Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 32 (2007) 599–605
Expert Systems with Applications
www.elsevier.com/locate/eswa

Treatment of multi-dimensional data to enhance neural network estimators in regression problems

H. Altun [a,*], A. Bilgil [b], B.C. Fidan [c]

[a] Department of Electrical and Electronics Engineering, Nigde University, 51200 Nigde, Turkey
[b] Department of Civil Engineering, Nigde University, 51200 Nigde, Turkey
[c] Yıldız Technical Univeristy, Faculty of Electrical and Electronics Engineering, Department of Electrical Engineering, 34349 Beşiktaş, İstanbul, Turkey

Abstract

This paper proposes and explains a data treatment technique to improve the accuracy of a neural network estimator in regression problems, where multi-dimensional input data set is highly skewed and non-normally distributed. The proposed treatment modifies the distribution characteristics of the data set. The prediction of the suspended sediment, which is an important problem in river engineering applications, will be considered as a case study. Conventional approaches lack in providing high accuracy due to the inherently employed simplicity in order to obtain empirical formulae. On the other hand, artificial neural networks are able to model the non-linear characteristics of the mechanism of the sediment transport and have a growing body of applications in diverse applications in civil engineering. It will be shown that a significant enhancement and superior score in accuracy, compared with the classical approaches, are obtainable when the proposed treatment is employed. The proposed technique is an extension to the understanding of the practical aspects of neural computing applications. Therefore the outcome of the present study is important as it is applicable to any scenario where neural network approaches are involved.

Keywords: Artificial neural networks; Regression; Multi-layered perceptron; Sediment transport; Skewness; Training data treatment

1. Introduction

Artificial neural networks (ANNs) are currently being used intensively in diverse engineering applications and become a regular method to provide solution in optimisation, regression and estimation problems. Despite of their simplified structure, ANNs have an ability to mimic human characteristics of problem solving via learning and generalization. Among various kind of ANNs, Multi-Layered Perceptron (MLP) has become the most widely used network architecture in neural network application.

In most applications where neural networks are expected to model highly non-linear and multi-dimensional functions, experimental data shows a non-uniform distribution. This fact is in line with the central limit theorem which states that data from an experiment approaches to normal distribution as the number of sample taken approaches infinity. However, in real life problems this assumption is not always realistic as data set might show non-normal and highly skewed distribution. The phenomenon is also mostly true in engineering problems. The common practice adopted by neural network practitioners is to scale the data set linearly into a small range before training process. As a result of this treatment, the data is scaled into a predefined range but the distribution characteristic of the data is preserved.

A recent study by Kumar (2005) which compares neural networks and classical regression methods shows that skewness in data set should be reduced using some transformation like power transformation before carrying out neural network analysis. It is indicated that skewness of

* Corresponding author. Tel.: +90 388 225 2281; fax: +90 388 225 1012.
E-mail addresses: haltun@nigde.edu.tr (H. Altun), abilgil@nigde.edu.tr (A. Bilgil), fidan@yildiz.edu.tr (B.C. Fidan).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.01.054

the dependent variables play a significant role in the predictive accuracy of the neural network models. The Box-Cox procedure is used to determine an appropriate transformation which reduces the skewness presented in data. Similar research was conducted by SubbaNarasimha, Arinze, and Anandarajan (2000) to see whether neural network models are better than multiple regression at prediction in samples which are highly skewed. These studies, though shed light on the practical use of neural network, are rather exploratory. Furthermore, these studies lack in considering and explaining the affect of the distribution of the target data set which is as much decisive as input data on neural learning. In literature, through an analytical and statistical analysis, it has been shown that learning in MLP neural network is susceptible to statistical characteristics of input and as well as that of target data (Altun & Curtis, 1999a, 1999b; Altun, Curtis, & Yalcinoz, 2003). The present article contributes to this line of research and makes use of the results to improve the accuracy of a neural network in the estimation of sediment transport.

2. Definition of the problem

Estimation of sediment is very important in hydraulic engineering applications such as the control and management of watersheds, the design of water constructions and reservoirs and the prediction of economical life of investments, etc. Several works on the mechanism of sediment transport were carried by many researchers. The conventional approaches to the problem can be classified broadly into two classes, namely indirect and direct methods. Indirect methods attempt to predict the total sediment discharge from the sum of bed load and suspended load functions. These are generally based on the pioneering research of Einstein (1950). Several formulae were developed to determine the incipient motion of non-uniform sediment mixtures by Zanke (1978), van Rijn (1984), Şentürk (2000), Misri, Ranga Raju, and Garde (1984), Samaga, Ranga Raju, and Garde (1986), Patel and Ranga Raju (1996), Cheng and Chiew (1999) and Wu, Wang, and Yafei (2000). The outcome of these studies are the empirical formulae which estimate the fractional transport rates of suspended load and bed material load. However, it has been shown that the indirect methods are difficult to apply and do not produce fairly high fidelity in prediction of sediment discharge. On the other hand, direct methods attempt to predict the total sediment discharge directly without making distinction between suspended and bed load.

In the present, the number of independent parameters and the unclear relation between total sediment load and these parameters are the open problems to be tackled. Several simplicities are assumed in order to obtain a relation between measured parameters and sediment discharge. As a consequence of these simplifications, the formulae obtained by the conventional methods in the prediction of sediment discharge are incapable of being completely satisfactory. It has been shown in literature that Artificial Neural Networks (ANNs) are highly capable in non-linear regression problems. In problem solving using ANNs, there is no need to suppose unrealistic assumption for the sake of simplicity. As a result, it is a reasonable expectation to have a more accurate sediment prediction from measured parameters.

There is a growing body of neural network applications in literature in civil engineering such as in the design of coastal sewage system by Sanchez, Arroyo, Garcia, Koev, and Revilla (1998), in the assessment of water quality by Haas (2004) and by Kuo, Liu, and Lin (2004), in the prediction of the flow by Cigizoglu (2003a, 2003b) and other hydrologic applications. ANNs have also been successfully applied to the sediment discharge prediction by Nagy, Watanabe, and Hirano (2002) and Cigizoglu (2002). In the present paper further improvement in the prediction of suspended sediment discharge using neural network approach will be introduced by a novel treatment of training data sets. It will be shown that the performance of neural networks can be considerably improved when training data is appropriately modified according to the findings reported. The performance of the proposed neural network approach will be compared with the methods proposed by Zanke (1978), van Rijn (1984), Şentürk (2000), Samaga et al. (1986), Wu et al. (2000) and EIE (2002).

3. Artificial Neural Networks

ANN is a highly promising emerging technology in modeling non-linear systems due to their ability to learn the system behavior under inspection from samples. In the training process, experimental data are introduced to neural network in order to establish a relationship between the input and output of the system under inspection. The objective of the training algorithm is to reduce the total error between predicted output and given target values. Gradient based algorithms are mostly employed in supervised learning process. In the proposed technique, improvements will be acquired for MLP neural network with backpropagation (BP) algorithm proposed by Rumelhart, Hinton, and Williams (1986).

3.1. MLP neural networks with backpropagation algorithm

MLP neural network is organized in layers, i.e. input layer, one or more hidden layers and output layer, namely. In each layer a specific number of neurons with an activation function are placed. Together with the number of neurons, the number of the layers determines the topology of MLP neural network. A MLP neural network mostly contains one or two hidden layers. The number of neurons in the input and output layers is set to match the number of input and target parameters of the process under investigation. The number of neurons in the hidden layers, on the other hand, is determined by experience and some rule of thumb. Therefore, there could be more than one topology for an MLP to model a process successfully.