Classification Using Decision Trees

Shaun Pritchard

Florida Atlantic University

CAP 6778

September -17-2021

M. Khoshgoftaar

Part 1: Initial tree

The data set provides 23 instances of ACL features and 72 instances of nonACL features. Approximately 31.94 % instance variance between the classes. Also note, when there are a large number of samples in different classes (unbalanced target), accuracy is not a very reliable metric for the true performance of a classifier.

In the initial classification model based on J48 with 10 k-fold cross validation, 78 % of instances of the model were correctly correlated and 22 % incorrectly correlated. For this model, the MAE was calculated at 0.2178, which is a very good baseline. Despite comparing up to 78% of the correctly classified instances, it does not seem to be very strong in comparison.

The misclassification error rates for both types of misclassifications from the confusion matrix are as follows in Table 1.1.

a	**b**	
14	9	**A =ACL**
12	60	**B = NonACL**

Table 1.1

The (ROC) Receiver Operator Characteristic shows the weighted average below the curve at 0.732, nonACL at 0.731, and ACL at 0.736. This confirms the precision of the model inference. The data shows ACL classification 14 correct classification compared to 60 correct classification for nonACL. There are 12 Type I classifications and 9 Type II classifications.

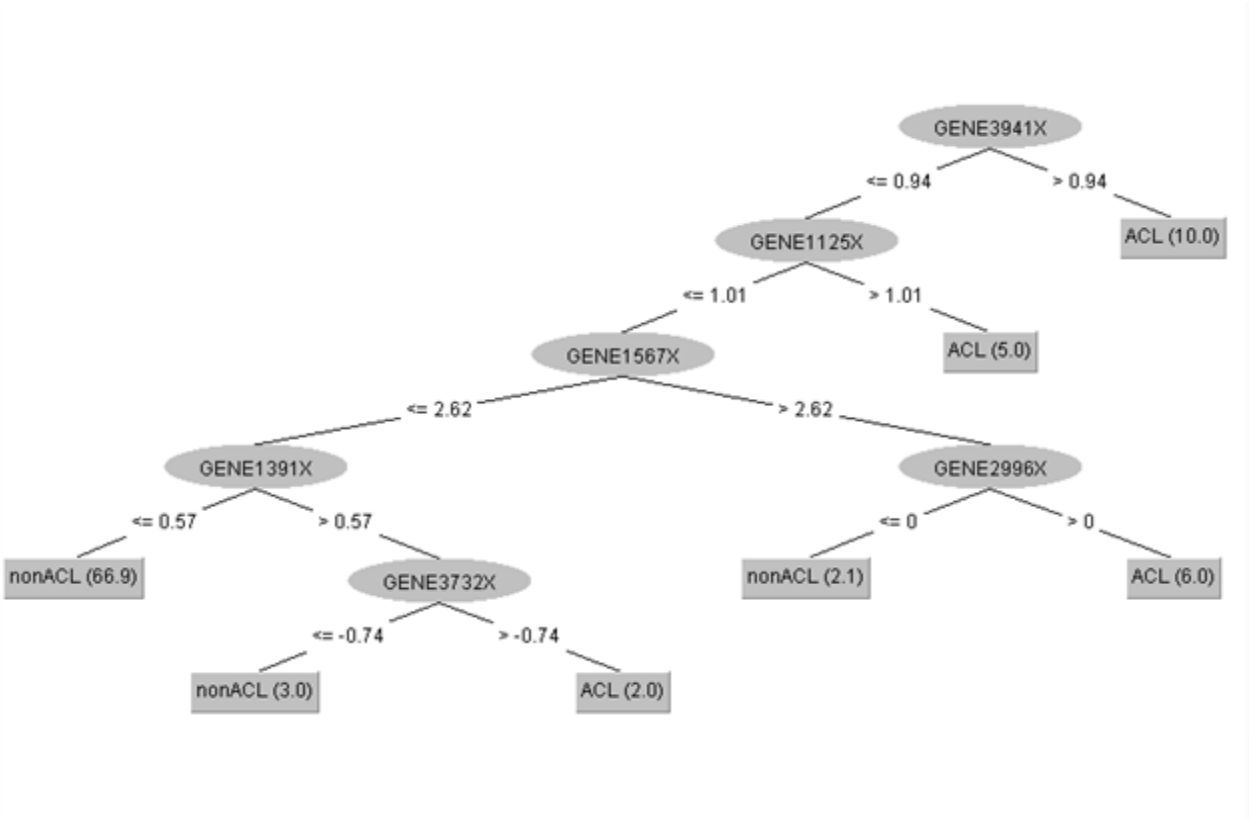
- Type I (False Positive): a nonACL module is classified as ACL

- Type II (False Negative): an ACL module is classified as nonACL

According to the J48 tree classification in Figure 1.2 below, there are 13 sizes and 7 leaves.

Below is a list of 7 leaves more suited for the nonACL classification.

Figure 1.1



Part 2: Unpruned tree

Setting the J48 10 k-fold cross validation with the attribute to unpruned. The observations show 80% correctly classified instances and 20% incorrectly classified instances. The MAE is 0.209 which is slightly lower than the initial J48 tree classification above.

The misclassification error rates for both types of misclassifications from the confusion matrix are as follows in Table 2.1.

a	**b**	
14	9	**A =ACL**
10	62	**B = NonACL**

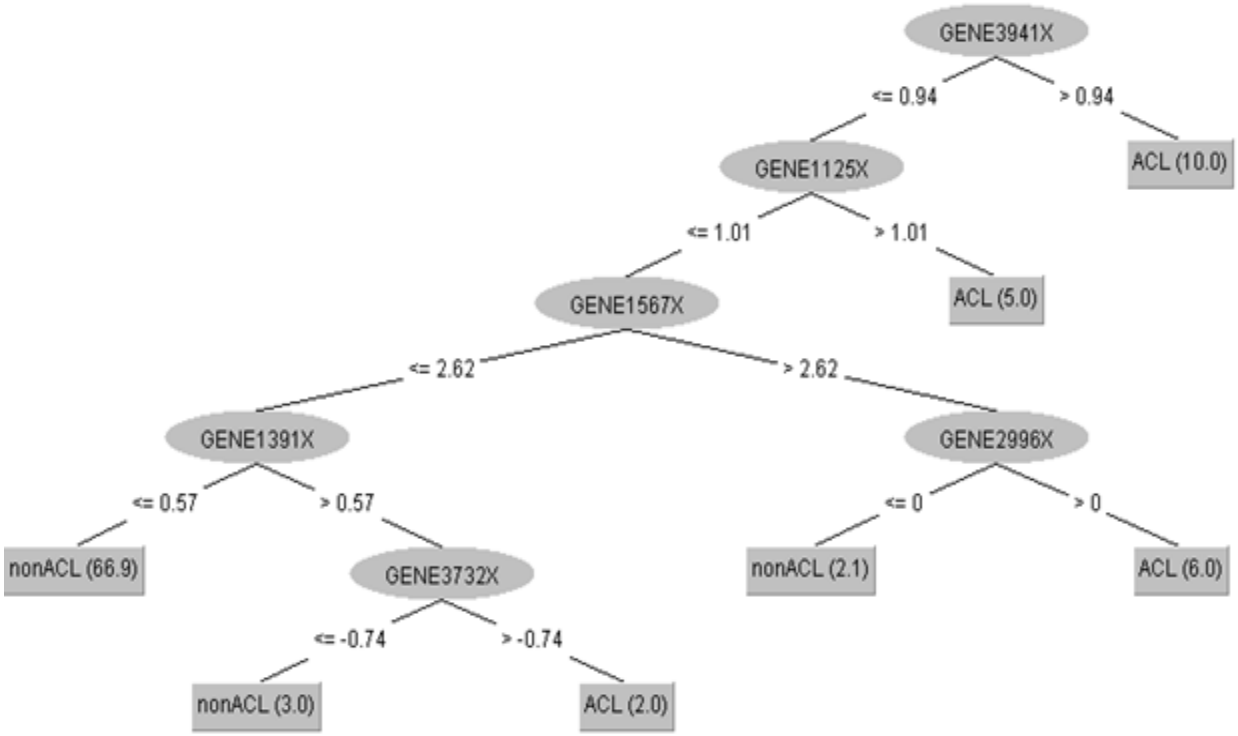
Table 2.1

The (ROC) Receiver Operator Characteristic shows the weighted average below the curve at 0.732, nonACL at 0.731, and ACL at 0.736. This is consistent with the previous results for the initial inference. There is a reduction of the True Positive rate for nonACL from 0.833 to 0.861 as well as a reduction in the False Positive rate in ACL from 0.167 to 0.139.

The data shows ACL classification 14 correct classification compared to 62 correct classification for nonACL. There are 10 Type I classifications and 9 Type II classifications.

According to the J48 tree classification in Figure 2.2 below, there is a 13-node tree with 7 leaves consistent with some variation.

Figure 2.2



Part 3: Confidence Factor

When a confidence factor is changed in the J48 with 10 k-fold cross validation, the tree size did not reduce nor was it pruned any more than it was originally in the initial tree. No pruning occurred due to the accuracy of J48 confidence factor which applies a statistical test value not significant enough to cause change or reduction in tree nodes alone. The only way for a strong prune would be to use and increase *minNumObj* attribute with a smaller confidence factor such as 0.01.

The size of each leaf is allowed to grow the output number of leaves and also decreases. lowering the confidence factor is supposed to decrease the amount of post-pruning. Therefore, a confidence factor of 0.01 with the same model accuracy as 0.25 is not significant enough to cause reduction.

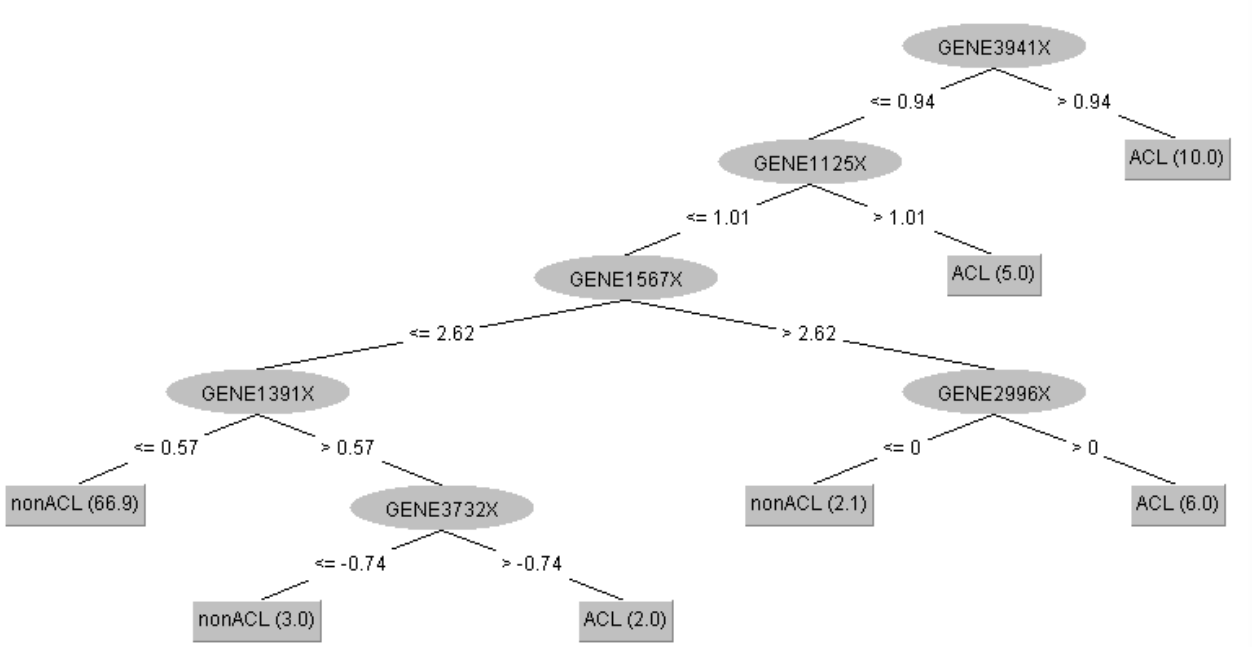
The table below shows the calculated accuracy of the confidence factor [1]. Essentially, pruning is the process of comparing the amount of error in a decision tree and then deciding on the best way to minimize it to avoid error. This error cannot be determined solely on the confidence error given. Each confidence factor was tested independently on the model by me [2].

Confidence Factor J48	Accuracy
0.005	73.0769
0.05	75.5245
0.1	75.5245
0.25	75.5245
0.5	73.4266

Table 3.1

Following is the J48 with confidence factor set to 0.01 below in Figure 3.1

Figure 3.1



Part 4: Cost sensitivity

The distinction between a Type I and a Type II error drastically changed when implementing the cost sensitive classifier with the J48. I increased the Type II classifier by a maximum of 2.5 and a minimum of 1.5 the results of the Type II errors are as follows.

For the increase in the cost matrix we see the results become worse with increased Type II errors.

- The Type II number of leaves decreased to 5, and the tree size decreased to 9
- The correctly classified instances where 70 at 74%
- Incorrectly classified instances where 25 at 26%
- The MAE increased to 0.2582

· The ROC for both classes was 0.591, this suggests a naïve or un-trustworthy precision in the classifier. 0 .5 is typically the minimum limit for a good classification.

· The confusion matrix showed ACL classification 7 correct classification compared to 63 correct classification for nonACL. There are 9 Type I classifications and 16 Type II classifications. The tree for the Type II cost classifier of cost 2.5 reduces the tree by 2 stumps as follows in Figure 4.1.

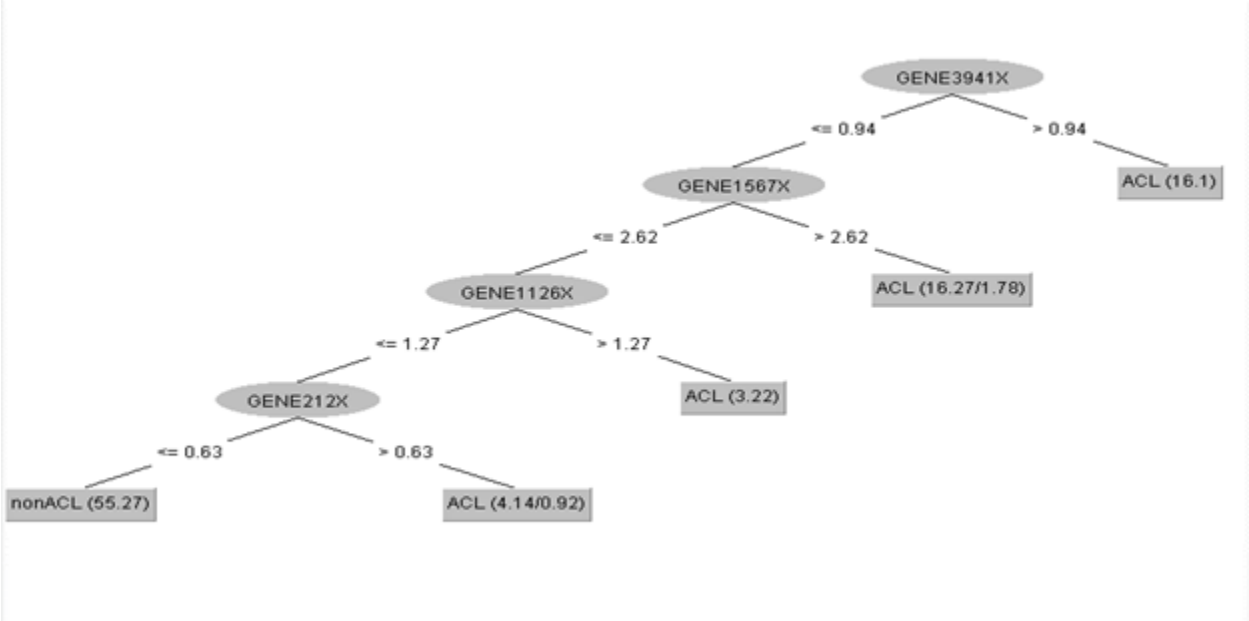


Figure 4.1

For the decrease in the cost matrix type II of 0.5 we see the results become best case with decreased Type II errors.

· The Type II number of leaves decreased to 6, and the tree size decreased to 11

· The correctly classified instances where 82 at 86%

· Incorrectly classified instances where 13 at 14%

· The MAE decreased to 0.1403(lowest)

· The ROC for ACL class 0.830, nonACL class 0.831 this suggest a good classification

· The confusion matrix showed ACL classification 16 correct classification compared to 66 correct classification for nonACL. There were 6 Type I classifications and 7 Type II classifications. The tree for the Type II cost classifier of cost 0.5 reduces the tree by 1 stump as follows in Figure 4.2.

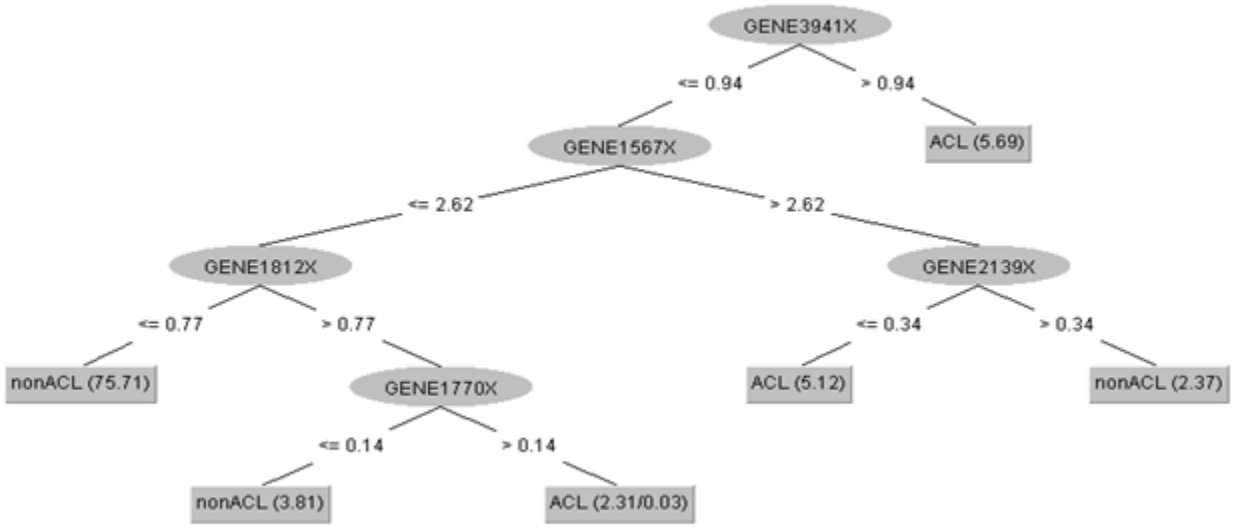


Figure 4.2

Overall, the reduction in the cost matrix through the cost sensitive classifier yielded the best classification for the model.

Weka Data Analysis Output

I was not sure if the assignment required the full output from my analysis? Regardless, here are all the classification instances I built in Weka with the assignment data.

Part 1 Initial Tree analysis

=== Run information J48 Initial Tree ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: figure1

Instances: 95

Attributes: 4027

[list of attributes omitted]

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

GENE3941X <= 0.94

| GENE1125X <= 1.01

| | GENE1567X <= 2.62

| | | GENE1391X <= 0.57: nonACL (66.9)

| | | GENE1391X > 0.57

| | | | GENE3732X <= -0.74: nonACL (3.0)

| | | | GENE3732X > -0.74: ACL (2.0)

| | GENE1567X > 2.62

| | | GENE2996X <= 0: nonACL (2.1)

| | | GENE2996X > 0: ACL (6.0)

| GENE1125X > 1.01: ACL (5.0)

GENE3941X > 0.94: ACL (10.0)

Number of Leaves : 7

Size of the tree : 13

Time taken to build model: 0.34 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 74 77.8947 %

Incorrectly Classified Instances 21 22.1053 %

Kappa statistic 0.4232

Mean absolute error 0.2178

Root mean squared error 0.4496

Relative absolute error 58.8069 %

Root relative squared error 104.8305 %

Total Number of Instances 95

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.609	0.167	0.538	0.609	0.571	0.425	0.736	0.562	ACL
	0.833	0.391	0.870	0.833	0.851	0.425	0.731	0.853	nonACL
Weighted Avg.	0.779	0.337	0.789	0.779	0.783	0.425	0.732	0.782	

=== Confusion Matrix ===

a	b	<-- classified as
14	9	| a = ACL

12 60 | b = nonACL

Part 2 Initial tree with Pruning set to true:

=== Run information Initial tree with Pruning ===

Scheme: weka.classifiers.trees.J48 -U -M 2

Relation: figure1

Instances: 95

Attributes: 4027

[list of attributes omitted]

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 unpruned tree

GENE3941X <= 0.94

| GENE1125X <= 1.01

| | GENE1567X <= 2.62

| | | GENE1391X <= 0.57: nonACL (66.9)

| | | GENE1391X > 0.57

| | | | GENE3732X <= -0.74: nonACL (3.0)

| | | | GENE3732X > -0.74: ACL (2.0)

| | GENE1567X > 2.62

| | | GENE2996X <= 0: nonACL (2.1)

| | | GENE2996X > 0: ACL (6.0)

| GENE1125X > 1.01: ACL (5.0)

GENE3941X > 0.94: ACL (10.0)

Number of Leaves : 7

Size of the tree : 13

Time taken to build model: 0.15 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	76	80 %
Incorrectly Classified Instances	19	20 %
Kappa statistic	0.463	
Mean absolute error	0.209	
Root mean squared error	0.436	
Relative absolute error	56.4239 %	
Root relative squared error	101.6764 %	
Total Number of Instances	95	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.609	0.139	0.583	0.609	0.596	0.463	0.736	0.562	ACL
	0.861	0.391	0.873	0.861	0.867	0.463	0.731	0.853	nonACL
Weighted Avg.	0.800	0.330	0.803	0.800	0.801	0.463	0.732	0.782	

=== Confusion Matrix ===

a	b	<-- classified as

14 9 | a = ACL

10 62 | b = nonACL

Part 3 with confidence factor set to 0.01

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.01 -M 2

Relation: figure1

Instances: 95

Attributes: 4027

[list of attributes omitted]

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

GENE3941X <= 0.94

| GENE1125X <= 1.01

| | GENE1567X <= 2.62

| | | GENE1391X <= 0.57: nonACL (66.9)

| | | GENE1391X > 0.57

| | | | GENE3732X <= -0.74: nonACL (3.0)

| | | | GENE3732X > -0.74: ACL (2.0)

| | GENE1567X > 2.62

| | | GENE2996X <= 0: nonACL (2.1)

| | | GENE2996X > 0: ACL (6.0)

| GENE1125X > 1.01: ACL (5.0)

GENE3941X > 0.94: ACL (10.0)

Number of Leaves : 7

Size of the tree : 13

Time taken to build model: 0.15 seconds

=== Stratified cross-validation ===

=== Summary ==

Correctly Classified Instances 74 77.8947 %

Incorrectly Classified Instances 21 22.1053 %

Kappa statistic 0.4232

Mean absolute error 0.2178

Root mean squared error 0.4496

Relative absolute error 58.8069 %

Root relative squared error 104.8305 %

Total Number of Instances 95

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.609	0.167	0.538	0.609	0.571	0.425	0.736	0.562	ACL
0.833	0.391	0.870	0.833	0.851	0.425	0.731	0.853	nonACL
Weighted Avg. 0.779	0.337	0.789	0.779	0.783	0.425	0.732	0.782	

=== Confusion Matrix ===

a	b	<-- classified as
14	9	a = ACL

12 60 | b = nonACL

Part 4-1 Cost sensitive classification with J48 and cost matrix variation Type II set to 2.5

=== Run information CSC with J48 Type II at 2.5 ===

Scheme: weka.classifiers.meta.CostSensitiveClassifier -cost-matrix "[0.0 2.5; 1.0 0.0]" -S 1 -W

weka.classifiers.trees.J48 -- -C 0.25 -M 2

Relation: figure1

Instances: 95

Attributes: 4027

[list of attributes omitted]

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

CostSensitiveClassifier using reweighted training instances

weka.classifiers.trees.J48 -C 0.25 -M 2

Classifier Model

J48 pruned tree

GENE1610X <= -0.77: nonACL (30.08)

GENE1610X > -0.77

| GENE3781X <= -0.77: nonACL (11.74)

| GENE3781X > -0.77

| | GENE3332X <= 2.16

| | | GENE1879X <= 0.36: ACL (43.65/1.47)

| | | GENE1879X > 0.36: nonACL (2.84)

| | GENE3332X > 2.16: nonACL (6.69)

Number of Leaves : 5

Size of the tree : 9

Cost Matrix

0 2.5

1 0

Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 71 74.7368 %

Incorrectly Classified Instances 24 25.2632 %

Kappa statistic 0.2445

Mean absolute error 0.2482

Root mean squared error 0.4934

Relative absolute error 67.0348 %

Root relative squared error 115.0524 %

Total Number of Instances 95

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.348	0.125	0.471	0.348	0.400	0.249	0.616	0.347	ACL
0.875	0.652	0.808	0.875	0.840	0.249	0.616	0.804	nonACL
Weighted Avg. 0.747	0.525	0.726	0.747	0.733	0.249	0.616	0.694	

=== Confusion Matrix ===

a b <-- classified as

8 15 | a = ACL

9 63 | b = nonACL

Part 4-2 Cost sensitive classification with J48 and cost matrix variation Type II set to 0.5

=== Run information CSC with J48 Type II at 0.5===

Scheme: weka.classifiers.meta.CostSensitiveClassifier -cost-matrix "[0.0 0.5; 1.0 0.0]" -S 1 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2

Relation: figure1

Instances: 95

Attributes: 4027

[list of attributes omitted]

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

CostSensitiveClassifier using reweighted training instances

weka.classifiers.trees.J48 -C 0.25 -M 2

Classifier Model

J48 pruned tree

GENE3941X <= 0.94

| GENE1567X <= 2.62

| | GENE1812X <= 0.77: nonACL (75.71)

| | GENE1812X > 0.77

| | | GENE1770X <= 0.14: nonACL (3.81)

| | | GENE1770X > 0.14: ACL (2.31/0.03)

| GENE1567X > 2.62

| | GENE2139X <= 0.34: ACL (5.12)

| | GENE2139X > 0.34: nonACL (2.37)

GENE3941X > 0.94: ACL (5.69)

Number of Leaves : 6

Size of the tree : 11

Cost Matrix

0 0.5

1 0

Time taken to build model: 0.15 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	82	86.3158 %
Incorrectly Classified Instances	13	13.6842 %
Kappa statistic	0.6215	
Mean absolute error	0.1403	
Root mean squared error	0.3572	
Relative absolute error	37.894 %	
Root relative squared error	83.2814 %	
Total Number of Instances	95	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.696	0.083	0.727	0.696	0.711	0.622	0.830	0.686	ACL
	0.917	0.304	0.904	0.917	0.910	0.622	0.831	0.906	nonACL
Weighted Avg.	0.863	0.251	0.861	0.863	0.862	0.622	0.831	0.853	

=== Confusion Matrix ===

a	b	<-- classified as
16	7	a = ACL
6	66	b = nonACL

References

[1] DataCadamia, "Data Mining - Pruning (a decision tree, decision rules)," https://datacadamia.com , 17 sept 2021. [Online]. Available: https://datacadamia.com/data_mining/pruning#confidence_factor. [Accessed 17 sept 2021].

[2] S. D. a. M. Montag, "Decision Tree Analysis using Weka," University of Miami, Miami, 2006.