Assignment3

Student Name: Hao Bai

Student Number: 180223545

1-2-7:

When the Tree pruning is switch on,

Number of Leaves: 61

Size of the tree: 93

Correctly Classified Instances: 210 90.5172 %

When the Tree pruning is switch off,

Number of Leaves: 121

Size of the tree: 175

Correctly Classified Instances 201 86.6379 %

J48 tree in Weka is used to classify and generate a pruned or unpruned C4.5 decision tree. As can be seen in the summary generated from Weka, the unpruned tree size and number of leaves is larger than pruned tree. The principle of the pruning algorithm is to examine and decide that the branches can be removed without affecting the performance too much.

1-2-8:

The summary generated from Weka shows that the pruned tree's correctly classified instances is 210 which is more than unpruned's correctly classified instances. It demonstrated that the pruned tree's performance is better than unpruned tree. According to Mitchell (1997), the reason is the pruning approach can reduce the risk of overfitting to the training data. The principle of the pruning algorithm is to examine and decide that the branches can be removed without affecting the performance too much.

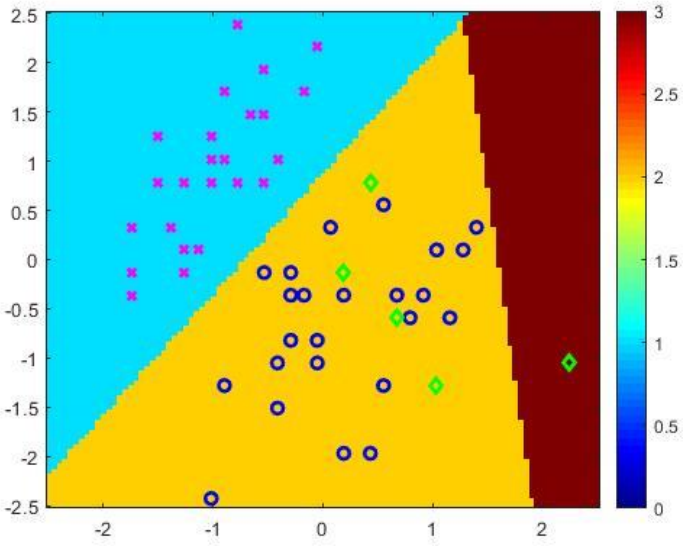
2-4-5:

confusion matrix (flower type 3 is 5 times less common)

25	0	0
0	25	0
0	4	1

Normalise confusion matrix

1.0000	0	0
0	1.0000	0
0	0.8000	0.2000



Train and test accuracy:

- Train accuracy: 0.93
- Test accuracy: 0.93

After normalizing the confusion matrix, third class is heavily mis-classified, the third-class accuracy is 0.2 and other class is 1.0.

The reason is that the third flower type is 5 times less common. In the unbalanced dataset, the classifier will classfy all the third class as other classes.

2-4-6:

The accuracy of the third-class is 20%, the confusion class of the third class is the second class.

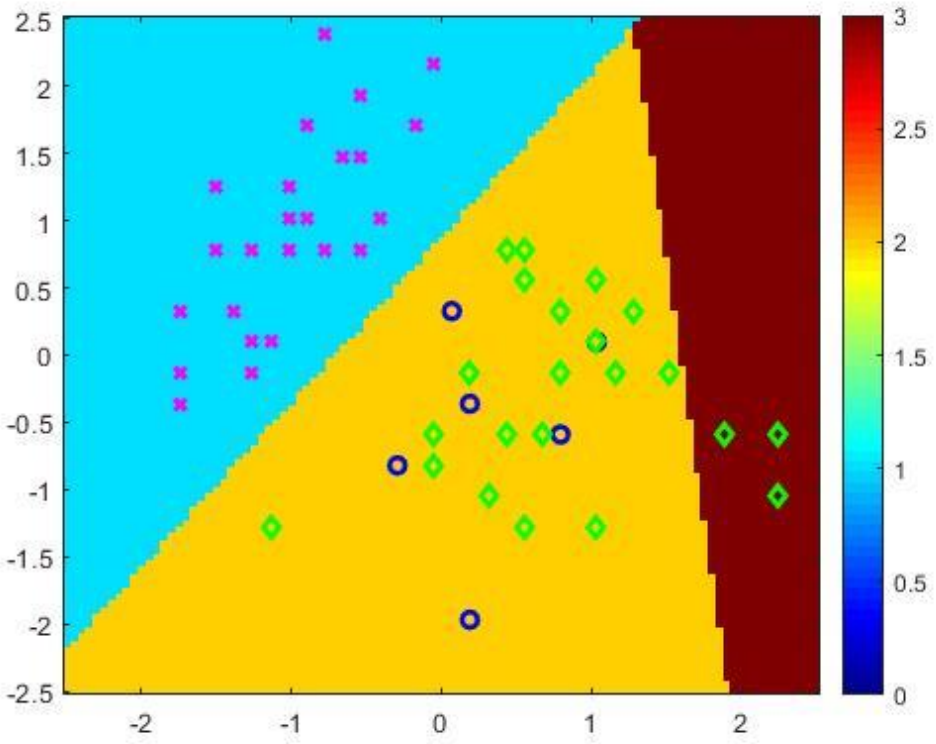
2-5-4:

confusion matrix (flower type 2 is 5 times less common)

25	0	0
0	6	0
0	22	3

Normalise confusion matrix

1.0000	0	0
0	1.0000	0
0	0.8800	0.1200



Train and test accuracy:

- Train accuracy: 0.61
- Test accuracy: 0.68

The new accuracy for class 3 is 0.12, which is less than the previous one. The reason causes that is because of the weight is used the previous weight. The weight is not calculated again. The previous weight will classfy 80% of the third class to the second class (in pervious train dataset, the flower type 3 is 5 times less common). So, the wrong weight will lead the classifier poorly recogenize the flower type 3. However, the new train dataset the amount of the flower type 3 is more than previous one.

2-8-2:

Adjust the priors:

```
S.prob = [1/3,1/3,1/3];
S.prob = [2/4,1/4,1/4];
S.prob = [3/5,1/5,1/5];
```

The three priors have the same train accuracy and test accuracy which is all 0.80.

confusion matrix:

24	1	0
0	6	0
0	10	15

Normalize confusion matrix:

0.9600	0.0400	0
0	1.0000	0
0	0.4000	0.6000

3-5-4:

Adjust the dim to 2:

```
dim = 2;
```

dim = 2, LR Train and test accuracy:

- LR Train accuracy: 0.82
- LR Test accuracy: 0.70

dim = 2, NB Train and test accuracy:

- NB Train accuracy: 0.82
- NB Test accuracy: 0.70

Adjust the dim to 200:

```
1.  dim = 200;
```

dim = 200, LR Train and test accuracy:

LR Train accuracy: 1.00

LR Test accuracy: 0.86

dim = 200, NB Train and test accuracy:

NB Train accuracy: 1.00

NB Test accuracy: 1.00

After change the number of the attributes (dimensions). The train and test accuracy are increase. The large number attributes (dimensions) create a complex model to classifier training data and test data. In lab4_2, the code created a complex model without doing regularize. That might lead to overfitting.

Reference

Mitchell, T. (1997). *Machine learning*. New York: McGraw Hill, pp.69-70.