Matthew Barnes  
0555121

**COIS 4400 Data Mining | Assignment 2**

**Question 1.**

**Given the following training and test instances classify each test instance using the k nearest neighbor classifier for k values of 1, 2, 4 and 8. Use Euclidean distance as the distance measure. Given your results, calculate the precision, recall, and f1 measure for each value of k. Which value of k performed better? Justify your answer in terms of the metrics you calculated making certain to indicate what each metric means from a performance perspective.**

|  |  |  |  |
| --- | --- | --- | --- |
| K=1 | | **Predicted Class** | |
| A | B |
| **Actual Class** | A | 0 | 2 |
| B | 0 | 2 |
|  |  |  |  |
| K=2 | | **Predicted Class** | |
| A | B |
| **Actual Class** | A | 0 | 2 |
| B | 1 | 1 |
|  |  |  |  |
| K=4 | | **Predicted Class** | |
| A | B |
| **Actual Class** | A | 0 | 2 |
| B | 1 | 1 |
|  |  |  |  |
| K=8 | | **Predicted Class** | |
| A | B |
| **Actual Class** | A | 1 | 1 |
| B | 1 | 1 |

**TP = True Positive  
FP = False Positive  
FN = False Negative**

Precision calculation:  
 TP/(TP+FP)

Recall calculation:  
 TP/(TP+FN)

F­1 calculation:  
 2 \* (Precision \* Recall) / (Precision + Recall)

**K=1**Precision: 0/(0+0) = 0  
Recall: 0/(0+2) = 0  
F1: 2 \* (0 \* 0) / (0 + 0) = 0

**K=2**Precision: 0/(0+1) = 0  
Recall: 0/(0+2) = 0  
F1: 2 \* (0 \* 0) / (0 + 0) = 0

**K=4**Precision: 0/(0+1) = 0  
Recall: 0/(0+2) = 0  
F1: 2 \* (0 \* 0) / (0 + 0) = 0

**K=8**Precision: 1/(1+1) = 0.5  
Recall: 1/(1+1) = 0.5  
F1: 2 \* (0.5 \* 0.5) / (0.5 + 0.5) = 0.5

Given the above results, it’s apparent that the kNN classifier performed best when K = 8. This was the only case in which the precision, recall, and the F1 measure did not equal zero. In these first three cases, although some of the data was accurately classified, there were no True Positives for class A.   
  
When K = 8, the precision was 0.5, that means, half of the data that was classified as A, was actually class A. The recall was 0.5, this means that half of the class A data, was correctly classified as class A. The F1 measure can be interpreted as a weighted average of the precision and recall. This tells us how effective our classifier was in classifying A correctly.

**Question 2.**

**It is difficult to assess accuracy based on class membership when data may belong to more than one class at a time. Propose and discuss three criteria that you would use to compare the performance of different classifiers on such data.**

In the case that records belong to more than one class, some of the possible criteria that could be used to evaluate different classifiers are: adding weight to the classes, and then evaluating the precision, recall and f-measure, run a classifier for every possible multiple of classes, and the final, is to combine the outputs of multiple subsets of classifiers.

In the first case, weighting is added to a class if the specific classifier determines it to be the correct class. This way, classes with larger probability are given a heavier weight, and then we can determine the classifier’s accuracy compared to others, using this metric. The performance of the classifier would be determined based on the weighting it assigns either class. Once the weighting is established, you can go a step further and use the weighting, incorporating it into your new classifiers, and judging their performance on the precision, recall and f-measure that they provide. If it results in imbalanced data, that is not proportional to the given set, then obviously the classifier is flawed.

The next possible way to determine the accuracy of classifiers would be to split all the classes into as many classes as there are combinations. After doing this, the classifiers can be run and data collected. For a class to be correctly identified, it should be classified as all the classes that make up its’ parent class. A classifier may also work much like a layered system, in which it determines that one class is a component of the parent, and then continues on to evaluate the rest of the component classes. We can still use f-measure, precision and recall to determine the accuracy of the classifier and the classified data.

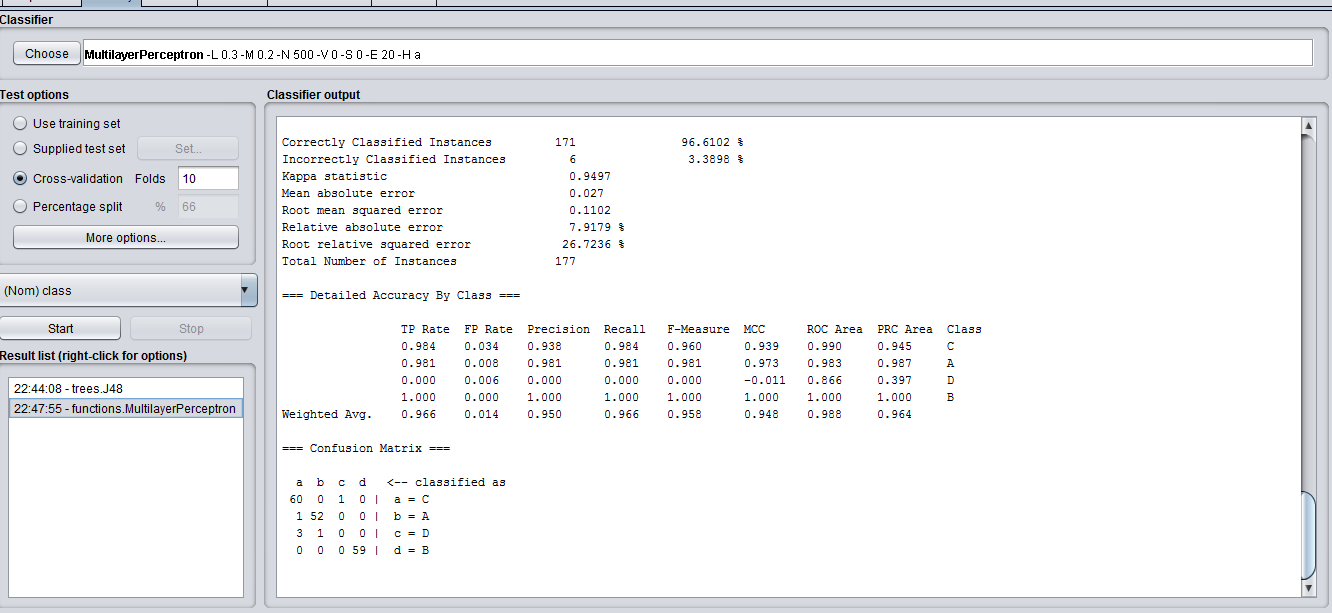
The final way that I mentioned would be to group classifiers together, into subsets, and view their output together. This is also called stacking the classifiers together. Working through the multiple classifiers’ predictions could allow us to more accurately determine the compound classes. The f-measure, recall and precision metrics can still be used to determine the worth of these subsets of classifiers.

**Question 3.**   
**Given a decision tree, you have the option of a) converting the decision tree to rules and then pruning the resulting rules, or b) pruning the decision tree and then converting the pruned tree to rules. Which approach do you think should be preferred and why?**

If you were to prune a decision tree and then convert the pruned tree to rules, you would be limiting a lot of the possible combinations of pruning that could be done, if you used the reverse order. If you prune a decision tree before converting it to rules, you remove the entire query on that stage of the tree. The path through the queries would have to be reconstructed, and possibly the entire order of your queries. If you were to convert a decision tree to rules and then remove a rule, it would most likely be much easier to rebuild the tree afterwards, if it is even necessary. Removing a single rule leaves a lot more options open to the person who is pruning the tree, such as replacing the rule, assimilating the path into another rule, etcetera, whereas removing the entire query is a little trickier. The removal of an entire query would prompt the rebuilding of the structure, and then determining if the removal of said query is more effective and efficient than leaving it in.

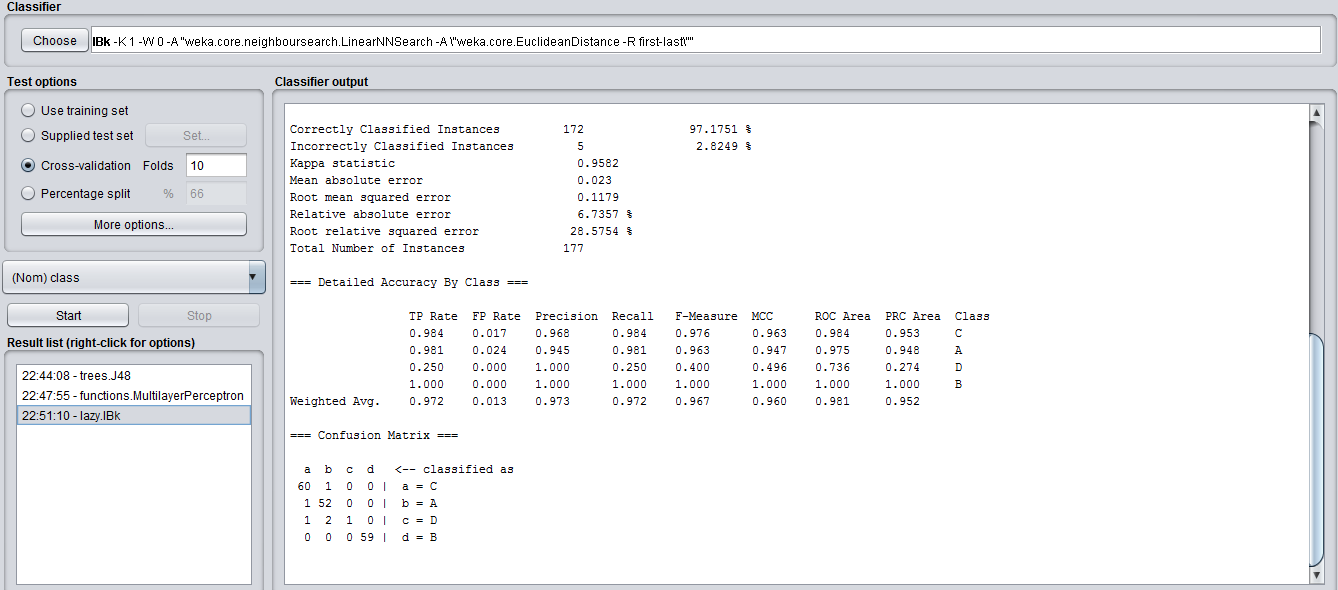
**Question 4.  
Using Weka, analyze the dataset posted on WebCT and discuss the results (include screen shots). Use the following classifiers (using the default configurations with the exception of the classifier of your own choice) and 10-fold Cross Validation:**

**MultilayerPerceptron**

The multilayer perceptron correctly identified classes 96.6102% of the time, incorrectly classifying only 6 records out of the 177.

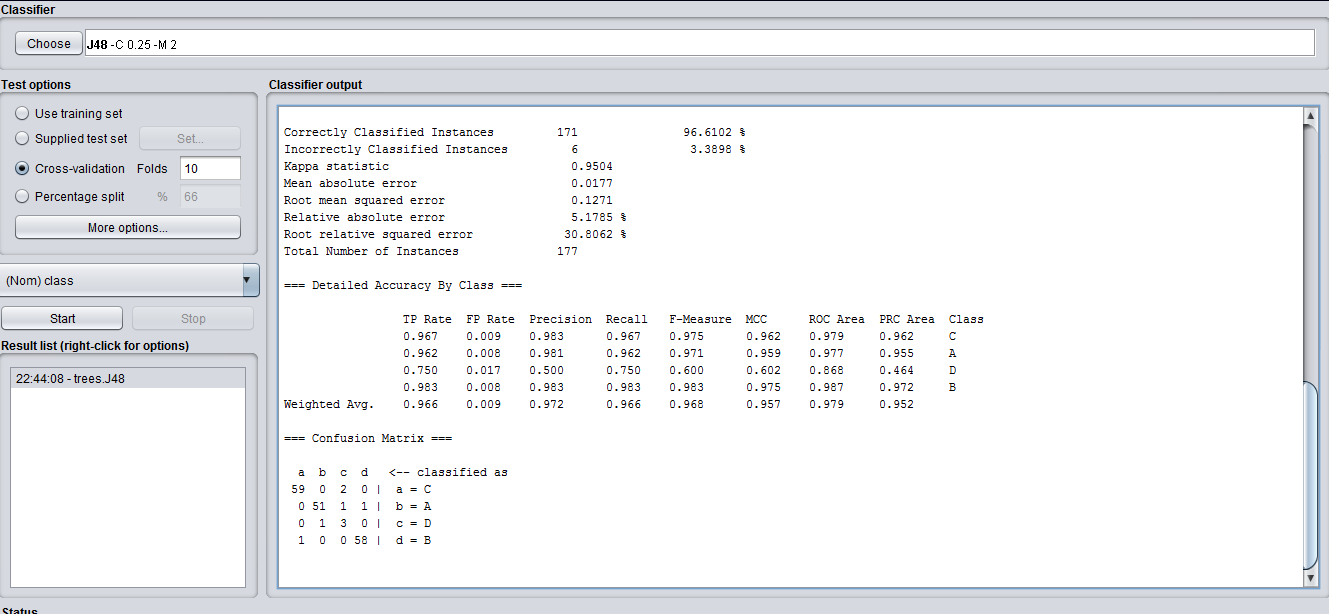
The multilayer perceptron did really well to classify class C with only 1 incorrectly classified instance, with the same happening for class A. Class D was classified completely incorrectly, with the classifier believing that the D class was actually either class C or class A.  
Class B was classified the most accurately, with a recall and precision of 1.

**IBk**

The IBk classifier correctly identified classes 97.1751% of the time, with only 5 tuples being incorrectly classified.

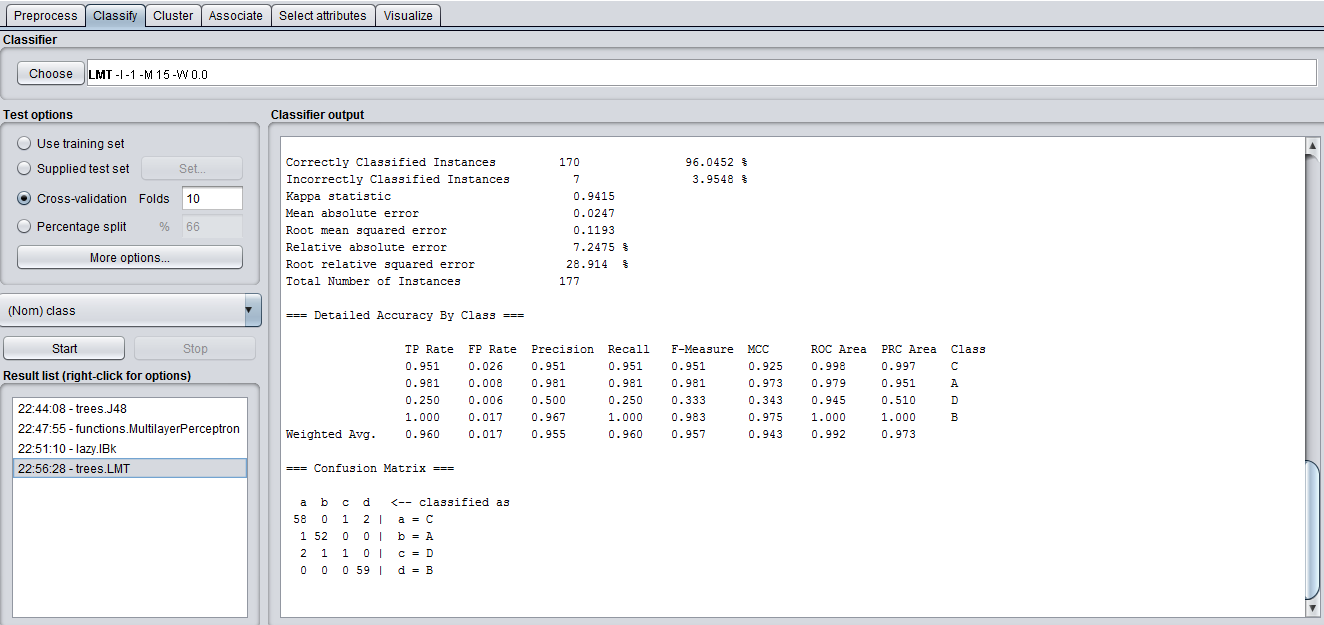
There are only a few differences between the results of the IBk model and the multilayer perceptron, and they are threefold.

IBk labeled a single class C record as class A, whereas the multilayer perceptron identified it as class D  
IBk labeled a class D record as class A, versus the multilayer perceptron’s classification of class C.  
and finally, but most importantly, the IBk classifier was able to correctly classify one class D instance, where the multilayer perceptron marked another class C.

**J48**

The J48 classifier falls short of the IBk’s 97.1751% correctly classified instances, matching the multilayer perceptron at 96.6102%

It correctly classified 75% of the class D records, which seemed to be the issue for the previous two models. With a recall of .75 on class D, that also pushes the J48’s overall F-Measure to 0.968, which is the highest of the four models. Because J48 most accurately represents all four classes, it has the best F-measure.

**LMT:**

The LMT is not nearly as good as the previous three models, correctly classifying only 96.0452% of the records.

It’s classification of class B is comparable to the other three models, correctly identifying every instance of the class. It failed however, to be as precise with class C, incorrectly identifying three of the instances. Along with incorrectly identifying three of the four instances of class D, and incorrectly identifying one instance of class A. This model performed the worst out of the four models.

**Based on the results of the above, answer the following questions:**

**a) Which of the classifiers performed better in terms of the underrepresented class? Justify your answer.   
b) Consider your results from the IBk classifier, given the default configuration why might this classifier be a poor fit for such an unbalanced classification problem?   
c) In general terms suggest two approaches you might take to improve upon the classification of the underrepresented class. Discuss the advantages and disadvantages of each approach.**

**a)**

The J48 classifier performed best in terms of the underrepresented class. It’s recall value of 0.75 places it leagues ahead of the multilayer perceptron (with a recall of 0), along with the LMT, (0.25) and the IBk classifier (0.25 as well). It does however, have a precision of 0.5, which means that it incorrectly classified records as class D as much as it correctly classified them. It also has the highest F-measure for the class (0.6), which is more than the IBk(0.4) and the LMT (0.333) classifiers.

**b)**

Because the IBk problem uses values of K, with the default value being one, we are using only the single nearest neighbor to determine classes. Because class D is not represented as well as the others, it is more likely that the nearest neighbor will in fact, not belong to D, and records will be assimilated by the other classes. If a single class A is closer than a group of class D records, no matter how close the group is, then the piece of data will be classified as A.

**c)**

To improve on the classification of the underrepresented class, we can use a technique discussed in class, oversampling. We can even out the balance between classes by randomly pulling the same four class D records over and over, until we manage to balance out the classes. The issue is that we will have very similar records, meaning some models results may not change (IBk), and that the memory requirements are larger, because the size of our dataset has now increased.   
 In contrast to that, we can always undersample, which means randomly selecting records from the other three classes, in order to create a balanced data set. This would decrease memory requirements, as our dataset would be extremely small, but in this case, it would make it almost useless, as the dataset would contain 16 records. For determining the classes of future, unknown instances, these 16 records don’t provide us with enough information. Oversampling may be a better bet, perhaps with some minor jitter, to introduce some variation on the data.