Họ tên sinh viên: Lâm Xuân Thư
Mã số sinh viên: 20153709 lớp: KSTN-CNTT-K60

Fuzzy Systems with Sigmoid-Based Membership Functions as Interpretable Neural Networks

Lâm Xuân Thư

Viện công nghệ thông tin & Truyền thông, Trường Đại học Bách Khoa Hà Nội,

Giảng viên hướng dẫn: Phạm Văn Hải

TÓM TẮT— *Bài báo cáo này sẽ trình bày một kiến trúc mạng nơ-ron khả hiểu mới. Một mạng nơ-ron với hàm kích hoạt sigmoid được biến đổi thành một toán tử xấp xỉ sigmoid-based, sau đó toán tử này được xấp xỉ bởi một hệ mờ kiểu Takagi-Sugeno. Tất cả quá trình này thể hiện rằng một mạng nơ-ron với các hàm kích hoạt sigmoid có thể được xấp xỉ bởi một hệ mờ Takagi-Sugeno. Do hệ mờ Takagi-Sugeno cung cấp khả năng khả hiểu, trong khi mạng nơ-ron cung cấp khả năng xấp xỉ, nên ta thu được một kiến trúc mạng nơ-ron khả hiểu mới. Tính khả hiểu của hệ mờ Takagi-Sugeno có thể cung cấp khả năng hiểu sâu sắc dữ liệu theo một cách mới, nâng cấp việc ra quyết định.*

I. GIỚI THIỆU

Các tập mờ, hệ mờ và các mạng nơ-ron được kết nối với nhau như là các hệ thuyết khác nhau với Soft Computing và Approximate Reseaning [14-16]. Kết nối chặt chẽ giữa hệ mờ và mạng nơ-ron bị ẩn đi trong các thuộc tính xấp xỉ của chúng. Cả hệ mờ và mạng nơ-ron đều có các thuộc tính xấp xỉ phổ biến [1, 4, 12, 13]. Ngoài ra, còn có nhiều kiến trúc lai, tổng hợp hệ mờ với mạng nơ-ron [5, 6, 8], trong đó thì ANFIS là một trong những kiến trúc được biến đến rộng rãi nhất [6]. Tính khả hiểu của Deep Learning của mạng nơ-ron là một bài toán đang được xuất hiện rất nhiều trong các nghiên cứu. Mạng Radial Basis Function có thể được xem như là vừa là hệ mờ vừa là mạng nơ-ron khả hiểu [7]. Trong bài báo tìm hiểu thì tác giả đề xuất dựa trên sự khác nhau của các hàm sigmoid. Một hàm sigmoid đơn có thể mô hình các quan hệ khác nhau, nhưng chúng ta muốn xem xét tính khả hiểu bằng cách cục bộ hoá các giá trị. Dễ thấy sự khác biệt của hai hàm sigmoid với cùng tham số độ dốc có dạng hình chuông, có thể được xem như một hàm fuzzy membership.

Một ý tưởng khác mà tác giả muốn khám phá là sự xấp xỉ giữa các hệ mờ kiểu Takagi-Sugeno với các mạng nơ-ron. Ý tưởng này xuất phát từ việc chúng ta biết rằng cả hệ mờ và mạng nơ-ron đều có các thuộc tính xấp xỉ phổ biến. Hệ quả là chúng ra có thể tạo ra hệ mờ xấp xỉ mạng nơ-ron và ngược lại. Tác giả cũng nghiên cứu các kiến trúc mạng nơ-ron và các kiến trúc deep learning khả hiểu phù hợp dựa trên phương pháp trên.

II. TỔNG QUAN

Một hệ mờ kiểu Takagi-Sugeno với đầu ra hằng số piece-wise được dựa trên các luật mờ kiểu:

if x is A_i then y is w_i

với i = 1, …, n, Ai là tập mờ, wi là giá trị rõ, x là đầu vào và $x \in R^m$.

Một hệ Takagi-Sugeno có thể được xem như một hàm dạng:

$$TS(\mathbf{x}) = \frac{\sum_{i=1}^{n} A_i(\mathbf{x})\, w_i}{\sum_{i=1}^{n} A_i(\mathbf{x})}$$

Một Adaptive Network-based Fuzzy Inference Systems (ANFIS) là một hệ mờ TS với các thuật toán học được them vào.

Tính khả hiểu của hệ mờ TS được dựa trên luật mờ được nhắc đến ở trên. Dễ dàng nhận thấy rằng các biến thể và các hệ mờ có tính khả hiểu của một xấp xỉ. Nói một cách khác, giả sử ta có hàm f, và biết rằng f(xi) = yi, khi đó một hàng xóm của xi thì sẽ cho giá trị của f gần với yi. Biểu diễn dưới dạng luật mờ ta có:

Nếu x là khoảng xi, thì y là khoảng yi, i = 1, …, n

Mạng nơ-ron được xem như một toán tử xấp xỉ. Xét một mạng nơ-ron với hàm kích hoạt sigmoid $\varphi(x) = \frac{1}{1+e^{-x}}$ như sau:

$$NN(x) = w_0 + \sum_{i=1}^{n} w_i \varphi(a_i(x - b_i))$$

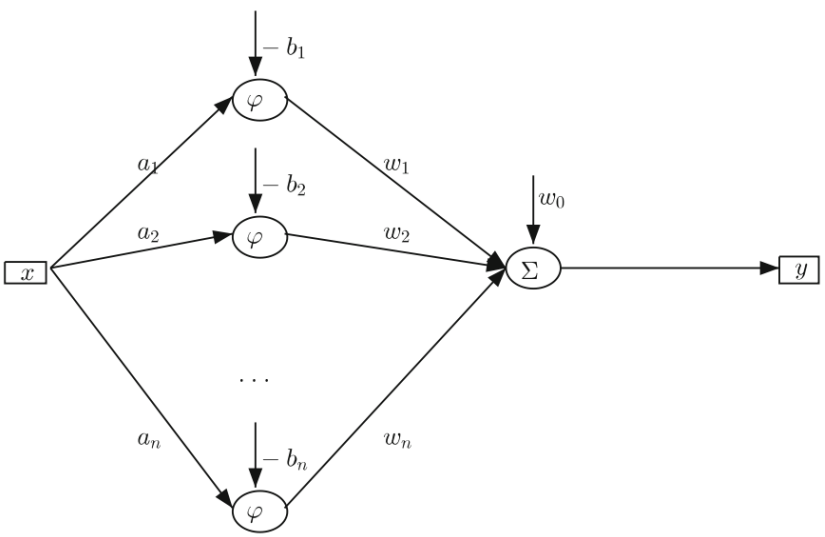
Kiểu mạng nơ-ron này có thể xem như một hệ mờ :

$$\text{if } x \text{ is } A_i \text{ then } y+ = w_i,\ i = 0, \ldots, n$$

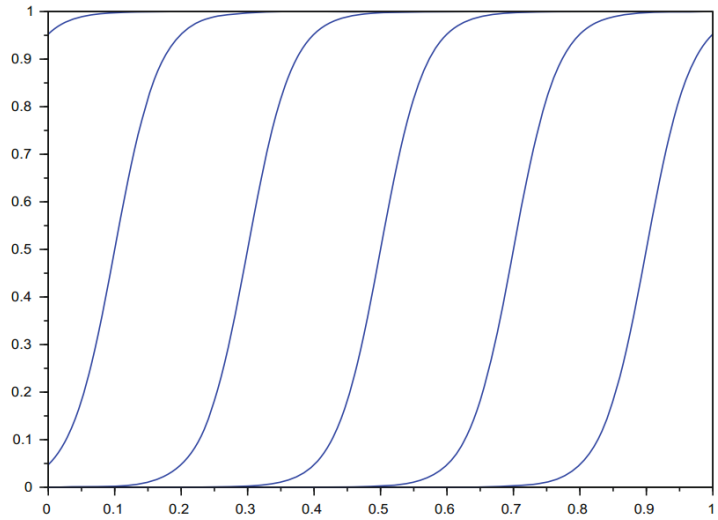
III. XẤP XỈ TAKAGI-SUGENO CỦA MẠNG NƠ-RON

Xét một mạng nơ-ron với một đầu vào, một đầu ra, một lớp ẩn và hàm kích hoạt sigmoid $\varphi(x) = \frac{1}{1+e^{-x}}$

$$NN(x) = w_0 + \sum_{i=1}^{n} w_i \varphi(a_i(x - b_i))$$



Hình 1 Kiến trúc của một mạng nơ-ron



Hình 2 Membership functions of sigmoid type

Chúng ta giả sử b1 < b2 < … < bn. Ngoài ra, do $\varphi(-x) = 1 - \varphi(x)$, chúng ta có thể giả sử a1, …, an > 0. Thật vậy, nếu chúng ta có ai < 0, chúng ta có thể viết như sau:

$$\varphi(a_i(x - b_i)) = \varphi(-|a_i|(x - b_i)) = 1 - \varphi(|a_i|(x - b_i))$$

và sau khi sắp xếp lại các giá trị a chúng ta thu được một mạng nơ-ron với tất cả ai > 0.

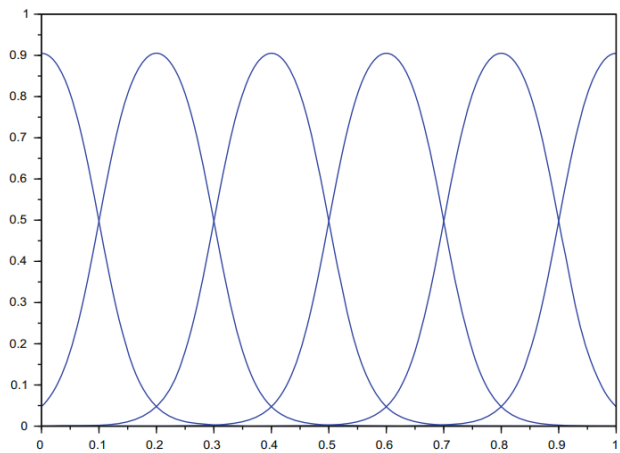
Chúng ta xây dựng một toán tử xấp xỉ với sigmoid-based membership functions.

$$A_0(x) = 1 - \varphi(a_1(x - b_1)),$$

$$A_i(x) = \varphi(a_i(x - b_i)) - \varphi(a_{i+1}(x - b_{i+1})),\ i = 1, ..., n - 1$$

$$A_n(x) = \varphi(a_n(x - b_n)).$$

Dạng của sigmoid-based membership functions thu được từ hiệu các hàm sigmoid được biểu diễn ở hình 3.



Hình 3 Sigmoid-based membership functions of a Takagi Sugeno fuzzy system

Chúng ta có thể dễ dàng thấy rằng Ai(x) có dạng chuông, nhưng không cần thiết phải dương với tất cả x ∈ R , do đó chúng không hẳn là tập mờ trong hầu hết các trường hợp. Hơn nữa, chúng ta nhận thấy rằng:

$$\sum_{i=0}^{n} A_i(x) = 1$$

Toán tử xấp xỉ có thể được biểu diễn như một sigmoid-based approximation (SA):

$$SA(x) = \sum_{i=0}^{n} A_i(x) \cdot y_i$$

Định lí 1. Mạng nơ-ron và hệ mờ TS miêu tả bên trên là tương đương nếu:

i. Cho hệ mờ TS như trên, nó có thể được viết như một mạng nơ-ron trên, với các trọng số:

$$w_0 = y_0,$$

$$w_i = -y_{i-1} + y_i, \ i = 1, ..., n,$$

ii. Cho mạng nơ-ron như trên, nó có thể được viết như một hệ mờ TS trên, với các trọng số:

$$y_0 = w_0,$$

$$y_i = \sum_{j=0}^{i} w_j, \ i = 1, ..., n$$

Chứng minh:

Chúng ta thấy rằng SA(x) và NN(x) là tổ hợp tuyến tính của cùng các hàm

$$\{1, \varphi(a_1(x - b_1)), \varphi(a_2(x - b_2)), ..., \varphi(a_n(x - b_n))\}$$

Xét đẳng thức:

$$SA(x) = NN(x), \forall x \in \mathbf{R}$$

Ta thu được:

$$\sum_{i=0}^{n} A_i(x) \cdot y_i = w_0 + \sum_{i=1}^{n} w_i \varphi(a_i(x - b_i))$$

Viết lại Ai dưới dạng các hàm sigmoid ta thu được:

$$y_0(1-\varphi(a_1(x-b_1))) + \sum_{i=1}^{n-1} y_i(\varphi(a_i(x-b_i)) - \varphi(a_{i+1}(x-b_{i+1}))) + y_n\varphi(a_n(x-b_n))$$

$$= w_0 + \sum_{i=1}^{n} w_i \varphi(a_i(x - b_i).$$

Do hệ

$$\{1, \varphi(a_1(x - b_1)), \varphi(a_2(x - b_2)), ..., \varphi(a_n(x - b_n))\}$$

là độc lập tuyến tính, do đó:

$$y_0 = w_0, \ \ - y_0 + y_1 = w_1, ..., -y_n + y_n = w_n,$$

Và (i) đã được chứng minh.

Chứng minh (ii) tương tự.

Khi mạng nơ-ron được viết dưới dạng một xấp xỉ sigmoid-based, không có sự mất mát thông tin trong bước khởi tạo. Bây giờ chúng ta xây dựng xấp xỉ Takagi-Sugeno của mạng nơ-ron.

Chúng ta bắt đầu xấp xỉ sigmoid-based (SA):

$$SA(x) = \sum_{i=0}^{n} A_i(x) \cdot y_i$$

Chúng ta xây dựng các luật mờ với hệ quả piece-wise constant:

$$\text{if } x \text{ is } A_i \text{ then } y = y_i, \quad i = 0, \ldots, n$$

Chúng ta giả sử b1 < b2 < … < bn. Đồng thời, giả sử a1, …, an > 0. Đặt:

$$c_0 = a_0$$

$$c_i = \frac{a_i + a_{i+1}}{2}, i = 1, .., n - 1,$$

$$c_n = a_n$$

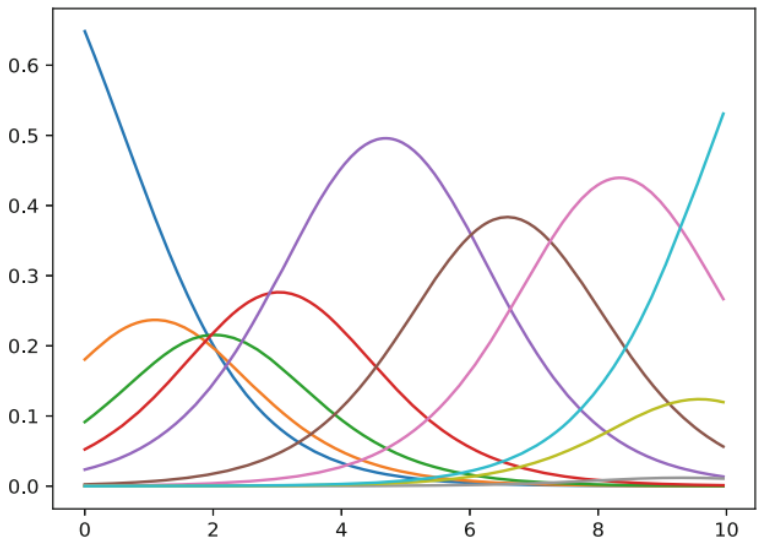
Chúng ta định nghĩa các tập mờ:

$$B_0(x) = 1 - \varphi(c_1(x - b_1)),$$
$$B_i(x) = \varphi(c_i(x - b_i)) - \varphi(c_i(x - b_{i+1})),\ i = 1, ..., n - 1$$
$$B_n(x) = \varphi(c_n(x - b_n)),$$

Và chúng ta định nghĩa hệ mờ TS:

$$TS(x) = \frac{\sum_{i=0}^{n} B_i(x) \cdot y_i}{\sum_{k=0}^{n} B_k(x)}$$

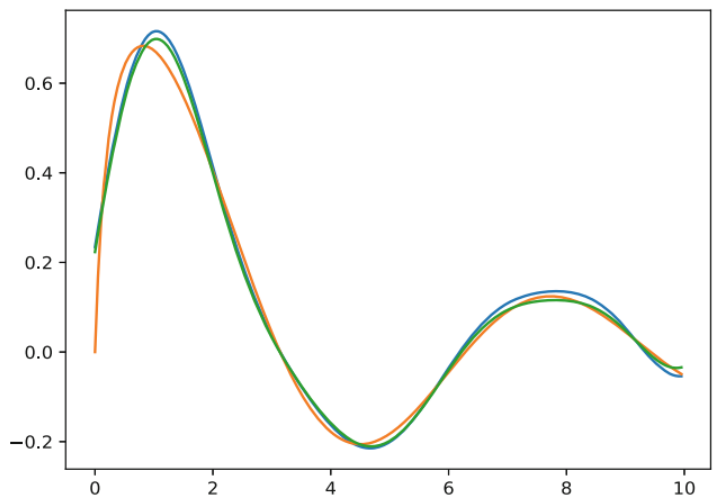
Hình 4 biểu diễn hình dạng của các hàm phụ thuộc sigmoid-based.



Hình 4 Antecedents of a TS fuzzy approximation of a Neural Network

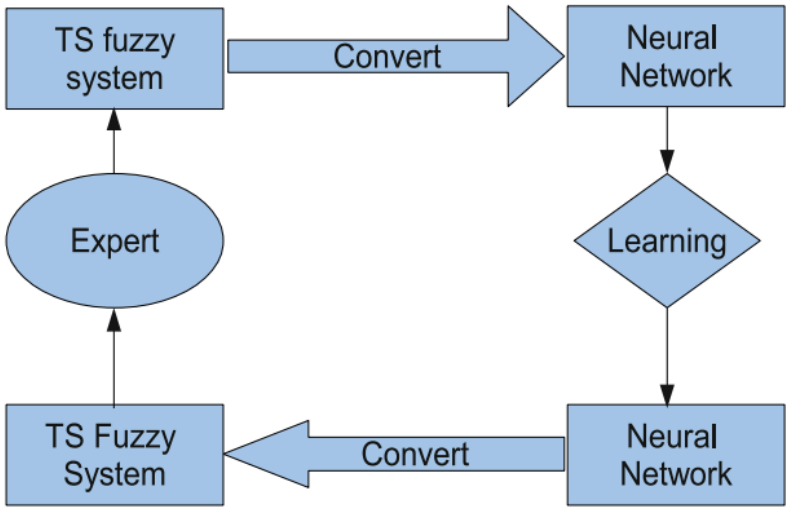
IV. Ví dụ minh hoạ

Chúng ta sẽ xét một ví dụ về bài toán xấp xỉ một hàm. Hàm f(x) = sin x / (x + 0.25) sẽ được xấp xỉ bởi một mạng nơ-ron và xấp xỉ TS tương ứng. Kết quả được minh hoạ trên hình 5.



Hình 5 Hàm f (màu cam), mạng nơ-ron (màu green), hệ mờ TS (màu blue)

Ý tưởng xấp xỉ một mạng nơ-ron bằng một hệ mờ TS cho phép chúng ta chuyển từ mạng nơ-ron sang hệ mờ TS và ngược lại. Điều này cung cấp một phương pháp thiết kế các hệ mờ khả hiểu. Quá trình này được minh hoạ ở hình 6. Một chuyên gia có thể thiết kế và tương tác với hệ mờ TS, trong khi các thành phần của mạng nơ-ron của hệ thống có thể học hiệu quả từ dữ liệu. Tổng hợp lại hệ thống này cho phép một hệ mờ khả hiểu.



Hình 6 Xấp xỉ hệ mờ TS của mạng nơ-ron là đề xuất cho mạng nơ-ron khả hiểu

V. TRƯỜNG HỢP NHIỀU CHIỀU

Để mở rộng kết quả bên trên cho trường hợp nhiều chiều chúng ta có thể sử dụng một vài cách tiếp cận. Trong cách tiếp cận mà chúng tôi chọn chúng tôi cho phép đầu vào nhiều chiều, với các trọng số nhiều chiều.

Xét hàm $f: R^n \to R$. b1, b2, …, bn là các số thực. ui ∈ R^n , i = 1, …, n. Chúng ta có thể định nghĩa các hàm nhiều chiều:

$$A_0(\mathbf{x}) = 1 - \varphi(\mathbf{u}_1 \cdot \mathbf{x} - b_1)$$
$$A_i(\mathbf{x}) = \varphi(\mathbf{u}_i \cdot \mathbf{x} - b_i) - \varphi(\mathbf{u}_{i+1} \cdot \mathbf{x} - b_{i+1}), i = 1, ..., n-1$$
$$A_n(\mathbf{x}) = \varphi(\mathbf{u}_n \cdot \mathbf{x} - b_n).$$

Xem xét các luật:

$$\text{if } \mathbf{x} \text{ is } A_i \text{ then } y = y_i, \ i = 0, \ldots, n.$$

Tương tự trường hợp một chiều chúng ta có:

$$\sum_{i=0}^{n} A_i(\mathbf{x}) = 1$$

Xấp xỉ sigmoid-based tương ứng với các luật này là:

$$SA(\mathbf{x}) = \sum_{i=0}^{n} A_i(\mathbf{x}) \cdot y_i.$$

Một mạng nơ-ron với hàm kích hoạt sigmoid có thể viết dưới dạng sau:

$$NN(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i \varphi(\mathbf{u}_i \cdot \mathbf{x} - b_i)).$$

Các định lí sau có thể đạt được tương tự như trường hợp một chiều.

Đính lí 2. Mạng nơ-ron và xấp xỉ SA mô tả ở trên là tương đương trong trường hợp:

1. Cho một xấp xỉ SA, nó có thể được viết dưới dạng một mạng nơ-ron, với các hệ số:

$$w_0 = y_0,$$
$$w_i = -y_{i-1} + y_i, \ i = 1, ..., n,$$

2. Cho trước một mạng nơ-ron, nó có thể được viết dưới dạng một xấp xỉ SA, với các hệ số:

$$y_0 = w_0,$$
$$y_i = \sum_{j=0}^{i} w_j, \ i = 1, ..., n.$$

Tương tự như trường hợp một chiều, các hàm $A_i(x)$ cho ở trên không nhất thiết phải là tập mờ do chúng có thể nhận giá trị âm.

Xét hàm $f : R^n \to R^n$, các số thực $b_1, \ldots, b_n$; $v_i \in R^n$, $i = 1, \ldots, n$. Chúng ta có thể định nghĩa các hàm nhiều chiều:

$$B_0(\mathbf{x}) = 1 - \varphi(\mathbf{v}_1 \cdot \mathbf{x} - b_1)$$
$$B_i(\mathbf{x}) = \varphi(\mathbf{v}_i \cdot \mathbf{x} - b_i) - \varphi(\mathbf{v}_i \cdot \mathbf{x} - b_{i+1}), i = 1, ..., n-1$$
$$B_n(\mathbf{x}) = \varphi(\mathbf{v}_n \cdot \mathbf{x} - b_n).$$

Lưu ý rằng cùng tham số độ dốc v_i trong các hàm sigmoid bị trừ ở B_i sẽ định nghĩa nó như là một tập mờ hợp lệ. Tuy nhiên tính chất partition of one bị mất trong trường hợp này. Chúng ta quan tâm các luật mờ TS:

$$\text{if } \mathbf{x} \text{ is } B_i \text{ then } y = y_i, \quad i = 0, \ldots, n.$$

Hệ mờ Takagi-Sugeno tương ứng với các luật này là

$$TS(x) = \frac{\sum_{i=0}^{n} B_i(\mathbf{x}) \cdot y_i.}{\sum_{i=0}^{n} B_i(\mathbf{x}).}$$

Chúng tôi quan sát thấy rằng có thể đơn giản hoá hệ mờ nếu không chuẩn hoá hệ mờ của chúng ta. Trong trường hợp này, chúng ta có thể viết hệ mờ TS như sau:

$$TSL(x) = \sum_{i=0}^{n} B_i(\mathbf{x}) \cdot z_i.$$

Hệ mờ này có thể được sử dụng như là một hệ mờ khả hiểu đơn giản. Các hệ số z_i được tính thông qua thuật toán học.

VI. HỆ MỜ TS DEEP LEARNING VỚI SIGMOID ANTECEDENTS

Các quan sát trong trường hợp nhiều chiều ở mục trước cung cấp cơ hội để khám phá kiến trúc deep learning. Chúng ta sẽ xây dựng lớp L của mạng nơ-ron được đề xuất như là một hệ mờ TS. Chúng ta xem xét lớp L với n + 1 nơ-ron có các hàm kích hoạt sau:

$$B_0(\mathbf{x}) = 1 - \varphi(\mathbf{v}_1 \cdot \mathbf{x} - b_1)$$

$$B_i(\mathbf{x}) = \varphi(\mathbf{v}_i \cdot \mathbf{x} - b_i) - \varphi(\mathbf{v}_i \cdot \mathbf{x} - b_{i+1}), i = 1, ..., n-1$$

$$B_n(\mathbf{x}) = \varphi(\mathbf{v}_n \cdot \mathbf{x} - b_n).$$

Khi đó lớp này có thể được viết dưới dạng một hệ mờ TS:

$$TSL(x) = \sum_{i=0}^{n} B_i(\mathbf{x}) \cdot \mathbf{z}_i.$$

Như vậy chúng ta có cả input và ouput là đa chiều, chúng ta có thể tổ hợp nhiều lớp kiểu như thế này để tạo thành một mạng nơ-ron. Trong các kiến trúc như vậy, deep learning trở thành khả thi.

VII. TỔNG KẾT

Bài báo này trình bày về một hệ thống cho việc học khả hiểu dựa trên hệ mờ TS xây dựng sử dụng các hàm sigmoid khác nhau. Hệ thống đề xuất hứa hẹn có tính chất khả hiểu và có thể học được. Chứng minh khái niệm cho hệ thống đề xuất được đưa ra trong một ví dụ xấp xỉ hàm. Các nghiên cứu sâu hơn sử dụng dữ liệu thực tế sẽ là bước tiếp theo để đánh giá tính hiệu quả của hệ thống đề xuất. Một mở rộng cho trường hợp đa chiều cũng được đưa ra, cùng với những hướng dẫn cho các nghiên cứu xa hơn liên quan đến deep learning.

VIII. TÀI LIỆU THAM KHẢO

[1] Bede, B.: Mathematics of Fuzzy Sets and Fuzzy Logic. Springer, Heidelberg (2013)

[2] Bonanno, D., Nock, K., Smith, L., Elmore, P., Petry, F.: An approach to explainable deep learning using fuzzy inference. In: Next-Generation Analyst V, SPIE Proceedings, vol. 10207, International Society for Optics and Photonics (2017)
[3] Costarelli, D., Spigler, R.: Approximation results for neural network operators activated by sigmoidal functions. Neural Netw. 44, 101–106 (2013)
[4] Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. 2(4), 303–314 (1989)
[5] Horikawa, S.-I., Furuhashi, T., Uchikawa, Y.: On fuzzy modeling using fuzzy neural networks with the back-propagation algorithm. IEEE Trans. Neural Netw. 3(5), 801–806 (1992)
[6] Jang, J.-S.R.: ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybern. 23(3), 665–685 (1993)
[7] Jin, Y., Sendhoff, B.: Extracting interpretable fuzzy rules from RBF networks. Neural Process. Lett. 17(2), 149–164 (2003)
[8] Kasabov, N.K.: Learning fuzzy rules and approximate reasoning in fuzzy neural networks and hybrid systems. Fuzzy Sets Syst. 82(2), 135–149 (1996)
[9] Montavon, G., Samek, W., Muller, K.R.: Methods for interpreting and understanding deep neural networks. Digit. Signal Process. 73, 1–15 (2018)
[10] Sugeno, M.: An introductory survey of fuzzy control. Inf. Sci. 36, 59–83 (1985)
[11] Tanaka, K., Sugeno, M.: Stability analysis and design of fuzzy control systems. Fuzzy Sets Syst. 45(2), 135–156 (1992)
[12] Wang, L.-X., Mendel, J.M.: Fuzzy basis functions, universal approximation, and orthogonal least-squares learning. IEEE Trans. Neural Netw. 3(5), 807–814 (1992)
[13] Ying, H.: General SISO Takagi-Sugeno fuzzy systems with linear rule consequent are universal approximators. IEEE Trans. Fuzzy Syst. 6(4), 582–587 (1998)
[14] Zadeh, L.A.: Fuzzy Sets. Inf. Control 8, 338–353 (1965)
[15] Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning - I. Inf. Sci. 8(3), 199–249 (1975)
[16] Zadeh, L.A.: Fuzzy logic, neural networks, and soft computing. Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems, pp. 775–782 (1996). Selected Papers by Zadeh, L.A