Classifying Forest Tree: Comparing Genetic Algorithm, Decision Tree & Neural Network

Wenhong Ma u6977108@anu.edu.au
Research School of Computer Science
Australia National University, Canberra Australia

Abstract. Geographical data mapping becomes more and more popular in Forrest management in recent years, dominating the market due to its cost-benefit advantage. However, the accuracy of the result published is not promising, more research is required to produce better results. I attempt to solve this task from different methods in term of tree classification from the image data available.

I am searching for another method to confirm the existing method used in my last attempt and compare the finding with the result generated from my previous methods, for a higher level of confidence in the prediction I make.

I engage the false positive and false negative decision tree classification method for my work on this paper and compare the result generated by the genetic algorithm, as well as the prediction I make engaged neural network at my last attempt.

In this paper, I focus on data dimensional reduction, or so call features selection, using genetic algorithm, to reduce the noise element, for better prediction outcome.

The reason for the selections of methods lays in the requirement of the task I undertook. The task requires me to extract rules from observations and apply the rule on the given dataset, compare the result of the rule extracted model with another specified model such as genetic algorithm and my previous work such as the neural network.

However, the result is similar at around 65%, I could not produce a better prediction than the method already published.

1 Introduction

A cost-effective system will benefit the Forrest management system, due to the higher cost to map large area where the area itself could undergo significant change over time. I intend to deliver one of such system.

I have 190 observations relate to five trees species, each tree(target) has 16 features associate with it.

My goal is to select the best feature set among the 16 feature for tree prediction against the baseline prediction, divide the 190 observation into the training set and the testing set, in the ratio of 80;20 respectively. The provided features are altitude, aspect, slope, geology, topographic position, rainfall, temperature and Landsat TM band 1 to 7(L.K. et al.1995). The five classes of species to predict are scrub(SC=0), dry sclerophyll(DS=1), wet-dry sclerophyll(WD=2), wet sclerophyll(WS=3), rainforest(RF=4). The hypothesis is that the new model will produce better prediction in the absence of noisy features.

I attempt to achieve my goal by engaging the genetic algorithm and extract rules from the 190 observations, apply the model from the extracted rule and compare results, taken into account the result of my previous neural network model.

However, classifying Forrest spicy could not be done from satellite data alone without supplement data such as photographs(L.K.Milne Citi Skidmore and Turner,1988), if a more accurate result is sought. (L.K.Milne Citi Skidmore et, al,1994)

I search and compare method that produces a more accurate result for species prediction. Among many methods, I select the neural network, decision tree and genetic algorithm, after thorough and careful data processing and data normalisation.

2 Data Preparation

Data used in this article come from the geographical data of Nullica State Forest, the south coast of New South Wales, an area occupying 20 by 10 km, divided into 179831 pixels in the shape of 30 by 30 m in size. Data analyse include satellite image, soil map and aerial photographs, the latent could be used to produce a terrain model. (L.K. et al. 1995)

The data is un-even distributed, the majority of observations relate to two trees only. It would be desirable to normalise those data before further processing to achieve a satisfactory prediction result.

The statistic of the data as follows:

	count	mean	std	min	25%	50%	75%	max
AS	190.0	43.368421	23.687137	0.0	20.00	40.0	60.00	80.0
SA	190.0	56.657895	29.836660	0.0	35.00	50.0	85.00	99.0
CA	190.0	56.652632	31.761935	0.0	35.00	50.0	85.00	99.0
AL	190.0	33.984211	13.268284	7.0	23.25	33.5	43.00	71.0
TP	190.0	65.768421	23.983880	16.0	48.00	64.0	96.00	96.0
SL	190.0	53.473684	11.525763	10.0	50.00	50.0	60.00	80.0
GE	190.0	63.105263	22.262015	10.0	50.00	70.0	90.00	90.0
RA	190.0	39.984211	15.622354	19.0	22.00	39.0	55.75	79.0
TE	190.0	54.315789	17.885742	0.0	30.00	60.0	60.00	90.0
T1	190.0	58.242105	3.389107	53.0	56.00	58.0	59.00	91.0
T2	190.0	19.936842	2.433499	16.0	19.00	20.0	20.00	46.0
T3	190.0	19.805263	4.390824	14.0	18.00	19.0	21.00	66.0
T4	190.0	53.952632	9.677491	25.0	48.00	54.0	60.00	80.0
T5	190.0	34.894737	11.311345	13.0	27.00	33.0	41.00	92.0
T6	190.0	38.921053	8.190786	15.0	36.00	39.0	42.00	63.0
T7	190.0	30.557895	9.680850	16.0	24.50	28.0	32.00	90.0
Specy	190.0	1.747368	1.093206	0.0	1.00	1.0	3.00	4.0

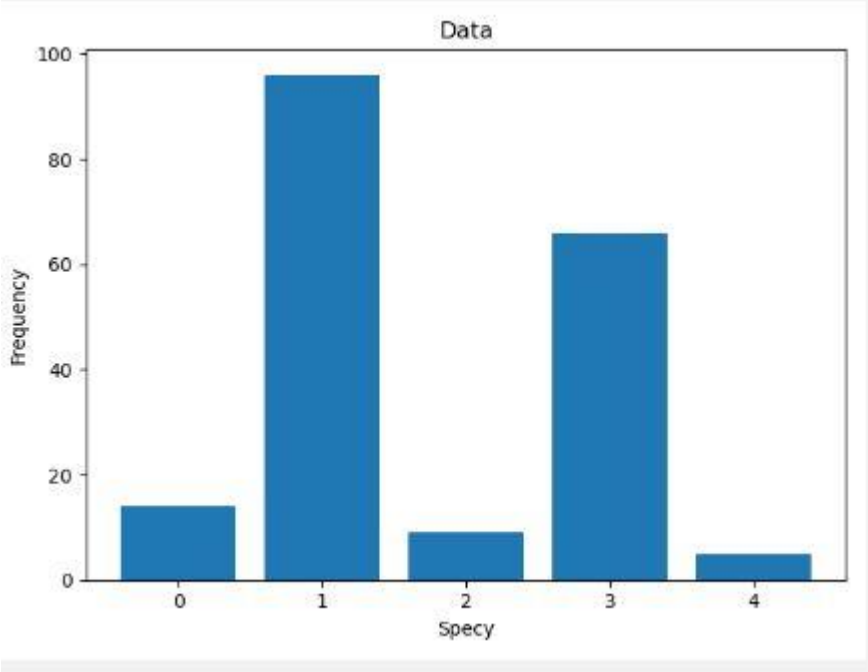
Unbalanced data need to be balanced before processing. The data had been preprocessed using the normalisation method, the spread of data had been smooth out between 0 and 1.

```
x_norm = np.linalg.norm(x_array, ord=1, axis=1, keepdims=True)

x = x_array/x_norm

x_array are features used to predict the target.
```

The distribution of the data collected is not normal, rather, there are many instances in two species, and few instances for the other three. Show as below:



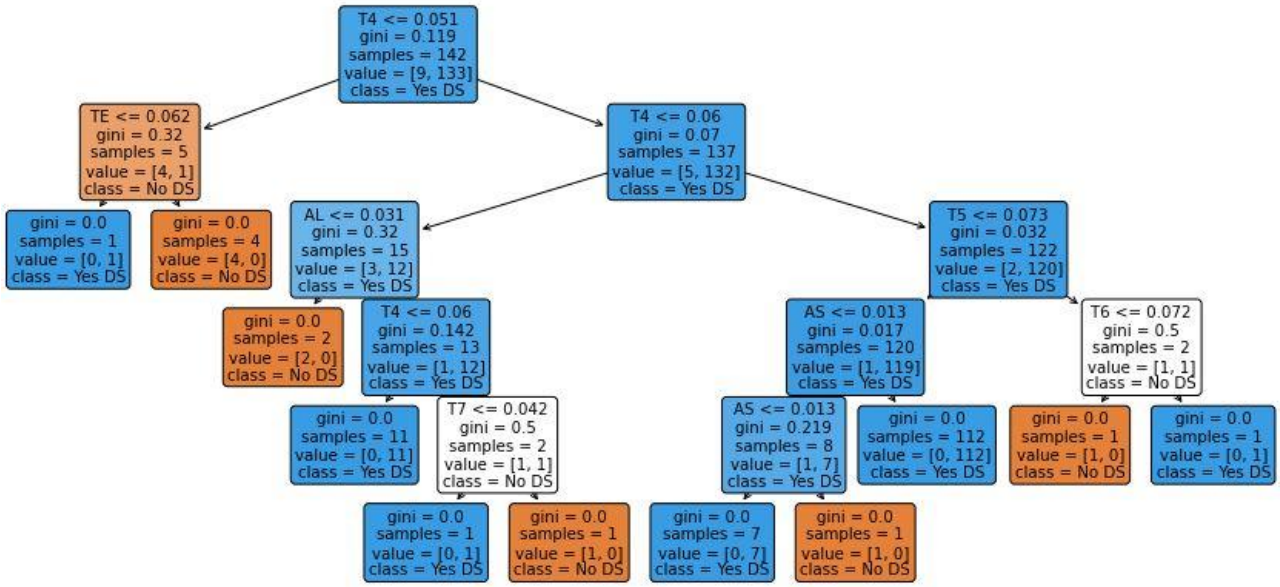
I produce 2 sets of data from raw data for comparison purpose, one set with the target as an integer, named as gis-data-2, another set the target is an array, named as gis-data, features in both sets of data are arrays.

3 Method

3.1 Decision Tree Classification

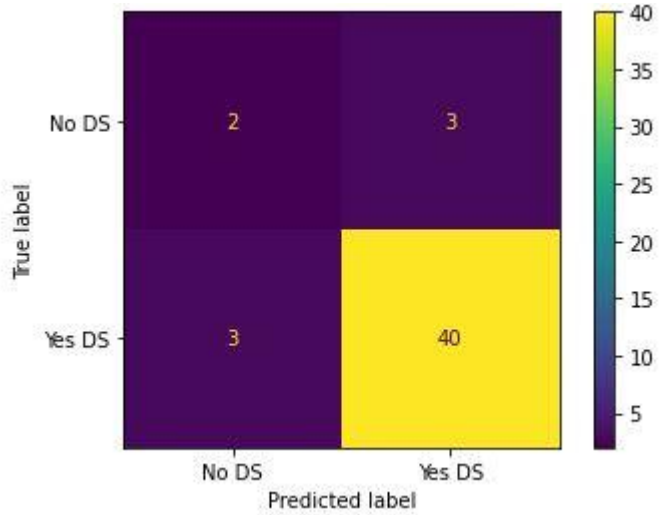
I engage the decision tree algorithm to gain knowledge from the observations, which extract rules from observations and apply the rules to products the true and false table, for comparison. I only extract the rule relate to DS, the most occurrence tree, then compare the result from the neural network I produced at my last attempt, the neural network could only predict DS and WS because there are reasonable data available for those two trees, as shown at the Data graph above. The neural network fails to predict SC(0), WD(2) and RF(4) in the absence of sufficient data available.

I employ the DecisionTreeClassifier and plot_tree from Sklearn_Tree for the decision tree classification, set DS as 1 and the rest of the trees as 0, to extract rules from the 190 observations. The rules extracted as follow



Apply the rules extracted from the dataset, using plot_confusion_matrix from the Sklearn_Metrics, I generate the following predicted outcomes, after randomly allocate the training set and the testing set of the full set of data:

There is a total of 48 predictions made. Out of them, 40 predictions correctly predict DS is DS, which is 83%, 3 out of 48, or 6% predict DS as No DS, 6% predict No DS as DS, and 2 out of 48, or 4% correctly predict No DS as No DS, a total of 87% predictions are correct.



It is not easy to obtain a higher accuracy rate for some species (No. 0,2 and 4), because the data is not normally distributed, and some species (No. 0, 2, 4) do not occur often, as shown as the data distribution graph, make the data available is relatively small.

DS is the most occur observation, I use it to compare with the result generated from other methods.
The result of 87% correct prediction is far better than the neural network at my last attempt, which is 79% at the best, the combination of precision, recall and f1-score for DS, however, the genetic algorithm produce a general result of 63%, not just the prediction relating to DS, which makes them not comparable to each other.

3.2 Neural Network

I keep my last attempt using the method of the neural network, for comparison purpose.

The Neural network model is a sequential model using softmax as activation function and categorical cross-entropy as loss function, adam as the optimizer. I use the softmax function and cross-entropy function to compute multi-classification output, rather than binary output, as the targets are 0,1,2,3, or 4. I use adam as the optimizer for speedy calculation with acceptable accuracy. The neural network model takes 16 input features, through 2 hidden layers, to produce the output which is the target being predicted. The full 190 sets of observation had been divided into the training set and the testing set at the ratio of 70: 30, using the train_test_split function imported from the Sklearn. model-selection.

The outcome of the classification report and confusion matrix as the following:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	4
1	0.54	0.79	0.64	28
2	0.00	0.00	0.00	1
3	0.44	0.32	0.37	22
4	0.00	0.00	0.00	2
accuracy			0.51	57
macro avg	0.19	0.22	0.20	57
weighted avg	0.43	0.51	0.46	57

On the table, it is easy to observe the species No.1 and 3 could be reasonably predicted, species No.0,2,4 could not be predicted, due to their low occurrence, being 4, 1 and 2 occurrences in the testing set.

3.3 Genetic Algorithm

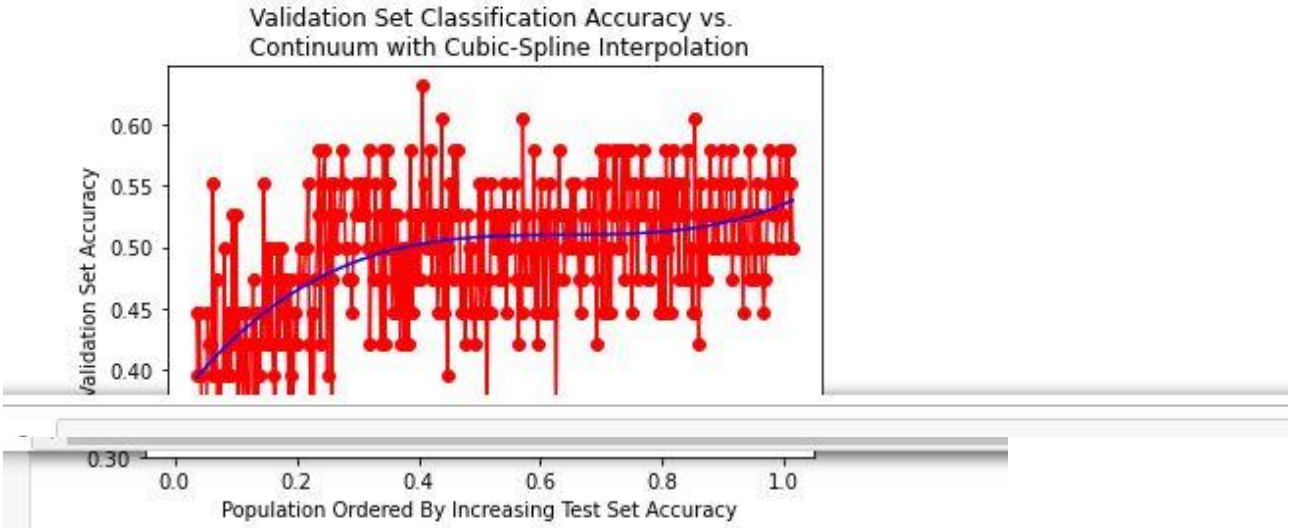
I allocate 20% of data for testing and 80% of data for training, use train_test-split from the sklearn_model_selection method.

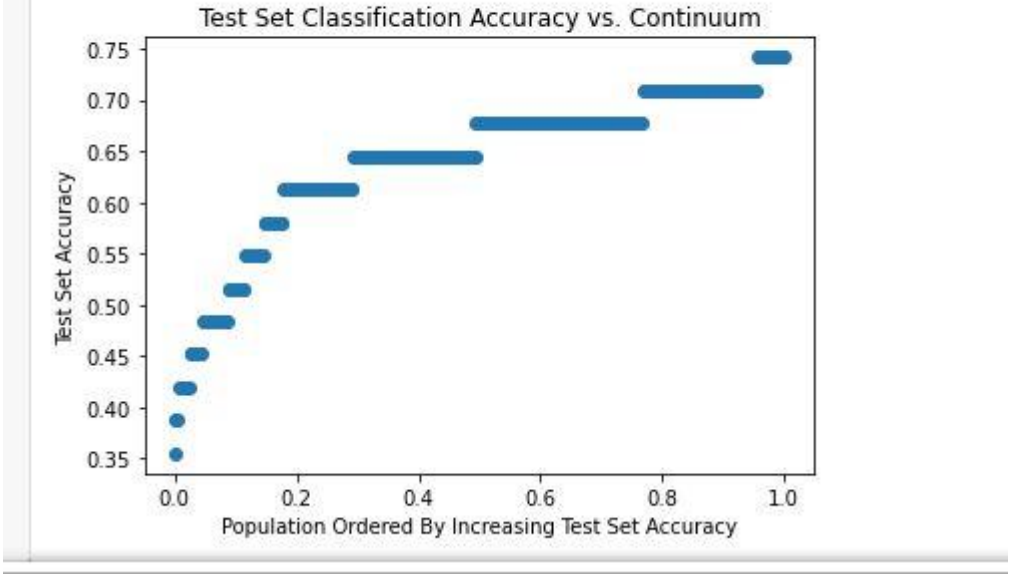
I import creator, base, tools and algorithms from the Deap and use its code as much as possible in term of creating individual and fitness, as well as creating toolbox such as the population, evaluate, mutate and selection; I choose Deap because of its rich library.

I programme the GA model in the following order. I first choose the fittest parent by applying the one-hot encoding to the features, then apply logistic regression on the data, and calculate accuracy; I then create Individual and toolbox, set population to 100 and generation to 10, then apply the genetic algorithm to choose the best individuals; afterwards, I obtain the list of percentiles in the best individual and the fitness data from each percentile; finally, I apply logistic regression using all the features to acquire a baseline accuracy, then apply a genetic algorithm to choose a subset of features that gives better accuracy than the baseline model, obtain a list of subsets that performed best on validation data; In conclusion, I plot the test and validation classification accuracy to see how these numbers change as I move from my worst feature subsets to the best feature subsets found by the genetic algorithm, calculate the best fit line for validation classification accuracy.

The output of the GA model as follow:

```
Percentile:                    0.3787528868360277
Validation Accuracy:           0.631578947368421
Individual:     [1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0]
Number Features In Subset:     10
Feature Subset: ['AS', 'SA', 'CA', 'TP', 'SL', 'GE', 'RA', 'T2', 'T5', 'T6']
```





From the graph, I observe the model produces the test set accuracy up to 75% and 63% on the validation set.

The reason to engage the GA model is to determine the features set that produce the most accurate prediction, to avoid the noise generated by the mass dataset.

The result of the GA model sits between the lower and the higher accuracy of the Neural network model, which is 54% and 79%, with a middle of 64%, not as good as the decision tree model; however, bear in mind, the decision tree model only be used to predicted DS species, the most occurrence one, it would not produce a similar result on the fewer occurrence species.

4 Result, Discussion, Limitation

I perform analysis on data of NSW forest augmented through the photograph. I apply neural network, classify decision tree and genetic algorithm.

Among the three methods, the results are similar. The neural network has the worst performance, with a 51% accuracy; the genetic algorithm performs slightly better at 63%, while the decision tree outperforms the rest, as high as 87% for its best prediction.

However, all methods are not good to predict the species which does not occur often, the lack of occurrence cause Deep learning to fail to learn from limited observation.

The genetic algorithm model employed is useful for complex matter such as this one, as many features element are just noise, they have little or no effect on the target, it is wise to remove them from the features used for prediction purpose, otherwise, PCA may need to reduce higher dimensional data to lower-dimensional one.

Comparing the result produced by the C4.5 model(L.K. et al.1995), the training set correctness is 109/(109+81) = 57%, the testing set correctness is 46/(46+24)= 66%, my relevant performance in Decision tree rules extracted model produces 87% accuracy on DS tree, well outperform their C4.5 model, however, this is not a direct comparison as my model only predict the most occurrence tree.

Comparing the result produced by the Maximum likelihood model(L.K. et al.1995), the training set correctness is 124/(124+24+42) = 65%, the testing set correctness is 42/(42+14+14)= 60%, my relevant performance in genetic algorithm model obtain 63% on the validation set and 75% on the testing set, slightly outperform his model.

Comparing the result produced by the neural network model(L.K. et al.1995), the training set correctness is 52%, the testing set correctness is 65.7%, my relevant performance in the neural network is 51% for the testing set, far worst than the result produced by his paper.

The loss performance of my neural network lay in the lack of complicated data processing and parameter fine-tune, as well as the model training itself.

The limitation of this research is the smaller sample size, machine learning relies on large data, a smaller and unnormal distributed sample is not a great way to learn.

4 Conclusion and future work

The result of the genetic algorithm model does not produce far more superior result than the Decision tree classification model in the scene because there is no direct comparison between them, where the Decision tree only predicts one tree and the genetic algorithm model predict five trees as a whole; however, the genetic algorithm model does outperform the neural network model by a reasonable margin.

The next stage of my work will be to train the genetic algorithm for better performance, such as add DNA size, Crossover rate and Mutation rate, adjust parameters.

My hypothesis, the genetic algorithm outperforms my previous neural network model, has been established.

I conclude the genetic algorithm outperforms the neural network model in term of tree prediction, on this particular dataset.

References

1. L.K. Milne , T.D.Gedeon and A.K.Skidmore , Classifing Dry Sclerophyll Forrest from Augmented Satellite Data: Comparing Neural Network, Decision Tree & Maximum Likelihood. Proceedings International Conf. on Artificial Neural Networks and Genetic Algorithms (ICANNGA), Alès, 1995.
2. Skidmore.A and Turner, BJ "Forest Mapping Accuracies are improved using supervised non-parametric classifier with spot data." Photogrammatic Engineering and Remote Sensing. Vol.54. no. 10, pp. 1415-1421,1988
3. Skidmore. AK.Brinkhof.W and Delancy.J "Using Neural Networks to Analyse Spatial Data." Processing 7th Australialasian Remote Sensing Conference. Melbourne, pp.235-246, 1994.