

ROC curves

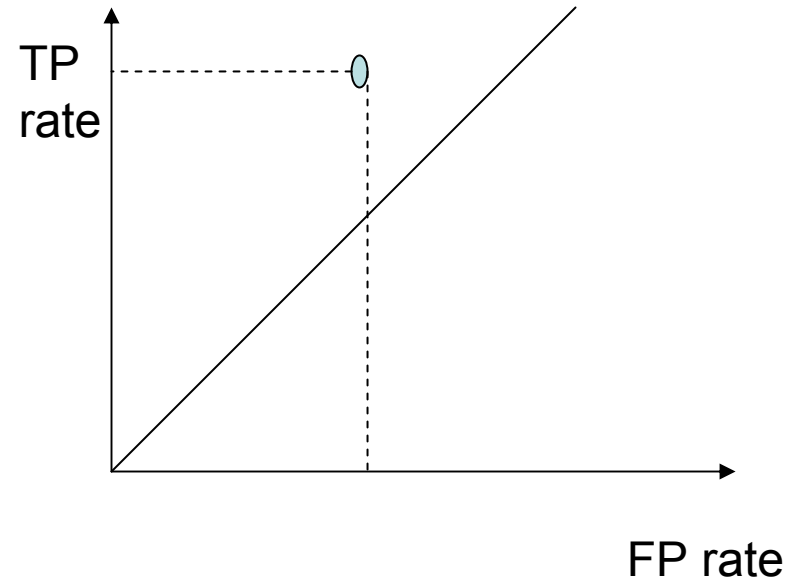
Data Mining Lab 5

Lab outline

- Remind what ROC curve is
- Generate ROC curves using WEKA
- Some usage of ROC curves

Point in ROC space

		TRUE CLASS	
		YES	NO
PREDICTED CLASS	YES	TP	FP
	NO	FN	TN
Total:		P	N



FP rate: FP/N TP rate: TP/P (recall)

FN rate: FN/N TN rate: TN/P

Classifier accuracy: $(TP+TN)/(P+N)$

Shows how good is classifier in discriminating positive instances from the negative ones

ROC curve of a probabilistic classifier

Naïve Bayes, for example, outputs the probability of an instance in a testing set to be classified as YES

Outlook	Temp	Windy	P(Y E)	Real class
overcast	mild	yes	0.95	YES
rainy	mild	no	0.80	YES
rainy	cool	yes	0.60	NO
sunny	mild	no	0.45	YES
sunny	cool	no	0.40	NO
sunny	hot	no	0.35	NO
sunny	hot	yes	0.25	NO

ROC curve of a probabilistic classifier

In a general case, we classify an instance as YES if the probability is more than 50%

Outlook	Temp	Windy	P(Y E)	Real class
overcast	mild	yes	0.95	YES
rainy	mild	no	0.80	YES
rainy	cool	yes	0.60	NO
sunny	mild	no	0.45	YES
sunny	cool	no	0.40	NO
sunny	hot	no	0.35	NO
sunny	hot	yes	0.25	NO

Classified as YES

Classified as NO

Operating threshold

ROC curve of a probabilistic classifier

We compute the confusion matrix

		TRUE CLASS	
		YES	NO
PREDICTED CLASS	YES	2 (TP)	1 (FP)
	NO	1 (FN)	3 (TN)
Total:		3 (P)	4 (N)

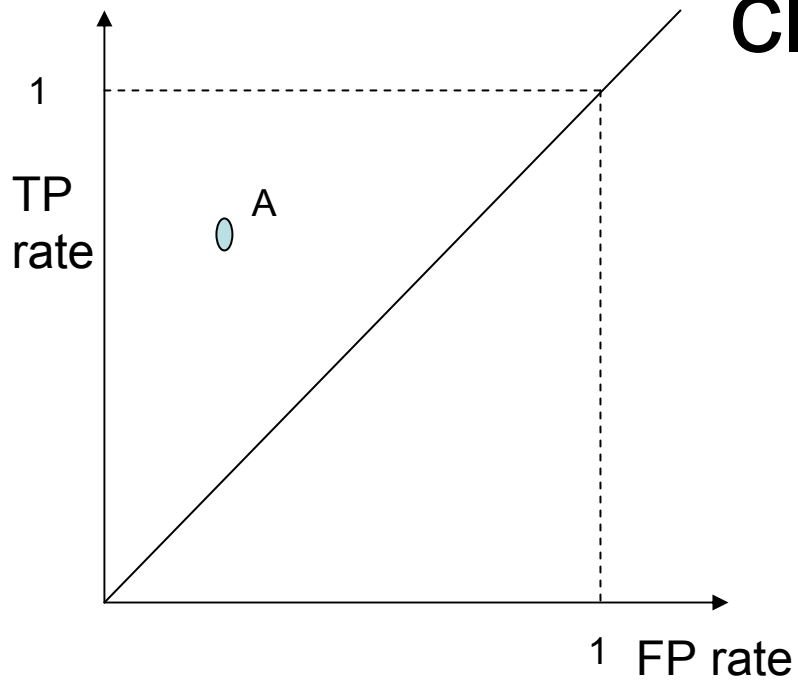
And the TP and FP rates:

TP rate: $TP/P=2/3\approx 0.7$

FP rate: $FP/N=1/4=0.25$

Outlook	Temp	Windy	P(Y E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	NO	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

ROC curve of a probabilistic classifier



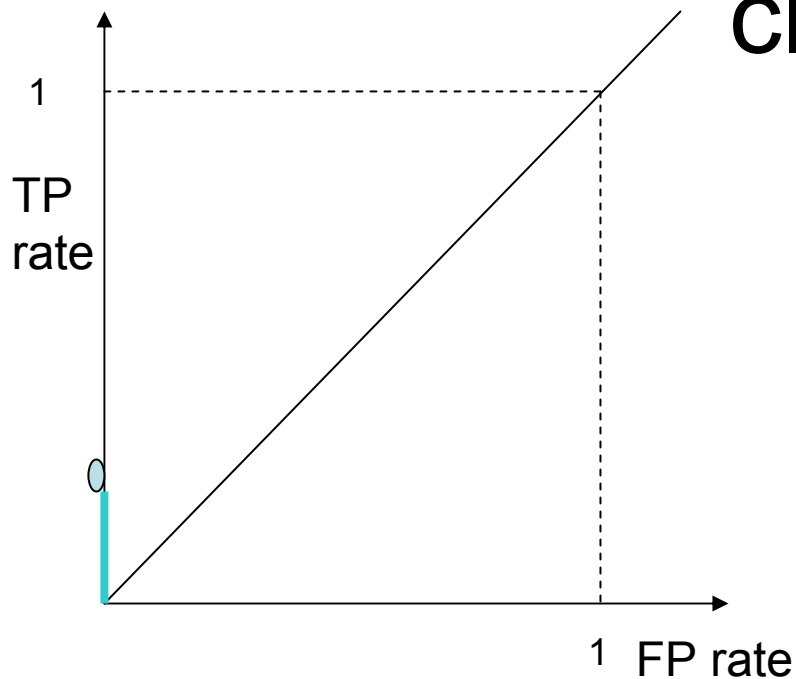
Outlook	Temp	Windy	P(Y E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	NO	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

This corresponds to point A in a ROC space

FP rate: $FP/N=1/4=0.25$

TP rate: $TP/P=2/3\approx 0.7$

ROC curve of a probabilistic classifier



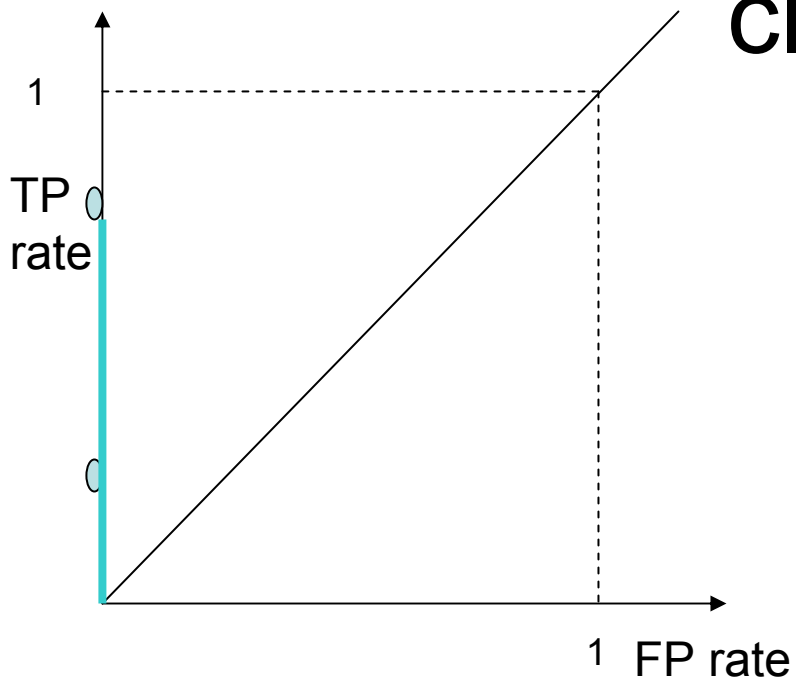
For different threshold values we get different points in the ROC space

Outlook	Temp	Windy	P(Y E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	NO	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

FP rate: $FP/N=0/4=0$

TP rate: $TP/P=1/3\approx 0.3$

ROC curve of a probabilistic classifier



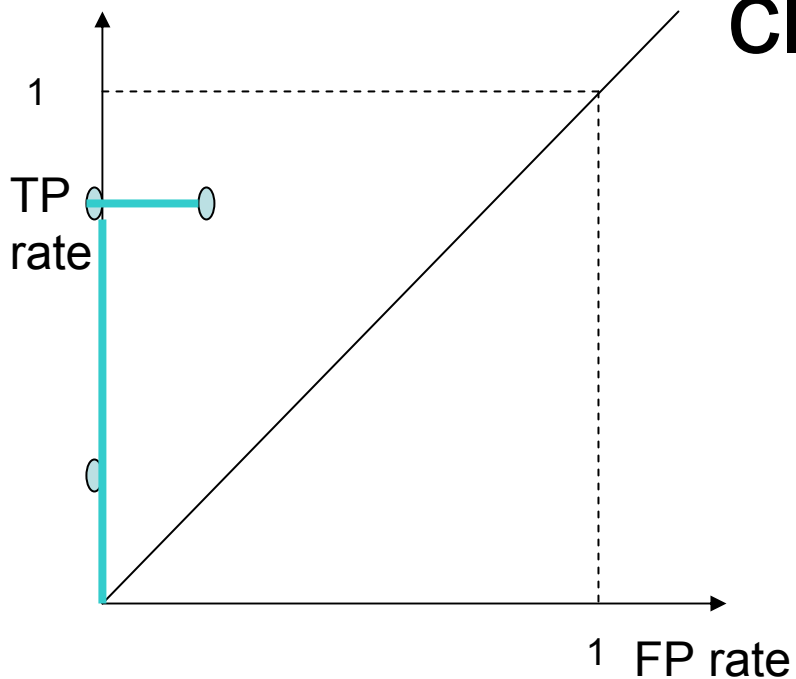
For different threshold values we get different points in the ROC space

Outlook	Temp	Windy	P(Y E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	NO	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

FP rate: $FP/N=0/4=0$

TP rate: $TP/P=2/3\approx 0.7$

ROC curve of a probabilistic classifier



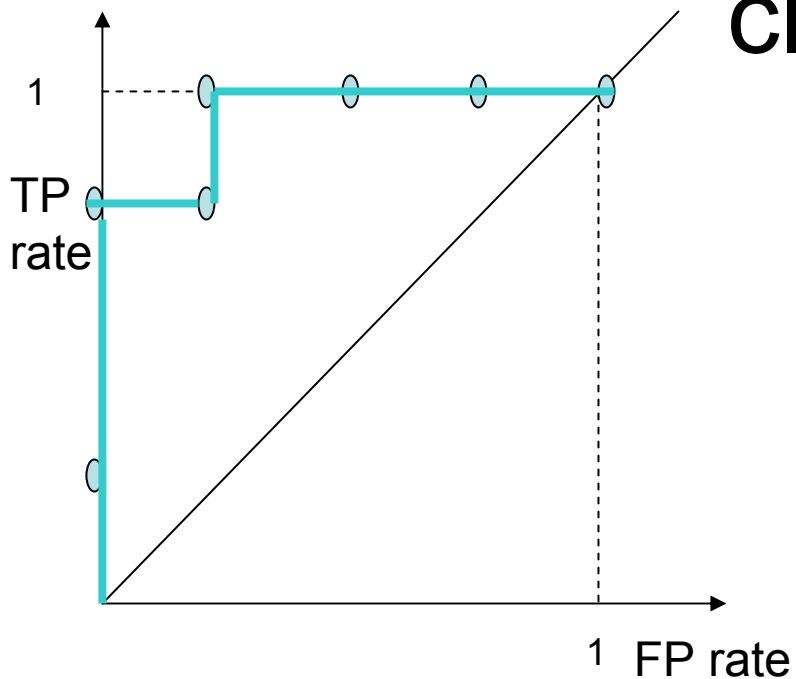
For different threshold values we get different points in the ROC space

Outlook	Temp	Windy	P(Y E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	NO	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

FP rate: $FP/N=1/4=0.25$

TP rate: $TP/P=2/3\approx 0.7$

ROC curve of a probabilistic classifier



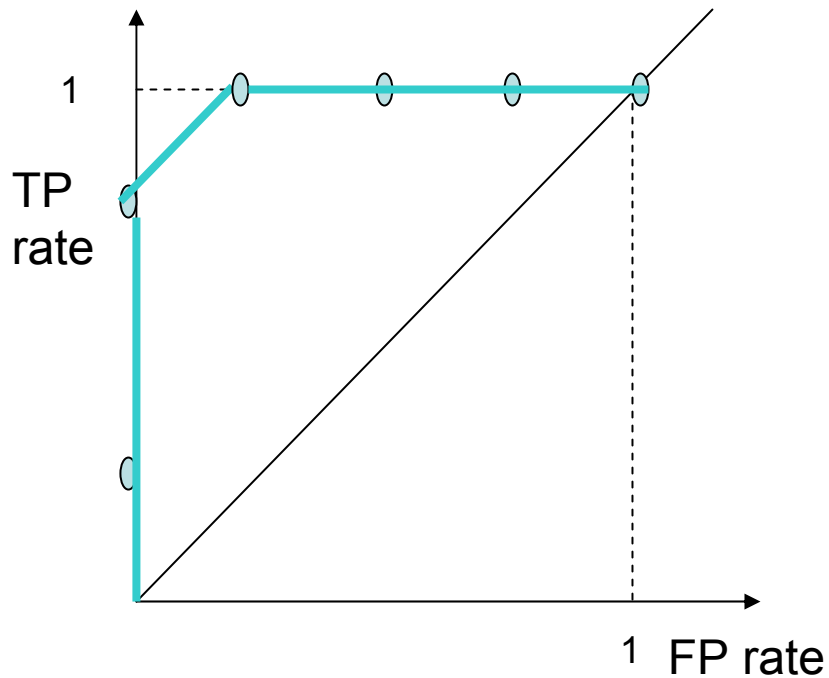
For different threshold values we get different points in the ROC space

Outlook	Temp	Windy	P(Y E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	YES	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

FP rate: $FP/N=1/4=0.25$

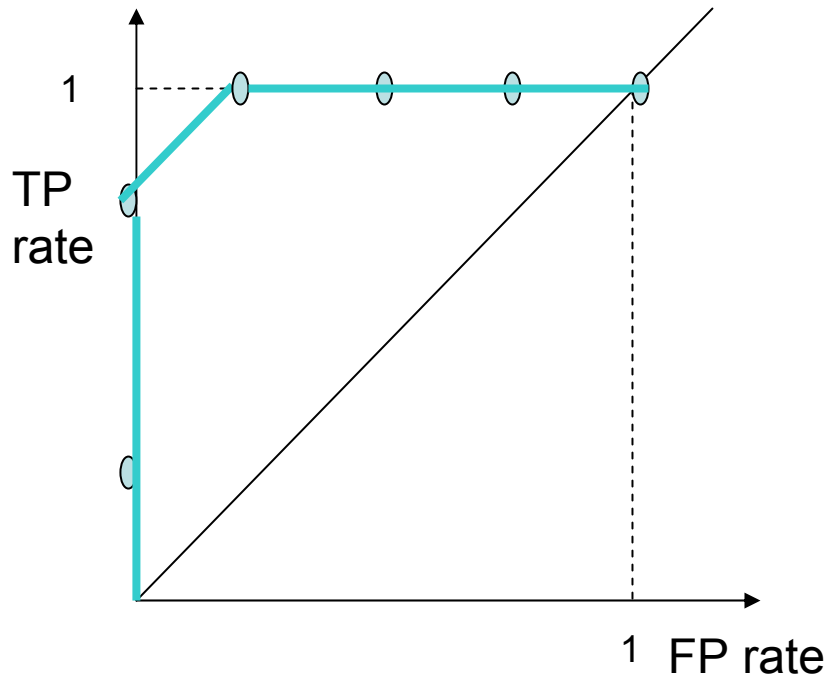
TP rate: $TP/P=3/3=1.0$, etc...

ROC curve of a probabilistic classifier

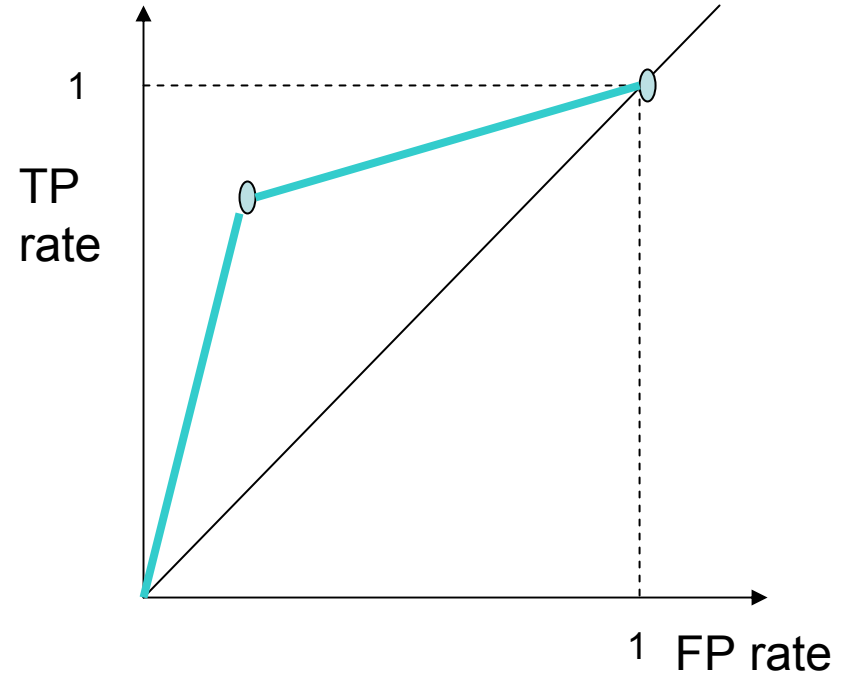


At the end we get the ROC curve for Naïve Bayes classifier

ROC curve of a probabilistic classifier vs discrete classifier



ROC curve for Naïve Bayes classifier
(probabilistic)



ROC curve for Decision Tree classifier
(discrete)

Lab outline

- Remind what ROC curve is
- Generate ROC curves using WEKA
- Some usage of ROC curves

Preparation

Step 1. Increase Java heap size

```
RunWeka.ini - Notepad
File Edit Format View Help
# Contains the commands for running weka either with a comma
# ("cmd_console") or without the command prompt ("cmd_default")
# One can also define custom commands, which can be used with
# launcher "Runweka.class". E.g., to run the launcher with a
# "custom1", you only need to specify a key "cmd_custom1" with
# command specification.
#
# Notes:
# - This file is not a DOS ini file, but a Java properties file
# - The settings listed here are key-value pairs, separated by
#   '='. A key can only be listed ONCE.
#
# Author: FracPete (fracpete at waikato dot ac dot nz)
# Version: $Revision: 1.3 $
#
# setups (prefixed with "cmd_")
cmd_default=java -Dfile.encoding=#fileEncoding# -Xmx#maxheap
cmd_console=cmd.exe /K start cmd.exe /K "java -Dfile.encoding=#fileEncoding# -Xmx#maxheap"
cmd_explorer=java -Dfile.encoding=#fileEncoding# -Xmx#maxheap
#
# placeholders ("#bla#" in command gets replaced with content
# Note: "#wekajar#" gets replaced by the launcher class, specified
# as parameter
maxheap=512m
# The MDI GUI
mainclass=weka.gui.Main
# The GUI Chooser
mainclass=weka.gui.GUIChooser
# The file encoding: use "utf-8" instead of "Cn1252" to display
```

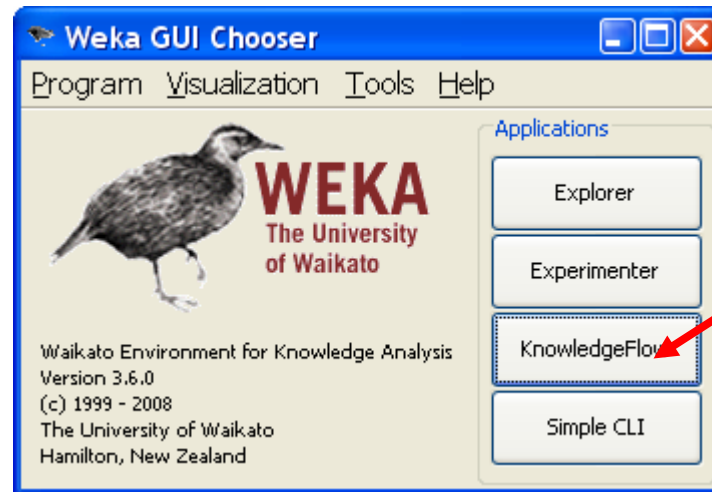
Step 2. Download input data file

adult_income.arff

into your home directory

Comparing classifiers.

Knowledge flow



Knowledge flow tabs

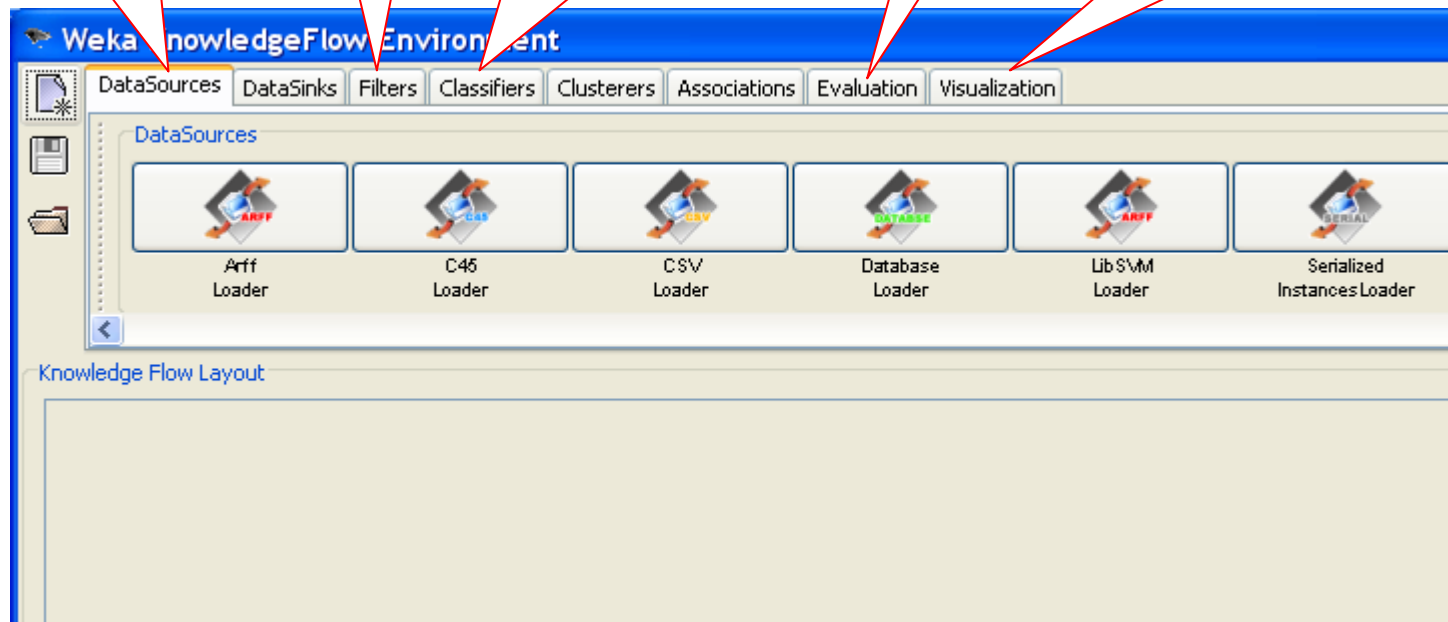
DATA
SOURCES

FILTERS

CLASSIFIERS

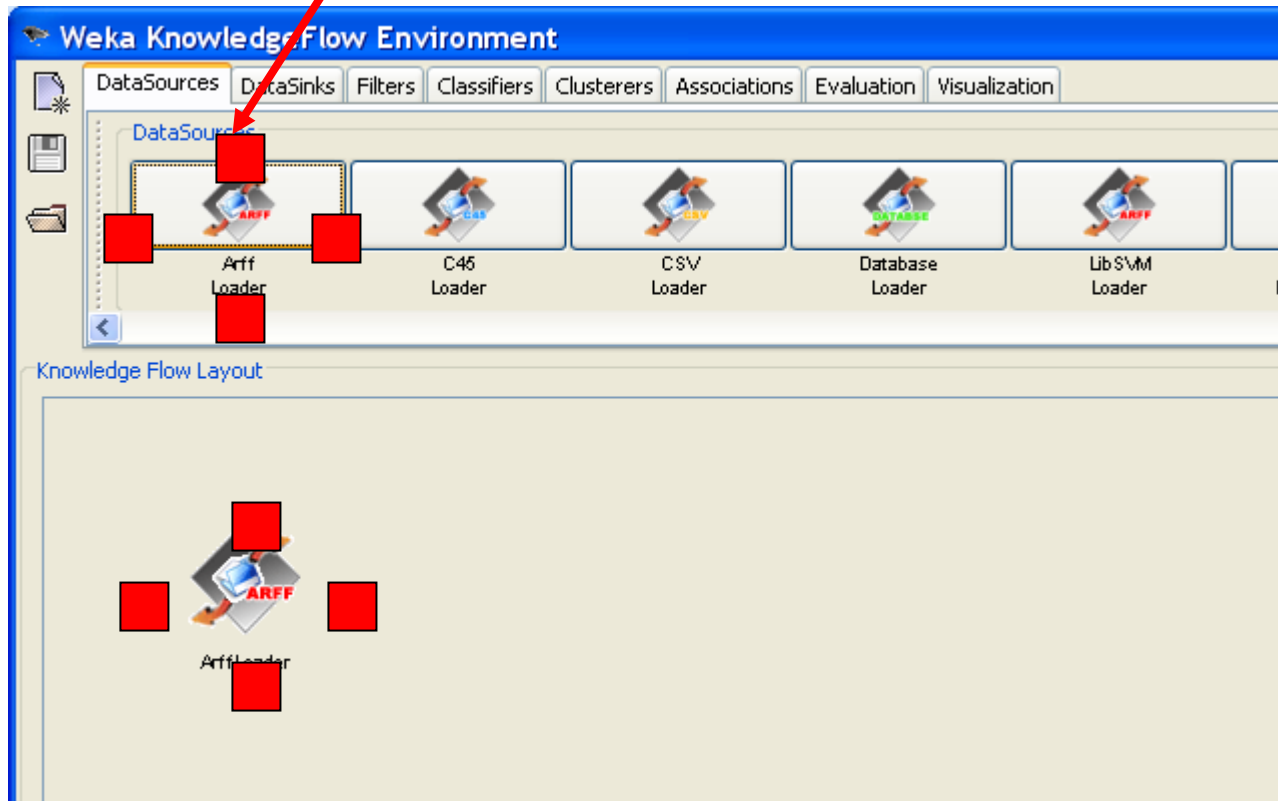
EVALUATION

VISUALIZATION

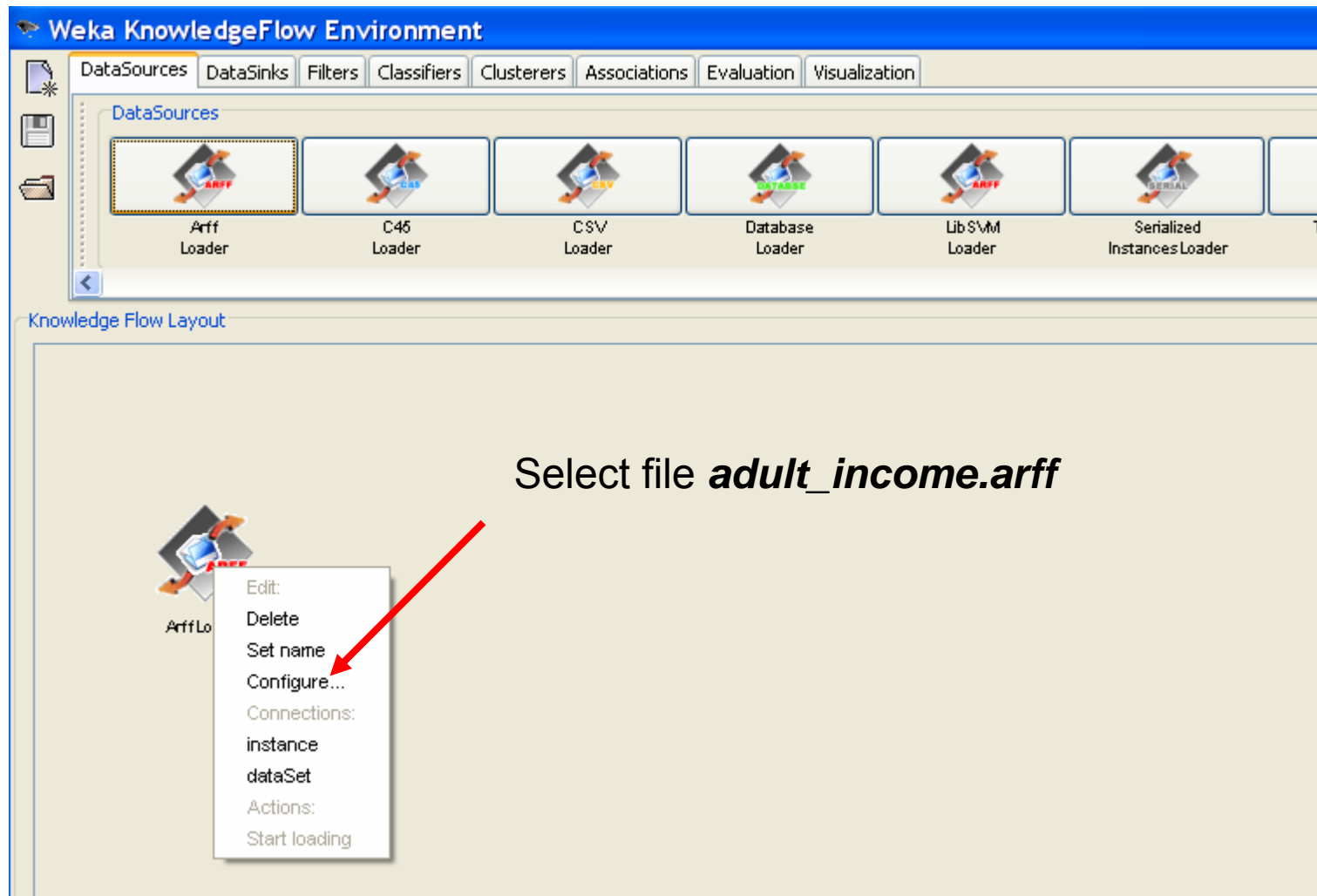


Loading the data

Click



Loading the data



Data file *adult_income.arff*

@relation adults

1. @attribute age numeric
2. @attribute workclass {Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked}
3. @attribute education real
4. @attribute marital_status {Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse}
5. @attribute occupation {Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces}
6. @attribute sex {Male, Female}
7. @attribute native_country {United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands}
8. @attribute class {>50K, <=50K}

Data from US census

Attributes of interest:

age, education,
class (income >50 K: YES,NO)

1. @attribute age numeric
3. @attribute education real
- last @attribute class {>50K, <=50K}

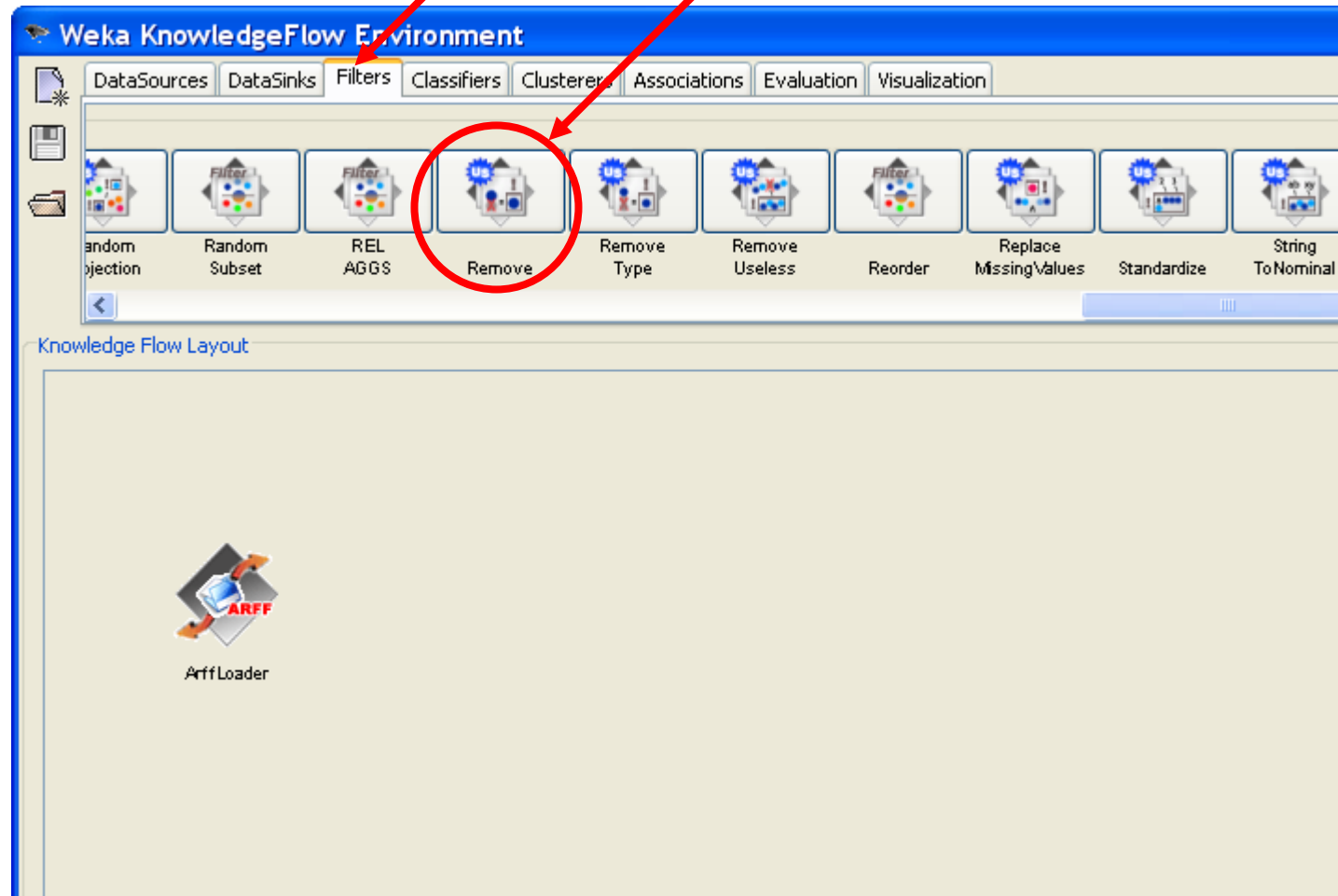
We remove all other attributes and leave only attributes 1,3, last – for simplicity

We build a classifier, which predicts income based on age and education level.

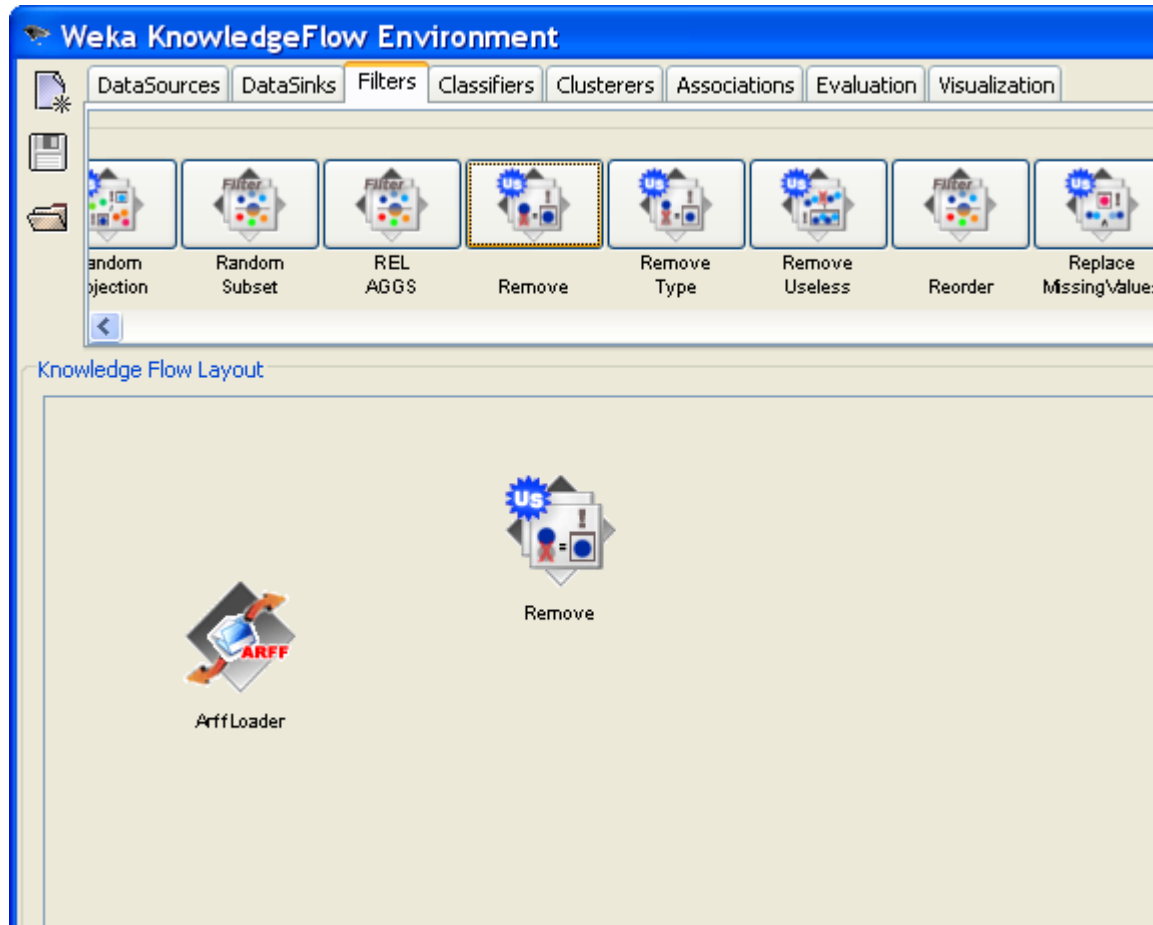
Numeric codes for education levels

Doctorate, 16
Prof-school, 15
Masters, 14
Bachelors, 13
Assoc-acdm, 12
Assoc-voc, 11
Some-college, 10
HS-grad, 9
11th, -7
12th, 8
10th, 6
9th, 5
7th-8th, 4
5th-6th, 3
1st-4th, 2
Preschool, 1

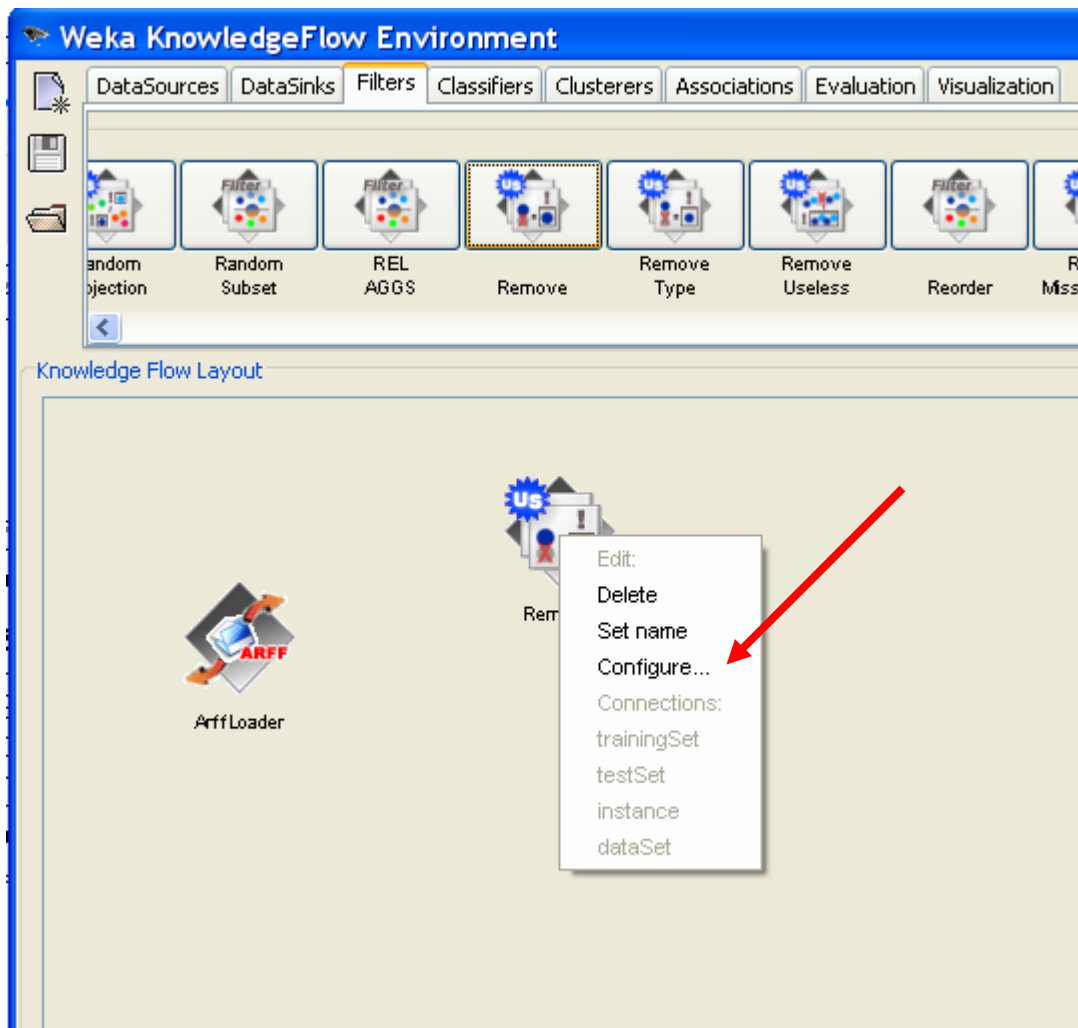
Removing attributes



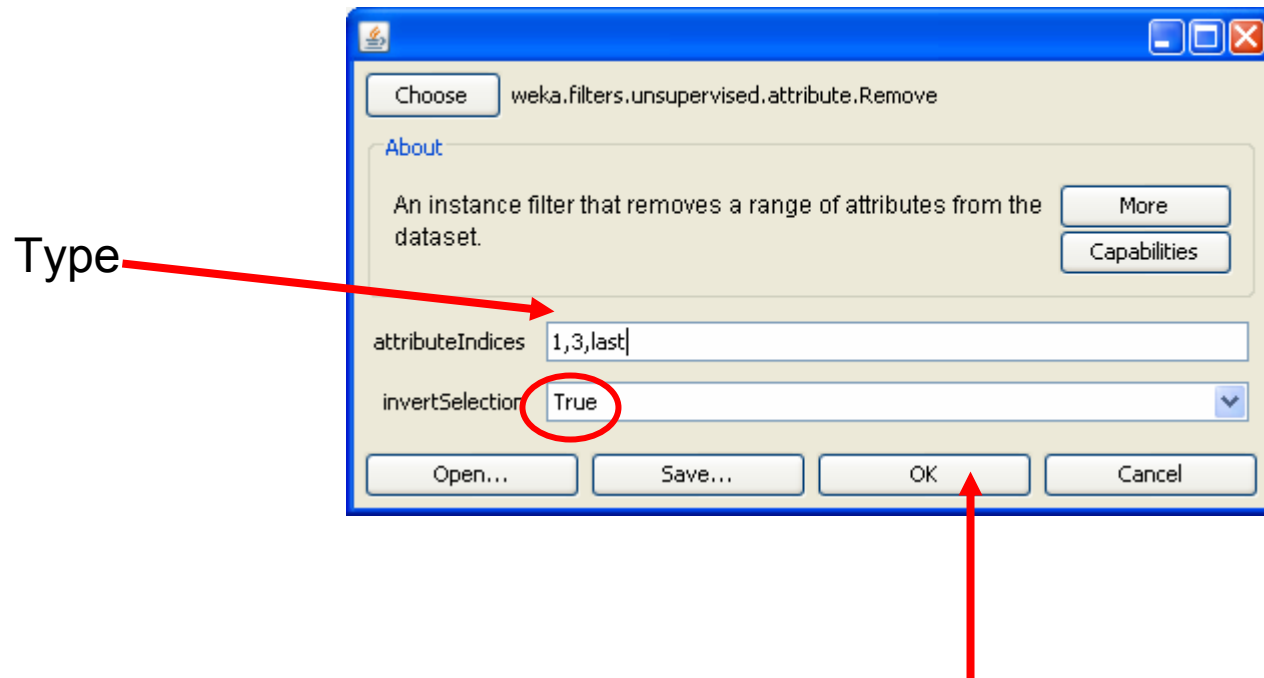
Removing attributes



Removing attributes

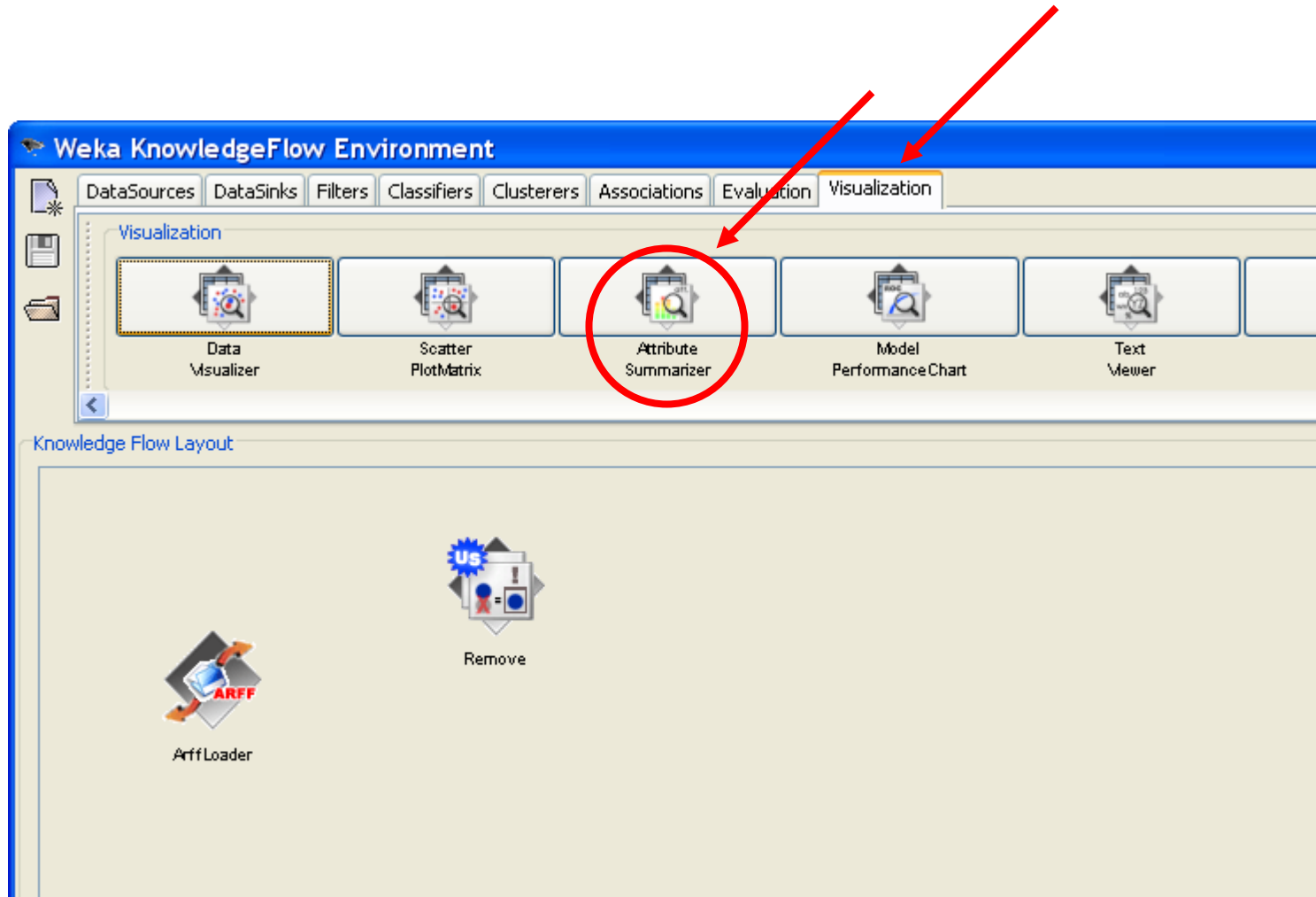


Removing attributes

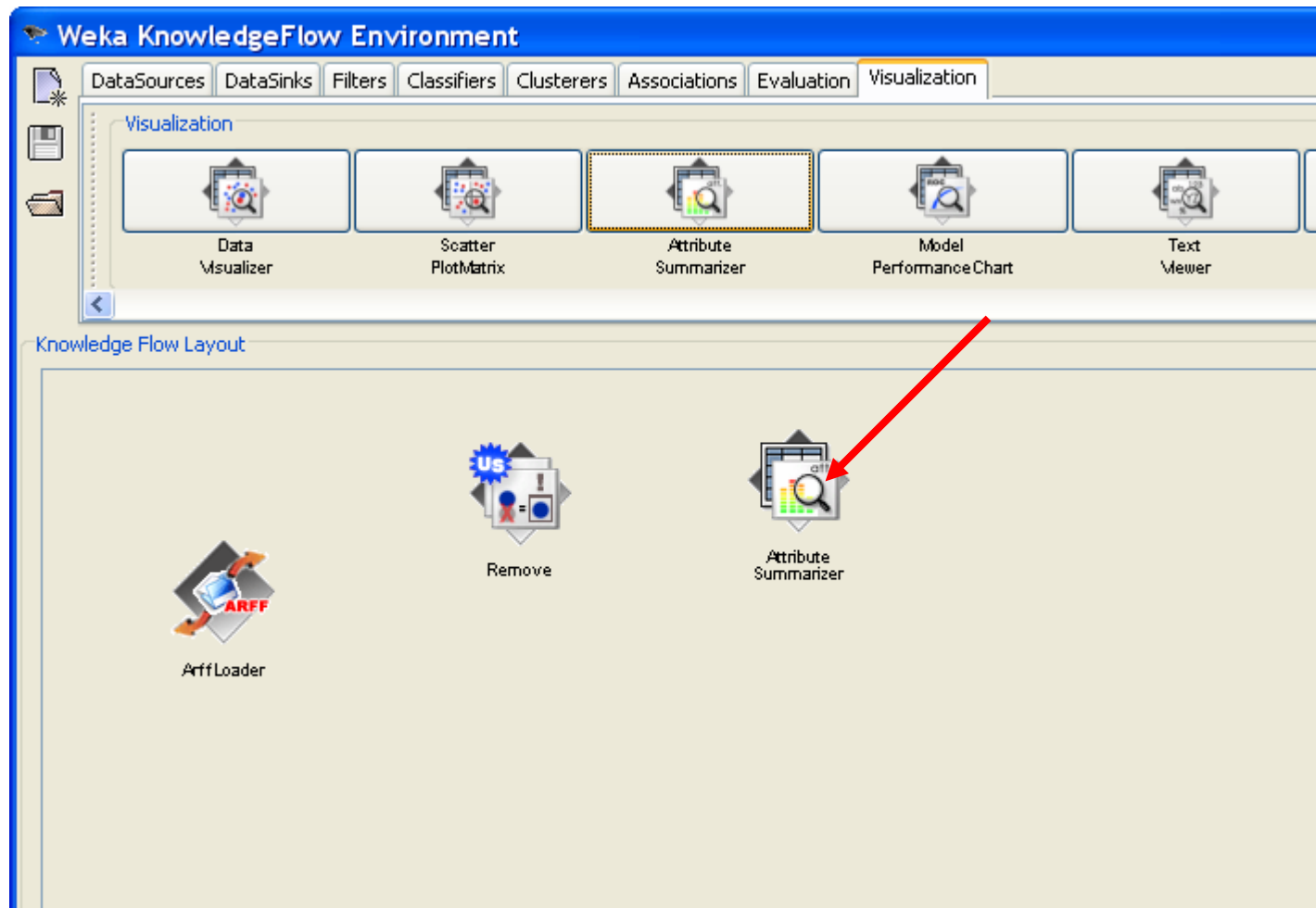


Means remove all except attributes 1,3,last

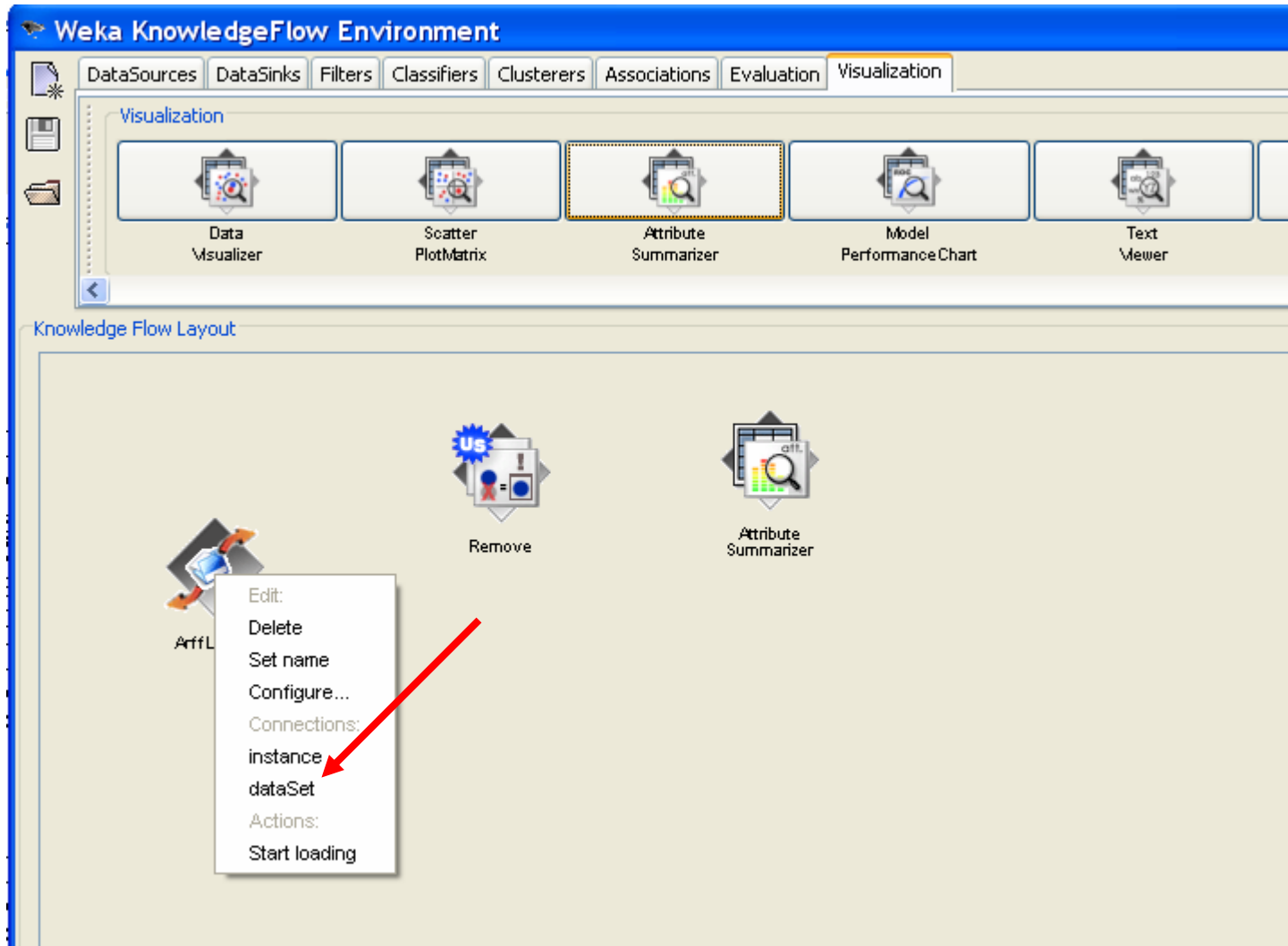
Visualize data



Visualize data



Connect the flow



Connect the flow:

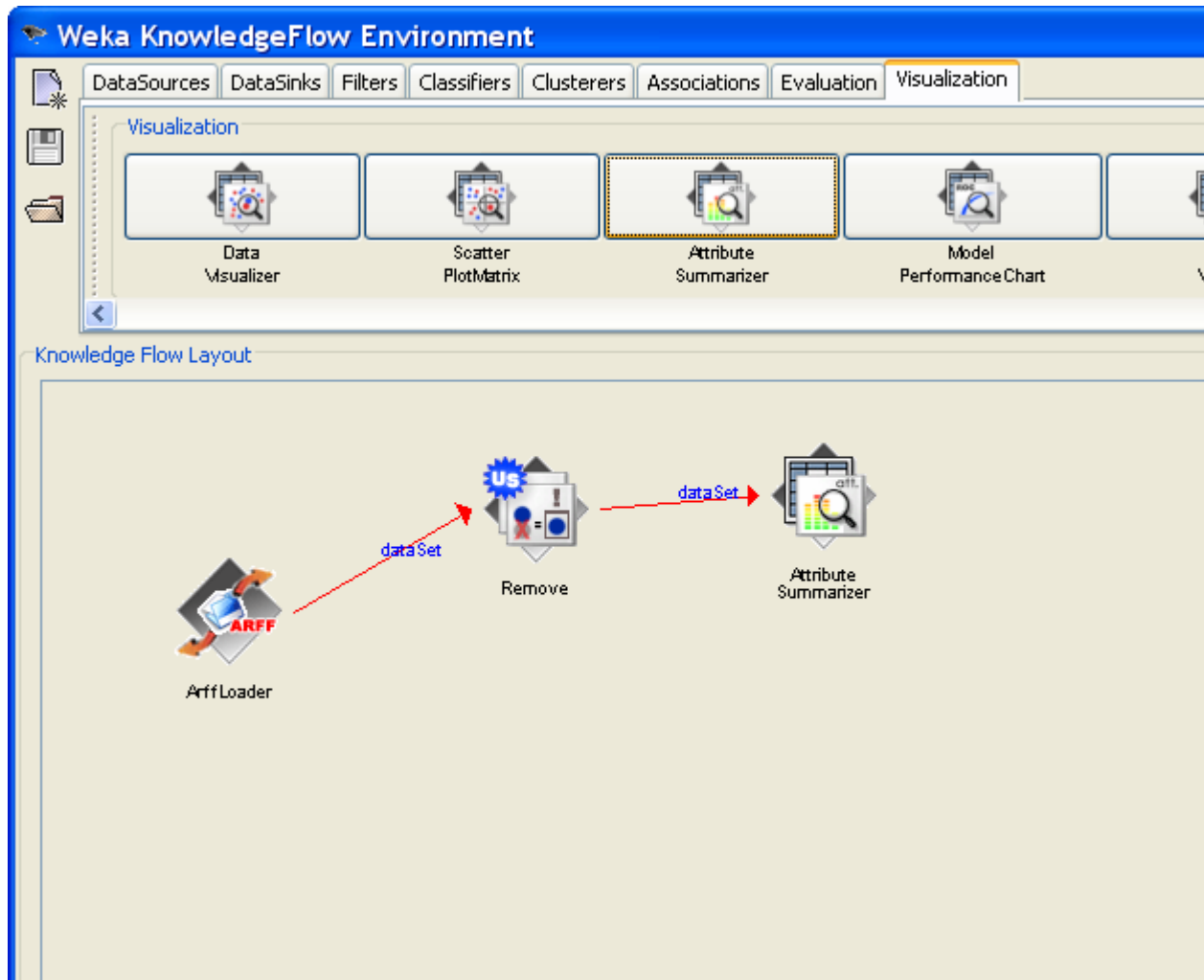
from data loader to attribute remover

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, a blue header bar contains the text "Weka KnowledgeFlow Environment". Below this, a horizontal menu bar includes tabs for "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Associations", "Evaluation", and "Visualization". The "Visualization" tab is currently selected. Underneath the menu, a "Visualization" panel shows several tool icons: "Data Visualizer", "Scatter PlotMatrix", "Attribute Summarizer" (which is highlighted with a dashed border), "Model Performance Chart", and "Text Viewer".

The main workspace, titled "Knowledge Flow Layout", contains a workflow diagram. On the left, an "Arff Loader" icon is connected by a red arrow labeled "data Set" to a "Remove" icon. The "Remove" icon features a blue starburst with the letters "US" and a red 'X' over a blue circle. To the right of the "Remove" icon is an "Attribute Summarizer" icon, which is not yet connected to the flow.

Connect the flow:

from attribute remover to summarizer



Start data flow

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, a blue title bar reads "Weka KnowledgeFlow Environment". Below it, a horizontal menu contains tabs for "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Associations", "Evaluation", and "Visualization". The "Visualization" tab is currently selected. Below the menu, a "Visualization" panel shows several tool icons: "Data Visualizer", "Scatter PlotMatrix", "Attribute Summarizer" (which is highlighted with a dashed border), and "Model Performance Chart".

The main area, titled "Knowledge Flow Layout", contains a workflow diagram. It starts with an "ArffL" node on the left. A red arrow labeled "data Set" points from "ArffL" to a "Remove" node. Another red arrow labeled "data Set" points from "Remove" to an "Attribute Summarizer" node. A context menu is open over the "ArffL" node, listing options: "Edit:", "Delete", "Set name", "Configure...", "Connections:", "instance", "dataSet", "Actions:", and "Start loading". A red arrow points from the "Start loading" option to the right.

Visualize the data

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, a blue header bar contains the text "Weka KnowledgeFlow Environment". Below this, a series of tabs are visible: "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Associations", "Evaluation", and "Visualization". The "Visualization" tab is currently selected and highlighted.

Under the "Visualization" tab, a row of five tool icons is shown, each with a label below it: "Data Visualizer", "Scatter PlotMatrix", "Attribute Summarizer", "Model Performance Chart", and "Text Viewer". The "Attribute Summarizer" icon is highlighted with a dashed border.

Below the tool icons, the "Knowledge Flow Layout" area is visible. It contains a workflow diagram with three nodes connected by red arrows labeled "data Set":

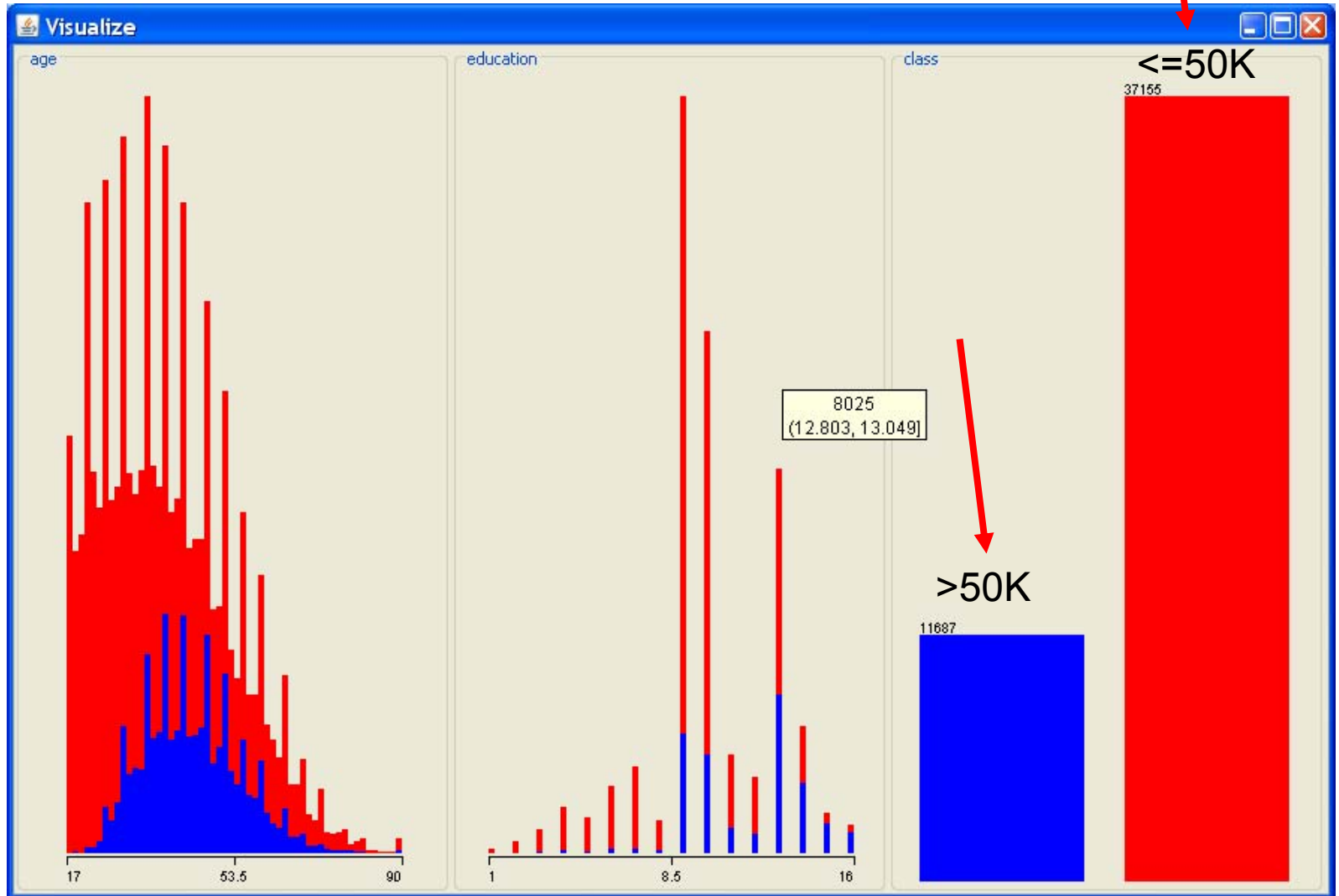
- An "ArffLoader" node on the left.
- A "Remove" node in the middle, which has a blue starburst icon with the letters "US" above it.
- An "Attribute Summarizer" node on the right, which has a magnifying glass icon over a bar chart.

A context menu is open over the "Attribute Summarizer" node. The menu items are:

- Edit:
- Delete
- Actions:
- Show summaries

A red arrow points from the right side of the image towards the "Show summaries" option in the context menu.

Visualize the data



Assigning the class

The screenshot displays the Weka KnowledgeFlow Environment interface. The top menu bar includes 'DataSources', 'DataSinks', 'Filters', 'Classifiers', 'Clusters', 'Associations', 'Evaluation', and 'Visualization'. The 'Evaluation' tab is selected, showing a toolbar with various tools: Training SetMaker, Test Set Maker, Cross Validation FoldMaker, Train Test SplitMaker, Instance Stream To BatchMaker, Class Assigner, Class Value Picker, and Classifier Performance Evaluator. The 'Class Assigner' tool is highlighted with a red circle. A red arrow points from the title 'Assigning the class' to this tool.

Below the toolbar is the 'Knowledge Flow Layout' area, which contains a workflow diagram. The workflow starts with an 'ArffLoader' node, which outputs 'data' to a 'Remove' node. The 'Remove' node outputs 'data Set' to two other nodes: 'Attribute Summarizer' and 'ClassAssigner'. The 'ClassAssigner' node is also highlighted with a red circle.

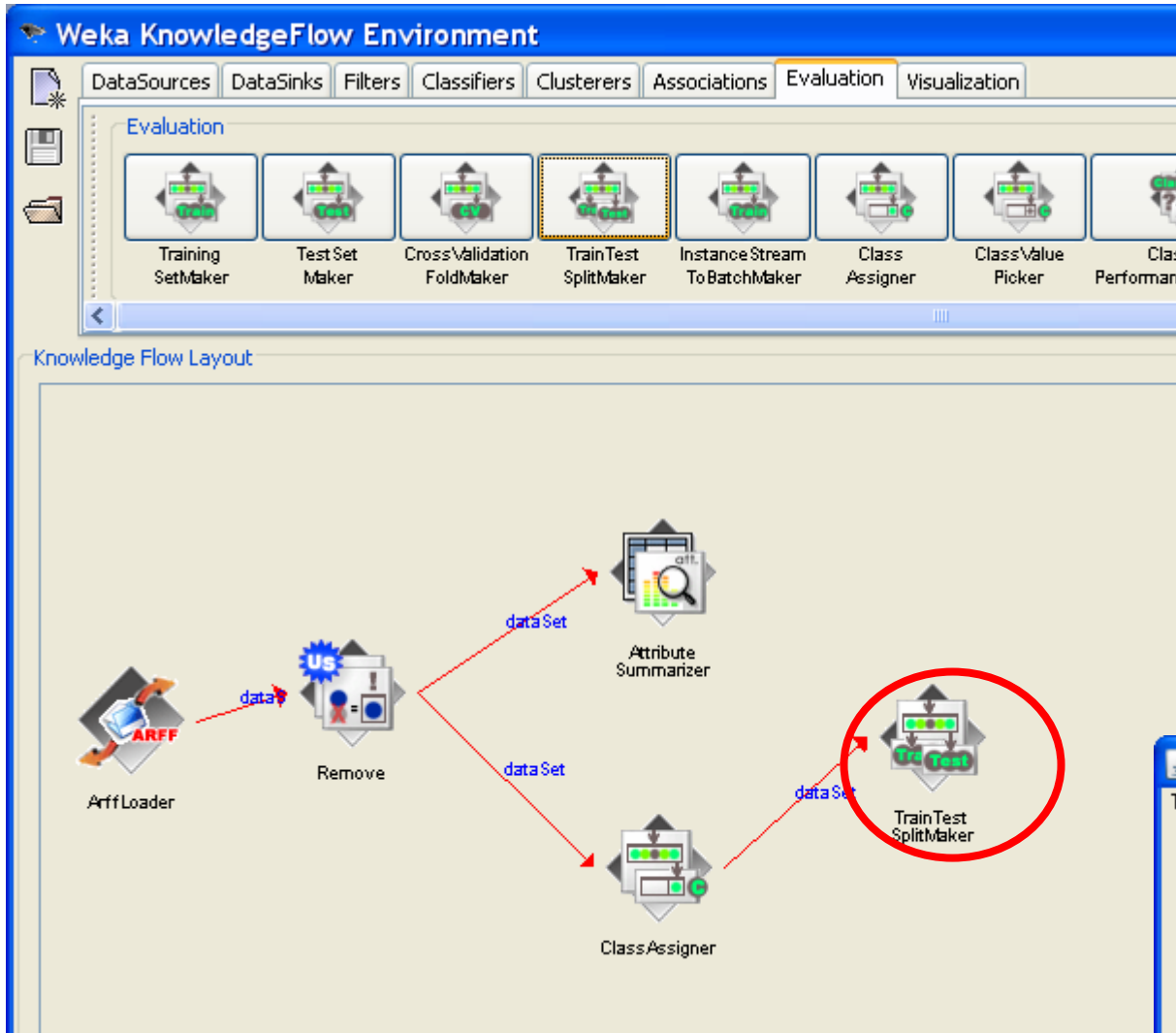
Configuring class assigner

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, there are tabs for DataSources, DataSinks, Filters, Classifiers, Clusters, Associations, Evaluation, and Visualization. The Evaluation tab is active, showing a sequence of icons representing a workflow. A dialog box titled "ClassAssignerCustomizer" is open, with the "About" tab selected. The dialog contains the text: "Designate which column is to be considered the class column in incoming data." Below this text is a text input field labeled "classColumn" with the value "last" entered. A red circle highlights the "classColumn" field. In the background, a workflow is visible with the following components: ArffLoader (outputting "data") -> Remove (outputting "data Set") -> Attribute Summarizer (outputting "data Set") -> ClassAssigner (outputting "data Set").

Subdivision of the dataset into “learning” and “test” set

The screenshot displays the Weka KnowledgeFlow Environment interface. The top menu bar includes tabs for DataSources, DataSinks, Filters, Classifiers, Clusters, Associations, Evaluation, and Visualization. The Evaluation tab is selected, and a red arrow points to it from the title above. Below the menu, a toolbar contains various evaluation-related icons: Training SetMaker, Test Set Maker, CrossValidation FoldMaker, Train Test SplitMaker (circled in red), Instance Stream To BatchMaker, Class Assigner, Class Value Picker, Classifier, and Incremental ClassifierEvaluator. The main workspace, titled "Knowledge Flow Layout", shows a workflow diagram. It starts with an ArffLoader icon connected to a Remove icon. The Remove icon outputs "data" to the ArffLoader and "data Set" to both an Attribute Summarizer and a ClassAssigner. The ClassAssigner outputs "data Set" to a Train Test SplitMaker icon.

Subdivision of the dataset into “learning” and “test” set



We want to build our prediction model on the 70% of the whole dataset, and compute the ROC curve on the remaining.

So, we set the TRAINTEST SPLIT MAKER (EVALUATION) in the diagram and configure its parameters.

TrainTestSplitMakerCustomizer

About

Split an incoming data set into separate train and test sets.

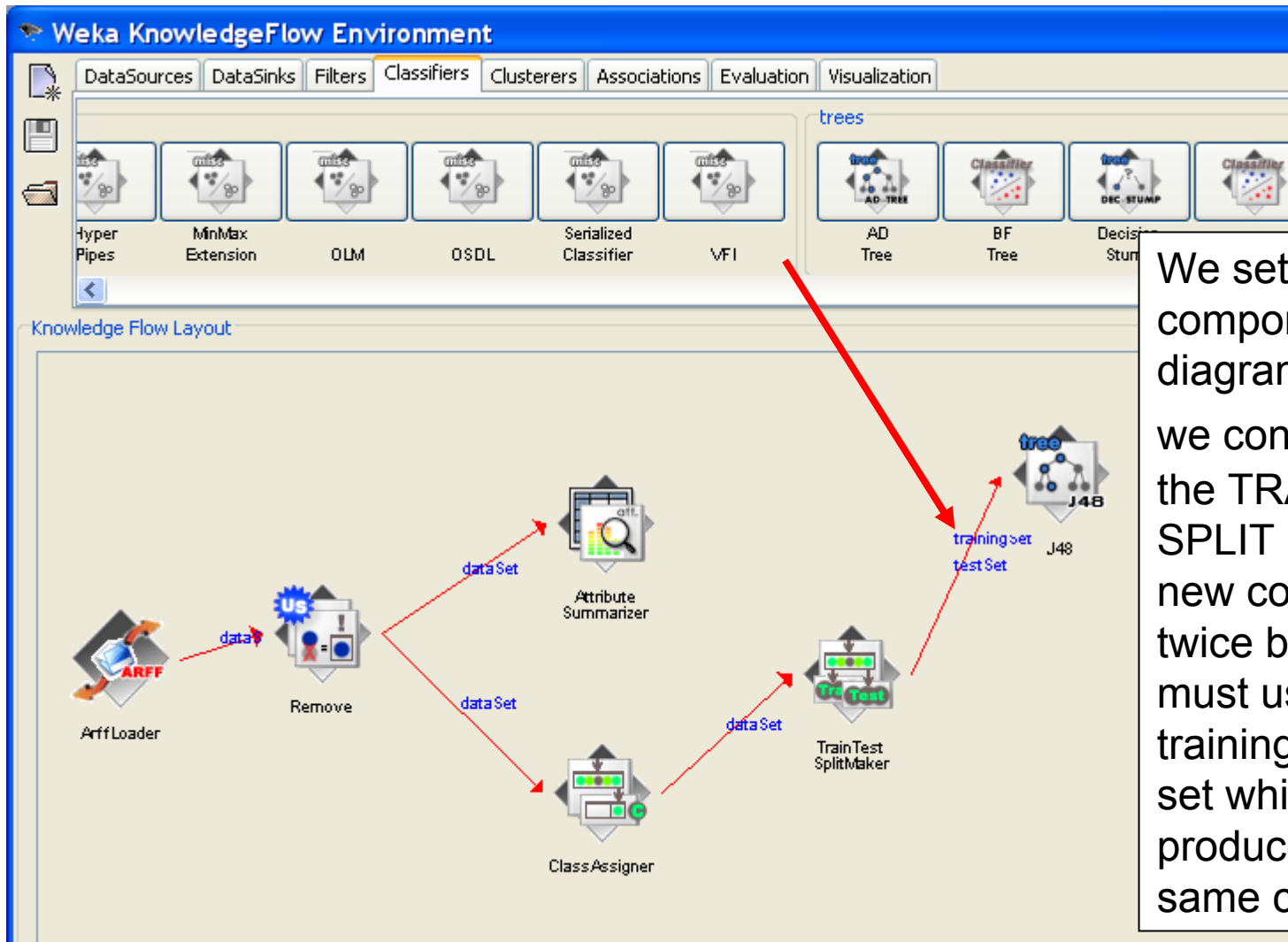
seed

trainPercent

Choosing discrete classifier – decision tree

The screenshot displays the Weka KnowledgeFlow Environment interface. The top menu bar includes 'DataSources', 'DataSinks', 'Filters', 'Classifiers', 'Clusters', 'Associations', 'Evaluation', and 'Visualization'. The 'Classifiers' tab is active, showing a toolbar with various classifier icons. A red arrow points from the title 'Choosing discrete classifier – decision tree' to the 'Classifiers' tab. Another red arrow points to the 'trees' sub-tab, which contains icons for 'AD Tree', 'BF Tree', 'Decision Stump', 'FT', 'Id3', and 'J48'. The 'J48' icon is circled in red. Below the toolbar, the 'Knowledge Flow Layout' area shows a workflow diagram. The workflow starts with an 'ArffLoader' node, followed by a 'Remove' node. From 'Remove', two 'data Set' arrows lead to 'Attribute Summarizer' and 'ClassAssigner' nodes. From 'ClassAssigner', a 'data Set' arrow leads to a 'Train Test SplitMaker' node. A 'J48' classifier node is also present in the layout area.

Connecting classifier to the data



We set J48 component in the diagram, we connect **twice** the TRAIN TEST SPLIT MAKER to this new component: twice because we must use together the training and the test set which are produced by the same component.

Adding visualizer to see the classification results

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, a blue header bar contains the text "Weka KnowledgeFlow Environment". Below this, a horizontal menu bar lists several categories: DataSources, DataSinks, Filters, Classifiers, Clusterers, Associations, Evaluation, and Visualization. The "Visualization" tab is highlighted with a red circle. Below the menu bar, a "Visualization" panel contains six icons representing different visualization tools: Data Visualizer, Scatter PlotMatrix, Attribute Summarizer, Model Performance Chart, Text Viewer, and Graph Viewer. The "Graph Viewer" icon is also highlighted with a red circle. A red arrow points from the "Graph Viewer" icon in the Visualization panel to the "Graph Viewer" node in the Knowledge Flow Layout below. The Knowledge Flow Layout shows a workflow starting with an "ArffLoader" node, followed by a "Remove" node. From the "Remove" node, two "data Set" arrows branch out to "Attribute Summarizer" and "ClassAssigner" nodes. From the "ClassAssigner" node, a "data Set" arrow points to a "Train Test SplitMaker" node. From the "Train Test SplitMaker" node, two arrows branch out: one labeled "training set" points to a "tree J48" node, and another labeled "test Set" points to a "Graph Viewer" node. The "tree J48" node is also labeled "J48".

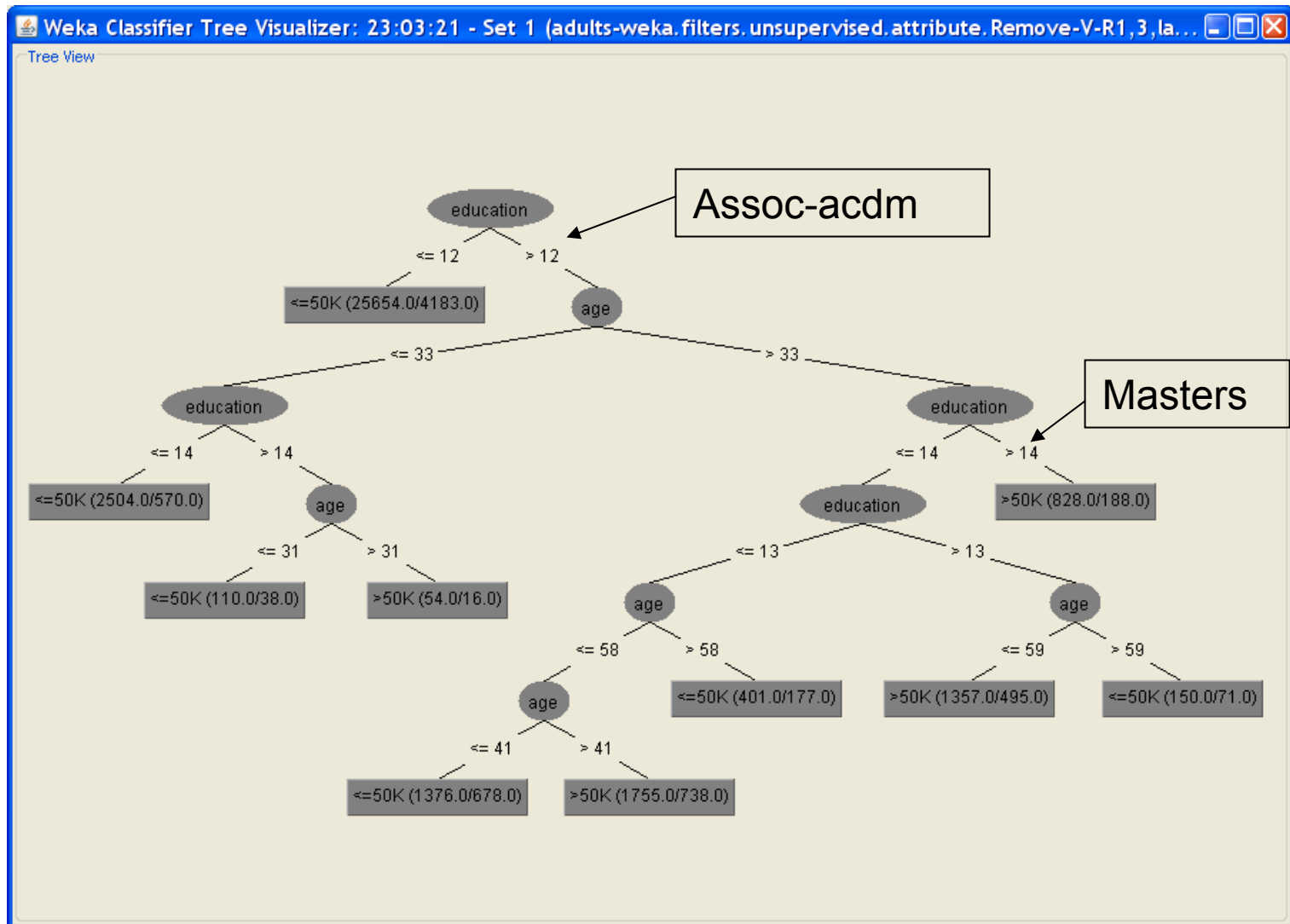
Perform classification

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, a blue header bar contains the text "Weka KnowledgeFlow Environment". Below this, a navigation bar includes tabs for "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Associations", "Evaluation", and "Visualization". The "Visualization" tab is currently selected.

Under the "Visualization" tab, a row of six visualization tools is shown: "Data Visualizer", "Scatter PlotMatrix", "Attribute Summarizer", "Model Performance Chart", "Text Viewer", and "Graph Viewer". The "Graph Viewer" tool is highlighted with a dashed border.

The main workspace, titled "Knowledge Flow Layout", contains a workflow diagram. The workflow starts with an "Arff" data source icon on the left. A red arrow labeled "data" points from the "Arff" icon to a "ClassAssigner" icon. A context menu is open over the "ClassAssigner" icon, listing options: "Edit:", "Delete", "Set name", "Configure...", "Connections:", "instance", "dataSet", "Actions:", and "Start loading". A red arrow points from the "Start loading" option to the "ClassAssigner" icon. From the "ClassAssigner" icon, a red arrow labeled "data Set" points to an "Attribute Summarizer" icon. Another red arrow labeled "data Set" points from the "ClassAssigner" icon to a "TrainTest SplitMaker" icon. From the "TrainTest SplitMaker" icon, a red arrow labeled "trainingset" points to a "J48" classifier icon, and another red arrow labeled "test Set" points to a "J48" classifier icon. Finally, a red arrow labeled "graph" points from the "J48" classifier icon to the "GraphViewer" icon.

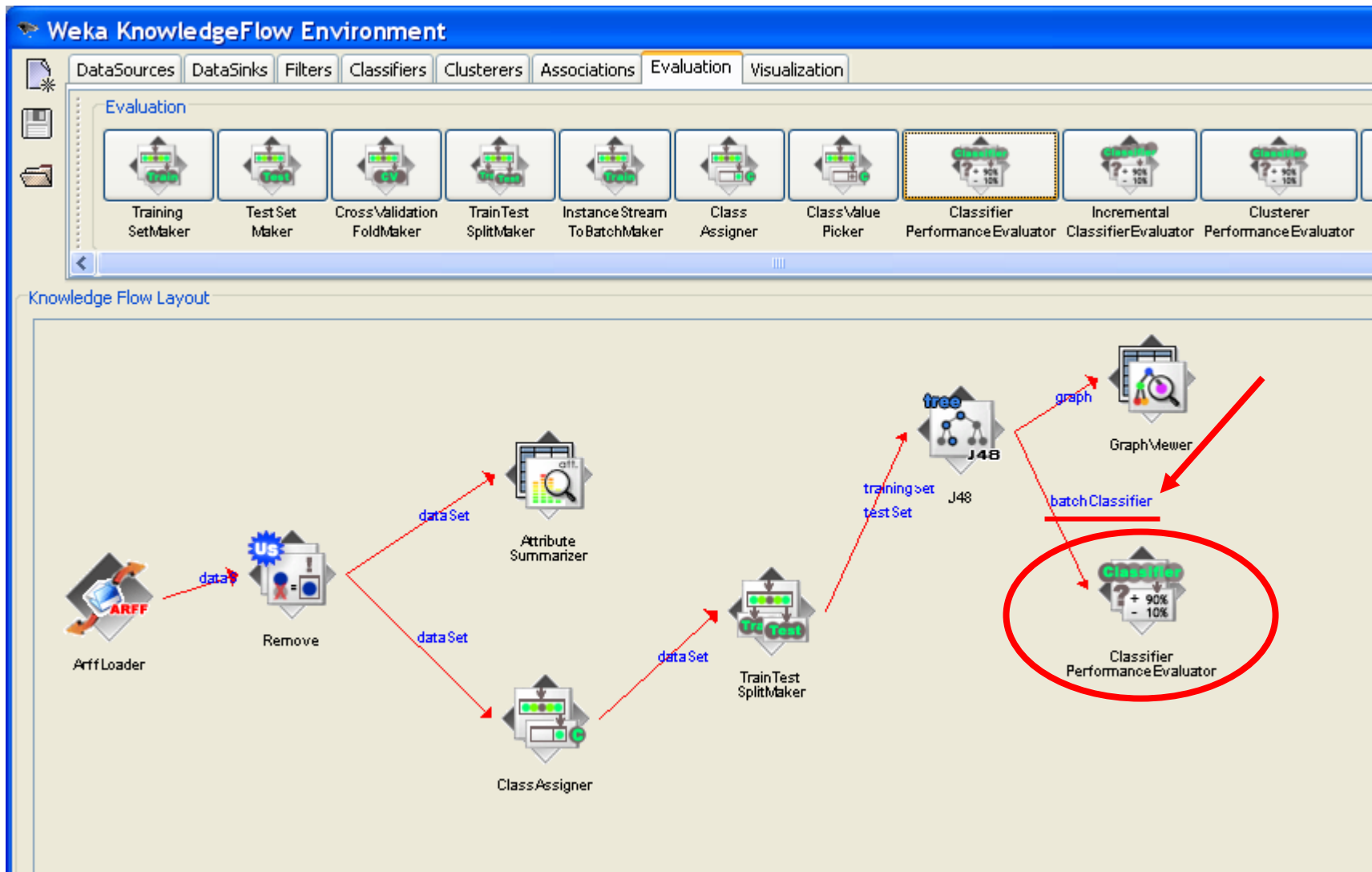
Show classification results (decision tree)



Classifier evaluation

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, the title bar reads "Weka KnowledgeFlow Environment". Below it, a menu bar includes "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Associations", "Evaluation", and "Visualization". The "Evaluation" menu is circled in red. Below the menu bar, a toolbar contains various workflow components: Training SetMaker, Test Set Maker, CrossValidation FoldMaker, Train Test SplitMaker, Instance Stream To BatchMaker, Class Assigner, Class Value Picker, Classifier Performance Evaluator (circled in red), Incremental Classifier Evaluator, and Clusterer Performance Evaluator. The main workspace, titled "Knowledge Flow Layout", shows a workflow diagram. It starts with an "ArffLoader" component connected to a "Remove" component via a "data" arrow. The "Remove" component has two outgoing "data Set" arrows: one to an "Attribute Summarizer" and another to a "ClassAssigner". The "ClassAssigner" is connected to a "Train Test SplitMaker" via a "data Set" arrow. The "Train Test SplitMaker" has two outgoing "data Set" arrows: one to a "tree J48" classifier and another to a "GraphViewer". The "tree J48" classifier is also connected to the "GraphViewer" via a "graph" arrow. The "tree J48" component is labeled with "J48" and "training set test set".

Connecting classifier to the evaluator



Selecting performance model: chart

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, a blue header bar contains the text "Weka KnowledgeFlow Environment". Below this, a series of tabs are visible: "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Associations", "Evaluation", and "Visualization". The "Visualization" tab is selected and highlighted with a red arrow pointing from the title above. Underneath the tabs, a row of visualization tool icons is shown. The "Model Performance Chart" icon is circled in red. Other icons include "Data Visualizer", "Scatter PlotMatrix", "Attribute Summarizer", "Text Viewer", "Graph Viewer", and "Strip Chart".

Below the visualization tools, the "Knowledge Flow Layout" section shows a workflow diagram. The workflow starts with an "ArffLoader" icon, which outputs "data" to a "Remove" icon. From "Remove", two "data Set" outputs lead to "Attribute Summarizer" and "ClassAssigner" icons. The "ClassAssigner" outputs "data Set" to a "Train Test SplitMaker" icon. The "Train Test SplitMaker" outputs "training set" and "test Set" to a "J48" icon. The "J48" icon outputs "graph" to a "Graph Viewer" icon and "batch Classifier" to a "Classifier Performance Evaluator" icon. The "Classifier Performance Evaluator" outputs "threshold Data" to a "Model Performance Chart" icon. A red arrow points from the "Model Performance Chart" icon in the visualization toolbar to the "Model Performance Chart" icon in the workflow diagram.

Running the model

Weka KnowledgeFlow Environment

DataSources | DataSinks | Filters | Classifiers | Clusterers | Associations | Evaluation | Visualization

Visualization

Data Visualizer | Scatter PlotMatrix | Attribute Summarizer | **Model Performance Chart** | Text Mewer | Graph Mewer | Strip Chart

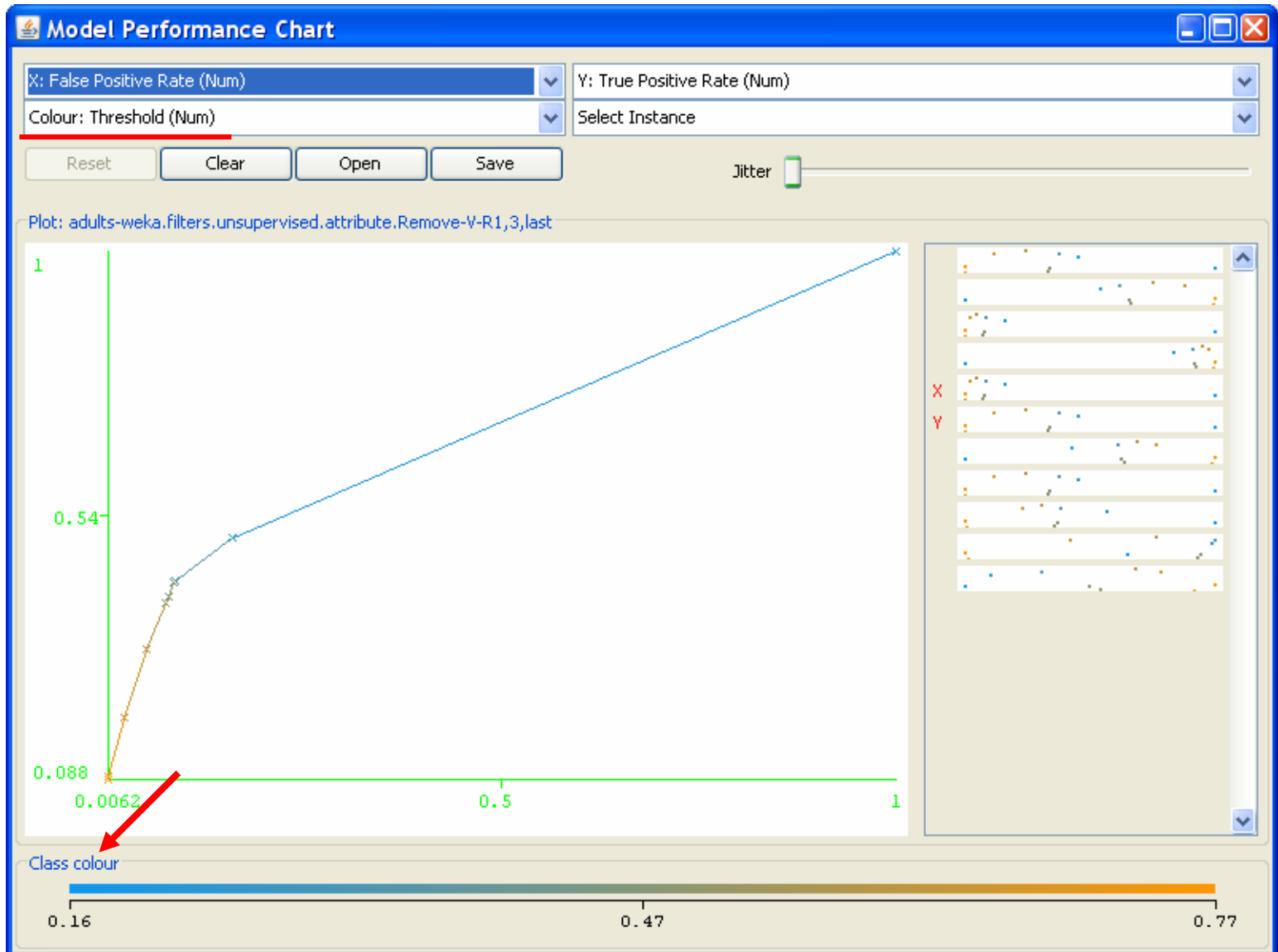
Knowledge Flow Layout

```
graph LR; AP[Affinity Propagation] -- data --> Us[Us]; Us -- dataSet --> AS[Attribute Summarizer]; Us -- dataSet --> CA[Class Assigner]; CA -- dataSet --> TTS[Train Test SplitMaker]; TTS -- training set --> J48[J48]; TTS -- test set --> J48; J48 -- graph --> GM[Graph Mewer]; J48 -- batchClassifier --> CPE[Classifier Performance Evaluator]; CPE -- threshold Data --> MPC[Model Performance Chart];
```

Context menu for 'Us' node:

- Edit:
- Delete
- Set name
- Configure...
- Connections:
- instance
- dataSet**
- Actions:
- Start loading**

View ROC curve



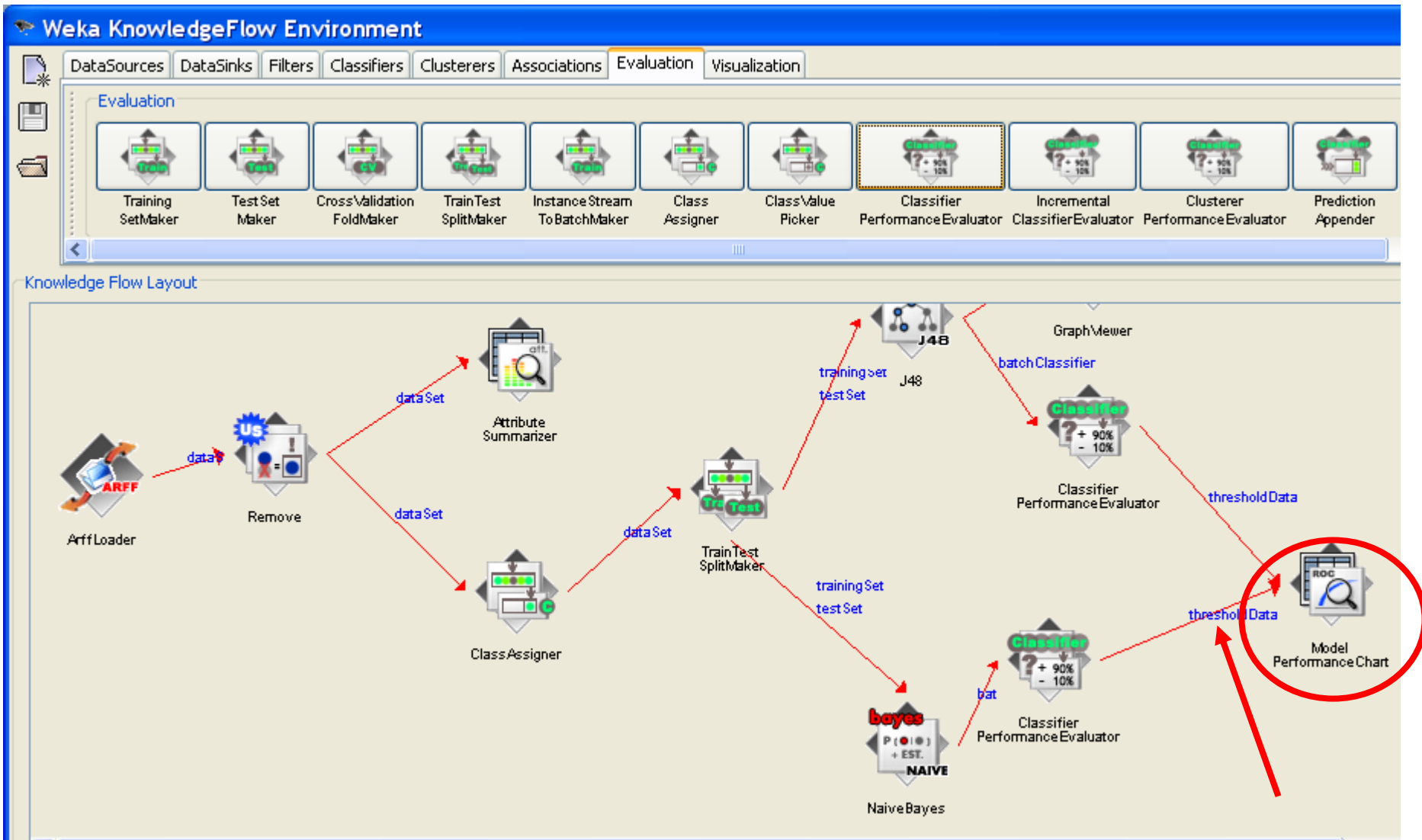
Adding Naïve Bayes classifier

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, the 'Classifiers' tab is selected, and the 'Naive Bayes' classifier icon is highlighted with a red circle. Below this, the 'Knowledge Flow Layout' shows a workflow starting with an 'ArffLoader' connected to a 'Remove' node. The 'Remove' node outputs 'data Set' to both an 'Attribute Summarizer' and a 'Class Assigner'. The 'Class Assigner' outputs 'data Set' to a 'Train Test SplitMaker'. The 'Train Test SplitMaker' outputs 'training set' and 'test Set' to a 'tree J48' node. The 'tree J48' node outputs 'graph' to a 'Graph Viewer' and 'batch Classifier' to a 'Classifier Performance Evaluator'. The 'Classifier Performance Evaluator' outputs 'threshold Data' to a 'Model Performance Chart'. The 'Naive Bayes' classifier icon is also highlighted with a red circle in the workflow layout.

Adding separate performance evaluator for Naïve Bayes classifier

The screenshot displays the Weka KnowledgeFlow Environment interface. The top menu bar includes 'DataSources', 'DataSinks', 'Filters', 'Classifiers', 'Clusterers', 'Associations', 'Evaluation', and 'Visualization'. The 'Evaluation' tab is selected, and the 'Classifier Performance Evaluator' icon is circled in red. Below the menu, the 'Knowledge Flow Layout' shows a workflow starting with an 'ArffLoader' connected to a 'Remove' filter. The data then splits into two paths: one through an 'Attribute Summarizer' and another through a 'Class Assigner'. Both paths lead to a 'Train Test SplitMaker', which outputs 'training set' and 'test set'. The 'training set' is used by a 'J48' classifier, and the 'test set' is used by a 'Naive Bayes' classifier. The 'Naive Bayes' classifier is connected to a 'Classifier Performance Evaluator', which is also connected to a 'Model Performance Chart'. A red arrow points from the title to the 'Classifier Performance Evaluator' icon in the menu and another red arrow points from the 'Naive Bayes' classifier to the 'Classifier Performance Evaluator' in the workflow.

Connecting second performance evaluator to the same Model Performance Chart



Run both classifiers

Weka KnowledgeFlow Environment

DataSourcees | DataSinks | Filters | Classifiers | Clusterers | Associations | Evaluation | Visualization

Evaluation

Training SetMaker | Test Set Maker | CrossValidation FoldMaker | TrainTest SplitMaker | InstanceStream ToBatchMaker | Class Assigner | ClassValue Picker | Classifier Performance Evaluator | Incremental Classifier Evaluator | Clusterer Performance Evaluator | Prediction Appender | Serial ModelS

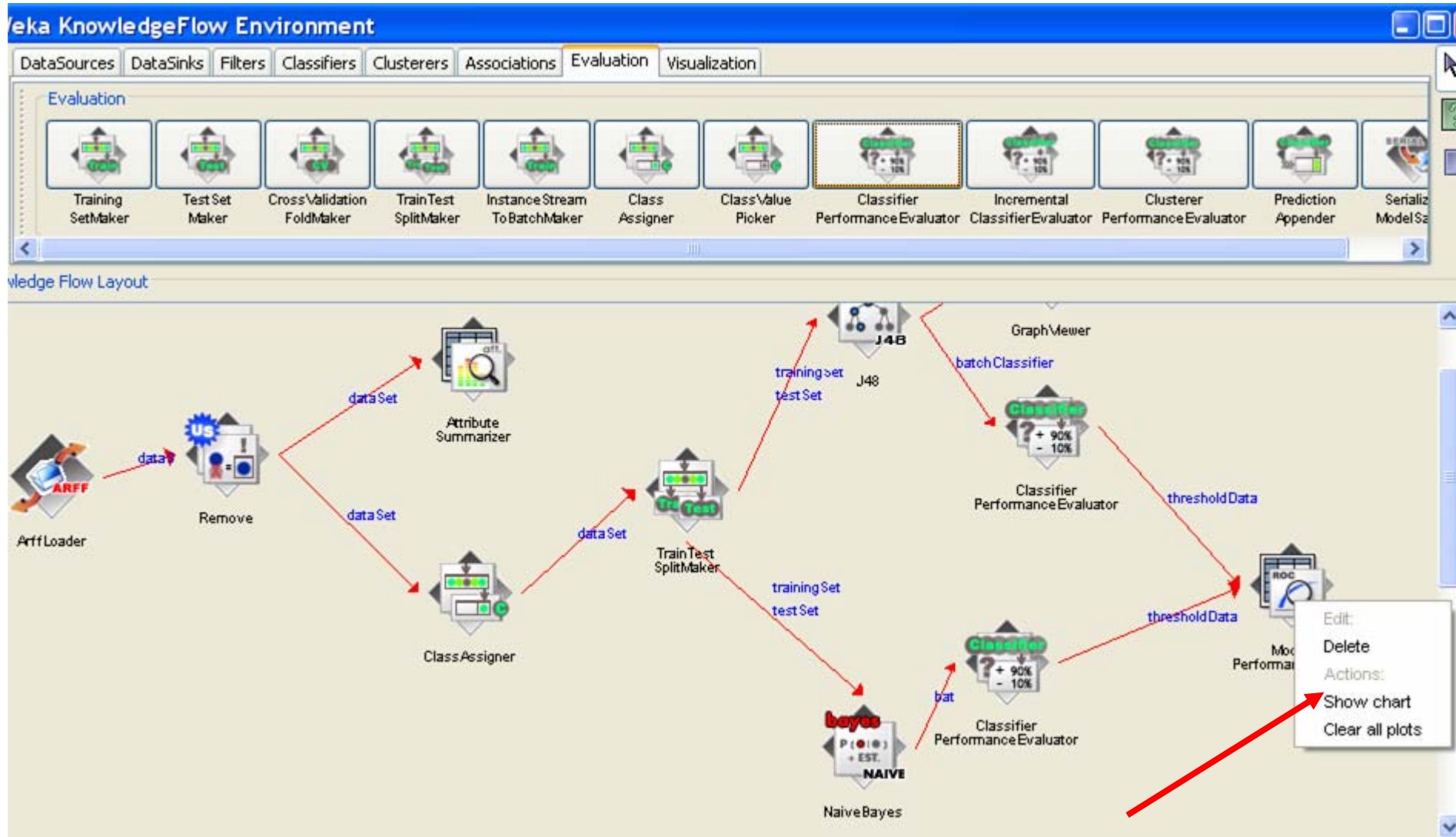
Knowledge Flow Layout

```
graph TD; ArffLoader[ArffLoader] -- data --> AttributeSummarizer[Attribute Summarizer]; ArffLoader -- data --> ClassAssigner[Class Assigner]; AttributeSummarizer -- data Set --> TrainTestSplitMaker[TrainTest SplitMaker]; ClassAssigner -- data Set --> TrainTestSplitMaker; TrainTestSplitMaker -- training set --> J48[J48]; TrainTestSplitMaker -- test set --> J48; TrainTestSplitMaker -- training set --> NaiveBayes[Naive Bayes]; TrainTestSplitMaker -- test set --> NaiveBayes; J48 -- batch Classifier --> J48Evaluator[Classifier Performance Evaluator]; NaiveBayes -- bat --> NaiveBayesEvaluator[Classifier Performance Evaluator]; J48Evaluator -- threshold Data --> ModelPerformanceChart[Model Performance Chart]; NaiveBayesEvaluator -- threshold Data --> ModelPerformanceChart;
```

ArffLoader context menu:

- Edit:
- Delete
- Set name
- Configure...
- Connections:
- instance
- dataSet
- Actions:
- Start loading

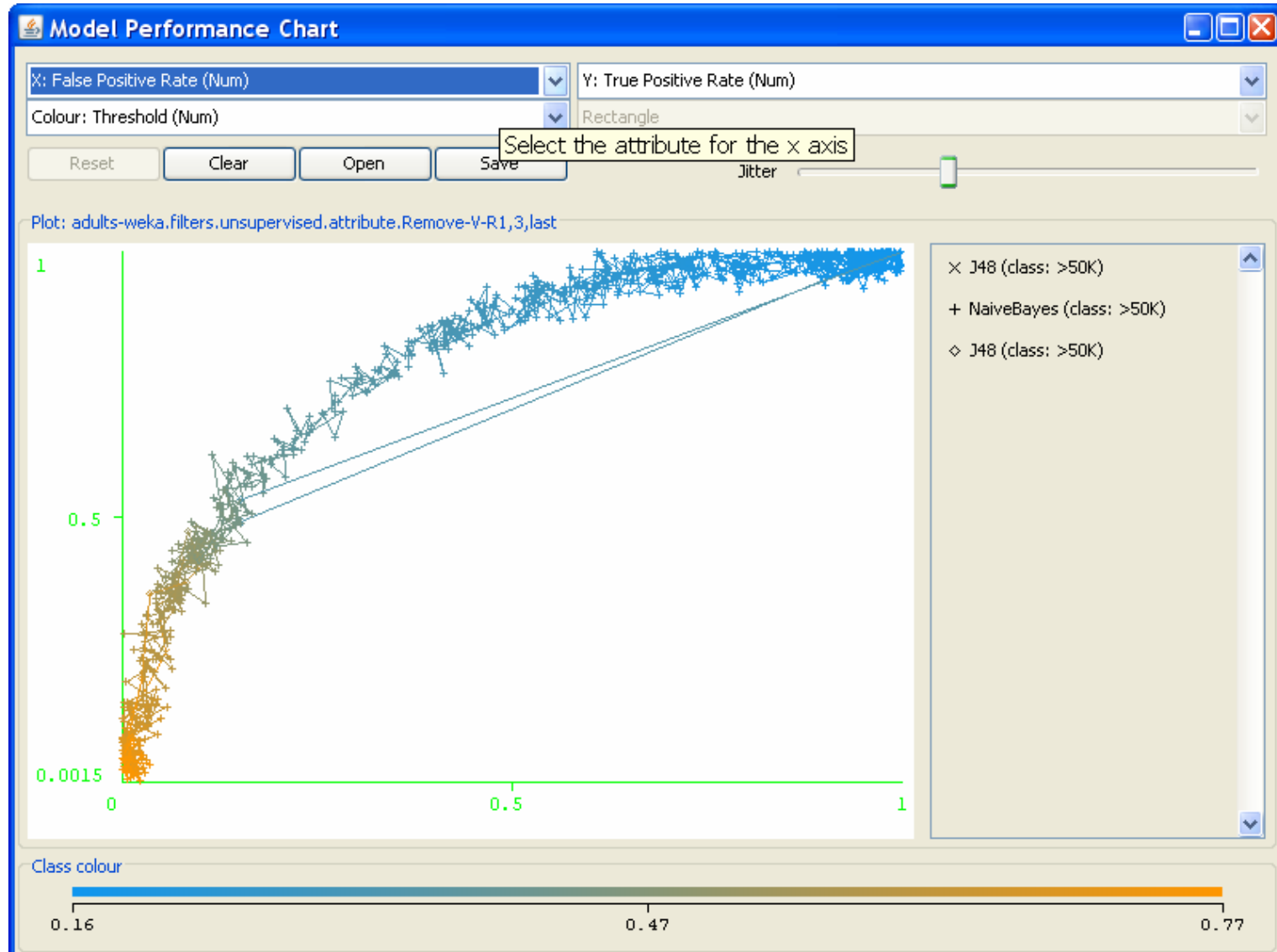
View ROC curves for both classifiers



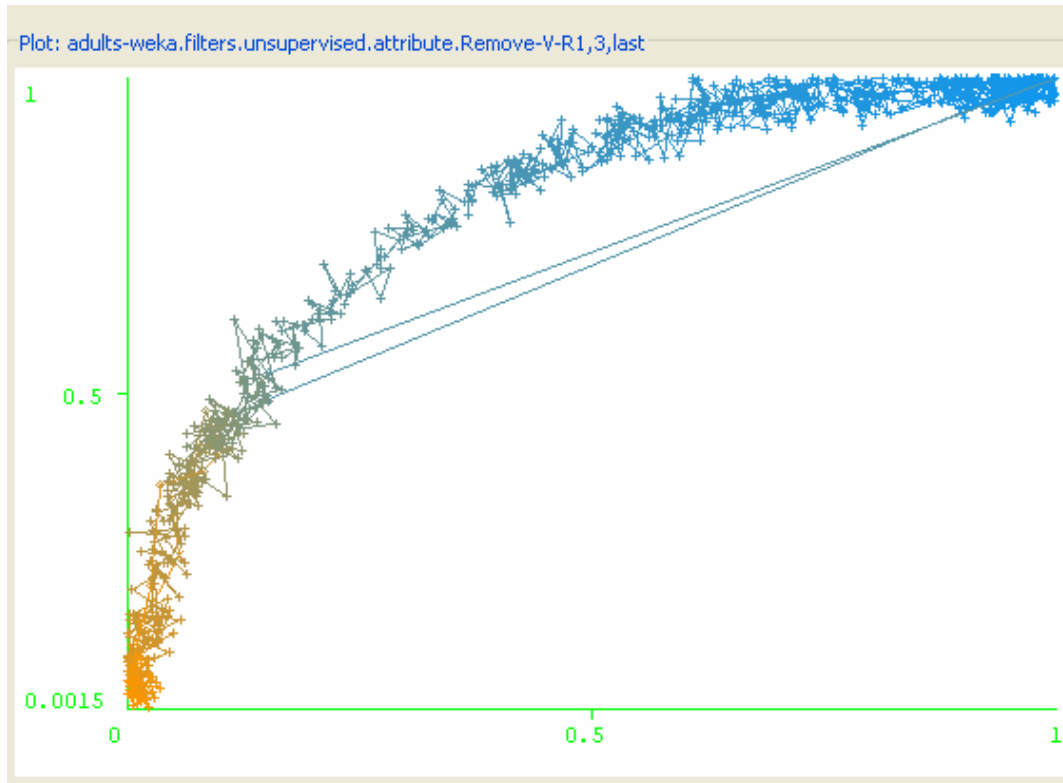
Lab outline

- Remind what ROC curve is
- Generate ROC curves using WEKA
- Some usage of ROC curves

Compare classifiers using their ROC curves



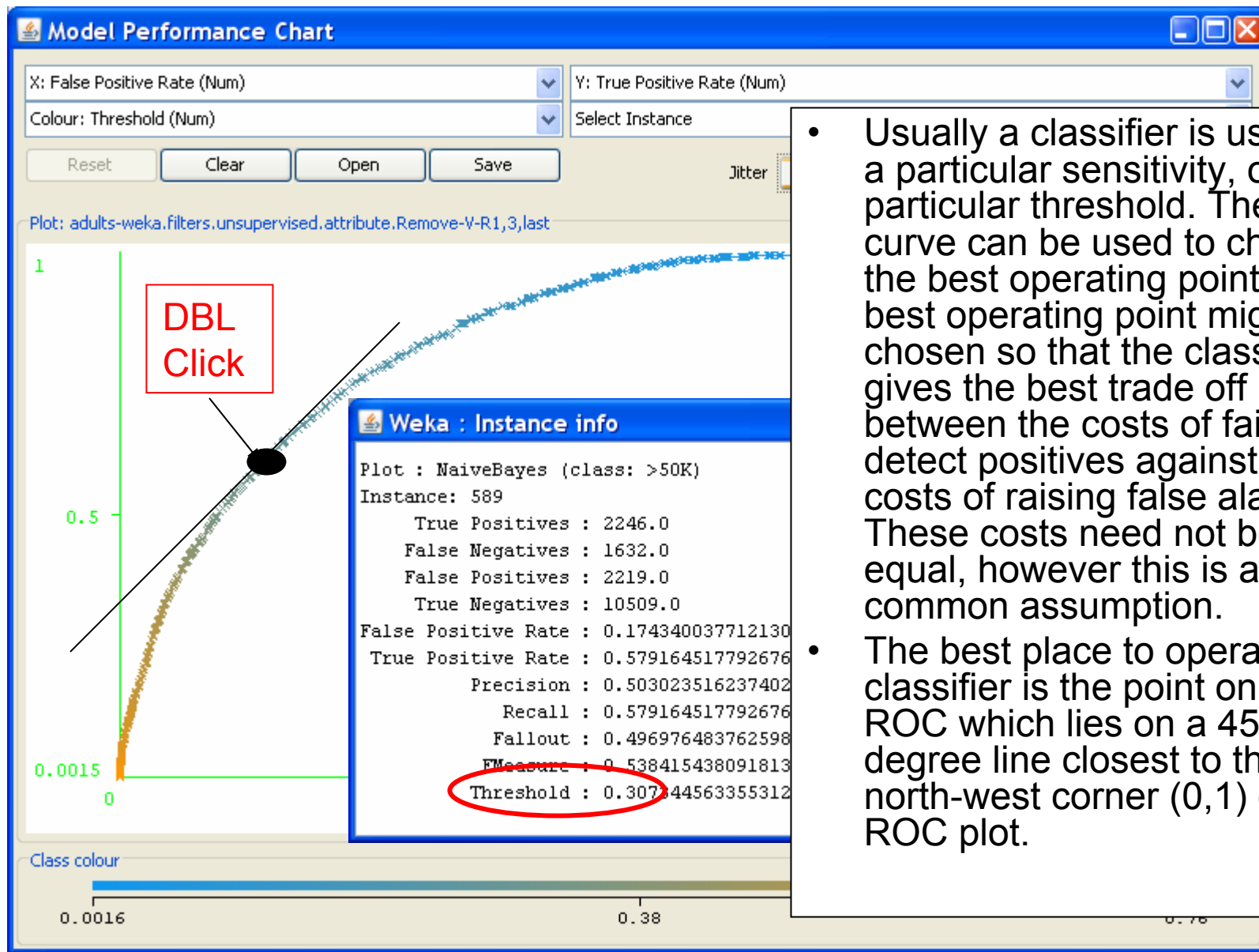
How good is the classifier



The area under the ROC curve shows the quality of a classifier – not accuracy, but the ability to separate between positive and negative instances.

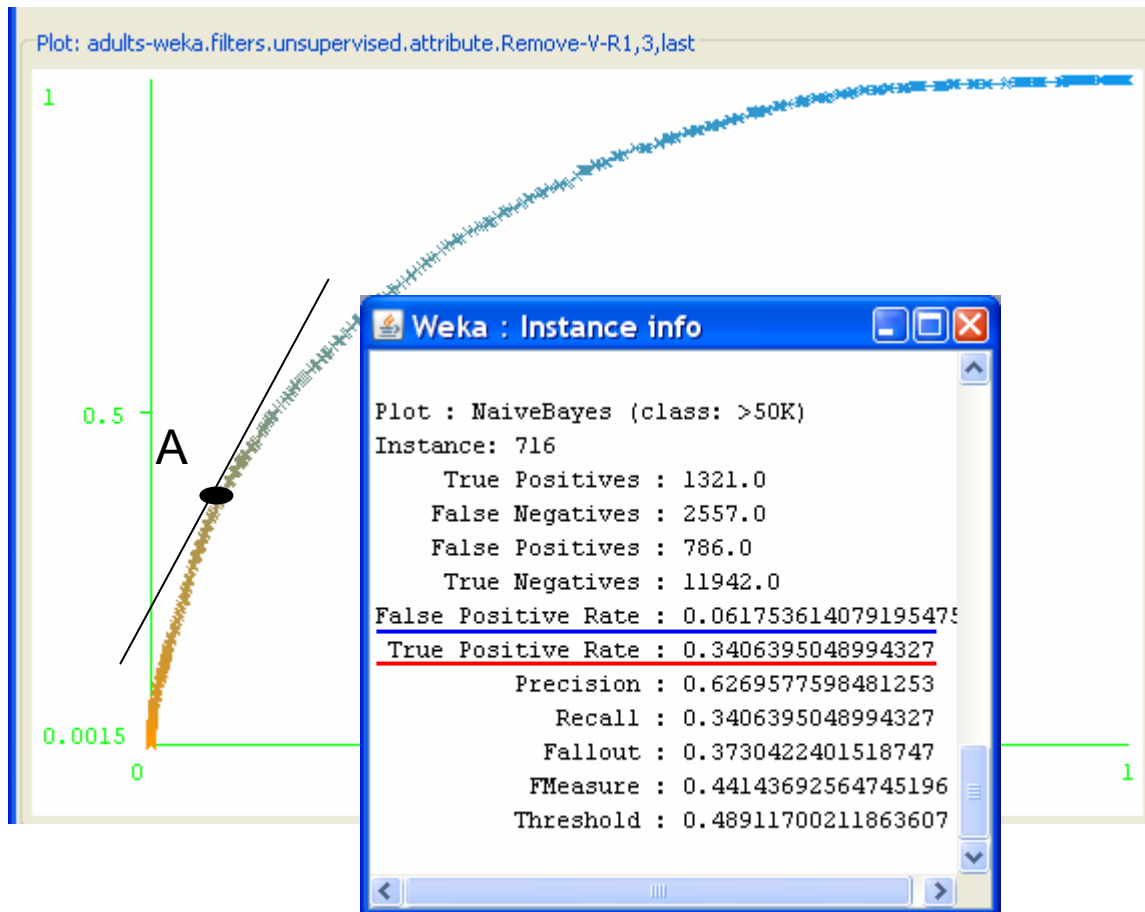
What classifier is better?

Choosing the Operating Point



- Usually a classifier is used at a particular sensitivity, or at a particular threshold. The ROC curve can be used to choose the best operating point. The best operating point might be chosen so that the classifier gives the best trade off between the costs of failing to detect positives against the costs of raising false alarms. These costs need not be equal, however this is a common assumption.
- The best place to operate the classifier is the point on its ROC which lies on a 45 degree line closest to the north-west corner (0,1) of the ROC plot.

Cost sensitive operating points

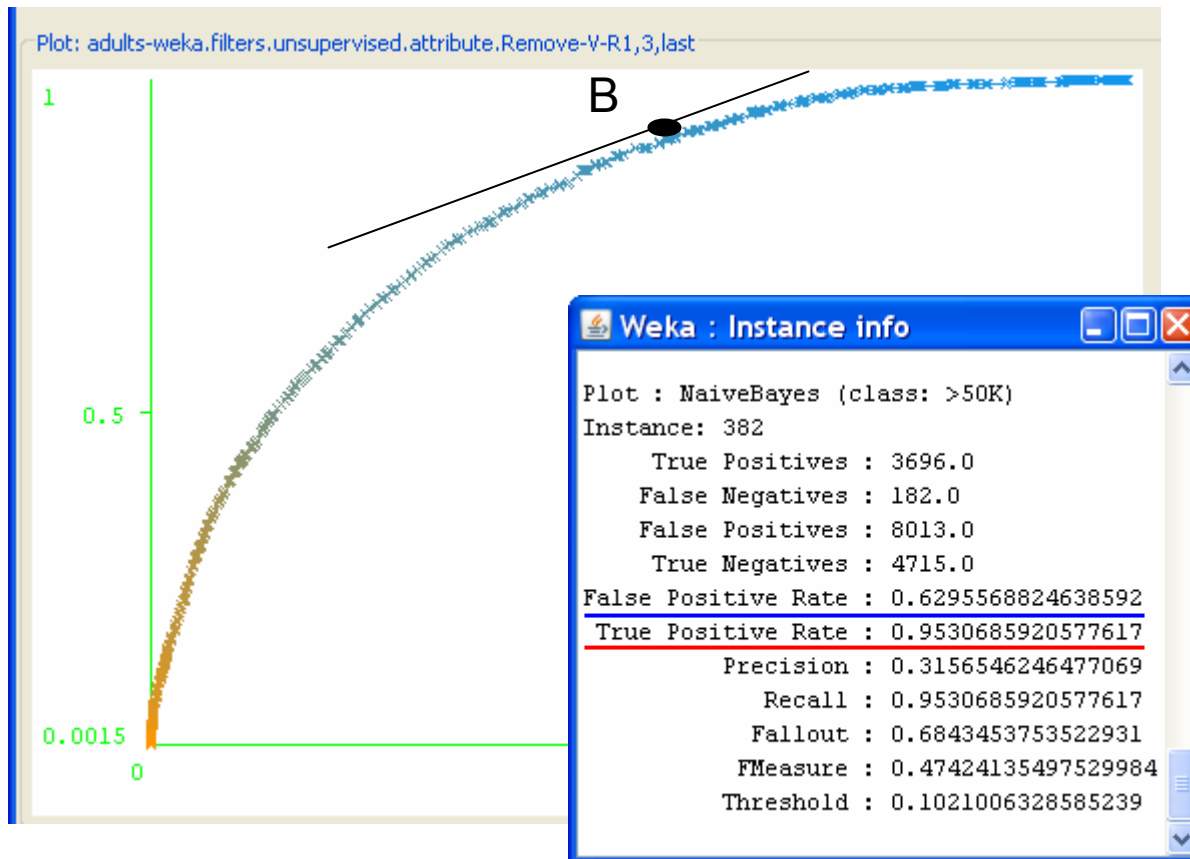


Is this threshold
good :

for cancer
detection?

for targeting
potential
customers?

Cost sensitive operating points



Is this threshold
good :

for cancer
detection?

for targeting
potential
customers?

Conclusions

- WEKA is a powerful datamining tool, but is not very easy to use



- There are other open source data mining tools, which are easier to use:
 - Orange:
 - <http://www.ailab.si/orange>
 - Tanagra:
 - <http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>