

Data Preparation

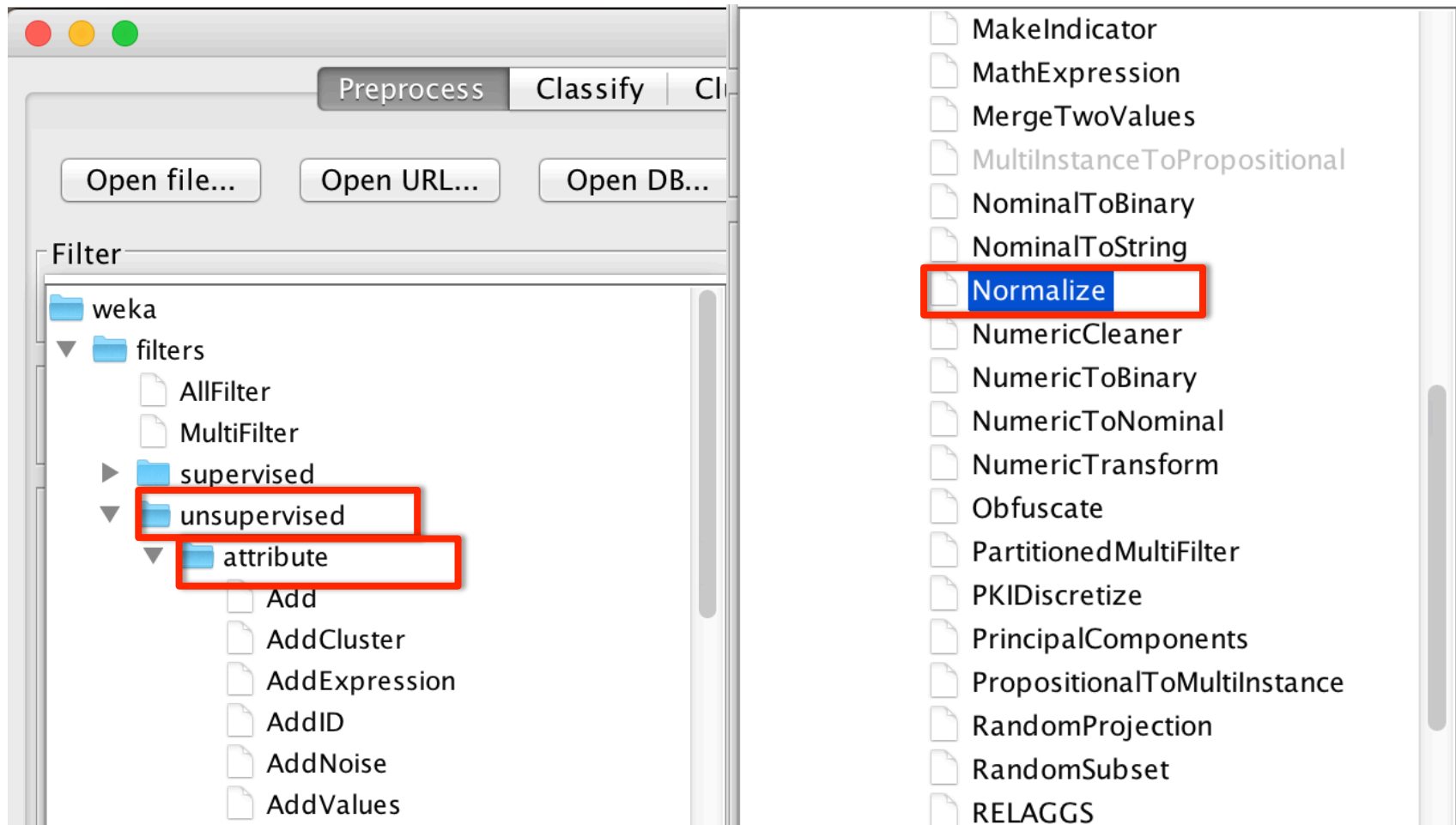
UROŠ KRČADINAC

EMAIL: uros@krcadinac.com

URL: <http://krcadinac.com>

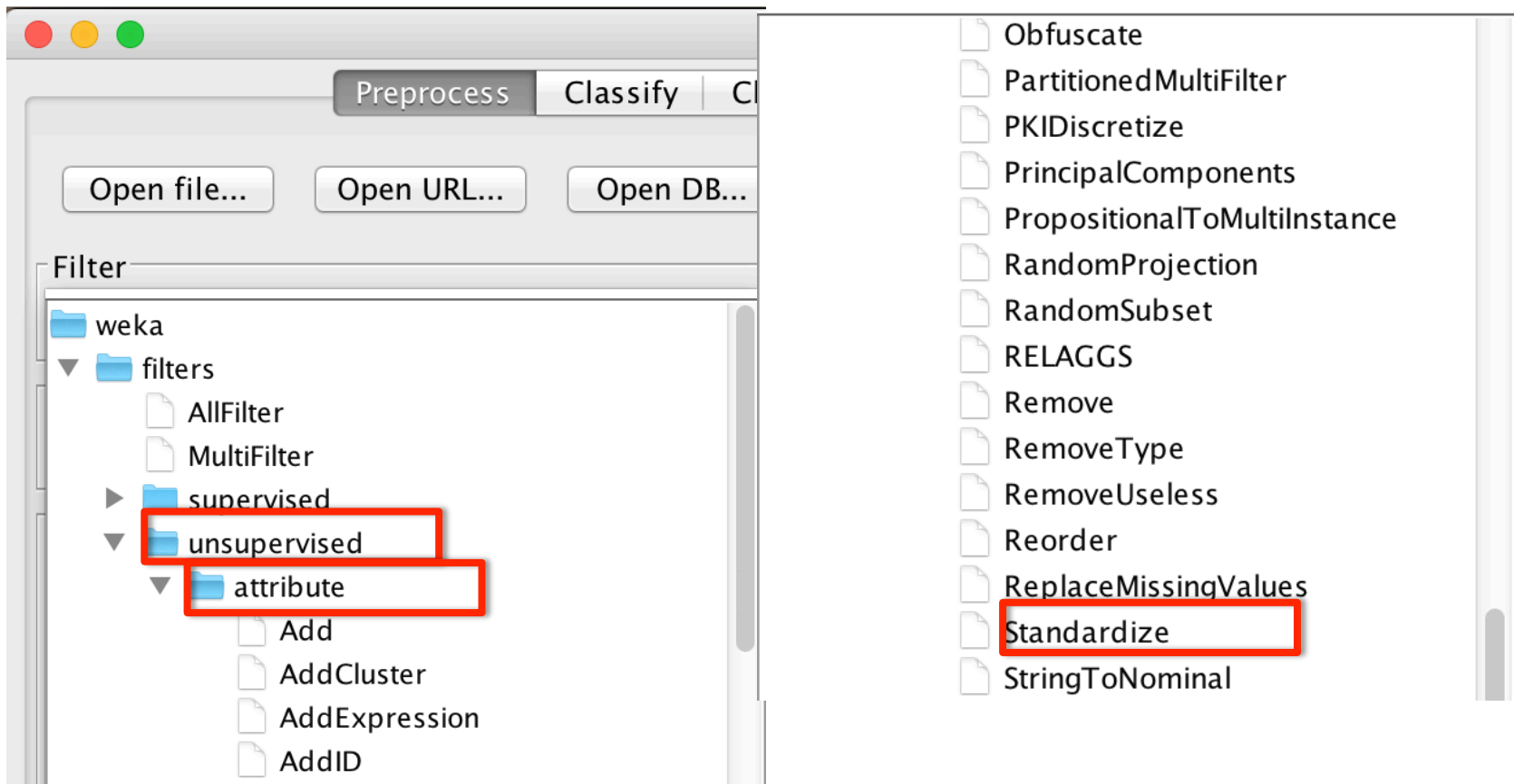
Normalization

Normalization is the process of rescaling the values to a specific value scale (typically 0 - 1)



Standardization

Standardization is a process of rescaling the values in order for the mean value to be 0, and standard deviation to have value of 1



Attribute discretization

Discretization is the process of transformation numeric data into nominal data, by putting the numeric values into distinct groups, which length is fixed.

Common approaches:

- Unsupervised:
 - Equal-width binning
 - Equal-frequency binning
- Supervised – classes are taken into account

Equal- Width Binning

Equal-width binning divides the scope of possible values into N subsopes (bins) of the same width.

$$\text{width} = (\text{max value} - \text{min value}) / N$$

Example: If the scope of the values is between 0 and 100, we should create 5 subsopes (bins) in the following manner:

$$\text{Width} = (100 - 0) / 5 = 20$$

Subsopes are: [0-20], (20-40], (40-60], (60-80], (80-100]

Usually, the first and the final subscope (bin) are being expended in order to include possible values outside the original scope.

Equal-frequency binning

Equal-frequency binning (or equal -height binning) divides the scope of possible values into N subsopes where each subscope (bin) carries the same number of instances.

Example: We want to put the following values in 5 subsopes (bins):

5, 7, 12, 35, 65, 82, 84, 88, 90, 95

So, each subscope will have 2 instances:

5, 7, | 12, 35, | 65, 82, | 84, 88, | 90, 95

Discretization in Weka

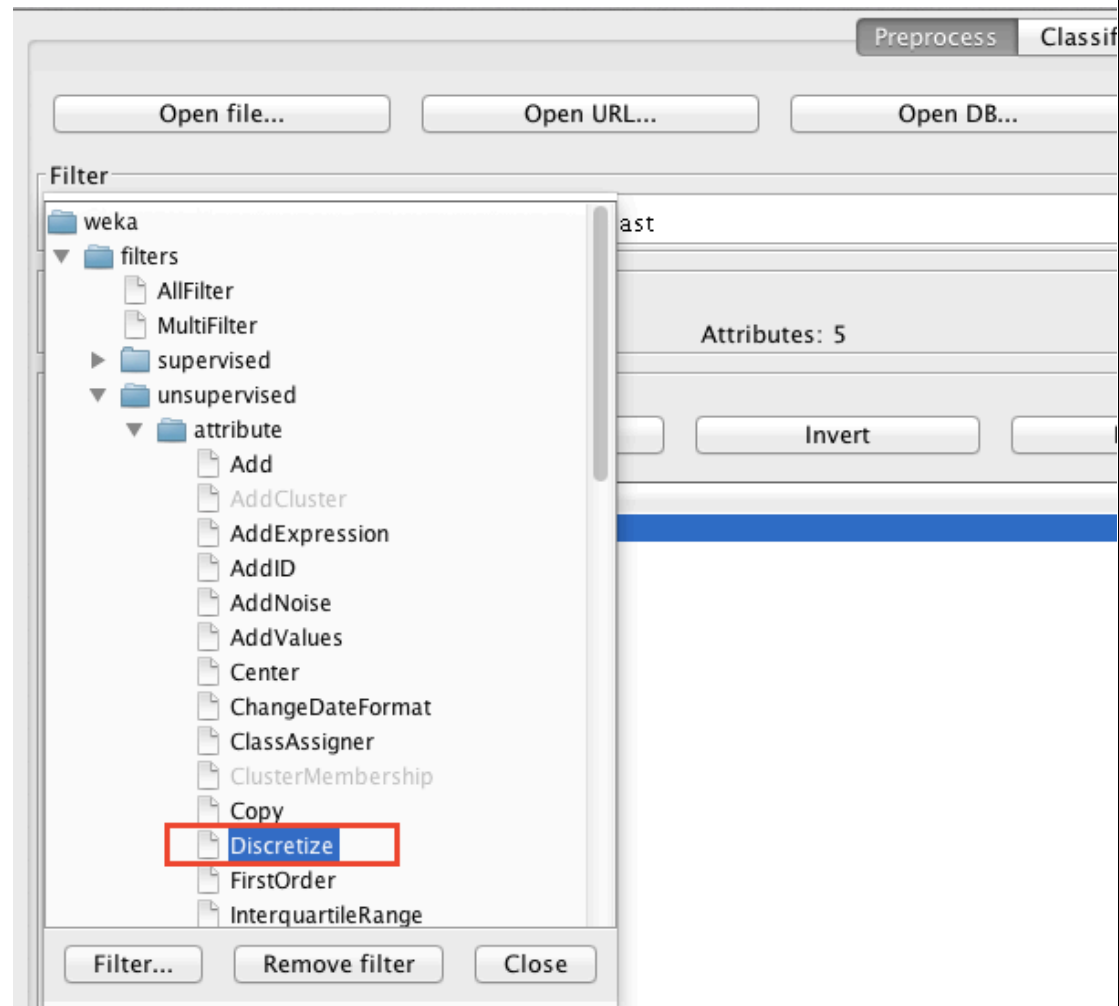
We apply certain *Filters* to attributes we want to discretize.

Preprocess tab

Option: *Choose -> Filter*

*filters/unsupervised/
attribute/Discretize.*

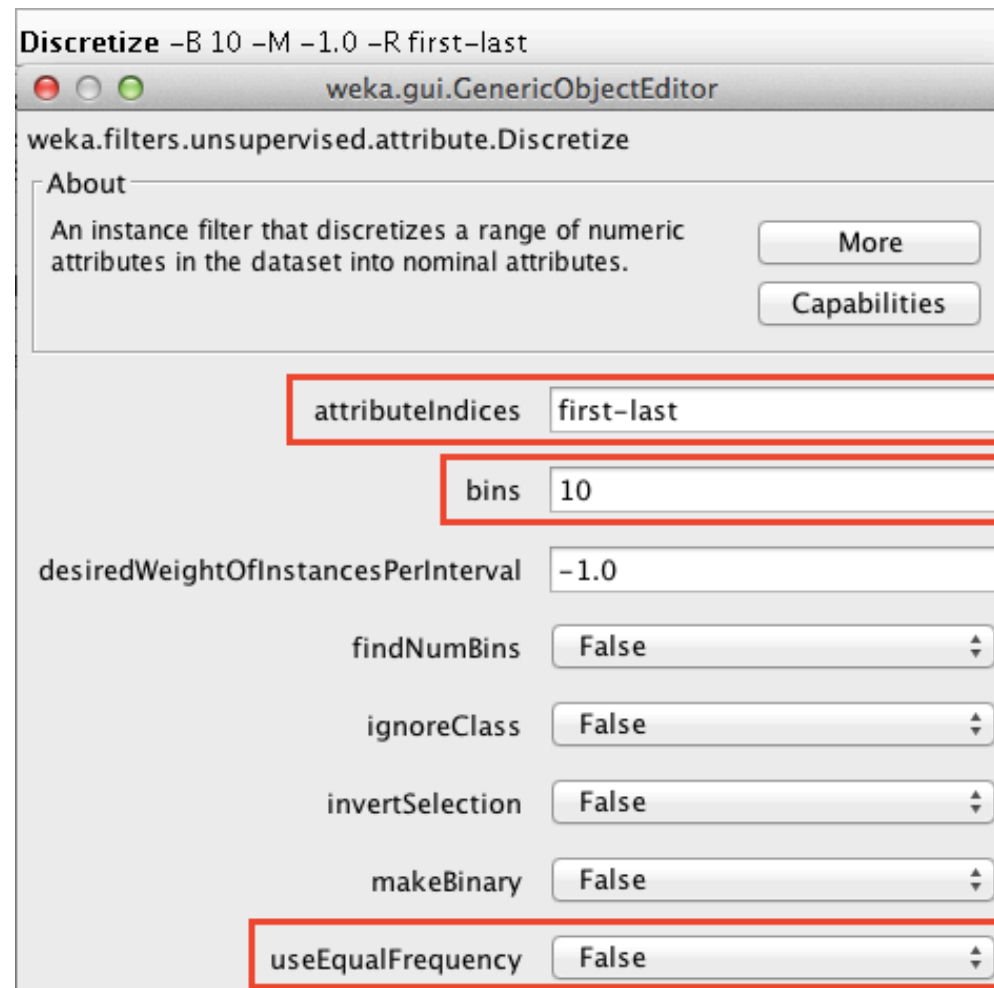
FishersIrisDataset.arff



Discretization in Weka

Equal-width binning is the default option.

- *attributeIndices* – the first-last value means that we are discretizing all values. We can also name the attribute numbers
- *bins* - the desired number of scopes (bins)
- *useEqualFrequency* - *false* by default; *true* if we use Equal Frequency binning



Discretization in Weka

Applying the filter

Filter: Choose **Discretize -B 10 -M -1.0 -R first-last** Apply

Current relation
Relation: FishersIrisDataset-weka.filters.unsupervised.attribute.Remove...
Instances: 150 Attributes: 5

Attributes

All None Invert Pattern

No. Name

☒ 1 Sepal Length

☐ 2 Sepal Width

☐ 3 Petal Length

☐ 4 Petal Width

☐ 5 Species

Remove

Selected attribute

Name: Sepal Length
Missing: 0 (0%) Distinct: 10 Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	'(-inf-4.66]'	9
2	'(4.66-5.02]'	23
3	'(5.02-5.38]'	14
4	'(5.38-5.74]'	27
5	'(5.74-6.1]'	22
6	'(6.1-6.46]'	20
7	'(6.46-6.82]'	18
8	'(6.82-7.18]'	6
9	'(7.18-7.54]'	5
10	'(7.54-inf]'	6

The resulting subscopes (bins)

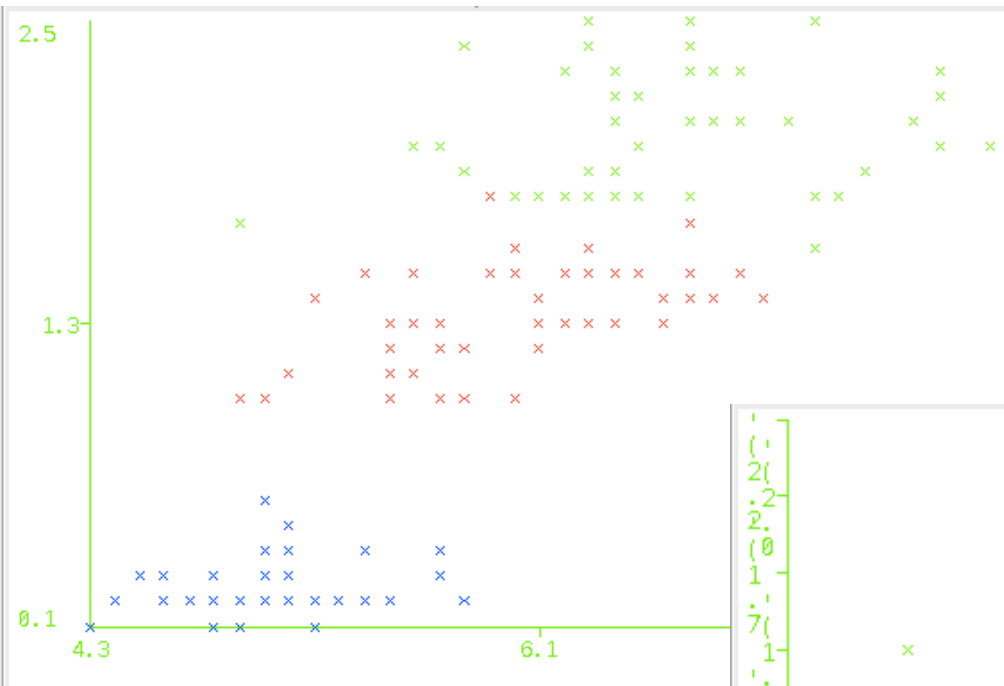
Class: Species (Nom) Visualize All

Bin	Setosa	Versicolour	Virginica	Total
1	9	0	0	9
2	23	0	0	23
3	14	0	0	14
4	27	0	0	27
5	22	0	0	22
6	20	0	0	20
7	18	0	0	18
8	6	0	0	6
9	5	0	0	5
10	6	0	0	6

Status: OK

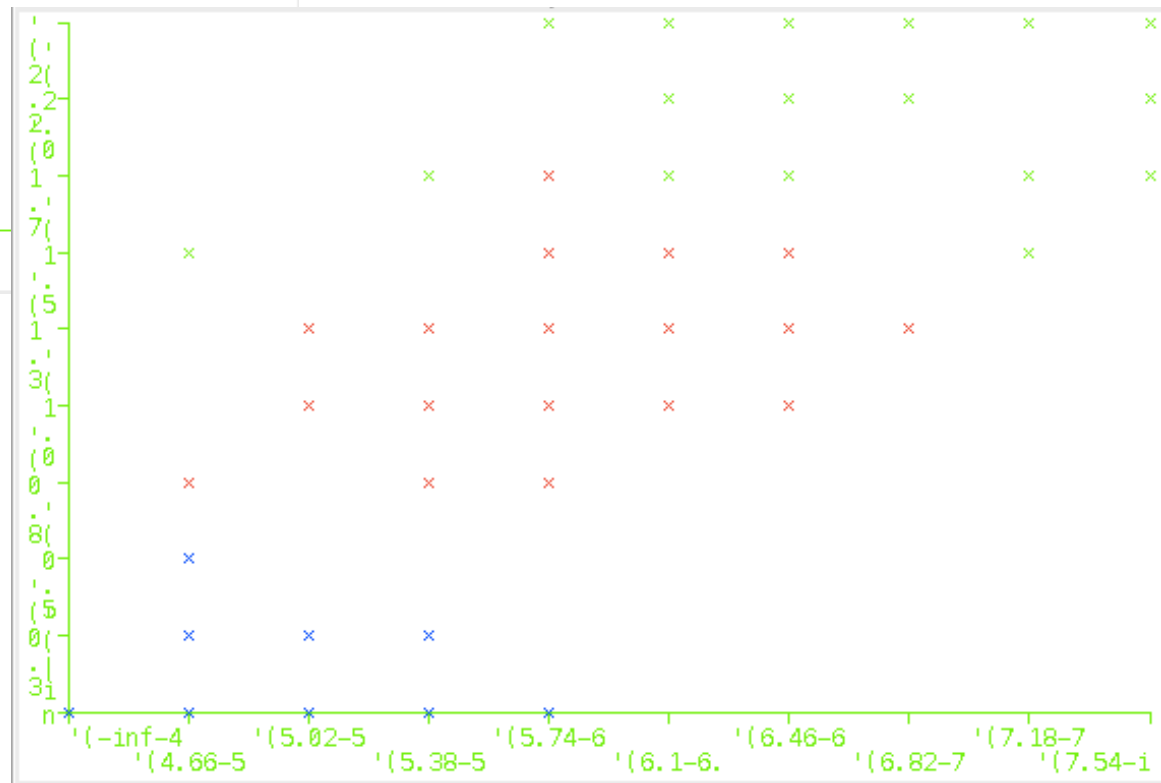
Log x 0

Data, before and after discretization



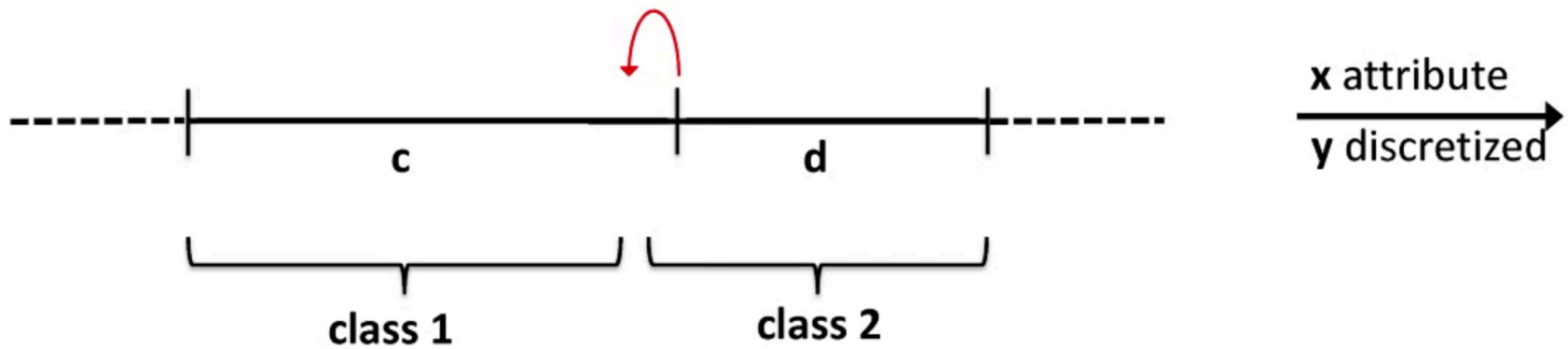
Before

After



Supervised discretization

- What if all instances in a bin have one class, and all instances in the next higher bin have another class except for the first, which has the original class?



- Supervised discretization takes the class values in account

Supervised discretization

- Use the entropy heuristic
- In the example *weather.numeric.arff*, the *temperature* attribute

64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no yes	yes yes	no	yes	yes	no
4 yes, 1 no					5 yes, 4 no						
entropy = 0.934 bits											

- Choose split point with smallest entropy (largest information gain)

64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no yes	yes yes	no	yes	yes	no

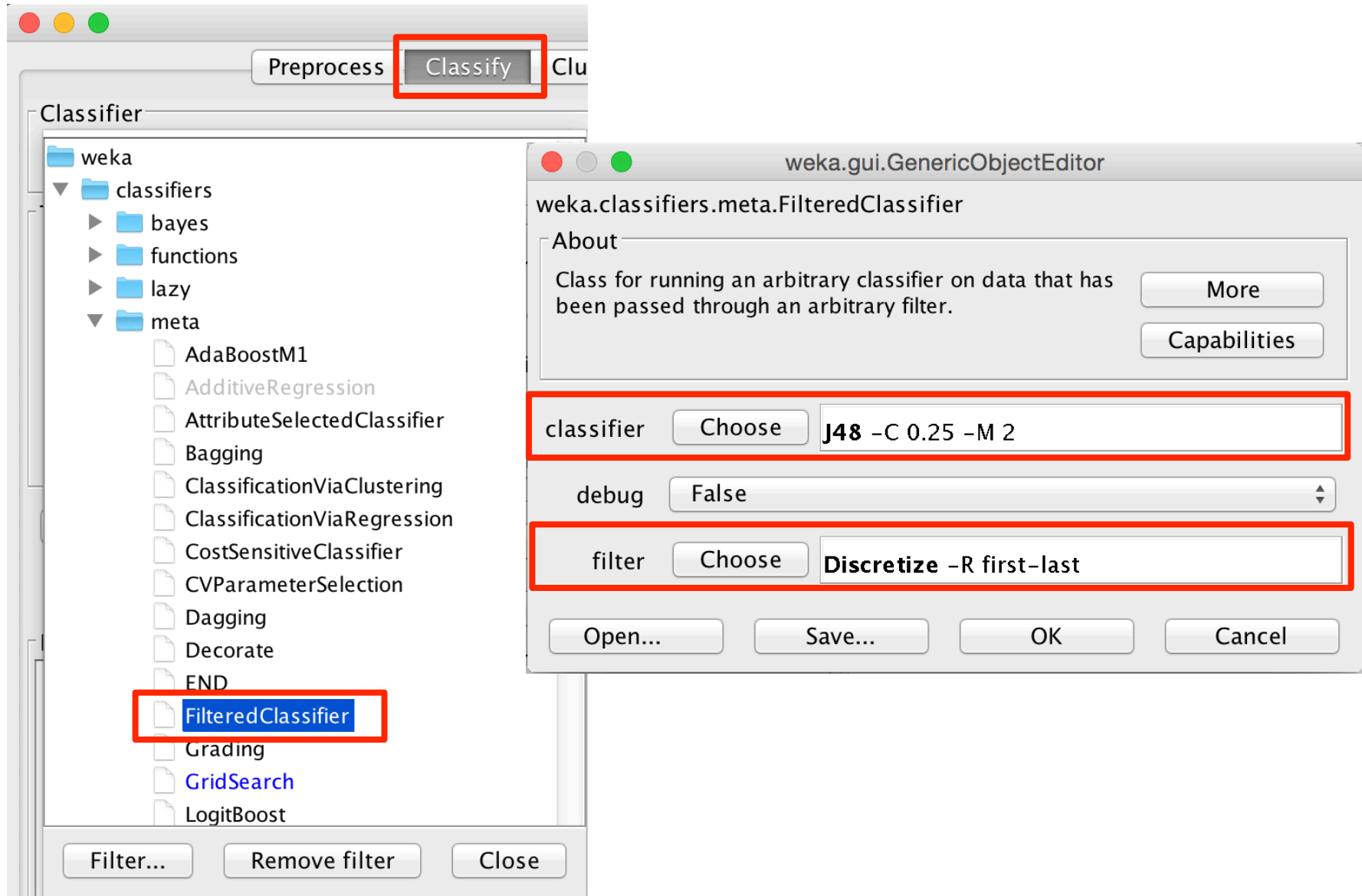
Supervised discretization in Weka

weather.numeric.arff

The image displays two side-by-side screenshots of the Weka software interface, specifically the 'Preprocess' tab. The left screenshot shows the 'Filter' dropdown menu with 'Choose' highlighted. The right screenshot shows the 'Filter' list with 'Discretize' highlighted. The 'weather.numeric.arff' dataset is loaded, showing 14 instances and 5 attributes.

The problem is that during supervised discretization we are using the data from the whole dataset, including the testing data. This will introduce an error when measuring classifier performance.

meta>FilteredClassifier



Attribute Selection

Attribute Selection (or Feature Selection) is the process of choosing a subset of relevant attributes that will be used during the further analysis.

It is being applied in cases where the dataset contains attributes which are redundant and/or irrelevant.

- Redundant attributes are the ones that do not provide more information than the attributes we already have in our dataset.
- Irrelevant attributes are the ones that are useless in the context of the current analysis.

Attribute Selection Advantages

Excessive attributes can degrade the performance of the model.

Advantages:

- Advances the readability of the model (because now the model contains only the relevant attributes)
- Shortens the training time
- Generalization power is higher because it lowers the possibility of overfitting

If the problem is well-known, the best way to select attribute is to do it manually. However, automated approaches also give good results.

Approaches to Attribute Selection

Two approaches:

- *Filter method* – use the approximation based on the general features of the data.
- *Wrapper method* – attribute subsets are being evaluated by using the machine learning algorithm, applied to the dataset. The name Wrapper comes from the fact that the algorithm is wrapped within the process of selection. The chosen subset of attributes is the one for which the algorithm gives the best results.

Attribute Selection Example

census90-income.arff

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter
Choose None Apply

Current relation
Relation: 1990census
Instances: 32561 Attributes: 15

Attributes
All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> workclass
3	<input type="checkbox"/> fnlwgt
4	<input type="checkbox"/> education
5	<input type="checkbox"/> education-num
6	<input type="checkbox"/> marital-status
7	<input type="checkbox"/> occupation
8	<input type="checkbox"/> relationship
9	<input type="checkbox"/> race
10	<input type="checkbox"/> sex
11	<input type="checkbox"/> capital-gain
12	<input type="checkbox"/> capital-loss
13	<input type="checkbox"/> hours-per-week
14	<input type="checkbox"/> native-country
15	<input type="checkbox"/> income

Remove

Selected attribute
Name: age
Missing: 0 (0%) Distinct: 73 Type: Numeric
Unique: 2 (0%)

Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

Class: income (Nom) Visualize All

Status
OK Log x 0

Attribute Selection Example

The screenshot shows the Weka GUI with the 'Preprocess' tab selected. The 'Filter' panel on the left displays a tree structure under 'weka' > 'filters' > 'supervised' > 'attribute'. The 'AttributeSelection' filter is highlighted with a red box. A yellow callout bubble points to it with the text: "We want to apply the attribute selection".

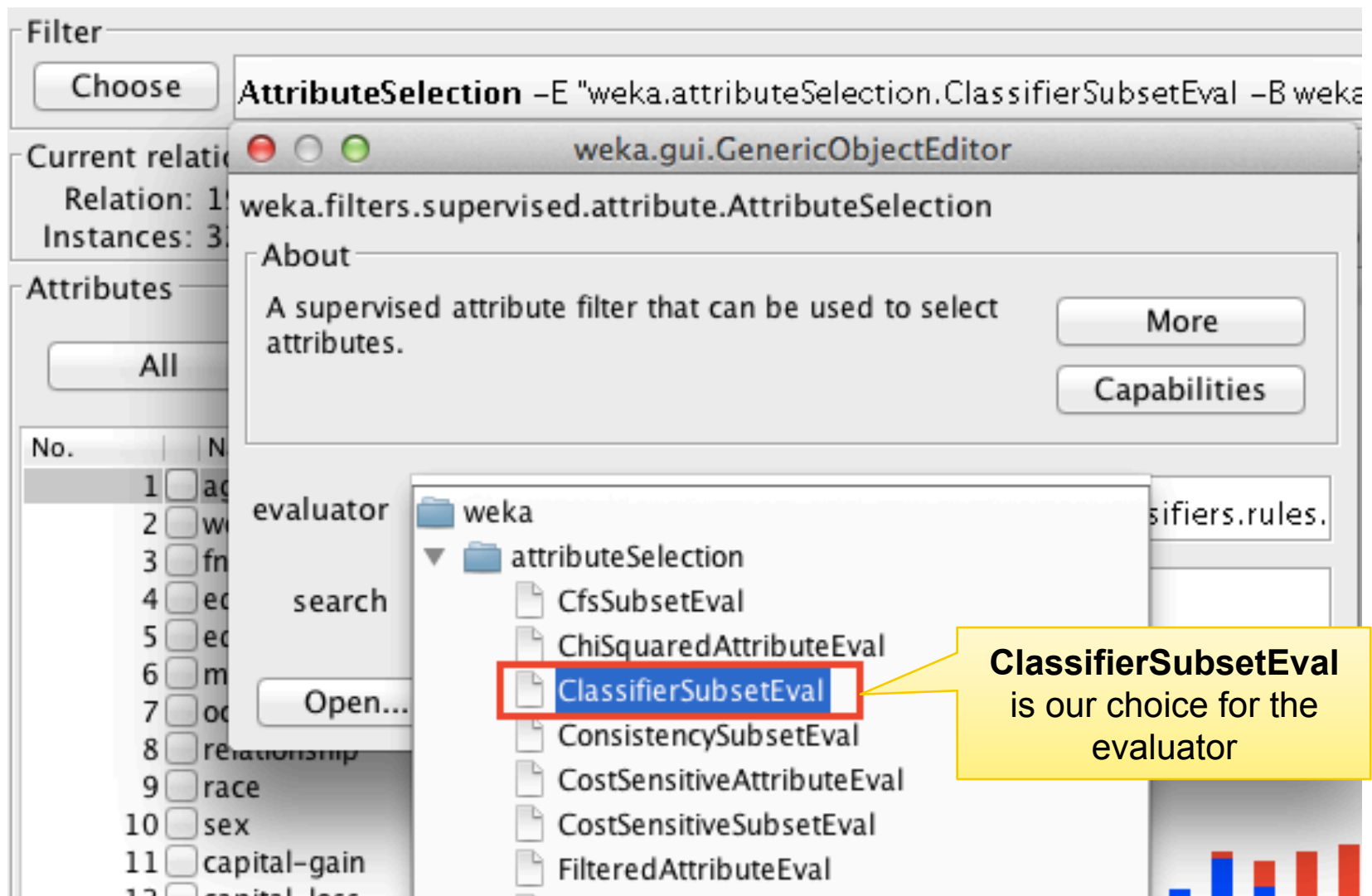
Below the filter list, the 'Class' dropdown is set to 'income (Nom)'. At the bottom of the filter panel are buttons for 'Filter...', 'Remove filter', and 'Close'.

The main window displays the 'Selected attribute' table for 'age'.

Selected attribute	
Name: age	Type: Unique
Missing: 0 (0%)	Distinct: 73
Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

At the bottom right, a histogram shows the distribution of the 'age' attribute, with blue bars for the selected attribute and red bars for the other attributes.

Attribute Selection Example



Attribute Selection Example

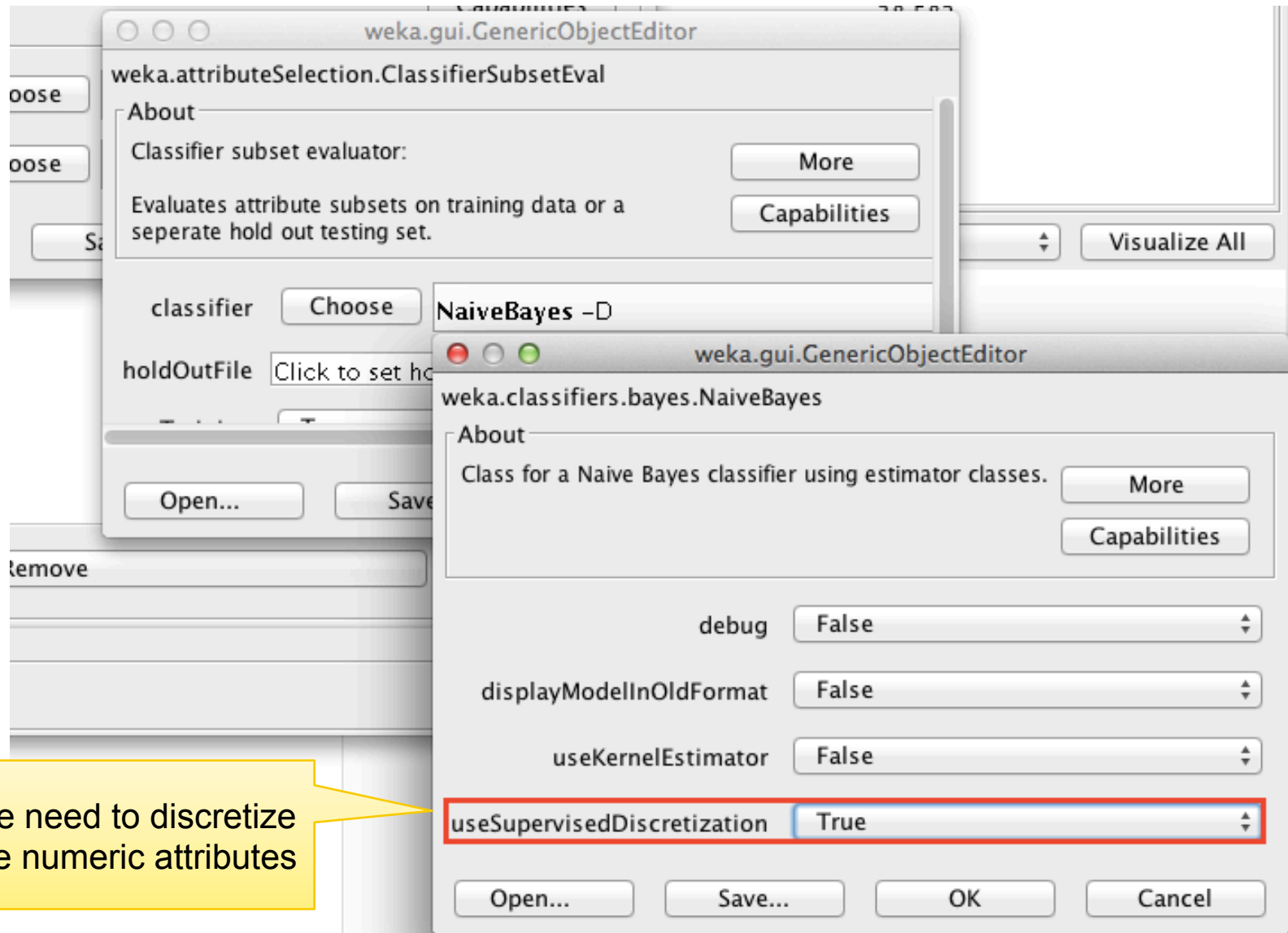
The screenshot displays the WEKA software interface during an attribute selection process. In the background, the 'AttributeSelection' window is open, showing the 'Filter' tab with 'weka.filters.supervised.attribute.AttributeSelection' selected. The 'Current relation' is 'Relation: 1' with 'Instances: 3'. The 'Attributes' list on the left includes: No., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15. The 'About' window for 'AttributeSelection' is also visible, describing it as a supervised attribute filter.

In the foreground, the 'ClassifierSubsetEval' window is open, showing the 'About' tab. The 'Classifier subset evaluator:' section states: 'Evaluates attribute subsets on training data or a separate hold out testing set.' The 'classifier' field is highlighted with a red box, and the 'Choose' button next to it is also highlighted. The 'NaiveBayes' classifier is selected in the dropdown menu. A yellow callout bubble points to the 'NaiveBayes' text, containing the text 'NaiveBayes classifier'.

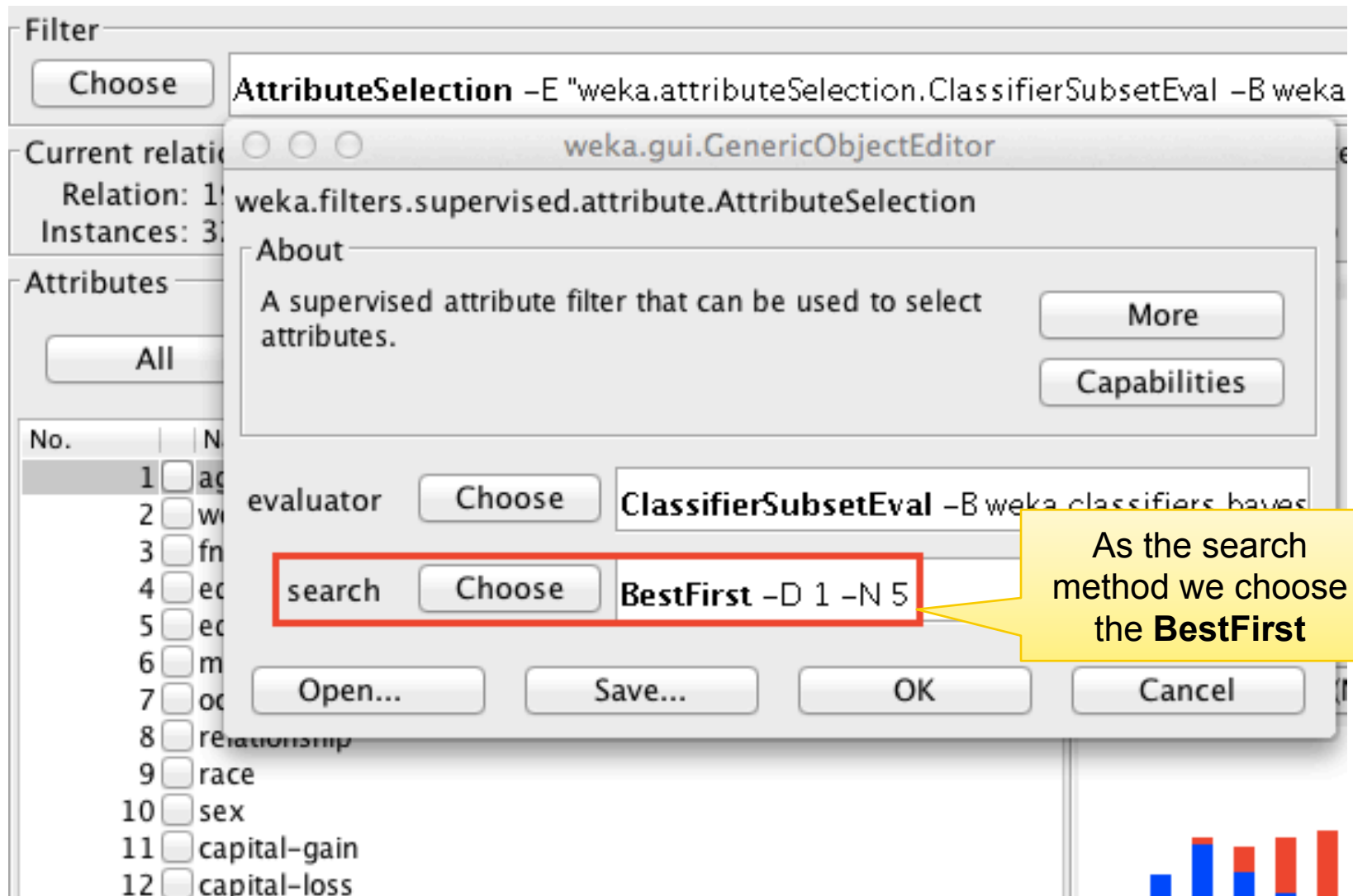
The 'holdOutFile' field is empty, and the 'useTraining' checkbox is checked. The 'Open...', 'Save...', 'OK', and 'Cancel' buttons are at the bottom.

No.	Value
1	17
2	90
3	38.582
4	13.64

Attribute Selection Example



Attribute Selection Example



Attribute Selection Example

The screenshot shows the Weka software interface with the 'Attribute Selection' tab active. The 'Filter' section at the top shows 'AttributeSelection' with a command line: `-E "weka.attributeSelection.ClassifierSubsetEval -B weka.classifiers.bayes.NaiveBayes -T -H \{Click`. The 'Apply' button is highlighted with a red box. A yellow callout box points to it with the text: "Filter is set and can be applied".

The 'Current relation' section shows 'Relation: 1990census' and 'Instances: 32561'. The 'Attributes' section lists 15 attributes, with 'age' selected (indicated by a blue highlight and a checked checkbox). The 'Selected attribute' section shows 'Name: age', 'Missing: 0 (0%)', 'Distinct: 73', and 'Type: Numerical'. The 'Statistic' section lists 'Minimum', 'Maximum', 'Mean', and 'StdDev'. The 'Class: income (Nom)' is selected, and the 'Visualize All' button is visible.

A histogram of the 'income' class is displayed at the bottom right, showing the distribution of income values. The x-axis ranges from 17 to 90, and the y-axis shows the frequency of instances. The histogram is a stacked bar chart with blue and red bars.

The 'Status' section at the bottom left shows 'OK'. The 'Log' button and a small icon are at the bottom right.

Attribute Selection Example

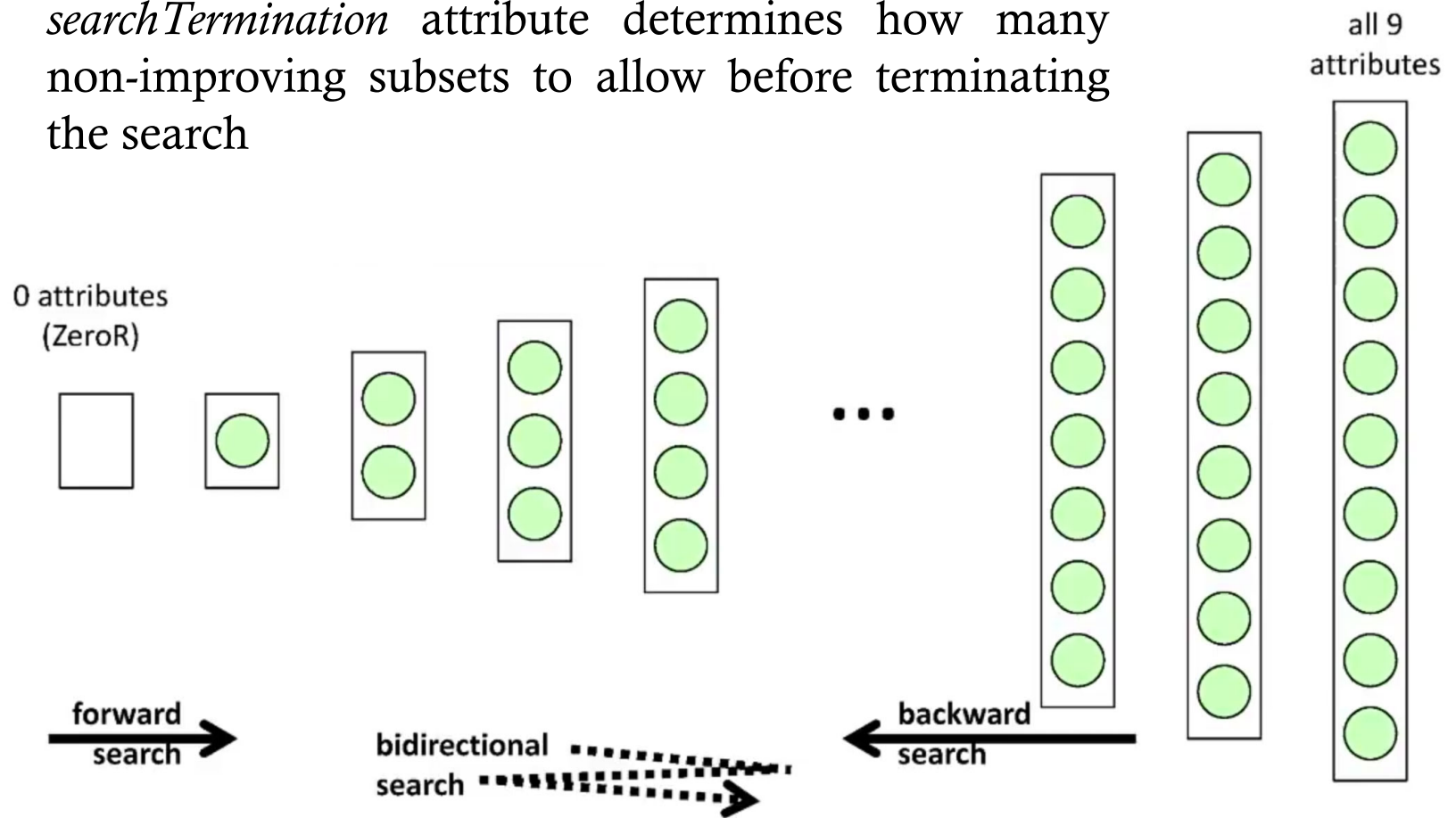
The screenshot displays the Weka software interface for attribute selection. The top menu bar includes 'Preprocess', 'Classify', 'Cluster', 'Associate', 'Select attributes', and 'Visualize'. Below this is a toolbar with buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Generate...', 'Undo', 'Edit...', and 'Save...'. The 'Filter' section shows the 'AttributeSelection' filter applied with the command: `-E "weka.attributeSelection.ClassifierSubsetEval -B weka.classifiers.bayes.NaiveBayes -T -H \\"Click"`. The 'Current relation' section indicates the dataset is '1990census-weka.filters.supervised.attribute.Attribute...' with 32561 instances and 7 attributes. The 'Attributes' section lists 7 attributes: 'age', 'education', 'relationship', 'race', 'capital-gain', 'capital-loss', and 'income'. A red box highlights the 'age' attribute, which is selected. A yellow callout box points to the 'age' attribute with the text: 'The number of attributes is reduced to 7'. The 'Selected attribute' section shows the 'age' attribute with statistics: Name: age, Missing: 0 (0%), Distinct: 73, Type: Numeric, Unique: 2 (0%). Below this is a table of statistics for 'age':

Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

The 'Class: income (Nom)' section shows a histogram of the 'income' class. The histogram has a blue base and red bars on top, showing a distribution of income values from 17 to 90. The status bar at the bottom shows 'Status OK' and a 'Log' button.

Search Method in Attribute Selection

- Exhaustive search (512 attribute subsets)
- Best First: *Forward, Backward, Bi-directional*
 - *searchTermination* attribute determines how many non-improving subsets to allow before terminating the search



Recommendations and credits

Weka Tutorials and Assignments @ The Technology Forge

- Link: <http://www.technologyforge.net/WekaTutorials/>

"Data Mining with Weka" and "More Data Mining with Weka": MOOCs from the University of Waikato. A self-paced session of "Data Mining with Weka" runs until 23 October June 2015.

- Link: <https://www.youtube.com/user/WekaMOOC/>

(Anonymous) survey for your
comments and suggestions:

<http://goo.gl/cqdp3l>

ANY QUESTIONS?

UROŠ KRČADINAC

EMAIL: uros@krcadinac.com

URL: <http://krcadinac.com>