Artificial Neural Networks as Statistical Tools in SAR/QSAR Modeling

H. G. Claycamp, N.B. Sussman, O. Macina and H.S. Rosenkranz
University of Pittsburgh
260 Kappa Drive, Pittsburgh, PA 15238

Abstract

There are two broadly-defined applications of artificial neural networks (ANNs) in SAR/QSAR modeling. The first is the use of networks as preprocessors to reduce the dimensionality of chemical descriptors for use in statistical or network models. The second is to create classification models for predictive toxicology. This report discusses the use of ANNs as classifiers in SAR/QSAR modeling and compares the approach to modeling using linear discriminant analysis (LDA) or logistic regression (LR).

Introduction

Artificial neural networks (ANNs) have been used for a variety of modeling and data processing applications in predictive toxicology. Applications of ANNs in predictive toxicology include their use as "pre-processors" to either transform data or reduce the dimensionality of e.g., Polanski et al., 1998), as a clustering tool, and as statistical classifiers (e.g., Huuskonen et al., 1998). In some applications, clustering could be considered to be classifications among several categories. Our initial efforts have focused on using ANNs for classification of chemicals by activity under a given toxicological end point. Presently we compare the use of ANNs with statistical classification techniques including the linear discriminant (LD) and logistic regression (LR). The classification studies generally use data bases derived for either QSAR or SAR studies.

There are a number of statistical approaches to classification including linear discriminant analysis and logistic regression. Neural networks as classifiers are often compared to logistic regression due to the similarity in mathematical expressions at the center of the processes. For example, logistic regression is of the form

$$\frac{p}{(1-p)} = \frac{1}{1+\exp(- \quad \beta_i x_{ij} + \beta_0)} \quad .$$

In this expression, the β_i are the coefficients, β_0 is the constant term and the left side of the equation is the odds ratio. The similarity between this equation and a common expression at the nodes of neural networks is obvious:

$$output = \frac{1}{1+\exp(- \quad w_i x_{ij} + \theta)} \quad .$$

While the two equations are a convenient starting point for comparison and to teach ANNs, the comparison with logistic regression is, in fact, coincidental. For example, an expression for node calculations that is often more useful than the logistic equation is the tanh function:

$$output = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$\text{where } z = \quad w_i x_{ij} + \theta$$

Perhaps the principal reason for the lack of interest in neural networks as statistical classifiers is the notion that the networks represent "black boxes" with respect to writing the final model equation. For example, a model consisting of (e.g.) 10 descriptors, or independent variables, would generate at least 10 coefficients and the offset term (B_0) in the logistic model. In an equivalent neural network model, perhaps 198 weights (w_i) would be used to model the data. Although the network weights are visible to the investigator, the "black box" notion refers to the fact that it is difficult to assign the magnitude of the network weights to the influence of single variables or interactions among several variables in the model. Further complicating this situation is that there are multiple solutions possible given a final true error rate (or predictivity).

Finally, the LDA and LR approaches can be readily adapted to best subsets analyses in order to reduce the number of variables in a model without sacrificing overall error rates (or predicitivity). Analogous methods for neural networks, such as network "pruning" techniques are typically more complicated and often are performed with respect to a given example (e.g., Bishop, 1995). Ultimately, investigators report supervised pruning of the networks, referring to repeated training of networks using manually selected subsets of variables. Presently, and in a companion paper (Sussman et al.), we

present simplified methods for reducing the number of variables in the model.

Methods

A data base of chemicals was prepared from the results of the produced Chernoff/Kavlock assay which uses end points of developmental toxicology (Chernoff and Kavlock, 1982). The specific developmental endpoints include reduced maternal weight, reduced maternal viability or embryotoxicity. The data base comprises 66 chemicals that were "active" under the assay criteria, and 59 chemicals for which no response was detectable. For the purpose of the present work, these chemicals were classified as "inactive."

A modified bootstrap approach was used. Ten subsets of 70 chemicals each were sampled randomly without replacement from the main data base. The remaining 37 chemicals after each selection was used as a testing set for that model.

Table 1. Physico-chemical variables used in the modeling.

Molecular Weight	Hydrogen Acceptor
Molecular Volume	Hydrogen Donor
Density	Hansen Hydrogen
Log P	%Hydrophilicity
Solubility Parameter	Water Solubility
Hansen's Dispersion	HOMO
Dipole Moment	LUMO
Hansen's Polarity	

Neural networks were implemented in either Microsoft™ VisualBasic 5.0 or NeuroSolutions™ Version 3.0. Networks were a multi-layer perceptron design that utilized a single hidden layer and a single output node. The equations at each network nodes was the *tanh* function unless otherwise specified. The network training method was backpropagation using a mean squared error (MSE) criterion for all examples in the batch. Progress was training was monitored using cross-validation, in which the current network model was applied to the corresponding testing set at the end of each backpropagation cycle. Training was stopped once that the MSE for the cross-validation set began to increase. In this manner, "over-fitting" of the networks was prevented.

The network design included fifteen variables from the data base which fed into 20 hidden layer nodes and one output node. For comparison on variable selections, the initial starting weights for each of the ten models were identical. Hereafter, this is referred to as "fixed initial weights." The ten subsets were modeled a second time using random initial weights (mean = 0 ± 0.5). For each of the ten sub models and in each set of starting weights, sensitivity, specificity and the distance parameter (Claycamp and Sussman, 1999), were calculated as a function of the decision level.

In each of the twenty models, the final weights were saved for statistical analyses. Each of the first layers yielded 15 x 20 matrices of weights, while the output layers were 20 x 1 matrices of weights.

Results and Discussion

The ten random models typically trained in less than 30 iterations ("epochs") each. The models yielded sensitivity and specificity that were comparable to the LDA and LR methods (Table 2). More interestingly, the three methods identified nearly identically the best three variables for the overall model (Table 3).

Table 2. Overall Sensitivities and Specificities for Ten Random Physico-Chemical Models

Modeling Tool	Sensitivity	Specificity
ANN: Random	0.521 ± 0.040	0.628 ± 0.098
ANN: Fixed	0.511 ± 0.033	0.611 ± 0.036
LDA	0.521 ± 0.039	0.606 ± 0.026
LR	0.516 ± 0.038	0.606 ± 0.027

ANN = artificial neural network
LDA = linear discriminant analysis
LR = logistic regression

Table 3. Variable selection using neural networks, linear discriminant and logistic regression

Neural Networks: Fixed	Neural Networks: Random	LDA	LR
WSOLUB	HDONOR	WSOLUB	WSOLUB
HANPOL	WSOLUB	HDONOR	HDONOR
LUMO	HANDISP	LUMO	LUMO
%HYDRO	LUMO	HOMO	HOMO

"Fixed" and "random" refer to the initial values of the weights in the first layer. "LDA"= linear discriminant analysis and "LR" = logistic regression.

The top four variables were identified using the sum of the absolute values of the weights for each variable. The top four variables are given in Table 3 for the ANNs, LR

and LDA techniques. Since the hidden layer contained 20 nodes, each variable's weight sum had twenty weights associated with it.

The rank order of the variables in Table 3 is semi-quantitative. For example, some adjacent variables in the ANN models differed by 0.7% in terms of the total weight. This difference is unlikely to be statistically significant. Furthermore, the comparisons are best kept among the top three variables for the LDA and LR models, since the fourth variable had only a minor influence on the models.

The use of identical starting weights for subsets of the data enabled selection of the dominant variables in the model (Table 3). By using the same starting weights, the uncertainty was reduced among in the final weights for the 10 models. This observation is best illustrated using plots of the signed weight sums for each of the 15 variables (Figures 1-2). The use of random starting weights resulted in apparently random sets of final

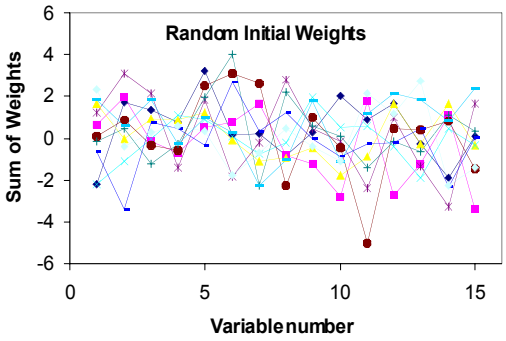

Figure 1. Final network weights sums using random initial weights. The sum of the final values of the first hidden layer weights are shown for fifteen variables and ten models. Each point represents a sum of 20 weights for that variable.

weights (Fig. 2), which illustrates the common criticism of neural network approaches, that there are multiple solutions inside of the black box. While this is clearly true for the random starting weight procedure; nevertheless, the data were still useful for variable selection when the sum of the absolute values of the network weights were used (Table 3).

Some of the reluctance to rely on neural networks by SAR/QSAR investigators is likely to arise from the apparent deficiency in descriptive statistics from the networks. In the present work, it is clear that the residual uncertainty among weights in the fixed initial weight approach represents uncertainty due to the data. We are currently studying methods for directly comparing the observed uncertainty with the uncertainty

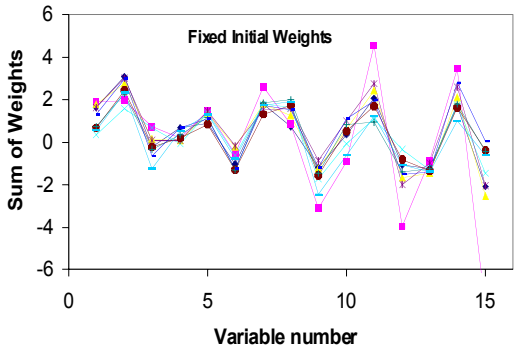

Figure 2. Final network weights sums using fixed initial weights. Ten models (networks) are shown that each began training with the same set of weights. The diminished spread of the weights compared to the random case (Fig. 1) was an aid in variable selection.

in the coefficients from the LDA and LR approaches. Similarly, the uncertainty among weights observed in the random initial weights approach can be used to estimate uncertainties among models (after subtracting the data uncertainty).

Finally, it has often been reported that about three times as many training examples as there are weights in the network should be used for training (e.g., Weiss and Kulikowski, 1991). The present work shows that the modified bootstrap approach can be used to recognize the controlling variables although the ratio was only (70 examples _ 320 weights =) 0.22. Variable selection comparable to the statistical techniques was accomplished in spite of the fact that, overall, the data base models were poorly predictive. Clearly, further exploration of this approach is warranted in order to determine its optimal utilization in SAR and QSAR

Conclusion

Valid comparisons among statistical and neural network techniques are possible with a level playing field. The present study shows that, once "level" comparisons are defined, neural networks perform well within the expectations of classical statistical methods. Finally, descriptive statistics for the characterization of coefficient and model uncertainty can be derived from experiments that fix and randomly vary initial starting weights.

References

Bishop CM (1995) Neural Networks for Pattern Recognition. Clarendon Press, Oxford England.

Chernoff N and Kavlock RJ (1982). An in vivo teratology screen utilizing pregnant mice. *J Toxicol Environ.Health* 10, 541-550.

Claycamp HG and Sussman NB (1999) A simple inter-class distance parameter for predictive SAR/QSAR models. *QSAR* (In press).

Huuskonen J, Salo M, and taskinen J (1998) Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J Chem Inf comput Sci* 38, 450-456.

Polanski J, Gasteiger J, Wagener M and Sadowski J (1998) The comparison of molecular surfaces by neural networks and its aplication to quantitative structure activity studies. *QSAR* 17, 27-36.

Weiss SM and Kulikowski CA(1991). *Computer Systems that Learn.* Morgan Kaufman Publishers, Inc., San Francisco.