# Cascade-Correlation Learning with Optimal Hyperplane Constraints

Mikko Lehtokangas
Tampere University of Technology
Signal Processing Laboratory
P.O.Box 553, FIN-33101 Tampere, Finland
Phone: +358-3-3653881, Fax: +358-3-3653095
email: mikkol@cs.tut.fi

ABSTRACT: The concept of optimal hyperplane has been recently proposed in the context of statistical learning theory. The important property of an optimal hyperplane is that it provides maximum margins to each class to be separated. Obviously, such a decision boundary is expected to yield good generalization. Considering neural network learning techniques, the majority of them do not make use of the optimal hyperplane concept. As a result, in many cases extensive tuning is required to reach good generalization. Also, with some techniques there is a trade-off between learning speed and generalization. One such a technique is the cascade-correlation learning. Its' main advantages are the abilities to learn quickly and to determine the network size. However, recent studies have shown that in many problems the generalization performance of cascade-correlation trained network may not be quite optimal. Moreover, to reach a certain performance level larger network may be required than with other training methods. In this study we describe modifications to the standard cascade-correlation learning that take into account the optimal hyperplane constraints. Experimental results demonstrate that with these modifications substantial performance gains can be obtained compared to the standard cascade-correlation learning. This includes better generalization, smaller network size and faster learning.

KEYWORDS: cascade-correlation learning, optimal hyperplane, classification, neural networks

## INTRODUCTION

Two important factors in neural network modelling are the generalization ability and training time. These are affected directly by the size and topology of a neural network. Too small a network will have difficulties in learning the training samples, while larger than necessary network tend to overfit resulting poor generalization, Geman et al. (1992). Also, a small network requires in general less computation than a larger one. As a result many recent studies have treated the topology of network as a trainable parameter and allow the network to adjust its structure according to the problem at hand, see reviews Kwok and Yeung (1997) and Reed (1993). One widely studied approach for structure learning is the constructive one, Kwok and Yeung (1997). There the training begins with minimal structure (no hidden units), and then more connections, neurons, layers are added to the network according to some predefined rule. One of the most well known constructive neural network technique is the cascade-correlation (CC) learning, Fahlman and Lebiere (1990). Its main advantages are the abilities to learn quickly and to determine the network size. However, recent studies have shown that in many problems the generalization performance of cascade-correlation trained network may not be quite optimal, or to reach a certain performance level larger network may be required than with other training methods, Hwang et al. (1996) and Lehtokangas (to appear).

Recently the concept of optimal hyperplane has been discussed in the context of optimal margin classification, Boser et al. (1992) and Cortes and Vapnik (1995). The aim of the optimal hyperplane learning methods is to maximize the margin between classes to be separated. Hence, hyperplane that provides maximum margin to each class is called the optimal hyperplane. Theoretical properties of optimal hyperplane have been extensively investigated in the context of statistical learning theory, Vapnik (1995). Obviously, with optimal hyperplane the generalization performance is expected to be maximized. Considering standard neural network learning techniques, the majority of them do not make use of the optimal hyperplane concept. As a result, in many cases extensive tuning is required to reach good generalization. Also, with some techniques, like cascade-correlation, there is a trade-off between learning speed and generalization. Obviously, incorporating the optimal hyperplane constraint to standard learning techniques can potentially yield substantial performance improvements. To investigate this potential, in this study we consider implementation of the optimal hyper-

plane constraints into the standard cascade-correlation learning. In practise this means, that several simple modifications need to be done to the standard cascade-correlation. Note that because cascade-correlation learning tends to produce hidden units that saturate, it has been found to be more suitable for classification tasks instead of regression tasks, Hwang et al. (1996). Therefore we shall consider here only classification problems. Experimental results demonstrate, that modifying cascade-correlation according to optimal hyperplane constraints can not only improve generalization performance but also reduce network size and training time.

## CASCADE-CORRELATION NETWORK AND OPTIMAL HYPERPLANE

The cascade-correlation network defines a hyperplane acting as the decision surface as follows

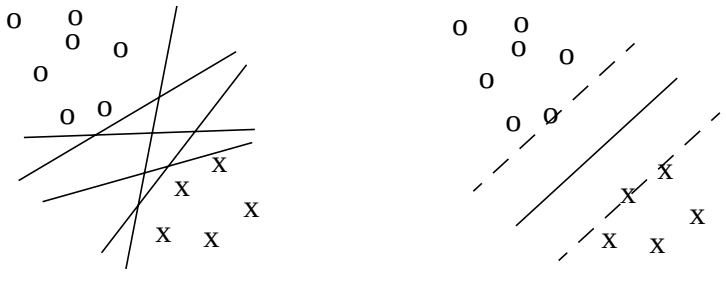
$$\sum_{j=1}^{q} w_j \varphi_j (\boldsymbol{x}, \varphi^j) + b = 0 \tag{1}$$

where $w_j$ is $j$:th adjustable weight and $b$ is an adjustable bias. In addition, $\varphi_j()$ denotes the $j$:th hidden unit function that provides input to the output unit, $\boldsymbol{x}$ is the network input vector and $\varphi^j$ is a vector consisting of outputs of the previous hidden units (previous units are the ones having indices from 1 to $j$-1). Normally the hidden unit function is a sigmoidal one, and as we see from equ. (1) it is a function of network inputs and previous hidden units. In cascade-correlation learning, the hidden units are frozen while training the output weights $w_j$. As a result, cascade-correlation learning has obvious connections to linear classifiers. Therefore, in the following we can consider the optimal hyperplane concept in the context of linear classifiers.

Let us consider a linear classifier and a linearly separable problem. Also, the problem can consist of two classes without loss of generality. In this case the aim is to separate the two classes by a linear classifier induced from available training data. The ultimate goal is to produce a classifier that will generalize well. Consider the example in Fig. 1a. There are many possible linear classifiers that can separate the data, but there is only one that maximizes margin for each class. The margin is the distance between the hyperplane and the nearest data point of each class. An example of hyperplane with maximum margin for each class is presented in Fig. 1b. Intuitively, we would expect this boundary to generalize well as opposed to the other possible boundaries shown in Fig. 1a. Hence, hyperplane that provides maximum margins to each class is called the optimal hyperplane. For theoretical justifications in using the optimal hyperplane as the decision boundary see Vapnik (1995). Now, the equation of a decision surface in the form of a linear hyperplane that does the separation is

$$\sum_{j=1}^{q} w_j \varphi_j + b = 0 \tag{2}$$

where $w_j$ and $b$ are as above. However, now is $\varphi_j$ fixed. For each input training example we have also the desired output response $d$. In two class problem, let the desired response for the first class be +1 and for the other class -1. Now the opti-



a) b)

Figure 1: a) Some of the possible linear hyperplanes that separate two linearly separable classes. b) Optimal hyperplane and respective margins (dashed lines).

mal hyperplane must satisfy the constraints, Haykin (1999),

$$\begin{aligned} \sum_{j=1}^{q} w_j \varphi_j + b \geq +1 \qquad & \text{for } d = +1 \\ \sum_{j=1}^{q} w_j \varphi_j + b \leq -1 \qquad & \text{for } d = -1 \end{aligned} \tag{3}$$

Note that the data points for which the first or second line of equ. (3) is satisfied with the equality sign are called support vectors. These vectors are those data points that lie closest to the decision surface and are therefore the most difficult to classify. As such, among training data they have the largest contribution in determining the location of the decision surface. However, the constraints in equ. (3) are not yet enough for learning optimal hyperplane. Namely with the previous settings it can be shown that the margin of separation between the two classes is, Haykin (1999),

$$\rho = \frac{2}{\|\boldsymbol{w}\|} \tag{4}$$

where $\boldsymbol{w}$ is the weight vector (note that it does not include the bias). Obviously, maximizing the margin of separation between classes is equivalent to minimizing the norm of the weight vector $\boldsymbol{w}$. As a result, the other constraint is that we should try to find a weight vector with as small norm as possible. This can be regarded as a sort regularization. Regularization has been widely studied in the context of network pruning, Reed (1993). However, there the bias terms are usually also included in the regularization scheme unlike here.

The above described optimal hyperplane constraints can be implemented by the use of SVM method, Cortes and Vapnik (1995). However, in cascade-correlation new hidden units are dependent on the previous hidden units, see equ. (1). Because of this dependency we prefer not to use the SVM method here. With SVM scheme the usage of connections between hidden units would lead to combinatorial explosion. Moreover, cascade-correlation learning is constructive in nature but the SVM is not. As a result, in the next section we implement the above optimal hyperplane constraints to the cascade-correlation learning in a different manner compared to the SVM approach.

## CASCADE-CORRELATION WITH OPTIMAL HYPERPLANE CONSTRAINTS

The detailed description about the standard cascade-correlation learning procedure can be found in Fahlman and Lebiere (1990). Therefore, we will concentrate here only in describing the modifications according to the optimal hyperplane constraints. First we should recall that the cascade-correlation learning has two distinct phases. These are the new hidden unit training phase and the output unit re-training phase. In the following modifications for both of the phases will be described.

Let us first consider the output unit re-training phase. There the goal is to train the output weights $w_j$ and bias $b$ to minimize the mean squared output error while keeping the hidden units fixed. This correspond exactly the training of a linear classifier for which the decision surface in the form of a linear hyperplane was given in equ. (2). Obviously, at this phase we can utilize both of the optimal hyperplane constraints described in the previous section. In fact, we could use here the SVM technique, Cortes and Vapnik (1995), for parameter estimation. However, since in the new hidden unit training phase gradient based training methods are used, we prefer to use gradient methods also in the output unit re-training phase. This is because using similar gradient routines in both training phase enables simple implementation. Moreover, considering possible adaptive extensions the gradient approach is more viable. Having made this choice it is rather simple to include the optimal hyperplane constraints to the gradient training. Let us first consider the constraint given in equ. (3). We can easily see, that for $d\varepsilon \leq 0$ ($d$ is the desired output $\{-1;+1\}$ and $\varepsilon = d - y$ is the output error, where $y$ is the network output) equ. (3) is satisfied and for $d\varepsilon > 0$ equ. (3) is not satisfied. Now, in optimal hyperplane perspective, at each iteration only those output errors are significant for which equ. (3) is not satisfied. On the other hand, those errors that satisfy equ. (3) are irrelevant and the corresponding training examples can be regarded as being correctly classified. Therefore, at each iteration all the output errors for which $d\varepsilon \leq 0$ can be set to zero. As a result the goal in the modified output unit re-training phase is to minimize the mean squared output error so that at each iteration the original non-zero error values are used only for those training examples for which $d\varepsilon > 0$. For all the other examples the error is zero. This modification is clearly very simple to implement. The other constraint requires that the norm of the weights $w_j$ should be as small as possible. This can be taken into account by modifying the cost function to be

$$C = MSE + \alpha \sum_j w_j^2 \quad (5)$$

where *MSE* is the usual mean square error function and $\alpha$ is a positive constant. Obviously, equ. (5) implements simple regularization to the output weights (excluding the bias) in which $\alpha$ controls the degree of regularization. Overall, we can conclude that the modifications for the output unit re-training phase are relatively simple to implement.

Considering the new hidden unit training phase, there the goal is to maximize the absolute covariance between the new hidden unit output $\varphi_j()$ and the network output error $\varepsilon$. In this learning phase we can only utilize the constraint given in equ. (3). Similarly as above, it can be implemented by examining the relevance of the output errors. That is, all the output errors for which $d\varepsilon \leq 0$ are set to zero. As a result the goal in the modified new hidden unit training phase is to maximize the absolute covariance between the new hidden unit output and the network output error so that the original non-zero error values are used only for those training examples for which $d\varepsilon > 0$. For all the other examples the error is zero. Also this modification is very simple to implement. Therefore we can say that the cascade-correlation learning with optimal hyperplane constraints effectively retains the simplicity of the original cascade-correlation. However, as the experiments in the next section demonstrate the described modifications yield substantial changes in the learning performance.

## EXPERIMENTS

In this section the performance of the cascade-correlation learning with optimal hyperplane constraints is empirically investigated. The performance of standard cascade-correlation learning, Fahlman and Lebiere (1990), is also presented for comparison purposes. We have used two benchmark problems. The first problem is the two spirals problem, Fahlman and Lebiere (1990), where the goal is to separate two interlocked spirals. This problem has been found relatively hard to solve with sigmoidal feedforward neural networks. The second problem is the channel equalization problem described in Kantsila et al. (to appear). There the goal is to equalize a burst of bits transmitted through a fixed communication channel having 20db signal-to-noise ratio. It should be emphasized that the two-spirals problem demonstrates only the convergence while the equalization problem demonstrates also generalization performance. In all the experiments the RPROP, Riedmiller (1994), optimization procedure was used for the mean square error minimization. The regularization constant $\alpha$ in equ. (5) was set to be 0.0001. As suggested in Fahlman and Lebiere (1990) we used 8 candidates during the new unit addition phase. Finally, each of the simulations were repeated twenty times.

The results are presented in Figs. 2 and 3. Both problems can be solved with fewer number of hidden units by utilizing the optimal hyperplane constraints. This speeds up the training about two times in both cases. Moreover, the equalization problem demonstrates also improved generalization performance. In some trials the generalization error was very close to zero. We conclude that the cascade-correlation learning with optimal hyperplane constraints can yield substantial improvements compared to the standard approach. It therefore provides a potentially very competitive constructive algorithm for classification type of problems.

## CONCLUSIONS

Cascade-correlation learning was considered. We first pointed out that in the standard method the resulted network size and generalization performance are not always quite optimal. Then we briefly introduced the optimal hyperplane scheme developed recently in statistical learning theory. The relation of cascade-correlation learning to optimal hyperplane scheme was also illustrated. Based on the optimal hyperplane constraints we described several modifications to the standard cascade-correlation learning. The modified scheme is usable only in classification problems, which is the main application area of cascade-correlation learning. Experimental results demonstrated, that cascade-correlation learning with optimal hyperplane constraints can not only reduce the network size and training time but also improve the generalization performance. Therefore we believe that the described (theoretically justified) modifications are useful and can substantially improve the cascade-correlation learning scheme.

## ACKNOWLEDGEMENTS

This work has been supported by the Academy of Finland.

## REFERENCES


Boser B., Guyon I. and Vapnik V., 1992, “A training algorithm for optimal margin classifiers,” Proceedings of 5th Annual Workshop on Computational Learning Theory, pp. 144-152.

Cortes C. and Vapnik V., 1995, “Support vector networks,” Machine Learning, vol. 20, pp. 273-297.

Fahlman S. and Lebiere C., 1990, "The cascade-correlation learning architecture," In D. Touretzky (Ed.), Advances in Neural Information Processing Systems 2, San Mateo, CA, Morgan Kaufman, pp. 524-532.

Geman S., Bienenstock E. and Doursat R., 1992, “Neural networks and the bias/variance dilemma,” Neural Computation, vol. 4, pp. 1-58.

Haykin S., 1999, Neural networks: a comprehensive foundation, 2nd edition, Prentice Hall, New Jersey.

Hwang J., You S., Lay S. and Jou I., 1996, “The cascade-correlation learning: a projection pursuit learning perspective,” IEEE Transactions on Neural Networks, vol. 7, no. 2, pp. 278-289.

Kantsila A., Lehtokangas M. and Saarinen J., to appear, “Burst adaptive equalization of binary data,” Journal of Intelligent Systems.

Kwok T. and Yeung D., 1997, “Constructive algorithms for structure learning in feedforward neural networks for regression problems,” IEEE Transactions on Neural Networks, vol. 8, no. 3, pp. 630-645.

Lehtokangas M., to appear, “Modelling with constructive backpropagation,” Neural Networks.

Reed R., 1993, “Pruning algorithms - a survey,” IEEE Transactions on Neural Networks, vol. 4, no. 5, pp. 740-747.

Riedmiller M., 1994, “Advanced supervised learning in multi-layer perceptrons - from backpropagation to adaptive learning algorithms,” International Journal on Computer Standards and Interfaces, Special Issue on Neural Networks, vol. 5.

Vapnik V., 1995, The nature of statistical learning theory, Springer-Verlag, New York.

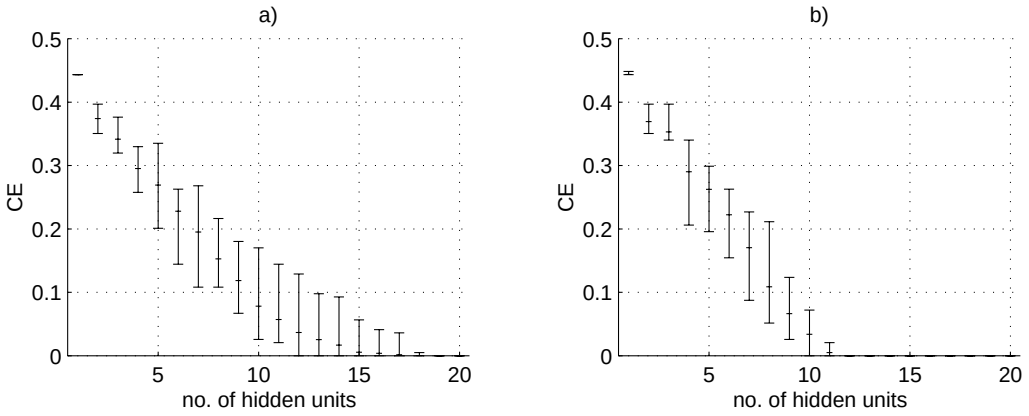




Figure 2: Classification error as a function of number of hidden units for the two-spirals problem; a) standard cascade-correlation, and b) cascade-correlation with modifications. The smaller horizontal line in the middle is located at the average of the repetitions, while the whiskers show the total range of values.

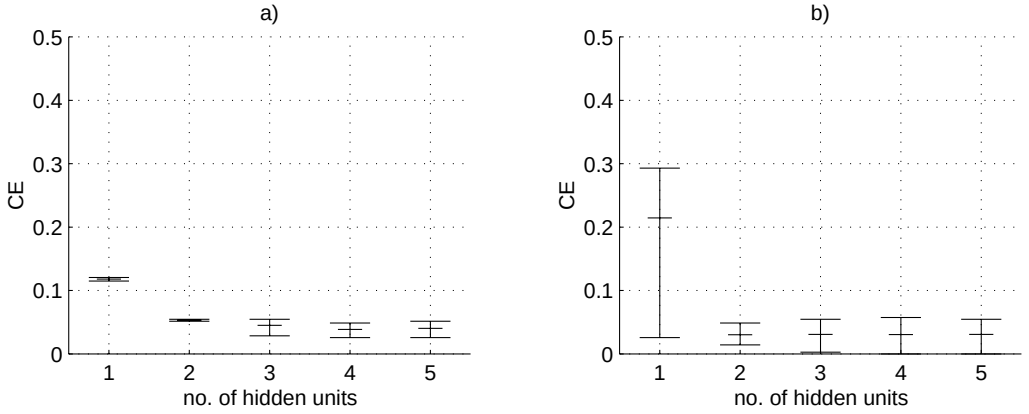



Figure 3: Classification error as a function of number of hidden units for the *independent test set* of the equalization problem; a) standard cascade-correlation, and b) cascade-correlation with modifications. The smaller horizontal line in the middle is located at the average of the repetitions, while the whiskers show the total range of values.