Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

ELSEVIER International Journal of Approximate Reasoning 33 (2003) 185–202

www.elsevier.com/locate/ijar

A survey on universal approximation and its limits in soft computing techniques

Domonkos Tikk [a,*], László T. Kóczy [a], Tamás D. Gedeon [b]

[a] *Department of Telecommunication and Telematics, Budapest University of Technology and Economics, H-1117 Budapest, Magyar Tudósok körútja 2, Hungary*

[b] *School of Information Technology, Murdoch University, 6150 Murdoch, WA, Australia*

Received 1 March 2001; accepted 1 November 2002

Abstract

This paper deals with the approximation behaviour of soft computing techniques. First, we give a survey of the results of universal approximation theorems achieved so far in various soft computing areas, mainly in fuzzy control and neural networks. We point out that these techniques have common approximation behaviour in the sense that an arbitrary function of a certain set of functions (usually the set of continuous function, C) can be approximated with arbitrary accuracy ε on a compact domain. The drawback of these results is that one needs unbounded numbers of "building blocks" (i.e. fuzzy sets or hidden neurons) to achieve the prescribed ε accuracy. If the number of building blocks is restricted, it is proved for some fuzzy systems that the universal approximation property is lost, moreover, the set of controllers with bounded number of rules is nowhere dense in the set of continuous functions. Therefore it is reasonable to make a trade-off between accuracy and the number of the building blocks, by determining the functional relationship between them. We survey this topic by showing the results achieved so far, and its inherent limitations. We point out that approximation rates, or constructive proofs can only be given if some characteristic of smoothness is known about the approximated function.

Keywords: Universal approximation performed by fuzzy systems and neural networks; Kolmogorov's theorem; Approximation behaviour of soft computing techniques;

* Corresponding author. Tel.: +361-4631758; fax: +361-4631763.

E-mail addresses: tikk@ttt.bme.hu (D. Tikk), koczy@ttt.bme.hu (L.T. Kóczy), tgedeon@murdoch.edu.au (T.D. Gedeon).

0888-613X/03/$ - see front matter © 2003 Elsevier Science Inc. All rights reserved.
doi:10.1016/S0888-613X(03)00021-5

Course of dimensionality; Nowhere denseness; Approximation rates; Constructive proofs

1. Introduction

In 1900, in his memorable lecture at the Second International Congress of Mathematicians in Paris, D. Hilbert, the famous German mathematician, listed 23 conjectures, hypotheses concerning unsolved problems which he considered would be the most important ones to be solved by the mathematicians of the 20th century. According to the 13th conjecture there exist continuous multivariable functions which cannot be decomposed as the finite superposition of continuous functions of fewer variables. (More precisely, Hilbert's assumption was formulated as a concrete minor hypothesis, namely, that there existed at least one such continuous function of three variables, exemplified by $f^7 + xf^3 + yf^2 + zf + 1 = 0$, which could not be decomposed as the finite superposition of continuous bivariate functions.) In 1957 Arnold disproved this hypothesis [1], moreover, in the same year, Kolmogorov [26] proved a general representation theorem with a constructive proof, where the functions in the decomposition were one-dimensional.

Theorem 1. *For all $n \geqslant 2$, and for any continuous real function f of n variables on the domain* [0,1], *$f : [0,1]^n \to \mathbb{R}$, there exist $n(2n+1)$ continuous, monotone increasing univariate functions on* [0,1], *by which f can be reconstructed according to the following equation*:

$$f(x_1, \ldots, x_n) = \sum_{q=0}^{2n} \phi_q \left(\sum_{p=1}^{n} \psi_{pq}(x_p) \right). \tag{1}$$

Here functions ψ_{pq} are universal for the given dimension n, and are independent of f. Only functions ϕ_q depend on f. However, functions ϕ_q and ψ_{pq} are often very complicated and highly nonsmooth, so their construction is difficult.

Kolmogorov's representation theorem was further improved by several authors. In 1965, Sprecher in [39] showed that the unknown approximated mapping could be generated by replacing ψ_{pq} by $\lambda^{pq}\psi_q$ in Eq. (1), where λ is a constant and ψ_q are monotonic increasing functions of the class Lip $[\ln 2/\ln(2n+2)]$ (see Section 2 for denotation). In 1966, Lorentz in [32] proved that functions ϕ_q could be replaced by only one function ϕ. However, these functions were still highly nonlinear and difficult to calculate with.

In 1980 De Figueiredo showed that Kolmogorov's theorem could be generalized for multilayer feedforward neural networks, and so these could be considered to be universal approximators [11]. From the late 1980s several

authors proved that different types of neural network possessed the universal approximation property (for further details see Section 3.1 and [4,17,28]).

Similar results have been established from the early 1990s in fuzzy theory. These results (see e.g. [7,27,45] and also Section 3.2) claim that different fuzzy reasoning methods are capable of approximating arbitrary continuous function on a compact domain with any specified accuracy.

It can be shown that the neural network models as well as the fuzzy ones have exponential complexity in terms of the number of variables of the original function. This means that in the neural network context, the number of units in the hidden layer(s), or in the fuzzy context, the number of rules in the rule base grows exponentially as the approximation error tends to zero. (In this paper we use the common term *building block* or *building unit* for the fuzzy sets in the fuzzy rules, and for the hidden neurons in neural networks.) This exponentiality cannot be eliminated, so the universal approximation property of these uncertainty based approaches cannot be straightforwardly exploited for practical purposes.

Moreover, for some special fuzzy systems (e.g. Sugeno and Takagi–Sugeno type controllers) [35,41] it is shown that if the number of the building blocks (rules here) is bounded, the resulting set of functions is *nowhere dense* in the space of approximated functions (for terminology see Section 2), i.e., this is an "almost" discrete set. According to the opinion of some researchers [23] analogous results should hold for most fuzzy and neural systems.

These mutually contradictory results naturally raise the question, to what extent the approximation should be accurate. From the practical point of view it is enough to have an "acceptably" good approximation, where the given problem determines the factor of acceptability in terms of the accuracy ε. Hence the task is to find a possible trade-off between the specified accuracy and the number of building units, which enables tractable approximation in time.

In the last decade a few results have been published in the neural network and fuzzy theory fields determining the number of building blocks as a function of the accuracy. In neural network theory Blum and Li [4] determined an upper bound for the number of McCulloch–Pitts (Mc–P) units of four-layer networks. For multilayer feedforward network Kůrková [28] gave an upper bound for the number of units having generalized sigmoidal activation functions. In fuzzy theory for some practical membership function shapes Kóczy and Zorat [24] determined the approximation error (worst-case) taking into account the parameterized cost functions of approximation inaccuracy and computation. For (Takagi)–Sugeno type controllers Ying and co-workers [10,53] and Zeng et al. [55] determined necessary and sufficient conditions on minimal system configuration for any specified accuracy. For the rather special stabilized KH controller even the optimal rate of convergence (saturation problem) is determined under certain conditions [40].

This paper is organized as follows. Section 2 recalls the mathematical notions used in the paper. Section 3 gives overviews on the "positive" universal approximation results achieved so far in the neural network field (Section 3.1), and in fuzzy theory (Section 3.2). Section 4 presents the negative results: the discontinuity of best approximation performed by certain neural models, and the nowhere denseness theorems for some fuzzy controllers if bounded numbers of rules are used. In Section 5 the performance of the approximation investigated is addressed in terms of the approximation rate. Section 6 introduces the results achieved on the number of building units as a function of accuracy. Finally, Section 7 presents some conclusions.

2. Preliminaries

In soft computing contexts, we are only interested in approximate realization of functions. The universal approximation property hence can be formulated in topological terms by means of the closure of a set and a dense subset. The symbol $\overline{Y}$ denotes the *closure* of a subset Y of a topological space X, which is the set of points in X having the property that every neighbourhood of such a point has a nonempty intersection with Y:

$$\overline{Y} = \{x \in X | \forall \varepsilon > 0,\ C_\varepsilon(x) \cap Y \neq \emptyset\}.$$

Here $C_\varepsilon(x)$ denotes the ε neighbourhood of the point $x \in X$. The points of $\overline{Y}$ are called *inner points* of Y. The subset Y is called *dense* in the topological space X, if $\overline{Y} = X$. The subset Y is *nowhere dense* in X, if $\overline{Y} = \emptyset$ (i.e. there is no inner point in Y), or equivalently, the complement of Y lies dense in X. By $C(X)$ we denote the set of all continuous real-valued functions on X.

In these contexts we are dealing with topologies defined by metrics. The most often used metrics are the supremum and the L_p norms. The *supremum norm* defined by $\|f\| = \sup\{|f(x)|, x \in X\}$ induces the supremum metrics $\|f - g\| = \sup\{|f(x) - g(x)|, x \in X\}$ and the derived topology is called the *topology of uniform convergence*. This is suitable for applications where the system has to perform simultaneously well for all input vectors from a set X. If some input environment measure μ, expressing the importance of various input vectors can be specified, then it is more convenient to use L^p *norms* ($p \geqslant 1$), defined for a measure μ (given on a set X of input vectors) on the set $L^p_\mu(X)$ of all real functions f on X for which the Lesbegue integral $\int |f|^p \, \mathrm{d}\mu$ is finite by $\|f\|_{p,\mu} = (\int |f|^p \, \mathrm{d}\mu)^{1/p}$ with the induced pseudometrics $\varrho_{p,\mu}(f, g) = (\int |f - g|^p \, \mathrm{d}\mu)^{1/p}$. The most popular choice is $p = 2$, which corresponds to the mean square error.

We remark that results achieved by using the supremum norm can be carried over for L^p norm as, e.g. for all reasonable measures on the n-dimensional unit cube (I^n) the space $C(I^n)$ lies dense in $L^p_\mu(I^n)$ with the topologies induced by $\varrho_{p,\mu}$.

Hence approximation capabilities with respect to the supremum norm guarantee the same capabilities with respect to all reasonable input environment measures. Henceforth, we shall use the supremum norm due to its simplicity, and without the loss of generality we restrict ourselves to the approximation of functions of I^n, because with a proper linear transformation every compact domain can be projected to each other.

The function $\omega_f(0,\infty) \to \mathbb{R}$ is called the *modulus of continuity* of the function $f : I^n \to \mathbb{R}$ if

$$\omega_f(\delta) = \sup\{|f(x_1,\ldots,x_n) - f(y_1,\ldots,y_n)| : (x_1,\ldots,x_n),(y_1,\ldots,y_n) \in I^n\},$$

where

$$|x_i - y_i| < \delta \quad \text{for every } i = 1,\ldots,n.$$

The function $f : I^n \to \mathbb{R}$ is called *Lipschitz continuous* with *Lipschitz coefficient* L (notation: $f \in \mathrm{Lip}L$) if

$$|f(x_1,\ldots,x_n) - f(y_1,\ldots,y_n)| \leqslant L(|x_1 - y_1| + \cdots + |x_n - y_n|) \quad \text{for all } (x_1,\ldots,x_n),(y_1,\ldots,y_n) \in I^n. \tag{2}$$

Finally, using the above-mentioned terminology the universal approximation property can be interpreted as the set of approximating functions lies dense in the set of approximated functions $C[X]$ w.r.t. a proper norm, where X is a compact domain. ($C[X]$ is usually the set of continuous functions.) It is worth remarking that the formalism of topology is convenient to prove only the existence of the approximation, and does not provide a way to construct the approximation itself.

We assume that the Reader is familiar with the basic neural network (multilayer feedforward network with various activation functions) and fuzzy controller types (Sugeno, Takagi–Sugeno, Mamdani), therefore we omit their definition and detailed description. For further details see e.g. [13,15,18].

3. Positive results on universal approximation

3.1. Universal approximation in neural networks

The first result in neural networks were based on Kolmogorov's representation theorem [11,14]. In the latter paper, Hecht-Nielsen reformulated Sprecher's theorem stating that any continuous function defined on I^n could be implemented exactly by a three-layered network with $2n+1$ units in the hidden layer with transfer function $\lambda^{pg}\psi_q$ ($p = 1,\ldots,n$; $q = 1,\ldots,2n+1$) from the input to the hidden units and ϕ from the hidden to the output layer. He showed that the universality of ψ_q could be exploited to approximate functions of higher dimension using Kolmogorov's representation theorem. Hence, any

functions fulfilling the above conditions can be approximated by a neural network having Kolmogorov functions as activation functions.

However, the functions ϕ and ψ_q $(q = 1, \ldots, 2n+1)$ are highly nonsmooth (they can have even fractal graphs—probably a reason of the failure of Hilbert's intuition), which are far from being typical activation function. According to Poggio and Girosi [37], Kolmogorov's result is not relevant for neural networks because in a Kolmogorov network units have "wild" and complex functions.

The first result claiming that a general class of neural networks with "normal" activation function are capable for universal approximation was published by Hornik et al. in [17]. They proved that any continuous function $f \in C(I^n)$ can be approximated arbitrarily well in the supremum norm by a three-layered feedforward network with semilinear hidden units using a threshold function and one linear output unit, formally,

$$\left| f(x) - \sum_{i=1}^{m} w_i g\left(\sum_{j=1}^{n} a_{ij} x_j + c_i \right) \right| < \varepsilon, \tag{3}$$

where real numbers w_i and a_{ij} are the weights and c_i the thresholds. Semilinear units have the form $g(L(x) - b)$, where $L(x)$ is linear in x, and the function g is a monotone real function with the limits

$$\lim_{x \to -\infty} g(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} g(x) = 1, \tag{4}$$

usually called hard limiter. Hence, g can be a sigmoidal function, e.g. $\sigma(z) = 1/(1 + e^{-z})$ or the Heaviside function

$$H(z) = \begin{cases} 1 & \text{if } z > 0, \\ 0 & \text{if } z \leqslant 0. \end{cases} \tag{5}$$

Additionally, in [17] it is also proved that any function $f \in L_2(I^n)$ (the space of L_2-integrable functions; see Section 2) can be approximated with arbitrary accuracy with respect to the L_2 norm by such three-layered feedforward networks.

The Stone–Weierstrass theorem was employed in both proofs. The results can be formulated for $f \in C(I^n)$ or $f \in L_2(I^n)$ as follows. Any such f can be approximated by a multivariate finite trigonometric sum, e.g., for dimension $n = 2$, such

$$\sum_{p,q=1}^{N} A_{pq} \cos px \cos qy. \tag{6}$$

The expression (6) can be written, by means of basic trigonometric expressions, as linear combination of terms $\cos Z_i$, where Z_i is a linear function of x and y. $\cos Z_i$, being continuous, can be approximated in the form (3), and hence, also

the original trigonometric sum, Eq. (6). From this, it is easy to see that any $f \in C(I^n)$ or $f \in L_2(I^n)$ can be approximated arbitrarily well by three-layered networks with respect to the supremum or the L_2 norm.

Similar results have been achieved concerning the approximation capabilities of feedforward networks by, e.g., Cybenko [9] and Funahashi [12] (Hornik's and Funahashi's results were slightly extended by Castro et al. in [8] for the case when also squashing function was used). They also used semilinear units, and, for the most part, monotone threshold functions. These results (including the one by Hornik et al.) are not constructive in a simple way. The proofs depend on existential theorems, e.g., on the Stone–Weierstrass [17], or on the Hahn–Banach theorem [9]. The first results providing constructive approximations were given in [4,28] to be discussed in Section 6.

In [16], Hornik substantially generalized the set of the activation functions employed so far. He proved that whenever the activation function g in (3) is bounded and nonconstant, then the multilayer feedforward network is capable of approximate every $f \in L^p_\mu(I^n)$, (and if additionally, g is continuous then every $f \in C(I^n)$), arbitrarily well with respect to the corresponding norm using sufficiently many hidden neurons. It can be concluded that it is not the specific choice of the activation function, but rather the multilayer feedforward architecture itself which gives neural networks the potential to be universal approximators.

3.2. Universal approximation in fuzzy systems

The first publication concerning approximation capabilities of fuzzy systems came from the practical side, showing experimentally that if an arbitrary continuous nonlinear function is given on a compact universe of discourse, it is possible to approximate it arbitrarily well by a fuzzy control system [22]. At the same time and soon after exact mathematical results were also proposed.

In [27], Kosko proved that Mamdani type controllers (more precisely, additive fuzzy systems, which in terms of structure and parameters are similar to the Mamdani method) could uniformly approximate $f \in C(X)$, where X is a compacta. (Here we remark that although the theorem is claimed to be valid for any compacta $X \subset \mathbb{R}^n$, the proof applies for one-dimensional input spaces.) The number of rules can be estimated by means of the minimal distance of the centers of two adjacent consequent sets: if they are denoted by y_i and y_{i+1} then for ε-approximation:

$$|y_i - y_{i+1}| < \frac{\varepsilon}{2p-1}, \tag{7}$$

where p is the number of the maximal overlapping antecedents over X, usually 2 (for one-dimensional input). From here, if the prescribed accuracy is ε, the number of rules in the base should be

$$|R| \geqslant \frac{|X|}{\varepsilon}$$

even in the one-dimensional input case. From here this is clear that for arbitrarily good approximation the number of the rules is *not bounded*, however, for a prescribed ε the number of rules can be estimated using inequality (7).

At the same time, the results of Wang [45,48] showed that a different type of fuzzy rule based system has similar properties. In [45], he investigated fuzzy systems with multiple input single output rules, with everywhere positive exponential (Gaussian) membership functions over all the input domain and also for the rule consequents, with Larsen inference algorithm, and finally with the centroid defuzzification method. He proved that the set of the described fuzzy controllers could approximate continuous functions with arbitrary accuracy with respect to the supremum norm. He also applied the Stone–Weierstrass theorem as Hornik et al. in [17] to show that the set of input–output functions of the above controllers lie dense in $C(I^n)$.

Unfortunately the same technical difficulty arises here as in the case of Kosko's theorem: the number of the rules in the base is not bounded. In addition, even the supports of the terms in the rules are not bounded (identical with the universe of discourse).

Next, numerous authors contributed to this topic showing that the universal approximation property holds for various type of fuzzy systems (see e.g. [5–7,36]). The most general paper was published by Castro [7]. It provides a proof for each fixed fuzzy logic belonging to a wide class of fuzzy logics (the widest among the results referred to), and for each fixed type of membership function belonging to a wide class of membership functions. The proof shows that fuzzy logic control systems of the aforementioned types, equipped with the center of area defuzzification are capable of approximating any real continuous function on a compact set to arbitrary accuracy. Although, as it is remarked in the conclusion of Castro's paper [7], neither a way of construction, nor the required number of rules are given; the latter remains unbounded.

Recently it was shown that even more general fuzzy systems that were not included in Castro's result, such as the one that applies the more general uninorm operation instead of t-norm and t-conorm in the inference engine, possess the universal approximation property [50]. Another example is the fuzzy system based upon genuine many-valued implications which was proved also to be universal approximator in [30,31].

The results on the approximation property of fuzzy controllers were criticized from another practical view point in [3]. It was pointed out that if we intended to design a controller, the above-mentioned results did not help much being purely existential in nature, and providing no hint at all as regards to which controller should have been chosen to approximate a particular function.

In [3] the authors dealt with Sugeno controllers, and investigated what input–output functions could be modelled such fuzzy systems. It turned out that in one dimension with just two rules having normalized fuzzy sets as antecedents any continuous function can be represented. (In the multi-dimensional case considerably more rules are required.)

Although this surprisingly powerful statement uses only two membership functions, the result has a weak point: the convexity of the terms are not guaranteed, although convexity plays a central role in natural (such as linguistic) reasoning. Hence, this method also does not offer reduction in the computational complexity as the nonlinearity of f is simply transferred into the membership functions. Moreover, the membership function of the antecedents must be constructed by a direct transformation of the function f. This means that if no exact information is available on the input–output function of the system under control, as in most practical cases, this result does not help in the design of the parameters of the fuzzy controller approximating an unknown function f.

While the first papers on universal approximation dealt particularly with Mamdani controllers, from the mid 1990s the same results were obtained for Takagi–Sugeno controllers [5,51–53,55]. In [5] two-input one-output TS fuzzy systems with a linear defuzzifier (weighted sum) are found to be universal approximators. In [51,52] TS fuzzy systems with proportional linear consequent functions, and in [53] with full-overlapped membership functions are proved to have the universal approximator property. While Buckley's result [5] is only purely existential, Ying [51–53] deals also with the design of a TS system in terms of the determination of membership functions and the number of the rules, and given the approximated continuous function and the desired accuracy. He calls these properties *sufficient conditions*. In [55] two other sufficient conditions are formalized, and a comparative study is carried out. For further details see Section 6.

In [46,47,49] certain hierarchical fuzzy systems were investigated: the additive n-input hierarchical fuzzy system that comprised $n-1$ two-input fuzzy systems invented by Raju et al. [38]. Such systems were proven also to be universal approximators, where the two-input fuzzy systems were implemented by TSK fuzzy systems. Other special hierarchical fuzzy systems were found also to be universal approximators, see e.g. [20].

Beside the fuzzy inference techniques on dense rule bases, identical results have been published recently for the general KH interpolation [40,44] and for its modification [43] operating on sparse rule bases. These results show that even if we relax the condition that ensemble of membership functions cover fully the possible input intervals in the rule base, the approximation capabilities of fuzzy systems remain unchanged, i.e. universal approximation property holds.

4. Negative results on universal approximation

4.1. Discontinuity of the best approximation

If for any function, there is a choice of network parameterizations (not necessary unique) producing an approximation with the minimum error, then we call this the *best approximation* property. In [29] the authors showed that the following two classes of one-hidden-layer feedforward networks possess this property. The first class contains perceptrons with an activation function in the hidden layer of the form:

$$\sum_{i=1}^{m} w_i \psi(\mathbf{a}_i \cdot \mathbf{x} + b_i), \quad \mathbf{a}_i, \mathbf{x} \in \mathbb{R}^n; \;\; w_i, b_i \in \mathbb{R}, \tag{8}$$

where m, the number of hidden units is bounded, and there also exists an upper bound, c for the network parameters, i.e. $|w_i| \leqslant c$, $|b_i| \leqslant c$ and $\|\mathbf{a}_i\| \leqslant c$ for every $i \in [1, m]$. The second class is similar with radial activation functions in the hidden units with the analog upper bounds.

However, in [21] they showed that the most of these network, e.g. Heaviside perceptrons and Gaussian radial-basis-networks, the best approximation with bounded number of hidden units can not be achieved in a continuous way, i.e. the best approximation operator is not continuous (for further details see the paper).

This has serious practical consequences: the stability of the computation cannot be guaranteed, even under "low amplitude" assumptions. On the other hand, the suitability of an approximation scheme can be measured by the worst-case approximation error. So as a theoretical consequence, the estimation of worst-case approximation cannot exploit the continuity of the approximation operator, and this machinery cannot be applied to neural networks.

4.2. Nowhere denseness theorems

It has also been investigated how the set of approximated functions changes if the number of rules is restricted in a fuzzy rule base. This condition apart from the computational reasons, can be motivated to keep one of the inherent features of fuzzy systems in the original sense of Zadeh [54], that is, that they can be characterized by a semantic relying on linguistic terms.

In [35], Moser proved that the set of Sugeno controllers with multiple inputs, having the number of rules restricted on each input space, lies *nowhere dense* in the space of continuous functions with respect to the supremum norm. In other worlds, it means that this set is "almost discrete" in the set of approximated function.

This result was later generalized for the set of, so-called, T-controllers [41]. Here T stands for tensor product consequents; i.e., when the consequents of rules are bounded tensor products of univariate functions. This set includes Takagi–Sugeno controllers of any order. In [41] the nowhere denseness is proved for this larger class of controllers. Recently, in [42] this result was carried over for control problem of dynamic systems being described by a polytopic model.

In [23] authors also question that universal approximation is the answer for the success of soft computing techniques: in the paper it is shown that simple crisp expert systems also possess the universal approximation property if the grid (or knot) points are positioned densely enough.

These results point out (in accordance with the presented results for neural networks by Hornik [16]) that not the great variety of design parameters but rather the number of rules used in the controllers is the reason of the universal approximation property of various fuzzy systems. It is expected that these results can be extended for most fuzzy and neural systems [23].

5. The rate of approximation

Having an approximation scheme, the natural question arises: how good is the actual scheme in terms of approximation speed, i.e. how fast the approximating function converges to the approximated one depending on the number of building blocks. The rate of approximation is usually measured in terms of the order of integrated squared error. The determination of the optimal approximation rate is called the saturation problem. We remark that classical approximation processes (polynomial, spline, trigonometric) are all saturated with order $O(m^{-1})$, m being the number of basis functions, with exponential numbers of parameters. Therefore we expect a similar saturation order for neural networks.

Barron pointed out in [2], that this rate of approximation can be achieved with feedforward one-hidden-layer neural networks and sigmoidal units (8). Moreover, the total number of parameters used in the network is $(n+2)m$, which is considerably smaller than in the classical case. It is assumed in [2], that the class of functions to be approximated possesses a smoothness property which is expressed in terms of the Fourier transform. In particular, the condition is the boundedness of the first moment of the magnitude distribution of the Fourier transform.

Jones [19] obtained the same approximation rate and properties for networks with sinusoidal units

$$\sin\left(\pi \sum_{i=1}^{m} w_i x + a_i\right)$$

and a similar smoothness restriction on the class of approximated functions. Other authors achieved similar results, proposing different conditions for the smoothness property. The common point in these papers is that some condition is posed on the class of approximated functions to achieve a good rate of approximation. The set of continuous functions is too general to obtain analog results.

In Barron's paper [2], the author gave a lower bound for the approximation rate (i.e. determined the theoretical saturation order) for every approximation scheme based on linear combinations of fixed type basis functions [33,34]. The approximation rate cannot be smaller order than $(1/m)^{2/n}$. This vanishingly small rate is the "curse of dimensionality" for the above class of approximation schemes.

6. Constructive results

6.1. Results for neural networks

The first results providing constructive approximations were given in [4,28]. In [4], the authors showed that four-layered feedforward networks with two-hidden layers of semilinear units and with a linear output unit were universal approximators. They used Mc–P units having Heaviside activation functions (5) in the hidden layers. A four-layered network with Mc–P units are able to implement any "simple" function (being a generalized notion for piecewise-constant functions on a compacta). Exploiting the fact that the set of simple functions lies dense in $C(I^n)$, the universality of the Mc–P network can be obtained straightforwardly. Moreover, an estimation for the upper bound of the number of Mc–P units required for a prescribed accuracy has been given in [4], if some global information about the approximated function is known. This can be either the modulus of continuity, $(\omega_f(\varepsilon))$, Lipschitz constant of f, L, or an upper bound for the Jacobian derivative, $\|f'\| \leqslant k$. According to their result, the number of hidden units can be estimated by $m^n + 2n(m+1)$, where n denotes the dimension, and m depends on ε as $m \geqslant \lceil 1/\omega_f(\varepsilon) \rceil$, $m \geqslant L/\varepsilon$, or $m \leqslant k/\varepsilon$, respectively.

In [28], Kůrková investigated the approximation capabilities of multilayer feedforward networks with sigmoidal activation functions. These networks implement functions of the form $\sum_{i=1}^{m} w_i \sigma(b_i x - c_i)$, which are called *staircase-like functions*. The author showed that, taking advantage of the fact that staircase-like functions are universal approximators in the space $C(I^n)$, Kolmogorov's representation theorem can be carried over for multilayer feedforward networks. So for any function $f \in C(I^n)$, it can be approximated with arbitrary accuracy by using sigmoidal functions ϕ and ψ_q $(q = 1, \ldots, 2n+1)$ in Eq. (1). Kůrková also determined the number of units needed for a prescribed

ε-approximation: in the first hidden layer $nk(k+1)$ and $k^2(k+1)^n$ in the second one. Here $k \geqslant 2n+1$, $n/(k-n)+v \leqslant \varepsilon/\|f\|$ and $\omega_f(1/k) \leqslant v(m-n)/(2m-3n)$ with some positive real v.

Not surprisingly, the number of hidden units needed for a good approximation is very large. It depends exponentially on the dimension of the approximated function. However, these results give a method for the construction of universally approximating multilayer feedforward networks. Moreover, in the latter case (Kůrková's result), the only adjustable weights correspond to the transfer from the second hidden layer to the output layer. Since these weights appear linearly in the parameterized expression, the problem of learning can be solved by linear regression.

6.2. Result for fuzzy systems

The results on the fuzzy systems field usually follow their neural network counterparts with a few years of delay. The first results were achieved in combination with a practical problem.

In [24,25] the authors determined the sufficient number of rules if practically important, nevertheless special shaped membership functions being triangular or trapezoidal, are applied. The optimal rule base is sought in terms of minimal calculation time in a target tracking problem (cat and mouse problem). The total tracking time consists of two components: the inference time where the hyperinterval of the current location of the target is predicted, and the action (search) time when that selected hyperinterval is searched. Intuitively, the finer the rule base (and as a consequence the smaller the determined hyperinterval), the greater the inference time and the smaller the action time. The optimal rule base size is determined for specific rule base models. For instance, when there are n inputs, one output, and the observation is crisp, then the optimal number of terms in each dimension is the closest integer number to

$$t = \sqrt[n+1]{\frac{2^n c_1}{c_0}} + 1,$$

where c_0 and c_1 are search cost parameters. Parameter c_1 involves the term of accuracy: this is the constant cost factor of searching the unit length of the output. Here the values of the constants depend strongly on the actual target tracking system. As an example the authors derived $t = 28$ for the following settings: $n = 2$, $c_0 = 1$, $c_1 = 5000$. This result in $t^2 = 784$ for the optimal number of rules.

Here the information on the approximated function is transferred to the proper selection of the search parameters. The determination of the unit length of the output presumes former knowledge about the behaviour of the target (mouse) and the seeker (cat). However, this practical example indicates that if

the approximation problem is placed in a special application framework, then the solution can be obtained more easily, because the significant features of the problem are known.

In [10] the authors give a constructive approximation with a MISO Mamdani fuzzy system. They pose a condition reasonable in practice on the class of approximated functions: it is assumed that the functions have finite number of extrema (M) on the compact domain they are approximating. The authors refer to the fuzzy system applied as of Mamdani type having a singleton output, therefore it should be considered, in fact, a Sugeno system. The further parameters of the system used are the following: very general shaped membership functions, product–sum inference, with parameterized defuzzification (including centroid and mean of maxima). They establish the connection between the prescribed accuracy and the minimal number of rules. The number of rules directly depends on M. This means that if ε is very small, keeping M small, then the number of rules remains relatively small. This insightful analysis provides an explanation for the fact that the majority of fuzzy models use just a small number of rules to achieve successful applications.

In [51–53,55] the authors deduce sufficient conditions on the number of rules for a prescribed accuracy in the case of Takagi–Sugeno and Sugeno controllers. In all papers they take advantage of a two-step approximation procedure, which is similar to that applied for two-hidden layer neural nets by Blum and Li [4] and Kůrková [28] (see previous subsection). The approximations exploit the Weierstrass theorem, which states that on a compact domain multivariate polynomials of a finite degree (q) can uniformly approximate any continuous function to an arbitrary accuracy (ε_1). Using this it is shown that polynomials can be approximated by TS and Sugeno controllers arbitrarily well ($\varepsilon - \varepsilon_1$), and finally the triangular inequality guarantees the required result. In [55] they obtained

$$n_0 > \frac{1}{\varepsilon - \varepsilon_1} \sum_{i=1}^{n} \left\| \frac{\partial P_q}{\partial x_i} \right\|_\infty - 1 \tag{9}$$

for the sufficient number of the rules in each dimension $i \in [1, n]$ in the case of Sugeno fuzzy systems, and

$$n_0 > \sqrt{\frac{1}{2(\varepsilon - \varepsilon_1)} \sum_{i=1}^{n} \sum_{k=1}^{n} \left\| \frac{\partial^2 P_q}{\partial x_i \partial x_k} \right\|_\infty} - 1 \tag{10}$$

for TS fuzzy systems. The x_i are the inputs, $i \in [1, n]$, and P_q is the approximating multivariate polynomial of degree q:

$$P_q(\mathbf{x}) = \sum_{d_1=0}^{m_1} \sum_{d_2=0}^{m_2} \cdots \sum_{d_n=0}^{m_n} \beta_{d_1 d_2 \ldots d_n} x_1^{d_1} x_2^{d_2} \cdots x_n^{d_n}.$$

These results were compared with the Ying's appropriate results, which are

$$n_0 \geqslant \frac{1}{\varepsilon} \sum_{d_1=0}^{m_1} \sum_{d_2=0}^{m_2} \cdots \sum_{d_n=0}^{m_n} \left(\beta_{d_1 d_2 \ldots d_n} x_1^{d_1} x_2^{d_2} \cdots x_n^{d_n} \sum_{i=1}^{n} d_i \right) \tag{11}$$

for Sugeno controllers, and

$$n_0 \geqslant \frac{|\beta_{1,0}| + |\beta_{0,1}| + \sum_{d_1=0}^{m_1} \sum_{d_2=0}^{m_2} |\beta_{d_1 d_2}| (2^{d_1+d_2} - 1)}{\varepsilon - \varepsilon_1} \tag{12}$$

for TS controllers (two-input one-output case).

In the comparison of (9) and (11) we can conclude that the former (Zeng's result) is advantageous when the dimension is relatively small, and the latter (Ying's result) is better in high-dimensional spaces, because (9) utilizes the extremum of the partial derivatives of a polynomial. Similar statements hold for the comparison of (10) and (12). To illustrate the difference $f(x_1, x_2) = \mathrm{e}^{x_1+x_2}$ can be approximated with 0.1 accuracy by 14 and 1369 rules in each dimension according to (9) and (11), respectively.

We should remark that (10) surpasses (12) asymptotically if ε_1 is very small. In this case the former yields approximation order $\mathrm{O}(\varepsilon^{-1/2})$ and the latter $\mathrm{O}(\varepsilon^{-1})$.

7. Conclusions

In this paper we summarized the state-of-the-art in the field of universal approximation property in soft computing techniques. The first results in both disciplines were purely existential, that is, neither the way of construction of a universal approximator, nor the required number of building blocks (hidden neurons; rules, resp.) were given. After some critical papers showing the results of the previous papers mainly rely on unbounded number of rules and not on the descriptive power of soft computing techniques, there was great demand among researchers for finding the functional dependency between the prescribed accuracy and the number of building units.

As a result, several papers were published (mostly in the neural network field) determining an upper bound for the building blocks. These bounds are based on elementary characteristics of the function being approximated, like the modulus of continuity, the Lipschitz constant or some smoothness property of the function or of its derivative(s). In the case when such characteristics are not known, we cannot estimate the number of building units.

Acknowledgements

Research funded by the Australian Research Council large grants scheme and partially supported by the Hungarian Scientific Research Fund Grants No. D34614, T34212 and T34233.

References

[1] V.I. Arnold, On functions of three variables, Dokl. Akad. Nauk USSR 114 (1957) 679–681.
[2] A.R. Barron, Universal approximation bounds for superpositions of sigmoidal functions, IEEE Trans. Inform. Theory 39 (3) (1993) 930–945.
[3] P. Bauer, E.P. Klement, A. Leikermoser, B. Moser, Modeling of control functions by fuzzy controllers, in: H. Nguyen, M. Sugeno, R. Tong, R.R. Yager (Eds.), Theoretical Aspects of Fuzzy Control, Wiley, New York, 1995, pp. 91–116.
[4] E.K. Blum, L.K. Li, Approximation theory and feedforward networks, Neural Networks 4 (4) (1991) 511–515.
[5] J.J. Buckley, Sugeno type controllers are universal controllers, Fuzzy Sets Syst. 53 (1993) 299–304.
[6] J.J. Buckley, System stability and the fuzzy controller, in: H. Nguyen, M. Sugeno, R. Tong, R.R. Yager (Eds.), Theoretical Aspects of Fuzzy Control, Wiley, New York, 1995, pp. 51–63.
[7] J.L. Castro, Fuzzy logic controllers are universal approximators, IEEE Trans. SMC 25 (1995) 629–635.
[8] J.L. Castro, C.J. Mantas, J.M. Benítez, Neural networks with a continuous squashing function in the output are universal approximators, Neural Networks 13 (2000) 561–563.
[9] G. Cybenko, Approximation by superposition of sigmoidal functions, Math. Control Signals Syst. 2 (1989) 303–314.
[10] Y.-S. Ding, H. Ying, S.-H. Shao, Necessary conditions on minimal system configuration for general MISO mamdani fuzzy systems as universal approximators, IEEE Trans. Syst. Man Cybernet. Part B 30 (6) (2000) 857–864.
[11] R.J.P. De Figueiredo, Implications and applications of Kolmogorov's superposition theorem, IEEE Trans. Automat. Control (1980) 1227–1230.
[12] K.-I. Funahashi, On the approximate realization of continuous mappings by neural networks, Neural Networks 2 (1989) 183–192.
[13] S. Haykin, Neural Networks: A Comprehensive Foundation, second ed., Prentice Hall, Upper Saddle River, NJ, 1999.
[14] R. Hecht-Nielsen, Kolmogorov's mapping neural network existence theorem, in: Proc. of the International Conference on Neural Networks, New York, vol. III, 1987, pp. 11–14.
[15] H. Hellendoorn, D. Driankov, M. Reinfrank, An Introduction to Fuzzy Control, Springer, Berlin, 1993.
[16] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Networks 4 (1991) 251–257.
[17] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (1989) 359–366.
[18] J.-S.R. Jang, C.-T. Sun, E. Mizutani, Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice Hall, Upper Saddle River, NJ, 1997.
[19] L.K. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, Ann. Stat. 20 (1992) 608–613.
[20] M.G. Joo, J.S. Lee, Universal approximation by hierarchical fuzzy system with constraints on the fuzzy rule, Fuzzy Sets Syst. 130 (2002) 175–188.

[21] P.C. Kainen, V. Kůrková, A. Vogt, Approximation by neural networks is not continuous, Neurocomputing 29 (1999) 47–56.
[22] S. Kawamoto, K. Tada, N. Onoe, A. Ishigame, T. Taniguchi, Construction of exact fuzzy system for nonlinear system and its stability analysis, in: Proc. of the 8th Fuzzy System Symposium, Hiroshima, 1992, pp. 517–520 (in Japanese).
[23] E.P. Klement, L.T. Kóczy, B. Moser, Are fuzzy systems universal approximators?, Int. J. Gen. Syst. 28 (2–3) (1999) 259–282.
[24] L.T. Kóczy, A. Zorat, Fuzzy systems and approximation, Fuzzy Sets Syst. 85 (1997) 203–222.
[25] L.T. Kóczy, A. Zorat, T.D. Gedeon, The cat and mouse problem—optimizing the size of fuzzy rule bases, in: Proc. of the Conf. on Current Issues of Fuzzy Technologies'95, Trento, Italy, 1995, pp. 139–151.
[26] A.N. Kolmogorov, On the representation of continuous functions of many variables by superpositions of continuous functions of one variable and addition, Dokl. Akad. SSSR 114 (1957) 953–956 (in Russian).
[27] B. Kosko, Fuzzy systems as universal approximators, in: Proc. of the IEEE Int. Conf. on Fuzzy Systems, San Diego, 1992, pp. 1153–1162.
[28] V. Kůrková, Kolmogorov's theorem and multilayer neural networks, Neural Networks 5 (1992) 501–506.
[29] V. Kůrková, Approximation of functions by perceptron networks with bounded number of hidden units, Neural Networks 8 (5) (1995) 745–750.
[30] Y.-M. Li, Z.-K. Shi, Z.-H. Li, Approximation theory of fuzzy systems based upon genuine many-valued implications—SISO cases, Fuzzy Sets Syst. 130 (2002) 147–157.
[31] Y.-M. Li, Z.-K. Shi, Z.-H. Li, Approximation theory of fuzzy systems based upon genuine many-valued implications—MIMO cases, Fuzzy Sets Syst. 130 (2002) 159–174.
[32] G.G. Lorentz, Approximation of Functions, Holt, Reinhard and Winston, New York, 1966.
[33] D.F. McCaffrey, A.R. Gallant, Convergence rate for single hidden layer feedforward networks, Technical report, RAND Corp., Santa Monica, CA and Department of Statistics, North Carolina State University, 1992.
[34] H.N. Mhaskar, C.A. Micchelli, Approximation by superposition of a sigmoidal function, Adv. Appl. Math. 13 (1992) 350–373.
[35] B. Moser, Sugeno controllers with a bounded number of rules are nowhere dense, Fuzzy Sets Syst. 104 (2) (1999) 269–277.
[36] H.T. Nguyen, V. Kreinovich, On approximations of controls by fuzzy systems, Technical Report TR 92-93/302, LIFE Chair of Fuzzy Theory, Tokyo Institute of Technology, Tokyo, 1992.
[37] T. Poggio, F. Girosi, Representation properties of networks: Kolmogorov's theorem is irrelevant, Neural Comput. 1 (4) (1989) 465–469.
[38] G.V.S. Raju, J. Zhou, R.A. Kisner, Hierarchical fuzzy control, Int. J. Control 54 (1991) 1201–1216.
[39] D.A. Sprecher, On the structure of continuous functions of several variables, Trans. Am. Math. Soc. 115 (1965) 340–355.
[40] D. Tikk, Notes on the approximation rate of KH controllers, Fuzzy Sets Syst., in press [doi:10.1016/S0165-0114(02)00387-1].
[41] D. Tikk, On nowhere denseness of certain fuzzy controllers containing prerestricted number of rules, Tatra Mountains Math. Publ. 16 (1999) 369–377.
[42] D. Tikk, P. Baranyi, R.J. Patton, Polytopic and TS model are nowhere dense in the approximation model space, in: Proc. of the Int. Conf. on Systems, Man and Cybernetics (SMC 2002), Hammamet, Tunisia, October 2002, p. 5.
[43] D. Tikk, P. Baranyi, Y. Yam, L.T. Kóczy, Stability of a new interpolation method, in: Proc. of the IEEE Int. Conf. on System, Man, and Cybernetics (IEEE SMC'99), Tokyo, Japan, vol. III, 1999, pp. 7–9.

[44] D. Tikk, I. Joó, L.T. Kóczy, P. Várlaki, B. Moser, T.D. Gedeon, Stability of interpolative fuzzy KH-controllers, Fuzzy Sets Syst. 125 (1) (2002) 105–119.
[45] L.X. Wang, Fuzzy systems are universal approximators, in: Proc. of the IEEE Int. Conf. on Fuzzy Systems, San Diego, 1992, pp. 1163–1169.
[46] L.X. Wang, Universal approximation by hierarchical fuzzy systems, Fuzzy Sets Syst. 93 (1998) 223–230.
[47] L.X. Wang, Analysis and design of hierarchical fuzzy systems, IEEE Trans. Fuzzy Syst. 7 (5) (1999) 617–624.
[48] L.X. Wang, J. Mendel, Generating fuzzy rules from numerical data with supplications, Technical Report TR USC-SIPI #169, Signal and Image Processing Institute, University of Southern California, 1991.
[49] C. Wei, L.X. Wang, A note on universal approximation by hierarchical fuzzy systems, Inform. Sci. 123 (2000) 241–248.
[50] R.R. Yager, V. Kreinovich, Universal approximation theorem for uninorm-based fuzzy systems modeling, Fuzzy Sets Syst., in press [doi:10.1016/S0165-0114(02)00521-3].
[51] H. Ying, General SISO Takagi–Sugeno fuzzy systems with linear rule consequents are universal approximators, IEEE Trans. Fuzzy Syst. 6 (4) (1998) 582–587.
[52] H. Ying, General Takagi–Sugeno fuzzy systems with simplified linear rule consequents are universal controllers, models and filters, J. Inform. Sci. 108 (1998) 91–107.
[53] H. Ying, Sufficient conditions on uniform approximation of multivariate functions by general Takagi–Sugeno fuzzy systems with linear rule consequents, IEEE Trans. SMC Part A 28 (4) (1998) 515–520.
[54] L.A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes, IEEE Trans. SMC 3 (1973) 28–44.
[55] K. Zeng, N.-Y. Zhang, W.-L. Xu, A comparative study on sufficient conditions for Takagi–Sugeno fuzzy systems as universal approximators, IEEE Trans. Fuzzy Syst. 8 (6) (2000) 773–780.