Dynamic Pattern Selection: Effectively training Backpropagation Neural Networks

A. Röbel
Department of Applied Computer Science, Technical University of Berlin
10587 Berlin, Germany

1 Introduction

It is well known that the generalization properties of neural networks used in function approximation are strongly affected by the size and distribution of the training set. However, the problem of selecting the optimal training set has not yet been solved. For a good generalization the training set has to contain enough information to fix the network function f_n not only at the training patterns, but on the domain $\mathbb{X}$ of the target function f_t.

To achieve this, we want to adapt the training set during training and employ the net function to decide which pattern should be chosen. Following this strategy the *active pattern selection* algorithm, which has been demonstrated to select surprisingly small training sets has been developed (Plutowski *et al.*, 1993). In contrast to this algorithm, the *dynamic pattern selection* proposed here, is constructed to effectively select the training sets by continually validating the generalization properties of the net (Röbel, 1993b).

2 Dynamic Pattern Selection

Assume that the task to solve consist of approximating a smooth target function f_t on its domain $\mathbb{X}$. For training a backpropagation network there are only a number of examples available which are in the set of available patterns $\mathbb{D}_a$.

As the distribution of the data in $\mathbb{D}_a$ is not adapted for neural network training, using all the data in $\mathbb{D}_a$ will, in general, not lead to optimal training results (Röbel, 1993b). We would like to select the training set $\mathbb{D}_t \subset \mathbb{D}_a$ such that minimizing the net error $N(\mathbb{D}_t)$ forces the network function f_n to regularly converge to f_t over the complete domain $\mathbb{X}$. The two questions to solve are: When should the new pattern be selected and which one should be chosen.

To answer the first question, we state that the training set shall be increased, whenever the generalization properties of the net becomes poor. As is known from cross validation, we divide $\mathbb{D}_a$ into the training and validation repertoire $\mathbb{D}_T$ and $\mathbb{D}_V$. The training data is selected from $\mathbb{D}_T$, while the validation set $\mathbb{D}_v$ is randomly chosen from $\mathbb{D}_V$ where $|\mathbb{D}_v| = |\mathbb{D}_t|$[1]. Based on theoretical investigations of the regular generalization of training sets, we define the generalization

[1] $|\mathbb{A}|$ denotes the cardinality of $\mathbb{A}$

training set size	training error $E(\mathbb{D}_s)$	generalization error $E(\mathbb{D}_u)$	generalization factor ρ_u	for/backward propagations
15	4.29e-3 ± 2.0e-3	2.50e-2 ± 1.7e-2	5.83	0.54e+6
25	3.77e-3 ± 1.8e-3	4.80e-3 ± 2.0e-3	1.27	0.90e+6
35	2.46e-3 ± 1.2e-3	2.90e-3 ± 1.4e-3	1.18	1.26e+6
50	3.02e-3 ± 1.6e-3	3.46e-3 ± 1.8e-3	1.15	1.80e+6
75	2.99e-3 ± 1.5e-3	3.14e-3 ± 1.7e-3	1.05	2.70e+6
100	2.90e-3 ± 1.6e-3	3.21e-3 ± 1.7e-3	1.10	3.60e+6
150	4.53e-3 ± 2.1e-3	5.08e-3 ± 2.4e-3	1.12	5.40e+6
69 ± 10	3.69e-3 ± 1.7e-3	3.06e-3 ± 1.1e-3	0.83	0.74e+6

Tab. 1. Predicting the Henon model. Comparison of the average training and generalization rms error using a 2-7-1 neural net and different training sets. The patterns in the fixed sets are approximately equally spaced on the henon attractor.

factor

$$\rho = \frac{N(\mathbb{D}_v)}{N(\mathbb{D}_s)}, \tag{1}$$

which is a normalization of the generalization error measured by means of the backpropagation error function N(). The generalization factor indicates the error we make in optimizing on $\mathbb{D}_s$ instead of $\mathbb{X}$. For regular generalization we demand that

$$\rho \leq 1.0. \tag{2}$$

As the selection overhead should be as small as possible, it is reasonable to select the training patterns based on an easily calculable criterion. We select the maximum error pattern out of $\mathbb{D}_T$ which is not in $\mathbb{D}_t$ such that the resulting training set resembles a worst case collection of patterns.

The training set may now be controlled to achieve a faithful generalization. The straightforward strategy, namely to enlarge $\mathbb{D}_t$ whenever the generalization factor is greater than one, results in very small training sets. For high precision training, however, the selection process turns out to be too slow. Estimating the tendency of the generalization factor and inserting the new training pattern whenever the generalization factor increases, fixes the problems (Röbel, 1993b).

3 Experimental Results

The properties of the *dynamic pattern selection* will be demonstrated by solving two nonlinear signal prediction tasks. For training, a batch mode backpropagation algorithm with adapted learning rate and momentum is used (Salomon, 1991).

In the first experiment the chaotic time series of the henon model (Grassberger and Procaccia, 1983)

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} y_n + 1 - a \cdot x_n^2 \\ b \cdot x_n \end{pmatrix} \quad \text{with} \quad (a = 1.4, b = 0.3) \tag{3}$$

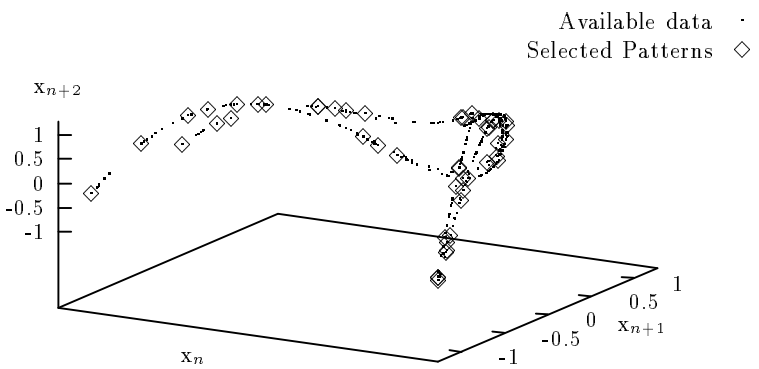

Fig. 1. A typical distribution of training patterns on the prediction function of the henon model. Depicted is the distribution of training patterns subsequent to 1000 training cycles.

is predicted. In table (1) the results for a number of fixed training sets are compared with the dynamic pattern selection. The normalized rms error $E()$ for the training set and an independent validation set $\mathbb{D}_u$ is given. The cost is estimated by the number of forward/backward propagations through the net. The fixed training sets with 35-100 patterns give the best results. In the case of dynamic training sets a mean of 69 patterns is selected. Due to the initially small training sets the cost for the dynamic selection training is considerably lower than for all appropriate fixed training sets.
Further investigation reveals that only the fixed training sets with more than 50 patterns achieve a stable generalization factor. Compared to this, the dynamically selected mean of 69 patterns is quite reasonable.

Along with the size of the training set, the distribution of the training patterns has a considerable effect on the generalization results. A typical distribution obtained by the dynamic pattern selection is shown in figure (1). The distribution is not uniform, but reflects the error distribution of the net function.

The second example is the prediction of the Mackey-Glass model

$$\dot{x}(t) = \frac{a \cdot x(t-\tau)}{(1+x(t-\tau)^{10})} - b \cdot x(t) \quad \text{with} \quad (a = 0.2, b = 0.1). \tag{4}$$

Lapedes and Farber (1987) demonstrated the prediction of the Mackey-Glass time series ($\tau = 30$) using a neural net with six input units, two hidden layers with ten units each and a linear output unit. These settings are used here, too. Lapedes and Farber used a fixed training set with 500 examples. Results for this training set and the dynamic pattern selection are shown in table (2). Achieving the same average generalization precision, the dynamically selected training sets contain a mean of 207 training patterns. The cost is lower by a factor of three. Due to the suboptimal distribution of the fixed training set, the generalization factor increases steadily.

training set size	training error $E(\mathbb{D}_s)$	generalization error $E(\mathbb{D}_u)$	generalization factor ρ_u	for/backward propagations
500	2.47e-2 $\pm$ 3.6e-3	2.73e-2 $\pm$ 3.4e-3	1.11	60e+6
207 $\pm$ 11	4.62e-2 $\pm$ 3.9e-3	2.74e-2 $\pm$ 1.4e-3	0.59	18e+6

Tab. 2. Predicting the Mackey-Glass model. Comparison of the average training and generalization rms error using a 6-10-10-1 neural net and different training sets.

4 Real world data and online training

The dynamic pattern selection algorithm has been successfully applied to optimize predictors of high quality musical signals (Röbel, 1993a). The experiments revealed that the dynamic pattern selection is not disturbed by noise, as long as the noise level is well below the prediction error.
It is widely accepted that, in the case of redundant data, the online mode training will yield superior results to training in batch mode. We therefore compared the computational expenses for online training with *Search-Then-Converge* learning rate schedule (Darken and Moody, 1991) and the dynamic pattern selection algorithm. The neural nets have been trained to predict a piano signal given 15000 training patterns.
Despite the preceding optimization for the online training algorithm and the fact that, in the case of dynamic training sets, only 50 out of the 15000 patterns have been selected (resulting in a considerable overhead for the selection procedure), the total expense for the dynamic pattern selection has been lower by a factor of three.

References

Christian Darken and John Moody (1991) Note on learning rate schedules for stochastic optimization. *Neural Information Processing Systems*, 3, 832–838.

P. Grassberger and I. Procaccia (1983) Estimation of the Kolmogerov entropy from chaotic signal. *Physical Review A*, 28(4),2591–2593.

A. Lapedes and R. Farber(1987) Nonlinear signal processing using neural networks: Prediction and system modelling. Technical Report LA-UR–87-2662, Los Alamos National Laboratory.

Mark Plutowski, Garrison Cottrell, and Halbert White (1993) Learning Mackey-Glass from 25 examples, plus or minus 2. *Preprint.*

Axel Röbel (1993a) Neuronal models of nonlinear dynamic systems and the application to musical signals. PhD thesis, Technical University of Berlin, (In German).

Axel Röbel (1993b) The dynamic pattern selection algorithm: Effective training and controlled generalization of backpropagation neural networks. Technical Report 93-23, Technical University of Berlin.

R. Salomon (1991) Improving connectionsts learning, based on gradient descend. PhD thesis, Technical University of Berlin, (In German).