TEACHING STRATEGIES FOR ARTIFICIAL NEURAL NETWORK LEARNING

S. Nordbotten*)

Abstract:

This paper presents an evaluation of the effects of variation of training set size, ordering of examples in the training set, adjustment (learning) rate, and reinforcement on pattern recognition in artificial single layer neural networks (ANNs) which use a learning algorithm based on the Widrow-Hoff principle. These parameters can be considered as alternative teaching strategies for ANNs.

The evaluation has been carried out as a set of simulation experiments on synthetic sets of patterns. The results indicate that for the type of pattern identification considered, learning in ANN is sensitive to the teaching strategy chosen.

1. Teaching and learning

A number of learning algorithms for different artificial neural networks have been developed in recent years [RUMELHART 1988]. The merits of different learning algorithms are currently being studied from theoretical as well as empirical perspectives [NORDBOTTEN 1990b].

In human training systems, there are two actors, the learner and the teacher, and material used for instruction. The complexity of the network topology and learning ability of a learner may be considered as inherited. How effectively a learner learns a given set of examples may depend, however, on the teacher's strategy with respect to the size of the training set used, the order in which the examples are presented, the intensity by which the examples are introduced, and the reinforcement of the teaching material. In the present study we consider the material used for instruction as predetermined.

In the development of an artificial network the designer will frequently define a network which can be trained by examples, choose an adequate learning algorithm suited for the network, and rely on a set of examples which serve as training material. The success of training can be evaluated in several ways:

1) during the training,
2) by evaluation tests or
3) by experience through practical application.

In the study reported in this paper, the impact of several teaching strategy factors has been investigated. The first factor investigated is the size of the training set used. A similar investigation for learning in stochastic knowledge bases used for consultation systems, has been reported in a previous paper [NORDBOTTEN 1991].

The second factor investigated, is the sequence in which the training examples are introduced to the learner. As for a human learner, we assume that the order in which training examples are presented may be important for ANN learning. Learning basic and simple examples before more complex is a frequently applied teaching strategy.

The size of the adjustment rate used by a training algorithm reflects how fast the learner adjusts to an error made. We assume that the adjustment rate can be controlled by the teacher and therefore also belongs to the teaching strategy.

A learner can be exposed repeatedly to each individual example in a training set a number of times before the teacher proceeds to the next example. We will call this strategy concentrated reinforcement. Alternatively, the learner can be exposed to each example of the training set sequentially, and then the exposition for the complete set is repeated a certain number of times. This strategy we will call dispersed reinforcement. Mixed strategies may also be designed by dividing the training set into partitions the examples of which are presented using concentrated reinforcement while the partitions are reinforced in a dispersed manner.

In reinforcement another important factor is the number of repetitive presentations of the training set to the learner. This factor we call the number of reinforcement cycles.

The performance of a trained network in our investigation is considered to be the ability of the network to correctly identify patterns. The overall aim of a teaching strategy is either to give the learner some predetermined performance level using a minimum of teaching investment, a maximum performance level by means of given teaching resources or a maximum of some weighted combination of performance level and teaching investment. We shall return in following
sections to the measurement of performance and teaching investment.

2. Teaching strategies

2.1 The Artificial Neural Network

The network model used in this study is a single layer network with M input sources providing simultaneous binary inputs denoted o[i, 0], i = 1 . . M, with value 0 or 1 to a set of N neurons which each generate an output o[j, 1], j = 1 . . N. The vectors of M input elements and N output elements are referred to respectively as a pattern and its identification.

A neuron is a processing unit characterized by its activation level and its output. The activation level is determined by the activation function:

\[ a[j, 1] = \text{SUM}[i] w[i, j] o[i, 0], \]

where \( w[i, j] \) denotes the weight between the input source i and the neuron j. The activation level is a real variable.

The output of the neuron is a binary variable:

\[ o[j, 1] = 1 \text{ if } a[j, 1] > a[k, 1], \]
\[ o[j, 1] = 0 \text{ if } a[j, 1] \leq a[k, 1] \]

for all \( k \neq j \), \( k,j = 1 . . N \).

This implies that a neuron never creates an output vector with more than one non-zero element. Usually an output vector will have only one non-zero element indicating the position of the pattern identification. In some situations, the output vector may have many zero elements, indicating that the network was unable to make an identification. One important aim is to train the ANN recognize the correct pattern identifiers for the input patterns.

In the real world applications we have in mind, there are frequently several different patterns associated with identical target vectors or pattern identifiers. This reflects the possibility of noise, uncertainty, or errors in the pattern.

2.2. The Learning algorithm

The learning algorithm used is based on the well known Widrow-Hoff algorithm [WIDROW 1960, 1985]. The core of the algorithm used in this investigation consists of two steps:

Step I. The forward computation by which the network, based on current knowledge, computes an activation vector \( a[j, 1] \), \( j = 1 . . N \), according to:

\[
\text{for } i := 1 \text{ to } M \text{ do } \\
\text{for } j := 1 \text{ to } N \text{ do } \\
a[j, 1] := a[j, 1] + w[i, j] o[i, 0];
\]

and.

Step II. The backwards computation of adjusted weights based on comparison of the target output vector \( t[j, 1] \), \( j = 1 . . N \), from the training set and the computed activation vector from Step I:

\[
\text{for } i := 1 \text{ to } M \text{ do } \\
\text{for } j := 1 \text{ to } N \text{ do } \\
w[i, j] := w[i, j] + \text{rate} * o[i, 0] * (t[j, 1] - a[j, 1]).
\]

Initially, all weights are set equal to zero. Rate is here the adjustment rate. During the learning process, the algorithm can be repeated in reinforcement cycles. Note that it is the computed activation vector, not the binary output vector, which is compared with the target vector.

The algorithm has the property that it adjusts the weights to minimize the sum of squares of the differences between the elements of the computed activation vector and the target vector. The sum of square errors (SSE) each pattern \( k \) of a training set of \( P \) patterns is:

\[
\text{for } k := 1 \text{ to } P \text{ do } \\
\text{for } j := 1 \text{ to } N \text{ do } \\
\text{SSE}[k] := \text{SSE}[k] + (t[j, 1] - a[j, 1])^2,
\]

while the mean square error (MSE) for a training set of \( P \) patterns after a learning cycle is:

\[
\text{for } k := 1 \text{ to } P \text{ do } \text{MSE} := \text{MSE} + \text{SSE}[k]/P
\]

A MSE decreasing in value from one reinforcement cycle to the next will indicate improved learning.

2.3. Strategy factors

2.3.1 The strategy vector

The teaching strategies can be represented in the vector space

\[ V = (S, O, L, C, R) \]

where \( S \) represents the size of a training set generated randomly from a probability distribution reflecting the patterns of the domain of interest, \( O \) represents training set order, \( L \) denotes the adjustment rate, \( R \) the reinforcement strategy, and \( C \) the number of reinforcement cycles.

2.3.2 Size of training set, \( S \)

A network's ability to learn a set of different patterns is a main property of the network. In evaluating
a network and its associated learning algorithm, the network can be exposed to and taught pattern sets randomly generated from the domain of interest. After learning, the same set of patterns can be presented to the network to determine how well the network has learned to identify the patterns.

The size of the training set can be easily varied. Two factors must be distinguished. One is the number patterns in the set. Keeping in mind that the training sets are random samples, subsets of identical patterns may exist. The second factor is the number of different target vectors or pattern identifiers present in the set. To each pattern identifier, a subset of different patterns may be associated. The differences are assumed to represent noises acting on the patterns. The ratio between the total number of patterns in the training set and the number of different pattern identifiers therefore indicates the average number of instances of each identifier within the training set. We would expect that the percent of different patterns correctly identified decreases as the number of patterns increases.

2.3.3 Order of presentation, \( O \)

In human teaching, a common strategy is to present to the learner simple patterns prior to more complex patterns. In our investigation a pattern is described as a binary vector with zero and one value elements. In all our patterns, the number of non-zero elements is less than the number of zero elements. The degree of complexity of a pattern is defined by the number of non-zero elements.

Our hypothesis is that learning will be more efficient if the network is exposed to a training set ordered by increasing complexity than if the network is exposed to the same set of patterns presented in a random order.

2.3.4 Adjustment rate, \( L \)

The adjustment rate can be considered as the intensity by which the learner is led or instructed to react to errors it makes in recalling a pattern when it learns the correct answer or the target vector. A adjustment rate of \( > 1 \) corresponds to an over-reaction while a adjustment rate of \( 0 \) corresponds to ignorance of errors, or inability to adjust knowledge to facts. We consider adjustment rates in the interval \( 0 < \text{rate} < 1 \) only.

A high rate should be expected to give a fast adjustment to the current pattern. Applied in a situation with many different patterns it can make harmful disturbance of the weights learned about other patterns. In such a situation, we would expect that a slower adjustment of the knowledge combined with more reinforcement cycles a more safe teaching strategy.

2.3.5 Reinforcement cycles, \( C \)

Like human learning, the artificial network's ability to identify a pattern is assumed to improve with reinforcement. We would expect that the ability of a network to correctly recognize a pattern increases with the number of times the pattern has been exposed to the network. However, improvement by repetition would be expected to approach asymptotically a limit, above which no further improvement can be expected.

9.3.6 Reinforcement dispersion, \( R \)

A related question is how reinforcement should be organized. We consider two alternative ways in which the reinforcement can be organized, and denote these as dispersed and concentrated reinforcement. In dispersed reinforcement, the different patterns are exposed to the network one by one in some sequence which is repeated in a prescribed number of cycles. In concentrated reinforcement, each pattern is exposed repetitionally to the network a prescribed number of times before the next pattern.

With a small number of simple patterns, we expect that concentrated reinforcement would be a wise strategy. However, with an increasing number of patterns, concentrated reinforcement might result in destruction of the knowledge of the first patterns before the last were learned.

3. Experimental design

3.1 Overall design

To study teaching strategies, a set of simulation experiments were carried out. Each experiment consisted of two steps, training and testing. Each training step was designed with a specific training set size, ordering of the patterns, adjustment rate, reinforcement cycle, and reinforcement organization. During this step the mean square error was also carried out. The second step of the experiment was performance tests of the network. In this step each pattern of the training set was recalled one by one to let the ANN compute the output vector, and the percentage of correctly identified patterns in the set was computed.

Two evaluations metrics were thus computed and used:

1) MSE of the training set after learning,
2) PCT of correctly identified patterns in recall from the set.

The mean square error is probably the more general indicator of how the network will work within the domain of interest. Decreasing MSE from one experiment to another indicates improving performance. The percentage of correctly identified patterns is more easily understood and directly interpretable measure. Increasing PCT from one experiment to another indicates superior performance in the second experiment.

NNW 1/91. 46—51
3.2 Training sets

The training sets of patterns used were obtained by random generation from a probability distribution [NORDBOTTEN 1990a]. In this distribution, which we assume represent our real world domain of interest, the different target vectors were assigned probabilities, and for each target vector, the element of the pattern vector was assigned a conditional probability. The generation process produced pattern identifiers according to their assumed frequency of occurrence. For each generated identifier, an associated pattern was generated with errors or noise included in accordance with the assumed error probability.

The mapping between pattern vectors and target vectors is therefore many-to-many. Different input vectors can be associated to the same target vector and vice versa. The latter is obviously the more serious for pattern identification.

The size of the pattern and target vectors were both 100 elements. The training sets have been used also in several other experiments and are described in detail in other papers [NORDBOTTEN 1989, 1990b, 1991].

Five statistically independent training sets were generated. Each set was replicated and then sorted by increasing complexity.

The results of the 22 experiments carried out. are summarized in table 9.1 to Table 9.5 attached at the end of this paper. The limitation of the experiments must be emphasized and clearly understood.

The topology of the network studied and the learning algorithm applied are only one pair out of many possible which might have been applied for the same purpose. Another pair may have given quite different results.

The training sets used are composed of synthetic patterns each classified in a simple target pattern. Even though the patterns may satisfy certain conditions for representability of a wide class of real world problems, the patterns cannot be claimed in any way to be universally representative. As representations of images, they have a low degree of resolution and complexity.

3.4 Implementation

The experiments were programmed in PASCAL and C, and the simulations carried out on an IBM AIX PS/2 and an IBM RS/6000 computer.

4. Discussion

4.1 Restrictions

The results of the 22 experiments carried out, are shown in Fig. 2 confirm the expected relations between performance indicators and size of training set. MSE has a increasing value while the PCT on the other hand has a decreasing value when the training set is increased. Bearing in mind the lo-
arithmetic scale used for presenting the size, the covari-

tion between the two performance indicators and

the size of the training set becomes less significant

when the set size is increased.

If we study Fig. 3 in which the horizontal axis of the

former figure is exchanged with the number of differ-

et pattern identifiers in each set from Table 1, quite

similar results are obtained, but the relations between

the performance indicators and the number of ident-

ifiers are linear.

The conclusion we may draw is that the ability of

an ANN with a given topology to learn and subse-

quently make correct identifications within a certain

domain decreases by the number of different pattern

identifiers existing in the domain. If the number of

training patterns is increased, more versions of each

patterns are exposed to the network. The network

learning will continue, but the improvement will be

decreasing and approach a state in which no further

improvement can be expected.

Table 2 gives the results of the simulations which

were carried out to evaluate the impact of ordering the

patterns before they are introduced to the ANN.

There is, however, no indication in the results of our

experiments that an ordering of the training set has

any serious impact on the learning results.

4.3 Adjustment rate

Table 3 presents the results as to the impact of the

adjustment rate on the performance indicators. The

experiments carried out were based on the training set

R800 with Dispersed reinforcement in 15 cycles. The

relationship between MSE and the value of the adjust-

ment rate is quite clear. The MSE value increases by

increasing adjustment rate value as illustrated in

Fig. 4. The interpretation must be that in our domain

of interest the adjustment rate should be given a small

value.

The performance expressed by PCT is less obvious.

It seems to indicate that the performance curve has a

minimum for an adjustment value in the interval

0.5—0.7. This may, however, be the result of random

variations in the training sets. Still, there is no indica-

tion that a higher adjustment value should be chosen

in preference for a low valued adjustment value. This

can be interpreted as support for the assumption that

a fast adjustment in a ANN may destroy previously

learned knowledge.

4.4 Reinforcement

Table 4 and Fig. 5 give the results of the reinforce-

ment cycle investigation. The experiments were all

based on the R800 training set and used an adjust-

ment rate of 0.1. The results support the assumption

of significant impact from reinforcement.

As was expected the performance improved rapidly

by number of reinforcement cycles. It is interesting to

note that already after 5 cycles the performance ob-
tained was relatively high and further gains in per-
formance by increasing the number of cycles up to 35

were not great.

Table 5 shows the results from experiments with dis-

persed and concentrated reinforcement in 15 cycles.

The experiments were based on the R800 set and the

adjustment rate was 0.1. A comparison between the
two reinforcement strategies indicates that in all ex-
periments carried out in this investigation, dispersed
reinforcement is the superior strategy. Mixed strate-
gies not investigated may, however, give better results
than the pure dispersed strategy.

For applications in which the cost of making an ad-
ditional reinforcement cycle is significant, it is a good
reason for considering this cost with the gain in per-
formance.

4.5 Further questions

This paper is one of a series reports on of empirical
studies in artificial intelligence and neural networks
using simulation. The present investigation was besed
on a certain set of assumptions about the functional
characteristics of the neurons included. Simulations
with non-linear activation functions will also be inves-
tigated and compared with the results presented in the present paper.

As pointed out in the introduction to this section, the practical value of the results from this investigation depends on how representative the patterns we have used are for the applications. If the patterns are not representative, are the results still valid? One way to proceed, which we plan to follow, will be to repeat the present simulations with more complex patterns than used in the present investigation and make comparisons between results from the different investigations.

The experiments and evaluations will also be extended to multi layer networks and learning algorithms for networks with hidden layers of neurons.

5. Conclusions

The investigation carried out indicates that the success of learning in artificial neural networks is sensitive to the teaching strategy applied. In particular, success depends on the number of patterns to be distinguished, the adjustment rate and the number of reinforcement cycles used.

However, the investigation did not give any results supporting the hypothesis that introduction of a sorted sequence of training examples would give better results than a random sequence of examples. Neither did the results support the hypothesis that a concentrated reinforcement would give better learning results than dispersed reinforcement.

Acknowledgement

This work was carried out as part of my research duties at the University of Bergen. I want to thank my colleagues Associate Professor Joan C. Nordbotten and Research Assistant Atilla E. Gunhan for constructive comments to draft versions of this paper.

References


