CLASSIFYING DRY SCLEROPHYLL FOREST FROM AUGMENTED SATELLITE DATA WITH NEURAL NETWORK, DECISION TREE AND SUPPORT VECTOR MACHINE

Siyu Du

School of Computer Science and Engineering, The Australian National University, The Australian National University Canberra ACT 2600 Australia
U6510288@anu.edu.au

Abstract. A detailed and accurate forest image comes increasingly valuable for forest management, such as forest fire and wild animal protection. This paper uses different current techniques to find a reliable and accurate model for forest classification which can be used to generate forest maps with additional available data. The raw data is collected by satellite, and data transportation is performed to fit into models. For each technique, we describe adjusting parameters to achieve a best model.

Keywords: Classification, Neural Network, Forest Type, Decision Tree, SVM.

1 Introduction

It is extremely expensive to generate map by surveying large size of lands. Therefore, it would save a huge amount of human resource and government budget to generate maps using data mining techniques based on current available data. In this paper, we explore the geographical data from NSW and use Neural Network, Decision Tree, SVM to perform classification. Rather than classifying five different forest types, this paper focus on the task to classify only Dry forest type. During the research, we experiment different parameters and methods to improve the stability and accuracy of the model, and show the result in figures and tables.

1.1 Data Introduction

The geographical data used in this paper is collected from Nullica State on the south coast of New South Wales. The size of land being analyzed is about 20 by 10 km and structured into a gird of 179831 pixels, 30m by 30m in size. The data sources include satellite imagery, soil maps and aerial photographs. The aerial photographs provides the possibility to derive a terrain model and based on this derive to extract several terrain features. Totally, 190 pixels have been surveyed in details as samples for training. For each pixel, there are 22 attributes including altitude, aspect, sin & cos of aspect, slope, geology, topographic position, rainfall, temperature, Landsat TM bands 1 to 7, and five forest supra-types.

1.2 Data Exploration and Preparation

The raw data has been preprocessed using cumulative histogram enhancement technique (Richards, 1986). For the purpose of building a Neural Network with high accuracy and performance, certain data preparations and encoding are necessary.

Aspect is represented as degrees from 0 as flat, 10 as north to 80 as northwest. From figure 1, we can see that the aspect degree distribution of the sample data is close to normal, and the samples with 0 is redundant as Slop Degree presents such information as well. It is a ‘circular’ column having values pointing to eight directions, and in this paper, we present aspect as a category variables. Four new variables A1, A2, A3 and A4 are created presenting each point of the compass showed in table 1.

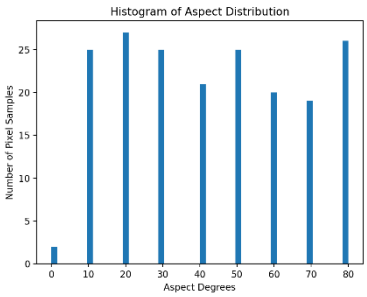

Fig. 1. Aspect degree distribution

As	Compass	A1	A2	A3	A4	Activ.
0	Flat	0	0	0	0	0
10	N	1	0.5	0	0.5	2
20	NE	1	1	0	0	2
30	E	0.5	1	0.5	0	2
40	SE	0	1	1	0	2
50	S	0	0.5	1	0.5	2
60	SW	0	0	1	1	2
70	W	0.5	0	0.5	1	2
80	NW	1	0	0	1	2

Table. 1. Aspect encoding

The range of Altitude is from 7 to 71, and its distribution is slightly skewed to the right. By contrast, the Slope appears to slightly skew to the left, but generally normally distributed. Temperature is equal to (degree – 11) * 30, and there are 4 values 0, 30, 60, 90 in the data set, which means surveyed temperature only ranges from 11 to 14 degree. The Rainfall is equal to (mm - 801) /5, and it is presented as a continuous variable with minimal value 19 to maximum value 79. The distribution of Rainfall is not normal having 40 percent of data with value 39, 25 percent with value 29 and 19 percent with value 59. The rest of samples only have significant low frequency closed to 0. For Altitude, Slope and Temperature, we normalize these attributes between 0 to 1 as continuous inputs for the network, since the logistic function is used.

Geology descriptor has category like data from 10 to 90 with only three major significant values 50, 70 and 90. There is no detailed information regarding this attribute, and it is difficult to know relation between these categories whether they are continuous or not. Therefore, we present the attribute by 4 inputs with value 50, 70, 90 and the rest. Similarly, the Topology position is also a category variable with mapping 32: gully, 48: lower slope, 64: mid-slope, 80: upper slope, 96: ridge. Considering the categories are related and continuous, we normalize the attribute from 0 to 1 as input. Regarding Landsat bands1 to 7, we simply linear squash to 0-1 as well.

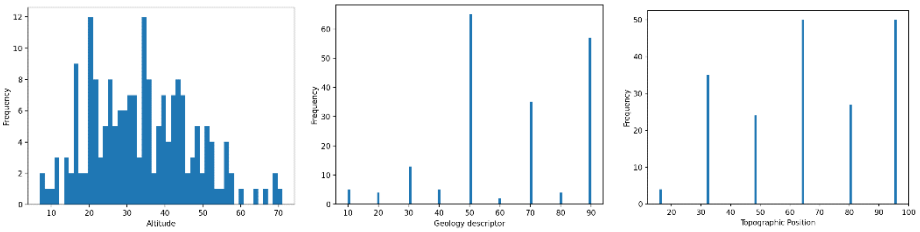

Fig. 2. Altitude, Geology descriptor and Topographic position distribution

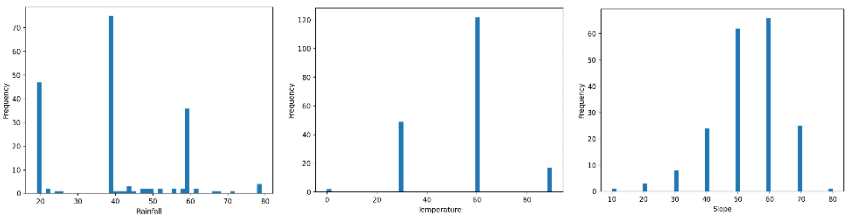

Fig. 3. Rainfall, Temperature and Slope distribution

The network is classifier to retrieve the forest supra-type. The raw data has five columns Scrub, Dry-sclero, Wet-dry, Wet-sclero and Rain forest with 90 value as true and 10 as false. It is difficult for network to learn using raw output, so this paper transforms five categories into four output units as table 2 (Lecture NN5 Hidden units). The final category is retrieved by calculating Euclidean distance between network output vector and four unit vector. However, due to unbalance category size, another task of the paper is to classify dry category, and new column created with 0 of non-dry and dry as 1.

Category	Unit 1	Unit 2	Unit 3	Unit 4
Scrub	0.184	0.317	0.371	0.4
Dry Sclerophyll	0.816	0.317	0.371	0.4
Wet-dry sclero.	0.5	0.865	0.371	0.4
Wet sclerophyll	0.5	0.5	0.887	0.4
Rain Forest	0.5	0.5	0.5	0.9

Table. 2. Output encoding

2 Methodology

2.1 NEURAL NETWORK

A number of neural network topologies are tested, and the result of network has no significant difference. Four different learning rates are tested from 0.05, 0.035, 0.005 and 0.0001. It shows that 0.035 has the best learning rate shape, when use 0.05 and higher the shape of loss spikes first and with the same performance with 0.035. However, this model is still suffering the stability and overfitting issue. To tackle overfitting, the weight delay technique is used. When we use AdamW optimizer and have tried weight decay as 0.1, 0.5 and 1, the 0.5 decay rate performs relatively better with other two. Showed in figure 9, the training accuracy 77% on average of 30 runs is slightly higher than test accuracy 69%, and also the learning rate on loss function appears faster.

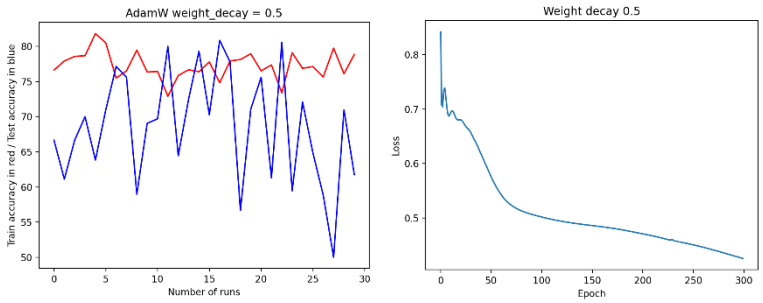

Fig. 4. Accuracy and Loss function of final model with weight decay = 0.5

2.2 DECISION TREE

The algorithm used to generate decision tree is C4.5 developed by Quinlan 1995. It is an extension of ID3 with capability to have categorical and numerical inputs, which is also called statistic classifier. In this method, information gain approach is used to decide the tree nodes and their property, thus we select geographic attribute with highest information gain (entropy reduction in the level of maximum).

There is a significant different on accuracy when use different data sample method. Firstly, we use fixed proportion of train and test data. The accuracy is 61% on test data, which is relatively lower than Neuronal Network. Moreover, we change the data sampling strategy to cross-validation with 10 training subsets and 1 test subset. The accuracy increases dynamically up to 90%. The result is very satisfying, but noticeably the miss classifications are all on dry forest type based the confusion table on figure 5.

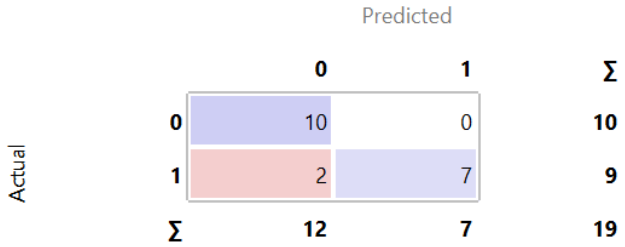

Fig. 5. Decision Tree confusion matrix

2.3 SUPPORT VECTOR MACHINE

Support Vector Machine was first proposed by Vapnik in 1992 to build a non-linear classifier. One important property of SVM is to minimize the error of classification error and maximal the geometric distance at the same time, so it is also called Maximal Margin Classifier. In our case, SVM needs to find a separating hyperplane in a higher dimension with 17 inputs. Different popular kernel functions have been tested and there is no significant different. Sigmoid kernel is used in the end, $K(xi, xj) = \tanh(\gamma\, xi^T xj + r)$.

The result of the model is 68% accuracy on test data and 72% on training data. Moreover, From the confusion matrix in figure 6, we can see that the recall and precision is same with 72%. As a result, we can see that the SVM provides a relatively balanced model.

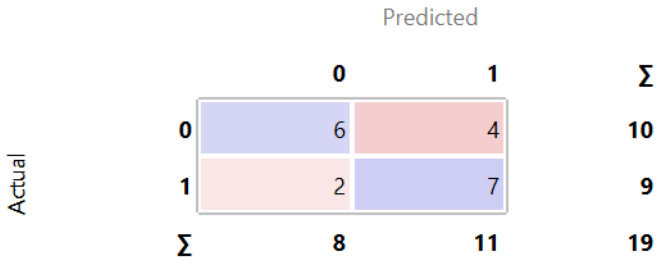

Fig. 6. SVM confusion matrix

3 Results and conclusion

The similar research has been done in previous report (L.K. Milne 1995). Three classification techniques including Decision Tree, Maximum Likelihood and Neural Network used to classify pixels containing dry sclerophyll forest. Statistically, there is no significant difference on Neural Network and SVM. However noticeably, the accuracy of Decision Tree is much higher with 90% accuracy, which can be considered a very successful model to classify forest type.

In this paper, we have used satellite imaged data of a NSW state forest augmented by ancillary data derived from aerial photography and other available information. Although there are five forest types available in the raw dataset, only pixel classification model on Dry has a relatively higher accuracy. This is due to lack of sample data for other forest categories. We have showed a number of methods including data transformation, hidden size adjustment and optimizer modification to tackle overfitting and under-fitting problems. The result is satisfying using Decision Tree with 90% accuracy on test data.

The next stage, we will use different validation methods to verify the model and make further action to improve the stability. Nosie removal and data generalization is also needed to increase the test accuracy, and we should collect more surveyed pixels if possible.

References:

L.K. Milne “CLASSIFYING DRY SCLEROPHYLL FOREST FROM AUGMENTED SATELLITE DATA: COMPARING NEURAL NETWORK, DECISION TREE & MAXIMUM LIKELIHOOD” 1995

Richards, J “REMOTE SENSING DIGITAL IMAGE ANALYSIS, SPRINGER” Verlad, 2nd ed., 1993

QUINLAN, J.R “C4.5 PROGRAMS FOR MACHINE LEARNING” SAN MATEO, CA: MORGAN KAUFMANN, 1992