INICIALIZAÇÃO DE REDES NEURAIS MLP BASEADA EM DIVISÕES SIMÉTRICAS

L. D. Tavares*, R. R. Saldanha*, D. A. G. Vieira†

*Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Minas Gerais
Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brasil

†ENACOM - Handcrafted Technologies
Parque Tecnológico de Belo Horizonte (BH-TEC)
Rua Professor José Vieira de Mendonça, 770, 31310-260, Belo Horizonte, MG, Brasil

Emails: tavares@dcc.ufmg.br, rodney@cpdee.ufmg.br, douglas.vieira@enacom.com.br

Abstract— The main goal of a learning machine is suitable mapping signal input-output presented. For such, a criterion is used for learning based on empiric error minimization. However, it is known that the initial values of the weights of the MLP (Multi-layer perceptron) neural network may influence the probability of convergence and learning rate. This study aims to investigate another criterion which is the conditioning number of generated space by the hidden layer and proposes a method that generates a low condition number.

Keywords— Weight initialization, conditioning number, neural network

Resumo— O principal objetivo de uma máquina de aprendizagem é o adequado mapeamento do sinal entrada-saída apresentado à ela. Para tal, é utilizado um critério de aprendizagem baseado em minimização do erro empírico. No entanto, é sabido que os valores iniciais dos pesos de uma rede neural MLP (Multi-layer perceptron)podem influenciar na probabilidade convergência bem como na taxa de aprendizagem. O presente trabalho visa investigar outro critério que é o número de condicionamento do espaço gerado pela camada oculta e propor um método que gere um número de condicionamento baixo.

Palavras-chave— Inicialização de pesos, número de condicionamento, redes neurais.

1 Introdução

O objetivo primário das redes neurais MLP é, antes de tudo, o aprendizado do mapeamento da entrada-saída que é apresentado durante a etapa de treinamento até que um determinado critério seja atendido. De modo geral é utilizada uma medida de erro no qual se compara a resposta gerada pela máquina de aprendizagem com a resposta do conjunto de dados de treinamento. Tal princípio é chamado de Minimização do Risco Empírico (ERM - *Empiric Risk Minimization*). No entanto, outros critérios também devem ser atendidos no momento da aprendizagem, que, no contexto do presente trabalho, se trata da qualidade do ponto de partida da aprendizagem, isto é, a adequada determinação dos valores iniciais dos parâmetros livres da máquina.

A motivação para que outros critérios, além da minimização do risco empírico, sejam atendidos surge das pesquisas na aprendizagem estatística nos anos 60, ganhando, porém popularidade nos anos 90 com o desenvolvimento das chamadas máquinas de vetor de suporte - SVMs. Vapnik (1995) apresenta o conceito de dimensão VC baseado na capacidade de expressão da máquina de aprendizagem em relação ao conjunto de dados apresentados a ela.

O modelo escolhido para o presente trabalho é uma rede neural do tipo MLP na forma:

$$f(\mathbf{x}, \mathbf{w}) = \varphi(\mathbf{x}^t \mathbf{w}_1)^t \mathbf{w}_2 \tag{1}$$

Onde $\mathbf{x}$ é o vetor de entrada, $\mathbf{w}_1$ e $\mathbf{w}_2$ são os pesos associados à camada oculta e à camada de saída, respectivamente, $(\cdot)^t$ é o operador de transposição e $\varphi(\cdot)$ é a função de ativação do tipo logística aplicada na camada oculta, na forma:

$$\varphi(u) = \frac{1}{1 + exp(-u)} \tag{2}$$

O modelo de rede neural apresentado em (1) é especialmente interessante uma vez que atende os requisitos do teorema da aproximação universal, conforme apresentado em (Cybenko, 1989) e (Hornik et al., 1989). Considere ainda que os termos de polarização estão presentes no modelo.

É possível observar que a camada de saída da rede neural apresentada em (1) é um sistema linear do tipo $\mathbf{A}\theta + \epsilon = \mathbf{b}$, sendo $\mathbf{A}$ a matriz de coeficientes, θ o vetor parâmetros e $\mathbf{b}$ o vetor de respostas reais. A partir de uma simples separação da expressão, substituindo os valores gerados por $f(\mathbf{x}, \mathbf{w})$ pelos valores que estamos buscando $\mathbf{y}$, e admitindo um resíduo ϵ ao tentar explicar o fenômeno de maneira linear, obtemos:

$$\underbrace{\overbrace{\varphi(\mathbf{x}^t \mathbf{w}_1)^t}^{\mathbf{A}} \overbrace{\mathbf{w}_2}^{\theta}}_{\hat{\mathbf{y}}} + \epsilon = \overbrace{\mathbf{y}}^{\mathbf{b}} \tag{3}$$

Nesse caso os parâmetros $\mathbf{w}_2$ que minimizam $||\epsilon||_2^2$ podem ser encontrados através do método de mínimos quadrados, da seguinte forma (Aguirre, 2004, pág 221), (Boyd and Vandenberghe, 2004, pág 293):

$$\mathbf{w}_2 = ((\varphi(\mathbf{x}^t\mathbf{w}_1)^t)^t(\varphi(\mathbf{x}^t\mathbf{w}_1)^t))^{-1}(\varphi(\mathbf{x}^t\mathbf{w}_1)^t)^t\mathbf{y} \tag{4a}$$

$$= (\varphi(\mathbf{x}^t\mathbf{w}_1)\varphi(\mathbf{x}^t\mathbf{w}_1)^t)^{-1}\varphi(\mathbf{x}^t\mathbf{w}_1)\mathbf{y} \tag{4b}$$

$$= \varphi(\mathbf{x}^t\mathbf{w}_1)^+\mathbf{y} \tag{4c}$$

Onde $\varphi(\mathbf{x}^t\mathbf{w}_1)^+$ é a pseudo-inversa do espaço gerado pelos neurônios da camada oculta. É possível constatar que a possibilidade de solução da equação (4c) depende fundamentalmente de como os parâmetros $\mathbf{w}_1$ são inicializados. Em certos casos o número de condicionamento do espaço gerado pela camada oculta seja tão alto que não seja possível invertê-lo, dependendo dos parâmetros $\mathbf{w}_1$.

Dessa forma, o objetivo principal do presente artigo é apresentar um método de inicialização dos pesos de redes neurais MLP, mais especificamente, da camada oculta do modelo (1), de modo que o número de condicionamento do espaço gerado pela camada oculta seja baixo.

A justificativa do presente estudo se dá pela influência que a inicialização possui, principalmente, em três aspectos (Fernández-Redondo and Hernández-Espinosa, 2000): i) desempenho da aprendizagem, ii) pela probabilidade de convergência e iii) pela generalização.

O trabalho está organizado como se segue: é feita uma breve contextualização e exposição do objetivo na presente seção e, em seguida, na seção 2, são apresentados os trabalhos relacionados com a inicialização de redes neurais MLP. Posteriormente, na seção 3, é descrito o método proposto de inicialização baseado divisões simétricas do espaço oculto. Por fim, nas seções 4 e 5 são apresentados os experimentos realizados e os trabalhos futuros, respectivamente.

2 Trabalhos relacionados

Do ponto de vista de otimização, inicializar uma rede neural MLP de forma a diminuir o número de épocas para a aprendizagem (visando utilizar o método de aprendizagem de retro-propagação) é um problema equivalente a encontrar o ponto de partida adequado para um método clássico de otimização baseado em gradiente. Nesse sentido, encontrar o conjunto ideal de valores para os pesos da rede neural, isto é, aquele que leva ao valor de mínimo global é um problema NP-Difícil (Bishop, 1996).

Diversos pesquisadores concentraram seus esforços de modo a desenvolver técnicas e mecanismos que possam acelerar o processo de aprendizagem, baseado na inicialização dos pesos da rede neural MLP.

Tais métodos podem ser, informalmente, categorizados em: i) métodos puramente probabilísticos, ii) métodos probabilísticos limitados pela estrutura da rede neural ou pelas características dos dados, iii) métodos determinísticos, iv) métodos híbridos e v) métodos baseados em computação evolucionária.

Os métodos puramente probabilísticos são aqueles que nada conhecem sobre a estrutura da rede e/ou como os dados estão distribuídos. Desse modo, os pesos da rede neural são inicializados através de uma distribuição aleatória em um dado limite ou configuração. Nessa categoria está o método ingênuo que inicializa os pesos através de uma distribuição uniforme $\mathbf{w} \approx U(-0.5, 0.5)$. Fahlman (1988) sugere os intervalos $\mathbf{w} \approx U(-1, 1)$ à $\mathbf{w} \approx U(-4, 4)$. Já Thimm and Fiesler (1997) sugerem $\mathbf{w} \approx U(-0.77, 0.77)$. Por outro lado Haffner et al. (1988) sugerem uma distribuição normal com média 0 e variância unitária, na forma $\mathbf{w} \approx N(0, 1)$.

Outros trabalhos, porém, tentam relacionar a estrutura da rede, isto é, número de entrada por neurônio, número de neurônios por camada escondida ou funções de ativação associadas. Nessa categoria dois trabalhos merecem destaque, sendo o primeiro deles o método descrito por Nguyen and Widrow (1990). O método realiza a inicialização em duas etapas: a primeira realiza uma inicialização probabilística e a segunda um ajuste conforme a estrutura da rede, na forma:

1. $\mathbf{w} \approx U(-1, 1)$
2. $\mathbf{w} = \frac{\beta\mathbf{w}}{\gamma}$

Onde $\beta = 0.7^{\frac{1}{d_{in}}}$, d_{in} é o número de entradas do neurônio e $\gamma = \sqrt{\sum(w)^2}$. A justificativa dada pelos autores é que todos os valores de entrada devem excitar no máximo 70% da região ativa da função logística. Tal método está implementado em grande parte das ferramentas de inteligência computacional.

O segundo método dessa categoria, também muito citado na literatura, é o chamado SCAWI (*Statistically Controlled Activation Weight Initialization*), descrito em Drago and Ridella (1992). A principal proposta do método é permitir que uma pesquena porcentagem, controlada de maneira estatística, estejam na região não ativa dos neurônios. Além disso, o método possui processos distintos para inicialização dos pesos da camada oculta e da camada de saída, na forma:

1. $\mathbf{w}_1^{i,j} = \frac{1.3}{\sqrt{1+d_{in}v^2}} r_{i,j}, \forall i = 1, 2, ..., m+2; j = 1, 2, ..., h$
2. $\mathbf{w}_2^i = \frac{1.3}{1+0.3.h} r_i, \forall i = 1, 2, ..., h$

Onde $v = \sum(w)^2/m$ e $r \approx U(-1, 1)$.

Na categoria de métodos determinísticos destacam-se aqueles que criam uma árvore de decisão e, posteriormente, a transforma em uma rede MLP, conforme apresentado em (Banerjee, 1994) e (Ivanova and Kubat, 1995). Destaca-se também a inicialização através da combinação de diversas combinações de neurônios previamente ajustados, chamados de protótipos, conforme descrito, pioneiramente, em (Denoeux and Lengellé, 1993).

Já para a categoria de métodos híbridos destaca-se o método descrito por Castillo et al. (2006), no qual a rede neural é transformada em uma função de sensibilidade e, uma vez resolvido o problema de sensibilidade a rede MLP não necessita de treinamento, segundo os autores.

Por fim, a categoria dos métodos baseados em computação evolucionária, como aqueles baseados em algoritmos genéticos (de Castro et al., 1998) e aqueles baseados em *simulated annealing* e algoritmos imunológicos (de Castro and Zuben, 2001).

É possível observar que os métodos revistos possuem duas preocupações principais: i) evitar que estejam em regiões saturadas da função logística e ii) minimização do número de épocas necessárias para a aprendizagem. A seguir será apresentado o método proposto que possui preocupações distintas das já citadas.

3 Método proposto

O método proposto possui como aspecto o central a divisão dos pontos do espaço de entrada entre os neurônios existentes na camada oculta da rede. No presente trabalho tal característica será apresentada a partir de uma interpretação geométrica. Para tal considere uma função geradora unidimensional que seja apresentada à uma rede que possua uma camada oculta com apenas 1 neurônio. É intuitivo que tal neurônio seja responsável por toda extensão do espaço de entrada, no caso o intervalo $x \in [-10, 10]$, conforme apresentado na figura 1.

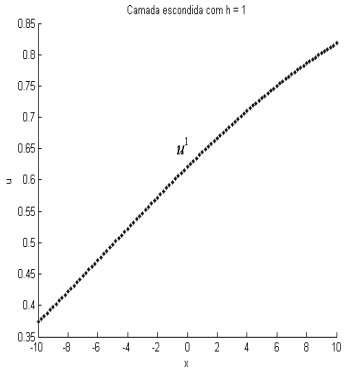

Figura 1: Representação da saída da camada escondida com $h = 1$ no espaço hipotético de entrada no intervalo $x \in [-10, 10]$.

Seguindo a ideia do método proposto, considerando uma rede neural com uma camada oculta com 5 neurônios, a distribuição do espaço de entrada deve estar conforme apresentado na figura 2.

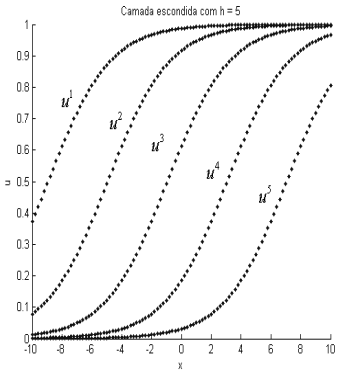

Figura 2: Representação da saída da camada escondida com $h = 1$ no espaço hipotético de entrada no intervalo $x \in [-10, 10]$.

É possível observar que o método proposto admite que existam pontos do espaço de entrada que estejam em regiões não ativas da função de ativação. É esperado com esse método que o espaço gerado pela camada oculta seja mais organizado e, com isso, o número de condicionamento, buscando tornar a solução de mínimos quadrados mais estável.

A descrição a seguir aponta as principais etapas para se obter tal efeito de distribuição simétrica dos pontos entre os neurônios, para o caso unidimensional:

1. Ordene os pontos no espaço de entrada;
2. Recupere o intervalo de pontos no qual cada neurônio ficará responsável;
3. Para cada neurônio da camada oculta:
   (a) Encontre os pontos que fazem parte do intervalo do dado neurônio;
   (b) Ajuste os pesos do neurônio de forma que o menor ponto se ajuste ao limite inferior da região ativa da função de ativação e que o maior ponto se ajuste ao limite superior da região ativa, os pontos intermediários devem ser ajustados via mínimos quadrados;

O procedimento para o caso multidimensional é facilmente generalizável. O resultado obtido com o método proposto, considerando um espaço bidimensional e uma rede neural com 5 neurônios na camada escondida no intervalo $\mathbf{x} \in [-10, 10]$, é apresentado na figura 3.

O objetivo do método é que ele seja capaz de gerar um espaço na camada oculta com baixo número de condicionamento, de modo que a solução dos pesos do espaço de saída seja possível.

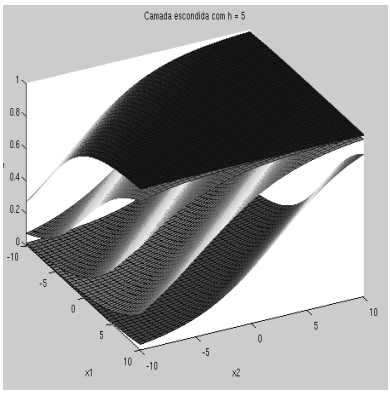

Figura 3: Distribuição espacial das saídas das funções de ativação de acordo com o método proposto. Nesse experimento existem 5 neurônios na camada oculta.

Uma característica importante do método que ele se comporta de maneira determinística.

A seção a seguir apresenta os resultados obtidos, em relação ao número de condicionamento, quando comparado á outros métodos de inicialização de pesos de redes neurais.

4 Experimentos realizados

O método proposto foi comparado com dois outros métodos bastante citados na literatura: Nguyen and Widrow (1990) e Drago and Ridella (1992). O modelo de rede neural utilizada para o experimento é no qual houve variação no número de neurônios na camada oculta no intervalo $h = 11, 2, ..., 90$ e com o propósito de regressão de função e predição de série temporal. Foram feitas 40 replicações para cada configuração de rede neural.

Foram utilizadas 2 funções, sendo:

1. Função de teste 01 - 1.000 amostras (figura 4):

$$f(x) = \frac{sin(x)}{x} + \xi, \xi \approx N(0, 0.1), x \approx U(-10\pi, 10\pi) \quad (5)$$

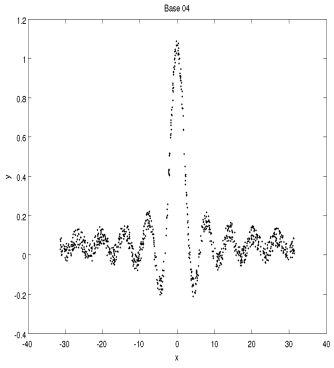

Figura 4: Função de teste 01

2. Função de teste 02 - Serie temporal de Mackey-Glass - 900 amostras (figura 5)(Wan et al., 2001):

$$f(x) = \frac{\partial x(t)}{\partial t} = -0.1x(t) + \frac{0.2x(t-\tau)}{1 + x(t-\tau)^{10}}, \tau = 17 \quad (6)$$

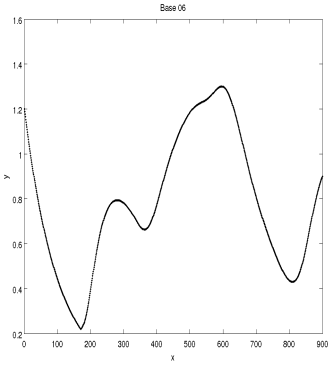

Figura 5: Função de teste 02

Nesse caso os regressores $x(-18)$, $x(-12)$, $x(-6)$ e $x(0)$ devem servir para predizer a variável $x(6)$.

As figuras 8, 7 e 8 apresentam, respectivamente, os resultados encontrados para o método Nguyen and Widrow (1990), Drago and Ridella (1992) e para o método proposto utilizando a base 1. Observe que os resultados são apresentados em escala logarítmica. Da mesma forma as figuras 11, 10 e 11 apresentam os resultados encontrados utilizando a base 2.

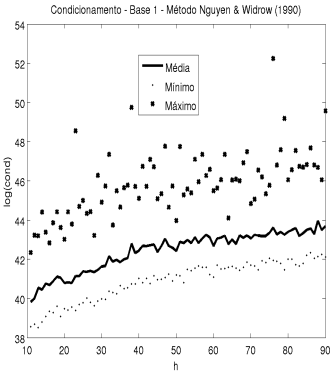

Figura 6: Número de condicionamento encontrado para o método de Nguyen & Widrow para a base 1. A linha contínua indica o resultado médio, a linha marcada por ponto "." representa os resultados mínimos e a linha marcada com "x" representa os resultados máximos encontrados.

Foram gerados 160 cenários de testes (80 configurações de neurônios ocultos e 2 bases verificadas). Os resultados demonstram que ao se comparar o número de condicionamento da camada

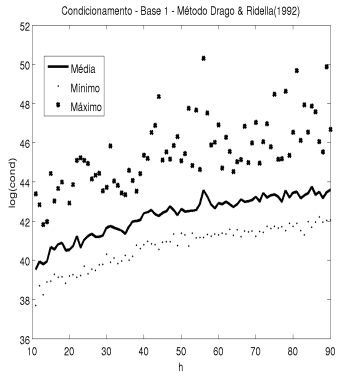

Figura 7: Número de condicionamento encontrato para o método de Drago & Ridella para a base 1.

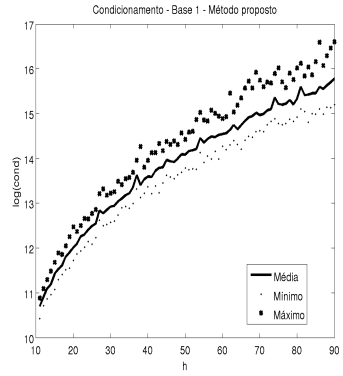

Figura 8: Número de condicionamento encontrado para o método proposto para a base 1.

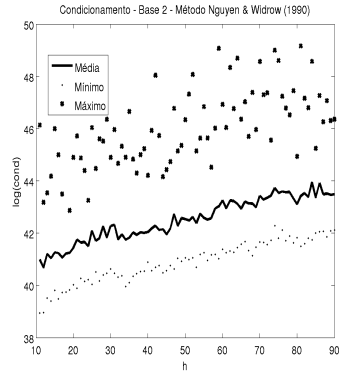

Figura 9: Número de condicionamento encontrado para o método de Nguyen & Widrow para a base 2.

oculta, o método proposto apresentou um resultado com diferenças significativas para todos os cenários, no qual, no melhor caso, a diferença chega a ser de 30 ordens de grandeza, para ambas as bases utilizadas para o teste.

Duas outras vantagens podem ser observadas: a primeira é que método é que ele possui baixa complexidade computacional e fácil implementação e a segunda é que o espaço gerado pela camada oculta é de fácil compreensão.

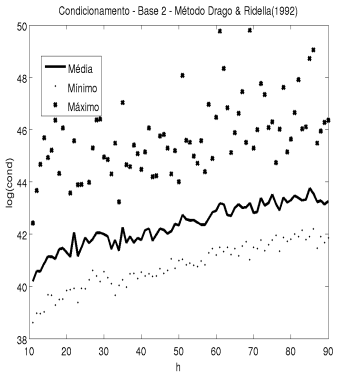

Figura 10: Número de condicionamento encontrato para o método de Drago & Ridella para a base 2.

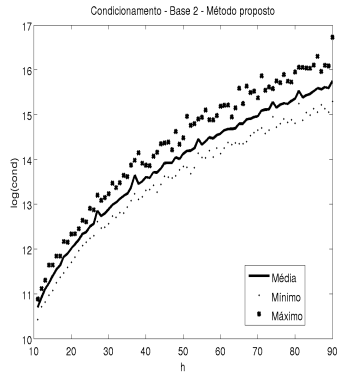

Figura 11: Número de condicionamento encontrado para o método proposto para a base 2.

Com isso, é possível vislumbrar novo métodos de inicialização e aprendizado que tenham foco não na minimização do risco empírico, e sim com o foco na organização dos dados na camada oculta e no controle de sensibilidade e generalização da rede neural como um todo. Nesse contexto, o controle do número de condicionamento do espaço gerado pela camada oculta pode ser mais um critério para ser levado em consideração, na etapa de aprendizagem.

5 Conclusão e trabalhos futuros

O presente trabalho apresentou um método de inicialização de redes neurais MLP baseado em separações simétricas do espaço de entrada entre os neurônios da camada de saída. O objetivo do método é tornar o espaço gerado pela camada oculta um espaço de tal forma que possua um número de condicionamento baixo.

Para tal, foi escolhido um modelo de rede neural MLP com 2 camadas, no qual a camada oculta possui uma função de ativação do tipo logística e camada de saída linear, conforme equação (1). Tal configuração de rede neural é especialmente interessante por ser simples e atender por completo o

teorema da aproximação universal.

Pela revisão da literatura apresentada é possível observar que o número de condicionamento do espaço gerado pela camada oculta não é uma preocupação para os métodos, e sim critérios como número de épocas para a aprendizagem e número pontos que estejam em regiões saturadas da função de ativação.

Como trabalhos futuros é desejável verificar o comportamento do método para dimensões mais elevadas, comumente utilizadas em problemas de aprendizagem de máquina, e para funções não sintéticas, onde haja ruídos de medição bem como dados faltosos.

Em um futuro trabalho esperamos investigar o ganho em tempo e/ou número de épocas necessárias para treinamento utilizando o método proposto em comparação aos principais métodos da literatura, bem como a aplicação do método proposto para outras funções de ativação.

Referências

Aguirre, L. (2004). *Introdução à Identificação de Sistemas – Técnicas Lineares e Não-Lineares Aplicadas a Sistemas Reais*, Editora UFMG.

Banerjee, A. (1994). Initializing Neural Networks using Decision Trees, *Proceedings of the International Workshop on Computational Learning and Natural Learning Systems*, MIT Press, pp. 3–15.

Bishop, C. M. (1996). *Neural Networks for Pattern Recognition*, 1 edn, Oxford University Press, USA.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*, Cambridge University Press.

Castillo, E., Guijarro-Berdinas, B., Fontenla-Romero, O. and Alonso-Betanzos, A. (2006). A Very Fast Learning Method for Neural Networks Based on Sensitivity Analysis, *Journal of Machine Learning Research*, Vol. 7, pp. 1159–1182.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals, and Systems (MCSS)* **2**(4): 303–314.

de Castro, L. N., Iyoda, E. M., Zuben, F. J. V. and Gudwin, R. (1998). Feedforward Neural Network Initialization: an Evolutionary Approach, *Neural Networks, Brazilian Symposium on* **0**: 43.

de Castro, L. N. and Zuben, F. J. V. (2001). An Immunological Approach to Initialize Feedforward Neural Network Weights, Conf. on Artificial Neural Networks and Genetic Algorithms, pp. 126–129.

Denoeux, T. and Lengellé, R. (1993). Initializing back propagation networks with prototypes., *Neural Networks* **6**(3): 351–363.

Drago, G. P. and Ridella, S. (1992). Statistically controlled activation weight initialization (SCAWI), *Trans. Neur. Netw.* **3**(4): 627–631.

Fahlman, S. E. (1988). An empirical study of learning speed in back-propagation networks, *Technical report*.

Fernández-Redondo, M. and Hernández-Espinosa, C. (2000). A Comparison among Weight Initialization Methods for Multilayer Feedforward Networks, *IJCNN (5)*, pp. 543–548.

Haffner, P., Shikano, K. and Waibel, A. (1988). Fast Back-Propagation Learning Methods for Neural Networks in Speech, Proceedings of the Fall Meeting of the Acoustical Society of Japan, pp. 619–624.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators, *Neural Networks* **2**(5): 359–366.

Ivanova, I. and Kubat, M. (1995). Initialization of Neural Networks by Means of Decision Trees, *Knowledge-Based Systems* **8**: 333–344.

Nguyen, D. and Widrow, B. (1990). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights, *International Joint Conference on Neural Networks, 1990., 1990 IJCNN* , Vol. 3, pp. 21–26.

Thimm, G. and Fiesler, E. (1997). High-order and multilayer perceptron initialization, *Trans. Neur. Netw.* **8**(2): 349–359.

Vapnik, V. N. (1995). *The nature of statistical learning theory*, Springer-Verlag New York, Inc., New York, NY, USA.

Wan, W., Hirasawa, K., Hu, J. and Murata, J. (2001). Relation between weight initialization of neural networks and pruning algorithms: case study on Mackey-Glass time series, *Proceedings. IJCNN '01. International Joint Conference on Neural Networks, 2001.* , Vol. 3, pp. 1750–1755.