Structure Minimization using Impact Factor in Neural Networks

Kap-Ho Seo, Jae-Su Song, and Ju-Jang Lee

Dept. of Electrical Engineering and Computer Science,
Korea Advanced Institute of Science and Technology,
373-1, Kusong-dong, Yusong-gu, Taejon, 305-701, Korea
Tel: 82-42-859-3432/5432; Fax: 82-42-869-3410;
E-mail: khseo@odyssey.kaist.ac.kr jssong@odyssey.kaist.ac.kr jjlee@ee.kaist.ac.kr

Abstract

The problem of determining the proper size of an neural network is recognized to be crucial, especially for its practical implications in such important issues as learning and generalization. Unfortunately, it usually is not obvious what size is best; a system that is too small will not be able to learn the data while one that is just big enough may learn very slowly and be very sensitive to initial conditions and learning parameters. One popular technique is commonly known as pruning and consists of training a larger than necessary network and then removing unnecessary weights/nodes. In this paper, a new pruning method is developed, based on the penalty-term methods. This method makes the neural networks good for the generalization and reduces the retraining time after pruning weights/nodes.

1. Introduction

Despite many advances, for neural networks to find general applicability in real-world problems, several questions must still be answered. One such open question involves determining the most appropriate network size for solving a specific task. There is dilemma that stems from the fact that both large and small networks exhibit a number of advantages. When a network has too many free parameters, not only is learning fast, but local minima are more easily avoided. Large networks can also form as complex decision regions as the problem requires and should exhibit a certain degree of fault tolerance under damage conditions. On the other hand, both theory and experience show that networks with few free parameters exhibit a better generalization performance, and this is explained by recalling the analogy between neural network learning and curve fitting. Moreover, knowledge embedded in small trained networks is easier to interpret and thus the extraction of simple rules can hopefully be facilitated.

To solve the problem of choosing the right size network, two different incremental approaches are often pursued. The first starts with a small initial network and gradually adds new hidden units or layers until learning takes place. Well-known examples of such growing algorithms are cascade correlation and others. The second, referred to as pruning, starts with a large network and excises unnecessary weights and/or units. This approach combines the advantages of training large networks (i.e., learning speed and avoidance of local minima) and those of running small ones (i.e., improved generalization). However it requires advance knowledge of what size is "large" for the problem at hand, but this is not a serious concern as upper bounds on the number of hidden units have been established. Among pruning algorithms there are methods that reduce the excess weights/nodes during the training process, such as penalty term methods and the gain competition technique.

Various techniques such as optimal brain surgeon (OBS) and optimal brain damage (OBD) have been proposed in literature to prune a fully connected feedforward artificial neural network (FANN). These techniques, however, require considerable additional computation as they require calculation of the Hessian matrix of the system. These post-