J48 Classifier Parameters

Overview

Very similar to the commercial C4.5, this classifier creates a decision tree to predict class membership. Decisions trees are also sometimes called classification trees when they are used to classify nominal target values, or regression trees when they are used to predict a numeric value. Decision trees are simple to understand and interpret, and require less data preparation than many other types of classifiers. They are also generally fast, even on large data sets, and can create non-linear decision boundaries that fit data very well. In fact, overfitting is a prime risk with decision trees so it is vital to cross validate your model before deploying it to the test set. As a decision tree is grown, it naturally tends to overfit the data. Pruning is a process by which the largest tree that is the most generalizable is selected, and all the branches below that level are pruned out. This improves performance on new data significantly.

Parameters

J48 has the following parameters that can be adjusted.

- **binarySplits**
  This specifies whether to use binary splits on nominal data. This is a process by which the tree is grown by considering one nominal value versus all other nominal values instead of considering a split on each nominal value individually. This results in a tree where there are only two branches from any node. Leave this false.

- **confidenceFactor**
  This determines how aggressive the pruning process will be. The higher this value, the more 'confident' you are that the data you are learning from is a good representation of all possible events, and therefore the less pruning that will occur. Smaller values induce more pruning. This significantly affects classifier performance.

- **debug**
  If set to true, this will output information to the results pane. Changing this will not affect classifier performance.

- **minNumObj**
  This determines what the minimum number of observations are allowed at each leaf of the tree. This is another way to control overfitting. Leave this at the default value.

- **numFolds**
  This determines how much of the data will be used to prune the tree. One of the folds is held out for pruning while the rest grow the tree. The default value of

three means one third of the data is used for pruning, while two thirds are used for growing the tree. Setting this number too low will increase overfitting. Leave this at the default value.

- **reducedErrorPruning**
  Reduced error pruning is an alternative algorithm for pruning that focuses on minimizing the statistical error of the tree, instead of the misclassification rate. Leave this false.

- **saveInstanceData**
  This option saves the training data for later visualization. This will not affect classifier performance, and can be left at false.

- **seed**
  This is a random number that is important in the reduced error pruning algorithm. This should be left false if reduced error is false.

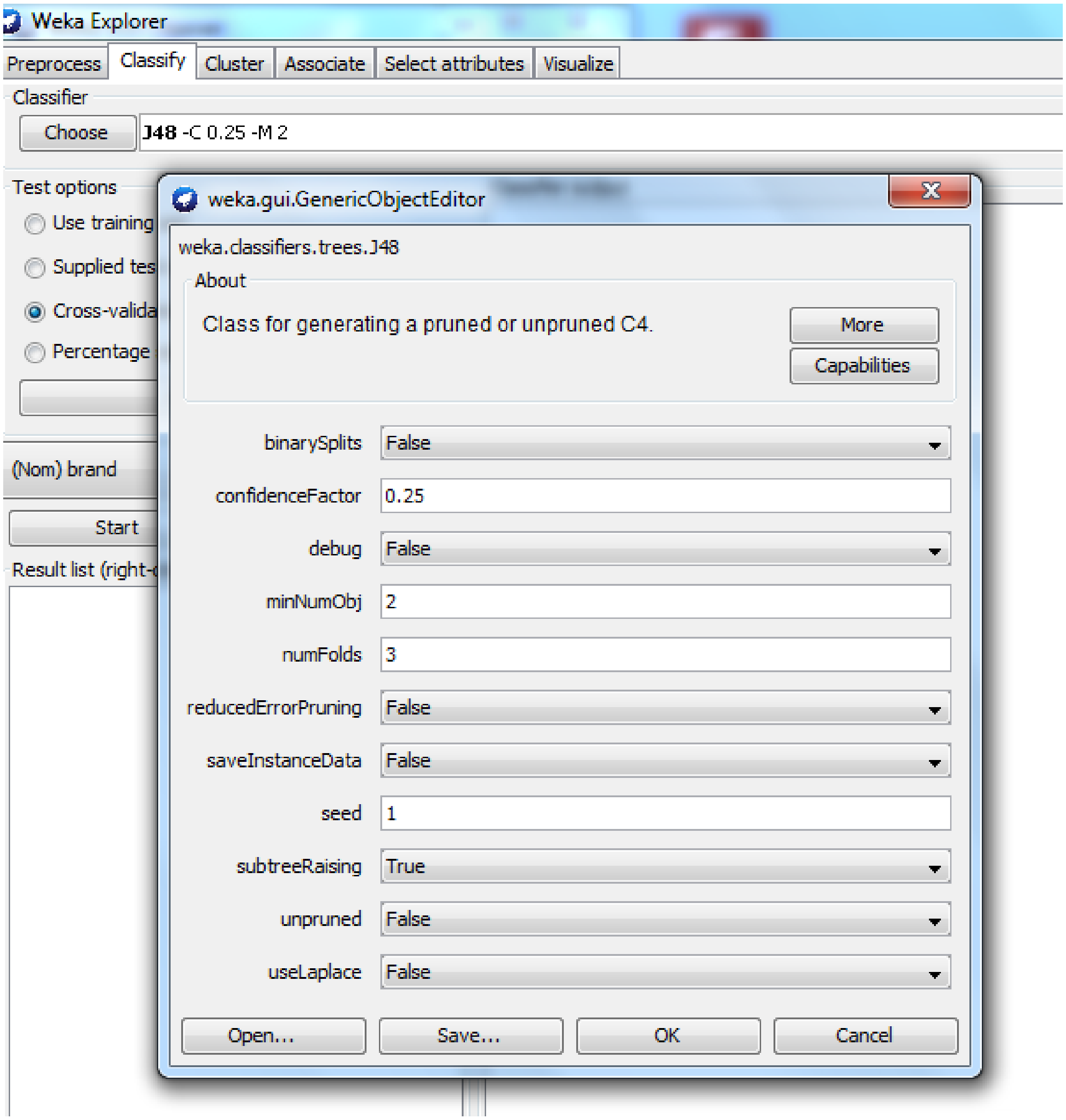
- **subtreeRaising**
  This is a specific method of pruning whereby a whole set of branches further down the tree are moved up to replace branches that were grown above it. It should remain true.

- **unpruned**
  This specifies if the tree should not be pruned. It should remain false.

- **useLaplace**
  This applies laplace smoothing to counts at the leaves. This is also sometimes called additive smoothing, and is a method by which a certain number is added to all instances in order to eliminate circumstances that are statistically undesirable, such as encountering the number zero. This is most useful when predicting probabilities, so should remain set to false here.

The parameters for all classifiers can be adjusted in the GenericObjectEditor which is accessed by clicking on the text of the classifier in the “Choose” window. Here is how the parameters described above will appear in the GenericObjectEditor.

J48 Classifier Parameters



J48 -C 0.25 -M 2

weka.classifiers.trees.J48

About

Class for generating a pruned or unpruned C4.

binarySplits	False
confidenceFactor	0.25
debug	False
minNumObj	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	True
unpruned	False
useLaplace	False