Statistics Using Neural Networks: Chance Effects

David J. Livingstone* and David T. Manallack

SmithKline Beecham Pharmaceuticals, The Frythe, Welwyn, AL6 9AR, Herts, U.K.

Received December 28, 1992

Perhaps one of the most exciting new developments in the field of artificial intelligence research is the emergence of useful applications of artificial neural networks. The appeal of these experiments is that they are attempts to simulate intelligence by the construction of models of the human brain which have similarities to the physical structure and organization of the brain. Artificial neural networks consist of simple processing units which are often arranged in interconnected layers. Typically, data are introduced to an input layer and passed through a so-called hidden layer to produce a response at an output layer.

One of the tasks which the human mind carries out so well is pattern recognition. If we consider that many statistical procedures are examples of pattern recognition, then the use of neural networks to perform these functions holds much promise. There have been several recent reports[1-3] of the application of networks to the analysis of chemical data sets in the field of quantitative structure–activity relationships (QSAR). Networks have also been applied to the prediction of other properties from a consideration of chemical structure, for example, the musk odor of a set of nitrobenzene derivatives[4] and the aqueous solubility of a diverse set of compounds.[5] We are involved in the search for an explanation of biological activity in terms of chemical structure and are thus interested in any new methods to achieve this end.

Standard statistical techniques such as regression and discriminant analysis are frequently used in QSAR, and the neural network analogues of these methods have shown interesting results. In the parlance of pattern recognition methodology these techniques are "supervised learning" since the data are used to supervise the learning, or training, of the algorithms. Since these supervised methods seek to fit a model to a data set, there is the potential for apparently good fits to occur by chance. The danger of chance correlations has been recognized for regression analysis[6] and discriminant analysis,[7,8] and guidelines have been proposed which will minimize the possibility of these chance effects happening. The situation is not so clear in the case of modeling using a neural network since the "fit" of a network is dependent on the network architecture as well as the data used to describe the biologically active molecules.

We have shown[9] that neural networks can be trained to carry out discriminant analysis using random numbers as input data. Network performance was dependent on the ratio of cases to connections (ρ) as proposed by Andrea and Kalayeh.[2] Since the networks were able to train using random numbers, it would appear that they are "memorizing" the data. Discriminant analysis requires a yes/no decision for classification which may lend itself to the way that neural networks operate. By contrast, regression involves training to values of a continuous target variable. We have examined the performance of networks designed to carry out multiple linear regression by using random numbers as input data and a single, random, continuous target.

Detailed results from these experiments are given in Table I, and Figure 1 summarizes these data with a plot of correlation coefficient (R^2) vs ρ for the four input unit networks. In order to recommend a critical ρ value for regression analysis, it is necessary to decide which is an acceptable risk of chance correlation. In other words, what value of R^2 by random fit can be tolerated without prejudicing the results of data modeling by regression. In an examination of chance effects using a standard regression package and random numbers, Topliss and Edwards demonstrated the importance of the number of variables considered.[6] As the number of observations in a data set was increased, for a given number of starting variables, so the average number of variables included in the resultant regression equations decreased, as did the average R^2 value. It is difficult to compare directly these results with ours since the experiments were designed to judge the probability of a chance fit based on the ratio of observations to descriptors screened. The numbers of terms in the reported regression equations, and hence adjustable parameters, are low compared with the equivalent quantity (connections) in the network models. However, our results are in broad agreement since R^2 increases as ρ decreases. A decrease in ρ is effectively a decrease in the ratio of the number of observations to adjustable parameters which might be likened to an increase in the number of variables considered.

Inspection of Figure 1 shows that the increase in the R^2 curve is at its steepest in the region of $1 < \rho < 3$, and at the mid-point of this region 74% of the variation in the "dependent" (random) variable is described by the network. Whatever level of chance correlation might be considered acceptable for a particular application, it seems unlikely that such a high figure would be reasonable. Real data, however, with its inbuilt structure, both dependent and independent, may not behave like this. A reported regression analysis by a network of a set of DHFR inhibitors[2] states that the optimum range for ρ was $1.8 < \rho < 2.2$. At smaller values than this range the network simply "memorized" the data while at higher values than 2.2 the network predictions were poor. The performance of these networks was assessed not just by fit, but also by prediction; using random numbers one would obviously expect prediction to be poor.

One potential reason for the observed poor predictions at ρ values higher than 2.2 may have been that too few connections were available to develop the linear, nonlinear, and/or complex cross-product terms required to relate biological to chemical properties. An alternative way of reducing the possibility of chance effects which allows a lower value of ρ was described recently by Weinstein et al.[10] In this study, the number of cases was randomly divided into 10 subsets. Ten networks were trained using nine-tenths of the data, and for each network a different test set was left out for prediction. Using such a scheme, a lower value of ρ may be employed as network performance can be monitored via cross-validation. Cross-validation has not been used to assess the performance of these fits to random data since it seems most likely that there is no "true" model in the data and thus a test of predictive ability should always give poor results. With real or structured

0022-2623/93/1836-1295$04.00/0

Table I. Effect of Varying Network Architecture on Regression Performance

network[a] architecture	connections[b]	ρ[c]	total RMS error	R^2 ± SEM
4,1,1	7	7.14	0.261 ± 0.005	0.214 ± 0.022
4,2,1	13	3.85	0.212 ± 0.006	0.434 ± 0.035
4,3,1	19	2.63	0.190 ± 0.008	0.542 ± 0.041
4,4,1	25	2.0	0.143 ± 0.006	0.743 ± 0.025
4,5,1	31	1.61	0.108 ± 0.009	0.852 ± 0.023
4,6,1	37	1.35	0.081 ± 0.004	0.915 ± 0.012
4,7,1	43	1.16	0.042 ± 0.004	0.977 ± 0.005
4,8,1	49	1.02	0.035 ± 0.008	0.985 ± 0.007
4,4,1	25 (15 cases)	0.6	0.020 ± 0.002	0.996 ± 0.001
4,4,1	25 (45 cases)	1.8	0.140 ± 0.007	0.770 ± 0.025
4,4,1	25 (55 cases)	2.2	0.163 ± 0.006	0.672 ± 0.023
4,4,1	25 (135 cases)	5.4	0.242 ± 0.005	0.318 ± 0.015
6,3,1	25	2.0	0.130 ± 0.006	0.779 ± 0.025
2,6,1	25	2.0	0.187 ± 0.010	0.555 ± 0.047

[a] Network architecture, giving the number of units in the input, hidden, and output layers, respectively. [b] The number of connections in the network. Results are the average of 10 experiments for each network architecture. New sets of random numbers were generated for each experiment. [c] Ratio of the number of cases (50, unless otherwise stated) to the number of connections.

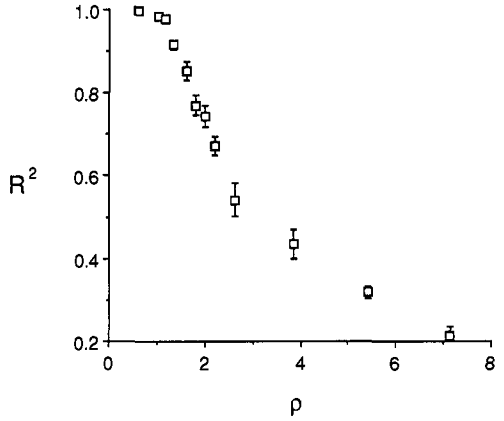

Figure 1. Plot of R^2 *vs* ρ for the first 12 network architectures listed in Table I. Random numbers (uniformly distributed) were generated using the RS1 data analysis package (BBN software, Staines, U.K.), and novel data sets were generated for each individual experiment. Neural networks were created using a commercial package, ANSIM (Science Applications International Corporation, San Diego, CA) and trained using the feed forward, back propagation algorithm.[12,13] Networks consisted of an input layer, an output layer, and a single layer of hidden units. The number of units in the hidden layer was altered to investigate network performance with respect to the number of connections. The data sets used for the regression analysis simulations typically had 50 cases (equivalent to 50 compounds) of five random variables, corresponding to four independent variables and one dependent variable. A limited number of networks were investigated by varying the number of cases, thus altering ρ, and two additional series of networks were run in which the number of independent input variables was changed. Training was halted when the maximum output unit error was less than 0.05 or the total RMS error was reducing at rate of less than 1 × 10^{-6} per cycle through the data. In the latter case, network weights and biases were perturbed by small random values, and training was allowed to continue until either of the above criteria was reached, and the network was then halted. This perturbation was applied in an attempt to ensure that the network had trained to the desired global minimum endpoint. Results from each trained network were assessed by comparison of the output and target values; since the output is a continuous variable, it is possible to calculate a correlation coefficient.

data, however, cross-validation is a useful measure of predictive ability and may also be used to judge how far network training should be carried out.

An additional factor concerning the choice of ρ involves the number of input units. In the first part of Table I and in Figure 1 we have concentrated our examination of chance effects using networks with four input units. Although we had previously found[9] that the number of input units did not appear to affect the results of discriminant analysis using two output units, regression by networks involves training to a continuous target using a single output neuron. We therefore investigated the dependence of regression performance on the number of input units. Table I demonstrates that at a ρ value of 2.0 the 2,6,1 network series does not perform to the same level as the 4,4,1 network, an R^2 value of 0.55 for the former and 0.74 for the latter. The other network with a ρ value of 2.0, the 6,3,1 network, gives an R^2 value of 0.78 which is a slight improvement on the 4,4,1 result but which perhaps suggests that network performance is reaching a maximum as a function of the number of input units. Additional work is being carried out to assess the performance characteristics of networks with differing numbers of units in the input layer. Furthermore, as random numbers do not adequately represent the type of data normally encountered in a QSAR study (i.e., structured), the experiments reported here will be repeated using structured data and a number of real QSAR data sets.

In conclusion, it has been shown that neural networks may be used to perform standard statistical tasks such as regression and discriminant analysis but that they suffer from the dangers of chance effects as shown here with randon number data. The ratio of observations to connections in a network has been shown to be an important determinant of performance, as has the number of units employed in the input (i.e. number of variables) layer. Some general guidelines concerning the ratio of observations to connections, ρ, can be stated:

For two unit discriminant networks ρ should exceed 2.0.[9]

For regression networks which do not employ a training/test set procedure to monitor overtraining, ρ should exceed 3 to keep chance correlations below R^2 = 0.5.

Alternatively, ρ values below 3 can be employed if some form of cross-validation scheme is implemented to examine predictive ability and thus avoid overtraining. The advantage of using a lower ρ value is that sufficient connections are available if complex nonlinear and cross-product terms are needed to solve the problem.

Finally, it appears that neural networks offer some advantage over standard statistical methods of modeling data since they can recognize complex relationships in the data without these having to be explicitly included in the analysis. One disadvantage to this form of modeling is that the importance of individual variables, as shown by the magnitude of their regression or discriminant coefficients, is not seen. It is possible to extract the connection weights from individual variables, but it has been suggested that contributions, the product of hidden unit activations and weights, is a more useful determinant of the "responsibility" of individual units in a network.[11]

References

(1) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure-Activity Relationship Analysis. *J. Med. Chem.* **1990**, *33*, 2583–2590.

(2) Andrea, T. A.; Kalayeh, H. Applications of Neural Networks in Quantitative Structure-Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824–2836.

(3) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Structure-Activity Relationships. *J. Med. Chem.* **1990**, *33*, 905–908.

(4) Chastrette, M.; de Saint Laumer, J. Y. Structure-odor Relationships Using Neural Networks. *Eur. J. Med. Chem.* **1991**, *26*, 829–833.

(5) Bodor, N.; Harget, A.; Huang, M.-J. Neural Network Studies. 1. Estimation of the Aqueous Solubility of Organic Compounds. *J. Am. Chem. Soc.* **1991**, *113*, 9480–9483.

(6) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.

(7) Whalen-Pedersen, E. K.; Jurs, P. C. The Probability of Dichotomization by a Binary Linear Classifier as a Function of Training Set Population Distribution. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 264–266.

(8) Stouch, T. R.; Jurs, P. C. Monte Carlo Studies of the Classifications Made by Nonparametric and Linear Discriminant Functions. 2. Effects of Nonideal Data. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 92–98.

(9) Manallack, D. T.; Livingstone, D. J. Artificial Neural Networks: Applications and Chance Effects for QSAR Data Analysis. *Med. Chem. Res.* **1992**, *2*, 181–190.

(10) Weinstein, J. N.; Kohn, K. W.; Grever, M. R.; Viswanadhan, V. K.; Rubinstein, L. V.; Monks, A. P.; Scudiero, D. A.; Welch, L.; Koutsokos, A. D.; Chiausa, A. J.; Paull, K. D. Neural Computing in Cancer Drug Development: Predicting Mechanism of Action. *Science* **1992**, *258*, 447–451.

(11) Sanger, D. Contribution Analysis: A Technique for Assigning Responsibilities to Hidden Units in Connectionist Networks. *Connection Sci.* **1989**, *1*, 115–138.

(12) Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing, Vol. 1*; MIT Press: Cambridge, MA, 1988.

(13) Salt, D. W.; Yildiz, N.; Livingstone, D. J.; Tinsley, C. J. The Use of Artificial Neural Networks in QSAR. *Pestic. Sci.* **1992**, *36*, 161–170.