Neural Networks 33 (2012) 127–135

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Computational properties and convergence analysis of BPNN for cyclic and almost cyclic learning with penalty[☆]

Jian Wang[a,b], Wei Wu[a], Jacek M. Zurada[b,*]

[a] School of Mathematical Sciences, Dalian University of Technology, Dalian, 116024, PR China

[b] Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40292, USA

ARTICLE INFO

Article history:
Received 8 September 2011
Received in revised form 25 April 2012
Accepted 27 April 2012

Keywords:
Weight decay
Backpropagation
Cyclic
Almost cyclic
Convergence

ABSTRACT

Weight decay method as one of classical complexity regularization methods is simple and appears to work well in some applications for backpropagation neural networks (BPNN). This paper shows results for the weak and strong convergence for cyclic and almost cyclic learning BPNN with penalty term (CBP-P and ACBP-P). The convergence is guaranteed under certain relaxed conditions for activation functions, learning rate and under the assumption for the stationary set of error function. Furthermore, the boundedness of the weights in the training procedure is obtained in a simple and clear way. Numerical simulations are implemented to support our theoretical results and demonstrate that ACBP-P has better performance than CBP-P on both convergence speed and generalization ability.

1. Introduction

A multilayer perceptron network trained with a highly popular algorithm known as the error backpropagation (BP) has been successfully applied to solve some difficult and diverse problems (Haykin, 1999; Rumelhart et al., 1986). This algorithm, based on the error-correction learning rule can be viewed as a generalization of the least-mean-square (LMS) algorithm. There are two main modes to implement it: batch learning, in which optimization is carried out with respect to all training samples simultaneously, and incremental learning, where it follows the presentation of each training sample (Saad, 1998).

There are three different incremental BP learning strategies: online learning, cyclic learning, and almost cyclic learning (Heskes & Wiegerinck, 1996). Incremental learning strategies require less storage capacity than batch mode learning. Due to the random presentation order of the training samples, incremental learning implementing the instant gradient of the error function is a stochastic process, whereas batch mode learning corresponds to the standard gradient descent method and is deterministic (Heskes & Wiegerinck, 1996; Nakama, 2009; Wilson & Martinez, 2003).

It is well known that the general drawbacks of gradient-based BPNN training methods are their more likely divergence and weak generalization. In real-world problems, the BP method is usually prone to require the use of highly structured networks of a rather large size (Haykin, 1999). Thus, it is requisite to reach an appropriate tradeoff between reliability of the training and the goodness of the model. Knowing that the network design is statistical in nature, the tradeoff can be achieved by minimizing the overall risk with regularization theory (Tikhonov, 1963). A general setting is to add an extra regularization term which is called the *penalty term* for BPNN (Haykin, 1999).

There are three classical different penalty terms for BPNN: *weight decay* (Hinton, 1989), *weight elimination* (Weigend, Rumelhart & Huberman, 1991) and *approximate smoother* (Moody & Rognvaldsson, 1997). In the weight decay procedure, the penalty term is stated as the squared norm of the weights in the BPNN (Hinton, 1989; Saito & Nakano, 2000). All the weights in the networks are treated equally. Some of the weights are forced to take values close to zero, while other weights maintain reasonably large values, and consequently improve the generalization of BPNN (Haykin, 1999). In the weight elimination procedure, the complexity penalty represents the complexity of the network as function of the weight magnitudes relative to a pre-assigned parameter (Reed, 1993). The approximate smoother approach is proposed in Moody and Rognvaldsson (1997) for BPNN with a hidden layer and a single output neuron. This method appears to be more accurate than weight decay or weight elimination for the complexity regularization of BPNN. However, it is much more computationally complex than its counterparts (Haykin, 1999).

Below we discuss the convergence of BPNN with penalty term from a mathematical point of view. Insofar as the satisfying

[☆] Project supported by the National Natural Science Foundation of China (No. 11171367) and China Scholarship Council (CSC).

[*] Corresponding author.
E-mail addresses: jmzura02@louisville.edu, jacek.zurada@louisville.edu (J.M. Zurada).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.04.013

performance in weight decay method, there are quantitative studies of the convergence property with different BP learning strategies (Shao, Wu & Li, 2005; Shao & Wu, 2007; Shao & Zheng, 2011; Wu, Shao & Li, 2006; Zhang, Wu & Yao, 2007; Zhang, Wu, Liu & Yao, 2009; Zhang & Wu, 2009).

For batch mode learning, the weak convergence and monotonicity are proved as a special case for the typical gradient descent method of optimization theory. A highlight in Wu et al. (2006) is that the boundedness of the weights between input and hidden layers are guaranteed. As an extension, the boundedness of the total weights in the BP feedforward neural networks based on batch learning has been proved in Zhang et al. (2007). For online learning, Zhang and Wu (2009) focuses on the linear output of BPNN, while an extension that the activation function satisfies twice continuously differentiable is proposed in Zhang et al. (2009). The main contribution of these two papers is to theoretically prove the boundedness of the weights and an almost sure convergence of the approach to the zero set of the gradient of the error function.

Assuming the training samples are supplied in random order in each cycle (almost cyclic), the monotonicity and weak convergence of the almost cyclic learning for BPNN with penalty term (ACBP-P) are guaranteed based on restricted conditions for activation functions and learning rates (Shao et al., 2005). Additionally, the results in Shao et al. (2005) are valid for BPNN without a hidden layer. On the basis of cyclic learning BPNN with penalty term (CBP-P), the convergence results are proved in Shao and Wu (2007) and Shao and Zheng (2011). A momentum term to speed up the training procedure is considered as well in Shao and Zheng (2011).

Within the framework of BPNN with cyclic and almost-cyclic learning, the latest convergence results concentrate on the regular BPNN (Wu, Wan, Cheng & Li, 2011) and on BPNN with momentum term (Wang, Yang, & Wu, 2011) under much relaxed conditions such as activation functions and learning rates. The training method of BPNN based on the common gradient descent without any additional term is considered in Wu et al. (2011). Furthermore, the strong convergence result was first proved which allows the stationary points of error function to be uncountable somehow. In Wang, Yang et al. (2011), weak and strong convergence results have been obtained for BPNN with a momentum term which performs much better than regular BPNN. None of the earlier studies focused on convergence results for similar learning modes with penalty term based on relaxed conditions. This paper attempts to fill this gap.

The aim of this paper is to present a comprehensive study for CBP-P and ACBP-P of weak and strong convergence with the identical relaxed training conditions (Wang, Yang et al., 2011; Wu et al., 2011), indicating that the gradient of the error function goes to zero and the weight sequence goes to a fixed point, respectively. In comparison to the convergence results which consider the CBP-P and ACBP-P (Shao et al., 2005; Shao & Wu, 2007; Shao & Zheng, 2011), quite simple and general conditions are formulated below for the learning rate and the activation functions to guarantee the convergence. The main points and novel contributions of this paper are as follows:

(1) The derivatives g', f' of the activation functions g, f are Lipschitz continuous on $\mathbb{R}$. This improves the corresponding conditions in Shao et al. (2005); Shao and Wu (2007) and Shao and Zheng (2011), which requires the boundedness of the second derivatives g'', f''.

From a mathematical point of view, we mention that different analytical tools are employed in Shao et al. (2005); Shao and Wu (2007), Shao and Zheng (2011) and Wu et al. (2006) and this study for the convergence analysis. The differential Taylor expansion in Shao et al. (2005); Shao and Wu (2007), Shao and Zheng (2011) and Wu et al. (2006) which requires the boundedness of the second derivative of the activation function g, is considered, while in this paper, we discuss the integral Taylor expansion and hence require the Lipschitz continuity of g', f' on $\mathbb{R}$ (Xu, Zhang & Jing, 2009).

(2) The condition on the learning rate in this paper is extended to a more general case: $\sum_{m=0}^{\infty} \eta_m = \infty; \sum_{m=0}^{\infty} \eta_m^2 \langle \infty, (\eta_m \rangle 0)$, which is identical to those in Wu et al. (2011) for cyclic learning without penalty.

Learning rate is an important criterion in the convergence analysis of BPNN. The convergence results in Shao and Zheng (2011) for cyclic learning with penalty and momentum term focus on no hidden layer feedforward neural networks, and require $\frac{1}{\eta_{k+1}} = \frac{1}{\eta_k} + \beta, (k \in \mathbb{N}, \beta > 0)$, where η_k is the learning rate of the k-th training cycle. Basically, this condition is equivalent to $\eta_k = O\left(\frac{1}{k}\right)$. It is easy to see that the conditions on the learning rate are more relaxed in this paper than those in Shao et al. (2005); Shao and Wu (2007) and Shao and Zheng (2011).

(3) The restrictive assumptions for the strong convergence in and Shao et al. (2005), Shao and Zheng (2011) and Wu et al. (2006) are relaxed such that the stationary points set of the error function is only required not to contain any interior point.

To obtain the strong convergence result, which means that the weight sequence converges to a fixed point, an extra condition is considered in Shao et al. (2005); Shao and Wu (2007), Shao and Zheng (2011) and Wu et al. (2006): the gradient of the error function has finitely many stationary points. Thus, this additional assumption is a special case in this paper (cf. (A3)).

(4) The deterministic convergence results are valid for ACBP-P as well.

We mention that CBP-P is typically a deterministic iteration procedure in that the updating fashion is deterministic for fixed order of samples. Due to the random order of samples in each training cycle, the experiment shows that ACBP-P behaves numerically better than CBP-P (Shao et al., 2005). In this paper, our convergence results are generalizations of both the results of Shao and Wu (2007), which considers CBP-P, and of the results of Shao et al. (2005) and Shao and Zheng (2011), which considers ACBP-P.

Remark. Considering the batch learning BPNN with penalty term we note that this method corresponds to the standard gradient descent algorithm. The convergence results are valid as well once the differential Taylor expansion in Wu et al. (2006) is replaced by the integral Taylor expansion in this paper. In addition, a simple and clear proof for the boundedness of the weights is presented.

(5) Illustrated experiments have been done to verify the theoretical results of this paper, such as boundedness of the weights, convergence property of BPFNN with penalty term.

Comparing to Wang, Wu, and Zurada (2011), three different simulations have been performed to demonstrate clearly the important properties of BPFNN with penalty term. Furthermore, one of the classification simulations shows that ACBP-P performs generally much better than CBP-P.

The rest of this paper is organized as follows: Section 2 introduces the two weights updating algorithms: CBP-P and ACBP-P. The main convergence results are presented in Section 3. The performance of the presented two algorithms are reported and discussed in Section 4. The detailed proofs of the main results are stated as Appendix for interested readers.

2. Algorithm description

Denote the numbers of neurons of the input, hidden and output layers of BPNN are p, n and 1, respectively. Suppose that the