Applied Mathematics and Computation 266 (2015) 404–413

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

A new class of nonmonotone conjugate gradient training algorithms

Ioannis E. Livieris*, Panagiotis Pintelas

Educational Software Development Laboratory, Department of Mathematics, University of Patras,GR 265-00, Greece

ARTICLE INFO

Keywords:
Artificial neural networks
Conjugate gradient algorithm
Nonmonotone line search
Global convergence

ABSTRACT

In this paper, we propose a new class of conjugate gradient algorithms for training neural networks which is based on a new modified nonmonotone scheme proposed by Shi and Wang (2011). The utilization of a nonmonotone strategy enables the training algorithm to overcome the case where the sequence of iterates runs into the bottom of a curved narrow valley, a common occurrence in neural network training process. Our proposed class of methods ensures sufficient descent, avoiding thereby the usual inefficient restarts and it is globally convergent under mild conditions. Our experimental results provide evidence that the proposed nonmonotone conjugate gradient training methods are efficient, outperforming classical methods, proving more stable, efficient and reliable learning.

1. Introduction

Artificial neural networks are parallel computational models comprised of densely interconnected, adaptive processing units, characterized by an inherent propensity for learning from experience and also discovering new knowledge. Their excellent capability of self-learning and self-adapting have established them as vital components of many systems and are considered as a powerful tool for pattern classification [4]. Moreover, they have been successfully applied in many applications of artificial intelligence [2–4,23–25,42] and are often found to be more efficient and accurate than other classification techniques [17].

The standard problem of *training* a neural network is to iteratively adjust the weights, in order to globally minimize a measure of difference between the actual output of the network and the desired output for all examples of the training set [36]. More mathematically, the training process can be formulated as the minimization of an error function $E(w)$ that depends on the connection weights w of the network. Conjugate gradient methods are probably the most famous iterative methods for efficiently training neural networks due to their simplicity, numerical efficiency and their very low memory requirements. These methods generate a sequence of weights $\{w_k\}$ using the iterative formula

$$w_{k+1} = w_k + \eta_k d_k, \quad k = 0, 1, \ldots, \tag{1.1}$$

where k is the current iteration usually called *epoch*, $w_0 \in \mathbb{R}^n$ is a given initial point, $\eta_k > 0$ is the learning rate and d_k is a descent search direction defined by

$$d_k = \begin{cases} -g_0, & \text{if } k = 0; \\ -g_k + \beta_k d_{k-1}, & \text{otherwise}, \end{cases} \tag{1.2}$$

* Corresponding author. Tel.: +302610997833.
E-mail address: livieris@upatras.gr, livieris@gmail.com, ioannis.livieris@hotmail.com (I.E. Livieris).

http://dx.doi.org/10.1016/j.amc.2015.05.053
0096-3003/

where g_k is the gradient of E at w_k which can be easily obtained by means of back propagation of errors through the network layers and β_k is a scalar. There have been proposed several choices for β_k which give rise to distinct conjugate gradient methods. The most well known conjugate gradient methods include the Fletcher–Reeves (FR) method [9], the Hestenes–Stiefel (HS) method [14] and the Polak–Ribière (PR) method [33] whose update parameters are respectively specified as follows:

$$\beta_k^{\mathrm{HS}} = \frac{g_k^T y_{k-1}}{y_{k-1}^T d_{k-1}}, \quad \beta_k^{\mathrm{FR}} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}, \quad \beta_k^{\mathrm{PR}} = \frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2},$$

where $y_{k-1} = g_k - g_{k-1}$ and $\|\cdot\|$ denotes the Euclidean norm. In the literature, there have been devoted many efforts to study and investigate the numerical efficiency and convergence properties of conjugate gradient methods (see [7,12,28] and the references therein).

Unfortunately these training methods, cannot guarantee to generate descent directions, therefore the use of restarts are employed in order to guarantee convergence. Nevertheless, there is also a worry with restart algorithms that restarts may be triggered too often; thus degrading the overall efficiency and robustness of the training process [27]. Furthermore, most classical conjugate gradient methods have the drawback of not being globally convergent for general functions and as a result they can cycle infinitely without presenting any substantial progress [34]. To overcome these difficulties, Chen and Liu [5] based on the works [45,46] proposed a new class of conjugate gradient methods by modifying the search direction (1.2) in the following way:

$$d_k = -\left(1 + \beta_k \frac{g_k^T d_{k-1}}{\|g_k\|}\right) g_k + \beta_k d_{k-1}.$$

An attractive property of this class is that the property $g_k^T d_k = -\|g_k\|$ always holds. Moreover, if β_k is specified by an existing conjugate gradient formula, we obtain the corresponding modified conjugate gradient method. Along this line, many related conjugate gradient methods have been extensively studied which possess global convergence for general functions and are also computationally competitive to classical methods [5,20,43,44]. On the basis of this idea, Livieris and Pintelas [19,21,22] proposed some descent conjugate gradient training algorithms providing some promising results. Based on their numerical experiments the authors concluded that the sufficient descent property led to a significant improvement of the efficiency of the training process.

However, the global convergence property ensures that starting from any initial point (weight) the training method will reach a minimizer but not necessarily a global minimum. This leads us to the conclusion that even if an algorithm is globally convergent, there is no guarantee that the method will efficiently explore the error surface since it may be trapped in a local minimum early. This is primarily caused by the fact that traditional conjugate gradient algorithms are monotone. In particular, enforcing monotonicity may considerably reduce the rate of convergence when the iteration is trapped near a narrow curved valley, which can result in very short steps [40]. Therefore, it might be advantageous to allow the iterative sequence to occasionally generate points with nonmonotone objective values. Grippo et al. [11] proposed a nonmonotone learning strategy that exploits the accumulated information with regard to the most recent values of the function. The philosophy behind nonmonotone strategy is that, many times, the first choice of a trial point by a minimization algorithm hides a lot of wisdom about the problem structure and that such knowledge can be destroyed by the decrease imposition. On the basis of this idea, many researchers [6,38,40] have proposed new nonmonotone schemes and some encouraging numerical results have been reported [29–32] when nonmonotone algorithms were applied to difficult nonlinear problems.

Motivated by the previous works, we propose a new class of conjugate gradient methods for training neural networks which ensures sufficient descent independent of the accuracy of the line search. Moreover, our proposed methods utilize a new modified nonmonotone strategy proposed by Shi and Wang [38] and they are globally convergent, under mild conditions. Our experimental results indicate that the proposed class of training methods are efficient and have a potential to significantly enhance the computational efficiency and robustness of the training process.

The remainder of this paper is organized as follows: In Section 2, we present the modified nonmonotone line search and our proposed class of conjugate gradient training algorithms and Section 3 presents the global convergence analysis. The numerical experiments are reported in Section 4 using the performance profiles Dolan and Morè [8]. Finally, Section 5 presents our concluding remarks and our proposals for future research.

2. Nonmonotone modified conjugate gradient training algorithm

Recently, Shi and Wang [38] proposed a modification of nonmonotone Armijo line search as follows:
Modified nonmonotone Armijo line search. Given a nonnegative integer M, the index $m(k)$ is defined by

$$m(0) = 0, \ 0 \le m(k) \le \min[m(k-1), M], \quad k \ge 1.$$

Set scalars δ_k, ρ, μ, $L_k > 0$ and σ with $\delta_k = -\frac{g_k^T d_k}{L_k \|d_k\|^2}$, $\sigma \in (0, \frac{1}{2})$, $\rho \in (0, 1)$ and $\mu \in [0, 2)$. Let η_k be the largest one in $\{\delta_k, \delta_k\rho, \delta_k\rho^2, \ldots\}$ such that

$$E(w_k + \eta_k d_k) - \max_{0 \le j \le m(k)} [E_{k-j}] \le \sigma \eta_k \left[g_k^T d_k + \frac{1}{2} \eta_k \mu L_k \|d_k\|^2 \right], \tag{2.1}$$