SIMULATION OF CORRELATION ACTIVITY PRUNING METHODS TO ENHANCE TRANSPARENCY OF ANNS

CHRISTOPHER ROADKNIGHT[1], DOMINIC PALMER-BROWN, DAVID AL-DABASS[2]

[1] B54/124, BT Labs, Martelsham Heath, Suffolk IP5 3RE.
christopher.roadknight@bt.com

[2] Dept of Computing, Nottingham Trent University
Nottingham NG1 4BU
david.al-dabass@ntu.ac.uk

Abstract: The use of ANNs as predictors of natural phenomena is an important application but equally important is any resulting explanation of the heuristics a network uses to achieve this prediction. The novel methods of equation synthesis and **c**orrelated **a**ctivation **p**runing (CAPing) are introduced and used to extract meaning from a trained ANN. Equation synthesis involves the incremental increase in the number of connections of the trained ANN used until satisfactory prediction is achieved. CAPing involves the identification of nodes that have similar effects on the desired output. Comparison of the inputs to these nodes can lead to useful dependency relationships. Several useful generalizations have been made in this project by using these methods. Generalizations have been made using ANNs, equation synthesis and CAPing. These techniques are applied to neural nets trained to simulate the pollutant/crop damage cause/effect relationship.

1. INTRODUCTION

The research set out in this paper was carried out with the aim of making the adoption of Neural Networks for real world problem solving more likely. It attempts to guide the reader in methods of application and provide novel tools for successful adoption. The testing ground for this work is a biological problem, but the findings of the research are applicable to any real world problem where the number and complexity of causative agents make deciding their actions complex.

Neural Networks.

The essential idea of a feed-forward ANN is that each neuron outputs a smoothly rising function of the sum of its weighted inputs, eg. F(a*w1+b*w2+c*w3). The weighted sum in the brackets also equals the scalar product of the data and weight vectors, d.w, which in turn equals D*W*cos(angle between d and w), where d is (a,b,c) and w is (w1, w2, w3). This is worth knowing because it demonstrates that the neuron is effectively detecting the feature w.

When networks are built using three layers, the middle layer is called 'the hidden layer'. Each unit within the hidden layer may act as a feature detector, responding to features appearing within the input data. This neural network structure is usually called a multi-layer perceptron (MLP). The MLP architecture is the most popular in real world applications. Each layer is fully connected to the next. The results of many authors working with MLPs over many years, including the authors of this paper, tends towards the optimistic view that simple monotonic functions like the sigmoid or the tan(h) are widely applicable.

The connection weights of a neural network need to be discovered for a correct solution to any problem and this is called training. Where the interpretation of a set of (training) data is known, it is appropriate to use supervised learning; whereas if there are no available interpretations for the data, supervised learning cannot be used and unsupervised learning can be useful.

Supervised learning involves presenting the network with target answers as well as inputs so the network learns by example. One algorithm used to adjust network weights correctly is back-propagation (Rumelhart *et al* 1986). This involves presenting input data to an ANN and comparing the output from the network with the desired output and adjusting the weights to minimise the error. Types of back-propagation are used in many neural network systems. For example, recurrent neural networks have feedback connections but the input and output patterns change with time. This is back-propagation through time. Here the network is expanded over time.

Equation Synthesis

Once an ANN has been successfully applied the solution manifests itself as a set of activation function parameters and connection weights. In all but the simplest of networks these have no apparent meaning. These weights and functions can be used to create an equation but while this will have more meaning to a scientist, it will only be comprehensible for trivial networks (Roadknight et al 1996, 1997a)

To synthesise useful equations from non-trivial networks some rationalisation is required to keep the size of the equation manageable. This is achieved by removing connections of low weight and testing the remaining, partially connected network. The algorithm for equation synthesis is as follows:

```
1. For all hidden nodes
        Set required threshold for weights from hidden units to output unit
        Remove all hidden units below this threshold
2. For all input nodes
        Set required threshold for weights from input units to hidden units
        Remove all inputs to hidden units below this threshold
3. Create equation by matching input nodes, approximate signed weights to source of input node value.
4. Test synthesised equation using appropriate partially connected ANN
5. Repeat with lower thresholds until partially connected ANN is sufficiently accurate
```

Initially, thresholds are set sufficiently high as to only include the most important input to the most important hidden unit. This was invariably too limited to solve the problem so the thresholds are reduced until the required accuracy is achieved. The resulting equation is a simplification of the network which acted as a simple model that could be extended to include more detailed information. This algorithm produces equations which are a simplified representation of the ANN, preserving the principle characteristics of the learnt model.

The performance of partially connected networks was usually worse than the fully connected network, but the relationship between number of connections, described by terms in an equation, and performance was not linear.

CAPing

The generalisation ability of an ANN is dependent on its architecture. An ANN with the correct architecture will learn the predictive task presented by the training set but also gather enough general rules to correctly predict outputs for unseen test set examples. To obtain this optimum network architecture it is often necessary to apply a labourious ‘trial and error' approach. One approach to achieving optimum network architecture in a more intelligent way is pruning.

Weight pruning is the most researched area, this involves the removal of connections based on the value of the connecting weights, these can be divided into two groups.

1. Sensitivity Calculation.
Once the full sized network is trained the sensitivity of the error function to zeroing of a weight is estimated and weights with a low impact are removed (Karin 1990).

2. Penalty-Term Methods
This involves the introduction of a new cost function to enforce weights of small magnitude to converge to zero during training, they can then be removed with no effect (Weigend *et al* 1991).

A third method is proposed here called Correlated activation Pruning (CAPing). Sietsma and Dow (1988) describe an interactive pruning method that uses several heuristics to identify units that fail to contribute to the solution and therefore can be removed with no degradation in performance. This approach removes units with constant outputs over all the training patterns as these are not participating in the solution. Also, units with identical or opposite activations for all patterns can be combined. The approach in this paper is rudimentary and a more comprehensive and mathematically robust approach would be much more useful.

The full equations for CAPing are available elsewhere (Roadknight *et al.* 1997b), for the purposes of this paper a brief description will suffice. Activation strengths at each hidden node are monitored for each pattern, these series of activations are then checked for degree of correlation and a correlation coefficient found for each pair of hidden nodes. The pair of nodes with the correlation coefficient nearest unity (1 or -1) are replaced by one node and the weights and biases changed accordingly. The ANN is then checked to see that generalisation has been retained. This cycle is repeated until generalisation is lost, usually when the correlation coefficient dips below +/- 0.85. Nodes with excess hidden nodes can therefore be pruned to a near optimum structure, if equation synthesis is then applied a simpler, more parsable equation is formed.

CAPing offers many benefits to the neural network practitioner. Reducing the size of an ANN without significantly reducing its performance can only lead to more transparency. This is aided by the fact that CAPing is very objective whereby simple rules are followed that don't claim to reduce every network, only those where too many hidden nodes are used. This objectivity is crucial, as there are already enough moveable goalposts in a researcher's drive for optimum performance.

2. RESULTS

Several pollution modelling tasks were undertaken, these took the form of manipulating the input data into a reasonable form to train a neural network to predict the onset and/or the amount of injury. An example of which is given in Figure 1. Here the value in the output unit would be a binary output: 1 if injury occurs on day 4, 0 if no injury occurs on day4.

CAPing and equation synthesis was applied to several pollution simulation neural networks. One such network predicted the onset of leaf injury from levels of pollutants and environmental condition during the 5 day period before injury was detected. Training and testing of the input data yield satisfactory prediction of injury onset (Figure 2,3)

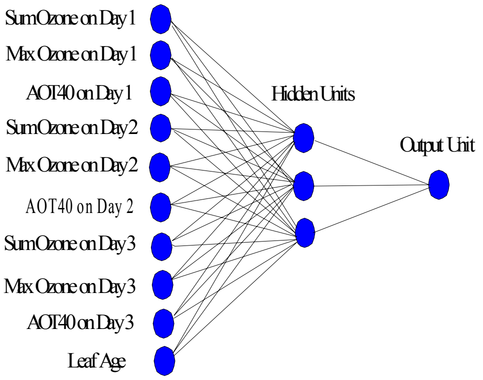

Figure 1. Typical injury simulating ANN

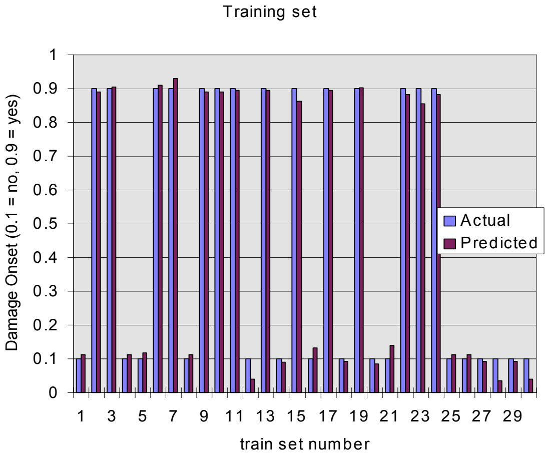

Figure 2. Training set for leaf damage simulation ANN

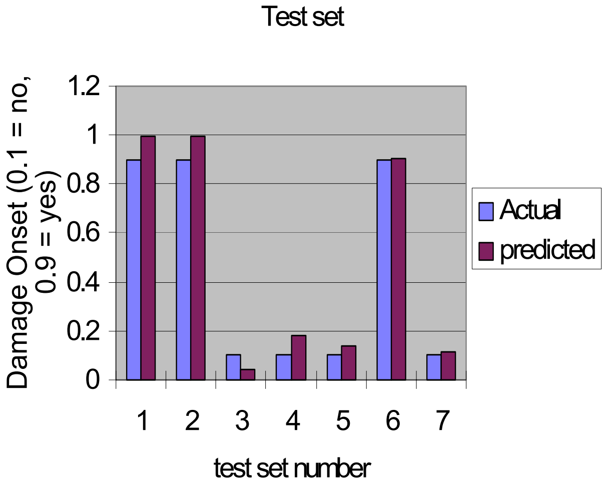

Figure 3. Test set for leaf damage simulation ANN

CAPing was applied to the trained network and it was pruned to network of only 2 hidden units, equation synthesis was then applied to the weights of this network and the following equation was produced:

Occurrence of injury =
F[Day2(7hmean + Max) - Day3(AOT40+Max0zone)] +
F[Day2(MaxOzone-AOT40) - Day1(AOT40 + 7hmean) + day3(7hmean)] - 1

The terms in this equation are pollutant factor levels eg. Mean of hourly readings of ozone levels between 10am and 5pm during day 2 is Day2(7hmean).

This equation contains the primary inputs for 2 hidden units, contained within an activation function (F), which can be approximated as a straight line between 0 and 1, with a value of 0.5 at F(0).

The first, and most influential, node shows that a fall from high to low levels of ozone precedes the onset of injury. This is apparent because ozone levels on day three are used in a negative way , ie higher levels of ozone give a lower output. Ideal conditions for a positive prediction of injury are therefore high levels of ozone on day 2 and low levels on day 3. This equation can be further simplified to:

Occurrence of injury = F[rise in ozone levels] + F[fall in ozone levels]

A visual examination of many graphs of ozone levels and injury agree with this conclusion (eg. Figure 4)

A 5 day profile of a damage causing ozone exposure was averaged and then entered into the trained network (Figure 5). The resulting output was less than 0.5, indicating that the ANN thought that no injury would occur after this theoretical exposure.

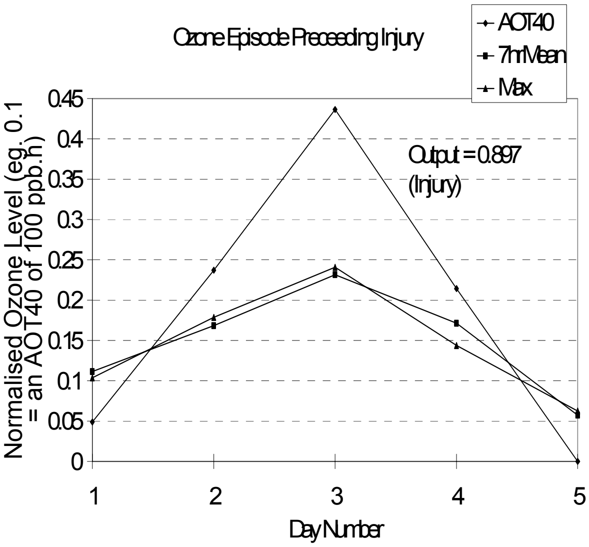

Figure 4. A typical pre-injury ozone episode.

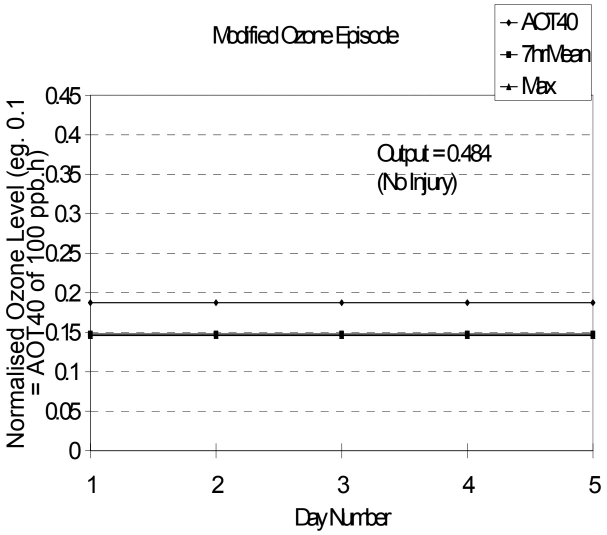

Figure 5. Equivalent quantity of ozone as Figure 4 giving no injury.

The behaviour of the hidden units has been an important focus. This is rightly so, in CAPing they provide the only good indicator of complex redundancy within a problem solving ANN. This chapter discusses the usefulness of analysing these activations for generating thresholds.

So far the problems of modelling ozone related leaf injury have largely been constrained to qualitative predictions. Now methods to give the researcher an idea on quantities of damage are examined, whether that be in the changes in yield, leaf area effected or giving quantitative rules of thumb for leaf injury expression.

States are analysed both visually then quantitatively, initial conditions can then be varied or new conditions added, the resulting effects on activations being the diagnostic tool. This is made possible by equation synthesis and CAPing enabling the accurate evaluation of the purpose of individual hidden nodes. For instance, 'sliding' variables and examining their effect gives an idea of a particular variable's impact, and the nature (non-linear, linear, thresholded) of its affect on the network's output. Some fixing of variables is sometimes required to enable one variable to be analysed at a time, but realistic values for secondary influencers are used.

The activations need to be put into relative and absolute context. The activations from one hidden node must be compared with other activations from the same node when presented with different inputs. Also, the importance of the node relative to the other nodes in the network must be gauged.

Each activation at each node within an ANN has a non-linear effect on the final output, this is due to the inherent non-linearity of the network. It is therefore difficult to take the activation of nodes and make extrapolations from these. The methods proposed in the earlier chapters of this thesis enable a minimised network to be created and for the salient rules by which that network performs its task to be alluded to. These methods are not specifically for the case study used. Taking a minimised network with a known set of equations generated by equation extraction, it was possible to look at the activations occurring at one node. This is possible because the most important node can be taken, or the node with strong equational bias towards factors of interest. In our example the most important node of the onset of leaf damage predicting ANN was taken. This approach could be applicable to any ANN ie. Identifying how the model is divided up over the nodes. All levels for the three ozone related variables were passed through the ANN, with their associated modifying factor levels. The results of this are shown in figure 6.

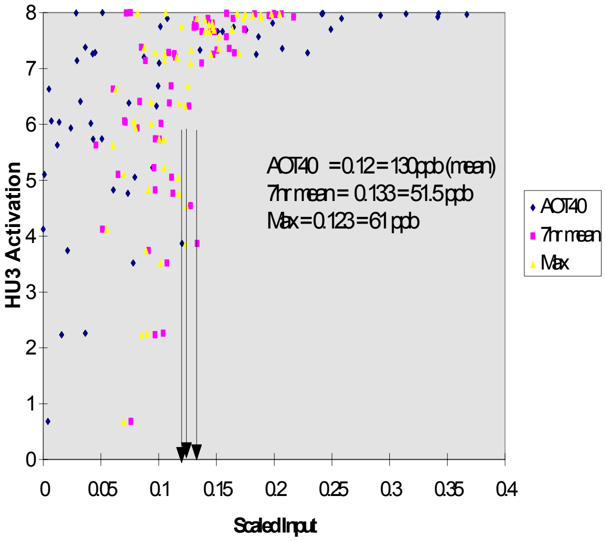

Figure 6. Thresholds for leaf damage due to ozone exposure.

It is encouraging to see a clear cut off point for all three variables at which it can be deemed that the highest level has been reached that did not cause a significant activation, these levels are shown within the figure. Notice how the activities vary between 0 and 8 but to the right of the line they are always above a value of 7.

This speculative approach can be applied to any ANN, and just involves looking for distinct boundaries within a set of activations and then making statements about the position of these boundaries. This is a useful method for finding general thresholds around which activation patterns change considerably.

3. DISCUSSION

These results show that once networks have been optimised and some meaning given to each hidden unit, the behaviour of an ANN can be explained. Anyone thinking of implementing an ANN for any problem solving task needs to be able to take sections of an ANN and explain what this part is doing, this transparency is one of the most important factors for ANN adoption. Making the ANN believable should be of as high a priority as making it predict well, as both are equally important if the ANN solution is to be ‘sold' to an implementor. Even for the purposes of self testing and evaluation, generating some meaningful rules of thumb, that make up part of the ANN's functional capabilities, can be very useful.

4. CONCLUSIONS

The use of ANNs as predictors of natural phenomena is an important application but equally important is any resulting explanation of the heuristics a network uses to achieve this prediction. The novel methods of equation synthesis and correlated activation pruning (CAPing) are introduced and used to extract meaning from a trained ANN. Equation synthesis involves the incremental increase in the number of connections of the trained ANN used until satisfactory prediction is achieved. CAPing involves the identification of nodes that have similar effects on the desired output. Comparison of the inputs to these nodes can lead to useful dependency relationships. Several useful generalizations have been made in this project by using these methods. Generalizations have been made using ANNs, equation synthesis and CAPing. For example, the temporal dynamics of an ozone exposure are evidenced as being more important than the quantity of the ozone exposure and light levels are shown to be the most important modifying factor in the injury process.

When the concentration of ground level ozone reaches significant levels, severe and economically important damage can occur to agricultural crop plants. Environmental modifying factors affect the expression of this injury. The use of artificial neural networks (ANNs) as tools for ozone damage prediction is investigated in this project and novel approaches to extracting meaning from these networks are examined.

ANNs have been further used to create a set of rules by which the onset of injury can be discriminated, the performance of these rules is only slightly inferior to the ANNs and their explicit nature makes them valuable.

The effect of Ozone, and its modifying factors, on yield is modeled using ANNs. The structures of these yield predicting models are explained. The importance of the nature of the ozone episode and modifying factors is analysed. By analysing the hidden unit weighting of CAPed neural networks it is possible to observe clear thresholds of effect, this is useful for giving clear indications of what the neural network has learnt.

The methods discussed in this paper could be applied to any suitably complex and multivariate data set. A degree of transparency, not generally expected from neural network approaches, would be apparent. Therefore the theories would be suitable for both neural network practitioners looking to add more transparency to their modelling and for the traditional data modeller looking for alternative techniques for their complex data set but fearful of ANNs traditional 'black box' downfall.

REFERENCES.

Karin ED. 1990. A simple procedure for pruning back-propagation trained neural networks. IEEE Trans. Neural Networks, vol.1 no.2. p239-242

Roadknight CM, Palmer-Brown D and Sanders GE. (1995). Learning the equations of data. Proceedings of 3rd annual SNN symposium on neural networks (eds. Kappen B and Gielen S) Springer-Verlag. 253-257.

Roadknight CM., Balls GR., Palmer-Brown D and Mills GE (1997a). Modelling of complex environmental data. IEEE Transactions on Neural Networks. Vol 8, No 4. P. 852-862

Roadknight CM., Palmer-Brown D and Mills GE (1997b). Correlated Activity Pruning. 5th Fuzzy Days, Dortmund April 28-30. 1997. Lecture notes in Computer Science. 591-592; Springer Verlag; ISBN 3-540-62868-1

Rumelhart DE and McClelland JL. 1986. Parallel Distributed Processing. The MIT Press, Cambridge, Mass. USA.

Sietsma J & Dow RJF. 1988. Neural net pruning - Why and how. Prc. IEEE Int. Conf. Neural Networks. Vol 1. p. 325-333.

Weigend AS, Rumelhart DE and Huberman BA. 1991. Generelization by weight elimination with applications to forecasting. In Advances in Neural Information Processing (3). Lippmann R, Moody J and Touretzky D. Eds. p. 875-882.

BIOGRAPHY.

Dr Chris Roadknight studied for his BSc and MSc at Manchester University in Applied Biology and Computer Sciences respectively. He then went on to study for a PhD entitled 'Transparent Neural Network Data Modelling' at Nottingham Trent University. Since 1997 he has been working at BT's research Labs in the programmable networks lab.

BIOGRAPHY: David Al-Dabass is Professor of Intelligent Systems, School of Computing & Mathematics, The Nottingham Trent University. He graduated from Imperial College in 1966 with BSc in Electrical Engineering, worked for Redifon Flight Simulation until 1972, completed a PhD in Parallel Processing at Staffordshire University in 1975 and held post-doctoral and advanced research fellowships (76-82) at the Control Systems Centre, UMIST. He joined The Nottingham Trent University in 1983. He is Fellow of the IEE, IMA and BCS and editor-in-chief of the International Journal of Simulation: Systems, Science and Technology; he currently serves as chairman of the UK Simulation Society and as Director of European Simulation Multi-conference (ESM) series for SCS Europe. For more details see his website: http://ducati.doc.ntu.ac.uk/uksim/dad/webpage.htm