How to Perform Feature Selection With Machine Learning Data in Weka

by **Jason Brownlee** on July 13, 2016 in **Weka Machine Learning**

Last Updated on December 13, 2019

Raw machine learning data contains a mixture of attributes, some of which are relevant to making predictions.

How do you know which features to use and which to remove? The process of selecting features in your data to model your problem is called feature selection.

In this post you will discover how to perform feature selection with your machine learning data in Weka.

After reading this post you will know:

- About the importance of feature selection when working through a machine learning problem.
- How feature selection is supported on the Weka platform.
- How to use various different feature selection techniques in Weka on your dataset.

Discover how to prepare data, fit models, and evaluate their predictions, all without writing a line of code in my new book, with 18 step-by-step tutorials and 3 projects with Weka.

Let's get started.

- **Update March/2018**: Added alternate link to download the dataset as the original appears to have been taken down.

How to Perform Feature Selection With Machine Learning Data in Weka
Photo by Peter Gronemann, some rights reserved.

Predict the Onset of Diabetes

The dataset used for this example is the Pima Indians onset of diabetes dataset.

It is a classification problem where each instance represents medical details for one patient and the task is to predict whether the patient will have an onset of diabetes within the next five years.

You can learn more about the dataset here:

- Dataset File.
- Dataset Details

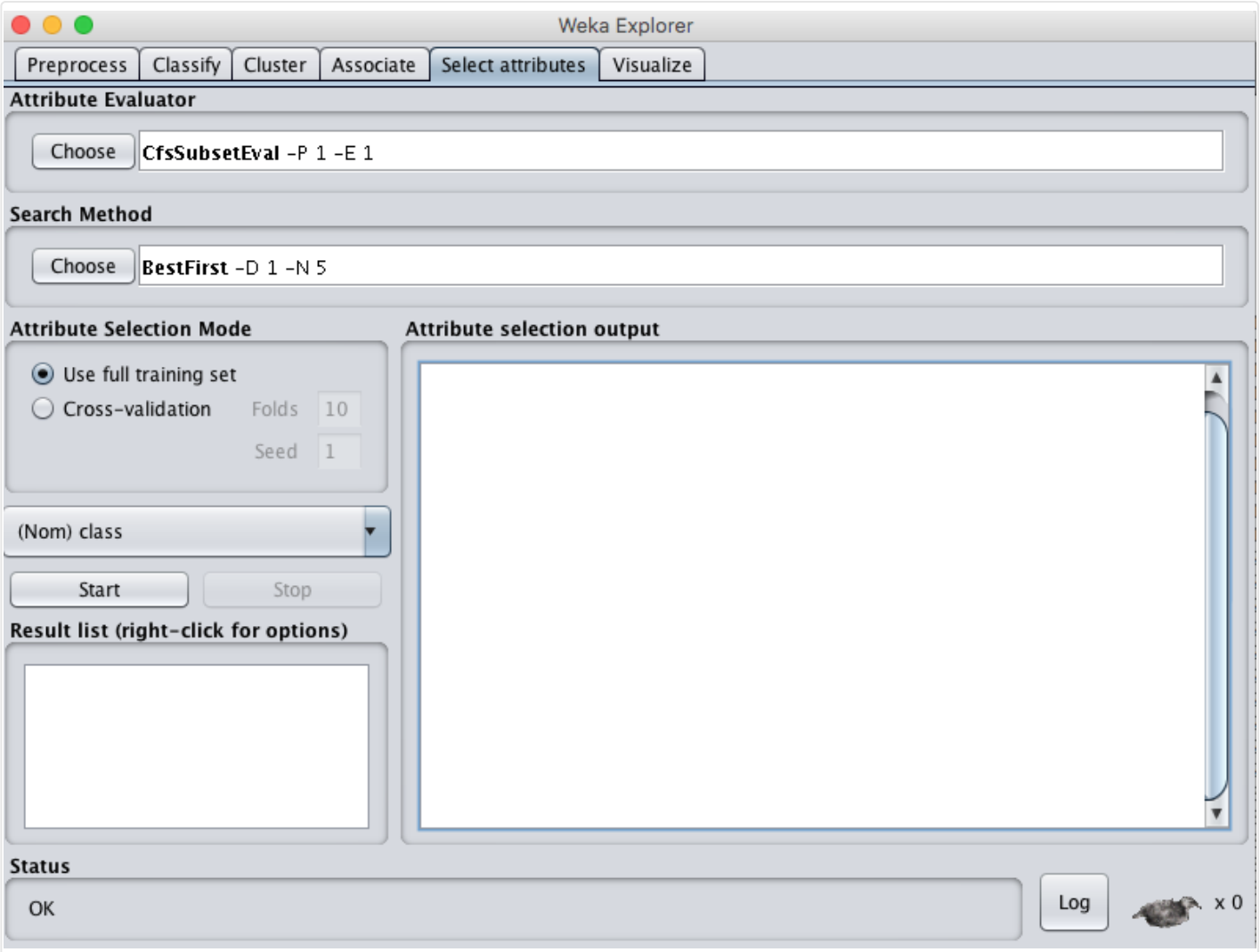
You can also access this dataset in your Weka installation, under the *data/* directory in the file called *diabetes.arff*.

Feature Selection in Weka

Many feature selection techniques are supported in Weka.

A good place to get started exploring feature selection in Weka is in the Weka Explorer.

1. Open the Weka GUI Chooser.
2. Click the “Explorer” button to launch the Explorer.
3. Open the Pima Indians dataset.
4. Click the “Select attributes” tab to access the feature selection methods.



Weka Feature Selection

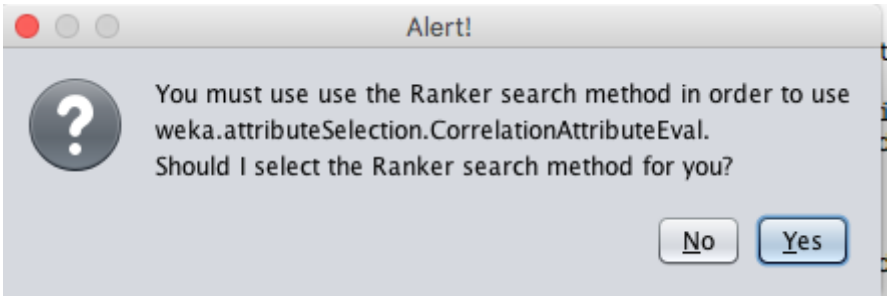
Feature selection is divided into two parts:

- Attribute Evaluator
- Search Method.

Each section has multiple techniques from which to choose.

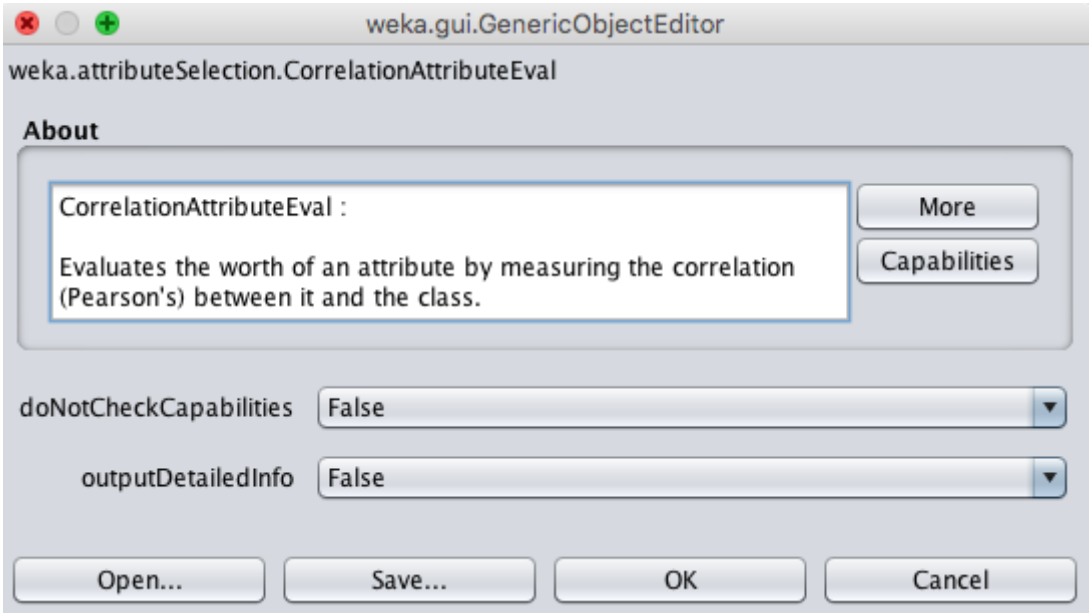
The attribute evaluator is the technique by which each attribute in your dataset (also called a column or feature) is evaluated in the context of the output variable (e.g. the class). The search method is the technique by which to try or navigate different combinations of attributes in the dataset in order to arrive on a short list of chosen features.

Some Attribute Evaluator techniques require the use of specific Search Methods. For example, the CorrelationAttributeEval technique used in the next section can only be used with a Ranker Search Method, that evaluates each attribute and lists the results in a rank order. When selecting different Attribute Evaluators, the interface may ask you to change the Search Method to something compatible with the chosen technique.



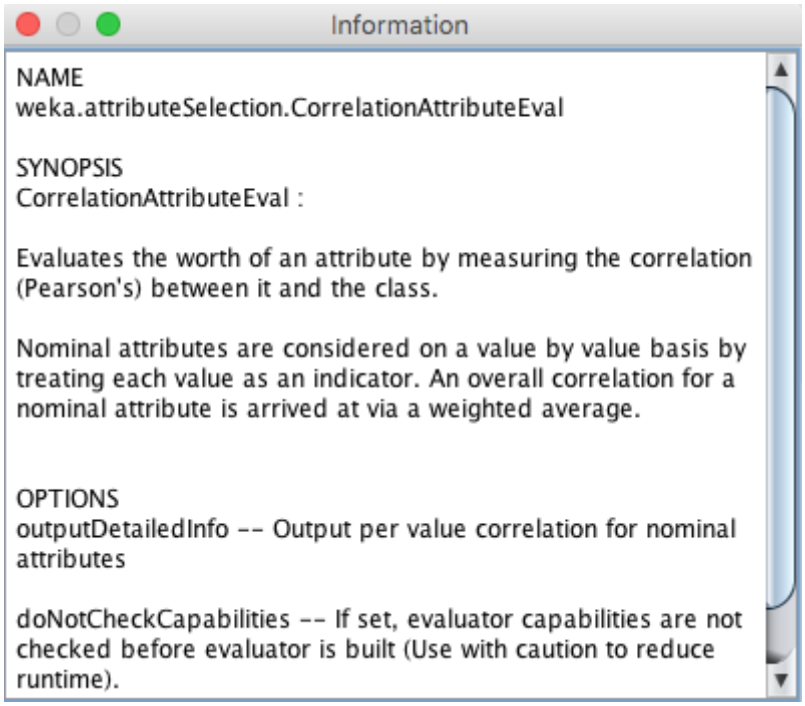
Weka Feature Selection Alert

Both the Attribute Evaluator and Search Method techniques can be configured. Once chosen, click on the name of the technique to get access to its configuration details.



Weka Feature Selection Configuration

Click the “More” button to get more documentation on the feature selection technique and configuration parameters. Hover your mouse cursor over a configuration parameter to get a tooltip containing more details.



Weka Feature Selection More Information

Now that we know how to access feature selection techniques in Weka, let's take a look at how to use some popular methods on our chosen standard dataset.

Correlation Based Feature Selection

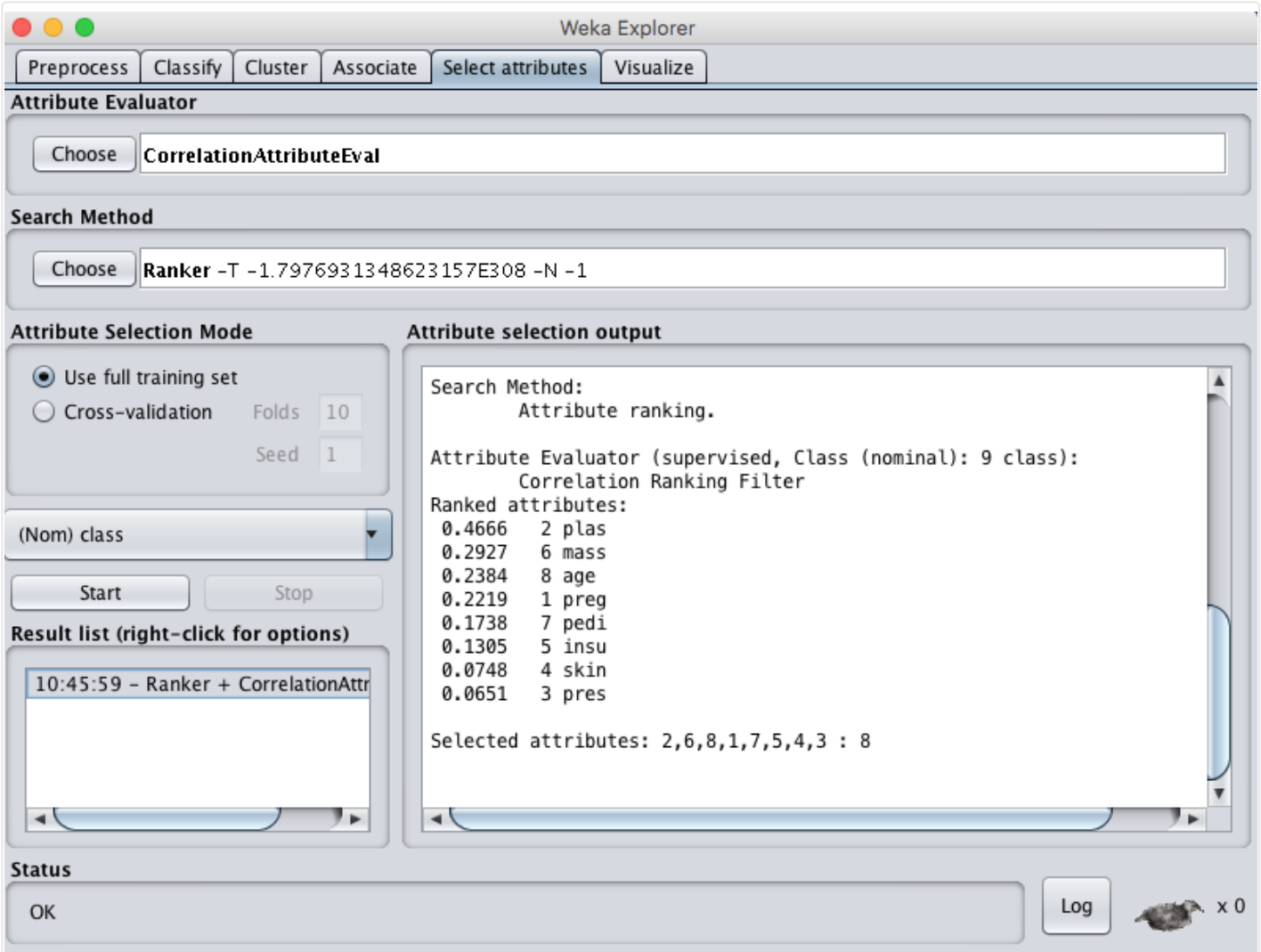
A popular technique for selecting the most relevant attributes in your dataset is to use correlation.

Correlation is more formally referred to as Pearson's correlation coefficient in statistics.

You can calculate the correlation between each attribute and the output variable and select only those attributes that have a moderate-to-high positive or negative correlation (close to -1 or 1) and drop those attributes with a low correlation (value close to zero).

Weka supports correlation based feature selection with the CorrelationAttributeEval technique that requires use of a Ranker search method.

Running this on our Pima Indians dataset suggests that one attribute (plas) has the highest correlation with the output class. It also suggests a host of attributes with some modest correlation (mass, age, preg). If we use 0.2 as our cut-off for relevant attributes, then the remaining attributes could possibly be removed (pedi, insu, skin and pres).



Weka Correlation-Based Feature Selection Method

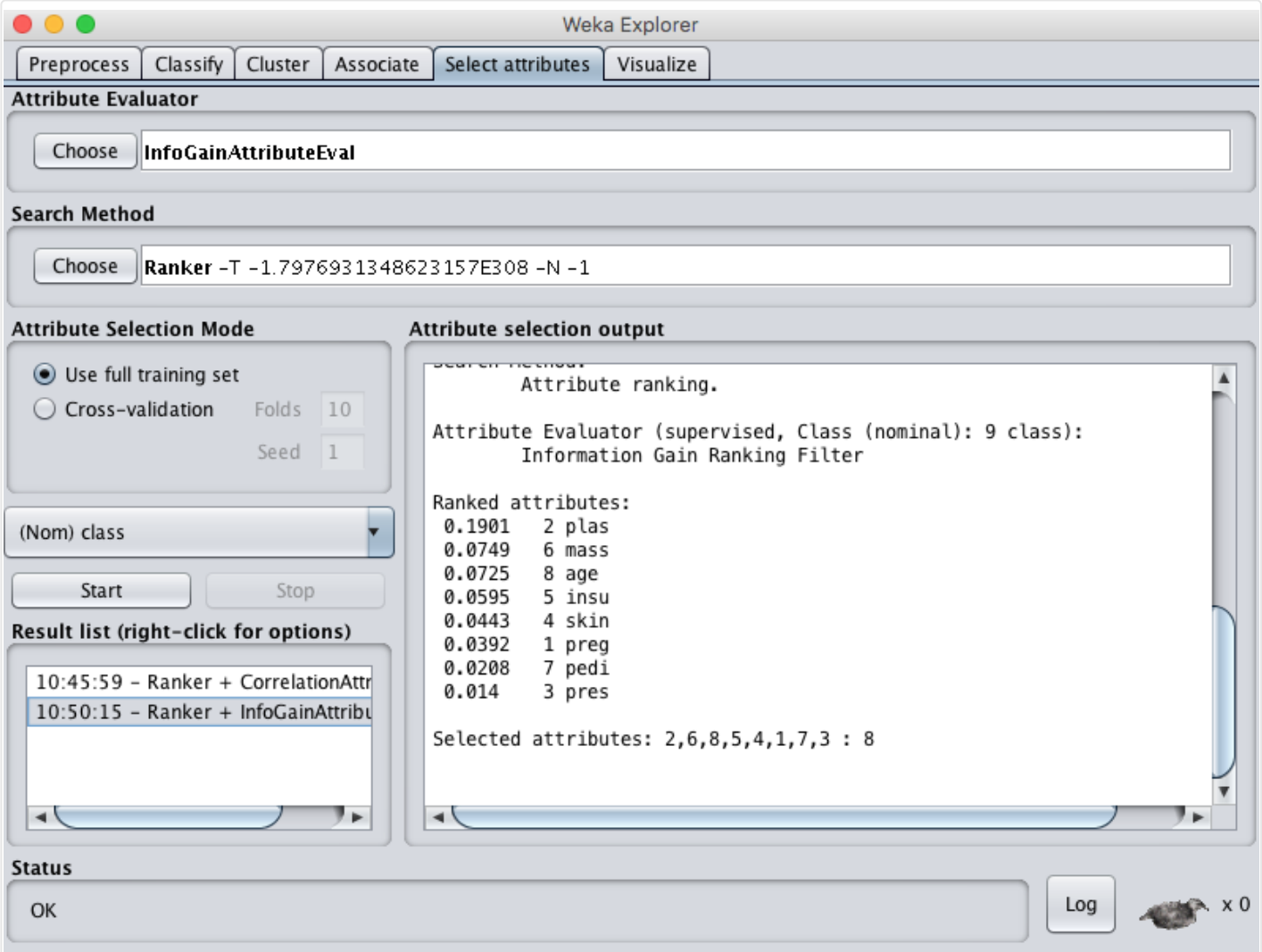
Information Gain Based Feature Selection

Another popular feature selection technique is to calculate the information gain.

You can calculate the information gain (also called entropy) for each attribute for the output variable. Entry values vary from 0 (no information) to 1 (maximum information). Those attributes that contribute more information will have a higher information gain value and can be selected, whereas those that do not add much information will have a lower score and can be removed.

Weka supports feature selection via information gain using the InfoGainAttributeEval Attribute Evaluator. Like the correlation technique above, the Ranker Search Method must be used.

Running this technique on our Pima Indians we can see that one attribute contributes more information than all of the others (plas). If we use an arbitrary cutoff of 0.05, then we would also select the mass, age and insu attributes and drop the rest from our dataset.



Weka Information Gain-Based Feature Selection Method

Learner Based Feature Selection

A popular feature selection technique is to use a generic but powerful learning algorithm and evaluate the performance of the algorithm on the dataset with different subsets of attributes selected.

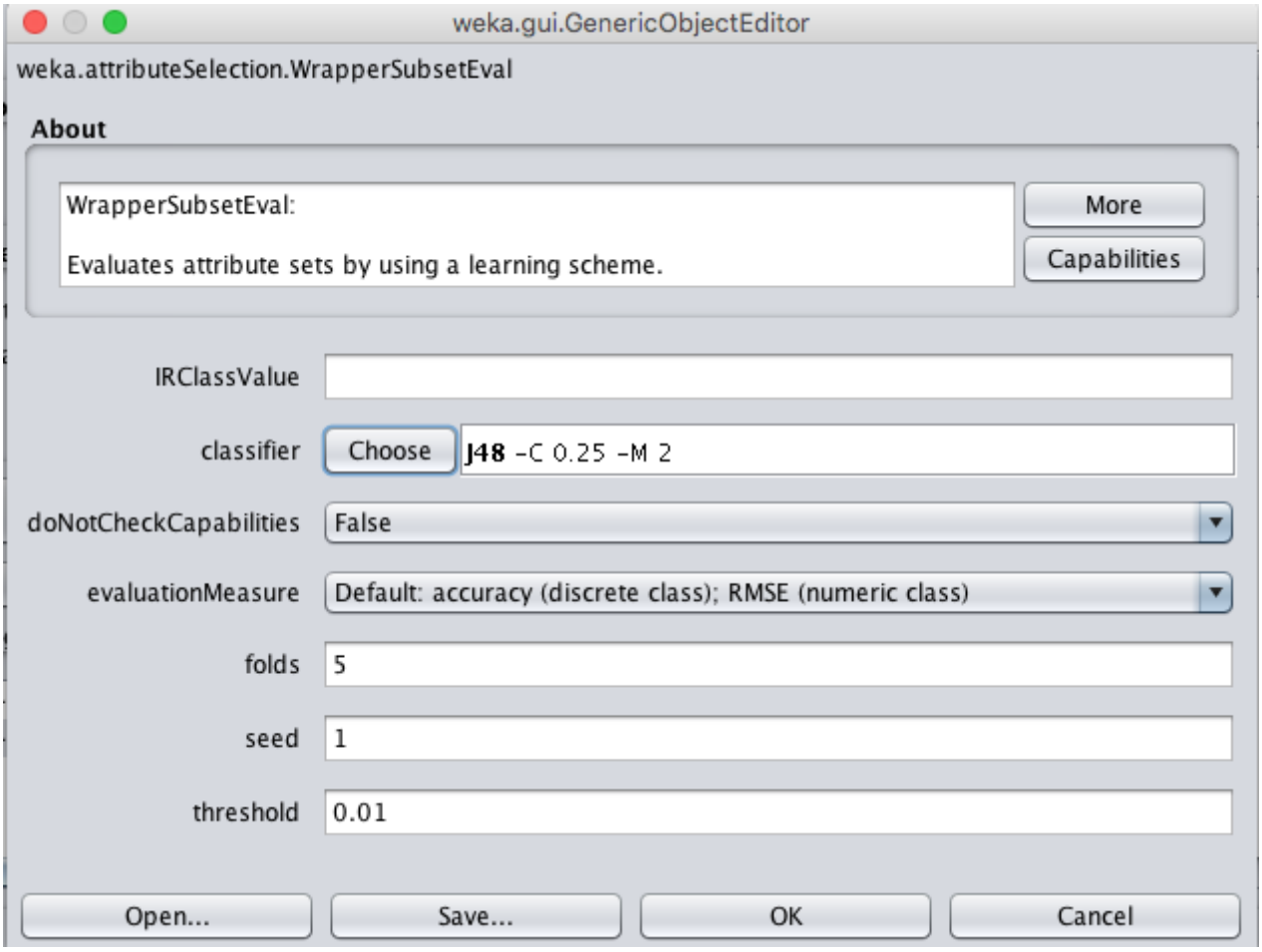
The subset that results in the best performance is taken as the selected subset. The algorithm used to evaluate the subsets does not have to be the algorithm that you intend to use to model your problem, but it should be generally quick to train and powerful, like a decision tree method.

In Weka this type of feature selection is supported by the WrapperSubsetEval technique and must use a GreedyStepwise or BestFirst Search Method. The latter, BestFirst, is preferred if you can spare the compute time.

1. First select the “WrapperSubsetEval” technique.

2. Click on the name “WrapperSubsetEval” to open the configuration for the method.

3. Click the “Choose” button for the “classifier” and change it to J48 under “trees”.



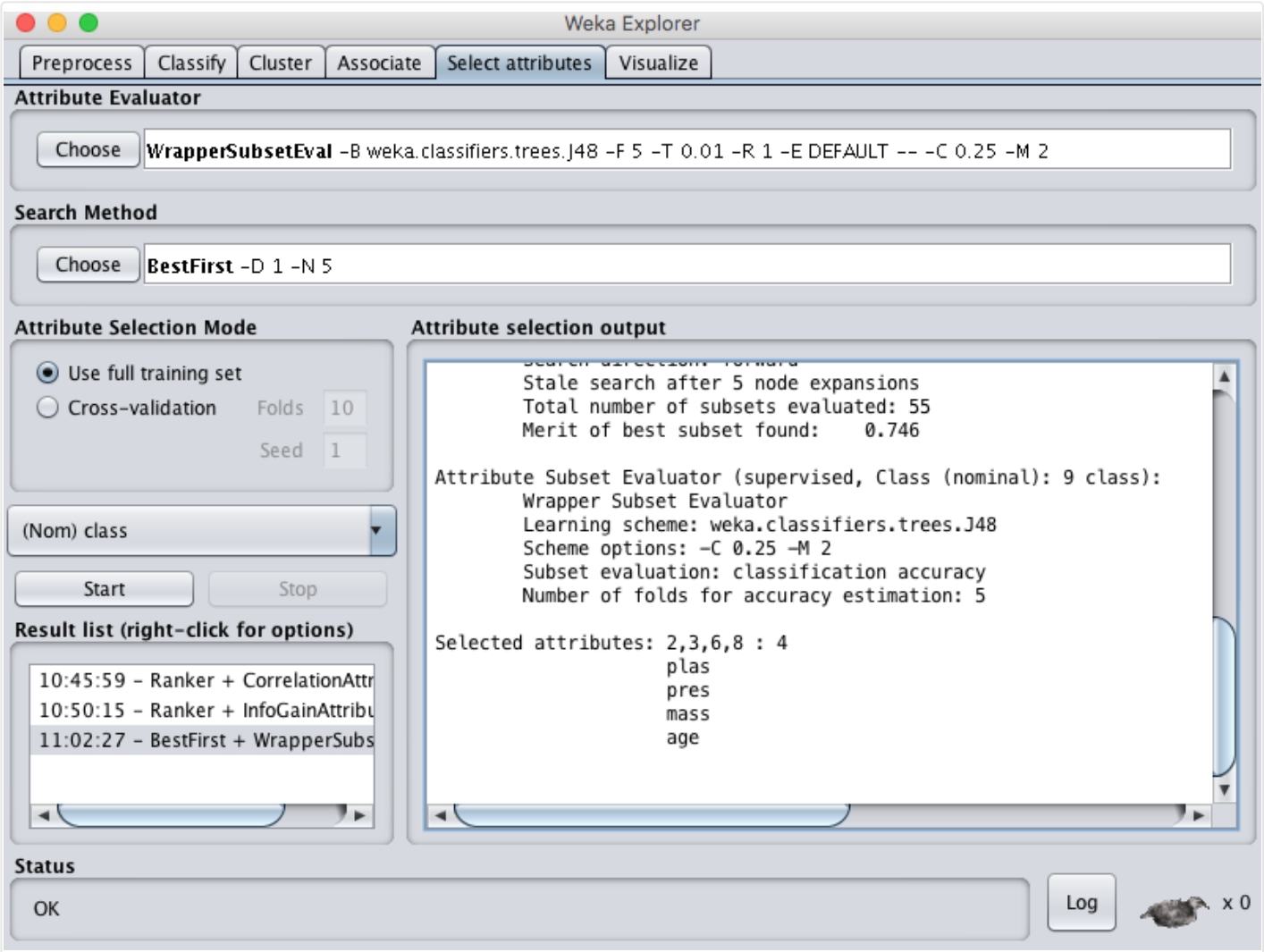
Weka Wrapper Feature Selection Configuration

4. Click “OK” to accept the configuration.

5. Change the “Search Method” to “BestFirst”.

6. Click the “Start” button to evaluate the features.

Running this feature selection technique on the Pima Indians dataset selects 4 of the 8 input variables: plas, pres, mass and age.



Weka Wrapper Feature Selection Method

Select Attributes in Weka

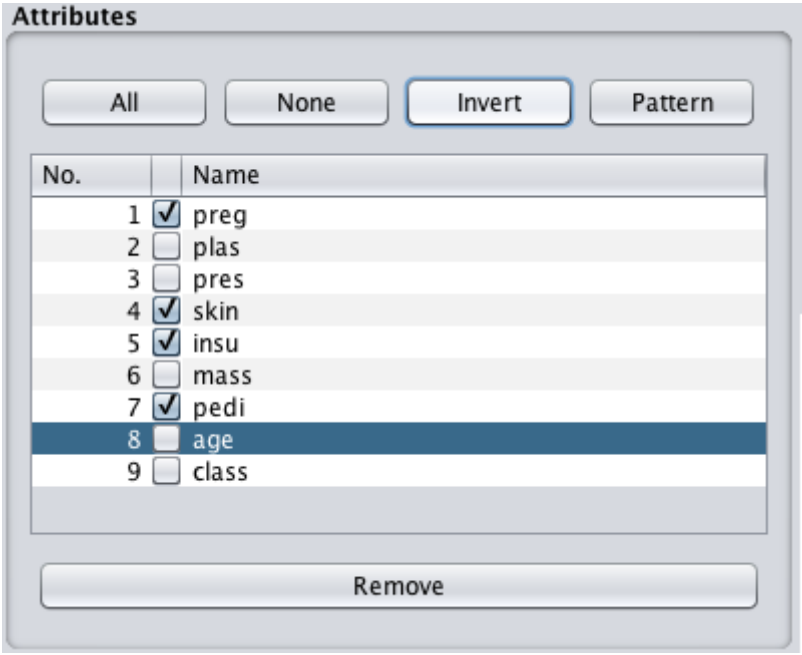
Looking back over the three techniques, we can see some overlap in the selected features (e.g. plas), but also differences.

It is a good idea to evaluate a number of different “views” of your machine learning dataset. A view of your dataset is nothing more than a subset of features selected by a given feature selection technique. It is a copy of your dataset that you can easily make in Weka.

For example, taking the results from the last feature selection technique, let’s say we wanted to create a view of the Pima Indians dataset with only the following attributes: plas, pres, mass and age:

1. Click the “Preprocess” tab.

2. In the “Attributes” selection Tick all but the plas, pres, mass, age and class attributes.

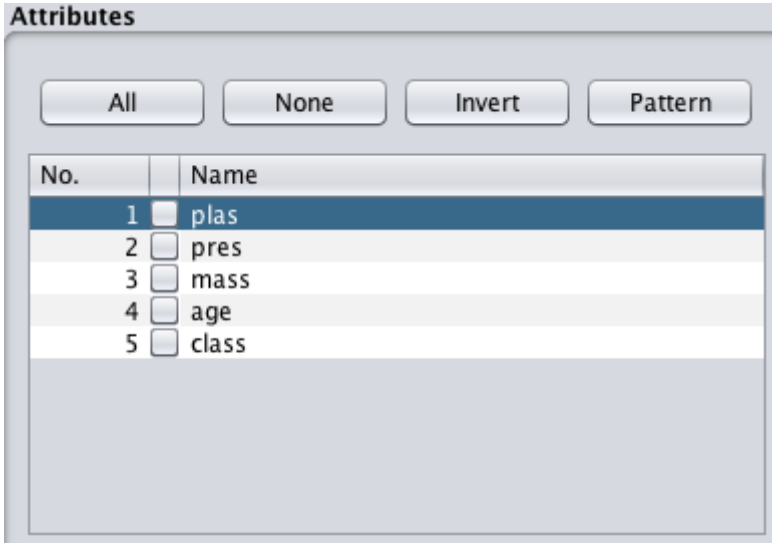


Weka Select Attributes To Remove From Dataset

3. Click the “Remove” button.

4. Click the “Save” button and enter a filename.

You now have a new view of your dataset to explore.



Weka Attributes Removed From Dataset

What Feature Selection Techniques To Use

You cannot know which views of your data will produce the most accurate models.

Therefore, it is a good idea to try a number of different feature selection techniques on your data and in turn create many different views of your data.

Select a good generic technique, like a decision tree, and build a model for each view of your data.

Compare the results to get an idea of which view of your data results in the best performance. This will give you an idea of the view or more specifically features that best expose the structure of your problem

to learning algorithms in general.

Summary

In this post you discovered the importance of feature selection and how to use feature selection on your data with Weka.

Specifically, you learned:

- How to perform feature selection using correlation.
- How to perform feature selection using information gain.
- How to perform feature selection by training a model on different subsets of features.