User Classifier Competition

One of WEKA's classifiers is special in that it is interactive and lets the user (i.e. you) construct their own decision tree classifier. This classifier is called UserClassifier.

To work with this classifier we want to use a new data set. Load up as training set the file 'segment-challenge.arff' in the Preprocess section. This file has 20 attributes and 7 distinct classes. All attributes except the class attribute are numeric.

For the UserClassifier it is best to have numeric attributes because they can be well represented in pixel plots. In the UserClassifier the nodes in the decision tree are not simple tests on attribute values, but are regions the user interactively selects in these plots. So if an instance lies inside the region it follows one branch of the tree, if it lies outside the region it follows the other branch. Therefore each node has only two branches going down from it.

We are running a competition to see who can build the classifier with the highest accuracy.

The aim is to come up with a tree where the final leaf nodes are as pure as possible. Remember however that you do not want to overfit the data i.e. try to select larger groups of instances.

When you decide that your tree is finished, your tree will be tested on the

test set. The accuracy measured on this test set will be your entry into the competition. If two accuracies are identical, the group first to finish, wins.

In the Classify section, set the classifier to 'trees.UserClassifier' (Click on 'Choose' button and select from the group 'trees' the classifier 'UserClassifier').

Choose as test set the file 'segment-test.arff' ('Test options' box, choose 'Supplied test set' option, click the 'Set' button, ...).

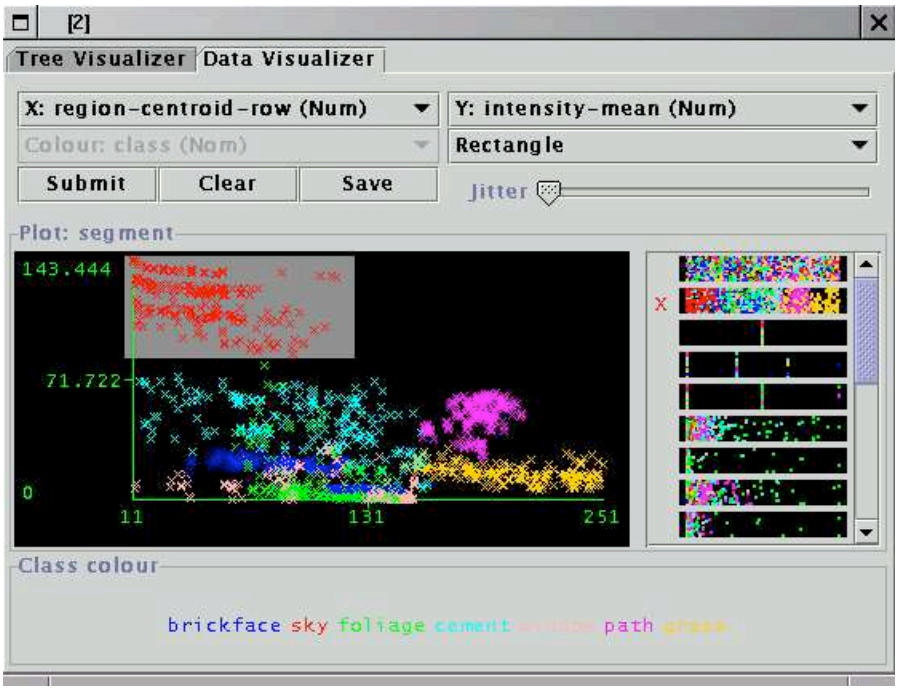
Click the **Start** button. This time a special window will appear and WEKA will wait for us to build our own classifier. You can use the tabs at the top of the window to switch between two views of your classifier. The **Tree visualizer** section shows the current state of your tree. Each node in your tree shows the counts of each type of class.

To begin with, there is only one node, the root node containing all the data. As you start to split your data, using the **Data visualizer** section, more nodes will appear in the tree.

Click the **Data visualizer** tab and you will see a 2D plot of the data. The points are colour coded by their class value. You can change the attributes used on the axes by either using the 'X:' and 'Y:' drop down menus at the top, or by left-clicking (for X) or right-clicking (for Y) on the horizontal strips to the right of the plot area. The horizontal strips show the spread of instances along each particular attribute.

You will want to try different combinations of X and Y axes to get the clearest separation you can find between the colours. Once you think you have found a good separation, it is time to select the region in the plot that will cause a branch in your tree.

Here is a hint to get you started: select 'region-centroid-row' for the X-axis and 'intensity-mean' for the Y-axis. You will see that the red class (sky) is nicely separated from the rest of the classes at the top of the plot.



You have three tools with which to select regions in the graph, which you can select in the drop down menu below the Y-axis selection menu:

1. **Rectangle** - allows you to drag rectangles around points in the graph to select them.
2. **Polygon** - allows you to build a free-form polygon with which to select points in the graph. Left-click to add vertices to the polygon, right-click to complete the polygon. The polygon will always be closed off by connecting the first point to the last.
3. **Polyline** - allows you to build a free-form polyline with which to select points in the graph. Left-click to add vertices to the polyline, right-click to complete the shape. The resulting shape will always be open as opposed to the polygon which is always closed.

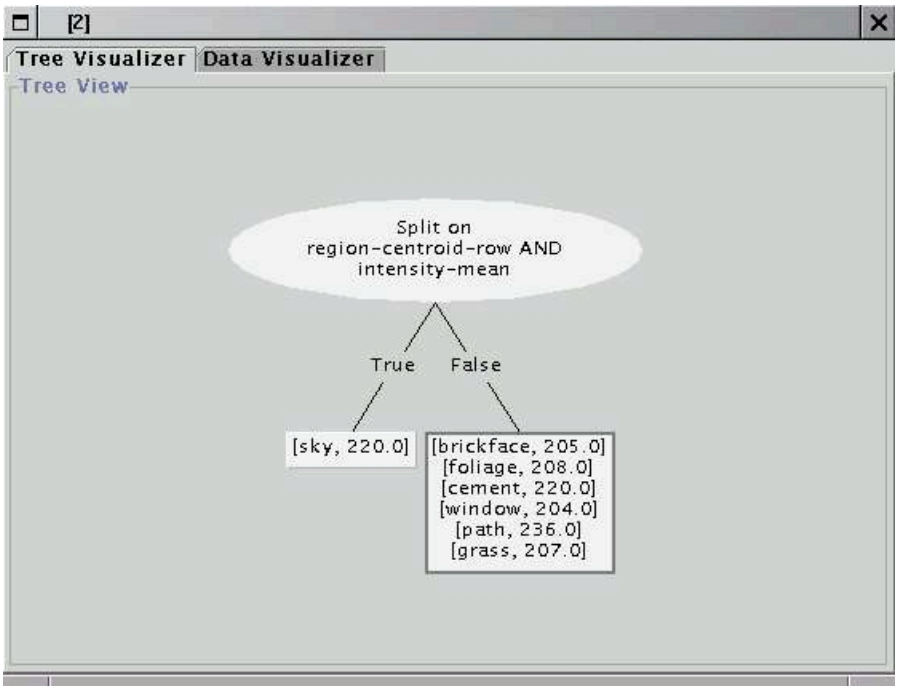
Once you have selected an area of the plot using the **Rectangle**, **Polygon** or **Polyline** tools, it will turn grey.

Clicking on the **Clear** button will remove any selection areas you have created without affecting the classifier.

When you are happy with your selection, clicking on the **Submit** button will create two new nodes in the tree—one that holds all of the instances in the selection, and one that holds all of the remaining instances.

Switch back to the **Tree visualizer** section to have a look at the changes in the tree. Clicking on different nodes on the tree will change the subset of data

that is shown in the **Data visualizer** section, and subsequently the data that you will split further by submitting a new region.



You will want to continue the process of adding nodes to the tree until you are happy that you have come up with the best separation of classes you can manage—that is, the leaf nodes in your tree are mostly pure.

When you have finished building your tree (i.e. the training of the model is finished), you will want to test the model. This you can do by right-clicking on any blank space in the **Tree visualizer** view and choosing **Accept The Tree** from the popup menu. WEKA will evaluate the tree you built against the test set and output statistics for how well you did.

Beware that as soon as you select 'Accept The Tree' you cannot go back to re-edit the same tree. You will have to start from scratch with a new tree.

4.1 The Competition

The competition is for the best accuracy score by a hand-built UserClassifier on the 'segment-challenge' data set tested on the 'segment-test' set. If two scores are ident the one finished earlier wins.

Try as many times as you like. When you think you have got a good score let one of the supervisors see it so that he or she can verify and record it (A good score is anything over 70% correct).