PERGAMON

Neural Networks 13 (2000) 561–563

Neural Networks

www.elsevier.com/locate/neunet

Neural Networks Letter

Neural networks with a continuous squashing function in the output are universal approximators

J.L. Castro*, C.J. Mantas, J.M. Benítez

Department of Computer Science and A.I., ETSI Informática, University of Granada, Avenida de Andalucia, 38, 18071 Granada, Spain

Received 30 January 1998; accepted 15 March 2000

Abstract

In 1989 Hornik as well as Funahashi established that multilayer feedforward networks without the squashing function in the output layer are universal approximators. This result has been often used improperly because it has been applied to multilayer feedforward networks with the squashing function in the output layer. In this paper, we will prove that also this kind of neural networks are universal approximators, i.e. they are capable of approximating any Borel measurable function from one finite dimensional space into $(0,1)^n$ to any desired degree of accuracy, provided sufficiently many hidden units are available.

Keywords: Feedforward networks; Universal approximation; Squashing functions; Continuous functions

1. Introduction

Since Hornik's and Funahashi's papers (Funahashi, 1989; Hornik, Stinchcombe & White, 1989) established the universal approximation capability of multilayer feedforward networks without the squashing function in the output layer, many papers (Bissessur & Naguib, 1995; Han, Xiu, Wang, Chen & Tan, 1997; Leyva, Martinez-Salamero, Jammes, Marpinard & Guinjoan, 1997; Moody & Antsaklis, 1996; Sietsma & Dow, 1991; Spall & Cristion, 1997) cite this result to justify that multilayer feedforward networks with the squashing function in the output layer are capable of approximating any continuous function F on $(0,1)^n$. Nevertheless, this has not been proved until now. We establish this result in this paper. In other works (Attali and Pages, 1997; Cardaliaguet and Euvrard, 1992; Hornik, 1991; Hornik, 1993; Kurkova, 1995; Leshno, Liu, Pinkus & Shocken, 1993; Meltser, Shoham & Manevitz, 1997), the universal approximation capability is also studied but a linear function is always considered in the output layer.

In the next section, it is proved that multilayer feedforward networks with strictly increasing squashing function in the output layer are capable of approximating any continuous function F on $(0,1)^m$. In Section 3, the multilayer feedforward networks with non-monotone, continuous, non-constant activation function in the output layer are studied. Finally, some conclusions are commented on.

2. Strictly increasing squashing function

The idea of the proof of the following theorem is to use Hornik's and Funahashi's results in order to prove that $\phi^{-1}\circ F$ can be approximated by a feedforward network without the squashing function in the output layer to any degree of accuracy, where ϕ is the squashing function of the output layer. Then, from the continuity of ϕ we conclude that F can be approximated by the same network with ϕ as the squashing function of the output layer. We begin with definitions and notations.

Definition 1. A function $\phi : \Re \rightarrow [0,1]$ is a squashing function if it is non-decreasing, $\lim_{\lambda\rightarrow\infty}\phi(\lambda) = 1$, and $\lim_{\lambda\rightarrow-\infty}\phi(\lambda) = 0$.

Theorem 1. *Let F be any Borel measurable or continuous function from $K \subset \Re^n$ on $(0,1)^m$, and let ϕ be any strictly increasing continuous squashing function. Then, for any $\epsilon > 0$ there exists a multilayer feedforward network N with the squashing function in the output layer and with only one hidden layer such that*

$$\|N(x) - F(x)\| < \epsilon \qquad \forall x \in K.$$

Proof (theorem 1). As ϕ is continuous and strictly increasing, there exists ϕ^{-1} and it is continuous. Then, let

* Corresponding author.

PII: S0893-6080(00)00031-9

G be defined by

$$G(x) = (\phi^{-1} \circ F(x)_1, \ldots, \phi^{-1} \circ F(x)_m).$$

Then G is a continuous function from K on $\Re^m$.

On the other hand, as G is continuous on K, $G(K)$ is a compact. Let $K' = [a, b]$ be the interval defined by

$$a = \min\{G(x)_i / x \in K, i = 1 \cdots m\} - 1,$$

$$b = \max\{G(x)_i / x \in K, i = 1 \cdots m\} + 1.$$

Then, as ϕ is uniformly continuous on every compact, there exists a $\delta > 0$ such that

$$|\phi(y) - \phi(z)| < \epsilon, \qquad \text{if } |y - z| < \delta \text{ and } y, z \in K'.$$

Let

$$\delta' = \begin{cases} \delta & \text{if } \delta \leq 1 \\ 1 & \text{otherwise} \end{cases}.$$

As G is a continuous function from K on $\Re^m$, there exists a multilayer feedforward network N' without the squashing function in the output layer and with only one hidden layer such that

$$\|N'(x) - G(x)\| = \max_{l \leq i \leq m} |N'(x) - G(x)_i| < \delta',$$

$$\forall x \in K,$$

hence, taking N as the neural network N' with the squashing function in the output, we have

$$\|N(x) - F(x)\| = \max_{l \leq i \leq m} |N(x)_i - F(x)_i|$$

$$= \max_{l \leq i \leq m} |\phi(N'(x)_i) - \phi((G(x)_i)| < \epsilon,$$

$$\forall x \in K,$$

since

$$a \leq G(x)_i - 1 \leq G(x)_i - \delta' < N'(x)_i < G(x)_i + \delta'$$

$$\leq G(x)_i + 1 \leq b,$$

$$a \leq G(x)_i - 1 < G(x)_i < G(x)_i + 1 \leq b,$$

and

$$|N'(x)_i - G(x)_i| < \delta' \leq \delta.$$

□

3. Activation function in the output is not monotone

This section exposes that the multilayer feedforward networks with non-monotone continuous non-constant activation function ($\varphi : \Re \to (0, 1)$) in the output layer, where φ is strictly monotone in an interval, are capable of approximating a linear transformation T of any function F on $(0, 1)^m$. The definition of the transformation T depends on the activation function φ, and is valid for approximating any function F with a neural network that uses the activation function φ in the output layer.

Definition 2. Let us consider a continuous non-constant function $\varphi : \Re \to (0, 1)$ that is strictly monotone in an interval $[a, b]$. Two cases:

1. Let us suppose that φ is strictly increasing in $[a, b]$, $\varphi(a) = u$ and $\varphi(b) = v$ with $0 < u < v < 1$. Hence, we can define the transformation $T : (0, 1) \to (u + \theta_1,\ v - \theta_1)$ as:

   $$T(x) = \frac{1 - (u + \theta_1)}{v - \theta_1} \cdot x + (u + \theta_1),$$

   where $\varphi(a + \theta_0) = u + \theta_1$, $\varphi(b - \theta_0) = v - \theta_1$ and θ_0, $\theta_1 > 0$.
2. If φ is strictly decreasing in $[a, b]$ then $\varphi(a) = v$ and $\varphi(b) = u$. Hence, we can define the transformation $T : (0, 1) \to (u + \theta_1,\ v\text{-}\theta_1)$. Now, $\varphi(a + \theta_0) = v - \theta_1$, $\varphi(b - \theta_0) = u + \theta_1$.

The definition of the function T depends on the values u and v. These values are established by the activation function φ.

Corollary 1. *Let F be any Borel measurable or continuous function from $K \subset \Re^n$ on $(0, 1)^m$, let φ be any non-monotone continuous non-constant activation function that is strictly monotone in an interval [a,b], and let T be the linear transformation associated with φ according to Definition 2. Then, for any $\epsilon > 0$ there exists a multilayer feedforward network N with activation function φ in the output layer and with only one hidden layer such that*

$$\|N(x) - (T \circ F)(x)\| < \epsilon \qquad \forall x \in K.$$

Proof (corollary 1). Without loss of generality, let us suppose that φ is strictly increasing in $[a, b]$, $\varphi(a) = u$ and $\varphi(b) = v$ with $0 < u < v < 1$. We build the function $\bar{\varphi}$ defined as

$$\bar{\varphi} : \Re \to \Re \qquad \bar{\varphi}(x) = \begin{cases} x + u - a & x < a \\ \varphi(x) & x \in [a, b] \\ x + v - b & x > b \end{cases}.$$

The function $\bar{\varphi}$ is continuous, strictly increasing and $\bar{\varphi}_{[a,b]}(x) = \varphi_{[a,b]}(x)$.

As $\bar{\varphi}$ is continuous and strictly increasing, there exists $\bar{\varphi}^{-1}$ and it is continuous. Then, let G be defined by

$$G(x) = (\bar{\varphi}^{-1} \circ T \circ F(x)_1, \ldots, \bar{\varphi}^{-1} \circ T \circ F(x)_m).$$

G is a continuous function from K on $\Re^m$.

On the other hand, as G is continuous on K, $G(K)$ is compact. By definition, $(T \circ F)(x) \in (u + \theta_1,\ v - \theta_1)\ \forall x \in K$.

Therefore, $G(x)_i \in (a+\theta_0, b-\theta_0)$, i.e. $G(x)_i \in [a,b]\ \forall x \in K, i=1,\ldots,m$.

Then, as $\bar{\varphi}$ is uniformly continuous on every compact, there exists a $\delta > 0$ such that

$$|\bar{\varphi}(y) - \bar{\varphi}(z)| < \epsilon, \qquad \text{if } |y-z| < \delta \text{ and } y,z \in [a,b].$$

Let $\delta' = \text{minimum}\{\theta_0, 1, \delta\}$.

As G is a continuous function from K on $\Re^m$, there exists a multilayer feedforward network N' without activation function in the output layer and with only one hidden layer such that

$$\|N'(x) - G(x)\| = \max_{l\le i\le m}|N'(x)_i - G(x)_i| < \delta',$$

$$\forall x \in K,$$

hence, taking $\bar{N}$ as the neural network N' with activation function $\bar{\varphi}$ in the output, we have

$$\|N'(x) - (T\circ F)(x)\| = \max_{l\le i\le m}|\bar{N}(x)_i - (T\circ F)(x)_i|$$

$$= \max_{l\le i\le m}|\bar{\varphi}(N'(x)_i) - \bar{\varphi}((G(x)_i)| < \epsilon,$$

$$\forall x \in K,$$

since

$$G(x)_i \in (a+\theta_0, b-\theta_0) \Rightarrow G(x)_i \in [a,b],$$

$$a \le a+\theta_0-\delta' \le G(x)_i - \delta' < N'(x)_i < G(x)_i + \delta'$$

$$\le b - \theta_0 + \delta' \le b, \Rightarrow N'(x)_i \in [a,b],$$

and

$$|N'(x)_i - G(x)_i| < \delta' \le \delta.$$

Finally, as $N'(x)_i \in [a,b]$ then $\bar{N}(x)_i = \bar{\varphi}(N'(x)_i) = \varphi(N'(x)_i) = N(x)_i,\ \forall x \in K,\ i=1,\ldots,m$.
Therefore

$$\|N(x) - (T\circ F)(x)\| = \|\bar{N}(x) - (T\circ F)(x)\| < \epsilon \qquad \forall x \in K.$$

□

4. Conclusion

The main result of this paper establishes that standard multilayer feedforward networks are capable of approximating any measurable function to any desired degree of accuracy. It is only a corollary of Hornik's and Funahashi's results, but it is necessary in order to apply the universal approximation capability to multilayer feedforward networks with the squashing function in the output layer correctly.

The second result establishes that the neural networks with non-monotone activation function (but strictly monotone inside an interval) in the output layer are capable of approximating the linear transformation of any measurable function. The majority of the non-monotone activation functions used in different works are strictly monotone inside an interval. Hence, the second result is important in practice.

References

Attali, J., & Pages, G. (1997). Approximations of functions by a multilayer perceptron. *Neural Network, 10* (6), 1069–1081.

Bissessur, Y., & Naguib, R. (1995). Buried plant detection: a Volterra series modelling approach using artificial neural networks. *Neural Networks, 9* (6), 1045–1060.

Cardaliaguet, P., & Euvrard, G. (1992). Approximation of a function and its derivative with a neural network. *Neural Networks, 5* (2), 207–220.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks, 2*, 183–192.

Han, Y., Xiu, L., Wang, Z., Chen, Q., & Tan, S. (1987). Artificial neural networks controlled fast valving in a power generation plant. *IEEE Transactions on Neural Networks, 8* (2).

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks, 4* (2), 251–257.

Hornik, K. (1993). Some new results on neural network approximation. *Neural Networks, 6* (8), 1069–1072.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*, 359–366.

Kurkova, V. (1995). Approximation of functions by perceptron networks with bounded number of hidden units. *Neural Networks, 8* (5), 745–750.

Leshno, M., Liu, V., Pinkus, A., & Shocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks, 6* (6), 861–867.

Leyva, R., Martinez-Salamero, L., Jammes, B., Marpinard, J., & Guinjoan, F. (1997). Identification and control of power converters by means of neural networks. *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications, 44* (8).

Meltser, M., Shoham, M., & Manevitz, L. (1997). Approximations of functions by a multilayer perceptron: a new approach. *Neural Networks, 10* (6), 1069–1081.

Moody, J., & Antsaklis, P. (1996). The dependence identification neural network construction algorithm. *IEEE Transactions on Neural Networks, 7* (1).

Sietsma, J., & Dow, J. (1991). Creating artificial neural networks that generalize. *Neural Networks, 4* (1), 67–79.

Spall, J., & Cristion, J. (1997). A neural network controller for systems with unmodeled dynamics with applications to wastewater treatment. *IEEE Transactions on Systems, Man and Cybernetics-Part B:Cybernetics, 27* (3).