Neurocomputing 000 (2017) 1–9

Contents lists available at ScienceDirect

Neurocomputing

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

A novel conjugate gradient method with generalized Armijo search for efficient training of feedforward neural networks☆

Jian Wang[a], Bingjie Zhang[a], Zhanquan Sun[b,*], Wenxue Hao[c], Qingying Sun[a]

[a] College of Science, China University of Petroleum, Qingdao 266580, China
[b] Shandong Computer Science Center (National Supercomputer Center in Jinan), Shandong Provincial Key Laboratory of Computer Networks, Jinan, Shandong 250014, China
[c] Zhiyuan Middle School, No. 888 Qianwangang Road, Economic & Technological Development Zone, Qingdao 266510, China

ARTICLE INFO

Article history:
Received 30 April 2017
Revised 17 July 2017
Accepted 22 August 2017
Available online xxx

Communicated by Dr Ding Wang

Keywords:
Feedforward neural networks
Backpropagation
Deterministic convergence
Conjugate gradient method
Generalized Armijo search

ABSTRACT

In this paper, a novel multilayer backpropagation (BP) neural network model is proposed based on conjugate gradient (CG) method with generalized Armijo search. The presented algorithm requires low memory and performs fast convergent speed in practical applications. One reason is that the constructed conjugate direction guarantees the sufficient descent behavior in minimizing the given objective function. The other stems from the fact that the generalized Armijo method can automatically determine a more suitable learning rate in each training epoch. As a theoretical contribution, two deterministic convergent results, weak and strong convergence, have been detailedly proved under more relaxed assumptions. The weak convergence means that the norm of gradient of the objective function tends to zero. For the strong convergence, it represents that the sequence of weight vectors approaches a fixed point. To support the theoretical results, some illustrated simulations have been done on various benchmark datasets.

1. Introduction

The backpropagation neural networks (BPNNs) have been an attractive topic in many research fields, such as pattern recognition, classification and computational intelligence [1–6]. As a supervised learning technique, the gradient descent method is applied to many instances during training BPNNs [7]. In terms of gradient descent method, an adaptive optimal control method has been put forward for solving the Hamilton–Jacobi–Bellman equation [8]. They rigorously prove the uniform boundedness of the closed-loop partial differential equations by using neural network approximator. For a model-free optimal control problem [9], the gradient descent scheme has also been employed to develop an adaptive optimal controller. And the convergent results of above method are gained with offline and online fed models of data.

We notice that the updating direction based on gradient descent method depends on the negative gradient of the objective function [10]. For accelerating the convergence rates of the gradient descent method, there have been extensive studies [2,11]. However, this method shows oscillatory behavior even with these changes in the training process while meeting steep valleys, which leads to poor efficiency. The essential reason of the poor efficiency is that this method has only first-order convergence.

For the purpose of accelerating the convergent rates, considerable reports have discussed the CG method and Newton method for BP algorithm [12,13]. Compared with the gradient descent method, Newton method is more effective, however it needs to calculate the Hessian matrix and its inverse. As a compromise algorithm between these two methods, the CG method has the quadratic convergence, so it is very efficient, and it is still easy to compute without the second derivatives [14,15].

Due to its special advantage, the CG method has been an interesting research topic. A specific CG method is described in [16] to solve some linear systems whose coefficient matrices are positive definite. For solving massive nonlinear optimization problems, the nonlinear CG method is first proposed in [17]. In terms of the diverse descent directions, there are three typical conjugate gradient methods which were proposed by Fletcher and Reeves (FR) [17], Polak-Ribière-Polyak (PRP) [20] and Hestenes and Stiefel (HS) [16]. To improve the convergent performance, more modified

☆ This work was supported in part by the National Natural Science Foundation of China (No. 61305075), the China Postdoctoral Science Foundation (No. 2012M520624), the Natural Science Foundation of Shandong Province (No. ZR2013FQ004, ZR2013DM015), the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20130133120014) and the Fundamental Research Funds for the Central Universities (No. 13CX05016A, 14CX05042A, 15CX05053A, 15CX08011A).

* Corresponding author.
E-mail addresses: sunzhq@sdas.org (Z. Sun), sunqingying01@163.com (Q. Sun).

http://dx.doi.org/10.1016/j.neucom.2017.08.037
0925-2312/

conjugate gradient methods have been presented through improving the learning step-size and the conjugate directions in [18,19]. To deal with nonsmooth convex minimization problems, a modified PRP conjugate gradient algorithm which combines with a nonmonotonic line search technique is proposed in [21]. By slightly modifying the search direction of the nonmonotone HS method, a variant HS conjugate gradient method is proposed which effectively satisfies the sufficient descent condition independent of any line search technique [22].

For purpose of solving the blind source separation problem, a PRP conjugate gradient method based on cyclic mode has been described in [23]. Its learning rate is obtained by exact line search method. In [24], three PRP updating methods are detailedly studied, which include batch mode conjugate gradient method (BCG), cyclic conjugate gradient method (CCG) and almost cyclic conjugate gradient method (ACCG). The learning rate of BCG is a positive constant in the training process. Unfortunately, the proposed algorithm with fixed learning rate is prone to leading to poor performance. As an improvement, the existing exact line search method contributes to obtain the optimal learning step for each iteration [25]. However, it is more time-consuming due to the strong dependence on many functions and its gradients. Especially when the iteration point is far away form the optimum point, this search method is not effective and reasonable. In order to make up for these defects, the inexact line search methods are presented, including Wolf search method [26], Armijo search method [27] and their improvements. For unconstrained optimization problems, a new three terms CG method with generalized Armijo step size rule is proposed in [19], and it provides a novel technique to prove the global convergence. To our best knowledge, the conjugate gradient network models with generalized Armijo search method is little referred to. We attempt to research a generalized Armijo search method in this paper to obtain a suitable learning rate in each iteration. As an inexact line search method, the generalized Armijo search method has many advantages. It not only can guarantee that the objective function has acceptable decrease, but also can make the eventual formed iterative sequence convergent.

The convergence properties attract wide attention in feedforward neural networks, which provides an significant guarantee for the actual applications. The existing convergent theories of BP algorithm proposed in many papers mostly consider the gradient descent method [28–32]. The convergent property of a fractional-order gradient descent method for BP neural networks with Caputo derivative is discussed in [33] and a learning method for sparse feedforward neural networks with group lasso regularization is researched in [34], respectively. However, the convergence properties of CG method need to be researched urgently. In [24], the weak and strong convergent theories of BCG, CCG and ACCG have been proven, respectively. In previous papers, most techniques were similar when we proved the convergence results for BPNNs.

Inspired by Sun and Liu [19], we propose a novel conjugate gradient method with generalized Armijo search to train a common three-layer BPNNs in this paper. For the presented algorithm, we also focus on its convergent behaviors, that is, the gradient of the error function tends to zero which results in the weak convergent behavior, and the sequence with respect to weights tends to a fixed point for strong convergence. Compared with the existing literature, the theoretical results are reached under more relaxed assumptions by employing a special technique. Specifically, the contributions are shown as follows:

A) In terms of a generalized Armijo search method, a novel conjugate gradient training algorithm for BP networks (BPCGGA) is proposed in this paper. The presented simulations observe the fast and effective performance. To speed up the training procedure, the optimal learning rate of BPCGGA is obtained by employing a generalized Armijo step size rule. A specific conjugate direction coefficient is constructed to guarantee the sufficiently decreasing property of the given error function. According to this descent direction, the monotonicity of the proposed algorithm is detailedly proved in the following section.
B) Compared with the existing literature, two deterministic convergent behaviors of BPCGGA, weak and strong convergence, are reached under some more relaxed assumptions. The restrictive assumptions (cf. [24,35]) which are employed to result in the convergence are weakened in this paper. For the activation functions, their first derivatives satisfying the local Lipschitz continuous condition are the necessary prerequisite to the convergence results in [24,35]. However, in this paper, these assumptions are greatly relaxed, that is, the derivatives of activation functions are confined to be uniformly continuous. By virtue of reduction to absurdity, the bounded assumption of weight sequence which is requisite in [24,35] is discarded during the proof of weak convergence.

The other sections of the paper are following: In Section 2, we describe the updating methods based on BPG, BPCG and BPCGGA. In Section 3, some theoretical results are proposed. Two supporting numerical experiments are shown in Section 4. Section 5 gives the proofs of the convergence results detailedly. Section 6 summarizes the paper.

2. Algorithms

We construct a common three-layer BP neural network model with p input, n hidden and one output neurons, respectively. The training sample set is given as $\left\{\mathbf{x}^j, O^j\right\}_{j=1}^{J} \subset \mathbb{R}^p \times \mathbb{R}$, where $\mathbf{x}^j$ and O^j are the input and ideal output of the jth sample. Denote the connecting weights between the input and hidden layers as $\mathbf{V} = \left(v_{i,j}\right)_{n\times p}$, and $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{ip})^T$ for $i = 1, 2, \ldots, n$. We write the vector $\mathbf{u} = (u_1, u_2, \ldots, u_n)^T \in \mathbb{R}^n$ to be the weights which connects the hidden and output layers. For simplicity, all of the weights among the input, hidden and output layers are combined as a total vector $\mathbf{w} = \left(\mathbf{u}^T, \mathbf{v}_1^T, \ldots, \mathbf{v}_n^T\right)^T \in \mathbb{R}^{n(p+1)}$. Two continuous differential functions, g and $f : \mathbb{R} \to \mathbb{R}$, are set to be the corresponding activations of the hidden and output layers, respectively. For convenience, a vector-valued function is introduced as follows

$$G(\mathbf{z}) = (g(z_1), g(z_2), \ldots, g(z_n))^T, \quad \forall\ \mathbf{z} \in \mathbb{R}^n. \tag{1}$$

The forward procedure during BP training can be depicted with the following process. Given any specified input $\mathbf{x} \in \mathbb{R}^p$, the responses of the hidden layer are activated with $G(\mathbf{Vx})$. Then the final output can be expressed by

$$y = f(\mathbf{u} \cdot G(\mathbf{Vx})). \tag{2}$$

Given a specific weight vector $\mathbf{w}$, the objective function of the network model can be obtained by

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}\sum_{j=1}^{J}(O^j - f(\mathbf{u} \cdot G(\mathbf{V}\mathbf{x}^j)))^2 \\ &= \sum_{j=1}^{J} f_j(\mathbf{u} \cdot G(\mathbf{V}\mathbf{x}^j)), \end{aligned} \tag{3}$$

where $f_j(t) = \frac{1}{2}(O^j - f(t))^2$, $j = 1, 2, \ldots, J$, $t \in \mathbb{R}$. We note that the gradient-based algorithms are very popular in training BP

Please cite this article as: J. Wang et al., A novel conjugate gradient method with generalized Armijo search for efficient training of feedforward neural networks, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.08.037