



República Federativa do Brasil
Ministério da Indústria, Comércio Exterior
e Serviços
Instituto Nacional da Propriedade Industrial

(11) PI 0403322-1 B1

(22) Data do Depósito: 19/08/2004

(45) Data de Concessão: 25/09/2018



(54) Título: FUNÇÃO BI-HIPERBÓLICA E SUA UTILIZAÇÃO COMO FUNÇÃO DE ATIVAÇÃO EM NEURÔNIOS COMPONENTES DE REDES NEURAIS ARTIFICIAIS

(51) Int.Cl.: G06N 3/02; G06F 15/18

(73) Titular(es): ADILSON ELIAS XAVIER

(72) Inventor(es): ADILSON ELIAS XAVIER

(85) Data do Início da Fase Nacional: 19/08/2004

Relatório Descritivo da Patente de Invenção para
"FUNÇÃO BI-HIPERBÓLICA E SUA UTILIZAÇÃO COMO FUNÇÃO DE
ATIVAÇÃO EM NEURÔNIOS COMPONENTES DE REDES NEURAIS
ARTIFICIAIS".

5 **Campo Técnico**

A invenção refere-se à utilização de uma especial
função de ativação de um neurônio, componente elementar na
construção de redes neurais artificiais.

Essas redes neurais têm o mais amplo espectro de
10 aplicações, nas mais diversas atividades técnicas. Ademais,
sua utilização tem mantido um expressivo crescimento a cada
ano.

A implementação dessas redes neurais é realizada
quer em software ou em circuitos hardwired, sob a forma de
15 inumeráveis variações.

Um elemento comum entre todas essas variações é a
presença ubíqua da função de ativação, como parte
protagônica no componente elementar de uma rede neural: o
neurônio.

20 As redes neurais integram sistemas utilizados em
áreas tais como: eletrônica, controle, telecomunicações,
dentre outras.

Descrição Geral de uma Rede Neural

As implementações de redes neurais associadas a
25 aplicações práticas se utilizam, de diversas funções de
ativação, conforme sobejamente registrado na literatura
científica sobre esse assunto.

Primeiramente, será uma descrição sucinta de uma
rede neural, do funcionamento de um neurônio e do papel de
30 uma função de ativação.

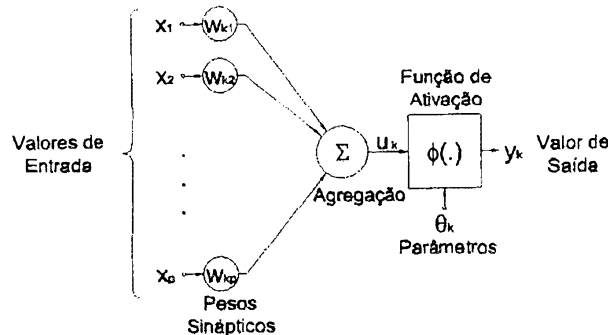
Uma rede neural é formada por um conjunto de
neurônios interligados. Assim, o funcionamento de um

07

neurônio constitui um elemento fundamental na operação dessas redes.

Haykin S., em seu livro "Neural Networks: A Comprehensive Foundation", publicado em 1994 pela editora Macmillan College Publishing Company, por exemplo, descreve claramente um modelo para um neurônio, conforme apresentado abaixo:

08



Dado um vetor de entradas $x = (x_1, x_2, \dots, x_p)$, a resposta y_k , saída de um neurônio genérico k , é calculada segundo três procedimentos básicos:

-Um conjunto de pesos associados a esse neurônio k , denominados também por sinapses, $w_k = (w_{k1}, w_{k2}, \dots, w_{kp})$, é aplicado às componentes de entrada;

-A seguir, é feita a agregação das componentes ponderadas pelas sinapses;

-Finalmente, esse valor agregado, transladado por um parâmetro θ_k , é fornecido a uma função de ativação $\phi(\cdot)$, que produz a resposta do neurônio.

Tais procedimentos compreendem a seguinte sequência de cálculos:

$$u_k = \sum_{j=1}^p w_{kj} x_j$$

$$v_k = u_k - \theta_k$$

$$y_k = \phi(v_k)$$

As funções de ativação $\phi(.)$ tipicamente fornecem a resposta dentro do intervalo $[0,1]$, que, sem perda de generalidade, será tomado, neste relatório, como referencial.

09

5 Técnicas Anteriores

Abaixo, são apresentadas as principais alternativas usuais de funções de ativação e suas correspondentes derivadas.

1- Função Degrau

10

$$\phi_1(v) = \begin{cases} 1, & \text{se } v \geq 0. \\ 0, & \text{se } v < 0. \end{cases}$$

Conforme representada na figura 1, a derivada da função degrau é $\phi'_1(v) = 0$, para qualquer $v \neq 0$, enquanto, não é definida no ponto $v = 0$.

A descontinuidade na origem e o valor nulo da derivada nos demais pontos praticamente anulam a utilidade prática dessa função $\phi_1(.)$.

2- Função Patamar

$$\phi_2(v, b) = \begin{cases} 0, & \text{se } v \leq -b; \\ (v + b)/2b, & \text{se } -b < v < b; \\ 1, & \text{se } v \geq b. \end{cases}$$

A função patamar assume a forma apresentada na figura 2. O valor do parâmetro b está associado ao ângulo α da inclinação da parte intermediária dessa função, de acordo com a seguinte expressão: $b = 1/(2 \tan \alpha)$.

A sua derivada não é definida nos pontos $v = -1/2$ e $v = 1/2$, e nos demais valores assume:

$$\phi'_2(v, b) = \begin{cases} 0, & \text{para } v < -b; \\ 1/2b, & \text{para } -b < v < b; \\ 0, & \text{para } v > b. \end{cases}$$

A insensibilidade da derivada fora do intervalo $(-b, b)$ limita consideravelmente o uso prático dessa função de ativação $\phi_2(\cdot)$.

3- Função Logística

- 5 A função de ativação mais amplamente utilizada é a função logística:

$$\phi_3(v, a) = \frac{1}{1 + e^{-av}}$$

A derivada da função logística apresenta a expressão:

10
$$\phi_3'(v, a) = a\phi_3(v, a)(1 - \phi_3(v, a))$$

A figura 3 mostra os gráficos dessa função para 3 diferentes valores do parâmetro a .

- A função logística oferece a importante flexibilidade dada por sua inclinação na origem,
15 $\phi_3'(0, a) = a/4$, ser variável com o parâmetro a , conforme ilustra a figura 3.

Além disso, a função logística apresenta propriedades de simetria e de completa diferenciabilidade, ou seja, pertence à classe de funções C^∞ .

- 20 Outra função também muito utilizada como função de ativação é a função tangente hiperbólica, que pode ser colocada sob a forma:

$$\phi_4(v, a) = \frac{1 - e^{-av}}{1 + e^{-av}}$$

- A função tangente hiperbólica, entretanto,
25 corresponde unicamente à transformação elementar da função logística:

$$\phi_4(v, a) = 2 \cdot \phi_3(v, a) - 1$$

- Dessa forma, a função tangente hiperbólica pode ser considerada como simplesmente uma derivação particular
30 da função logística.

Descrição Detalhada da Invenção

A função de ativação proposta, aqui denominada função bi-hiperbólica, e suas derivadas primeira e segunda, na sua forma mais geral, assumem as seguintes expressões:

$$5 \quad \varphi_x(v, \lambda, \tau_1, \tau_2) = \sqrt{\lambda^2 \left(v + \frac{1}{4\lambda}\right)^2 + \tau_1^2} - \sqrt{\lambda^2 \left(v - \frac{1}{4\lambda}\right)^2 + \tau_2^2} + \frac{1}{2}$$

$$\varphi'_x(v, \lambda, \tau_1, \tau_2) = \frac{\lambda^2 \left(v + \frac{1}{4\lambda}\right)}{\sqrt{\lambda^2 \left(v + \frac{1}{4\lambda}\right)^2 + \tau_1^2}} - \frac{\lambda^2 \left(v - \frac{1}{4\lambda}\right)}{\sqrt{\lambda^2 \left(v - \frac{1}{4\lambda}\right)^2 + \tau_2^2}}$$

$$\varphi''_x(v, \lambda, \tau_1, \tau_2) = \frac{\lambda^2 \tau_1^2}{(\lambda^2 \left(v + \frac{1}{4\lambda}\right)^2 + \tau_1^2)^{3/2}} - \frac{\lambda^2 \tau_2^2}{(\lambda^2 \left(v - \frac{1}{4\lambda}\right)^2 + \tau_2^2)^{3/2}}$$

O gráfico dessa função proposta é apresentada na
10 figura 4.

A função $\varphi_x(., \lambda, \tau_1, \tau_2)$ apresenta a desejada propriedade de possuir diferenciabilidade infinita, ou seja, pertence à classe de funções C^∞ .

A função $\varphi_x(., \lambda, \tau_1, \tau_2)$ apresenta as seguintes
15 propriedades triviais, consentâneas às demais funções de ativação:

$$\lim_{v \rightarrow -\infty} \varphi_x(v, \lambda, \tau_1, \tau_2) = 0$$

$$\lim_{v \rightarrow \infty} \varphi_x(v, \lambda, \tau_1, \tau_2) = 1$$

$$\lim_{v \rightarrow -\infty} \varphi'_x(v, \lambda, \tau_1, \tau_2) = 0$$

$$\lim_{v \rightarrow \infty} \varphi'_x(v, \lambda, \tau_1, \tau_2) = 0$$

Diversamente das funções de ativação ortodoxas, quando $\tau_1 \neq \tau_2$, a função $\varphi_x(., \lambda, \tau_1, \tau_2)$ apresenta as
20 diferenciadas propriedades de não oferecer simetria, bem como, de possuir uma imagem que ultrapassa o tradicional espaço definido pelo intervalo $[0,1]$. Assim, quando $\tau_1 > \tau_2$, o valor da função $\varphi_x(., \lambda, \tau_1, \tau_2)$ ultrapassa o valor 1, situação retratada pelo gráfico da figura 4.

Inversamente, quando $\tau_1 < \tau_2$, o valor da função $\phi_x(\cdot, \lambda, \tau_1, \tau_2)$ assume valores menores do que 0.

Em sua forma mais particular, igualando os valores dos parâmetros $\tau_1 = \tau_2 = \tau$, a função $\phi_x(\cdot, \lambda, \tau_1, \tau_2) = \phi_x(\cdot, \lambda, \tau)$, assume uma forma mais consentânea a outras funções de ativação, tendo imagem no intervalo $[0, 1]$ e oferecendo a propriedade de simetria, conforme retratado pelos gráficos das figuras 5 e 6.

$$\phi_x(v, \lambda, \tau) = \frac{\sqrt{\lambda^2 \left(v + \frac{1}{4\lambda}\right)^2 + \tau^2} - \sqrt{\lambda^2 \left(v - \frac{1}{4\lambda}\right)^2 + \tau^2}}{2}$$

As derivadas da função $\phi_x(v, \lambda, \tau)$ assumem as expressões:

$$\phi'_x(v, \lambda, \tau) = \frac{\lambda^2 \left(v + \frac{1}{4\lambda}\right)}{\sqrt{\lambda^2 \left(v + \frac{1}{4\lambda}\right)^2 + \tau^2}} - \frac{\lambda^2 \left(v - \frac{1}{4\lambda}\right)}{\sqrt{\lambda^2 \left(v - \frac{1}{4\lambda}\right)^2 + \tau^2}}$$

$$\phi''_x(v, \lambda, \tau) = \frac{\lambda^2 \tau^2}{(\lambda^2 \left(v + \frac{1}{4\lambda}\right)^2 + \tau^2)^{3/2}} - \frac{\lambda^2 \tau^2}{(\lambda^2 \left(v - \frac{1}{4\lambda}\right)^2 + \tau^2)^{3/2}}$$

A função $[\phi_x(v, \lambda, \tau) - 1/2]$ é anti-simétrica, ou seja:

$$\phi_x(v, \lambda, \tau) - 1/2 = -[\phi_x(-v, \lambda, \tau) - 1/2]$$

No ponto $v=0$, são observados os seguintes valores para a função ϕ_x e para sua derivada:

$$\phi_x(0, \lambda, \tau) = 1/2,$$

$$\phi'_x(0, \lambda, \tau) = \frac{\lambda}{2\sqrt{1/16 + \tau^2}}$$

$$\lim_{\tau \rightarrow 0} \phi'_x(0, \lambda, \tau) = 2\lambda$$

Na figura 5, são mostrados gráficos para a forma simétrica da função proposta correspondentes a três diferentes valores para o parâmetro λ , mantendo-se o parâmetro τ constante. Pode-se ver um efeito similar àquele produzido pela variação do parâmetro a na função logística, conforme apresentado pela figura 3. Dessa forma,

pode-se associar o parâmetro λ à inclinação da função na origem.

A função $\phi_x(., \lambda, \tau)$, apresenta, ademais, os seguintes comportamentos assintóticos:

$$\lim_{\lambda \rightarrow \infty} \phi_x(v, \lambda, \tau) = \phi_1(v)$$

$$\lim_{\tau \rightarrow 0} \phi_x(v, \lambda, \tau) = \phi_2\left(v, \frac{1}{4\lambda}\right)$$

Podemos assim aproximar a função de ativação patamar $\phi_2(., .)$, por uma forma continuamente diferenciável usando a função $\phi_x(., ., .)$, podendo essa aproximação ser tão próxima quanto se queira. Adicionalmente, pode-se efetuar essa aproximação assintótica mantendo constante uma inclinação com especificado ângulo α na origem, $v=0$, bastando para isso se fazer a seguinte re-parametrização:

$$\lambda = 4\lambda' \sqrt{1/16 + \tau^2}$$

sendo $\lambda' = \frac{1}{2} \operatorname{tg} \alpha$. Na figura 6, são mostrados 2 gráficos da função proposta com uma inclinação constante na origem. Aí pode-se ver, que o parâmetro τ está associado ao afastamento da curva às duas assíntotas horizontais.

Em termos computacionais, através do cálculo de 100.000.000 valores da função e de sua derivada primeira, com argumentos aleatoriamente gerados, pode ser observado que a função $\phi_x(., ., .)$ é 24,8% mais rápida que a função logística. Esse resultado foi obtido através de simulação programada na linguagem FORTRAN, compilador COMPAQ, computador tipo IBM-PC, com 1000Mhz de velocidade.

A existência de dois parâmetros na função bi-hiperbólica $\phi_x(., \lambda, \tau)$, em sua forma mais particular quando $\tau_1 = \tau_2 = \tau$, um a mais que as demais funções de ativação, enseja à essa função uma maior flexibilidade para representar mais adequadamente os fenômenos modelados com

13

redes neurais implementadas com as funções de ativação ortodoxas.

5 Numa rede neural multicamadas, por exemplo, essa maior flexibilidade engendra, certamente à função de ativação ora proposta, o poder de aproximar qualquer função de uma forma mais sintética, com menor número de neurônios.

10 Através da manipulação conveniente de seus parâmetros, em particular do parâmetro τ , a função $\phi_x(.,\lambda,\tau)$, oferece ademais a possibilidade de poder enfrentar mais convenientemente o desastroso fenômeno de saturação, que não raramente ocorre no treinamento de redes neurais, (vide, por exemplo, Looney C. G., à pagina 161 de seu livro Pattern Recognition Using Neural Networks, publicado pela editora Oxford University Press em 1997, ou 15 à página 165 do livro de Haykin, acima especificado).

Igualmente, uma manipulação adequada dos parâmetros de $\phi_x(.,\lambda,\tau)$, em tese, pode ser usada para se evitar um indesejado mínimo local.

20 Em relação às questões de saturação e de mínimo local, acima referidas, a manipulação dos parâmetros τ_1 e τ_2 só pode oferecer à função assimétrica $\phi_x(.,\lambda,\tau_1,\tau_2)$ possibilidades adicionais para enfrentá-las.

REIVINDICAÇÕES

1- Método de utilização da função bi-hiperbólica como função ativação de neurônios artificiais em rede neural, em que dado um vetor de entradas $x = (x_1, x_2, \dots, x_p)$, a resposta y_k , saída de um neurônio genérico k , é calculada segundo três procedimentos básicos:

- Um conjunto de pesos associados a esse neurônio k , $w_k = (w_{k1}, w_{k2}, \dots, w_{kp})$, é aplicado às componentes de entrada;

- A seguir, é feita a agregação das componentes ponderadas pelos pesos;

- Finalmente, esse valor agregado, transladado por um parâmetro θ_k , é fornecido a uma função de ativação $\phi(\cdot)$, que produz a resposta do neurônio;

compreendendo a seguinte seqüência de cálculos:

$$\begin{aligned}u_k &= \sum_{j=1}^p w_{kj} x_j \\v_k &= u_k - \theta_k \\y_k &= \phi(v_k)\end{aligned}$$

em que a função $\phi(\cdot)$ na última expressão é caracterizada pela forma particular da função bi-hiperbólica

$$\phi_x(v, \lambda, \tau_1, \tau_2) = \sqrt{\lambda^2 \left(v + \frac{1}{4\lambda}\right)^2 + \tau_1^2} - \sqrt{\lambda^2 \left(v - \frac{1}{4\lambda}\right)^2 + \tau_2^2} + \frac{1}{2}$$

onde v é o argumento de entrada e λ, τ_1 e τ_2 são parâmetros de ajuste, e com as derivadas primeira e segunda desta função bi-hiperbólica utilizadas nos procedimentos de aprendizado da rede neural.

2- Utilização da função bi-hiperbólica, conforme descrita na reivindicação 1, caracterizada pelo uso como função de ativação de neurônios artificiais, seja a implementação da rede neural alvo por circuitos eletrônicos, moleculares ou óticos, analógicos ou digitais (hardware), ou combinações destas técnicas.

3- Utilização da função bi-hiperbólica, conforme descrita na reivindicação 1 e de acordo com a reivindicação 2, caracterizada pelo uso da relação de desigualdade entre os parâmetros τ_1 e τ_2 no controle do grau de simetria da função de ativação, de forma a permitir seu uso em ambas as formas, assimétrica, fazendo $\tau_1 \neq \tau_2$, ou simétrica, fazendo $\tau_1 = \tau_2$, quando assume a expressão mais simplificada:

$$\phi_x(v, \lambda, \tau) = \sqrt{\lambda^2 \left(v + \frac{1}{4\lambda}\right)^2 + \tau^2} - \sqrt{\lambda^2 \left(v - \frac{1}{4\lambda}\right)^2 + \tau^2} + \frac{1}{2}$$

4 - Utilização da função bi-hiperbólica, conforme descrita na reivindicação 1 e de acordo com a reivindicação 2, caracterizada pelo uso do parâmetro λ no controle do grau de inclinação da função de ativação, de forma a permitir o uso da função de ativação tanto como Função Degrau, Função Patamar, ou mesmo imitar a Função Logística e derivações elementares tais como a Função Tangente Hiperbólica, com qualquer grau de inclinação desejado entre 0° e 90° , enquanto mantendo a desejável característica de diferenciabilidade, normalmente requerida nos processos de aprendizado.

5 - Utilização da função bi-hiperbólica, conforme descrita na reivindicação 1 e de acordo com a reivindicação 2, caracterizada pelo uso do parâmetro τ no controle do grau de suavidade da função de ativação, de forma a permitir o uso da função de ativação tanto como aproximação da Função Patamar, ou como Função Logística e derivações elementares tais como a Função Tangente Hiperbólica, mantendo-se neste caso a inclinação na origem constante, enquanto mantendo a desejável característica de diferenciabilidade, normalmente requerida nos processos de aprendizado.

FIGURAS

58

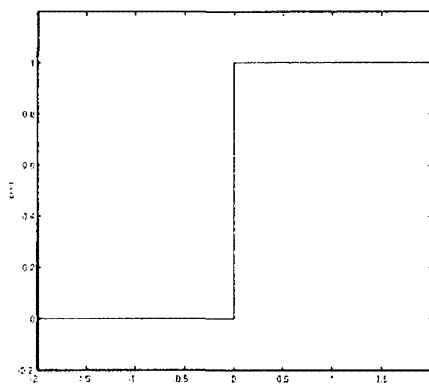


Figura 1

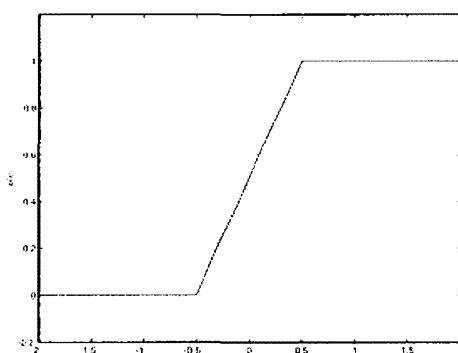


Figura 2

59

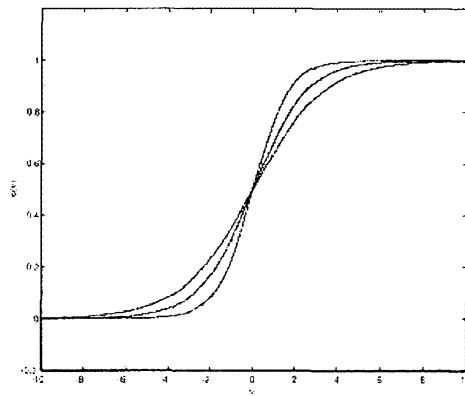


Figura 3

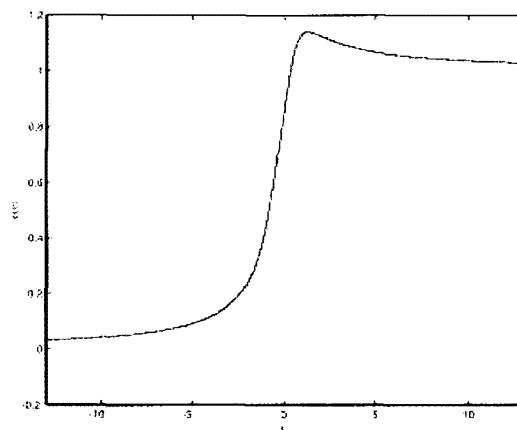


Figura 4

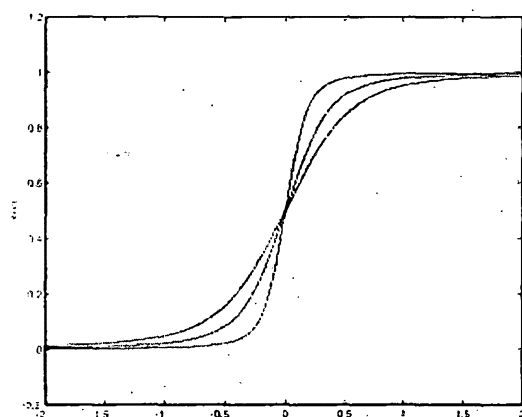


Figura 5

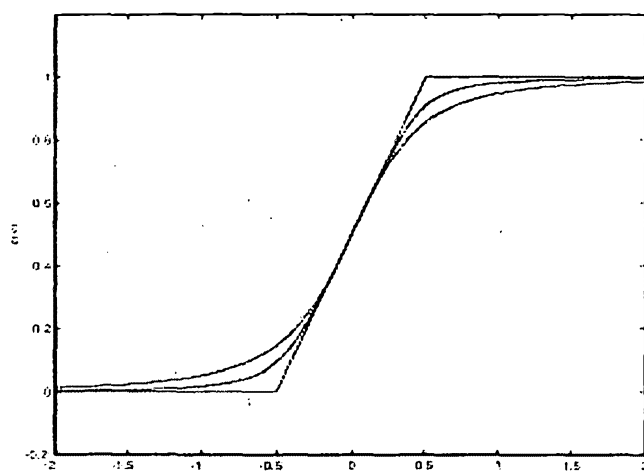


Figura 6