Indicators of Input Contributions: Comparing Functional Measures

Tamás D. Gedeon
School of Computer Science and Engineering
The University of New South Wales
Sydney 2052 AUSTRALIA
tom@cse.unsw.edu.au
Fax: +61 2 385 5995

ABSTRACT: *The problem of data encoding and feature selection for training back-propagation neural networks is well known. The basic principles are to avoid encrypting the underlying structure of the data, and to avoid using irrelevant inputs. In the real world we often receive data which has been processed by at least one previous user. The data may contain too many instances of some class, too few instances of other classes, ands often include many irrelevant or redundant fields.*

Previous approaches have focussed on the analysis of the weight matrix of trained networks to determine the magnitude of contribution particular inputs make to the output to determine which are less significant.

This paper examines measures to determine the functional contribution of inputs to outputs. Inputs which include minor but unique information to the network are more significant than inputs with higher magnitude contribution but providing redundant information also provided by another input.

This paper presents a novel functional analysis of the weight matrix based on a technique developed for determining the behavioural significance of hidden neurons. This is compared with the application of the same technique to the training and test data available.

Finally, a novel aggregation technique is introduced.

1. INTRODUCTION

The initial network topology was 12-7-1, being twelve inputs, seven hidden neurons, and one output neurons. The data for this study was acquired from a novel eye gaze detector developed at Westmead Hospital. The major advantage of eye gaze data is that when we know where the eye is looking, we know the contents of the major input channel to the brain.

For example, the point of first fixation for schizophrenic versus normal controls on a neutral affect face produces results which statistically separate the two cases. This work extends this classification process to reliably classify the individual cases based on multiple responses to a wire frame drawing, a neutral affect face, a happy face and a sad face. This initial trial used 10 schizophrenic and 10 normal individuals, with 4 responses of 10 seconds duration recorded at 50 Hz.

The detector uses infra-red to detect the difference between the angle of reflection from the front of the eye and the retina to determine where on screen the subject is looking. The data used in this paper makes use only of the summary statistics of the entire data stream, with respect to fixations of gaze of 200 msec or longer.

The twelve inputs are: x and y co-ordinates; overall distance, horizontal, and vertical distance to previous fixation point; distance to previous fixation point relative to scan distance; pupil area; pupil area relative to pre and post-stimulus pupil areas; dwell time; and relative dwell time compared to the average dwell time; and finally, which image in being looked at.

The single output classifies by values above/below 0.5 whether the particular patterns belongs to a normal control or schizophrenic patient. Note the this problem is particularly hard, as the network needs to determine a classification based on the current eye gaze location and the difference from the previous one.

Previous work has investigated the use of related soft computing techniques, being vector quantisation and simulated annealing in the classification of schizophrenic versus medicated schizophrenic patients versus normal controls [1].

0-7803-3679-8/96/$5.00

The network was trained using error-backpropagation [2]. All connections are from units in one level to units in the next level, with no lateral, backward or multi-layer connections. Each unit is connected to each unit in the preceding layer by a simple weighted link. The network is trained using a training set of input patterns with desired outputs, using the back-propagation of error measures.

The network is tested using a validation set of patterns which are never seen by the network during training and thus can provide a good measure of the generalisation capabilities of the network. Thus all results quoted in this paper are for the test set.

2. MAGNITUDE MEASURES OF CONTRIBUTIONS

Garson [3] proposed the a measure for the proportional contribution of an input to a particular output based on the size of input to hidden weights relative to the sum of all inputs to that hidden, and weighted by the magnitude of the connection to the output neuron concerned. This value was then normalised to account for the overall magnitudes of such contributions from all inputs.

$$G_{ik} = \frac{\sum_{j=1}^{nh} \frac{w_{ij}}{\sum_{p=1}^{ni} w_{pj}} \cdot w_{jk}}{\sum_{q=1}^{ni} \left(\sum_{j=1}^{nh} \frac{w_{qj}}{\sum_{p=1}^{ni} w_{pj}} \cdot w_{qj} \right)}$$

A disadvantage of this approach is that during the summation process, positive and negative weights can cancel their contribution which leads to inconsistent results.

Wong, Gedeon and Taggart [4] used a measure for the contribution of an input to a hidden layer neuron which used the absolute values of the weights.

$$P_{ij} = \frac{|w_{ij}|}{\sum_{p=1}^{ni} |w_{pj}|}$$

Milne [5] commented that the sign of the contribution is lost, and modified Garson's measure by taking the absolute values of the two normalisation terms in Garson's formula.

$$M_{ik} = \frac{\sum_{j=1}^{nh} \frac{w_{ij}}{\sum_{p=1}^{ni} |w_{pj}|} \cdot w_{jk}}{\sum_{q=1}^{ni} \left| \left(\sum_{j=1}^{nh} \frac{w_{qj}}{\sum_{p=1}^{ni} |w_{pj}|} \cdot w_{qj} \right) \right|}$$

Gedeon [6] introduced an extension of the absolute value technique [4], by defining a measure P_{jk} for the contribution of a hidden neuron to an output neuron similar to the measure P_{ij} used previously, and combine the two measures by aggregating the contributions using all possible connections between the desired input and the output.

$$P_{jk} = \frac{|w_{jk}|}{\sum_{r=1}^{nh} |w_{rk}|} \qquad Q_{ik} = \sum_{r=1}^{nh} (P_{ir} \times P_{rk})$$

The benefit of this approach is that the magnitude of the contribution is disentangled from the sign of the contribution. The magnitude of contributions is significant in indicating whether an input is important, while the sign of contribution is largely irrelevant in the decision to remove or retain an input, and is recoverable in any case from the raw data by simple statistical methods.

Each of the above techniques could be extended to networks with larger numbers of hidden layers than the topology used in this experiment.

3. FUNCTIONAL MEASURES

The technique of distinctiveness analysis [7] uses hidden neuron activations over a training set to determine similarity using the angle between the multi-dimensional vectors thus formed.

The technique has been extended for examining the functionality of hidden neurons using the weight matrix [8], and is adapted here to determine the functional differences between inputs as represented by the pattern of input to hidden weights.

For comparison purposes, the pattern of values of inputs in the labelled set available (being both

training and test sets) is analysed in an analogous fashion, That is, the values for a particular input in all of the instances in the labelled set are used to construct a multi-dimensional vector.

This 1,334 dimensional vector is then compared to the vectors for the other 11 inputs. Note that this is essentially a first-order correlation measure, and does not incorporate the possible higher order features that a neural network could learn and incorporate into its weight matrix. Hence we would expect this measure to be less reliable than the distinctiveness approach applied to the weight matrix.

The distinctiveness analysis technique was initially developed for pruning hidden neurons, and provides ranking in which pairs of similar neurons are listed together. During pruning, most often only single neurons are removed at a time, in the process of fine-tuning the generalisation of a trained network.

For eliminating inputs, however, we would wish to remove larger numbers of inputs at one time, as the elimination of a single input produces relatively little savings on the time taken to train a network. As a compromise, in this paper two inputs are removed together. Since we wish to remove more than one input, some aggregate measure is required, which is provided here by the average angle to all other

In the following section the results for both of the above forms of functional analysis, together with the aggregated rankings are provided, listing the order of significance of inputs.

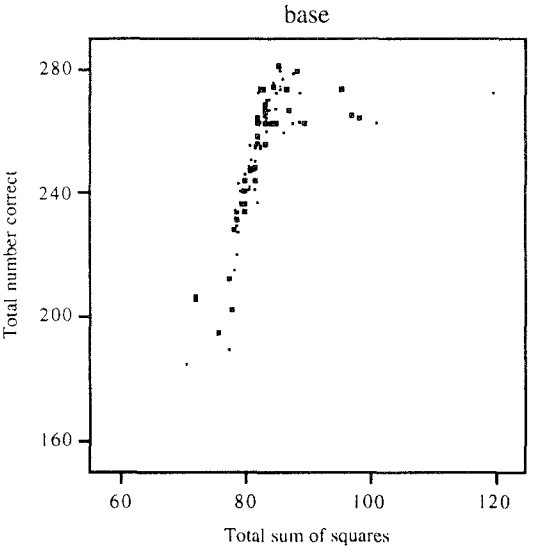
To maintain comparability with the previously used magnitude measures, the three measures discussed earlier are used to provide a joint ranked list based on the individual lists, which can be assumed to be representative of such magnitude ranking techniques and less affected by the computational peculiarities of the specific measures. Thus, the input which is highest in all three lists is clearly the most important by magnitude techniques and so on.

4. RESULTS AND COMPARISONS

In the following table, model *I* is the distinctiveness of inputs over the labelled set of patterns, of which *C* is the aggregated form, *W* is the weight distinctiveness, of which *U* is the aggregated form. The combined ranking for the three magnitude measures is given in the last column.

model	I	C	W	U	Mag.
Least signif.	12	4	1	11	9
	7	5	12	12	12
	9	12	7	8	7
	2	6	10	9	15
	5	9	5	10	21
	6	2	2	1	2
	1	1	4	5	1
	4	7	11	7	10
	8	11	3	2	6
	3	8	6	3	11
	10	10	8	4	5
Most signif.	11	3	9	6	4

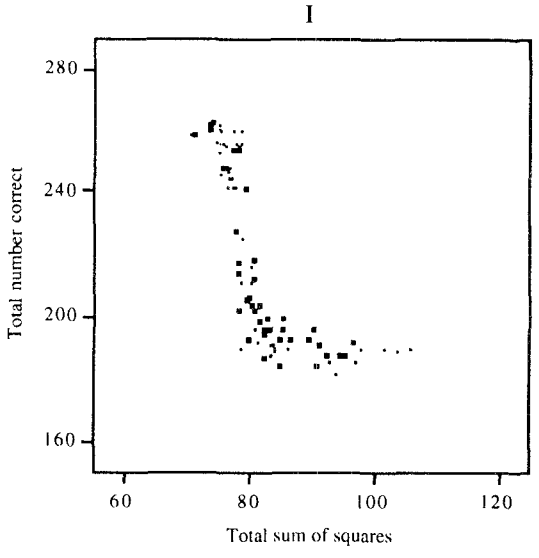
The following diagram show the base case, using the full complement of inputs as analysed for the above table.



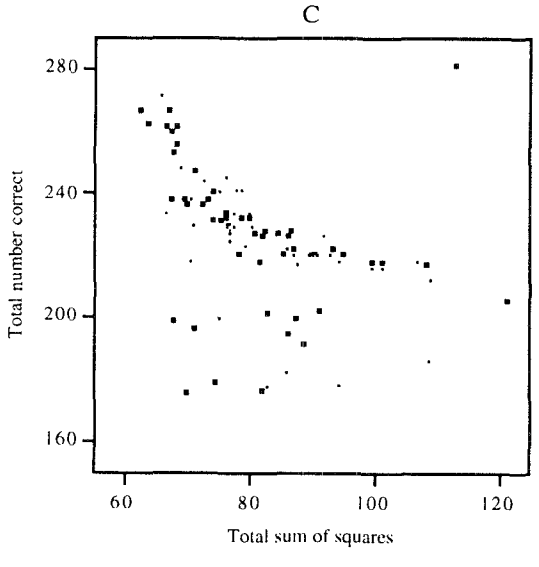
The anti-correlation of the total sum of squares (tss) value and the number of patterns correctly classified demonstrates the degree of difficulty of the classification problem for the network. There are a number of inputs which are providing irrelevant information, and the network was trained using sum squared error measure, when the network provides a better result in terms of low tss, the number of

correctly classified patterns is reduced.

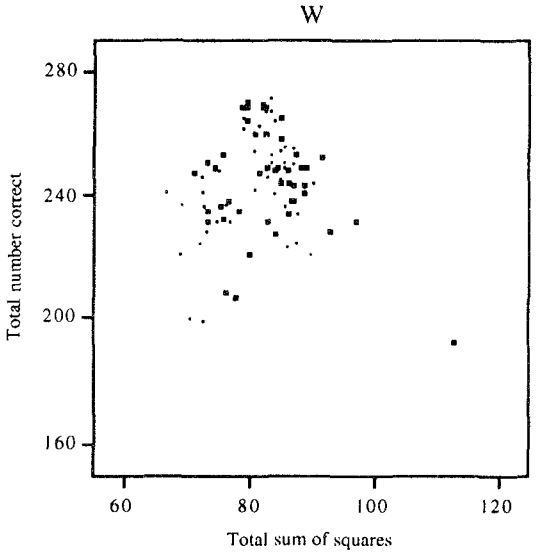
The following diagrams show the results on using the above table to eliminate some pairs of inputs. These are the top two inputs from the table, being the least significant inputs.



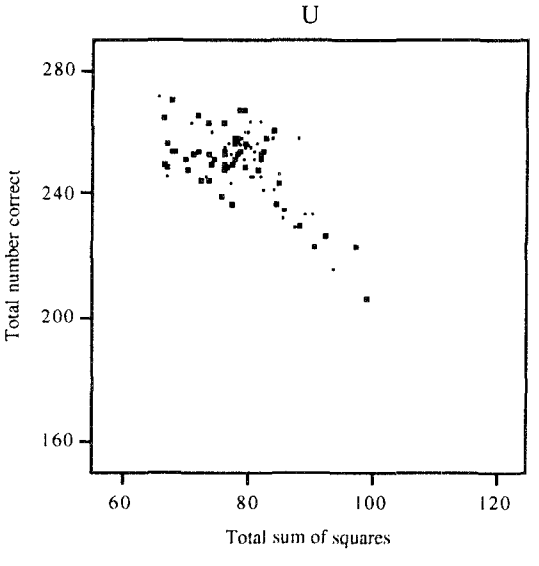
There is now a better correlation between tss and total correct, however overall the number correct has decreased indicating some significant information has also been lost.



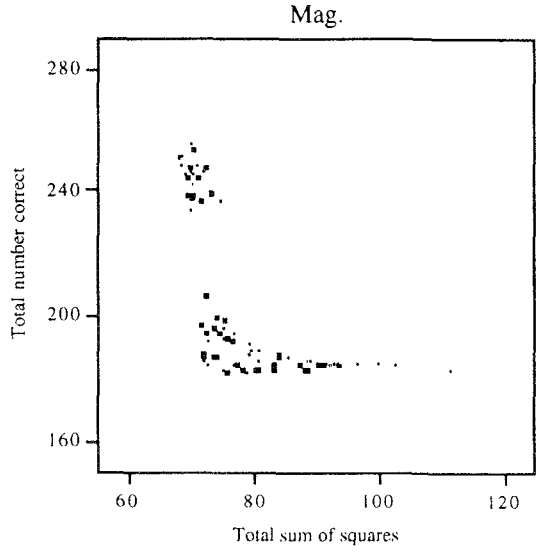
The correlation is much better, and less significant information has been lost. This demonstrates that the aggregated measure is a better predictor of significance, at least for this functional measure. The *I* and *C* measures depend only on the labelled pattern set, and input elimination using these measures has reduced the number of correct classifications overall.



This diagram has shown that the anti-correlation has been removed, and a slight correlation introduced, similar to *I*. Note that the overall number correct is significantly improved in *W* over *I*, which indicates the former is a better indicator of significance.

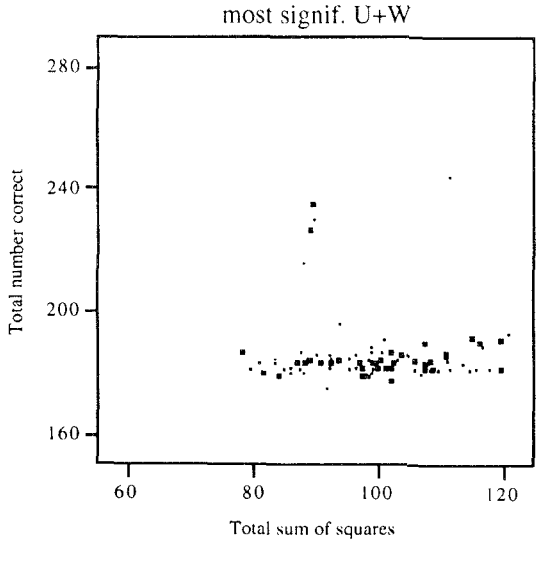


The above diagram demonstrates the correlation hoped for, with higher total correct classifications than the original base case. This again demonstrates the advantage of the aggregated measure. The diagram also shows that the network learnt weights provide a better indication of the significance of inputs than the simple statistical properties of the labelled set.



The correlation is slightly improved, but there is now a discontinuity in the quality of results produced, and the number correct are overall significantly lower. This indicates that at least one of the inputs removed was significant, notwithstanding the measure. This significant input is probably input 9, as the other input (12) is in the least significant pair in three of the functional measures used above. To test this, a further pair of inputs were eliminated, using the most significant inputs as determined by the U, and W measures.

The observed discontinuity possibly indicates that there is now a 'local minimum' which is easier to find than the best minimum that the network is otherwise able to find given its inputs.



The above diagram demonstrates that the effect of removing the two most significant inputs has a catastrophic effect on network performance. The lowest tss value is now higher, and the total number correct is no longer related to the tss value. This suggests that the number correct is now due to chance.

5. CONCLUSION

In this paper a number of functional measures for determining the significance of inputs were introduced. They were contrasted with with the traditional magnitude based input significance measures in both qualitatively by means of dicussion, and quantitatively, by experimentation.

The experimental work demonstrated that the functional measures, particularly based on the analysis of the network and not just the data, produce better indicators of the significance of particular inputs, as shown exhaustively by the elimination of pairs of inputs judged to be least significant by each measure. The effect of eliminating the two most significant inputs was to destroy network performance which serves as extra validation of the utility of functional measures introduced.

6. REFERENCES

[1] Haig, AR, Gordon, E, Rogers, G and Anderson, J "Classification of single-trial ERP sub-types: application of globally optimal vector quantization using simulated annealing," *Evoked Potentials*, 31 pages, 1995.

[2] Rumelhart, DE, Hinton, GE, Williams, RJ, "Learning internal representations by error propagation," in Rumelhart, DE, McClelland, *Parallel distributed processing*, vol. 1, MIT Press, 1986.

[3] Garson, GD "Interpreting Neural Network Connection Weights," *AI Expert*, pp. 47-51, April, 1991.

[4] Wong, PM, Gedeon, TD and Taggart, IJ "An Improved Technique in Porosity Prediction: A

Neural Network Approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, n. 4, pp. 971-980, 1995.

[5] Milne, LK "Feature Selection Using Neural Networks with Contribution Measures," *Proceedings Australian Conference on Artificial Intelligence AI'95*, Canberra, 1995.

[6] Gedeon, TD "Indicators of Input Contributions: Analysing the Weight matrix," *Proceedings ANZIIS'96 International Conference*, 4 pages, Adelaide, 1996.

[7] Gedeon, TD and Harris, D "Network Reduction Techniques," *Proceedings International Conference on Neural Networks Methodologies and Applications*, AMSE, vol. 1, pp. 119-126, San Diego, 1991.

[8] Gedeon, TD "Indicators of Hidden Neuron Functionality: Static versus Dynamic Assessment," invited paper, *Australasian Journal of Intelligent Information Systems*, vol., 10 pages, June, 1996.