Egyptian Informatics Journal (2011) **12**, 197–209

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

ORIGINAL ARTICLE

Shannon Entropy and Mean Square Errors for speeding the convergence of Multilayer Neural Networks: A comparative approach

Hussein Aly Kamel Rady

El Shorouk Academy, Computer Science Department, El Shorouk – Cairo, P.O. Box 3, El Shorouk, Egypt

Received 18 May 2011; revised 22 September 2011; accepted 26 September 2011
Available online 9 November 2011

KEYWORDS

Shannon Entropy;
Mean Square Error;
Activation function;
Learning rate;
Backpropagation Neural Network

Abstract Improving the efficiency and convergence rate of the Multilayer Backpropagation Neural Network Algorithms is an active area of research. The last years have witnessed an increasing attention to entropy based criteria in adaptive systems. Several principles were proposed based on the maximization or minimization of entropic cost functions. One way of entropy criteria in learning systems is to minimize the entropy of the error between two variables: typically one is the output of the learning system and the other is the target. In this paper, improving the efficiency and convergence rate of Multilayer Backpropagation (BP) Neural Networks was proposed. The usual Mean Square Error (MSE) minimization principle is substituted by the minimization of Shannon Entropy (SE) of the differences between the multilayer perceptions output and the desired target. These two cost functions are studied, analyzed and tested with two different activation functions namely, the Cauchy and the hyperbolic tangent activation functions. The comparative approach indicates that the Degree of convergence using Shannon Entropy cost function is higher than its counterpart using MSE and that MSE speeds the convergence than Shannon Entropy.

E-mail address: dr_Hussein_Rady@yahoo.com

Peer review under responsibility of Faculty of Computers and Information, Cairo University.
doi:10.1016/j.eij.2011.09.002

Production and hosting by Elsevier

1. Introduction

Artificial Neural Networks (ANNs) has been a hot topic in recent years in cognitive science, computational intelligence and intelligent information processing [1–7]. They have emerged as an important tool for classification. The recent vast research activities in neural classification have established that neural networks are a promising alternative to various conventional classification methods [8,9]. On the other hand, a Neural Network is a well known as one of powerful computing tools to solve optimization problems. Due to massive computing unit neurons and parallel mechanism of neural network

approach it can solve the large-scale problem efficiently and optimal solution can be obtained [10]. The advantage of neural networks lies in the following theoretical aspects. *First*, neural networks are data driven self-adaptive methods in that they can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model. *Second*, they are universal functional approximators in that neural networks can approximate any function with arbitrary accuracy. *Third*, neural networks are nonlinear models, which makes them flexible in modeling real world complex relationships. Finally, neural networks are able to estimate the posterior probabilities, which provides the basis for establishing classification rule and performing statistical analysis.

The feedforward neural network [11–15] is the simplest (and therefore, the most common) ANN architecture in terms of information flow direction. Many of neural network architectures are variations of the feedforward neural network [16]. Backpropagation (BP) is the most broadly used learning method for feedforward neural networks [17,11,18,14]. There are two practical ways to implement the Backpropagation algorithm: batch updating approach and online updating approach. Corresponding to the standard gradient method, the batch updating approach accumulates the weight correction over all the training samples before actually performing the update. On the other hand, the online updating approach updates the network weights immediately after each training sample is fed [1,19].

Information theory is commonly used in coding and communication applications and more recently, it has also been used in classification. In information theoretic classification, a learner is viewed as an agent that gathers information from some external sources. Information theoretic quantities have been widely used for future extraction and selection [20]. As defined in information theory, entropy is a measure of the uncertainty of a particular outcome in a random process [1,21]. The entropy of a random variable is a measure of the uncertainty of the random variable; it is a measure of the amount of information required on the average to describe the random variable. Entropy is a nonlinear function to represent information we can learn from unknown data. In the learning process, we learn some constraints on the probability distribution of the training data from their entropy.

Usually error backpropagation for neural network learning is made using MSE as the cost function [22]. During the learning process, the ANN goes through stages in which the reduction of the error can be extremely slow. These periods of stagnation can influence learning times. In order to resolve this problem, the MSE are replaced by entropy error function [23,8,24]. Simulation results using this error function shows a better network performance with a shorter stagnation period. Accordingly, our purpose is the use of the minimization of the error entropy instead of the MSE as a cost function for classification purposes. Let the error $e(j) = T(j) - Y(j)$ represent the difference between the target T of the j output neuron and its output Y, at a given time t. The MSE of the variable $e(j)$ can be replaced by its EEM counterpart.

MSE has been a popular criterion in the training of all adaptive systems including artificial neural networks. The two main reasons behind this choice are analytical tractability and the assumption that real-life random phenomena may be sufficiently described by second-order statistics. The Gaussian probability density function (pdf) is determined only by its first- and second-order statistics, and the effect of linear systems on low order statistics is well known. Under these linearity and Gaussianity assumptions, further supported by the central limit theorem, MSE, which solely constrains second-order statistics, would be able to extract all possible information from a signal whose statistics are solely defined by its mean and variance [25]. On the other hand, MSE can extract all the information in the data provided that the dynamic system is linear and the noise is Gaussian distributed. However, when the system becomes nonlinear and the noise distribution is non-Gaussian, MSE fails to capture all the information in the error sequences. In this case an alternative criterion is needed in order to achieve optimality. Entropy is a natural extension beyond MSE since entropy is a function of probability density function (pdf), which considers all high order statistics [26]. Various optimization techniques were suggested for improving the efficiency of error minimization process or in other words the training efficiency [27,28].

The rest of the paper is organized as follows. Related work is outlined in Section 2. Section 3 introduces the Multilayer Backpropagation Neural Networks. Section 4 introduces the Mean Square Error. Shannon Entropy was discussed and analyzed in Section 5. Simulated results were discussed in Section 6 for Shannon Entropy and in Section 7 for Mean Square Error. Section 8 compares Shannon Entropy and MSE. Finally Conclusions are outlined in Section 9.

2. Related work

Entropy, which is introduced by Shannon, is a scalar quantity that provides a measure for the average information contained in a given probability distribution function. By definition, information is a function of the pdf; hence, entropy as an optimality criterion extends MSE. When entropy is minimized, all moments of the error pdf (not only the second moments) are constrained. The entropy criterion can generally be utilized as an alternative for MSE in supervised adaptation, but it is particularly appealing in dynamic modeling [25]. MSE can extract all the information in the data provided that the dynamic system is linear and the noise is Gaussian distributed. However, when the system becomes nonlinear and the noise distribution is non-Gaussian, MSE fails to capture all the information in the error sequences. Entropy is a natural extension beyond MSE since entropy is a function of probability density function (pdf), which considers all high order statistics [26].

Many researchers introduces the theoretical concepts of using Error Entropy Minimization as a cost function for artificial Neural Networks. In [26], Xu et al. discusses the information theoretic learning and states that entropy, which measures the average information content in a random variable with a particular probability distribution was previously proposed as a criterion for supervised adaptive filter training and it was shown to provide better neural network generalization compared to MSE. In [22], Alexandre and Sa introduces the Error Entropy Minimization approach to replace the MSE, as the cost function of a learning system, with the entropy of the error. They discusses the theoretical basis of the Renyi's quadratic entropy. In their experimental results, they used three values of Learning rates which are 0.1, 0.2, and 0.3 with MSE and EEM for different smoothing parameters and they