Attribute Selection in Neural Networks used to Classify Remotely Sensed Data.

Linda Milne
Computer Science and Engineering
University of New South Wales
Sydney NSW 2052
linda@cse.unsw.edu.au
tel +61-2-9385-3979
fax +61-2-9385-1814

Abstract

As more remotely sensed data becomes available there is an increasing need for automated image processing techniques. In particular, there is a need for the selection of relevant attributes used in a given classification problem. Neural networks are widely used for classification of image data, but few practitioners achieve optimal results. In part, this is due to the use of noisy or irrelevant data. This paper compares a new attribute selection method specifically for use with neural networks, namely contribution analysis, with the more general wrapper method of attribute selection.

1 Introduction

Remotely sensed and supplementary data can be particularly noisy or contain irrelevant information. Induction algorithms, such as C4.5, generalise poorly if allowed to use all available attributes as compared with using a relevant subset of the attributes [CF92]. Neural networks have been perceived as a technique that can unproblematically deal with noisy or irrelevant data. This is true to a limited extent, however, neural networks too will perform better if given a set of relevant attributes.

One of the main problems with supervised classification in a remote sensing domain is the small amount of training data. Classification accuracy will start to decrease as the ratio of training cases to attributes decreases [JR96]. So, if we have a large number of irrelevant attributes our ability to produce a reliable classification is reduced. On the other hand, the more classes a classifier needs to identify the more attributes that are needed [JR96]. Thus, there is a trade off between the number of attributes and the number of training classes.

In general, an exhaustive search for the best subset of attributes is not possible and so a number of heuristic search techniques have been developed (see [Lan94]). Such techniques may not find the best possible subset of attributes for a given classification task but typically they find a good subset.

Other work has also shown that different classifiers perform better with different attribute subsets [Mil95, JKP94, JR96]. John et al [JKP94] argue that the attributes selected should depend not only on the features and the classification task but also on the classification scheme used. To this end, they proposed the wrapper method. The wrapper method works on the basic assumption that selecting appropriate attributes for a given classification task is done by the classifier to be used.

The wrapper method, however, is not ideal when applied to neural networks as the training times very quickly become intractable, for most practical purposes. Milne [Mil95] proposed a new method, contribution analysis, which uses the contribution of input attributes to the output to select the most relevant attributes for a given neural network classification.

2 Attribute Selection Using Contribution Analysis.

Milne [Mil95] outlined a technique for selecting relevant attributes for a neural network classification task using the weights from a trained neural network. The neural networks used were multi-layer perceptrons trained using back-propagation.

The networks used consist of 3 layers, an input layer (with *ninputs* inputs), a single hidden layer (with *nhidden* nodes) and an output layer (with *noutputs* nodes). 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}.

The contribution of an input i to the output o is given as

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

It was suggested that attributes that have a contribution *close* to zero or those that have a *large* variation in contribution are irrelevant and can be left out of the training. This was used to remove irrelevant attributes from a classification of remotely sensed data and resulted in an increase in classification accuracy. In addition to this training times were reduced due to a smaller number of attributes being used. However, [Mil95] does not show if this method chooses a good set of attributes.

The contributions of each attribute are calculated for five training runs of the neural network. These values are then plotted (as shown in Figure 1) and the irrelevant attributes are chosen manually. For example, in Figure 1 the contribution of the `holding` attribute is close to zero and the contribution of the `body_shape` attribute varies and so they may be considered irrelevant.

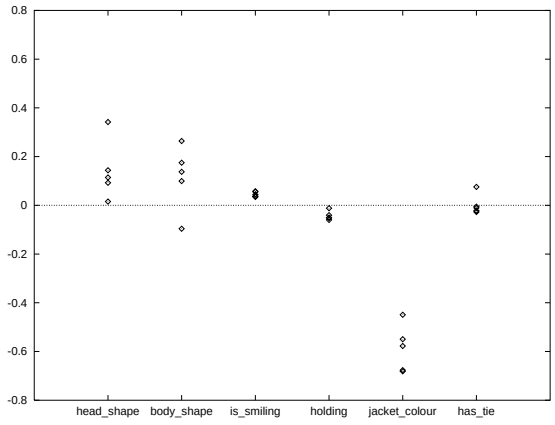

Figure 1: Contribution of each attribute to the output of a neural network (attribute vs contribution).

Once the set of irrelevant attributes is chosen they are removed from the training set and the neural network is retrained using only the relevant attributes.

3 The Wrapper Method of Attribute Selection.

The wrapper method [JKP94] was proposed as a way of determining a good set of attributes for a given classifier and classification task.

The algorithm starts with an empty attribute set. Each iteration adds each of the attributes is individually the set and the error rate is determined[1]. After all of the attributes have been considered the attribute that reduces the error rate by the largest amount is added to the set. Each iteration continues in this manner, considering single attributes that are not already in the set of good attributes and adding the one that reduces the error rate. Attributes are added until the error rate on the test set stops improving.

4 Comparison of Techniques.

The aim of this work is to determine if irrelevant attributes can be identified using their contribution to the output of a neural network. This is done firstly by comparing the error rates of classifiers trained on attribute subsets using the two attribute selection methods with classifiers trained on the entire dataset. The results of the neural network classification were also compared with the classification using C4.5, a decision tree induction algorithm [Qui93]. Next, the effects of noise on the contributions of attributes were investigated.

4.1 Description of the Datasets Used.

Four datasets were selected from the UCI Machine Learning Repository (see `http://www.ics.uci.edu/AI/ML/MLDBRepository.html`).

The iris dataset (*iris*) has been widely used for evaluating classification algorithms. Four attributes are used to described three types of iris – `setosa`, `virginica` and `versicolor`.

The mushroom (*mushroom*) dataset was drawn from *The Audubon Society Field Guide to North American Mushrooms*, G.H. Lincoff, 1981. It has 22 attributes and two possible classes – `poisonous` or `edible`.

The MONK's dataset was generated to compare the performance of different learning algorithms. The problem describes an artificial robot domain using 6 attributes (`head_shape`, `body_shape`, `is_smiling`,

[1]The error rate of the classifier for a given set of attributes is determined using n-fold cross validation [Koh95].

`holding`, `jacket_colour` and `has_tie`). The first problem (*monks1*) was to train a classifier to give a true or false value for `(head_shape = body_shape) or (jacket_colour = red)`. That is, only the attributes `head_shape`, `body_shape` and `jacket_colour` are required for the classification.

The solar flare dataset (*flare*) gives the number of solar flares in a 24hr period and the conditions under which they occurred. Ten attributes were used to describe three flare types – `M-class`, `C-class` and `X-class`.

Each classification task was formulated as a yes/no problem – an input is in a given class or not in that class. The training cases that are not in the given class are grouped into the single not in class. For each dataset the following classes were used.

dataset	class to be recognised
iris	`setosa`
mushroom	`edible`
monks1	`true`
flare	`M-class`

For each classification task 5 networks were trained to obtain an average error rate. Each network consisted of three layers with n inputs (the number of attributes for a given problem), $n/2$ hidden nodes and 1 output.

The cases for each dataset was split into three sets – a training set, a stopping set (used only for the neural network), and the test set. The stopping set is not used to train the neural network, but when the error on this set is at a minimum the training is stopped. The test set is not used at any time during training and is used to give an unbiased estimate of the error for the classifier. Attribute values were mapped to values between 0 and 1, and outputs to 0.1 (not in the given class) or 0.9 (in the given class).

4.2 Results.

For each of the datasets five training runs, as outlined below, were carried out.

`nn` A neural network trained using all possible attributes.

`c4` C4.5 trained using all possible attributes.

`ca` A neural network trained on a subset of the attributes chosen using contribution analysis.

`nnw` A neural network trained on a subset of the attributes chosen using the wrapper method.

`c4w` C4.5 trained on a subset of the attributes chosen using the wrapper method.

All three attribute selection methods chose a similar subset of attributes for a given dataset. In all but one case attribute selection did not decrease classification accuracy, as seen in the overall error rates in Table 1.

	attribute selection method				
dataset	nn	c4	ca	nnw	c4w
iris	0%	2%	0%	0%	2%
mushroom	0%	0%	3%	0%	0%
monks1	11%	0%	0%	0%	0%
flare	17%	19%	17%	17%	3%

Table 1: Overall classification error rates.

However the datasets used here are fairly well behaved datasets - attributes have been chosen by experts because they are useful and the cases given are generally low in noise. Thus, the effects of noise on the contributions were investigated in the following way.

Firstly, an additional noise attribute was added to each dataset. That is, an additional attribute was added to each dataset that had random numbers for values. In all but the *flare* dataset the contribution of the noise attribute was close to zero (see Figure 2) for the contributions of the *iris* dataset). Thus, if the contribution of an attribute is close to zero it can be considered irrelevant. However, not all irrelevant attributes will necessarily have a contribution that is close to zero. It is also interesting to note that the contributions of the original attributes did not change and so in these cases the neural networks are indeed able to distinguish between relevant and irrelevant attributes.

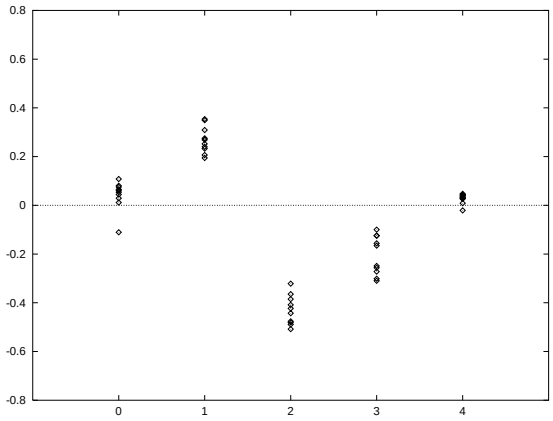

Figure 2: Contribution of each attribute to the output of a neural network trained with an additional noise attribute (attribute number 4).

Next, the effects of noise on a relevant attribute were investigated. Noise levels of 1%, 5%, 10%, 20%, 50% and 100% were added to one of the relevant attributes in each of the datasets and the contribution for that attribute only were plotted (see Figure 3 for the change in contribution for the `petal-length` attribute from the *iris* dataset). In all cases the absolute value of the contribution of the attribute with the added noise decreased towards zero as the noise increased.

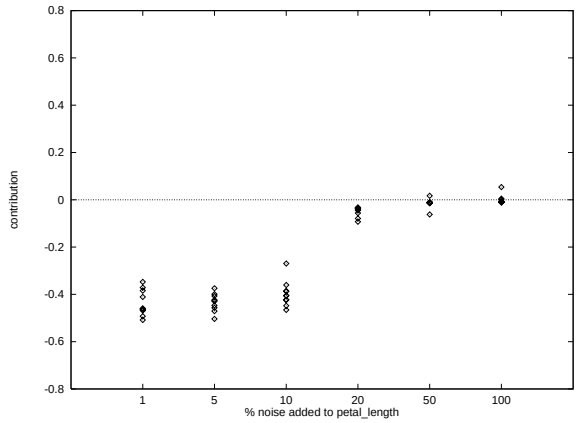

Figure 3: Contribution of a single attribute with increasing levels of noise to the output of a neural network (%noise vs contribution).

This leads us to conclude that the relevance of an attribute is mainly determined by the absolute value of the contribution. If the contribution of an attribute is *close* to zero the attribute is irrelevant for the classification. It does not seem to be important if there is variation in the contribution of an attribute if the absolute value of that contribution is *large*[2].

5 Contribution Analysis of Remotely Sensed Data.

Remotely sensed data is problematical in that attributes used for classification can be inaccurate or noisy. For example, soil and climate maps are generated from a small number of survey points and can not possibly include information about local variations. Although these attributes can be useful for human interpretation of remotely sensed images they may not have much of a role to play in automated classification. It is also possible to generate a large number of additional attributes using, for example, unsupervised classification and other remote sensing techniques[3]. Thus, we investigate the use of contribution analysis for attribute selection in a remote sensing domain.

A remotely sensed image was initially classified into four classes – `grass`, `trees`, `urban` and `water`. However, it was found that areas of `grass` were misclassified as `trees`. An additional 147 attributes were generated from the original 4 spectral bands using a variety of techniques [Mil97] in an attempt to distinguish between these two classes. A training set containing `grass` and `not-grass` cases was generated and the desired classification can be seen in Figure 4.

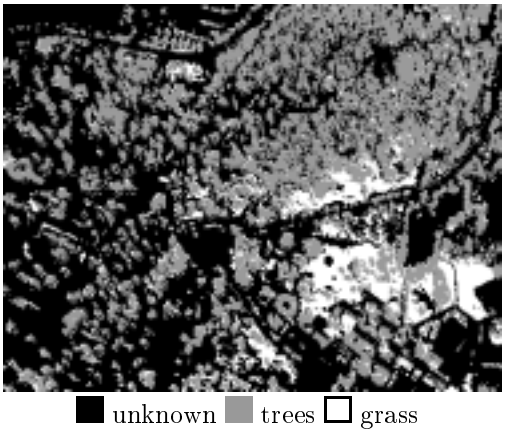

Figure 4: Classification showing areas of `grass` and `trees`.

Classification using all 151 attributes (*all*) resulted in no distinction being made between the two classes. The average error rate on the test set was 38%, but varied from 24% to 53%. This tends indicates that the neural network has not been able to distinguish between relevant and irrelevant attributes.

Attribute selection was then carried out on the 151 attributes using contribution analysis only. The magnitude of the contributions varied from around -0.04 to 0.03 (a sample of which can be seen in Figure 5).

Only three of the 151 attributes had a large absolute contribution and seven had a *zero* contribution. The remainder of the attributes could not be clearly distinguished as relevant or irrelevant. Thus, attributes were removed by thresholding the contributions – attributes whose contributions were between ± 0.01 (*thr0.01*), ± 0.02 (*thr0.02*) and ± 0.03 (*thr0.03*) were successively removed from the set of attributes used in the classification. The error rates for these classifications can be seen in Table 2.

[2] At this stage no attempt is made to quantify how close to zero or how large the variation may be.

[3] Such as principal components analysis and generation of vegetation indices (see [Ric94]).

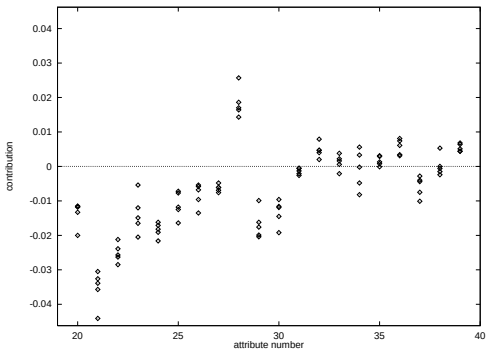

Figure 5: Contributions for 20 of the attributes used in the remotely sensed image classification.

dataset	no. attributes	av. error
all	151	38%
thr0.01	53	5%
thr0.02	12	7%
thr0.03	3	14%

Table 2: Overall classification error rates for the remotely sensed image.

The average error rates for the classifications after irrelevant attributes have been removed show significant improvement. Also, the error rates on individual neural networks only differed from the average by at most a few percent. However, in the case of the *thr3* dataset the error is due to all `grass` cases being misclassified as `not-grass`.

It is not possible to generate a classification of the entire image using all the attributes, however the classifications of the entire image after attribute selection can be seen in Figures 6- 8. The *thr0.02* classification is obviously the best with large areas of grass clearly distinguished (Figure 7). The *thr0.01* classification still seems to contain too many attributes to clearly distinguish the two classes, while the *thr0.03* classification has removed too many.

Further work will need to be done to determine exactly how to threshold the contributions to remove irrelevant attributes. It is however clear that contribution analysis can at least serve as a guide to which attributes may be irrelevant for a particular classification task.

6 Conclusions.

This paper presented the results of applying contribution analysis for relevant attribute selection to

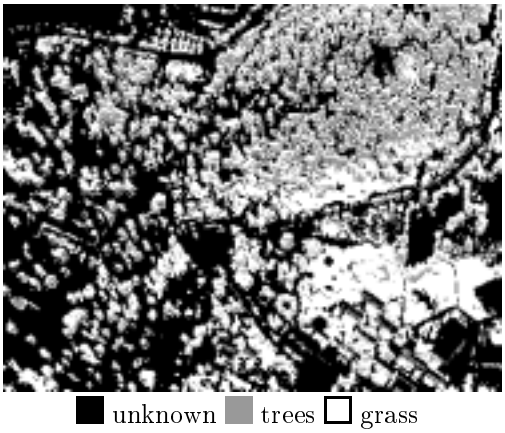

Figure 6: Classification of image using the *thr0.01* dataset.

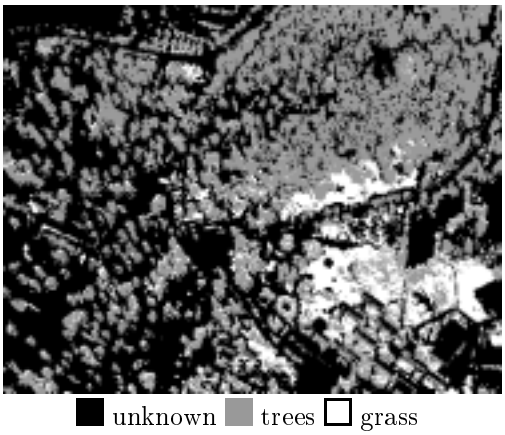

Figure 7: Classification of image using the *thr0.02* dataset.

five different domains. These results were compared with classification from the entire attribute set and with classification using the attributes selected from the wrapper method.

Both attribute selection methods provide similar accuracies on unseen data sets. This leads to improved classification of remotely sensed imagery, from which better maps can be drawn. However, the advantage of contribution analysis is that it only requires one additional training run of the neural network while the wrapper method needs at least n additional training runs, for n attributes.

Acknowledgements.

Many thanks to Spatial Analysis Unit at Charles Sturt University who have provided the data for this

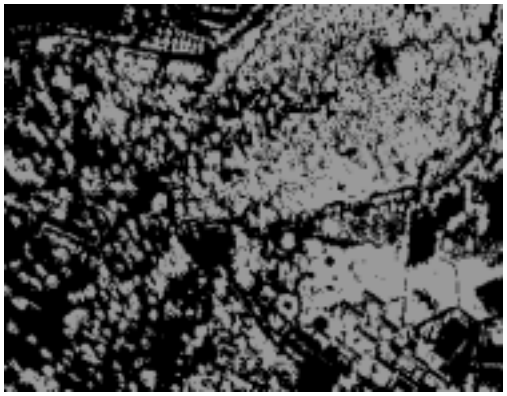

Figure 8: Classification of image using the *thr0.03* dataset.

work. Thanks also to Dr David Lamb (Charles Sturt University), Dr Andrew Taylor (UNSW), Dr Tom Gedeon (UNSW), Micheal Harries (UNSW), Ashesh Mahidadia (UNSW) and Dr Graham Mann (UNSW) for their valuable suggestions.

References

[CF92] R. Caruana and D. Freitag. Greedy attribute selection. In *Proc. Machine Leaning Conf.*, pages 249–256, 1992.

[JKP94] G.H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proc. Machine Leaning Conf.*, pages 121–129, 1994.

[JR96] X. Jia and J.A. Richards. Feature reduction using a supervised hierarchical classifier. In *Proc. Aust. Rem. Sens. Conf.*, pages 108–114, 1996.

[Koh95] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. Int. Joint Conf. AI*, pages 1137–1143, 1995.

[Lan94] P. Langley. Selection of relevant features in machine learning. In *Proc. AAAI Symposium on Relevance*, pages 379–382, 1994.

[Mil95] L.K. Milne. Feature selection using neural networks with contribution measures. In *AI'95*, 1995.

[Mil97] L.K. Milne. *Machine Learning for Classification of Remotely Sensed Data.* PhD thesis, University of New South Wales, School of Computer Science and Engineering, 1997. In preparation.

[Qui93] J.R. Quinlan. *C4.5 : Programs for Machine Learning.* Morgan-Kaufmann, 1993.

[Ric94] J.A. Richards. *Remote Sensing Digital Image Analysis : An Introduction.* Springer–Verlag, 1994.