Pattern Recognition 43 (2010) 1984–1992

Contents lists available at ScienceDirect

Pattern Recognition

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

A new convex objective function for the supervised learning of single-layer neural networks

Oscar Fontenla-Romero *, Bertha Guijarro-Berdiñas, Beatriz Pérez-Sánchez, Amparo Alonso-Betanzos

Laboratory for Research and Development in Artificial Intelligence (LIDIA), Department of Computer Science, Faculty of Informatics, University of A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain

ARTICLE INFO

Article history:
Received 8 June 2009
Received in revised form
27 October 2009
Accepted 26 November 2009

Keywords:
Single-layer neural networks
Global optimum
Supervised learning method
Least squares
Convex optimization
Incremental learning

ABSTRACT

This paper proposes a novel supervised learning method for single-layer feedforward neural networks. This approach uses an alternative objective function to that based on the MSE, which measures the errors before the neuron's nonlinear activation functions instead of after them. In this case, the solution can be easily obtained solving systems of linear equations, i.e., requiring much less computational power than the one associated with the regular methods. A theoretical study is included to proof the approximated equivalence between the global optimum of the objective function based on the regular MSE criterion and the one of the proposed alternative MSE function.

Furthermore, it is shown that the presented method has the capability of allowing incremental and distributed learning. An exhaustive experimental study is also presented to verify the soundness and efficiency of the method. This study contains 10 classification and 16 regression problems. In addition, a comparison with other high performance learning algorithms shows that the proposed method exhibits, in average, the highest performance and low-demanding computational requirements.

1. Introduction

For a single-layer feedforward neural network, with linear activation functions, the weight values minimizing the mean-squared error function (MSE) can be found in terms of the pseudo-inverse of a matrix [1,2]. Furthermore, it can be demonstrated that the MSE surface of this linear network is a quadratic function of the weights [3]. Therefore, this convex hyperparaboloidal surface can be easily traversed by a gradient descent method. However, if nonlinear activation functions are used then local minima can exist in the objective function based on the MSE criterion [4–6]. In [7] it was shown that the number of such minima can grow exponentially with the input dimension. Only in some specific situations it is guaranteed the lack of local minima. In the case of linearly separable patterns and a threshold MSE criterion, it was proved the existence of only one minimum in the objective function [8,9]. Nevertheless, this is not the general situation.

The contribution of this work is to present a new convex objective function, equivalent to the MSE, that does not contains local minima and the global solution is obtained using a system of linear equations. This system can be solved, for each output, with a complexity of $O(N^2)$, where N is the number of parameters of the network.

The problem of local minima for one-layer networks was rigorously demonstrated in [5], where an example with a sigmoid transfer function, for which the sum of squared errors presents a local minimum, is given. They pointed out that the existence of local minima is due to the fact that the error function is the superposition of functions whose minima are at different points. In this situation, a closed form solution is no longer possible.

Previous approaches, during the last decades, have been presented to overcome the problems emerged by the presence of these stationary points in single-layer neural networks. In [10], a globally convergent natural homotopy mapping is defined for single-layer perceptrons by deformation of the node nonlinearity. This homotopy tracks a possibly infinite number of weights by transforming coordinates and characterizing all solutions by a finite number of distinct and unique solutions. Although this approach ensures computation of a solution, it does not provide global optimization [1]. At the same time, these authors proposed in [11] a method for both the a posteriori evaluation of whether a solution is unique or globally optimal and for a priori scaling of desired vector values to ensure uniqueness, through analysis of the input data. Although these approaches are potentially helpful for evaluating optimality and uniqueness, the minima are characterized only after training is complete. In addition, other authors have proposed methods for different criteria from the MSE to avoid the problem of local minima in the objective

[*] Corresponding author.
E-mail addresses: ofontenla@udc.es (O. Fontenla-Romero), cibertha@udc.es (B. Guijarro-Berdiñas), bperezs@udc.es (B. Pérez-Sánchez), ciamparo@udc.es (A. Alonso-Betanzos).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.11.024

function to minimize. In this sense, in [12] it was proposed an on-line additive learning method for matching cost functions based on the Bregman divergence.

Pao [2] proposed the functional link approach that obtains an analytical solution of the weights establishing a system of linear equations $Xw = z$, where X is a matrix formed by the input patterns, w is the weight vector and z is another vector formed by the inverse of the activation function applied over the desired output. The dimensions of matrix X are $S \times N$, where S is the number of training patterns and N is the number of weights. As Pao mentions in his work, if $S = N$ and the determinant of X is not zero then the solution can be obtained by $w = X^{-1}z$. However, this is not the common situation, because in real data sets usually $S > N$ or $S < N$. For these last cases, Pao analyzes the situation separately. In the case $S < N$, a large number of solutions could be obtained (perhaps an infinite number of them) which is not obviously desired. He proposed a partition of X to avoid in some way this problem. In the other case, $S > N$, an infinitely large number of orthonormal functions could be generated, and then a method based on the pseudoinversion is proposed ($w = (X^TX)^{-1}X^Tz$). However, as he already mentions, this formulation could be often unacceptable as is indicated by the high error value at the end of the learning process.

Some studies for multilayer feedforward neural networks have used similar results to the one proposed in [2] for the backpropagation of the desired output or for the learning of the weights of the output layer. Specifically, there have been heuristic proposals of least-squares initialization and training approaches [13–17]. Of special interest is [15], where three least-squares initialization schemes were compared for speed and performance. Nevertheless, these methods did not rigorously consider the transformation of the desired output through the nonlinear activation functions as they did not take into account the scaling effects of the slopes of nonlinearities in the least squares problem. This is an important issue as it will be discussed later.

Lastly, in a previous paper [18], a new learning method, for single-layer neural networks, based on a system of linear equations was presented. This approach is possible due to the use of a new objective function that measures the sum of the squared errors *before* the nonlinear activation functions instead of *after* these functions, as it is usually done. Although the experimental results presented in this previous work support the validity and soundness of the proposed method, some theoretical research was still necessary to proof the equivalence between the global optimum of the objective function based on the MSE after the nonlinearities and the proposed objective function (minimization of the MSE before the nonlinear functions). This paper completes the mentioned research presenting a theoretical analysis and considering in the objective function the scaling effects of the slope of the nonlinear transfer function. Besides, a new set of linear equations, to obtain the optimal weights for the problem, are derived.

2. Description of the proposed method

The architecture of the considered neural network is shown in Fig. 1. The inputs are denoted as x_{is} and outputs as y_{js} being $i = 0, 1, \ldots, I$; $j = 1, 2, \ldots, J$ and $s = 1, 2, \ldots, S$. The numbers I, J and S represent the number of inputs, outputs and training samples, respectively. The network contains only a single layer of J output neurons with nonlinear activation functions $f_1, f_2, \ldots, f_J$. The set of equations relating inputs and outputs is given by

$$y_{js} = f_j(z_{js}) = f_j\left(\sum_{i=0}^{I} w_{ji}x_{is}\right), \quad j = 1, 2, \ldots, J, \;\; s = 1, 2, \ldots, S, \tag{1}$$

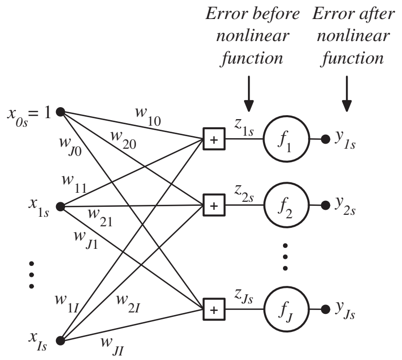

Fig. 1. Architecture of a single-layer feedforward neural network.

where w_{j0} and w_{ji}, $i = 1, 2, \ldots, I$, are, respectively, the bias and the weights associated with neuron j (for $j = 1, 2, \ldots, J$). The system presented in (1) has $J \times S$ equations and $J \times (I+1)$ unknowns. In practice, since the number of data is large ($S \geqslant I+1$), this set of equations does not have a solution, and consequently, it cannot be solved analytically.

Thus, the widely employed approach to obtain the optimal weights is based on the optimization, by means of an iterative procedure, of an objective function that measures the errors obtained by comparing the real output of the network and some desired response.

2.1. Regular objective function: mean-squared error after nonlinearities

Currently, different objective functions have been proposed being one of the most used that based on the mean-squared error (MSE) criterion. This is the function considered in this work. Thus, the usual approach is to consider some errors, ε_{js} measured *after* the nonlinearities. Therefore, the set of equations relating inputs and outputs is now defined as

$$\varepsilon_{js} = d_{js} - y_{js} = d_{js} - f_j\left(\sum_{i=0}^{I} w_{ji}x_{is}\right), \quad j = 1, 2, \ldots, J, \;\; s = 1, 2, \ldots, S, \tag{2}$$

where d_{js} is the desired output for neuron j and the training pattern s. To estimate (learn) the weights, the sum of squared errors defined as

$$MSEA = \sum_{s=1}^{S}\sum_{j=1}^{J} \varepsilon_{js}^2 = \sum_{s=1}^{S}\sum_{j=1}^{J}\left(d_{js} - f_j\left(\sum_{i=0}^{I} w_{ji}x_{is}\right)\right)^2 \tag{3}$$

is minimized. There exists many gradient descent methods that can be used to obtain a saddle point of this function.

It is important to note that, due to the presence of the nonlinear activation functions f_j, the function in (3) is nonlinear in the weights. In this situation, the absence of local minima in *MSEA* is not guaranteed, as was demonstrated in [5]. Therefore, a gradient descent method can be stuck in a local minimum instead of achieving the global optimum of the objective function.

2.2. New objective function: mean-squared error before nonlinearities

In order to avoid the problems mentioned in the previous section, a new approach for the supervised learning of single-layer feedforward neural networks is proposed. This method is based on the use of an alternative objective function that measures the