Optimization of Neural Network Topology and Information Content Using Boltzmann Methods

O. M. Omidvar, University of the District of Columbia
Washington, DC 20008
C L. Wilson, National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

Reduction in the size and complexity of neural network applications are the driving force behind the current research in network optimization. Most of the known optimization methods heavily rely on weight sharing concepts for pattern separation and recognition. The method used in this research focuses on network topology and information content for optimization. We have studied the change in the network topology and its effects on information content dynamically during the optimization of the network. The changes in the network topology were achieved by altering the number of weights. The primary optimization was scaled conjugate gradient and the secondary method of optimization a Boltzmann method. Our findings demonstrate that for a difficult character recognition problem the number of weights in a fully connected network can be reduced by 90.3% with a temperature of 0.55 while achieving training and testing of identical accuracies.

1. Introduction

The size and the complexity of neural network applications has grown rapidly. The search for small networks with large information content and generalization capability is ongoing. Most of the optimization strategies are a trade-off between error and network complexity. The known optimization schemes[1,2,3] have used this trade-off to minimize the cost function. Among various complexity measures, Vapnic-Chervonenkis (VC) dimensionality [4], concentrates on information content and distribution of information in the network. The Boltzmann machine has also been used as a statistical method for combinatorial optimization and for the design of learning algorithms[5,6]. Current weight removal strategies risk pruning connections which are critical to the performance of the network. The strategy used in this research is to remove the weights using Boltzmann criteria during the training process. We also use information content as a measure of network complexity for evaluation of the generalization performance of the network.

2. Pruning via Boltzmann Methods

In this paper a fully connected network is optimized using the Scaled Conjugate Gradient method (SCG) developed by Moller [7] and modified by Blue and Grother. The SCG method is used as a starting network for the Boltzmann weight pruning algorithm. The network has an input layer with thirty two input nodes, a variable size hidden layer with sixteen, thirty two or sixty four nodes and an output layer with ten nodes. This is a fully connected network. The pruning was carried out by selecting a normalized temperature, T and removing weights based on a probability of removal:
2 PRUNING VIA BOLTZMANN METHODS

\[ P_i = (1 - \exp(|w_i|/T)) \]

The values of \( P_i \) are compared to the set of uniformly distributed random numbers, \( R_i \), on the interval \([0, 1]\). If the probability \( P_i \) is greater than \( R_i \) then the weight is set to zero. The process is carried out for each iteration of the SCG optimization process and is dynamic. If a weight is removed it may subsequently be restored by the SCG algorithm, the restored weight may survive if it has sufficient magnitude in subsequent iterations.

The dynamic effect of this is shown in figure 1 for five temperatures between 0.1 and 0.5 at 0.1 intervals. As the temperature increases the number of weights removed initially increases but the effect of later iterations of optimization and pruning decreases. The number of weights in the initial network was 1389. At all temperatures the initial iterations are very effective in reducing the weights.

![Figure 1: Weights removed as a function of iteration and temperature](image)

Figure 1: Weights removed as a function of iteration and temperature

The effect of the number of hidden nodes can be seen in figures 2, 3 and 4. Figure 2 shows the effect on the network with 32 hidden nodes used in figure 1. As the temperature is increased the accuracy of the network for recognition decreases slowly for temperatures up to 0.4. As the temperature approaches 0.5 the rate of weight removal shown in figure 1 slows and the rate of accuracy decay accelerates. The two curves plotted are the training set and testing set accuracy of the network. The training set accuracy is initially greater than the testing accuracy. At a critical temperature, \( T_c \), the testing and training accuracies are identical. In figure 2, at the critical temperature of 0.581 chaotic behavior sets in due to a critical effect from removing each weight.

The behavior of the 32-64-10 network in figure 4 is similar to the 32-32-10 network. The 32-16-10 network shows an increase in temperature, \( T_c \), and a decrease in accuracy at \( T_c \). This increase in \( T_c \) is caused by the reduced set of possible pruned configurations in the 32-16-10 network.
Figure 2: Change in testing and training accuracy as a function of temperature for a 32-32-10 network

Figure 3: Change in testing and training accuracy as a function of temperature for a 32-16-10 network
Figure 4: Change in testing and training accuracy as a function of temperature for a 32-64-10 network.

Figure 5: Weight distribution below $T_c$. 

IV-597
3. Weight Reduction and Information Content

The effect on the information content of the network can be evaluated by examining the distribution of weights in the network as a function of temperature. Figure 5 shows the distribution of the absolute value of the weights at a temperature near, but below, $T_c$. Figure 6 shows the distribution of the absolute value of the weights at a temperature near, but above, $T_c$.

These distributions illustrate the mechanism involved in the collapse of testing and training accuracy near $T_c$. The cause is the large increase in weights near zero. In a given training cycle some weights are removed. If these weights are redundant they will be compensated for by other weights in the network. If these weights are critical they will be restored by the SCG optimization. The peak in the distribution near zero in both figures 5 and 6 is caused by this process. At $T_c$ the SCG creation process is balanced by the Boltzmann pruning.

The effect of the near zero weights is more important when viewed as information content. The VC dimension and the information content are both approximately $\sum (\log_2(|w_i|) + 1)$. When large numbers of near-zero weights exist, their contribution to the sum dominates the network information. Under these conditions the network is dominated by recently created weights which have not been optimized by SCG iterations. This lowers network accuracy without reducing VC dimension.

4. Conclusions

A method of network optimization has been developed which reduces the number of weights required for accurate character recognition between 80% to 90%. The method is based on achieving equilibrium between the information in the training set and the number of network weights. These reductions allow both smaller training sets and smaller classification networks to be used.
References


