John Hancock

Florida Atlantic University

Advanced Data Mining and Machine Learning

CAP-6778

September 18th, 2014

Assignment 1: Modeling assignment: Classification using decision trees

Summary

In this assignment we use Weka tool for data mining for classifying instances of a dataset. The instances of the dataset are comprised of biometric information. We investigate the effect of changing algorithm parameters on classification results.

Introduction

We use the Weka implementation of the J48 (C4.5) algorithm for this assignment. Over the four parts of this assignment we add, and change parameters to the J48 algorithm. We apply the J48 algorithm to the Lymphoma96x4026.arff dataset. We report on the results of applying the algorithm to the dataset for each part of the assignment below.

Part 1: Initial Tree

Methodology

For this part of the assignment we use Weka to build a classification model using the J48 (C4.5) classifier algorithm. We import the data into Weka, and run J48 on the data using default settings to generate the data in the results below. Please see the next section for results.

Note: In light of parts 2 and 3 of this assignment, for the results in this section, J48 options in Weka tool include unpruned set to false and confidence factor set to 0.25.

Results

Misclassification Rates

For the Lymphoma96x4026.arff dataset, using J48(C4.5) classifier, Weka produces the following confusion matrix:

classified as	ACL	nonACL
actual class ACL	14	9
actual class nonACL	7	66

Table 1: Part 1 Confusion Matrix

The Assignment 1 requirements document states, "Type I (False Positive): a nonACL module is classified as ACL." Hence, ACL must be the positive class, because this statement says that something is classified as positive and it is classified in the ACL class. Since ACL must be the positive class, nonACL must be the negative class.

Type I / False Positive Rate = 9.60% Assignment 1 tells us a false positive classification is classifying a nonACL module as an ACL module. 7 nonACL modules are classified as ACL module.

There are a total of $7 + 66 = 73$ true nonACL module in the data, therefore the Type I error rate is $\frac{7}{73} \approx 0.096$.

Type II / False Negative Rate = 39.1% Assignment 1 tells us a false negative error is classifying an ACL module as nonACL.

We have a total of $14 + 9 = 23$ truly ACL modules in our data. 9 of the truly ACL modules are classified as nonACL. So the Type II error rate is $\frac{9}{23} \approx 0.391$.

Area Under ROC Curve

Weka produces 3 results for the area under the receiver operating characteristic (ROC) curve. Here are the ROC area numbers when we run the J48 classifier with unpruned set to false and confidence factor set to 0.25 on the Lymphoma96x4026.arff dataset:

	ROC Area
ACL	0.781
nonACL	0.777
Weighted Average	0.778

Table 2: Part 1 Areas under ROC Curves

The assignment defines a Type I / false positive error as classifying nonACL as ACL. This means that ACL is the positive class. Hence nonACL is the negative class. In the text we see an ROC curve presented as a plot of false positive rates to true positive rates [1]. In that context we are interested in the ROC curve associated with the positive class, ACL. In the Weka classifier output, the number is 0.781.

See the section titled “Area under ROC Curve Comparison,” for a comparison of ROC curve areas for the three different parameter settings we set for the J48 algorithm.

Decision Tree Nodes/Leaves and Representation

Part 1 of the assignment requires us to record the number of leaves and nodes in the selected tree, and to produce a representation of the tree the same way as the text book.

There is an image of a decision tree in a discussion of using Weka to apply the J48 algorithm to data in the text. One may locate this image using the bibliographic information in reference [1].

Weka has a feature for drawing the decision trees it computes for data sets. Weka produces the diagram in figure 1 below when we use it to apply the J48(C4.5) algorithm to the Lymphoma96x4026.arff dataset. This digram is like the one in the text noted in reference [1]. Please see the next page for the diagram.

13 Nodes and 7 Leaves: We can count the number of nodes and leaves in figure 1 to record the number of nodes and leaves. There are 13 nodes and 7 of these are leaves.

Representation of Tree

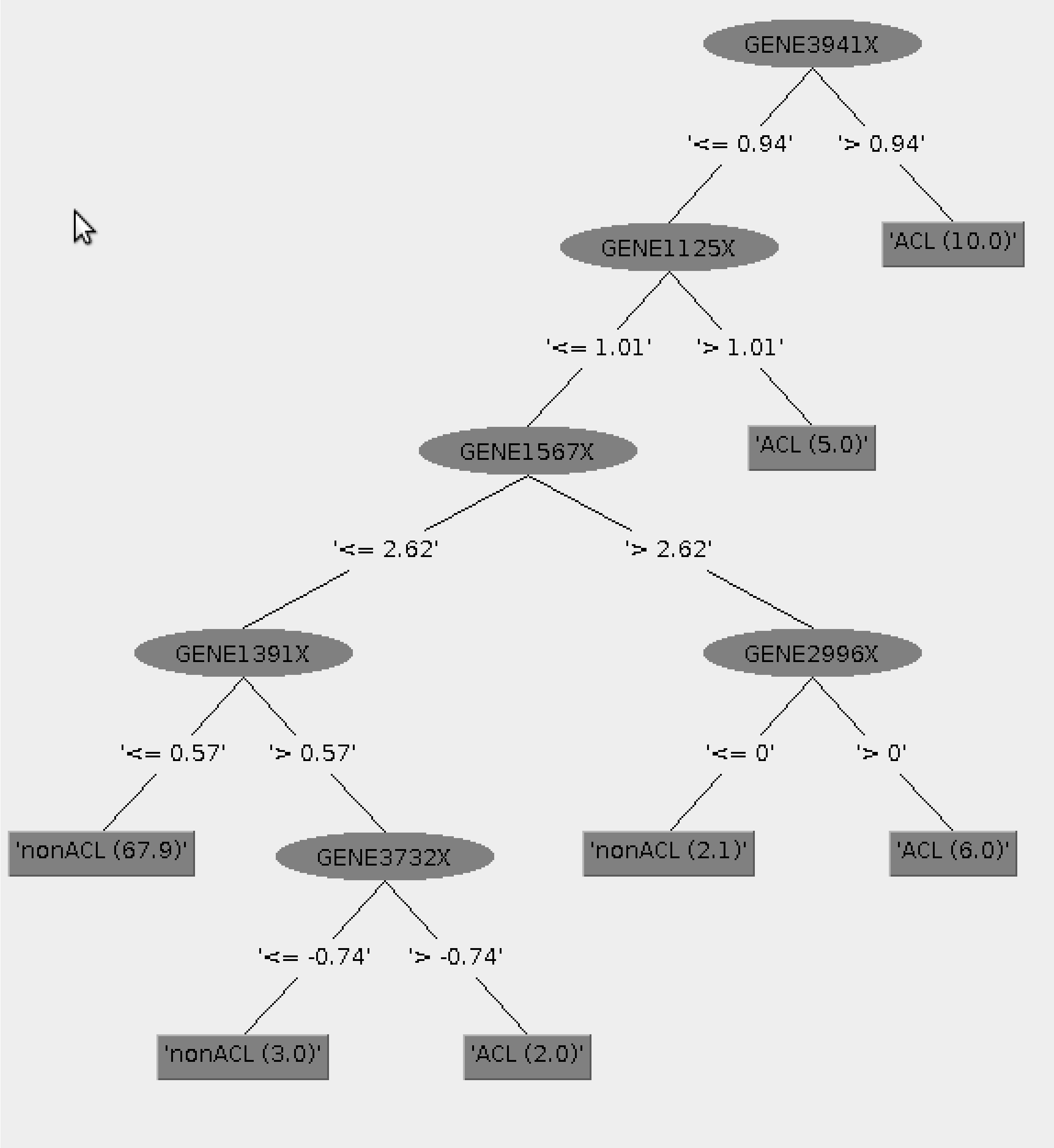

Figure 1: Decision tree for J48 Algorithm, unpruned=false, confidenceFactor=0.25

Part 2: Unpruned Tree

Methodology

For Part 2, in the Weka tool, we set the unpruned option to true for the J48 algorithm and apply it to the Lymphoma96x4026.arff again.

Results

Misclassification Rates

Although we set the unpruned option to true, there is no change to the confusion matrix that Weka produces. It is identical to the one we have in Part 1:

classified as	ACL	nonACL
actual class ACL	14	9
actual class nonACL	7	66

Table 3: Part 2 Confusion Matrix

Type I / False Positive Rate = 9.60% Because the confusion matrix is unchanged, the false positive error rate is also unchanged.

Type II / False Negative Rate = 39.1% The false negative rate is also unchanged from what we have in Part 1.

Area Under ROC Curve

Here are the ROC area numbers when we run the J48 classifier with unpruned set to true. on the Lymphoma96x4026.arff dataset:

	ROC Area
ACL	0.773
nonACL	0.769
Weighted Average	0.77

Table 4: Part 2 Areas under ROC Curves

It is interesting that these numbers are different from the area under ROC curve numbers Weka produces for J48 with unpruned equal to false in Part 1.

See the section titled "Area under ROC Curve Comparison," for a comparison of ROC curve areas for the different parameter settings we use for the J48 algorithm.

Decision Tree Nodes/Leaves and Representation

The resulting tree is identical to the one Weka generates in Part 1

13 Nodes and 7 Leaves: We can count the nodes and leaves in figure 2 below.

Representation of Tree

The tree Weka produces is identical to the tree in figure 1. We present a larger image of the tree in figure 2, and then smaller images of both trees side by side to make it obvious that the trees are identical. Please see the next page for the representation of the tree.

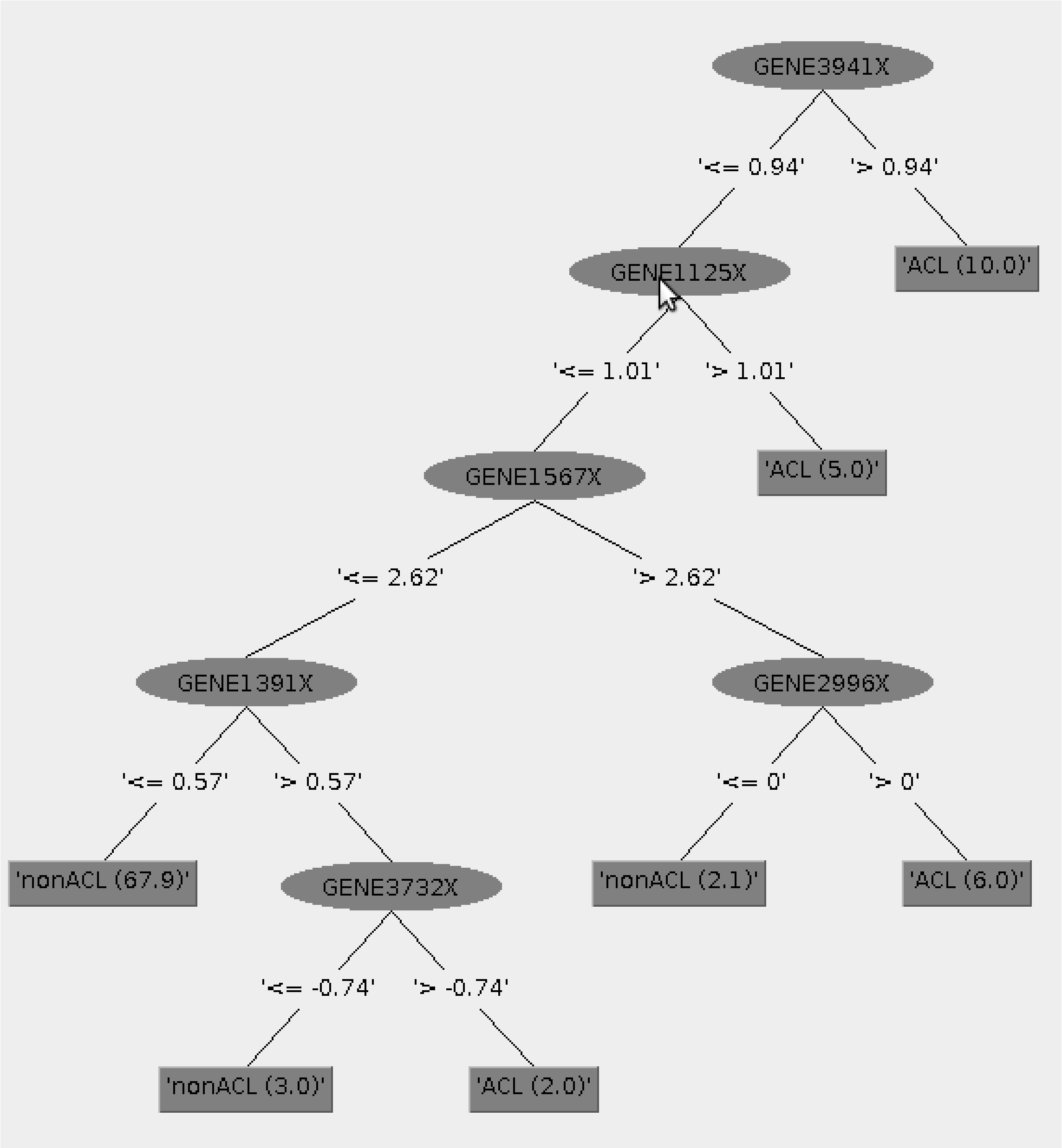

Figure 2: decision tree with unpruned set to true.

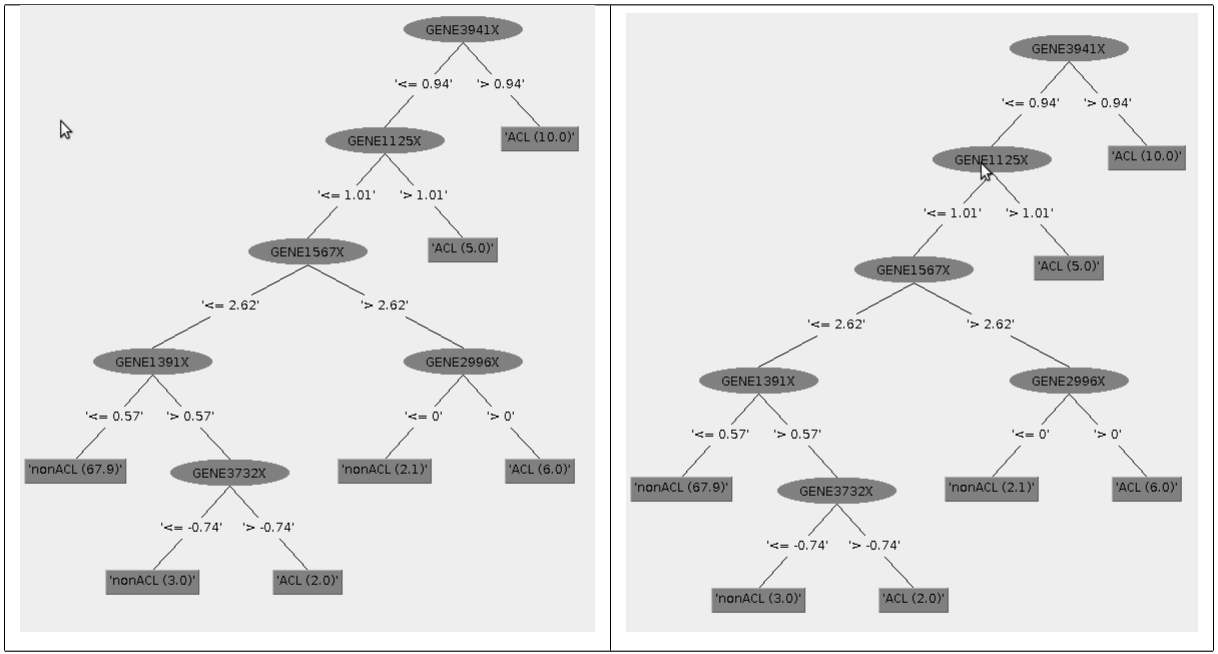

Figure 3: comparison of generated trees.

Part 3: Confidence Factor

Methodology

For part 3, in the J48 options we set the confidence factor to 0.01, unpruned to false, and follow the same steps in Part 1 to generate the data in the results below.

Results

Overall, we do not see any difference in results when we change the confidence factor to 0.01. There is no change to the decision tree Weka produces.

The assignment 1 requirements document specifies that we must discuss some facts about the decision tree Weka produces. The list below addresses these requirements.

- "How does the size of the new tree compare to one built in Part 1?"
    - The size of the tree is unchanged.

- After the question, there is a requirement to explain why the size of the tree is different.
  - Since the size of the tree is unchanged, we are unable to make an explanation.
- What part was pruned?
  - Since we do not see a change in the size or structure of the tree, the answer to this question is that no part is pruned.

Misclassification Rates

The confusion matrix after changing the confidence factor is unchanged from parts one and 2:

classified as	ACL	nonACL
actual class ACL	14	9
actual class nonACL	7	66

Table 5: Part 3 Confusion Matrix

Type I / False Positive Rate = 9.60% The false positive rate is unchanged because the confusion matrix is unchanged.

Type II / False Negative Rate = 39.10% The false negative error rate remains what we have in Part 1 and Part 2, also because the confusion matrix here in Part 3 is identical to what we have in Parts 1 and 2.

Area Under ROC Curve

Here are the ROC area numbers when we use run Weka's J48 classifier with unpruned set to false and confidenceFactor set to 0.01 on the Lymphoma96x4026.arff dataset:

	ROC Area
ACL	0.773
nonACL	0.769
Weighted Average	0.77

Table 6: Part 3 Areas under ROC Curves

The areas under the ROC curves are the same here as they are in Part 2. As we see in Part 2, changing the unpruned option causes the area under the ROC curve that Weka reports to change. Here there is no change.

See the section titled "Area under ROC Curve Comparison," for a comparison of ROC curve areas for the three different parameter settings we set for the J48 algorithm.

Decision Tree Nodes/Leaves and Representation

13 Nodes and 7 Leaves The decision tree has the same number of nodes and leaves as the previous sections. Please see the next page for the representation of the tree.

Representation of Tree

As required, here is a representation of the tree. It is not different from the tree in Parts 1 and 2:

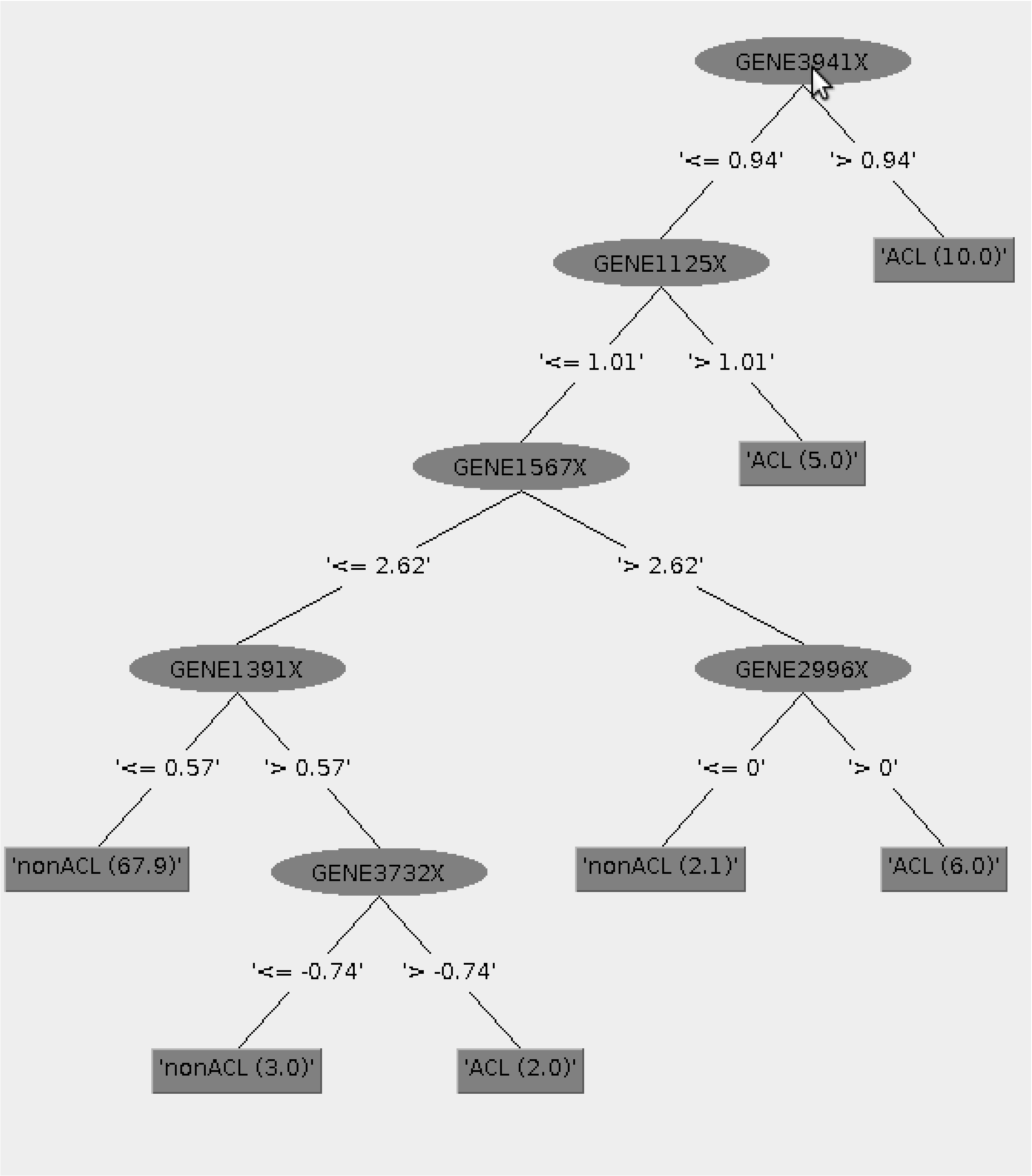

Figure 4: decision tree for J48 Algorithm, unpruned=false, confidenceFactor=0.01

Area under ROC Curve Comparison

In this section we present a summary comparison of the changes Weka reports in areas under ROC curves

J48 Setting	Part 1	Part 2	Part3
unpruned	false	true	false
confidenceFactor	0.25	n/a	0.01
ACL	0.781	0.773	0.773
nonACL	0.777	0.769	0.769
weighted Average	0.778	0.77	0.77

Table 7: Comparison of Areas Under ROC curves

An area under ROC curve value closer to one implies better performance [3].

We see that Weka reports the highest area under ROC curve values when we use it with J48 with options set in Part 1 (unpruned=false, confidenceFactor=0.25).

Furthermore, we would like to point out that Weka reports the same change in ROC curves' areas when we change the unpruned option to true, or when we lower the confidenceFactor setting to 0.01.

Part 4: Cost Sensitivity

Methodology

In this section we configure Weka to use a a cost sensitive classifier to determine the optimal cost ratio. We configure the cost sensitive classifier to use a J48 classifier with unpruned set to false and confidenceFactor set to 0.25. These are the same settings we use in Part 1.

We leave the cost of a Type I error fixed at 1, and change the cost of a Type II error to the values 0.5, 1, 2, 4, 8, and 16. We run the cost sensitive classifier on the Lymphoma96x4026.arff dataset for each cost of a Type II error.

Results

In the results below we present the trend in misclassification rates.

Please note: for every value of a Type II error below, the value used for the cost of a Type I error is 1.

We then answer the question, “What happens when the cost of a Type II error decreases/increases?”

Misclassification Rates

The table below summarizes how error rates change as the cost of a Type II error increases. We base Type I percentages on 23 instances of the positive class. We base Type II percentages on 73 instances of the negative class. We used a binary search technique to home in on regions of interest to choose Type II error costs in the tables on the next page.

Type II Error Cost	0	0.5	1	2	4	8	16	20	21	21.5	22	22.5	23	24	28	32	34	36	40	64	128
Type I / False Positive Rate	0%	12.3%	9.6%	20.5%	19.2%	20.5%	31.5%	27.4%	28.8%	28.8%	32.9%	28.8%	30.1%	37.0%	34.2%	37.0%	43.8%	46.6%	49.3%	64.4%	100.0%
Type II / False Negative Rate	100%	30.4%	39.1%	47.8%	39.1%	39.1%	39.1%	39.1%	30.4%	30.4%	34.8%	34.8%	30.4%	30.4%	34.8%	26.1%	34.8%	21.7%	21.7%	17.4%	0%

Table 8: Changes in Error Rates as Cost of Type II Error Increases

Effect of Increasing/Decreasing Type II Error Cost

With the table above, we can answer the question, “What happens when the cost of a Type II error increases/decreases. The table shows that as the cost of a Type II error increases, the false positive rate increases to 100%, and the false negative rate goes to 0. When we set the cost of a Type II error to 0, we get a false positive error rate of 0 and a false negative error rate of 100%.

The graph below is a plot of Type II error cost versus Type I and Type II errors. We exclude the point with Type II error cost of 0 because we use a logarithmic scale for the x-axis.

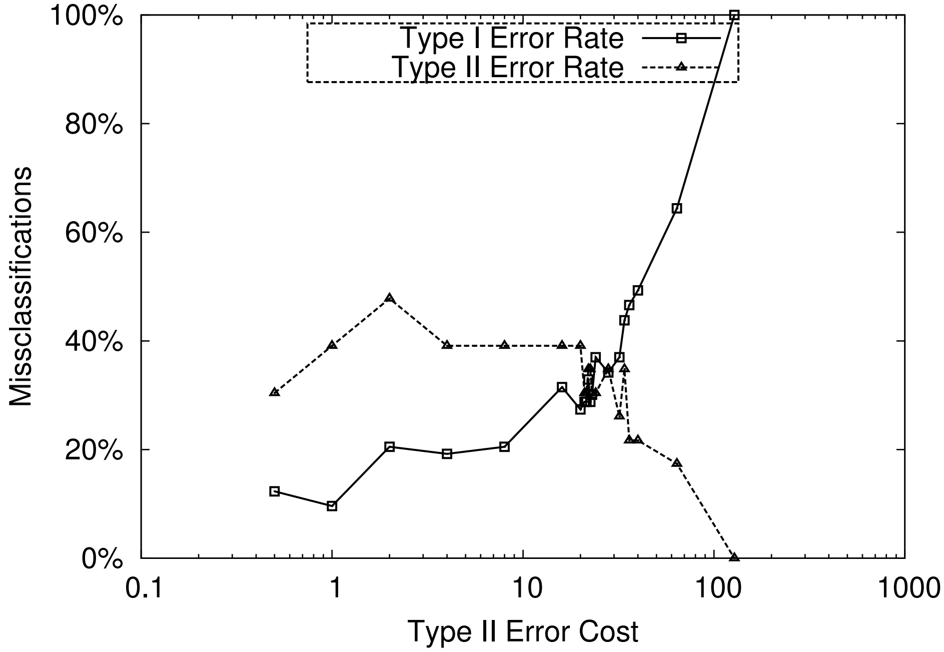

Figure 5: False Negative and False Positive Rates for Increasing Type II error cost.

The graph below is a restriction of the above with Type II error costs restricted to the range 8-32:

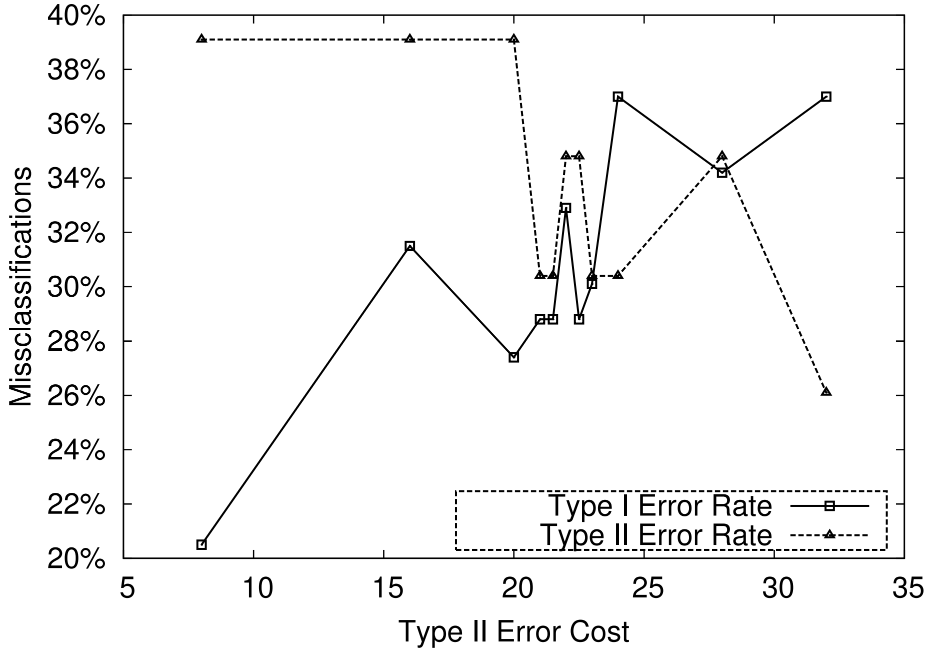

Figure 6: Detail of figure 5, x-axis restricted to the range 8-32.

Optimal Cost of Type II Error=21.5

Assignment 1 requires us to determine a balanced misclassification rate, with Type II errors as low as possible. With the cost of a Type II error set to 28, we see that the error rates are very close to one and other. At a Type II error cost of 28, the false negative rate is 34.2% and the false positive rate is 34.8%. However, if we lower the cost of a Type II error to 21 or 21.5, we get a Type II error rate of 30.4%, and the Type I error rate of 28.8%. Therefore, at a Type II error cost of 21 or 21.5, we have approximately equal error rates, but the Type II error rates are lowest in this neighborhood of Type II error costs. We feel that the higher Type II error cost, 21.5, is best. For data different from what we have in the test data, we feel there would be a better chance of not making a Type II error with the higher error cost. We would like to emphasize that this optimal Type II error cost is associated with a Type I error cost of 1.

Future Research

The results we present in this assignment are based on one data set, Lymphoma96x4026.arff. It is possible that the optimal cost ratio we find is dependent on this data set. In order to determine whether or not this is the case, we would require more datasets. For future research we would use these datasets to determine multiple optimal cost ratios. Furthermore, we would do ANOVA on the cost ratios to determine the significance of using a particular dataset.

We also noticed that for certain error costs, 21, 21.5, 23 and 24 the Type II error rate is 30.4%. This is interspersed with Type II error rates of 34.8% for Type II error costs of 22, 22.5, and 28. More data would help determine whether or not this is a coincidence, or there is some kind of periodic phenomenon happening as we increase the cost of Type II errors in the range from 21to 28.

Conclusions

The results we present allow us to draw four conclusions.

1. Changing the unpruned option from true to false, or changing the value of the confidenceFactor from 0.25 to 0.01 int the J48 classification algorithm options has no effect on the confusion matrix Weka produces for the Lymphoma96x4026.arff dataset. The confusion matrices we show in Parts 1-3, Tables 1, 3, and 5, confirm this conclusion.

2. Changing the unpruned option true has the effect of lowering the areas under the ROC curves Weka reports. See Table 7 above for evidence supporting this conclusion.

3. Weka reports the same drop in ROC curves' areas when we change the unpruned option to true, or when we lower the confidenceFactor setting to 0.01. We base this conclusion on table 7 above.

4. The optimal cost of a Type II error for the Lymphoma96x4026.arff dataset is 32, when the cost of a Type I error is 1. See Figure 5 for the basis of this conclusion.

References

[1] I. Witten and E. Frank, Data Mining (second edition). San Francisco: Elsevier, 2005, ch. 5 p.169 fig. 5.2.

[2] I. Witten and E. Frank, Data Mining (second edition). San Francisco: Elsevier, 2005, ch. 10 p.379 fig. 10.6(a).

[3] *The Area Under an ROC Curve* [Online]. Available: http://gim.unmc.edu/dxtests/roc3.htm