ON THE ANALYSIS OF PYROLYSIS MASS SPECTRA USING ARTIFICIAL NEURAL NETWORKS. INDIVIDUAL INPUT SCALING LEADS TO RAPID LEARNING.

Mark J. Neal, Royston Goodacre & Douglas B. Kell

Institute of Biological Sciences, University of Wales, Aberystwyth, Dyfed SY23 3DA, U.K.
MJN@ABER.AC.UK RRG@ABER.AC.UK DBK@ABER.AC.UK Fax: +44 970 622354

Abstract

Pyrolysis mass spectra were obtained from various mixtures containing known amounts of glycogen and casamino acids. Feedforward neural networks were trained using the standard backpropagation algorithm to predict the percentage of casamino acids in unseen mixtures from their pyrolysis mass spectra. By scaling the input nodes *individually*, the variation between the spectra could be maximised and the convergence rate (as judged by the RMS error on test sets) increased by more than 100-fold compared with training runs in which the scaling was over the whole dataset.

INTRODUCTION

There is a continuing need for more rapid, precise and accurate analyses of the chemical composition of fermentor broths and the organisms which they contain. An ideal method would permit the simultaneous estimation of multiple determinands, would have negligible reagent costs, and would run under the control of a PC, to allow flexible operation of the sample handling, instrument calibration, and data analysis and visualisation routines. Our present work is directed towards the development of exactly such an instrument.

Pyrolysis is the thermal degradation of a material in an inert atmosphere, and leads to the production of volatile fragments from non-volatile material such as microorganisms or other biological samples (Irwin 1982). Curie-point pyrolysis is a particularly reproducible and straightforward version of the technique, in which the sample, dried onto an appropriate ferromagnetic metal or alloy, is rapidly heated (0.5s is typical) to the Curie point of the metal, which may itself be chosen (358, 480, 510, 530, 610 and 770°C are common temperatures). The volatile fragments (pyrolysate) resulting from the Curie-point pyrolysis may then be separated and analysed in a mass spectrometer (Meuzelaar *et al* 1982), and the combined technique is then known as pyrolysis mass spectrometry or PyMS.

Almost all biological materials will produce pyrolytic degradation products such as methane, ammonia, water, methanol and H_2S, whose mass:charge (m/z) ratio < 50, and fragments with m/z > 200 are rarely analytically important in microbiology (Berkeley *et al* 1990) unless very special conditions are employed (Smith & Snyder 1992). The analytically useful data are thus constituted by a set of 150 intensities (normalised to the total ion count) versus m/z in the range 51-200.

Within microbiology and biotechnology, PyMS has been used as a taxonomic aid in the *identification* and *discrimination* of different microorganisms (Gutteridge 1987). To this end, the reduction of the multivariate data generated by the PyMS system (and indeed of those generated by other arrays of sensors; Gardner & Bartlett 1991) is normally carried out using principal components analysis (PCA), whihc is a well-known technique for reducing the dimensionality of multivariate data whilst preserving most of the variance. Whilst PCA does not take account of any groupings in the data, neither does it require that the populations be normally distributed, i.e. it is a non-parametric method. (In addition, it permits the loadings of each of the m/z ratios on the principal components to be determined, and thus the extraction of at least *some* chemically significant information.) The closely-related canonical variates analysis technique then separates the samples into groups on the basis of the principal components and some *a priori* knowledge of the appropriate number of groupings (MacFie *et al* 1978). Provided that the data set contains "standards" (i.e. type or centro-strains) it is evident that one can establish the closeness of any unknown samples to a known organism, and thus effect the identification of the former. An excellent example of the discriminatory power of the approach is the demonstration (Goodacre & Berkeley 1990) that one can even use it to distinguish 4 strains of *E. coli* which differ only in the presence or absence of single antibiotic-resistance plasmids.

More recently, we (Goodacre *et al*. 1992, 1993b, 1994a) and others (Chun *et al*. 1993) have exploited artificial neural networks (ANNs) in supervised learning mode for the very successful identification of a variety of biological samples from their pyrolysis mass spectra, training fully interconnected multilayer perceptrons (MLPs) with one hidden layer on known standards using binary-encoded outputs and the standard backpropagation algorithm, and testing on spectra from unseen samples. We have also exploited Kohonen's self-organising feature map (Kohonen 1989) succesfully to carry out unsupervised learning, and hence the classification of microorganisms, from their pyrolysis mass spectra (Goodacre *et al*. 1994a).

Of perhaps more general chemical interest is the ability to use PyMS and ANNs for the *quantification* of substances in complex biological samples. The strategy is to obtain pyrolysis mass spectra from appropriate samples of interest and train ANNs to recognise the relative concentration of a chemical substance (as measured by wet chemistry) from the PyMS. We again demonstrated for the first time that ANNs could indeed be trained to give accurate values for the concentration of indole in *Escherichia coli* cultures (Goodacre & Kell 1993), and for the concentrations of individual compunds in a variety of binary, ternary and more complex mixtures (e.g. Goodacre *et al*. 1993a, 1994b).

Given that any non-volatile biological material can be pyrolysed, and that it has been established that MLPs with sigmoidal activation functions and at least one hidden layer of arbitrary size can effect any nonlinear mapping of a continuous function to an arbitrary degree of accuracy (e.g. Hornik *et al*. 1989), our interest is focussed on improving both the learning speed and the ability to generalise of ANNs trained on pyrolysis mass spectral data. In the case of PyMS data, each input is of a similar *character* (in that they are all chemical fragments), but some inputs may contain more noise than others (in that lower ion counts will have a greater percentage of electronic noise); in the worst case the lowest inputs may simply be noise, whose presence would both harm learning and without a rather robust cross-validation method would likely lead to overtraining. Since the data are normalised to the total ion count, any increase in a given mass is necessarily accompanied by a concomitant decrease in all of the others. However, it is known from the statistical literature (as the 'parsimony principle') that much better predictions can often be obtained when only the most relevant input variables are considered (e.g. Rawlings 1988, Miller 1990, Seasholtz & Kowalski 1993), it was therefore of interest to analyse the effects of varying the methods of scaling the *input* variables on the performance of our ANNs.

There are of course within the connectionist literature a multitude of articles which describe optimal growth or pruning of feedforward networks, designed to effect a sparse representation of the input-to-output mapping and thence improve generalisation (see e.g. LeCun *et al.* 1989, Mozer & Smolensky 1989, Fahlman & Lebiere 1990, Weigend *et al.* 1991, Finoff *et al.* 1993, Hassibi & Stork 1993). However, most of these growth/ skeletonisation algorithms have been devised to work on the creation or destruction of *individual weights*, particularly those to and from the hidden layer(s), and at all events make no attempt to distinguish the physically meaningful inputs from the latent variables represented by the nodes in the hidden layers (cf. Moody 1992). Since obtaining extra variables not only tends to cause overfitting but also normally costs more, it is *more generally* desirable to minimise the number of inputs used in the formation of the connectionist representation. The present study therefore addresses, and serves to illustrate, the substantial importance of optimal scaling of the inputs for the speed of learning and, to some extent, the ability to generalise.

EXPERIMENTAL SYSTEM

The experimental system studied consisted of mixtures of casamino acids and glycogen, as a model for the complex proteins and carbohydrates to be found in typical biological samples. Mixtures containing different percentages of each component were made up gravimetrically, and pyrolysed at 530°C as described (Goodacre *et al* 1993a). Typical pyrolysis mass spectra are shown in Fig 1, where it can be seen that they are not easy to distinguish by eye, and one may construe that such data constitute ideal material for analysis *via* computer/AI/neural methods.

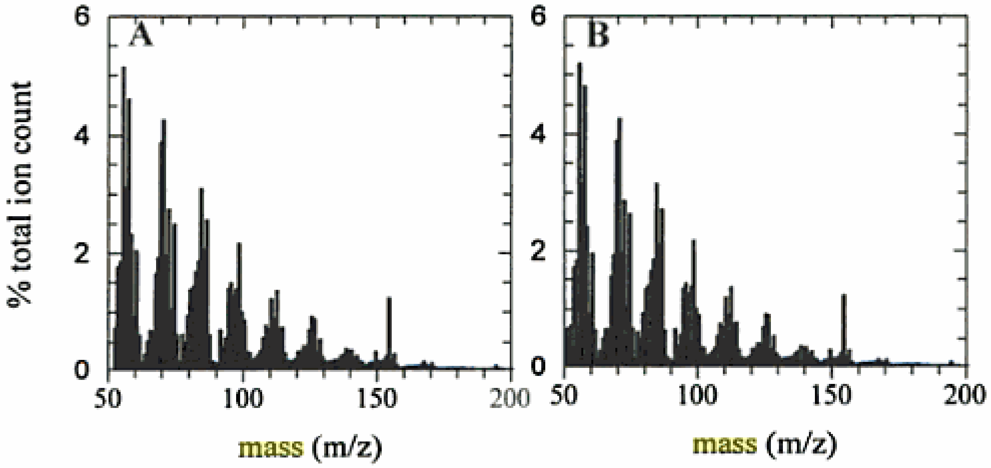

Fig 1. Normalised pyrolysis mass spectra of (A) 20 µg glycogen *plus* 100 µg casmino acids, and (B) 20 µg glycogen *plus* 90 µg casmino acids

The training set consisted of normalised spectra from mixtures containing 20 µg glycogen *plus* 10, 20, 30...100 µg casmino acids whilst the test set were spectra from mixtures containing 20 µg glycogen *plus* 5, 10, 15, 20, 25...100 µg casmino acids. To avoid the well-known problem of the sensitivity of backpropagation to initial conditions (Kolen & Pollack 1990), each run was done in sextuplicate and the data median-averaged. All neural networks were of the fully interconnected feedforward MLP type with a 150-8-1 architecture, and trained using the standard backpropagation algorithm with a logistic activation function, a learning rate of 0.1 and a momentum term of 0.9. Inputs were scaled as described in the text, whilst outputs were scaled between 0.1 and 0.9.

RESULTS AND DISCUSSION

Individual scaling of inputs

Whilst the usual backpropagation methods scale all inputs and outputs to lie between 0 and 1, this leaves open the question of how the scaling is done throughout the (columns of the) *population* of examples of interest. In particular, in the present case, some input columns contribute far more numerically to the inputs to the hidden layer than others (Fig 1). It is common in some other supervised multivariate calibration methods such as partial least squares to normalise the inputs in proportion to the reciprocal of their standard deviations (see e.g. Martens & Næs 1989). We therefore studied the effect of scaling the inputs on the basis of the highest ion counts throughout the entire dataset *versus* scaling the inputs of each m/z independently over the dataset. In the latter case, this means that the range of each input in the population is made equal.

TRAINING SET % RMS ERROR	EPOCHS UNTIL CONVERGENCE TO STATED % RMS ERROR		TEST SET % RMS ERROR	
	Scaled individually	Scaled on whole dataset	Scaled individually	Scaled on whole dataset
2	90	805	2.95	2.61
1	335	9770	2.44	2.02
0.50	725	84060	2.00	2.57
0.25	1470	217640	2.12	2.65
0.125	2240	>500000	2.08	-

Table 1. Effect of scaling inputs individually on the speed of convergence of backpropagation learning on an MLP.

It is evident from the data in Table 1 that *individual* scaling of the input nodes can effect a dramatic speed-up, of more than 100-fold, in the convergence of a neural network learning algorithm. This indicates that when all the scaled inputs to the net are of approximately the same magnitude the error value from a single input is less likely to dominate the error value at a given node, and therefore is less likely to swamp smaller error values associated with other connections to that node. This allows the reduction of error values in many dimensions in the input space to occur simultaneously. Put another way, by scaling the inputs *individually* in this way we are maximising the variance in the training set data, which therefore makes the discriminating features in the data easier (quicker) to learn.

From the data in Table 1 it is clear that although the convergence of the learning algorithm on the training set data is much quicker, there is a slight reduction in the accuracy of the predictions on the unseen data. This can however be improved by allowing the network to train to a slightly lower RMS error on the training set. The trade-off is such that individual scaling is still markedly superior when the criterion of training is the RMS error on the *test* data.

Pruning input variables

Given the dramatic speed-up that could be obtained by scaling the inputs individually, it was also of interest to see whether generalisation could be affected by removing masses whose numerical contribution to the total ion count over the population of samples was the lowest. The results of such a study are shown in Table 2, where it may be seen that removal of the numerically least significant masses had little effect on generalisation and a slightly unfavourable effect on the number of epochs needed for convergence to a given RMS error on the test set. This is consistent with the conclusion above that maximising the overall variance in the dataset leads to faster learning.

NUMBER OF EPOCHS UNTIL STATED % RMS ERROR OF TRAINING SET				
Training set % RMS error	Zero inputs removed	Remove m/z if <0.025%	Remove m/z if <0.05%	Remove m/z if <0.1%
2	90	125	90	95
1	335	280	345	460
0.5	725	850	950	1060
0.25	1470	1785	2105	2670
0.125	2240	2605	3250	4760
% RMS ERROR ON TEST SET				
2	2.95	3.07	3.17	3.02
1	2.44	2.33	2.37	2.18
0.5	2	2.08	2.27	2.44
0.25	2.12	2.21	2.19	2.75
0.125	2.08	2.26	2.43	3.53

Table 2. Effect of removal of masses with the lowest contribution to the total ion count over the population on the speed of learning and generalisation. Input nodes were scaled individually.

CONCLUSION

Individual scaling of the inputs of an artificial neural network maximises the variance in a given dataset and can effect a dramatic speed-up in the rate of convergence to a given RMS error on both training and test data. In the examples displayed, this speed-up could be more than 100-fold.

Acknowledgments

We thank the Biotechnology Directorate of the UK SERC for support of this work, under the terms of the LINK scheme in Biochemical Engineering, in collaboration with Horizon Instruments, Neural Computer Sciences and Zeneca plc.

References

Berkeley, RCW, Goodacre, R, Helyer, R & Kelley, T (1990) *Lab. Pract.* **39** (10) 81-83.

Chun, J, Atalan, E, Ward, AC & Goodfellow, M (1993) *FEMS Microbiol. Lett.* **107**, 321-326.

Fahlman, SE & Lebiere, C (1990) *The cascade-correlation learning architecture*. Report CMU-CS-90-100, Carnegie-Mellon University.

Finoff, W., Hergert, F. & Zimmermann, H.G. (1993) *Neural Networks* **6**, 771-783.

Gardner, JW & Bartlett, PN (1991) in *Techniques & Mechanisms in Gas Sensing*, ed. Mosley, PT, Norris, JOW & Williams, DE, pp. 347-380. Adam Hilger, Bristol.

Goodacre, R & Berkeley, RCW (1990) *FEMS Microbiol. Lett.* **71**, 133-138.

Goodacre, R & Kell, DB (1993) *Anal. Chim. Acta* **279**, 17-26.

Goodacre, R, Kell, DB & Bianchi, G (1992) *Nature* **359**, 594.

Goodacre, R Edmonds, AN & Kell, DB (1993a) *J. Anal. Appl. Pyrol.* **26**, 93-114.

Goodacre, R, Kell, DB & Bianchi, G (1993b) *J. Sci. Food Agric.* **63**, 297-307.

Goodacre, R, Neal, MJ, Kell, DB, Greenham, LW, Noble, WC & Harvey, RG (1994a) *J. Appl. Bacteriol.*, **76**, 124-134.

Goodacre, R, Neal, MJ & Kell, DB (1994b) *Anal. Chem.*, in the press.

Gutteridge CS (1987) *Meth. Microbiol.* **19**, 227-272.

Hassibi, B & Stork, DG (1993) in Hanson, SJ, Cowan, JD & Giles, CL, eds., *Advances in Neural Information Processing Systems 5*, 164-171, Morgan Kaufmann, San Mateo, CA.

Hornik, K, Stinchcombe, M.& White, H (1989) *Neural networks* **2**, 359-366.

Irwin, WJ (1982) *Analytical Pyrolysis: A Comprehensive Guide.* Marcel Dekker, New York.

Kohonen, T. (1989) *Self-Organization and Associative Memory.* Springer-Verlag, Berlin.

Kolen, JF & Pollack, JB (1990) *Complex Systems* **4**, 269-280.

Le Cun, Y., Denker, J.S. & Solla, S.A. (1989) in Touretzky, DS (ed) *Advances in Neural Information Processing Systems* Vol 2, 598-605. Morgan Kaufmann, New York.

MacFie, HJH, Gutteridge, CS & Norris, JR (1978) *J. Gen. Microbiol.* **104**, 67-74.

Martens, H. & Næs, T. (1989) *Multivariate Calibration*. Wiley, New York.

Meuzelaar, HLC, Haverkamp, J and Hileman, FD (1982) *Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials.* Elsevier, Amsterdam.

Miller, AJ (1990) *Subset selection in regression.*Chapman & Hall, London.

Moody, J. (1992) in Lippmann, RP (ed) *Advances in Neural Information Processing Systems* 4, 847-854. Morgan Kaufmann, San Mateo, CA.

Mozer, MC & Smolensky, P (1989) in Touretzky, DS (ed) *Advances in Neural Information Processing Systems* Vol 1, 107-115. Morgan Kaufmann, New York.

Reed, R (1993) *IEEE Trans. Neural Networks*, **4**, 740-747.

Rawlings, JO (1988) *Applied Regression Analysis*. Wadsworth & Brooks, Pacific Grove, CA.

Seasholtz, MB & Kowalski, BR (1993) *Anal. Chim. Acta* **277**, 165-177.

Smith, PB & Snyder, AP (1992) *J. Anal. Appl. Pyrol.* **24**, 23-38.

Weigend, AS, Rumelhart, DE & Huberman, BA (1991) in Lippmann, RP, Moody, E & Touretzky, DS (Eds.), *Neural Information Processing Systems 3,* pp. 875- 882. Morgan Kaufmann, San Mateo, CA.