The need for stochastic replication of ecological neural networks

Colin R. Tosh* and Graeme D. Ruxton

*Division of Environmental & Evolutionary Biology, Institute of Biomedical and Life Sciences,
Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, UK*

Artificial neural networks are becoming increasingly popular as predictive statistical tools in ecosystem ecology and as models of signal processing in behavioural and evolutionary ecology. We demonstrate here that a commonly used network in ecology, the three-layer feed-forward network, trained with the backpropagation algorithm, can be extremely sensitive to the stochastic variation in training data that results from random sampling of the same underlying statistical distribution, with networks converging to several distinct predictive states. Using a random walk procedure to sample error–weight space, and Sammon dimensional reduction of weight arrays, we demonstrate that these different predictive states are not artefactual, due to local minima, but lie at the base of major error troughs in the error–weight surface. We further demonstrate that various gross weight compositions can produce the same predictive state, suggesting the analogy of weight space as a ‘patchwork’ of multiple predictive states. Our results argue for increased inclusion of stochastic training replication and analysis into ecological and behavioural applications of artificial neural networks.

Keywords: alternate states; ecology; multiple attractors; neural networks; stochastic models

1. INTRODUCTION

Artificial neural networks are increasingly being used by ecosystem, behavioural, and evolutionary ecologists. A particularly popular model is the three-layer feed-forward network, trained with the backpropagation algorithm (e.g. Arak & Enquist 1993; Ghirlanda & Enquist 1998; Manel *et al.* 1999; Spitz & Lek 1999; Holmgren & Getz 2000; Kamo *et al.* 2002; Beauchard *et al.* 2003). The use of this design (especially if, as is common, the output layer consists of a single node) is that for a given set of input data, the network can be trained to make decisions, and this decision apparatus can subsequently be applied to inputs that are novel to the network. For example, an ecosystem ecologist with a finite set of ecological, biochemical and bird-occurrence data for a river environment can train a network to produce a predictive tool that will determine the likelihood of bird occurrence through sampling of the environment (Manel *et al.* 1999). Or in behavioural and evolutionary ecology, a network can be trained to distinguish between a ‘resident animal’ signal and ‘background’ signals, and subsequently used to determine how stimulating a mutant animal signal is, and hence how signals can evolve to exploit receiver training (Kamo *et al.* 2002). Reasons for the popularity of the backpropagation training method (Rumelhart *et al.* 1986) include: its computational efficiency, robustness, and flexibility with regard to network architecture (Haykin 1999).

Despite the sensitivity of many other nonlinear modelling methods to variation in initial system state (Scott 1999), and the inherent tendency of ecologists to ‘replicate’, detailed investigation of the nature of variation in network properties following stochastic replication of training data and/or starting weight composition is rarely reported in ecological applications using the aforementioned network design. Treatment of replication in published studies varies greatly, from cases of apparent non-replication (a single sequence of training data and a single starting weight array) to studies where genuine, stochastic variation of starting weights and/or training data is clearly evident. It is also common for research procedures to be reported in insufficient detail to be able to gauge treatment with respect to replication of training. Even in studies where considerable effort has been made to replicate training procedures, the precise nature of resultant variation in network properties is not usually a significant theme in presentation of results. Hence, it is not always clear whether the variation in network properties that might result from, for example, training networks with different samples of the same variables from the same environment, or random sampling of the same set of environmental data, is trivial, or can lead to substantial differences in the predictive properties of networks.

The issue of stochastic replication of artificial neural networks is also important from more than a technical viewpoint. Ghirlanda & Enquist (2007) term the phenomenon in which different histories of experience (paths) may at first seem to produce the same behavioural effects yet reveal important differences when further examined, ‘path dependence’. Using a two-layer neural network with a single output node and the so-called δ training rule, the authors demonstrate that the path dependence can contribute to such

* Author for correspondence (c.tosh@bio.gla.ac.uk).

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rstb.2006.1973> or via <http://www.journals.royalsoc.ac.uk>.

One contribution of 15 to a Theme Issue ‘The use of artificial neural networks to study perception in animals’.

phenomena as lack of peak shift after ‘errorless discrimination learning’, decrease of peak shift during extinction testing, and the shift of generalization gradients towards the average of the test stimuli. Thus, examination of the properties of individual neural networks that differ subtly in training regime may inform on fundamental behavioural phenomena and also potentially relate to individual differences in the behaviour of real organisms.

As a part of ongoing work into the sensory basis of predator–prey interactions, we have investigated the consequences of stochastic variation in training regime, resulting from random sampling of the same underlying statistical distribution, for the predictive properties of a three-layer feed-forward, backpropagated neural network. Like Ghirlanda & Enquist (2007), we show that the consequences of subtle variation in the training dataset can be far from trivial, and can lead to networks with quite different, discrete, predictive properties.

2. MATERIAL AND METHODS

(a) *The neural network and training procedure*

We are using three-layer feed-forward neural networks as caricatures of the sensory surface/interneuron/sensory map system found in a range of taxa from flies to primates (Ashley & Katz 1994; Cohen & Newsome 2004; Shipp 2004). The first layer consists of 20 nodes, and subsequently ‘bottlenecks’ into a hidden layer of 10 nodes (neural bottlenecks are often seen as nerves leave sensory organs). Architecture then differs from the common single-node output architecture and the 20-node architecture of the input layer is reconstructed to represent an isomorphic spatial map. All layers are fully connected, hyperbolic tan (tanh) activation functions (Krakauer 1995; Ghirlanda & Enquist 1998) are used throughout, and training rate is set at 0.2.

During training, we presented networks with 4000, 20-element vectors consisting of ‘–1s’ (representing empty space) and ‘1s’ (an object), whose relative numbers per vector followed a normal random distribution with mean = 10 and s.d. = 2. Positioning within vectors followed a uniform random distribution. Starting weight arrays were constructed from a uniform random distribution between –1 and 1. Fifty-two such training input/weight array sets were created, and each was used to train a network over a maximum of 1000 epochs. The task of the backpropagation training algorithm was to reconstruct each input pattern in the output layer. Epoch for termination of training was chosen according to an early stopping procedure (Hecht-Nielsen 1990) in which a 4000 × 20 test input array was created as above, error in reconstructing its constituent vectors calculated after each epoch, and stopping point is defined as the epoch where error was at a minimum. Weight updating was sequential and input vectors were presented in the same order in each epoch.

(b) *Object targeting*

We are particularly interested in how object targeting by organisms is affected by the number of distracter objects (this is sometimes investigated under the guise of the ‘confusion effect’, e.g. Landeau & Terborgh 1986). Following training, we therefore chose an arbitrary network position (‘position 10’ in a two-dimensional network input, if the reader wishes to visualize the procedure) into which a target object of value equal to 1 was input, and the number of distracter objects (also equal to 1) varied from 0 to 19 (empty space again equal to –1). For each number of distracters, positioning around

the target object was varied 50 times according to a uniform random distribution, and targeting accuracy gauged as the proportion of occasions in which an object (gauged as output more than 0.9) was reconstructed at *output* position 10 (the equivalent position in the spatial map to the input unit stimulated by the target object). The same test arrays as just described were applied to all trained networks (hence, the following results are not due to stochastic variation in *test* data).

(c) *Characterizing network error–weight surfaces*

We initially suspected the discrete network predictive states which resulted from stochastic replication of training (see §3) to be artefactual, due to settling of the backpropagation algorithm in different local minima in the network error–weight hyperspace (Haykin 1999). We therefore characterized error–weight surfaces for one training data–starting weight set that terminated in each of the predictive states. For each of these four training data–starting weight sets, 14 random walks in weight space were initiated from both the end and start point of training. Each walk was started by randomly changing each element in the weight arrays by plus or minus three and subsequently determining total error in reconstructing the appropriate 4000 training vectors, measured using the total summed, squared measure employed within the backpropagation algorithm (see Rumelhart *et al.* 1986). This was repeated 50 times in order to characterize the region immediately surrounding the start and end points of training. Incremental change to weight elements was then increased to plus or minus 10 and the process repeated a further 50 times in order to characterize the wider weight space, and finally increments were increased to plus or minus 50 and the process repeated 50 more times, to characterize gross weight space. Thus, each walk consisted of 150 samplings of weight space. As sampling of gross weight space was relatively coarse, the complete sampling procedure of 28 walks was repeated eight times for each training data–starting weight set. While this is a thorough sampling of weight space (33 600 samples for each unique surface) we still cannot entirely exclude the possibility of extremely narrow error minima, especially in the area of gross sampling. The 400 weight values in each vector of the set of 4200 samples of weight space (28 walks of 150 samples each) were reduced to single points in two-dimensional space using Sammon mapping (Sammon 1969) implemented using the ‘sammon’ function of the SOM Toolbox for MATLAB (<http://www.cis.hut.fi/projects/somtoolbox/>), run for 200 iterations with a step size of 0.2. This dimensional reduction procedure incrementally reduces the discrepancy between total Euclidean interpoint distance in multi- and lower-dimensional space using a pseudo-Newton error minimization method, and tends to associate points with similar weight compositions. It was used in the present application in preference to more familiar procedures such as PCA because the first two components in the latter method (and allied dimensional reduction procedures) usually capture only a small proportion of the original variation in data with high order datasets such as those used presently. Following two-dimensional mapping, surfaces were produced by Delauney triangulation and plotted against network training input reconstruction error.

(d) *Mapping the relationship between end-weight composition and network predictive properties*

We further employed Sammon mapping to characterize weight arrays at the start and end of training for two of the

Figure 1. The predictive properties of 52 neural networks created through stochastic replication of input training dataset and starting weight composition. Replicates with similar properties are presented together in each graph. Refer to §2 for the treatment of object targeting by neural networks, including the definition of ‘network targeting accuracy’ (y -axis).

distinct network predictive states obtained (those shown in figure 1*a,b*; results and conclusions presented applied to the other predictive patterns, but data were not shown to ease visualization). Specifically, we wanted to know if similar predictive patterns arose from similar weight compositions, or if quite different weight compositions can produce the same predictive property.

3. RESULTS AND DISCUSSION

Figure 1 shows the object targeting versus group size relationships obtained following 52 stochastic training runs. It can be seen clearly that stochastic variation in training procedure was sufficient to create at least four discrete network predictive states. Object targeting was either (i) always accurate, (ii) more accurate for smaller numbers of input objects, (iii) more accurate for small

and large numbers of input objects, less for intermediate sizes, or (iv) more accurate for larger numbers of input objects.

Generally, error-weight surfaces of representative training data-starting weight sets that terminated in each of the predictive states consisted of a plateau with a single major error minimum, to the bottom of which the backpropagation algorithm lead weight arrays (figure 2). We found no example of a lower error value than the end point of training in any of the 33 600 sample points for each unique error surface, although there did appear to be examples of major alternative minima in some surfaces (figure 2*a7,b5,c5-6,d1,d7,d8*, for example). These regions should be viewed with caution, however, as they could be produced through misplacement of points in Sammon mapping. We conclude that the end weight values for networks

Figure 2. Neural network error surfaces for one example training data-starting weight set associated with each network predictive pattern illustrated in figure 1. Parts (a,b,c,d) refer to the pattern with corresponding letters in figure 1. Multiple random walks were initiated from the start and end point of backpropagation training (shown by crosses on the surface) and surfaces produced by Sammon mapping of weight vectors, Delaunay triangulation of Sammon map values in two dimensions, and subsequent plotting against network training input reconstruction error. Numbers 1–8 refer to repeats of the random walk procedures for each training data-starting weight set. The figure may be viewed in more detail online, where it is available as electronic supplementary material. Further details can be found in §2.

predicting each of the object targeting patterns (figure 1) are major error minima (if not the principal minima) in the error surface and there is little evidence to suggest that the various network predictive patterns result from suboptimal convergence of networks into local regions of intermediate error reduction.

Figure 3 indicates that quite different weight compositions at the end of network training can produce the same predictive property and vice versa. This suggests that the weight space of this neural network can be considered something of a patchwork of predictive states and the eventual properties of a network will depend on the particular ‘patch’ the training algorithm converges on. The observation from figure 3 that the system can converge to different weight compositions and predictive properties from similar

starting weight arrays, indicates the primacy of stochastic variation of training input data rather than starting weights in the phenomena discussed. This was confirmed by training with the same starting weight arrays but different training sets: networks still converge to different predictive states (data not shown).

(a) *The need for stochastic replication of ecological neural networks*

A point we wish to make from the preceding demonstrations is that a researcher investigating ecological or behavioural phenomena using the described system in a non-replicated fashion might come to a quite different biological conclusion relative to the researcher who has stochastically replicated the system and fully investigated predictive properties

Figure 3. Sammon map of weight composition at the start and end of network training, for networks that converge to the predictive states shown in figure 1*a,b*. The fact that there is no pronounced same-symbol aggregation within the figure indicates that multiple networks with quite different weight array compositions can have the same predictive properties, and networks with relatively similar weight compositions can have quite different properties: the weight space appears to be a patchwork of different network predictive states.

therefrom. The stochastic variation in network training data that results from random sampling of the same statistical distribution can lead to dramatic differences in network predictive states in three-layer feed-forward, backpropagated neural networks. Although the network we have used differs both architecturally and in training objective to the typical decision-making network used in ecology, pronounced effects on the behaviour of artificial neural networks consequent upon subtle differences in training regime have also been demonstrated in a two-layer feed-forward network with a single output node trained with the gamma rule (Ghirlanda & Enquist 2007) and a five-layer feed-forward network with a single output node trained with a genetic algorithm (C. Tosh & G. Ruxton 2006, unpublished work). In the latter system, networks were trained to be very specialized with regard to choice of plant-like resource objects projected onto the input layer and varied randomly in starting weight composition and the position and order of objects projected onto the input layer. When the ability of networks to choose an appropriate resource object while being 'distracted' by different numbers of inappropriate objects was analysed, some networks showed decreasing discrimination accuracy with increasing numbers of distracters and some showed increasing discrimination. Thus, path-dependent effects may be observed in a variety of artificial neural network systems. We do not think that ecologists and behavioural biologists who have applied networks with similar designs without a full appreciation of path-dependent effects should be

unduly alarmed by our results, however. We suspect that many conscientious researchers in these disciplines do in fact replicate networks and analyse the range of resultant properties routinely in research procedures (presumably excluding this as a significant source of variation in their system) prior to presenting work for publication. We do, however, encourage researchers who have not considered the variation in neural network properties that can arise from stochastic replication of training procedures, to include such replication as matter of course in research procedures. Procedures that ecologists might consider include obtaining multiple training datasets from the same environment, random subsampling from the same dataset prior to network training, and random ordering of input vectors from the same dataset for presentation to the network. These procedures may involve considerable additional effort on the part of the researcher and may make results more or less interesting for subsequent publication, but workers who do not apply such procedures should be aware that the research results they present may represent only one of a range of possible predictive network states.

What is the biological significance of path-dependence? Ghirlanda & Enquist (2007) show that subtle variations in an artificial neural network training regime can contribute to important phenomena in animal behaviour, such as lack of peak shift after errorless discrimination learning, decrease of peak shift during extinction testing, and the shift of generalization gradients towards the average of the

test stimuli. Could path-dependent effects also contribute to inter-individual variation in behaviour? Certainly, individuals of a species from the same population will experience a similar set of stimuli during neural network training, but with subtle differences in the temporal order and content of stimulus sets. Could such variation contribute to behavioural syndromes, the consistent differences in the behavioural tendencies of individuals of the same species (Sih *et al.* 2004)? While speculative, the demonstration that many artificial neural networks are inherently sensitive to subtle variation in training regime indicates that path dependence will provide a useful framework to tackle such questions.

This work was funded by the UK Biotechnology and Biological Sciences Research Council (Grant no: BBS/B/01790).

REFERENCES

- Arak, A. & Enquist, M. 1993 Hidden preferences and the evolution of signals. *Phil. Trans. R. Soc. B* **340**, 207–213.
- Ashley, A. A. & Katz, F. N. 1994 Competition and position-dependent targeting in the development of the *Drosophila* R7 visual projections. *Development* **120**, 1537–1547.
- Beauchard, O., Gagneur, J. & Brosse, S. 2003 Macroinvertebrate richness patterns in North African streams. *J. Biogeograph.* **30**, 1821–1833. (doi:10.1111/j.1365-2699.2003.00954.x)
- Cohen, M. R. & Newsome, W. T. 2004 What electrical microstimulation has revealed about the neural basis of cognition. *Curr. Opin. Neurobiol.* **14**, 169–177. (doi:10.1016/j.conb.2004.03.016)
- Ghirlanda, S. & Enquist, M. 1998 Artificial neural networks as models of stimulus control. *Anim. Behav.* **56**, 1383–1389. (doi:10.1006/anbe.1998.0903)
- Ghirlanda, S. & Enquist, M. 2007 How training and testing histories affect generalization: a test of simple neural networks. *Phil. Trans. R. Soc. B* **362**, 449–454. (doi:10.1098/rstb.2006.1972)
- Haykin, S. 1999 *Neural networks: a comprehensive foundation*. Upper Saddle River, NJ: Prentice Hall.
- Hecht-Nielsen, R. 1990 *Neurocomputing*. New York, NY: Addison-Wesley.
- Holmgren, N. M. A. & Getz, W. M. 2000 Evolution of host plant selection in insects under perceptual constraints: a simulation study. *Evol. Ecol. Res.* **2**, 81–106.
- Kamo, M., Ghirlanda, S. & Enquist, M. 2002 The evolution of signal form: effects of learned versus inherited recognition. *Proc. R. Soc. B* **269**, 1765–1771. (doi:10.1098/rspb.2002.2081)
- Krakauer, D. J. 1995 Groups confuse predators by exploiting perceptual bottlenecks: a connectionist model of the confusion effect. *Behav. Ecol. Sociobiol.* **36**, 421–429. (doi:10.1007/s002650050165)
- Landeau, L. & Terborgh, J. 1986 Oddity and the confusion effect in predation. *Anim. Behav.* **34**, 1372–1380. (doi:10.1016/S0003-3472(86)80208-1)
- Mamel, S., Dias, J. M., Buckton, S. T. & Ormerod, S. J. 1999 Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *J. Appl. Ecol.* **36**, 734–747. (doi:10.1046/j.1365-2664.1999.00440.x)
- Rumelhart, D., Hinton, G. & Williams, R. 1986 Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition*, vol. 1 (eds D. Rumelhart & J. McClelland). *Foundations*, pp. 318–363. Cambridge, MA: MIT Press.
- Sammon, J. W. 1969 A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **C-18**, 401–409.
- Scott, A. 1999 *Nonlinear science*. Oxford, UK: Oxford University Press.
- Shipp, S. 2004 The brain circuitry of attention. *Trends Cogn. Sci.* **8**, 223–230. (doi:10.1016/j.tics.2004.03.004)
- Sih, A., Bell, A. M., Johnson, J. C. & Ziemba, R. E. 2004 Behavioral syndromes: an integrative overview. *Q. Rev. Biol.* **73**, 241–278.
- Spitz, F. & Lek, S. 1999 Environmental impact predication using neural network modelling. An example of wildlife damage. *J. Appl. Ecol.* **36**, 317–326. (doi:10.1046/j.1365-2664.1999.00400.x)