DESEMPENHO DO ALGORÍTMO DE BACKPROPAGATION COM A FUNÇÃO DE ATIVAÇÃO BI-HIPERBÓLICA

Geraldo Miguez
COPPE / PESC, Universidade Federal do Rio de Janeiro, Brasil
geraldomiguez@gmail.com

Nelson Maculan
COPPE / PESC, Universidade Federal do Rio de Janeiro, Brasil
maculan@cos.ufrj.br

Adilson Elias Xavier
COPPE / PESC, Universidade Federal do Rio de Janeiro, Brasil
adilson@cos.ufrj.br

RESUMO

A utilização mais ampla do algoritmo de Backpropagation na solução de problemas práticos do mundo real tem esbarrado na sua relativa lentidão. Muitas técnicas têm sido discutidas para acelerar o seu desempenho e a técnica apresentada neste trabalho é a utilização de uma nova função de ativação, a função Bi-hiperbólica, que proporciona melhor desempenho computacional, possibilitando fugir do problema de saturação dos neurônios e exigindo uma arquitetura mais leve, com um menor número de neurônios em sua camada oculta. A maior flexibilidade, eficiência e capacidade de discriminação desta função é demonstrada através de um conjunto de experimentos computacionais com problemas tradicionais da literatura.

PALAVRAS CHAVE. Redes Neurais; Backpropagation; Função Bi-hiperbólica;

Área principal. Programação Matemática

ABSTRACT

The use of Backpropagation algorithm in real world problems solutions has been blocked by its slow performance. Many techniques have been discussed to speed up its performance and in this paper a new strategy is presented based on the use of a new activation function, the Bi-hyperbolic function, that offers more flexibility, avoids saturation's problem e needs a smaller architecture with less neurons and shows a faster evaluation time. The efficiency and the discrimination capacity of the proposed methodology are shown through a set of computational experiments with traditional problems of the literature.

KEYWORDS. Neural Networks; Backpropagation; Bi-hyperbolic Function;

Main area. Mathematical Programming

Introdução

A utilização de Redes Neurais Artificiais (RNA) vem se destacando na construção de sistemas para uso em diversas áreas do conhecimento humano. Isto se deve, em grande parte, pela utilização das redes do tipo Perceptron de Múltiplas Camadas (Multilayer Perceptrons – MLP).

Por suas características de obter soluções através do aprendizado do comportamento do ambiente no qual ela está inserida, as redes neurais artificiais dependem de um eficiente algoritmo de treinamento. Segundo informes encontrados na literatura especializada, o algoritmo mais utilizado tem sido o Backpropagation. É um método computacionalmente eficiente para o treinamento de redes MLPs e que resolve o problema de realizar a propagação reversa do erro em RNAs com múltiplas camadas. Apesar disto, ele ainda apresenta algumas limitações na sua utilização que impedem a sua aplicação de uma forma mais ampla em problemas do mundo real. Sendo um método baseado no uso de gradientes, existe sempre a possibilidade de convergência para um mínimo local, falhando na localização do mínimo global. Outro problema constantemente relatado diz respeito ao caso em que, mesmo nos casos em que consegue atingir o seu objetivo e apresentar um erro dentro dos limites desejados, a lentidão muito grande no seu processamento pode chegar a impedir o seu uso. Esta demora no processamento dificulta a sua utilização em uma gama maior de aplicações práticas, em especial em aplicações de médio e grande porte (SCHIFFMANN et al, 1994), (OTAIR et al, 2005).

Um dos fatores possivelmente responsável pela lentidão deste processo de convergência é a função de ativação usada em seus neurônios. Sendo o processo de aprendizagem da rede essencialmente iterativo, uma função mais lenta para ser calculada torna todo o procedimento demorado. Acredita-se que a razão para isto seja a saturação da função de ativação usada nos neurônios em suas diversas camadas. Isto se deve ao fato de que, uma vez que a saturação de uma unidade ocorre, o gradiente descendente tende a assumir valores muito pequenos, mesmo nos casos em que o erro de saída ainda é grande.

Para otimizar a eficiência e a taxa de convergência do algoritmo de backpropagation, é proposto neste trabalho a utilização de uma nova função de ativação, a Função Bi-hiperbólica Simétrica. Ela apresenta características que atendem às necessidades do algoritmo de backpropagation, além de possibilitar uma maior flexibilidade na representação dos fenômenos modelados. Ela conta com o uso de dois parâmetros, um a mais do que nas funções tradicionalmente utilizadas para esta finalidade. Isto implica na possibilidade de melhor enfrentar o problema da saturação, além de permitir melhor tratamento para evitar os mínimos locais. Por suas características próprias apresenta, ainda, a vantagem de ser computacionalmente muito mais rápida de ser avaliada do que a função logística. (XAVIER, 2005).

Além disso, o uso desta Função Bi-Hiperbólica possibilita, por sua maior flexibilidade, a capacidade de poder aproximar qualquer função de uma forma mais sintética, permitindo a utilização de um menor número de neurônios, melhorando ainda mais o desempenho do algoritmo backpropagation, agindo diretamente na topologia da rede (XAVIER , 2005).

Para possibilitar a avaliação computacional destas características foi desenvolvido um protótipo em MATLAB que, através de uma interface gráfica, permitiu a obtenção de resultados altamente favoráveis, apresentados posteriormente neste trabalho.

Redes Neurais Artificiais

Uma Rede Neural Artificial funciona pela criação de ligações entre suas unidades de processamento matemático, chamados de neurônios artificiais. O conhecimento é codificado na rede pela força destas conexões entre diferentes neurônios, chamada de peso, e pela criação de camadas de neurônios que trabalham em paralelo. O sistema aprende através de um processo de determinação do número de neurônios, ou nós, e pelo ajuste dos pesos dessas conexões com base nos dados usados para o treinamento. O poder computacional de uma RNA é devido basicamente à sua estrutura paralela pesadamente distribuída e à sua habilidade de aprender e, conseqüentemente, generalizar (HAYKIN, 2001).

Neurônios Artificiais são unidades de processamento das RNAs. Eles são simplificações do conhecimento que se tinha do neurônio biológico, feitas por McCulloch e Pitts (KÓVACS, 1996). O modelo desenvolvido apresenta vários terminais de entrada, representando os dendritos, e um terminal de saída, representando o axônio. As sinapses têm seu comportamento simulado pelo acoplamento de pesos a cada terminal de entrada do neurônio artificial e podem assumir valores positivos ou negativos, emulando sinapses excitatórias ou inibitórias. A saída do neurônio artificial é obtida através da aplicação de uma função de ativação que pode ativar ou não esta saída, dependendo da soma ponderada dos valores de cada entrada, submetida a esta função, atingir um limiar pré-determinado. A função de ativação limita a faixa de amplitude permitida do sinal de saída a algum valor finito. Tipicamente, a amplitude normalizada da saída de um neurônio é restrita ao intervalo unitário fechado [0, 1] ou, alternativamente, [-1, 1]. O modelo neural usado inclui uma polarização externa (*bias*), que tem o efeito de aumentar ou diminuir o argumento da função de ativação (φ), que define a saída do neurônio em termos do potencial de ativação.

O neurônio pode ser descrito, em termos matemáticos da seguinte forma:

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

$$v_k = u_k - \theta_k$$

$$y_k = \varphi(v_k)$$

Onde x_1, x_2, ... , x_p são os sinais de entrada; w_{k1}, w_{k2}, ... , w_{kp} são os pesos sinápticos do neurônio *k;* u_k é a saída proveniente da combinação linear dos sinais de entrada e pesos; θ_k é o bias; $\varphi(\bullet)$ é a função de ativação; e y_k é o sinal de saída do neurônio (HAIKIN, 2001).

Funções de Ativação

Um dos componentes mais importantes do neurônio artificial é a sua função de ativação ou transferência. Ela tem por objetivo limitar a amplitude válida do sinal de saída do neurônio a um valor finito. Normalmente, esta amplitude normalizada se encontra em um intervalo fechado unitário [0, 1] ou, em alguns casos, [-1, 1].

As funções de ativação mais comumente utilizadas e disponibilizadas na literatura são apresentadas abaixo. Também são descritas as suas derivadas, que têm grande importância no método de treinamento de redes neurais artificiais conhecido como Backpropagation (XAVIER, 2005), (HAYKIN,2001).

a) Função Degrau

$$\varphi_1(v) = \begin{cases} 1 & se \quad v \geq 0 \\ 0 & se \quad v < 0 \end{cases}$$

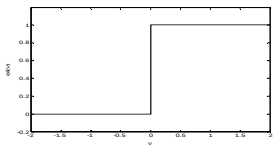
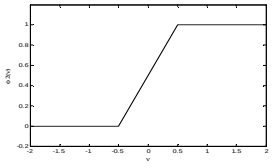


Figura 1: Função Degrau

A derivada desta função é $\varphi_1'(v) = 0$ para $\forall \quad v \neq 0$ e não é definida para $v = 0$. A descontinuidade na origem associada ao valor nulo da derivada nos demais pontos restringe muito a utilidade prática desta função. Ela está representada na Figura 1.

b) Função Patamar

$$\varphi_2(v,b) = \begin{cases} 0, & se \quad v \leq -b; \\ (v+b)/2b, & se \quad -b < v < b; \\ 1, & se \quad v > b; \end{cases}$$



Sendo $b = 1/2\tan\alpha$, onde α é o ângulo de inclinação. Figura 2: Função Patamar

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

$$\varphi_2'(v,b)=\begin{cases}0, & para \quad v<-b;\\ 1/2b, & para \quad -b<v<b;\\ 0, & para \quad v>b;\end{cases}$$

A insensibilidade da derivada fora do intervalo (-*b*, *b*) limita consideravelmente o uso prático dessa função de ativação $\varphi_2(\bullet)$.

c) Função Logística

Esta é a função de ativação mais utilizada na construção de redes neurais artificiais. Ela é definida como uma função estritamente crescente que exibe um balanço entre o comportamento linear e o comportamento não-linear, sendo definida por:

$$\varphi_3(v,a)=\frac{1}{1+e^{-av}}$$

onde ***a*** é o parâmetro de declividade da função logística.

A derivada da Função Logística é definida por:

$$\varphi_3'(v,a)=a\varphi_3(v,a)(1-\varphi_3(v,a))$$

Segundo Xavier (XAVIER, 2005), a Função Logística oferece a importante flexibilidade dada por sua inclinação na origem, $\varphi'_3(0,a)=a/4$, ser variável com o parâmetro ***a***. Através da variação deste parâmetro ***a*** foram obtidos os gráficos de funções Logísticas com diferentes declividades, que podem ser vistos na Figura 3. Além disso, a Função Logística apresenta propriedades de simetria e de completa diferenciabilidade, ou seja, pertence à classe de funções c^{∞}. Variando o parâmetro a foram obtidas as derivadas da função logística de diferentes declividades, apresentadas na Figura 4.

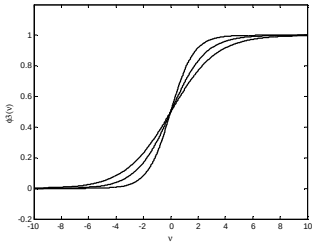


Figura 3: Função Logística – Efeito da variação do parâmetro ***a***

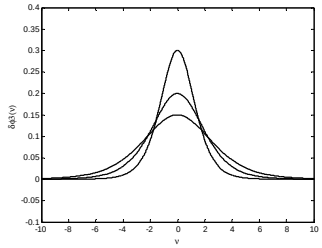


Figura 4: Derivadas da Função Logística variando o parâmetro ***a***

d) Função de Elliott (ELLIOTT,1993) (XAVIER, 2005)

Esta função, apresentada na Figura 5, é definida por:

$$\varphi_4(v)=\left(\frac{v}{1+|v|}+1\right)/2$$

A sua derivada, apresentada na Figura 6, é definida por:

$$\varphi_4'(v)=\frac{1}{2(1+|v|)^2}$$

Ela apresenta a inclinação de sua derivada na origem invariante, $\varphi'_4(0)=1/2$, independente de qualquer transformação de escala, fato que limita fortemente a flexibilidade dessa função e seu decorrente uso prático.

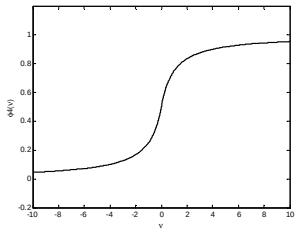


Figura 5: Função de Elliot

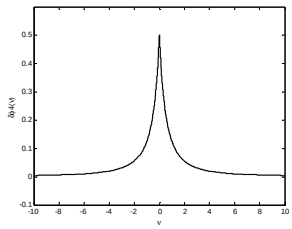


Figura 6: Derivada da Função de Elliot

e) Função Bi-Hiperbólica (XAVIER, 2005)

A função Bi-Hiperbólica Assimétrica em sua forma mais geral é definida por:

$$\varphi_5(v,\lambda,\tau_1,\tau_2) = \sqrt{\lambda^2(v+1/4\lambda)^2+\tau_1^{\,2}}) - \sqrt{\lambda^2(v-1/4\lambda)^2+\tau_2^{\,2}}) + 1/2\,,$$

Sendo sua derivada é definida por:

$$\varphi_5'(v,\lambda,\tau_1,\tau_2) = \frac{\lambda^2(v+1/4\lambda)}{\sqrt{\lambda^2(v+1/4\lambda)^2+\tau_1^{\,2}}} - \frac{\lambda^2(v-1/4\lambda)}{\sqrt{\lambda^2(v-1/4\lambda)^2+\tau_2^{\,2}}}$$

A função $\varphi_5(\bullet,\lambda,\tau_1,\tau_2)$ apresenta a desejada propriedade de possuir diferenciabilidade infinita, ou seja, pertence à classe de funções c^{∞}, o que permitirá a sua utilização em algoritmos de otimização mais robustos, além de apresentar as seguintes propriedades triviais consentâneas às demais funções de ativação:

$$\lim_{v\to-\infty} \varphi_5(v,\lambda,\tau_1,\tau_2) = 0$$

$$\lim_{v\to\infty} \varphi_5(v,\lambda,\tau_1,\tau_2) = 1$$

$$\lim_{v\to-\infty} \varphi'_5(v,\lambda,\tau_1,\tau_2) = 0$$

$$\lim_{v\to\infty} \varphi'_5(v,\lambda,\tau_1,\tau_2) = 0$$

Se considerarmos o caso particular obtido igualando-se os valores dos parâmetros $\tau_1 = \tau_2 = \tau$, a função $\varphi_5(\bullet,\lambda,\tau_1,\tau_2) \underset{=}{\Delta} \varphi_5(\bullet,\lambda,\tau)$, 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 da Figura 7 e Figura 8.

$$\varphi_5(v,\lambda,\tau) = \sqrt{\lambda^2(v+1/4\lambda)^2+\tau^2}) - \sqrt{\lambda^2(v-1/4\lambda)^2+\tau^2}) + 1/2$$

$$\varphi'_5(v,\lambda,\tau) = \frac{\lambda^2(v+1/4\lambda)}{\sqrt{\lambda^2(v+1/4\lambda)^2+\tau^2}} - \frac{\lambda^2(v-1/4\lambda)}{\sqrt{\lambda^2(v-1/4\lambda)^2+\tau^2}}$$

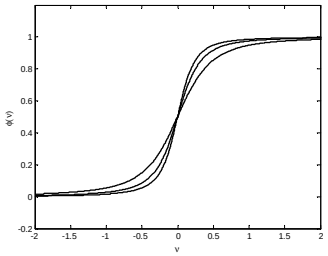


Figura 7: Curvas Bi-Hiperbólica variando λ com τ fixo

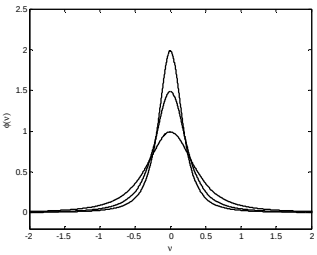


Figura 8: Derivadas das Curvas Bi-Hiperbólica variando λ com τ fixo

A função $[\varphi_5(v,\lambda,\tau) - 1/2]$ é anti-simétrica, ou seja:

$$\varphi_5(v,\lambda,\tau) - 1/2 = -[\varphi_5(-v,\lambda,\tau) - 1/2]$$

No ponto $v = 0$, são observados os seguintes valores para a função φ_5 e sua derivada:

$$\varphi_5(0, \lambda, \tau) = 1/2$$

$$\varphi'_5(0, \lambda, \tau) = \frac{\lambda}{2\sqrt{1/16 + \tau^2}},$$

$$\lim_{\tau \to 0} \varphi'_5(0, \lambda, \tau) = 2\lambda$$

Na figura 7, são mostrados gráficos da forma simétrica da função Bi-Hiperbólica correspondentes a três valores diferentes 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. Dessa forma pode-se associar o parâmetro λ à inclinação da função na origem.

A função $\varphi_5(\bullet, \lambda, \tau)$ apresenta, ademais, os seguintes comportamentos assintóticos:

$$\lim_{\lambda \to \infty} \varphi_5(v, \lambda, \tau) = \varphi_1 x(v)$$

$$\lim_{\tau \to 0} \varphi_5(v, \lambda, \tau) = \varphi_2\left(v, \frac{1}{4\lambda}\right)$$

Como bem ressalta Xavier (XAVIER, 2005), a existência de dois parâmetros, um a mais que as demais funções de ativação, possibilita a essa função dispor de uma maior flexibilidade para representar mais adequadamente os fenômenos normalmente modelados com redes neurais.

Numa rede neural multicamadas, por exemplo, essa maior flexibilidade certamente fornece à função de ativação Bi-Hiperbólica o poder de aproximar qualquer função de uma forma mais sintética, com um menor número de neurônios. Através da manipulação conveniente de seus parâmetros, a função $\varphi_5(\bullet, \lambda, \tau)$, oferece também a possibilidade de poder enfrentar mais convenientemente o desastroso fenômemo de saturação, além de poder evitar um indesejado mínimo local. Um forte indicador destas possibilidades pode ser observado no gráfico de sua derivada, na Figura 8, onde se pode ver que ela apresenta uma taxa de variação do crescimento bem mais acentuada do que o das derivadas das demais funções.

Arquitetura da rede neural

O projeto de uma rede neural artificial começa com a seleção de uma arquitetura apropriada e com o treinamento através dos exemplos e de um algoritmo específico. Esta fase é a chamada de aprendizagem. Em seguida é feita a avaliação com dados não usados no treinamento para determinar o seu desempenho nesta tarefa específica. Esta fase é a chamada de generalização. O projeto de uma rede neural artificial é baseado diretamente nos dados do mundo real, fazendo com que a rede forneça um modelo implícito do ambiente no qual está inserida, além de realizar a função de processamento de informações.

Para uma rede neural do tipo MLP o dimensionamento das camadas de entrada e de saída será sempre determinado pela natureza do próprio problema, enquanto que a determinação de quantas camadas ocultas e de quantos neurônios estas devem possuir, não é uma tarefa que permita uma resposta exata. Existem soluções aproximadas ou heurísticas, que procuram estimar estes valores. Estas heurísticas expõem sempre o compromisso entre a convergência e a generalização da rede. Considera-se Convergência a capacidade da rede de aprender todos os padrões de entrada usados no seu treinamento. Uma rede muito pequena em relação ao problema em análise não será capaz de aprender os dados de treinamento do problema, ou seja, a rede não possuirá parâmetros ou pesos sinápticos suficientes (HECHT-NIELSEN, 1989) (HAYKIN, 2001).

Generalização é a capacidade da rede neural de responder adequadamente a padrões fora dos usados no treinamento. Uma rede muito grande, com número de neurônios muito superior ao necessário, não responderá corretamente a estes novos padrões e perderá a capacidade de generalizar, uma vez que, durante o processo de treinamento o ajuste dos pesos sinápticos da rede

a levarão a memorizar especificamente estes vetores de entrada além do ruído presente nestes dados de treinamento.

A capacidade de generalização de uma rede neural é afetada pelo tamanho e eficiência dos dados de treinamento, pela arquitetura da rede e número de processadores nas camadas ocultas e pela complexidade do problema. Na prática, as heurísticas são utilizadas em conjunto com séries de tentativas e ajustes na arquitetura e definições da rede. O principal objetivo é obter uma rede que generalize, ao invés de memorizar os padrões usados no treinamento (STATHAKIS, 2009), (HORNIK, 1989) e (HECHT-NIELSEN, 1989).

Aprendizagem

A propriedade mais importante de uma Rede Neural Artificial é a sua capacidade de aprender a partir do seu ambiente e melhorar seu desempenho através deste aprendizado. Isto se resume no problema de obter um conjunto de parâmetros livres que permita à rede atingir o desempenho desejado. Neste processo a rede é estimulada pelo ambiente e sofre mudanças em seus parâmetros livres como resultado deste estímulo. Devido às mudanças ocorridas em sua estrutura interna, ela passa a responder de uma nova forma ao ambiente.

O tipo de aprendizagem é determinado pela forma através da qual é efetuada a mudança nos parâmetros. Os dois paradigmas básicos de aprendizagem são o aprendizado através de um tutor (Aprendizado Supervisionado) e o aprendizado sem um tutor (Aprendizado Não-Supervisionado). Uma terceira forma chamada de Aprendizagem por Reforço utiliza um crítico.

No Aprendizado Supervisionado, uma série de padrões, representados pelos vetores de entrada, é associada com os resultados desejados como resposta e é apresentado à rede. Os parâmetros internos da rede, chamados de pesos sinápticos, são alterados sistematicamente de forma a aproximar os resultados obtidos aos das respostas desejadas. Este procedimento consiste em minimizar os erros obtidos na comparação entre os resultados desejados e os calculados para os padrões usados no treinamento**.** (HAYKIN, 2001).

Algoritmo de Backpropagation

O treinamento de um Perceptron de Múltiplas Camadas (MLP) consiste em ajustar os pesos e os thresholds (bias) de suas unidades para que a classificação desejada seja obtida. Quando um padrão é inicialmente apresentado à rede, ela produz uma saída e, após medir a distância entre a resposta atual e a desejada, são realizados os ajustes apropriados nos pesos de modo a reduzir esta distância. Este procedimento é conhecido como Regra Delta.

Esse tipo de rede apresenta soluções para funções linearmente não-separáveis e necessita de um algoritmo de treinamento capaz de definir de forma automática os pesos. O algoritmo mais utilizado para o treinamento destas redes MLP é uma generalização da Regra Delta denominada de Algoritmo de Backpropagation.

Durante o treinamento com o algoritmo Backpropagation, a rede opera em uma seqüência de dois passos. No primeiro, um padrão é apresentado à camada de entrada da rede e o sinal resultante flui através dela, camada por camada, até que a resposta seja produzida pela camada de saída. No segundo passo, a saída obtida é comparada à saída desejada para esse padrão particular. Se esta não estiver correta, o erro é calculado. Este erro é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo modificados à medida que o erro é retropropagado.

Na Regra Delta padrão é implementado um gradiente descendente no quadrado da soma do erro para funções de ativação lineares. Entretanto, como a superfície do erro pode não ser tão simples, as redes ficam sujeitas aos problemas de mínimos locais.

A Regra Delta Generalizada, ou Backpropagation, funciona quando são utilizadas na rede unidades com uma função de ativação semi-linear, que é uma função diferenciável e não decrescente.

A Taxa de Aprendizado é uma constante de proporcionalidade no intervalo [0,1], pois este procedimento de aprendizado requer apenas que a mudança no peso seja proporcional à meta. Entretanto, como o verdadeiro gradiente descendente requer que sejam tomados passos

infinitesimais, quanto maior for essa constante, maior será a mudança nos pesos, aumentando a velocidade do aprendizado. Tal situação pode levar a uma oscilação do modelo na superfície de erro. Procura-se, então, utilizar a maior taxa de aprendizado possível que não leve a uma oscilação, resultando em um aprendizado mais rápido. O treinamento das redes MLP com backpropagation pode demandar muitos passos no conjunto de treinamento, resultando em um tempo de treinamento consideravelmente longo. Se for encontrado um mínimo local, o erro para o conjunto de treinamento pára de diminuir e estaciona em um valor maior que o aceitável (HAYKIN, 2001).

A utilização da função de ativação Bi-Hiperbólica apresenta uma vantagem grande por possuir dois parâmetros que ajudam a obter um ajuste mais preciso. Outro fator que beneficia este uso é dado pela mudança maior na inclinação de sua derivada, conforme pode ser visto na Figura 8, o que contribui para diminuir o problema da saturação, que ocorre muitas vezes no treinamento das redes neurais (XAVIER, 2005).

Estudo comparativo

A proposta apresentada neste trabalho para o problema de otimizar a eficiência e a taxa de convergência do algoritmo de Backpropagation, prevê a utilização de uma nova função de ativação, a Função Bi-Hiperbólica, com características que atendem às necessidades do algoritmo de backpropagation. Ela oferece a vantagem de possibilitar maior flexibilidade na representação dos fenômenos modelados. Conta com o uso de dois parâmetros, um a mais do que nas funções tradicionalmente utilizadas para esta finalidade. Isto implica em melhor enfrentar o problema da saturação, além de permitir tratamento para evitar os mínimos locais. Outra vantagem, observada empiricamente, é a de ser computacionalmente muito mais rápida de ser avaliada do que a função logística, pois apesar de também poder ser considerada uma função sigmoidal, o seu cálculo é feito diretamente, enquanto que a função logística tem como seu principal algoritmo de cálculo usado em sistemas computacionais ser o de expansão da Série de Taylor, conforme apresentado na seguinte equação (HAHN, 1993):

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

Além disso, a sua maior flexibilidade possibilita a capacidade de poder aproximar qualquer função de uma forma mais sintética, permitindo a utilização de um menor número de neurônios. Isto melhora ainda mais o desempenho do algoritmo backpropagation, agindo diretamente na topologia da rede (XAVIER, 2005).

Para permitir uma avaliação computacional das características descritas, comparando-as com a função de ativação tradicionalmente utilizada, foi desenvolvido um protótipo em MATLAB, que através de uma interface gráfica, apresenta as funções necessárias aos treinamentos e testes, permitindo a execução do ciclo de treinamento e a verificação dos resultados obtidos pela comparação de desempenho com o modelo usando a função logística.

Foi adotada no protótipo uma rede neuronal artificial do tipo MLP, progressiva e completamente conectada com a arquitetura apresentada a seguir. O número de nós fonte na camada de entrada da rede é determinado pela dimensionalidade do espaço de observação, que é responsável pela geração dos sinais de entrada. O número de neurônios na camada de saída é determinado pela dimensionalidade requerida da resposta desejada. A existência de camadas ocultas se deve para permitir a extração de estatísticas de ordem superior de algum desconhecido processo aleatório subjacente, responsável pelo "comportamento" dos dados de entrada, processo sobre o qual a rede está tentando adquirir conhecimento. Este é um valor arbitrário e pode variar em função da análise do desempenho do modelo. A decisão de utilizar uma única camada oculta é baseada na demonstração feita por Robert Hecht-Nielsen de que, teoricamente, o uso de três camadas é sempre suficiente para a aproximação de qualquer função. Entretanto, ele também resalta que, em se tratando de problemas do mundo real, esta aproximação por apenas três camadas poderá resultar na necessidade de uma quantidade de neurônios na camada oculta extremamente grande, fazendo com que seja mais prático o uso de um maior número de camadas para a obtenção de uma solução tratável. Para a determinação do número de neurônios na camada

oculta foi adotada a heurística proposta por Hecht-Nielsen (HECHT-NIELSEN, 1989). Este é um dos parâmetros mais importantes na definição de uma RNA, uma vez que, quanto maior for esse valor, maior será o número de pesos a serem ajustados.

Para possibilitar a avaliação comparativa da função proposta, o protótipo desenvolvido faz o treinamento em duas redes distintas, com os mesmos parâmetros básicos e com o uso de funções de ativação diferenciadas. Uma das redes utiliza como função de ativação a Função Logística e, a outra rede, utiliza como função de ativação a Função Bi-Hiperbólica.

Para a execução dos testes foram utilizados os seguintes parâmetros em comum nos dois modelos:

a) Topologia inicial da Rede Neural:
   - Uma camada externa com 10 nós, um para cada uma dos nove atributos descritivos das características observadas e mais um para o controle do bias;
   - Uma camada escondida com 21 nós, definida com base na heurística proposta por Hecht-Nielsen (HECHT-NIELSEN, 1989).
   - Uma camada de saída com um nó;
b) Nível de Erro Médio Quadrático considerado: menor que 0,001;
c) Taxa de aprendizado: 0,05
d) Amostra usada no treinamento: 200 instâncias;
e) Amostra usada para avaliação do modelo: 483 instâncias;

Base de Dados para teste do modelo

Para possibilitar a obtenção de dados comparativos, foi utilizada a base de dados conhecida como "Wisconsin Breast Cancer Data", disponível no site *http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)* da *University of Wisconsin-Madison*. Ela tem sido bastante utilizada em artigos publicados na área médica e de reconhecimento de padões, facilitando as comparações com os resultados a serem obtidos (PRECHELT, 1994).

É uma base com um razoável número de amostras, atributos e padrões bem definidos, sendo formada por dados de amostras obtidas através da realização de biópsias em massas com suspeitas de malignidade, encontradas em exames de mamas humanas. Cada amostra apresenta um identificador e nove atributos descritivos das características observadas, que utilizam uma escala numérica padronizada. A cada amostra está associado o resultado da avaliação feita por especialistas, classificando-as como benignas (resultado negativo) ou malignas (resultado positivo). Foram utilizadas 683 amostras, sendo 444 classificadas como benignas (65 %) e 239 classificadas como malignas (35 %) (MANGASARIAN, 1990), (WOLBERG, 1990).

Resultados computacionais

Para a preparação inicial dos dados, foi feita uma aleatorização das instâncias, para evitar alguma tendência não conhecida devido a, por exemplo, a temporalidade da obtenção das amostras. Foi feita, também, uma normalização dos atributos originais para uma escala de valores entre zero e um. Nenhum destes procedimentos altera as características das amostras, visando apenas facilitar a visualização dos dados.

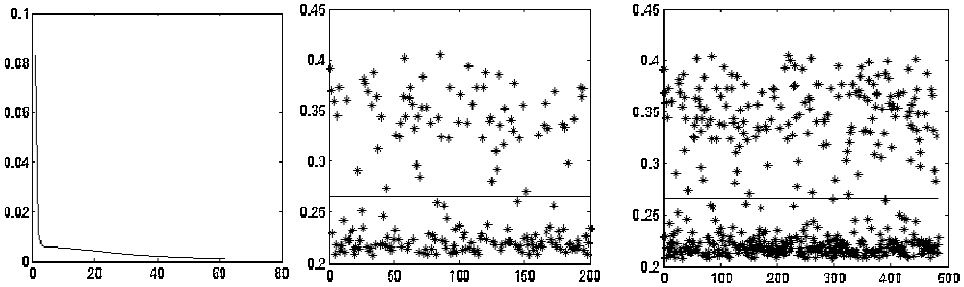
Para avaliar a influência do número de neurônios na camada oculta sobre o erro quadrático médio, foram feitos treinamentos independentes da rede utilizando arquiteturas contendo de desde 21 neurônios na camada oculta, valor obtido pelo uso da heurística proposta por Hecht-Nielsen (HECHT-NIELSEN, 1989), até o limite arbitrário de cinco neurônios ocultos.

Para evitar a influência da inicialização dos pesos feita com valores aleatórios, todos os testes forma realizados com os mesmos valores iniciais para os pesos das ligações entre os neurônios das diversas camadas.

Foi feito um teste de sensibilidade para os parâmetros em ambos os modelos. Assim, estão apresentados abaixo os parâmetros que ofereceram o melhor resultado em termos de acertos e de número de épocas (iterações usadas no treinamento com o conjunto de amostras destacado para tal fim).

Modelo com Função de Ativação usando a Curva Logística

O parâmetro variável da curva Logística que apresentou o melhor resultado, em termos de acertos, foi ***a*** igual a 0,3 que convergiu em 62 épocas, apresentando sete diagnósticos errados, o que corresponde a um percentual menor que 1,5% de respostas erradas. Os gráficos obtidos neste processamento estão apresentados na Figura 9.



(a) EQM (b) amostras-treinamento (c) amostras – validação

Figura 9: Saídas obtidas no processamento do modelo

Foram utilizadas arquiteturas contendo de 21 neurônios na camada oculta, até o limite arbitrário de cinco neurônios. Considerando-se apenas as redes que apresentaram o melhor resultado obtido, com sete diagnósticos errados em 483 instâncias avaliadas, obtivemos os valores apresentados na Figura 10. Podemos verificar que, em alguns casos, o mesmo resultado foi obtido por uma mesma arquitetura, mas com a utilização de parâmetros diferentes.

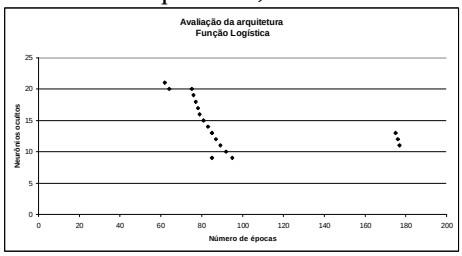
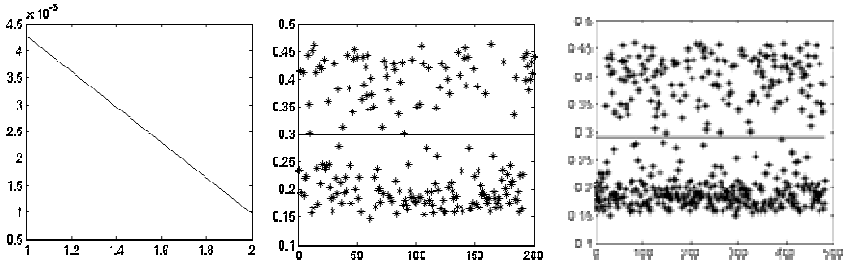


Figura 10: Avaliação das arquiteturas para a Função Logística

Modelo com ativação pela Função Bi-Hiperbólica

Para possibilitar uma avaliação comparativa do desempenho do modelo que utiliza a Função Bi-Hiperbólica, foram feitos testes variando conjuntamente o parâmetro λ, que pode ser associado com a inclinação da curva na origem, e o parâmetro τ, pode ser associado com o afastamento da curva às duas assíntotas horizontais.

Foram feitos os treinamentos e avaliações combinando entre si estes parâmetros. O melhor resultado obtido, em termos de acertos, foi a obtenção de sete diagnósticos errados. Isto foi obtido em 371 combinações no total, sendo que em 21 delas este resultado foi obtido com apenas duas épocas, e em 25 destas com apenas sete neurônios na camada oculta. Isto demonstra o enorme poder de convergência do modelo, bem como a sua capacidade de operar com uma rede de arquitetura com menos neurônios. Isto facilita o seu uso em ambientes computacionais com menos recursos disponíveis. Os gráficos obtidos neste processamento estão apresentados na Figura 11.



(a) EQM (b) amostras-treinamento (c) amostras – validação

Figura 11: Saídas obtidas no processamento do modelo com 7 neurônios ocultos

O processamento deste modelo com a variação dos parâmetros citada anteriormente permitiu, também a obtenção de outros resultados muito interessantes, como por exemplo, considerando como melhor resultado o número de épocas, importante no caso de sistemas computacionais mais lentos ou com necessidade de treinamento mais rápido, foram obtidas 68 combinações que convergiram em apenas uma época e que apresentaram entre oito e nove diagnósticos errados. Considerando-se apenas as redes que apresentaram o melhor resultado obtido, com sete diagnósticos errados em 483 instâncias avaliadas, obtivemos os valores apresentados na Figura 12 que mostra o número de épocas associado ao número de neurônios na camada oculta. Podemos verificar que, em alguns casos, o mesmo resultado foi obtido por uma mesma arquitetura, mas com a utilização de parametrização diferente.

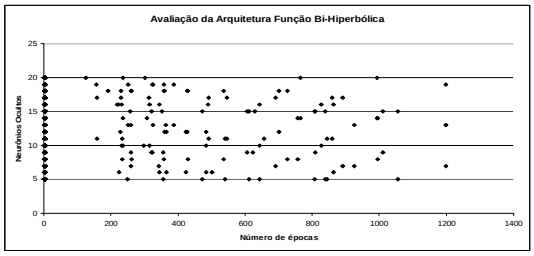


Figura 12: Avaliação das arquiteturas para a Função Bi-Hiperbólica

Estas características podem ser atribuídas a uma maior taxa de variação da derivada da função bi-hiperbólica em comparação com a da outra função de ativação usada. Isto pode ser observado na comparação a seguir com a derivada da função logística. Foram usadas curvas com a mesma inclinação na origem, apresentadas na Figura 13 e as respectivas derivadas na Figura 14, onde as linhas pontilhadas correspondem à função bi-hiperbólica.

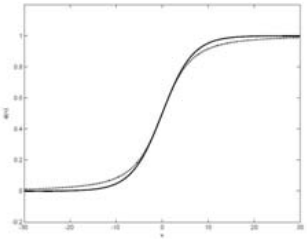


Figura 13: Curvas com a mesma inclinação na origem

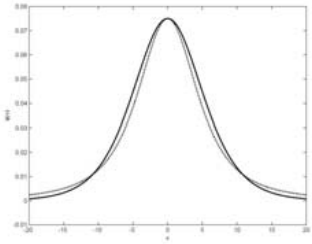
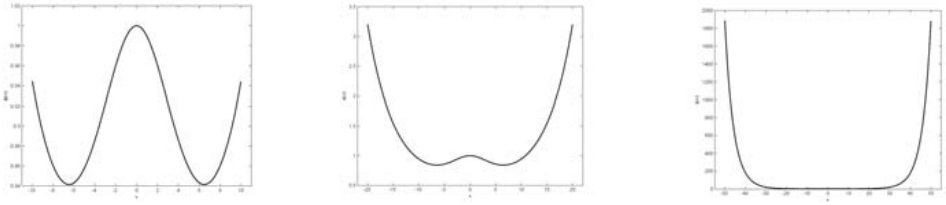


Figura 14: Derivadas das curvas com mesma inclinação na origem

Mantendo os mesmos parâmetros básicos e variando a amplitude dos dados de entrada foram obtidas as razões entre as derivadas das funções bi-hiperbólica e da logística, apresentadas na Figura 15. Pode ser visto que a razão entre estas derivadas mostra uma tendência que justifica as proposições feitas neste trabalho pois, depois de período de pequenas variações, esta razão cresce exponencialmente. Isto apoia nossa afirmação de que assim é evitada a saturação da função, acelerando o processo de convergência da rede neural.



(a) amplitude -10 to 10 (b) amplitude -20 to 20 (c) amplitude -50 to 50

Figura 15: Razão entre as derivadas das funções bi-hiperbólica e logística

Conclusões

Os resultados obtidos demonstraram a grande viabilidade de utilização da função de ativação Bi-Hiperbólica, ecoando as previsões de maior capacidade de generalização, convergência mais rápida (convergiu em um número de iterações de aproximadamente 3% do observado para o outro modelo), maior velocidade de cálculo e arquitetura de rede com menor número de neurônios (utilizou 1/3 do número de neurônios ocultos utilizado no modelo com a Função Logística).

Outro fator importante que se pode inferir dos resultados é que a atividade de configuração da arquitetura da rede com o uso desta função, que normalmente é obtida através de processos heurísticos e de tentativas e erros, fica facilitada uma vez que uma ampla combinação de parâmetros diferentes possibilita a obtenção dos resultados desejados.

Referências

ELLIOTT, David L. A Better Activation Function for Articial Neural Networks, Institute for Systems Research, ISR Technical Report TR 93-8, 1993.

FYFE, Colin. Artificial Neural Networks, Departament of Computing and Information Systems. The University of Paisley, 2000.

HAHN, Brian D., *Fortran 90 for Scientists and Engineers*. University Press, Cambridge, London. 1993

HAYKIN, S. *Redes Neurais: princípios e prática*. 2. ed. Porto Alegre, Bookman, 2001.

HECHT-NIELSEN , R. Theory of the Backpropagation Neural Network; Neural Networks, 1989. IJCNN., International Joint Conference . pp 593 – 605. Washington, USA

HORNIK, K. Multilayer Feedforward Networks are Universal Approximatorz, Neural Networks, Vol. 2, pp. 359-366, 1989.

KÓVACS, Z. L., *Redes neurais artificiais: fundamentos e aplicações*. São Paulo, Edição Acadêmica, 1996.

MANGASARIAN, O. L., Wolberg, W. H., "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

OTAIR, M. A., SALAMEH, W. A., Speeding Up Back-Propagation Neural Networks, in Proceedings of the 2005 Informing Science and IT Education Joint Conference, Flagstaff, Arizona, USA.

PRECHELT, L.,Proben1 - A Set of Neural Network Benchmark Problems and Benchmarking Rules, Universit at Karlsruhe, Technical Report 21/94, 1994

SCHIFFMANN W., JOOST M., WERNER, R., Optimization of the Backpropagation Algorithm for Training Multilayer Perceptrons, University of Koblenz, Institute of Physics, Koblenz, 1994.

STATHAKIS, D. How many hidden layers and nodes? International Journal of Remote Sensing Vol. 30, No. 8, 20 April 2009, 2133–2147

WOLBERG, W. H., MANGASARIAN, O. L., "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

XAVIER, Adilson Elias, Uma Função de Ativação para Redes Neurais Artificiais Mais Flexível e Poderosa e Mais Rápida. Learning and Nonlinear Models – Revista da Sociedade Brasileira de Redes Neurais (SBRN), Vol. 1, No. 5. PP. 276-282, 2005.