Feature Selection Using Neural Networks with Contribution Measures

Linda Milne
Computer Science and Engineering
University of New South Wales
Sydney NSW 2052
linda@cse.unsw.edu.au
tel +61-2-385-3979
fax +61-2-313-7916

Abstract

There still seems to be a misapprehension that neural networks are capable of dealing with large amounts of noise and useless data. This is true to a certain extent but it is also true that the cleaner and more descriptive the data is the better the neural networks will perform, especially when dealing with small data sets. A method for determining how useful input features are in giving correct classifications using neural networks is discussed here.

1 Introduction

There still seems to be a degree of unrealistic optimism regarding the abilities of neural networks. To push them to the limits of their capabilities we need to recognise that the cleaner our data sets and using only the significant features in training and classification will give us the best possible results.

Many data sets contain large amounts of noise. It is tempting to throw as many features at the neural network as possible in the hope that it will be able to work out what is significant and give a classification. As we do not have a good understanding of how neural networks work it is difficult to know which of the features available are the most useful in describing the key properties of the input vectors class, this will also depend on the domain the data comes from.

It is possible that using the same method for feature selection and classification produces better results [1]. A neural network may use a different set of features to a machine learning algorithm so to use some other feature selection method may remove data that is useful and keep data that is not. So we use the neural network to help us decide which are the most useful features in giving a classification.

By giving a measure of the contribution each input feature makes to the final output of the network we can select the features to use. This will reduce the noise and extraneous information that the network has to deal with as well as reducing training and classification times.

2 Calculating Proportion Contribution of Input Features to Outputs

The networks used here consist of 3 layers, an input layer, a single hidden layer and an output layer. Assuming the following values

ninputs	the number of inputs
nhidden	the number of hidden units
noutputs	the number of outputs

the input units are numbered from 1 to ninputs, the hidden nunits are numbered from ninputs+1 to ninputs+nhidden and the output units are numbered from ninputs+nhidden+1 to ninputs+nhidden+noutputs. The weight between unit i in layer n and unit j in layer *n+1* is given by w_{ji}.

Garson [2] proposed the following measure of proportion contribution. The contribution that input unit i makes to output unit o is

$$\frac{\sum_{j=1}^{nhidden} \frac{w_{ji}}{\sum_{l=1}^{ninputs} w_{jl}} . w_{oj}}{\sum_{k=1}^{ninputs} \left(\sum_{j=1}^{nhidden} \frac{w_{jk}}{\sum_{l=1}^{ninputs} w_{jl}} . w_{oj} \right)}$$

This method will not give a true proportion when there is a combination of positive and negative weights.

Another measure of contribution, used in [3], gives the proportion contribution of a unit in one layer to a unit in the next layer. For example, the contribution of input unit i to the hidden unit h would be

$$\frac{|w_{hi}|}{\sum_{l=1}^{ninputs} |w_{hl}|}$$

A possible disadvantage of this is that the sign of the contribution is lost.

A better measure of the proportion contribution of input i to output o would be

$$\frac{\sum_{j=1}^{nhidden} \frac{w_{ji}}{\sum_{l=1}^{ninputs} |w_{jl}|} . w_{oj}}{\sum_{k=1}^{ninputs} \left(\sum_{j=1}^{nhidden} \left| \frac{w_{jk}}{\sum_{l=1}^{ninputs} |w_{jl}|} . w_{oj} \right| \right)}$$

as it takes into account the fact that weights can be positive or negative, and gives a true proportion.

To use this measure for feature selection, inputs that have a contribution close to zero can be left out of the training. A large negative value tends to decrease the size of the output while a large positive value tends to increase the size of the output. It is possible that the features that make positive contributions give the overall characteristics for a particular class while the negative values adjust for special cases.

3 The Data

This work is part of a project investigating neural networks to generate maps of species occurrence in eucalypt forests. The area being studied is the Nullica State Forest on the south coast of New South Wales, Australia. The data consists of 190 training vectors and 70 test vectors that have been generated from surveys of the area. It is important to note that it is difficult to generate more training data as it would require further surveys of the area. The area is approximately $20km$x$10km$ of fairly rugged country so to generate large amounts of training data would be prohibitively expensive in time and money. As we have only a small data set it is even more crucial that we have the cleanest training set possible. Other methods are also being investigated to give the best possible classifications.

The the features available in the data are aspect, altitude, slope, topographic position, geology type, rainfall, temperature, and Landsat TM bands 1 to 7. These values are scaled to range between 0 and 1. The aspect is represented as a vector to reflect the circular nature of this feature.

It is known that some of the features are noisy. For example, the geology map of the area is estimated as only about 40% correct. Features such as rainfall and temperature are derived from data from a small number of weather stations close to the field area and models of weather patterns. Thus, it may be desirable to remove some of the features from the training.

Initially the aim is to produce a set of general classifications which will be used later in the actual species classifications. The possible classes are scrub (SC), dry sclerophyll forest (DS), wet/dry sclerophyll forest (WD), wet sclerophyll forest (WS) and rainforest (RF).

4 The Neural Network

The networks used consist of a 17 unit input layer, 14 unit hidden layer and a one unit output layer trained using back-propagation. Each network is trained to recognise a single class. Each input vector is given a class of 0.9 if it is in the given class and a class of 0.1 if it is not. To determine class membership of the output of the trained neural network a threshold is chosen so that the number of correct classifications is maximised [4]. To avoid over-fitting the data, cross validation is used to determine when to stop training.

The area being studied is predominantly dry sclerophyll forest. Of the 190 training vectors 99 are in the DS class and 46 of the test vectors are in the DS class. For this reason the data has been split into DS and not DS classes with alternative neural network methods being considered to produce the other classifications.

5 Contribution Calculation and Feature Selection

Five networks, with different initial weights, were trained for the DS class. The graph in figure 1 shows the contribution of each input for the trained networks.

As can be seen each feature shows a similar level of contribution to the output for each of the networks. This shows that, in general, the initial weights have no effect on the contribution of a feature for a single output/2 class network. The candidates for removal from training and classification could be *slope* and *tm7* as their contribution to the output is close to zero.

A “large” variation in the contributions for a particular feature over different networks may also indicate that the given feature is not important in the classification. For example, a feature with a large variation in the contribution is *tm2*. This is possibly a better candidate for removal than say altitude as its contribution is close to zero.

It is also possible that only the features with a large positive or negative contribution are useful. So the most significant features may only be *tp*, *alt*, *tm3* and *tm6*.

6 Removal of Features

A further four training runs were carried out. Firstly *slope* and *tm7* were removed, then *tm2* alone was removed, and then all three features were removed, from both the training

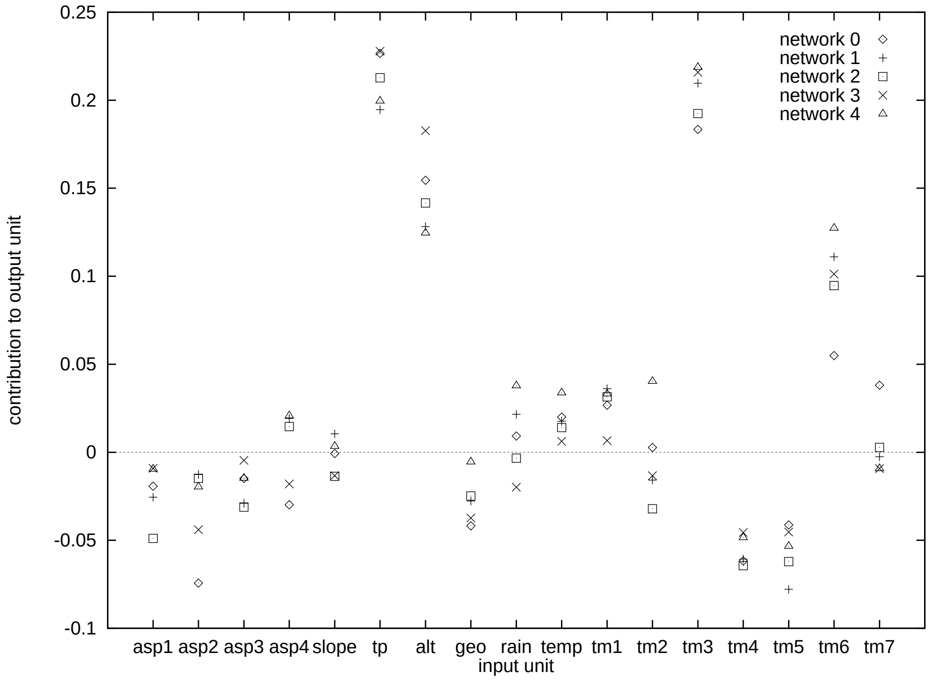

Figure 1: contributions of the input features for 5 networks trained to recognise dry sclerophyll forest

and the test sets. The final training run used only the *tp*, *alt*, *tm3* and *tm6* features, all of the other features were removed from the training and test sets. The average number of correct classifications for each of the five training runs are given in table 1.

7 Discussion and Conclusions

As can be seen the number of overall correct classifications does increase. Although the results for the not DS class in the test set are still reasonable the drop in the number of correct classifications will need to be investigated further. The numbers of correct classifications given by the individual networks can also be fine tuned by adjusting the class membership threshold.

Even with removing all but four of the features we can still get an improvement in the results. We get at least as good performance as using all the data, but not as good as removing only three of the features. This seems to indicate that the other features while not completely superfluous are very noisy or contain only small amounts of information, possibly even both. So, as expected, as we remove too many features the accuracy drops.

The variation of the contribution for different networks is also useful in determining the useful features for correct classifications, as can be seen with the significant improvement

	all data	slope, tm7 removed	tm2 removed	slope, tm2, tm7 removed	tp, alt, tm3, tm6 only
av DS correct	52%	64%	66%	65%	63%
av not DS correct	83%	77%	79%	76%	83%
av correct	66%	70%	72%	70%	72%

(a) training set classification

	all data	slope, tm7 removed	tm2 removed	slope, tm2, tm7 removed	tp, alt, tm3, tm6 only
av DS correct	43%	66%	63%	64%	54%
av not DS correct	86%	71%	70%	73%	67%
av correct	58%	55%	65%	67%	58%

(b) test set classifications

Table 1: average number of correct classifications

when removing the *tm2* data.

Removing features with contribution close to zero and those with high variation in contribution still produces good results. i.e. the removal of *slope*, *tm2* and *tm7*. It is yet to be determined at what point the contribution of a feature is close enough to zero and how much variation is required for a feature to be removed. It would not be as desirable to remove features that have a large contribution with variation as the fact remains it makes a large contribution. Of course, the number of input features that need to be removed will be determined by the data being used in a particular application.

John et al [1] gave the following definitions for the relevance of features in a classification and a method for feature selection using induction algorithms.

strongly relevant feature — the feature is necessary and can not be removed without decreasing the number of correct classifications
weakly relevant feature — the feature sometimes contributes to the classification
irrelevant feature — the feature will never contribute to the classification

This provides a possible explanation of the behaviour of the neural network feature selection.

i. non-zero contribution/little variation determine the strongly relevant features
ii. zero contribution/little variation determine the irrelevant features
iii. variation/close to zero contribution determine weakly relevant or irrelevant features
iv. variation/non-zero contribution determine weakly relevant features or strongly relevant features

The results from using only *tp*, *alt*, *tm3* and *tm6* seem to indicate that these features are strongly relevant. The results for the other networks with data removed indicate that the remaining features are either very noisy, weakly relevant or possibly even irrelevant. Methods of removing noise from the data are being investigated and so it may be possible to determine which of the features truly are irrelevant.

For networks with more than one output the problem of feature selection becomes more difficult. In the case where a network is to produce a number of classifications it may be that the optimal solution requires removal of different features for different classes. Thus, splitting up the classes over several networks, as is done here may be necessary. It will be necessary to develop methods of selecting features for removal that are meaningful for more general networks.

It would also be useful to incorporate information about how the outputs change over the range of input values. To this end WV-curves [5] as part of the feature selection process is currently being investigated.

It has been shown that proportion contribution is a useful measure when determining which of the possible features are the most useful for describing the class of a particular input vector. This method will, of course, need to be tested on further data sets. It is hoped that both the proportion contribution and WV-curves will provide a robust feature selection tool.

Acknowledgements

I gratefully appreciate the financial support provided by Andrew Skidmore (School of Geography, University of New South Wales) and Brian Turner (Dept of Forestry, School of Resource and Environmental Management, Australian National University).

References

[1] John GH, Kohavi R, and Pfleger K. Irrelevant features and the subset selection problem. In *Proc 11th International Conference on Machine Learning*, pages 121–129, 1994.

[2] Garson GD. Interpreting neural-network connection weights. *AI Expert*, pages 47–51, April 1991.

[3] Wong PM, Gedeon TD, and Taggart IJ. An improved technique in porosity prediction: A neural network approach. *IEEE Transactions on Geoscience and Remote Sensing*, (in press) 1995.

[4] Milne LK, Gedeon TD, and Skidmore AK. Classifying dry sclerophyll forest from augmented satelite data : Comparing neural network, decision tree & maximum likelihood. In *Proc. 6th Australian Conference on Neural Networks*, pages 160–163, 1995.

[5] Bischof H, Schneider W, and Pinz AJ. Multispectral classification of landsat images using neural networks. *IEEE Transactions on Geosciences and Remote Sensing*, 30(3):482–490, May 1992.