Data Preparation

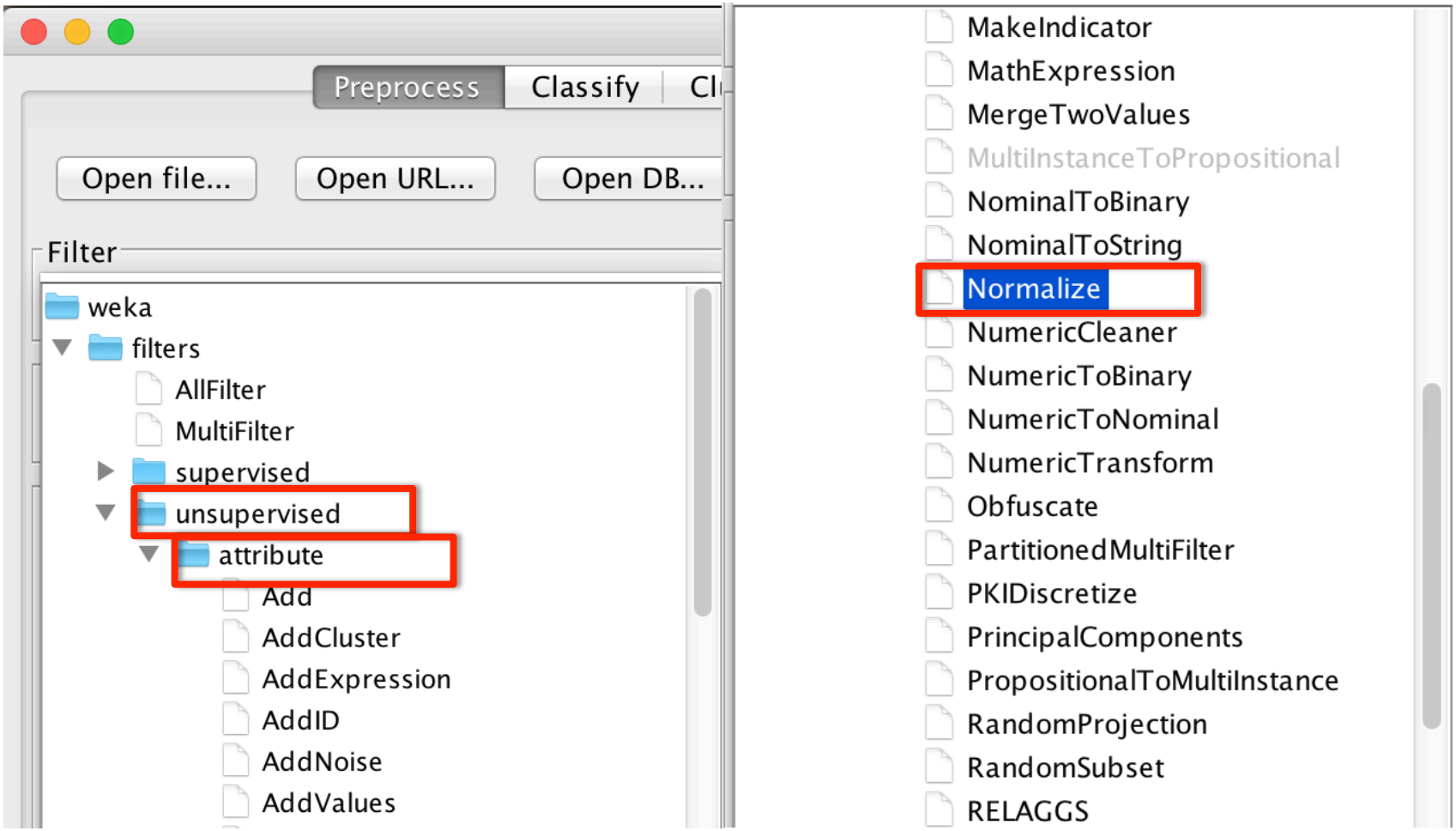
UROŠ KRČADINAC

EMAIL: uros@krcadinac.com

URL: http://krcadinac.com

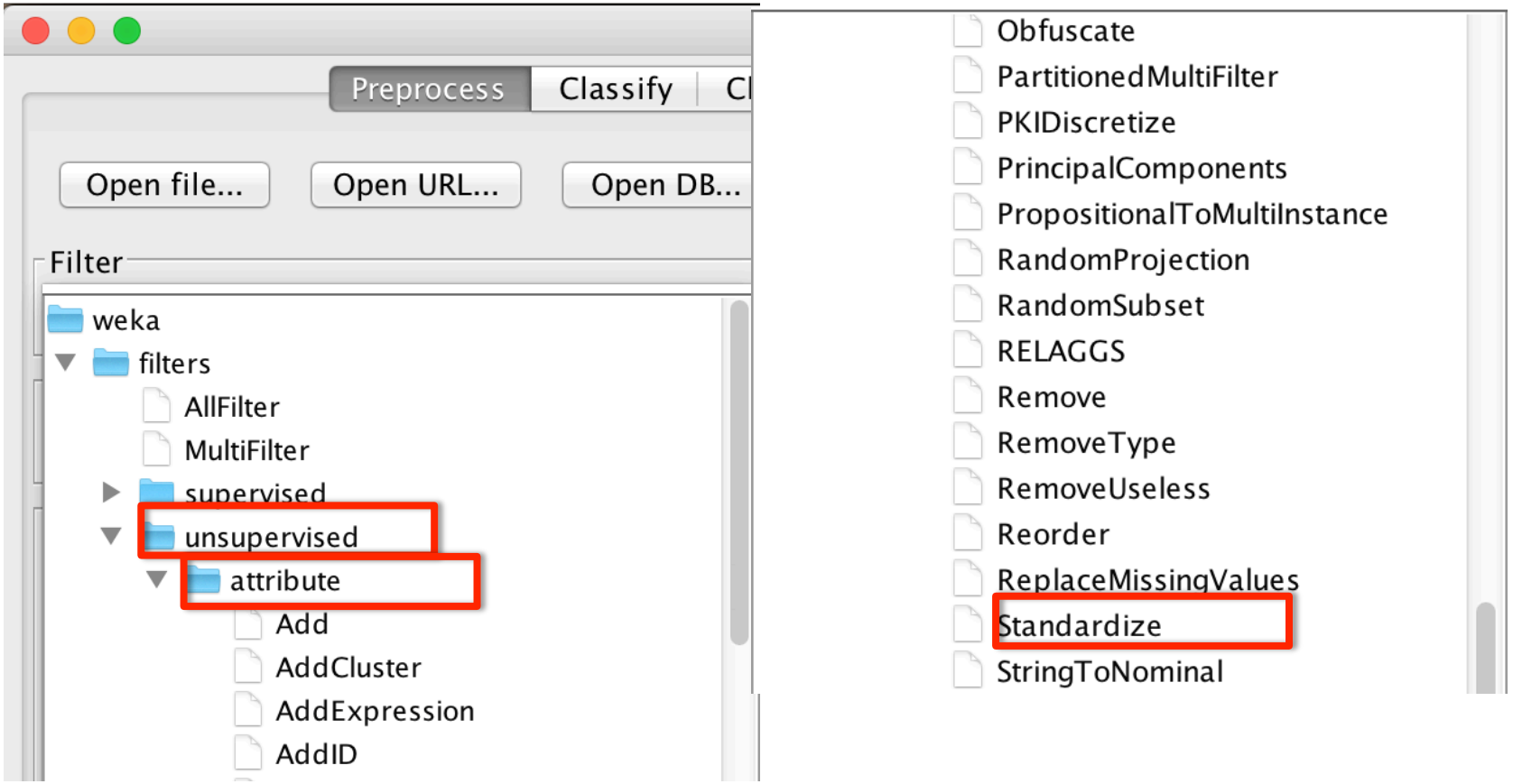
Normalization

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 subscopes (bins) of the same width.

width = (max value – min value) / N

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

Width = (100 – 0) / 5 = 20

Subscopes 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 subscopes where each subscope (bin) carries the same number of instances.

Example: We want to put the following values in 5 subscopes (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

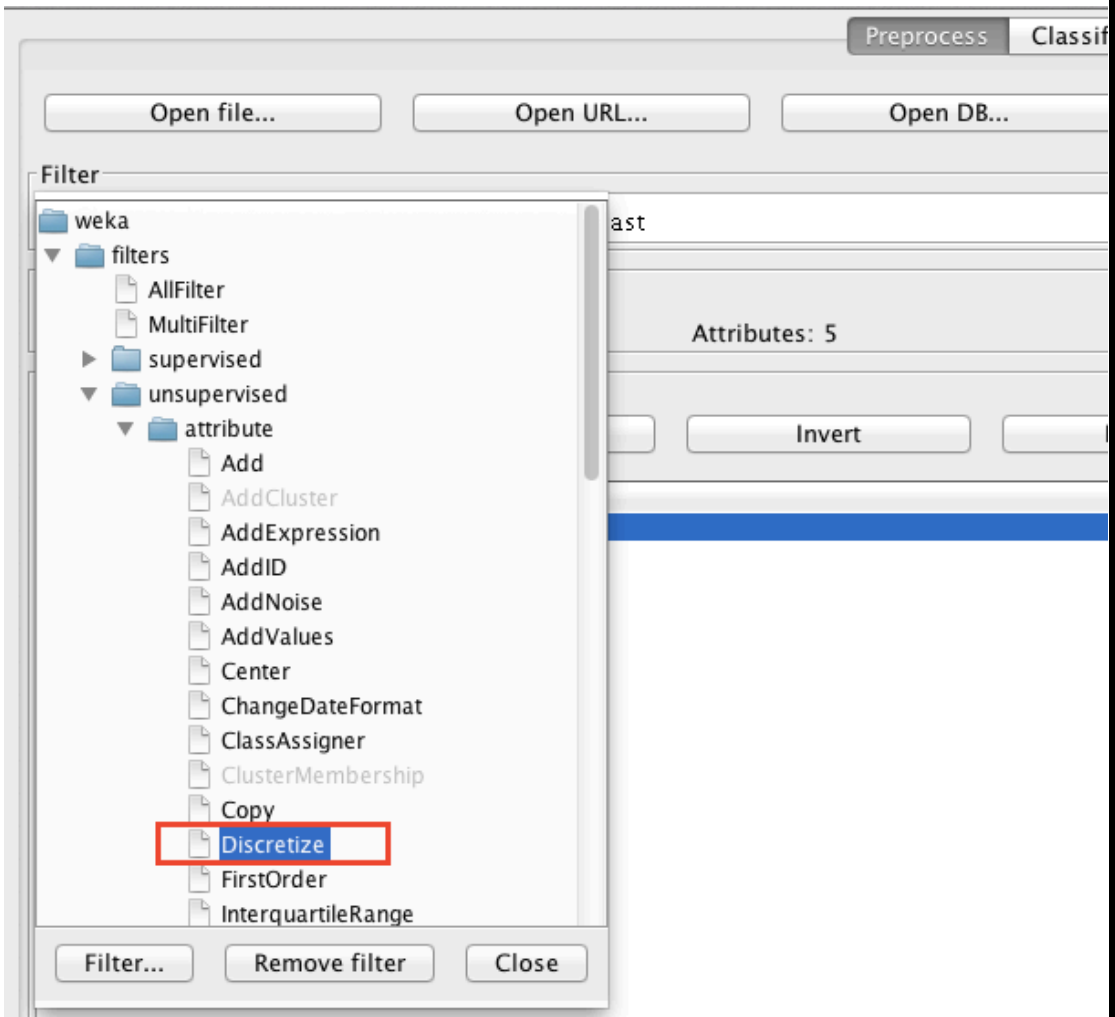
Discretization in Weka

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

Preprocess tab

Option: *Choose -> Filter*

*filters/unsupervised/attribute/**Discretize***.

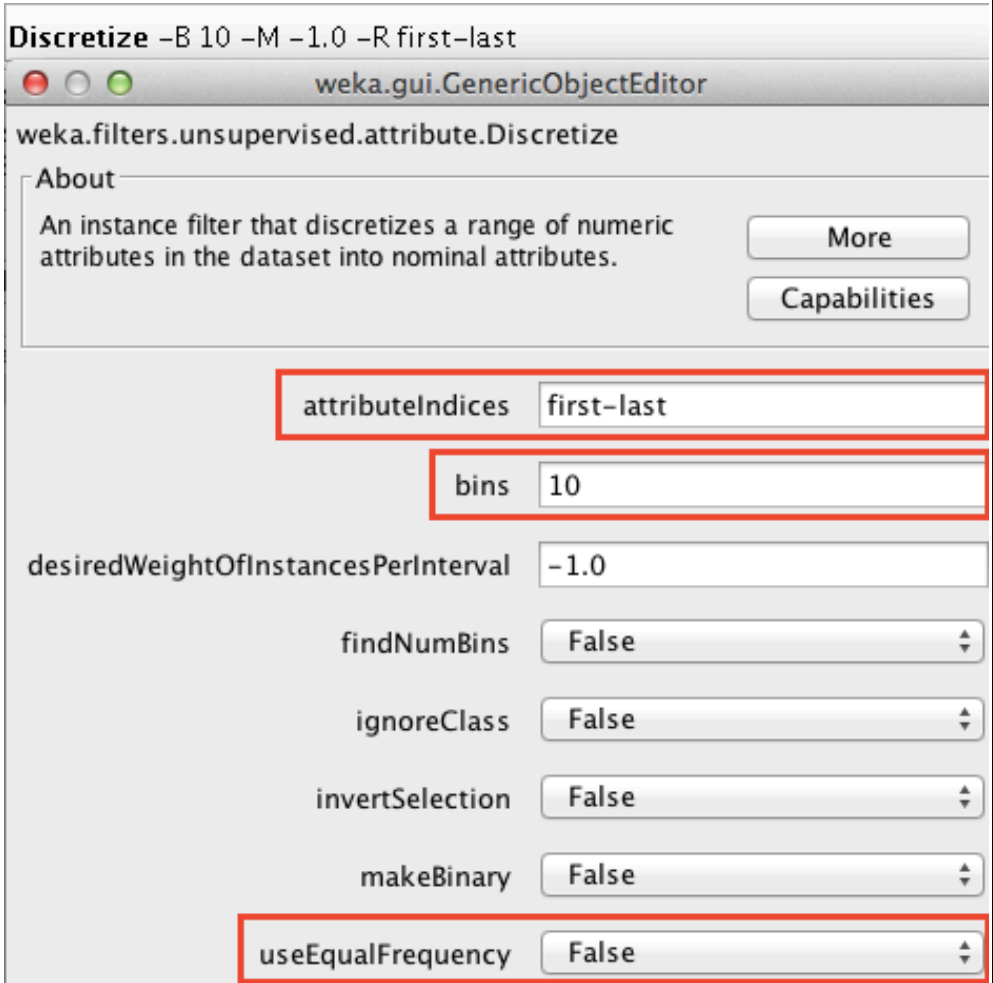


FishersIrisDataset.arff

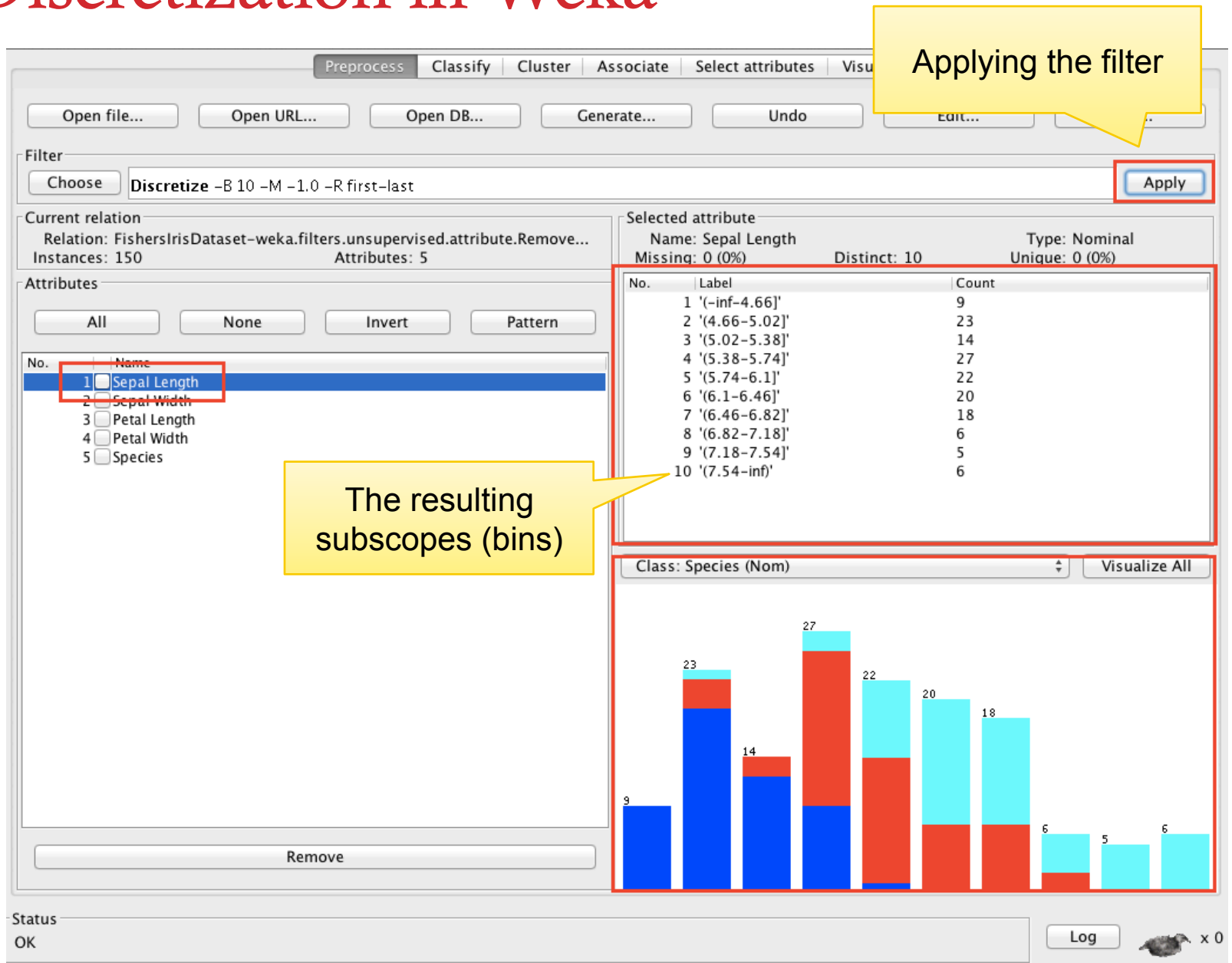
Discretization in Weka

Equal-width binning is the default option.

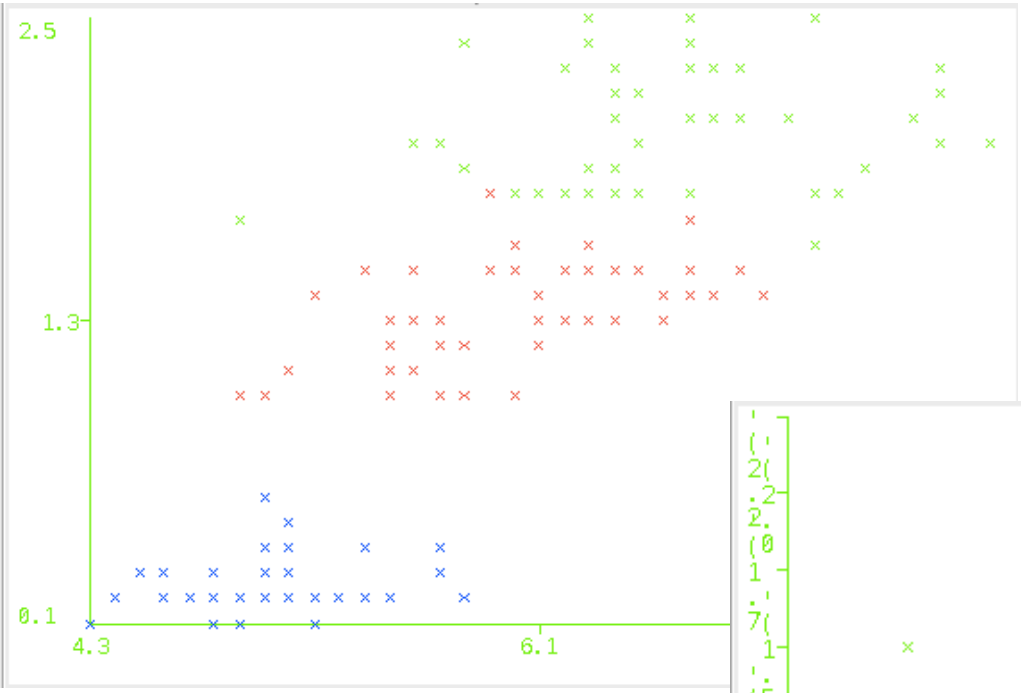
- ***attributeIndices*** – the first-last value means that we are discretizing all values. We can also name the attribute numbers
- ***bins*** - the desired number of scopes (bins)
- ***useEqualFrequency*** - *false* by default; *true* if we use Equal Frequency binning



Discretization in Weka

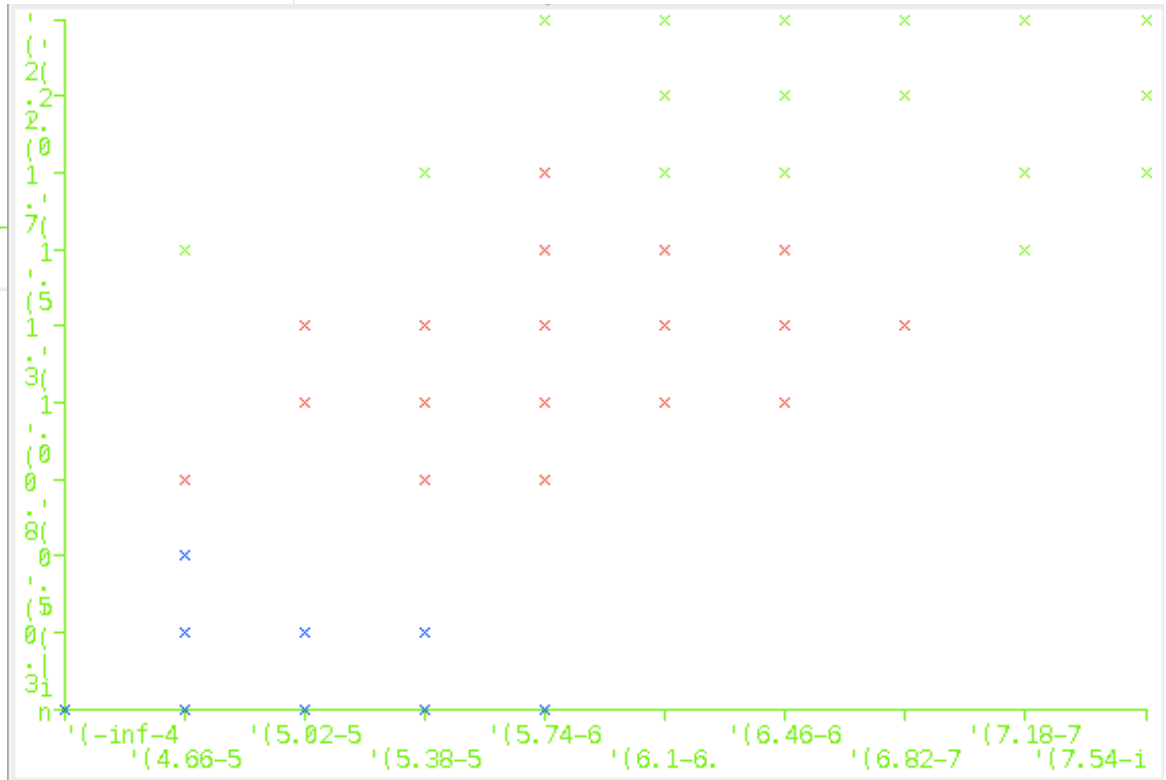


Data, before and after discretization



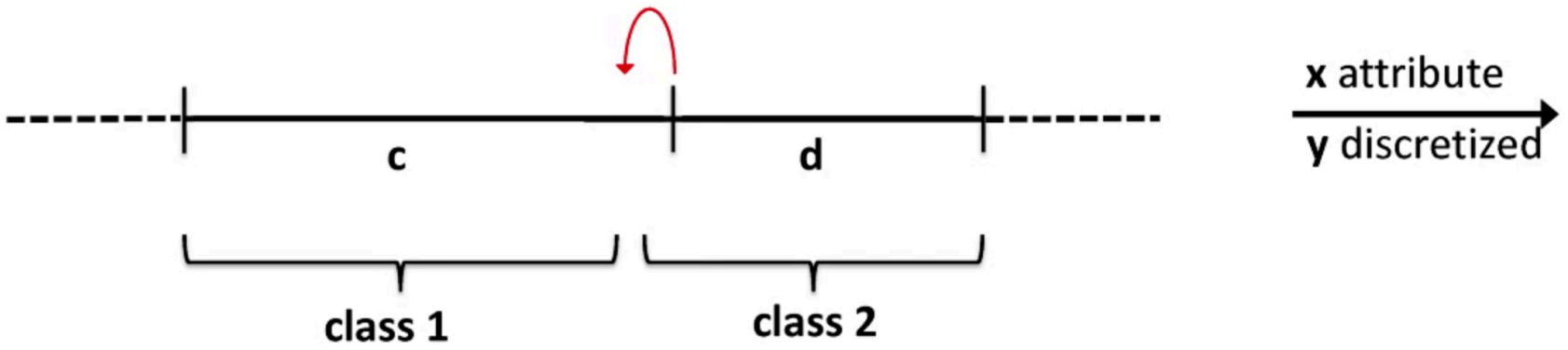
Before

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

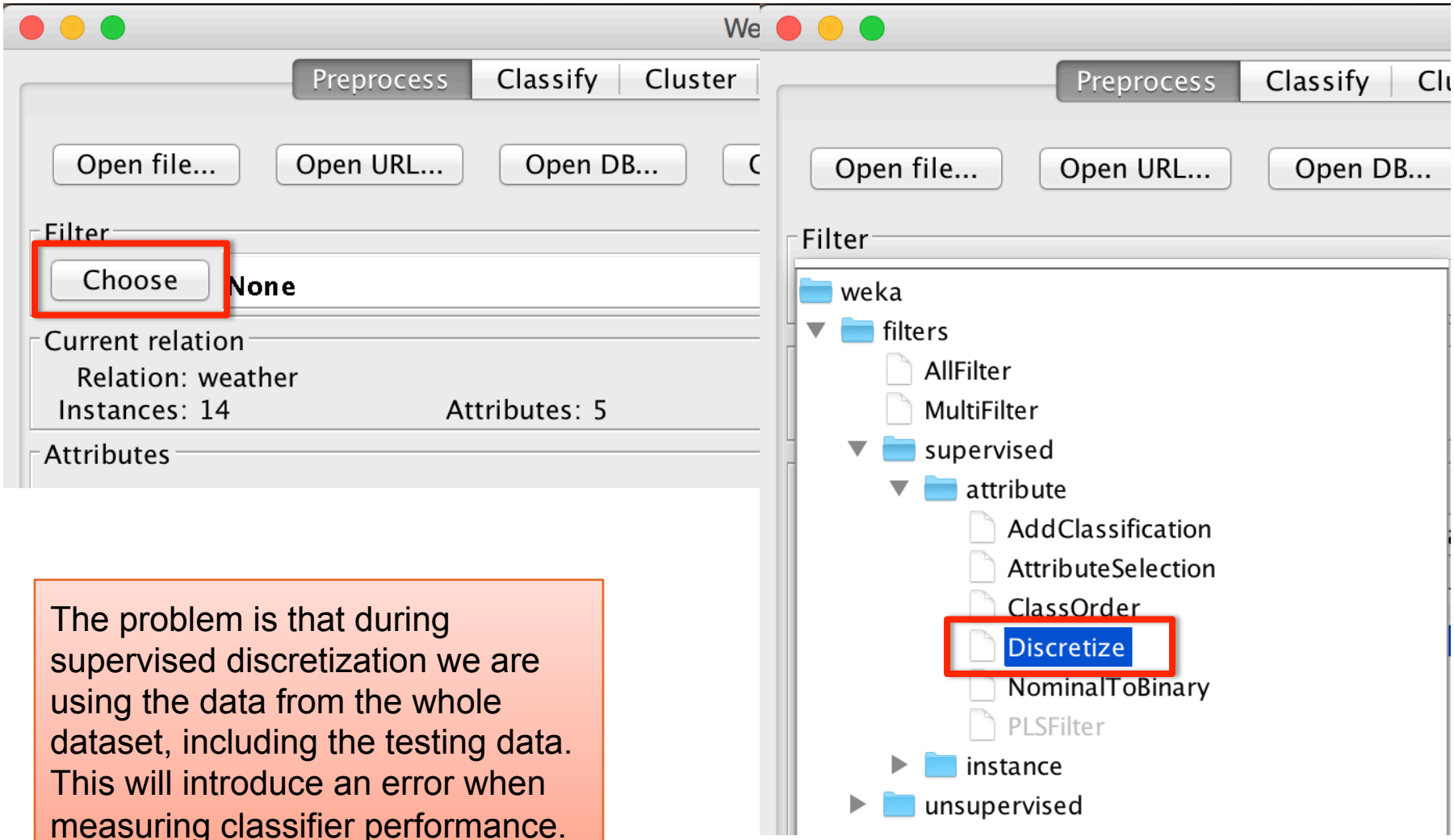
4 yes, 1 no 5 yes, 4 no
entropy = 0.934 bits

- 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

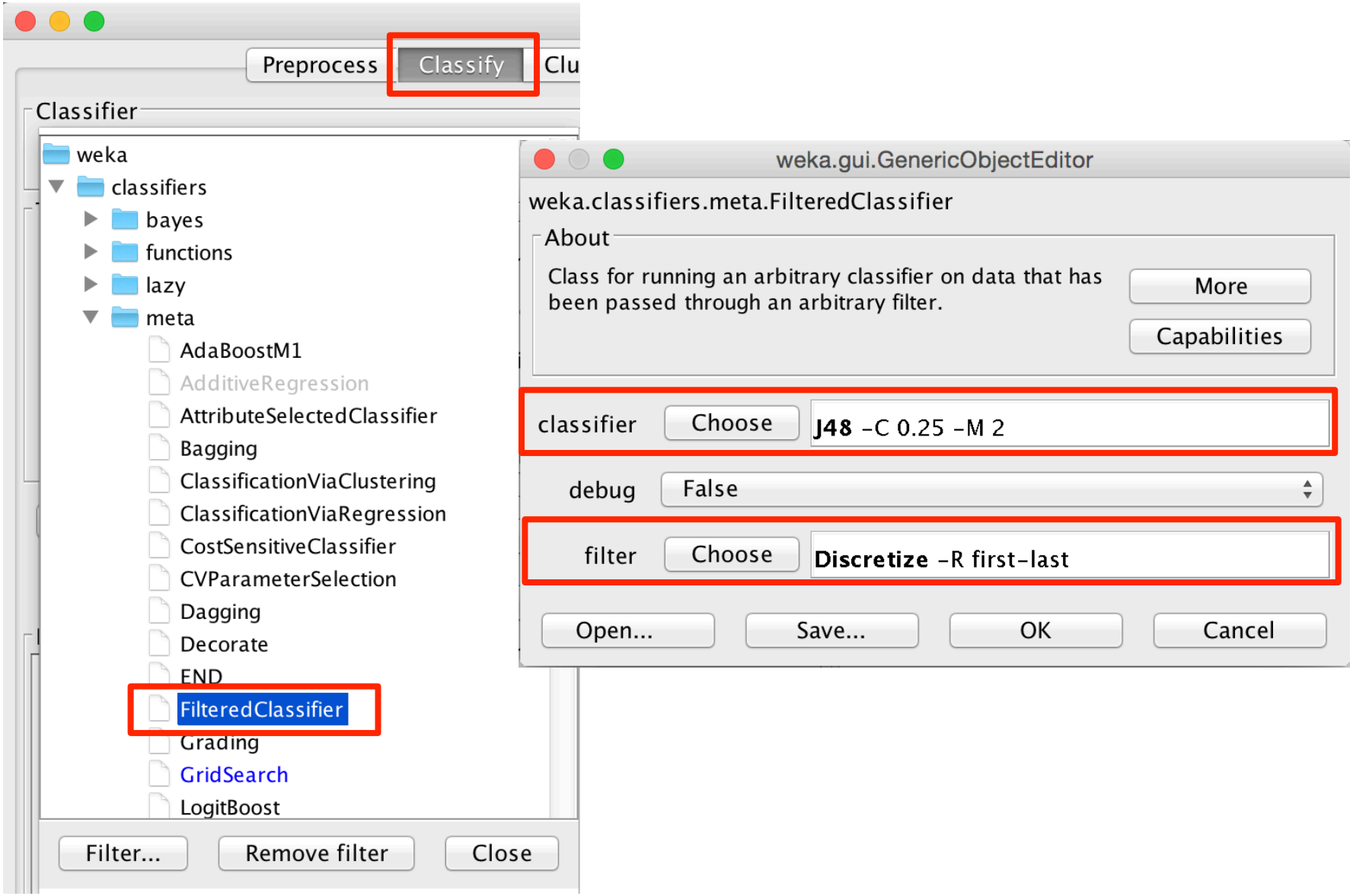
Supervised discretization in Weka

weather.numeric.arff



The 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.

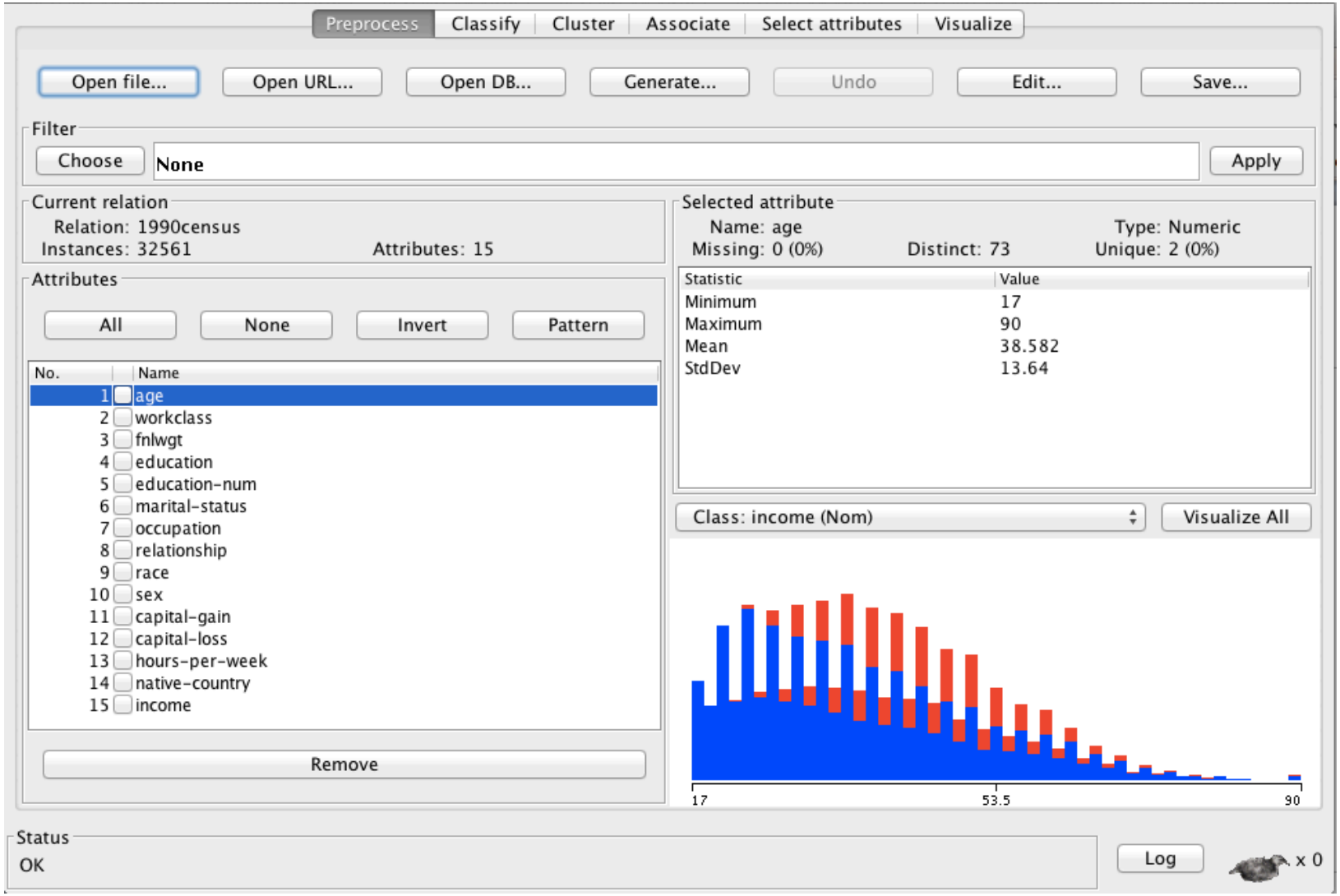
Approaches to Attribute Selection

Two approaches:

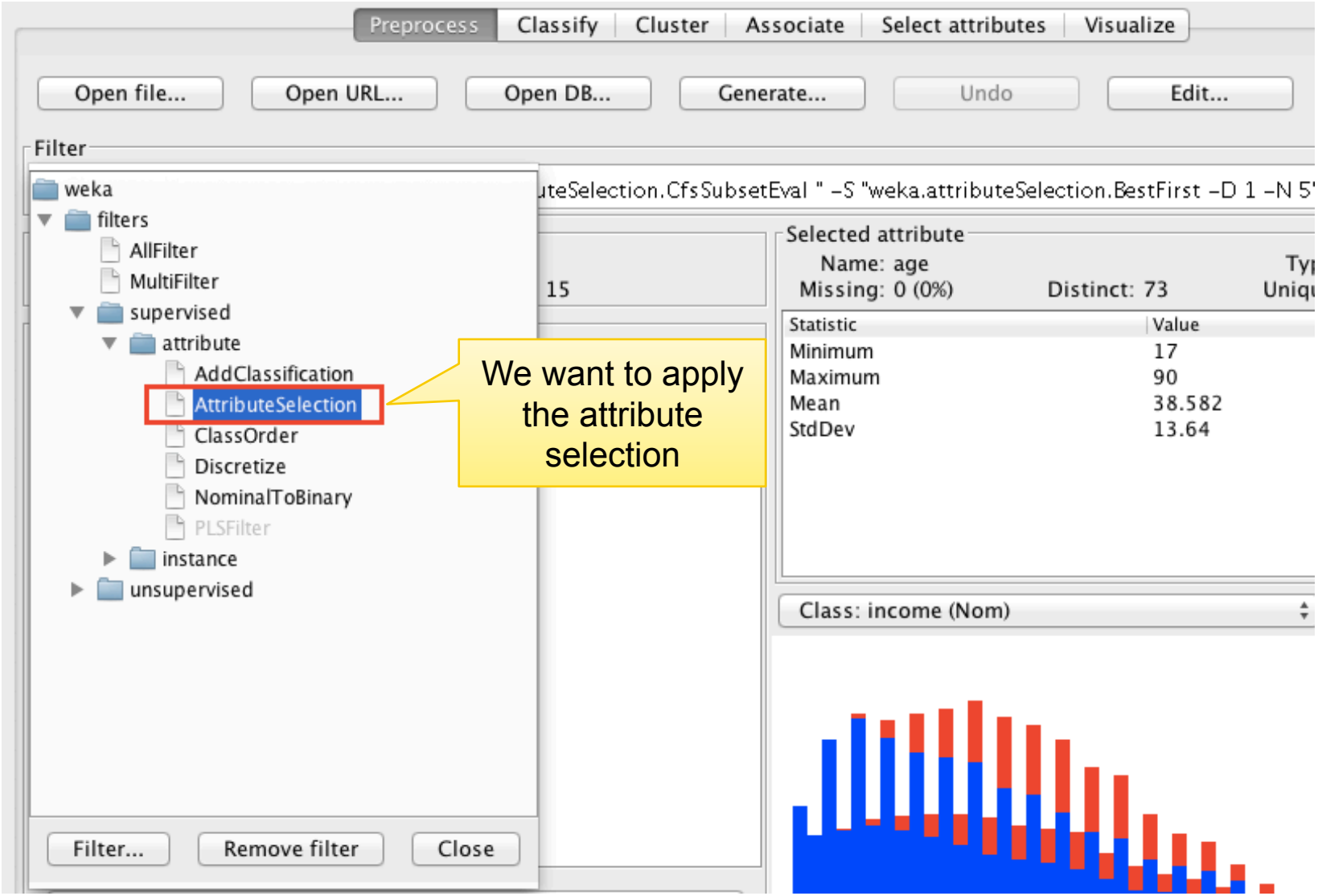
- *Filter method* – use the approximation based on the general features of the data.
- *Wrapper method* – attribute subsets are being evaluated by using the machine learning algorithm, applied to the dataset. The name Wrapper comes from the fact that the algorithm is wrapped within the process of selection.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

Attribute Selection Example

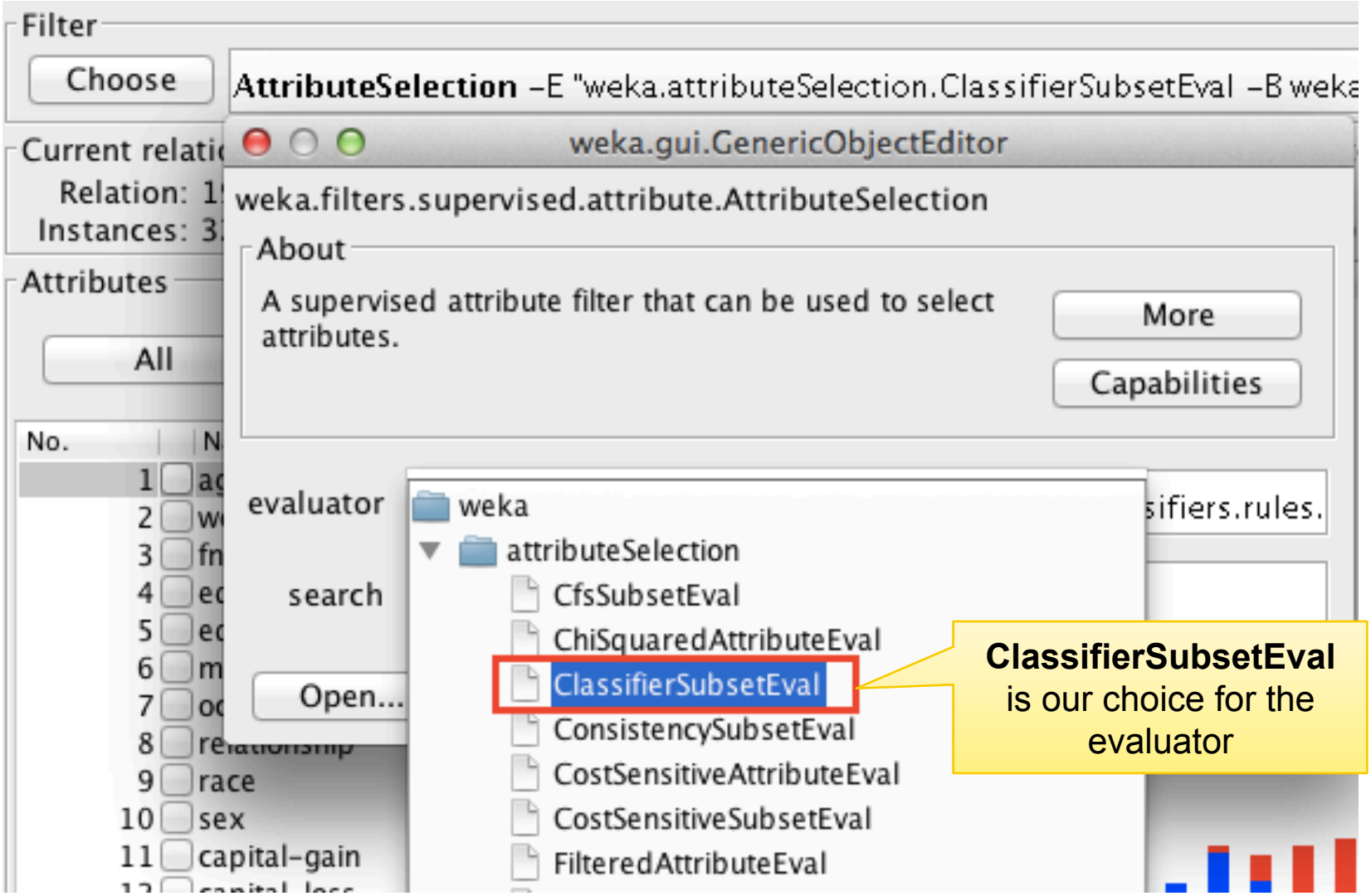
census90-income.arff



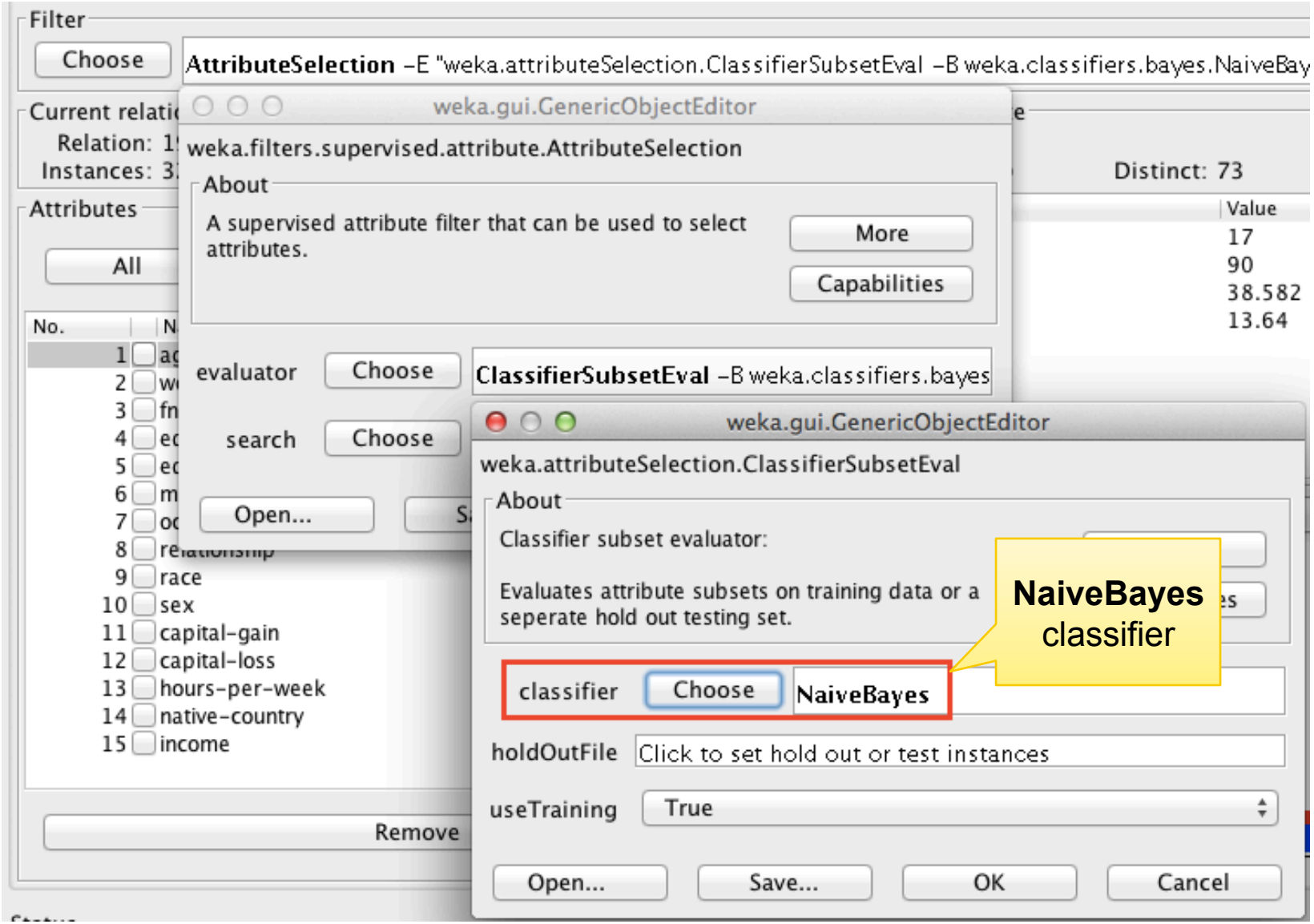
Attribute Selection Example



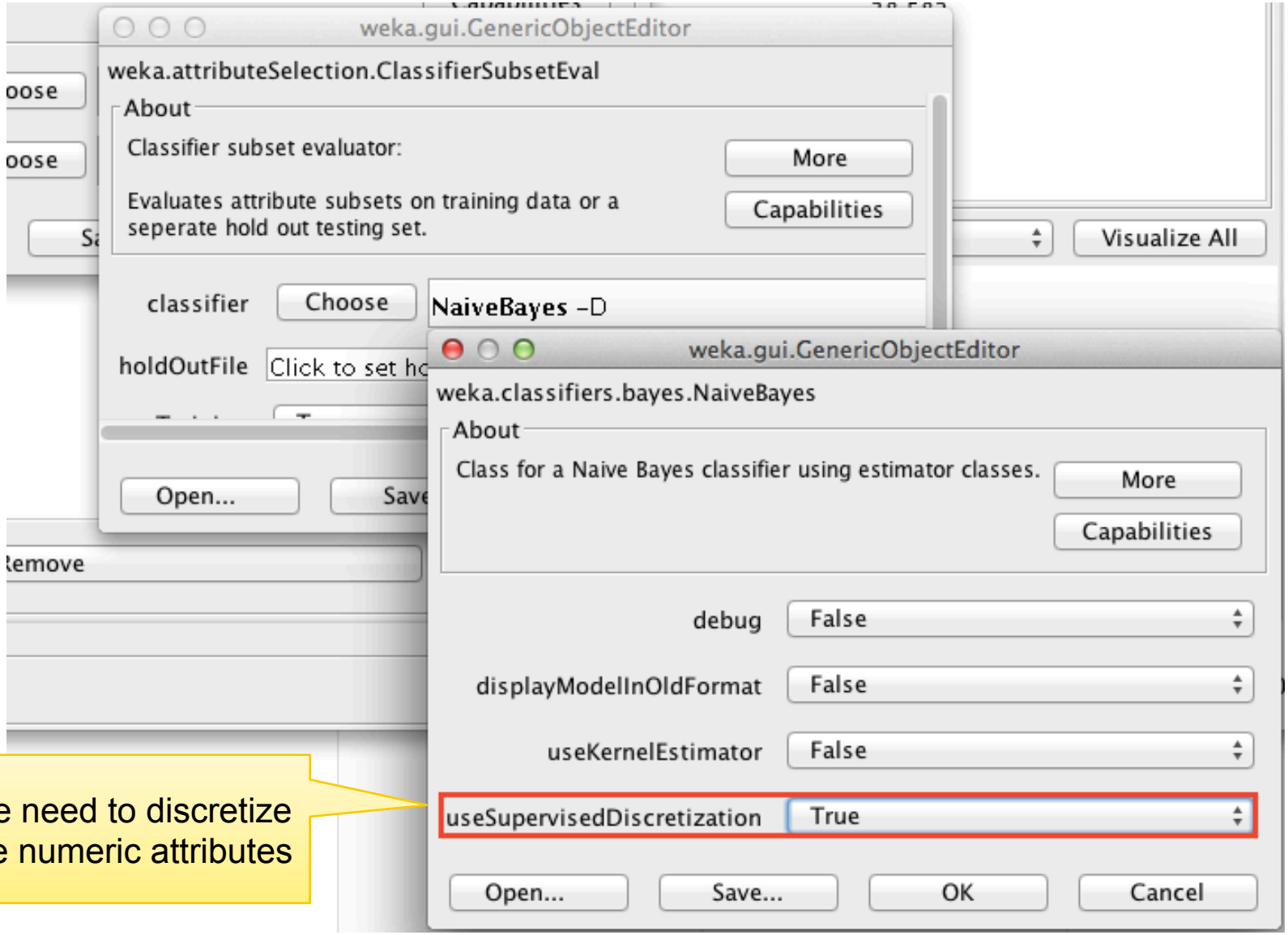
Attribute Selection Example



Attribute Selection Example

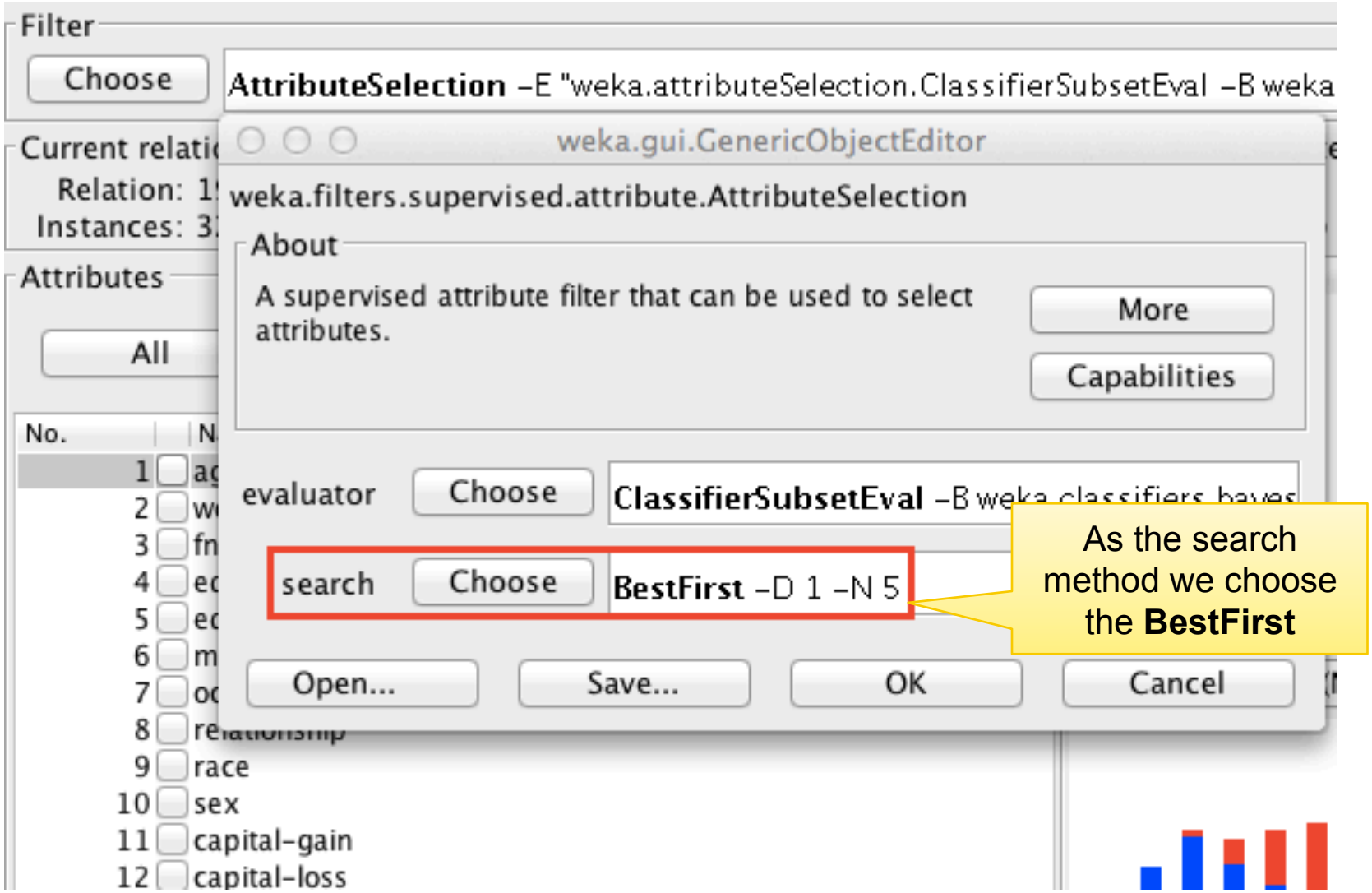


Attribute Selection Example

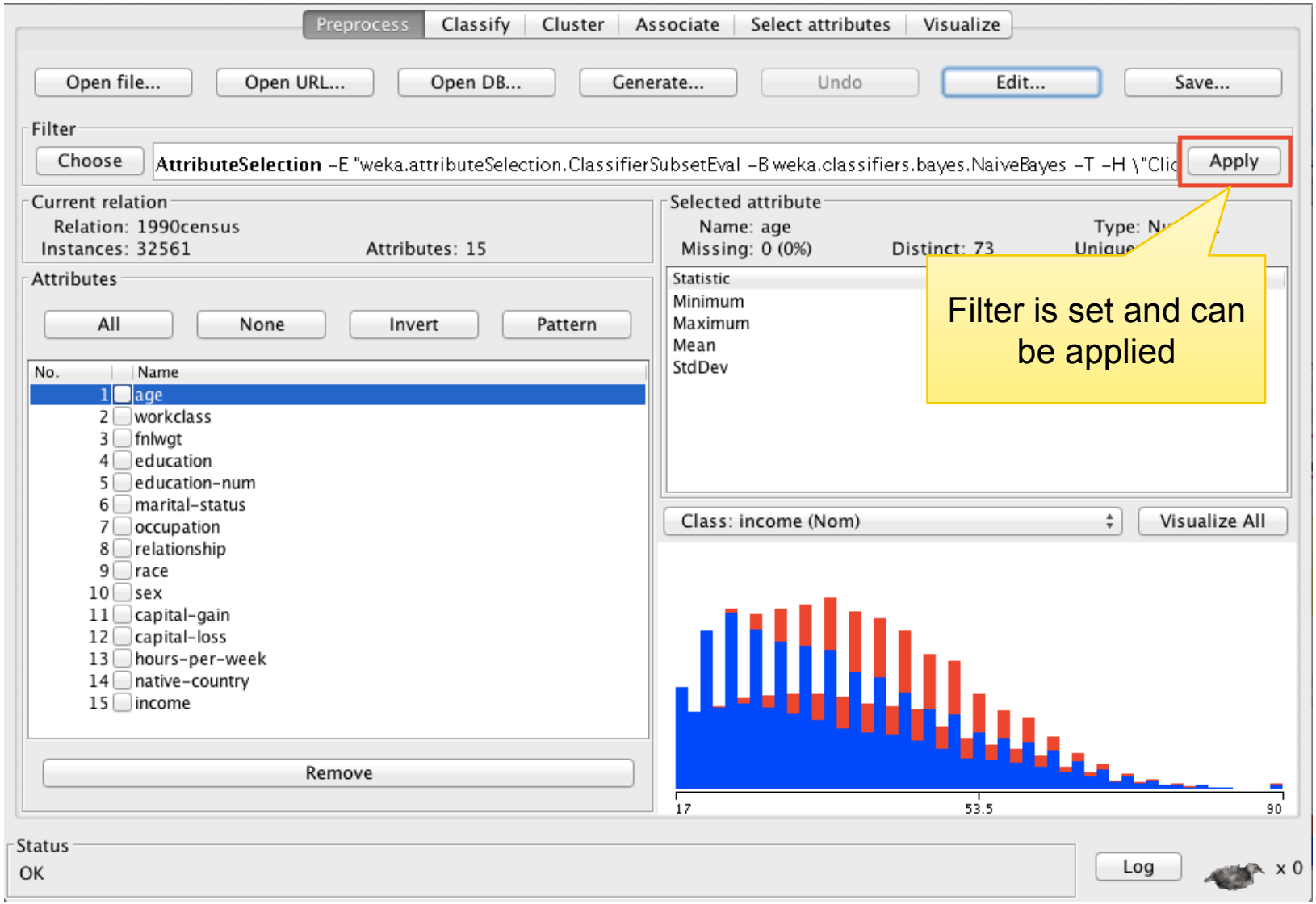


We need to discretize the numeric attributes

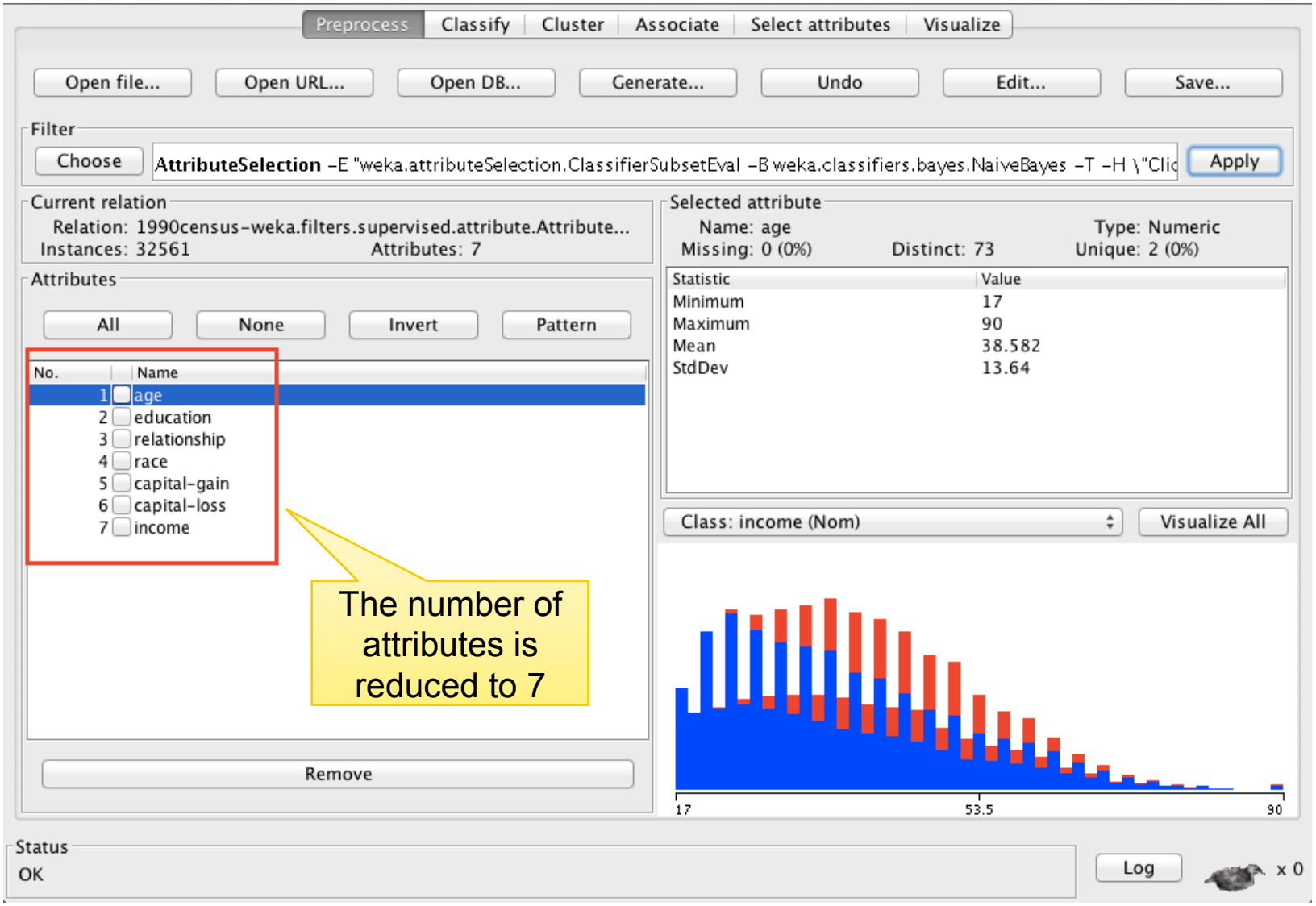
Attribute Selection Example



Attribute Selection Example

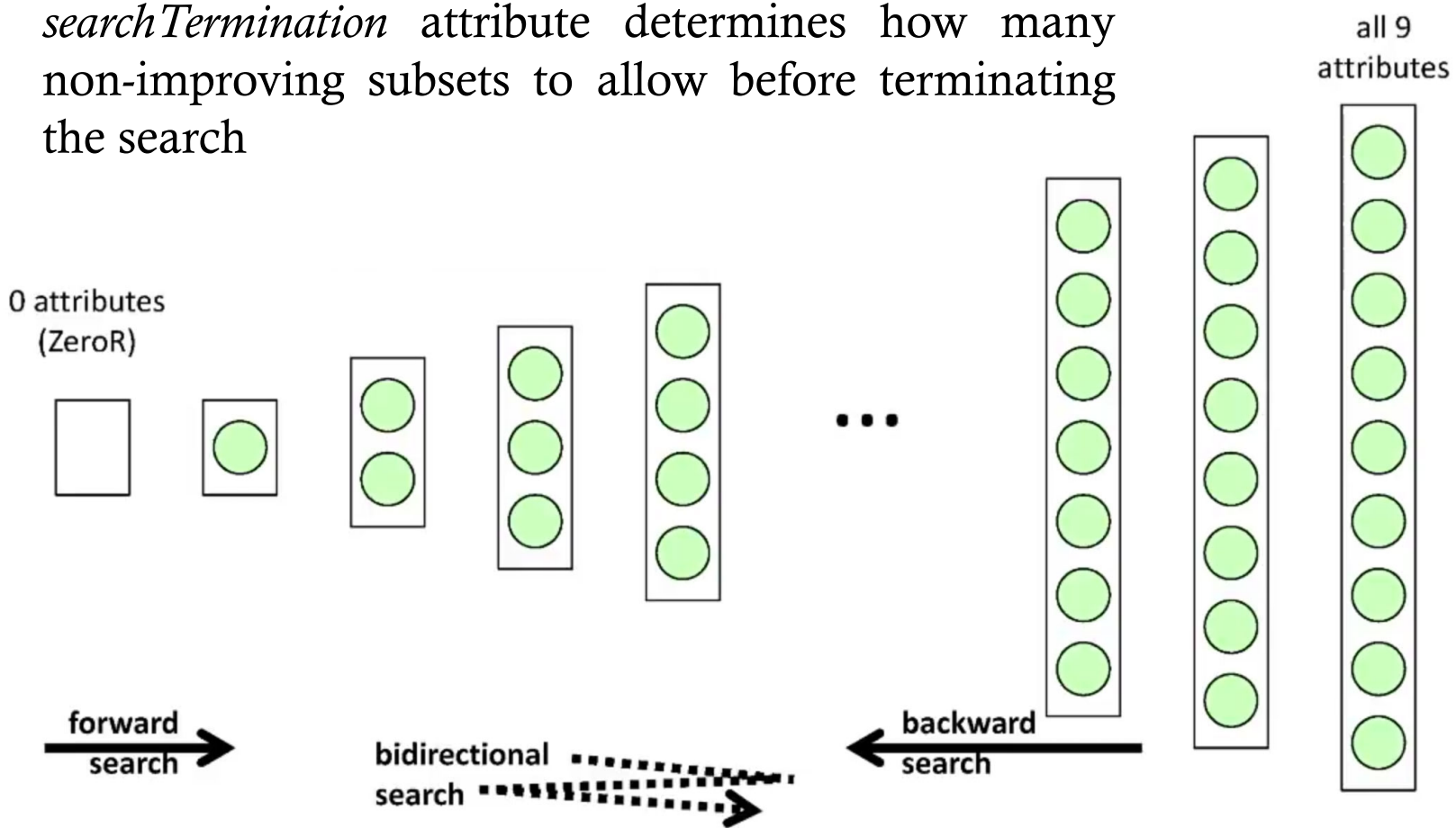


Attribute Selection Example



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