Neural Networks 99 (2018) 148–157

Contents lists available at ScienceDirect

Neural Networks

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

A novel type of activation function in artificial neural networks: Trained activation function

Ömer Faruk Ertuğrul

Department of Electrical and Electronics Engineering, Batman University, 72060 Batman, Turkey

ARTICLE INFO

Article history:
Received 3 April 2017
Received in revised form 16 December 2017
Accepted 18 January 2018
Available online 31 January 2018

Keywords:
Activation function
Trained activation function
Artificial neural network
Random weight artificial neural network

ABSTRACT

Determining optimal activation function in artificial neural networks is an important issue because it is directly linked with obtained success rates. But, unfortunately, there is not any way to determine them analytically, optimal activation function is generally determined by trials or tuning. This paper addresses, a simpler and a more effective approach to determine optimal activation function. In this approach, which can be called as trained activation function, an activation function was trained for each particular neuron by linear regression. This training process was done based on the training dataset, which consists the sums of inputs of each neuron in the hidden layer and desired outputs. By this way, a different activation function was generated for each neuron in the hidden layer. This approach was employed in random weight artificial neural network (RWN) and validated by 50 benchmark datasets. Achieved success rates by RWN that used trained activation functions were higher than obtained success rates by RWN that used traditional activation functions. Obtained results showed that proposed approach is a successful, simple and an effective way to determine optimal activation function instead of trials or tuning in both randomized single and multilayer ANNs.

1. Introduction

Although successful results have been reported in artificial neural networks (ANNs) in many cases, it is really hard or sometimes may be impossible to optimize the structure of an ANN (i.e., the number of neurons in the hidden layer and the activation function) and the learning parameters (Basheer & Hajmeer, 2000; Guoqiang Zhang & Eddy Patuwo, 1998). In randomized ANNs, the issue of determining optimal learning parameters has been solved by assigning the weights and biases in the hidden layer randomly and calculating the other weights and biases analytically via the Fisher method (Huang, Zhu, & Siew, 2004; Huang, Zhu, & Siew, 2006; Pao, Park, & Sobajic, 1994; Schmidt, Kraaijveld, & Duin, 1992; Zhang & Suganthan, 2016a, b). Due to this non-tuning approach, randomized ANNs do not require to use learning rate, the number of maximum epochs, or stopping criteria.

In literature, generally, ANNs that have different network structures are tested and the structure of one that yields the best success rate is assigned as the optimal ANN structure (Ertuğrul, 2016; Ertuğrul & Kaya, 2016). But, in this approach, the activation function and the number of neurons in the hidden layer that are used in tests must be determined by the user/researcher. As an alternative to this complex and time-consuming approach, pruning methods have been employed in order to automatically or adaptively optimize the number of neurons in the hidden layer (Huang & Chen, 2008; Huang, Chen, & Siew, 2006).

However, determining an optimal number of neurons in the hidden layer is still an open issue (Duch & Jankowski, 2001; Dudek, 2016; Huang, Bin Huang, Song, & You, 2015). The results reported in the literature showed that a relationship between the statistical properties of the dataset and the most suitable activation function could not be found. It can be said that the most suitable activation function for a dataset may not be the optimal one for another dataset (Duch & Jankowski, 1997). Based on the high relationship between the achieved success rate of an ANN with the employed activation function, the optimal activation function is generally searched via trials (Sharma & Venugopalan, 2014; Dorofki, Elshafie, Jaafar, & Karim, 2012; Ertuğrul, 2016) or by tuning (Li, Li, & Rong, 2013; Shen & Wang, 2004; Wu, Zhao, & Ding, 1997; Youshou, Mingsheng, & Xiaoqing, 1997).

There are still two important unanswered questions: in which activation functions of the optimal one must be searched, since, it was proven that a single hidden layer feed-forward ANN keeps up its approximation capability with arbitrary selected bounded and non-constant piecewise of any continuous function (Hornik, 1991; Huang et al., 2015, 2006) such as sigmoid (Schmidt et al., 1992), radial basis function (Broomhead & Lowe, 1988; Hernández-Aguirre, Koutsougeras, & Buckles, 2002; Lowe, 1989; Porwal, Carranza, & Hale, 2003) or trigonometric functions (e.g., sin and cos)

E-mail address: omerfarukertugrul@gmail.com.

https://doi.org/10.1016/j.neunet.2018.01.007
0893-6080/

Table 1
Properties of employed datasets.

Name	Type	Performed task	#Attributes	#Observations	#Classes	Source
Lithuanian	Synthetic	Classification	2	1 000	2	Duin et al. (2004)
Highleyman	Synthetic	Classification	2	1 000	2	Duin et al. (2004)
Banana shaped	Synthetic	Classification	2	1 000	2	Duin et al. (2004)
Spherical	Synthetic	Classification	2	1 000	2	Duin et al. (2004)
Liver	Real	Classification	7	345	2	Lichman (2013)
Pima Indian diabetes	Real	Classification	8	762	2	Lichman (2013)
Banana	Real	Classification	2	5 300	2	mldata.org
Image segmentation	Real	Classification	19	2 315	7	Lichman (2013)
Satellite image	Real	Classification	36	6 435	7	Lichman (2013)
Statlog (shuttle)	Real	Classification	9	58 000	7	Lichman (2013)
Wine	Real	Classification	13	178	3	Lichman (2013)
Cardiotocography	Real	Classification	22	2 126	3	Lichman (2013)
Skin segmentation	Real	Classification	3	245 057	2	Lichman (2013)
Seeds	Real	Classification	7	210	3	Lichman (2013)
Seismic bumps	Real	Classification	18	2 584	2	Lichman (2013)
Banknote authentication	Real	Classification	4	1 372	2	Lichman (2013)
Balance scale	Real	Classification	4	625	3	Lichman (2013)
Acute inflammations	Real	Classification	7	120	2	Lichman (2013)
Dermatology	Real	Classification	35	366	6	Lichman (2013)
Diabetic retinopathy debrecen	Real	Classification	19	1 151	2	Lichman (2013)
Fertility	Real	Classification	9	100	2	Lichman (2013)
Haberman	Real	Classification	3	306	2	Lichman (2013)
Hayes-Roth	Real	Classification	5	132	3	Lichman (2013)
QSAR biodegradation	Real	Classification	41	1 055	2	Lichman (2013)
Climate model simulation	Real	Classification	18	540	2	Lichman (2013)
Approximate sinc	Synthetic	Regression	1	5 000	–	Huang et al. (2004)
Istanbul stock exchange	Real	Regression	7	537	–	Lichman (2013)
Ailerons	Real	Regression	40	13 750	–	dcc.fc.up.pt
Delta ailerons	Real	Regression	5	7 129	–	dcc.fc.up.pt
Auto-price	Real	Regression	15	159	–	Lichman (2013)
Bank-8FM	Real	Regression	8	6 481	–	dcc.fc.up.pt
Breast cancer	Real	Regression	32	194	–	Lichman (2013)
Census-8L	Real	Regression	8	22 784	–	dcc.fc.up.pt
Census-8H	Real	Regression	8	22 784	–	dcc.fc.up.pt
Census-16L	Real	Regression	16	22 784	–	dcc.fc.up.pt
Census-16H	Real	Regression	16	22 784	–	dcc.fc.up.pt
CPU-small	Real	Regression	12	8 192	–	Lichman (2013)
CPU	Real	Regression	21	8 192	–	Lichman (2013)
Diabetes child	Real	Regression	2	43	–	dcc.fc.up.pt
Delta elevators	Real	Regression	6	9 517	–	dcc.fc.up.pt
Elevators	Real	Regression	18	16 599	–	dcc.fc.up.pt
Kinematics	Real	Regression	8	8 192	–	dcc.fc.up.pt
Puma-8NH	Real	Regression	8	6 677	–	dcc.fc.up.pt
Puma-32H	Real	Regression	32	4 938	–	dcc.fc.up.pt
Pyrimidines	Real	Regression	28	74	–	dcc.fc.up.pt
Servo	Real	Regression	4	167	–	Lichman (2013)
Stocks	Real	Regression	9	950	–	dcc.fc.up.pt
Energy efficiency: cooling	Real	Regression	8	768	–	Lichman (2013)
Energy efficiency: heating	Real	Regression	8	768	–	Lichman (2013)
Yacht hydrodynamics	Real	Regression	7	308	–	Lichman (2013)

(Das & Panda, 2004). As seen in the literature, even determining the activation functions that will be employed in trials is really hard (Sharma & Venugopalan, 2014). Therefore, generally, only some popular activation functions are employed in tests and the activation function of the one, which showed highest success, is selected as optimal activation function. It is clear that the selected activation function is only the optimal one in a small group of activation functions. The second question is: whether an optimal activation for a neuron in the hidden layer is also optimal one for the other neurons.

This paper was written in order to address a way to determine an optimal activation function for each particular neuron in the hidden layer based on a simple idea, why we do not train activation functions or in other words, why a trainable function was not employed as an activation function. In order to validate the proposed idea, the linear regression, which is a popular way to obtain a relationship between inputs and output, was employed as the activation functions in the hidden layer. The proposed approach was evaluated and validated by 50 benchmark datasets, where 25 of them are classification and the others are regression datasets. Obtained success rates showed that the proposed approach can be employed as an alternative to classical activation functions. In the rest of the paper, in Section 2, the employed datasets were described briefly. The proposed approach and its application to the random weight artificial neural networks (RWN) were given in Section 3. Obtained results were given and discussed in Section 4 and the paper was concluded in Section 5.

2. Employed datasets

To evaluate and validate the proposed approach 25 classification and 25 regression benchmark datasets were employed. Details of each used dataset and their sources are given in Table 1.

3. Proposed approach

3.1. Trainable activation function

In ANN, a sample is mapped into a new feature space according to the activation function and other network parameters. As seen in the literature review, generally, the optimal activation function for a specific dataset was determined by two different approaches: