clic derecho sobre el operador **Decision Tree** y seleccionar **Replace Operator → Modeling → Classification and Regression → Rule Induction → Rule Induction**. Después de ejecutar el proceso cambiado con este ejemplo, se presenta el Nuevo modelo:

```
IF Cielo = Cubierto THEN Sí
IF Temperatura ≤ 77.500 AND Ventoso = Falso AND Cielo = Lluvioso THEN
Sí
IF Cielo = Lluvioso THEN No
IF Humedad > 77.500 THEN No ELSE Sí
```

3. CLASIFICACIÓN CON VALIDACIÓN CRUZADA

- Ejemplo 9 del tutorial, pero para clasificación

<new process*> – RapidMiner 5.3.015 @ PC-164-148
File Edit Process Tools View Help
miperspectiva
Process XML
Process
Main Process
Retrieve
inp out
Validation
tra mod tra ave ave
res res
Parameters Context
Validation (X-Validation)
average performances only
leave one out
number of validations 10
sampling type stratified sampling
use local random seed
parallelize training
parallelize testing
Compatibility level 5.3.015
Help Comment
X-Validation (RapidMiner Core)
Synopsis
This operator performs a cross-validation in order to estimate the statistical performance of a learning operator (usually on unseen data sets). It is mainly used to estimate how accurately a model (learnt by a particular learning operator) will perform in practice.
Problems Log
One potential problem
Message Fixes Location
Parameter 'repository entry' accesses a repository by name (//Local Repository/data/oil... No quick fix available Retrieve
Double-click to zoom, drag to move.

<new process*> – RapidMiner 5.3.015 @ PC-164-148
File Edit Process Tools View Help
miperspectiva
Process XML
Process ▸ Validation ▸
Training
Decision Tree
tra
tra mod
exa
mod mod
thr tes
thr
Testing
Apply Model
mod lab
unl mod
Performance ...
lab per
per exa
ave
ave
Parameters Context
Validation (X-Validation)
average performances only
leave one out
number of validations 10
sampling type stratified sampling
use local random seed
parallelize training
parallelize testing
Compatibility level 5.3.015
Help Comment
X-Validation (RapidMiner Core)
Synopsis
This operator performs a cross-validation in order to estimate the statistical performance of a learning operator (usually on unseen data sets). It is mainly used to estimate how accurately a model (learnt by a particular learning operator) will perform in practice.
Problems Log
One potential problem
Message Fixes Location
Parameter 'repository entry' accesses a repository by name (//Local Repository/data/oil... No quick fix available Retrieve

4. ÁRBOL DE DECISIÓN Y KNN CON VALIDACIÓN CRUZADA

- Usar File -> Open **Template** -> Compare Learning Algorithms by Significance Test
- Reemplazar el primer clasificador por un árbol de decisión
- Reemplazar el segundo clasificador por KNN
- Reemplazar ambos "regression performance measurement" por "binomial performance measurement". Hacer que se impriman, aparte de "accuracy", también "AUC"

//Local Repository/processes/PRUEBAS* – RapidMiner 5.3.015 @ PC-164-148
File Edit Process Tools View Help
miperspectiva
Result Overview
PerformanceVector (Performance (2))
PerformanceVector (Performance)
Pairwise t-Test (T-Test)
Anova Test (Anova)
Repositories
Table / Plot View Text View Annotations
Criterion Selector
classification_error
AUC
false_positive
false_negative
true_positive
true_negative
Multiclass Classification Performance Annotations
Table View Plot View
classification_error: 7.15% +/- 0.82% (mikro: 7.15%)
true 2.000 true 1.000 class precision
pred. 2.000 6 32 15.79%
pred. 1.000 35 864 96.11%
class recall 14.63% 96.43%
Samples (none)
DB
Local Repository (Aler)
Log
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Testing
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Training
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Testing
Nov 27, 2013 8:24:19 PM INFO: Saving results.
Nov 27, 2013 8:24:19 PM INFO: Process //Local Repository/processes/PRUEBAS finished successfully after 2 s
System Monitor
Max: 10 GB
Total: 1024 MB

//Local Repository/processes/PRUEBAS* – RapidMiner 5.3.015 @ PC-164-148
File Edit Process Tools View Help
miperspectiva
Result Overview
PerformanceVector (Performance (2))
PerformanceVector (Performance)
Pairwise t-Test (T-Test)
Anova Test (Anova)
Repositories
Table / Plot View Text View Annotations
Criterion Selector
classification_error
AUC
false_positive
false_negative
true_positive
true_negative
Multiclass Classification Performance Annotations
Table View Plot View
classification_error: 4.69% +/- 1.97% (mikro: 4.70%)
true 2.000 true 1.000 class precision
pred. 2.000 12 15 44.44%
pred. 1.000 29 881 96.81%
class recall 29.27% 98.33%
Samples (none)
DB
Local Repository (Aler)
Log
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Testing
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Training
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Testing
Nov 27, 2013 8:24:19 PM INFO: Saving results.
Nov 27, 2013 8:24:19 PM INFO: Process //Local Repository/processes/PRUEBAS finished successfully after 2 s
System Monitor
Max: 10 GB
Total: 1024 MB

miperspectiva

Result Overview | PerformanceVector (Performance (2)) | PerformanceVector (Performance) | Pairwise t-Test (T-Test) | Anova Test (Anova)

T-Test Significance | Text View | Annotations

T-Test Significance

	0.047 +/- 0.020	0.071 +/- 0.008
0.047 +/- 0.020		**0.002**
0.071 +/- 0.008		

Probabilities for random values with the same result.
Bold values are smaller than alpha=0.050 which indicates a probably significant difference between the actual mean values!

Repositories

- Samples (none)
- DB
- Local Repository (Aler)

Log

Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Testing
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Training
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Testing
Nov 27, 2013 8:24:19 PM INFO: Saving results.
Nov 27, 2013 8:24:19 PM INFO: Process //Local Repository/processes/PRUEBAS finished successfully after 2 s

System Monitor

Max: 10 GB
Total: 1024 MB

File Edit Process Tools View Help

miperspectiva

Result Overview | PerformanceVector (Performance (2)) | PerformanceVector (Performance) | Pairwise t-Test (T-Test) | Anova Test (Anova)

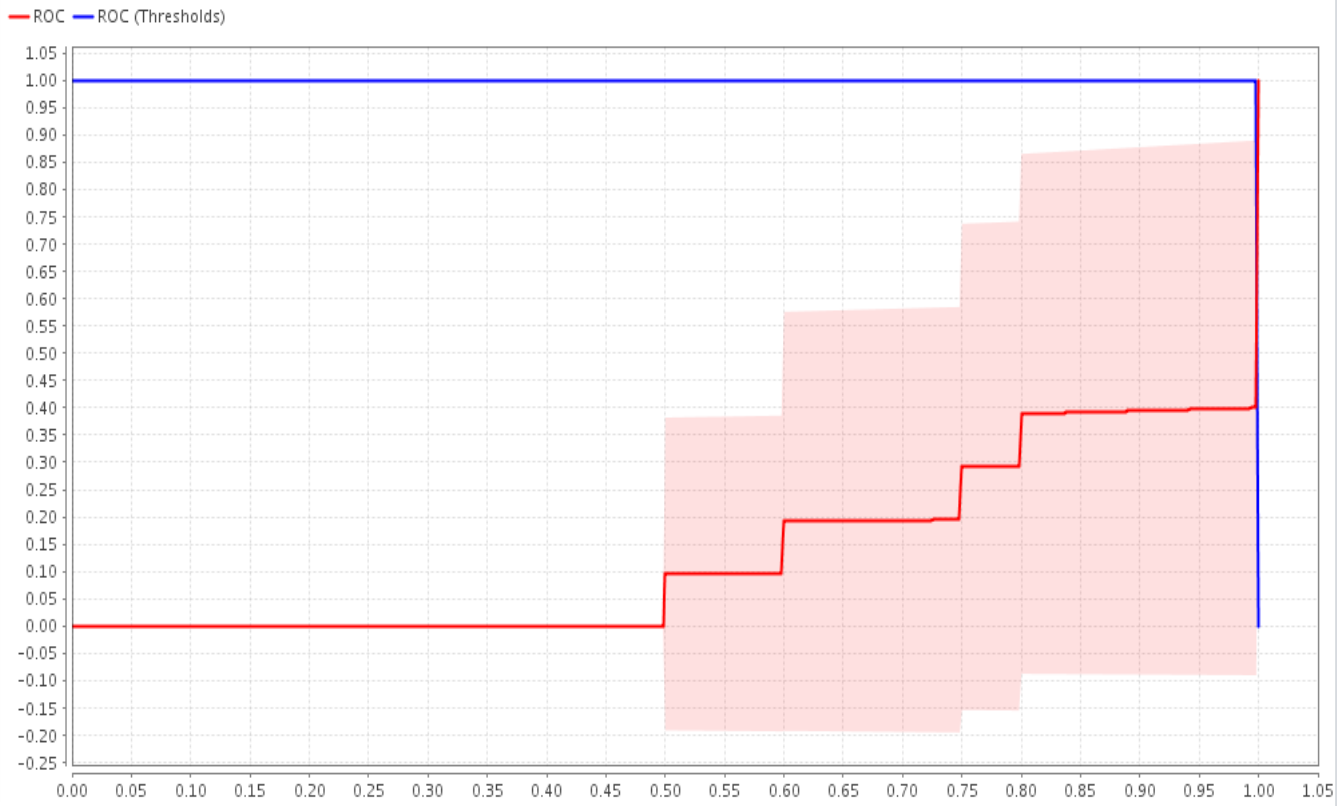
Table / Plot View | Text View | Annotations

Criterion Selector

- classification_error
- AUC
- false_positive
- false_negative
- true_positive
- true_negative

Area Under Curve | Text View | Annotations

AUC: 0.500 +/- 0.000 (mikro: 0.500) (positive class: 1.000)



Repositories

- Samples (none)
- DB
- Local Repository (Aler)

Log

Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Testing
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Training
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Testing
Nov 27, 2013 8:24:19 PM INFO: Saving results.
Nov 27, 2013 8:24:19 PM INFO: Process //Local Repository/processes/PRUEBAS finished successfully after 2 s

System Monitor

Max: 10 GB
Total: 1024 MB

//Local Repository/processes/PRUEBAS* – RapidMiner 5.3.015 @ PC-164-148

File Edit Process Tools View Help

miperspectiva

Result Overview | PerformanceVector (Performance (2)) | PerformanceVector (Performance) | Pairwise t-Test (T-Test) | Anova Test (Anova)

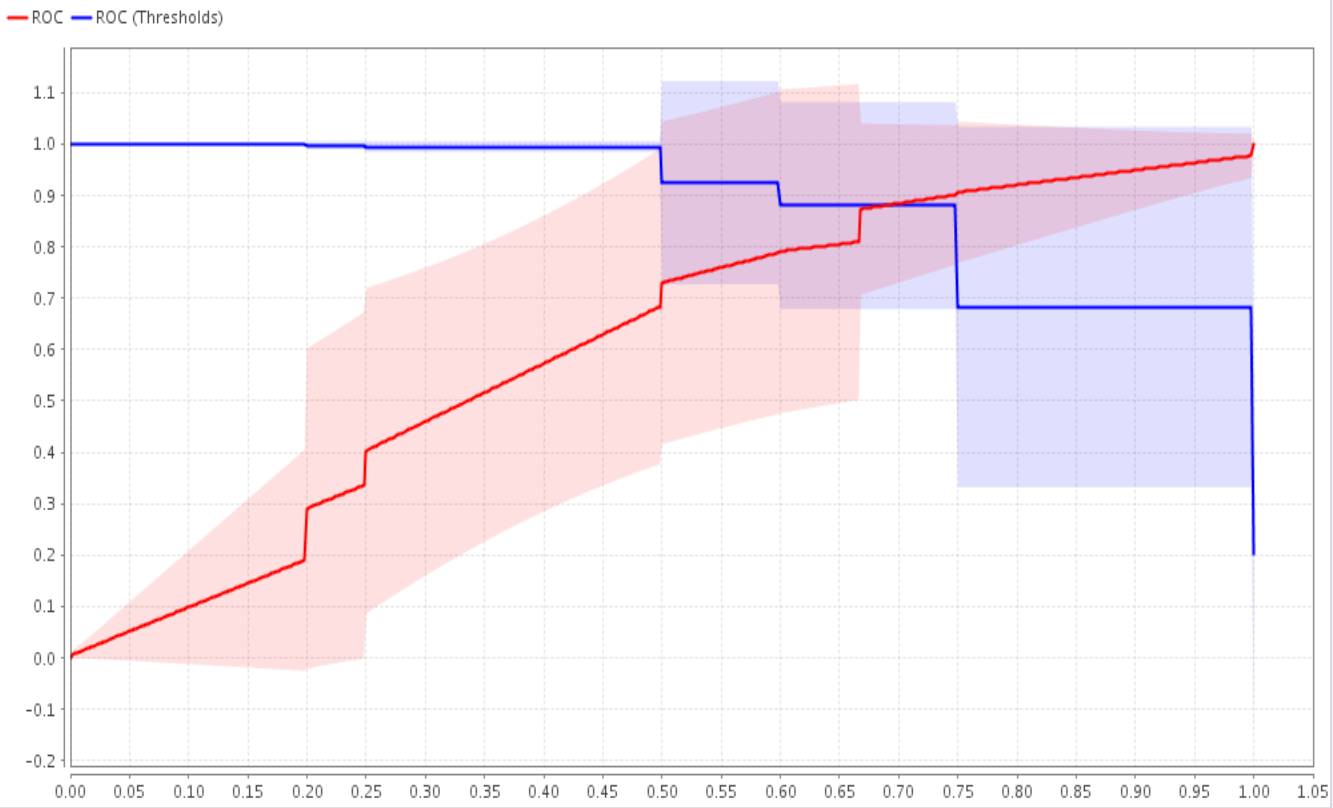
Table / Plot View | Text View | Annotations

Criterion Selector

- classification_error
- AUC
- false_positive
- false_negative
- true_positive
- true_negative

Area Under Curve | Text View | Annotations

AUC: 0.610 +/- 0.174 (mikro: 0.610) (positive class: 1.000)



Repositories

- Samples (none)
- DB
- Local Repository (Aler)

Log

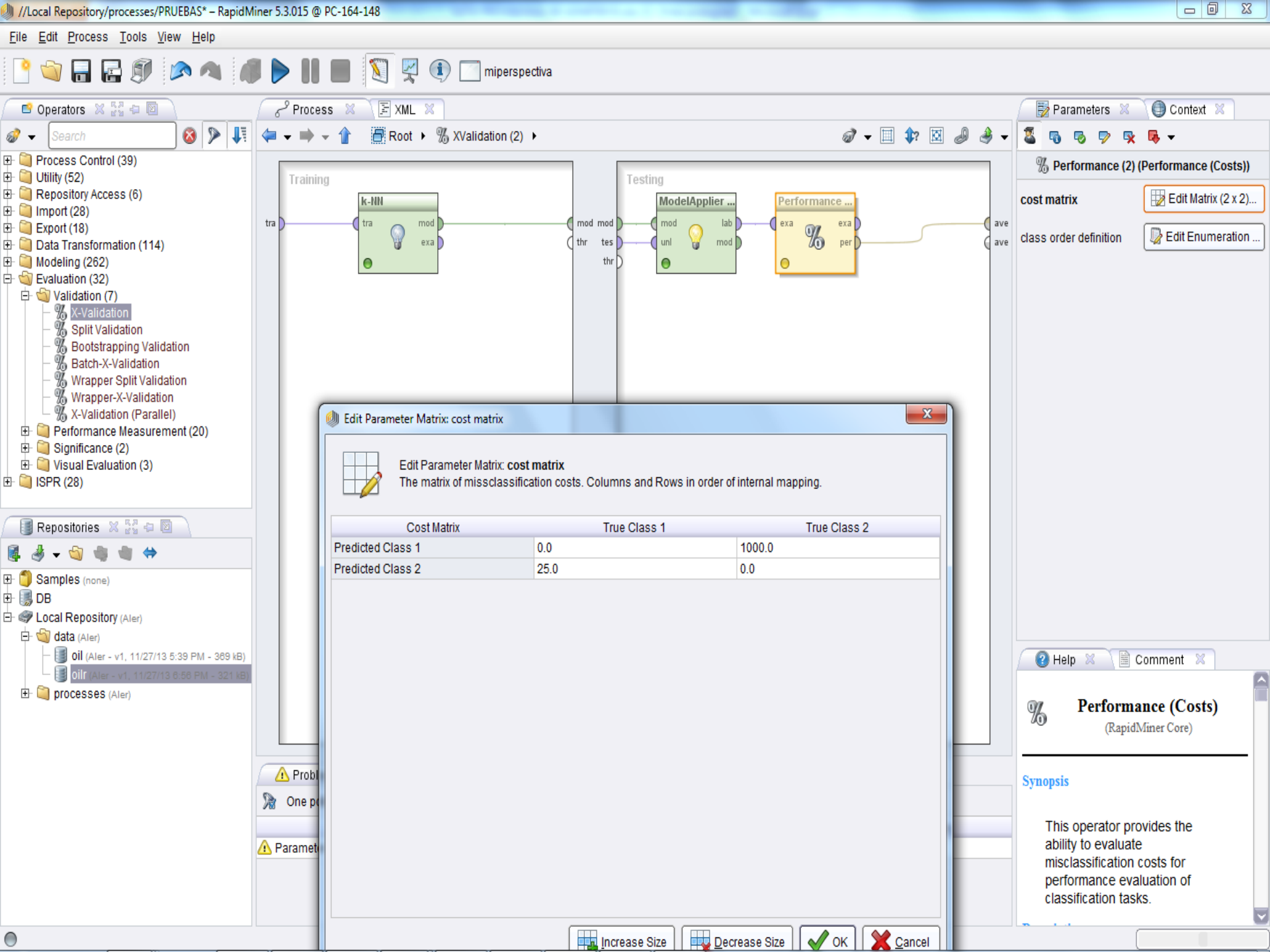
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Testing
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Training
Nov 27, 2013 8:24:19 PM INFO: Executing process concurrently: Testing
Nov 27, 2013 8:24:19 PM INFO: Saving results.
Nov 27, 2013 8:24:19 PM INFO: Process //Local Repository/processes/PRUEBAS finished successfully after 2 s

System Monitor

Max: 10 GB
Total: 1024 MB

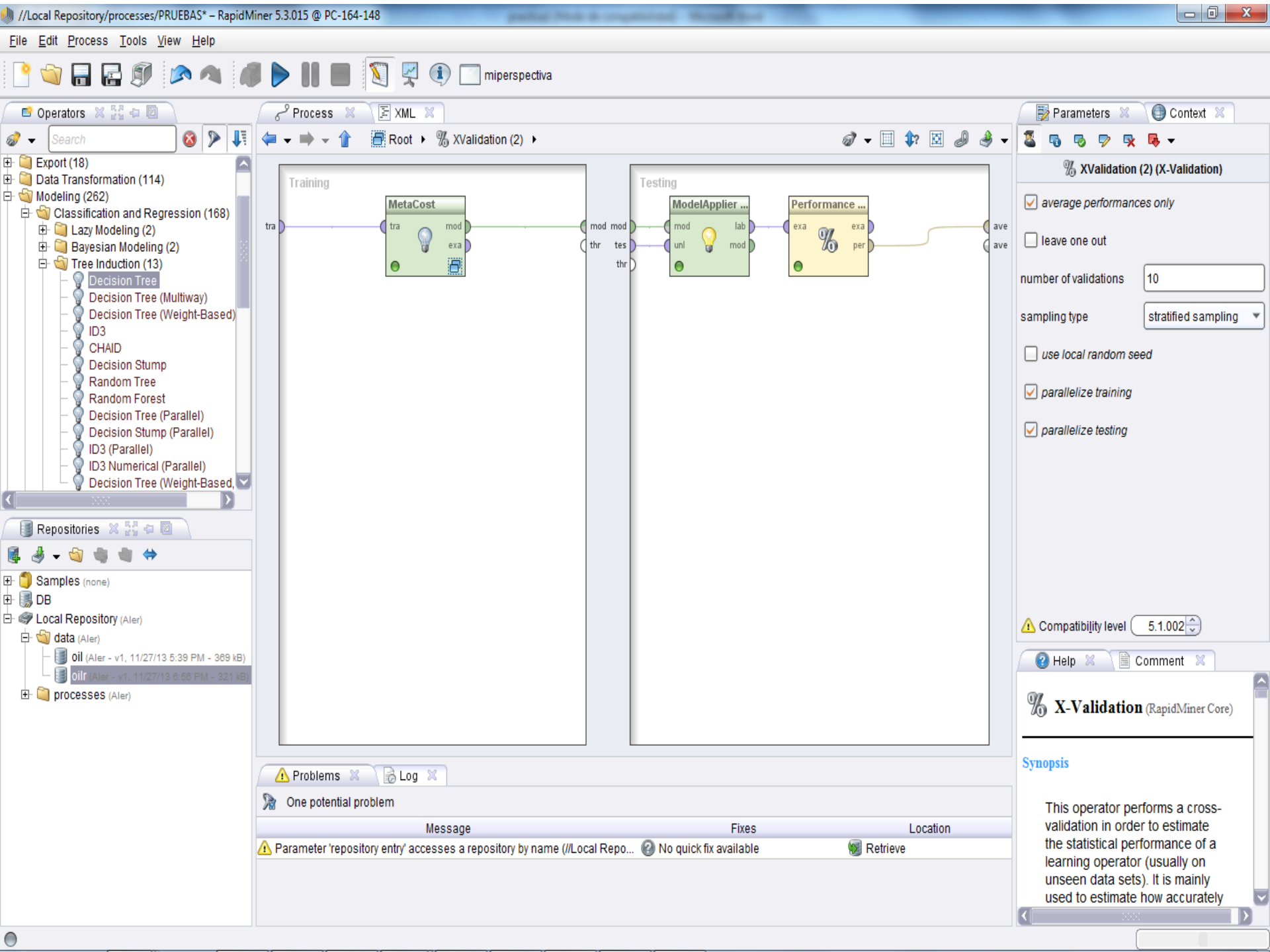
5. ÁRBOL DECISIÓN + CON MATRIZ DE COSTES)

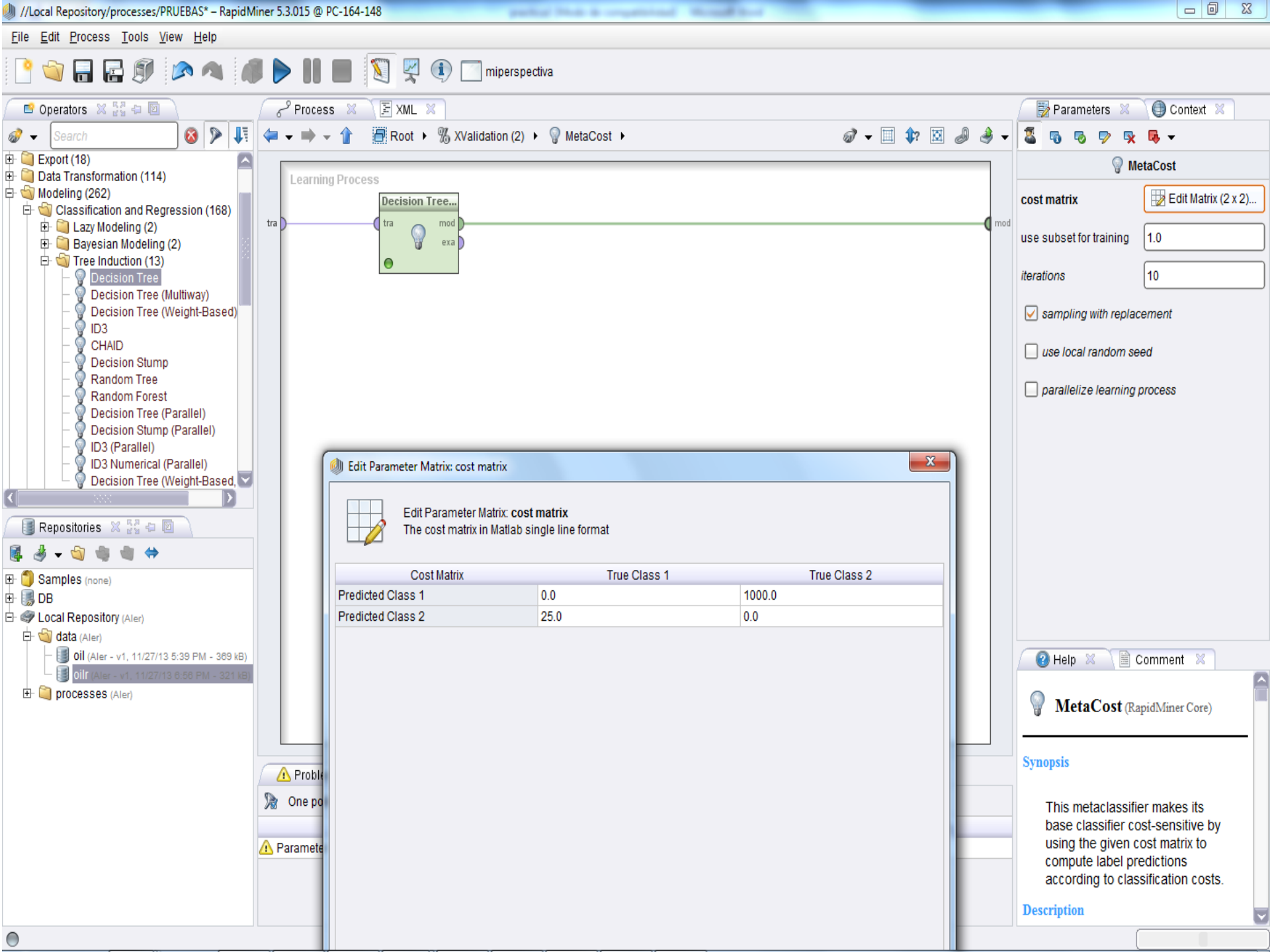
- Reemplazar ambos “binomial performance measurement” por “performance costs”.
- Introducir en ambos una matriz de costes, de tal manera que el coste de confundir la clase 1 por la dos sea de


//Local Repository/processes/PRUEBAS* – RapidMiner 5.3.015 @ PC-164-148
File Edit Process Tools View Help
miperspectiva
Operators
Search
Process Control (39)
Utility (52)
Repository Access (6)
Import (28)
Export (18)
Data Transformation (114)
Modeling (262)
Evaluation (32)
Validation (7)
X-Validation
Split Validation
Bootstrapping Validation
Batch-X-Validation
Wrapper Split Validation
Wrapper-X-Validation
X-Validation (Parallel)
Performance Measurement (20)
Significance (2)
Visual Evaluation (3)
ISPR (28)
Repositories
Samples (none)
DB
Local Repository (Aler)
data (Aler)
oil (Aler - v1, 11/27/13 5:39 PM - 369 kB)
oilr (Aler - v1, 11/27/13 6:56 PM - 321 kB)
processes (Aler)
Process
XML
Root
XValidation (2)
Training
k-NN
tra
mod
exa
Testing
ModelApplier ...
Performance ...
mod
lab
unl
exa
per
ave
Parameters
Context
Performance (2) (Performance (Costs))
cost matrix
Edit Matrix (2 x 2)...
class order definition
Edit Enumeration ...
Help
Comment
Performance (Costs)
(RapidMiner Core)
Synopsis
This operator provides the ability to evaluate misclassification costs for performance evaluation of classification tasks.
Edit Parameter Matrix: cost matrix
Edit Parameter Matrix: cost matrix
The matrix of missclassification costs. Columns and Rows in order of internal mapping.
Cost Matrix
True Class 1
True Class 2
Predicted Class 1
0.0
1000.0
Predicted Class 2
25.0
0.0
Increase Size
Decrease Size
OK
Cancel

6. ÁRBOL DE DECISIÓN VS. METACOST CON ÁRBOLES DE DECISIÓN Y MATRIZ DE COSTES

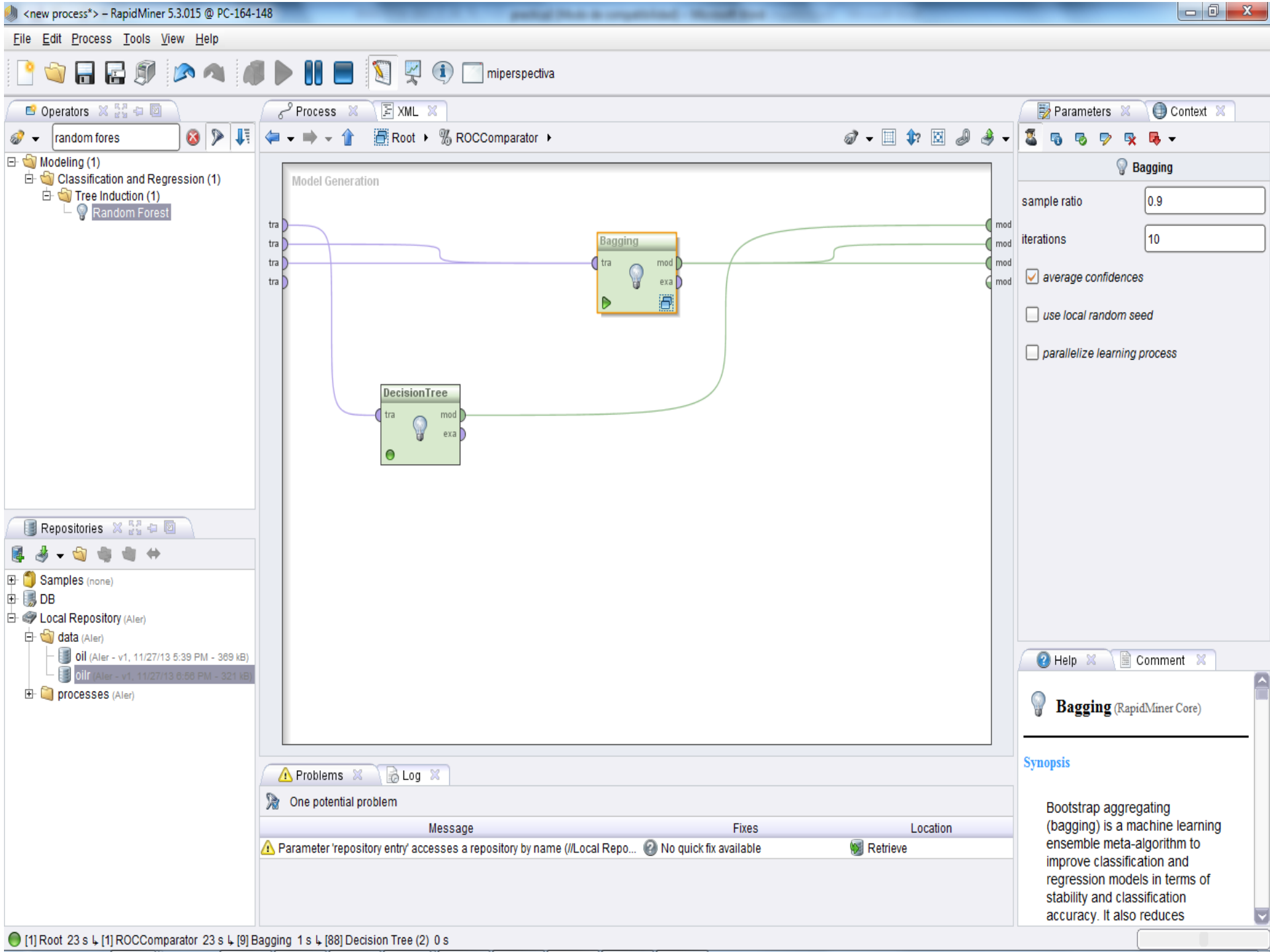
- Vamos a sustituir knn por metacost, que recordemos era capaz de estimar probabilidades y dada una matriz de costes, era capaz de devolver un clasificador óptimo para ella
- Metacost es un clasificador "meta" y trabaja con cualquier clasificador. En este caso utilizaremos como clasificador para metacost a los árboles de decisión


//Local Repository/processes/PRUEBAS* – RapidMiner 5.3.015 @ PC-164-148
File Edit Process Tools View Help
miperspectiva
Operators
Search
Export (18)
Data Transformation (114)
Modeling (262)
Classification and Regression (168)
Lazy Modeling (2)
Bayesian Modeling (2)
Tree Induction (13)
Decision Tree
Decision Tree (Multiway)
Decision Tree (Weight-Based)
ID3
CHAID
Decision Stump
Random Tree
Random Forest
Decision Tree (Parallel)
Decision Stump (Parallel)
ID3 (Parallel)
ID3 Numerical (Parallel)
Decision Tree (Weight-Based,
Repositories
Samples (none)
DB
Local Repository (Aler)
data (Aler)
oil (Aler - v1, 11/27/13 5:39 PM - 369 kB)
oilr (Aler - v1, 11/27/13 6:56 PM - 321 kB)
processes (Aler)
Process
XML
Root
XValidation (2)
Training
MetaCost
tra
mod
exa
Testing
ModelApplier ...
mod
lab
unl
Performance ...
exa
per
thr
tes
ave
Problems
Log
One potential problem
Message
Fixes
Location
Parameter 'repository entry' accesses a repository by name (//Local Repo...
No quick fix available
Retrieve
Parameters
Context
XValidation (2) (X-Validation)
average performances only
leave one out
number of validations
10
sampling type
stratified sampling
use local random seed
parallelize training
parallelize testing
Compatibility level
5.1.002
Help
Comment
X-Validation (RapidMiner Core)
Synopsis
This operator performs a cross-validation in order to estimate the statistical performance of a learning operator (usually on unseen data sets). It is mainly used to estimate how accurately


//Local Repository/processes/PRUEBAS* – RapidMiner 5.3.015 @ PC-164-148
File Edit Process Tools View Help
miperspectiva
Operators
Search
Export (18)
Data Transformation (114)
Modeling (262)
Classification and Regression (168)
Lazy Modeling (2)
Bayesian Modeling (2)
Tree Induction (13)
Decision Tree
Decision Tree (Multiway)
Decision Tree (Weight-Based)
ID3
CHAID
Decision Stump
Random Tree
Random Forest
Decision Tree (Parallel)
Decision Stump (Parallel)
ID3 (Parallel)
ID3 Numerical (Parallel)
Decision Tree (Weight-Based,
Repositories
Samples (none)
DB
Local Repository (Aler)
data (Aler)
oil (Aler - v1, 11/27/13 5:39 PM - 369 kB)
oilr (Aler - v1, 11/27/13 6:56 PM - 321 kB)
processes (Aler)
Process
XML
Root
XValidation (2)
MetaCost
Learning Process
Decision Tree...
tra
mod
exa
Edit Parameter Matrix: cost matrix
Edit Parameter Matrix: cost matrix
The cost matrix in Matlab single line format
Cost Matrix
True Class 1
True Class 2
Predicted Class 1
0.0
1000.0
Predicted Class 2
25.0
0.0
Parameters
Context
MetaCost
cost matrix
Edit Matrix (2 x 2)...
use subset for training
1.0
iterations
10
sampling with replacement
use local random seed
parallelize learning process
Help
Comment
MetaCost (RapidMiner Core)
Synopsis
This metaclassifier makes its base classifier cost-sensitive by using the given cost matrix to compute label predictions according to classification costs.
Description

7. COMPARACIÓN DE CURVAS ROC

- Usaremos la File -> Template “compare ROCs”
- Incluiremos decision tree y bagging con decision trees


<new process*> – RapidMiner 5.3.015 @ PC-164-148
File Edit Process Tools View Help
miperspectiva
Operators
random fores
Modeling (1)
Classification and Regression (1)
Tree Induction (1)
Random Forest
Process
XML
Root
ROCComparator
Model Generation
tra
tra
tra
tra
Bagging
tra
mod
exa
DecisionTree
tra
mod
exa
mod
mod
mod
mod
Parameters
Context
Bagging
sample ratio
0.9
iterations
10
average confidences
use local random seed
parallelize learning process
Repositories
Samples (none)
DB
Local Repository (Aler)
data (Aler)
oil (Aler - v1, 11/27/13 5:39 PM - 369 kB)
oilr (Aler - v1, 11/27/13 6:56 PM - 321 kB)
processes (Aler)
Problems
Log
One potential problem
Message
Fixes
Location
Parameter 'repository entry' accesses a repository by name (//Local Repo...
No quick fix available
Retrieve
Help
Comment
Bagging (RapidMiner Core)
Synopsis
Bootstrap aggregating (bagging) is a machine learning ensemble meta-algorithm to improve classification and regression models in terms of stability and classification accuracy. It also reduces
[1] Root 23 s ↳ [1] ROCComparator 23 s ↳ [9] Bagging 1 s ↳ [88] Decision Tree (2) 0 s

<new process*> – RapidMiner 5.3.015 @ PC-164-148
File Edit Process Tools View Help
miperspectiva
Result Overview
ROC Comparison (ROCComparator)
ExampleSet (/Local Repository/data/oilr)
ROC Comparison
Annotations
Bagging
DecisionTree
1.20 1.15 1.10 1.05 1.00 0.95 0.90 0.85 0.80 0.75 0.70 0.65 0.60 0.55 0.50 0.45 0.40 0.35 0.30 0.25 0.20 0.15 0.10 0.05 0.00
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05
Repositories
Samples (none)
DB
Local Repository (Aler)
data (Aler)
oil (Aler - v1, 11/27/13 5:39 PM - 369 kB)
oilr (Aler - v1, 11/27/13 6:58 PM - 321 kB)
processes (Aler)
Log
Nov 27, 2013 9:26:39 PM INFO: Executing process concurrently: Model Generation
Nov 27, 2013 9:26:41 PM INFO: Executing process concurrently: Model Generation
Nov 27, 2013 9:26:44 PM INFO: Executing process concurrently: Model Generation
Nov 27, 2013 9:26:47 PM INFO: Saving results.
Nov 27, 2013 9:26:47 PM INFO: Process finished successfully after 24 s
System Monitor
Max: 10 GB
Total: 1024 MB