

A: XXXIX-0000

TÉCNICAS ESTATÍSTICAS PARA O PRÉ-PROCESSAMENTO DE DADOS EXPERIMENTAIS PARA O USO EM ALGORITMOS INTELIGENTES

STATISTICAL TECHNIQUES FOR EXPERIMENTAL DATA PRE-PROCESSING FOR USE IN INTELLIGENT ALGORITHMS

Vanderci F. Arruda (A)(1); Gray F. Moita (2); Eliene P. Carvalho (2); Priscila F. S. Silva (1)

(1) Eng. Civil, Estudante D.Sc., Centro Federal de Educação Tecnológica, Belo Horizonte, Brasil.

(2) Dr. Prof., Centro Federal de Educação Tecnológica, Belo Horizonte, Brasil.

Endereço para correspondência: vanderci-engcivil@hotmail.com; (A) Apresentador

Área temática: Análise estrutural: métodos computacionais

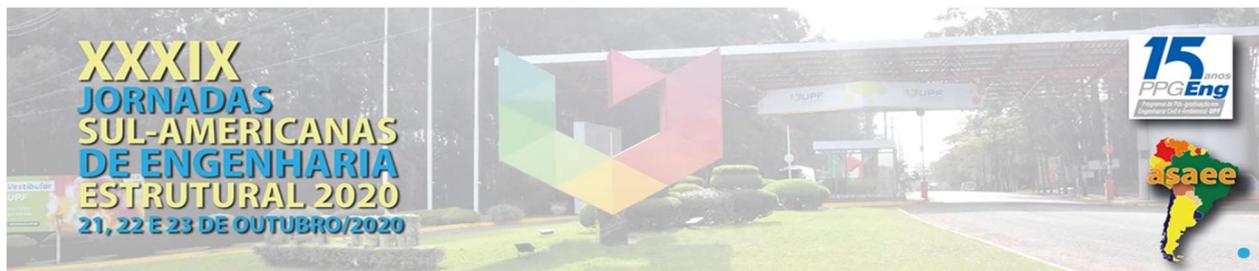
Resumo

A estatística é uma ferramenta essencial na análise quantitativa de dados, com a qual é possível extrair informações importantes. A experimentação traz consigo múltiplas medições da mesma grandeza que estão sujeitas a erros. Neste trabalho, os dados experimentais provêm de ensaios destrutivos da análise de aderência entre barras finas de aço e o concreto convencional. O objetivo deste trabalho é aplicar técnicas estatísticas com a finalidade de preparar os parâmetros de entrada para um sistema inteligente. Os sistemas inteligentes podem ser utilizados como maneira alternativa ao uso de testes destrutivos que demandam tempo e dinheiro. Modelos preditivos que utilizam métricas de desempenho e gradiente assumem que os dados estejam padronizados e, deste modo, técnicas estatísticas e de pré-processamento se mostram eficazes e podem melhorar o desempenho de modelos computacionais. Uma análise preliminar foi realizada visando a retirada de dados espúrios, ou seja, pontos fora da curva ou fora do conjunto de dados, considerando-se uma distribuição normal. Para essa análise, foram utilizados dois métodos: o teste de Grubbs e a avaliação dos valores médios e do desvio padrão de cada amostra. Adotando o teste de Shapiro-Wilk, foi verificado se os dados seguem distribuição normal ou gaussiana. Após serem tratados, os dados foram utilizados como parâmetro de entrada para as Redes Neurais Artificiais e o resultado foi avaliado, comparando-se os valores experimentais com os valores determinados pelo modelo preditivo. Também foram inseridos todos os dados, sem nenhum tratamento prévio, comparando-se os valores experimentais com os valores teóricos calculados pelo programa. Verificou-se que houve melhoria das métricas de desempenho das Redes Neurais Artificiais com a aplicação de uma análise estatística para retirada de valores espúrios.

Palavras-chave: análise estatística, dados experimentais, sistemas inteligentes, Rede Neural Artificial

Abstract

Statistics is an essential tool in quantitative data analysis with which important information can be extracted. Experimentation brings with it multiple measurements of the same magnitude that are subject to error. In this paper, the experimental data come from destructive tests of the bond analysis between thin steel bars and conventional concrete. The goal of this work is to apply statistical techniques in order to prepare the input parameters for an intelligent system. Intelligent systems can be used as an alternative way to the use of time- and money-consuming destructive testing. Predictive models using performance and gradient metrics assume that the data is standardized, thus statistical and preprocessing techniques prove effective and can improve the performance of computational models. A preliminary analysis was performed to remove spurious data, i.e. points outside the curve or outside the data set considering a normal distribution. Two methods were used for this analysis: Grubbs' test and the evaluation of the mean values and standard deviation for each sample. Adopting the Shapiro-Wilk test, it was checked whether the data follow a normal



or Gaussian distribution. After being treated, the data was used as input parameters for the Artificial Neural Networks and the result was evaluated, comparing the experimental values with the values determined by the predictive model. All the data was entered, without any previous treatment, and the experimental values were compared with the theoretical values calculated by the program. The performance metrics of the Artificial Neural Networks were improved by applying statistical analysis to remove spurious values.
Keywords: statistical analysis, experimental data, intelligent systems, Artificial Neural Network

1. INTRODUÇÃO

Com o desenvolvimento da tecnologia, o uso de sistemas inteligentes tem se tornado rotineiro. Desta maneira, sistemas preditivos têm feito uso de algoritmos inteligentes em sua formulação. Modelos preditivos que utilizam métricas de desempenho e gradiente assumem que os dados estejam padronizados, e, deste modo, técnicas estatísticas e de pré-processamento se mostram eficazes e podem melhorar o desempenho de modelos computacionais.

Dentre as ferramentas da estatística está a retirada de dados fora da curva do conjunto de dados. No presente trabalho foram utilizados um tratamento preliminar e o proposto por Grubbs (1950) com tal finalidade.

O conjunto de dados a ser preparado neste trabalho provém de dados experimentais de ensaios destrutivos da análise de aderência entre barras finas de aço e o concreto convencional, o ensaio *pull-out test*. Em suma, o ensaio propõe a extração de uma barra de aço de um bloco de concreto cúbico, sendo medidos a força de arrancamento e o deslizamento relativo entre aço/concreto. O esquema do ensaio é apresentado na Figura 1.

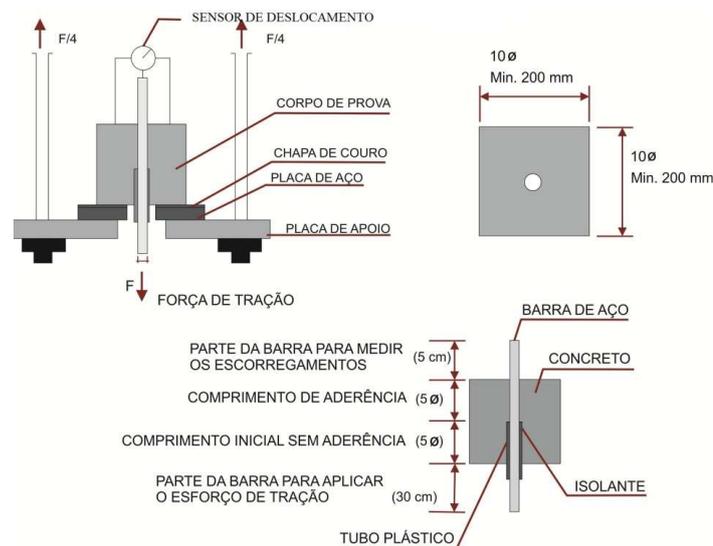
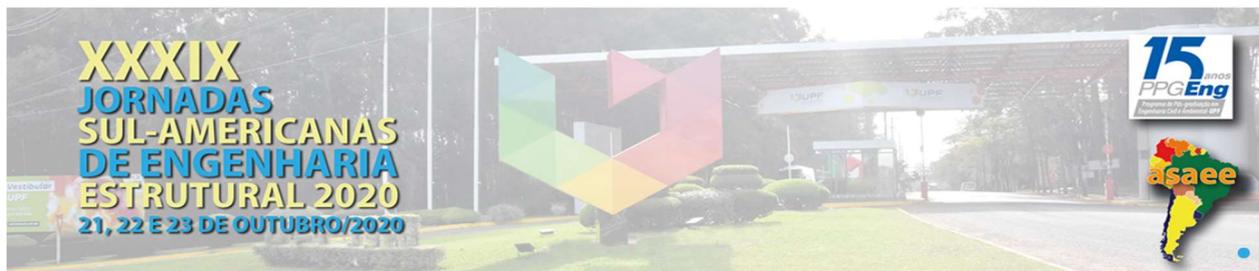


Figura 1 - Ensaio de arrancamento *Pull-Out*. Carvalho et al. (2017)

O objetivo deste trabalho é aplicar técnicas estatísticas com a finalidade de preparar os parâmetros de entrada para um sistema inteligente. Os sistemas inteligentes podem ser utilizados como maneira alternativa ao uso de testes destrutivos que demandam tempo e dinheiro. Em uma análise preliminar, serão retirados os valores fora da curva padrão do conjunto de dados, ou seja, os *outliers*. Depois, os conjuntos de dados antes e após essa análise serão utilizados em uma Rede



Neural Artificial, comprovando que o uso de técnicas estatísticas corrobora para a obtenção de melhores métricas de desempenho para um algoritmo inteligente.

2. METODOLOGIA

Neste item são tratados o conjunto de dados, pré-processamento e as redes neurais artificiais utilizados.

2.1 Banco de dados

Para este estudo, foi proposto o uso do conjunto de dados presente nos estudos de Carvalho et al (2017). O objetivo do estudo foi analisar a performance de barras de aço com diâmetros menores de 10mm nos ensaios *pull-out test*. Neste estudo, a resistência à aderência aço/concreto foi caracterizada pelos seguintes parâmetros de entrada: resistência à compressão do concreto, diâmetro da barra de aço, comprimento de ancoragem, e conformação superficial das barras, e, como parâmetro de saída, os valores de força de arrancamento obtidos no ensaio (em N). Este conjunto de dados contém 14 amostras, sendo estas compostas por barras de aço nervuradas e entalhadas, totalizando 98 elementos.

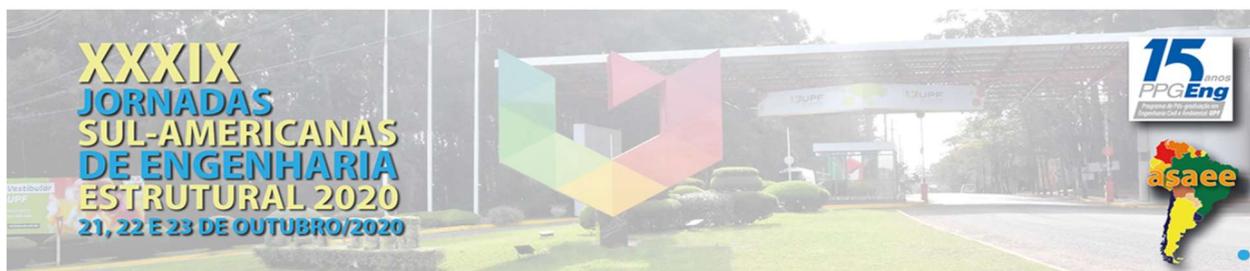
2.2 Pré-processamento de dados

Uma vez definido o conjunto de dados, foram utilizadas técnicas estatísticas com a finalidade de corroborar, no algoritmo inteligente, com a melhoria dos parâmetros de performance. A primeira técnica estatística utilizada teve como objetivo determinar se as amostras seguem distribuição gaussiana por meio do emprego do teste de Shapiro-Wilk com nível de significância de 5%. Em seguida, foram aplicadas técnicas estatísticas com a finalidade de retirar valores espúrios.

O primeiro método de critério de retirada de dados espúrios, chamado de tratamento preliminar, é o que se segue:

- A. os dados numéricos foram considerados suspeitos quando o valor absoluto da variável de resposta, subtraído da média de suas repetições, era maior do que o desvio padrão;
- B. após a identificação dos valores suspeitos, os mesmos foram desconsiderados e foram calculados os novos valores das médias e desvios padrões;
- C. a seguir, foram refeitos os cálculos, retirando a variável de resposta suspeita pela nova média calculada. O valor era considerado espúrio se o resultado, em valor absoluto, ultrapassasse duas vezes o novo desvio padrão.

O segundo método de retirada de *outliers* foi o teste de Grubbs. Ele é também conhecido como teste residual máximo normalizado, sendo usado para detectar *outliers* em um conjunto de dados. O teste assume uma amostra normalmente distribuída, ou seja, em uma primeira análise é necessário verificar se os dados seguem distribuição normal (teste de Shapiro-Wilk). Em seguida, a amostra é ordenada e, após esse procedimento, o teste detecta possíveis *outliers*, um de cada vez. Para isto, são verificados cada um dos seguintes itens: um ponto fora da curva presente nas extremidades, dois pontos em extremidades opostas, e, por fim, dois pontos fora da curva presente na mesma extremidade. São calculados coeficientes, que são comparados com os valores tabelados para uma significância de 5%.



2.3 Rede Neural Artificial

Segundo Silva (1998), o estímulo inicial que conduziu ao desenvolvimento de modelos matemáticos de redes neurais – as RNAs – foi um esforço para entender mais detalhadamente o funcionamento do cérebro humano. Segundo Haykin (1994), uma rede neural em camadas é composta por uma camada de entrada onde os nós fontes se projetam sobre uma camada de saída de neurônios.

Uma vez feita a retirada dos dados espúrios, foi feita a implementação de uma Rede Neural Artificial (RNA) para verificar a melhoria das métricas de performance com a retirada dos dados fora da curva (*outliers*).

As redes neurais artificiais são um método de aprendizado de máquinas bioinspirados, capazes de fazer o processamento de dados a partir de unidades simples. Por propor um modelo simples, elas podem solucionar problemas de diferentes áreas resolvendo problemas de categorização de dados, previsão e na tomada de decisão. As RNAs têm a capacidade reconhecer e extrapolar conhecimento, tendo destaque na modelagem de sistemas complexos. As redes neurais são compostas por neurônios artificiais e sua quantidade depende do problema e ser resolvido. Um neurônio artificial é um modelo que traz consigo as habilidades de um neurônio biológico.

Os parâmetros de performance utilizados nesse trabalho foram a minimização do erro quadrático médio e o coeficiente de determinação. No presente estudo, eles foram usados para verificar se técnicas estatísticas corroboram na melhoria das métricas de desempenho de um algoritmo inteligente.

Conforme descrito por Hagan et al. (1996), existe uma vasta gama de algoritmos de treinamento que podem ser baseados nos métodos de Jacobiano e Gradiente, entre eles GradientDescent, GradientDescent with momentum, Quasi-Newton e Levenberg-Marquardt.

Neste trabalho foi empregada uma RNA com a utilização do aprendizado supervisionado e com a utilização do algoritmo de treinamento proposto por Levenberg-Marquardt. Também foi utilizada uma topologia de uma camada sendo constituída de 10 neurônios artificiais. Como função de ativação, foi utilizada a tangente hiperbólica na camada de entrada e linear na camada de saída.

3 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

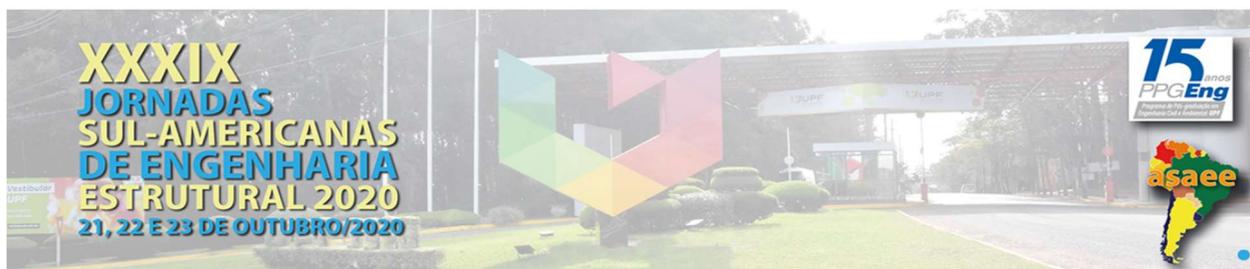
Este item resume os resultados obtidos pelo trabalho.

3.1 Teste Shapiro-Wilk

O teste de Shapiro-Wilk foi utilizado para verificar a normalidade das amostras do banco de dados, conforme Tabela 1, para barras entalhadas.

Tabela 1 – Valor-p Shapiro-Wilk para barras entalhadas

Amostra	valor da estatística	valor-p de Shapiro	Valor-p
E1	0,91	0,26	0,05
E2	0,82	0,02	
E3	0,94	0,59	
E4	0,96	0,86	



E5	0,95	0,76
E6	0,95	0,74
E7	0,68	0,00
E8	0,89	0,40
E2SW	0,85	0,07
E7SW	0,91	0,52

Na Tabela 1 as amostras E1, E3, E4, E5, E6 e E8 possuem valor-p de Shapiro maior que o valor de referência (valor-p), concluindo que as amostras seguem distribuição normal.

Das amostras E2 e E7 foram retirados os valores espúrios via tratamento preliminar, sendo renomeadas de E2SW e E7SW, e obtendo valor-p de Shapiro superior ao valor de referência, concluindo que estas seguem distribuição normal.

O teste de Shapiro-Wilk foi utilizado para verificar a normalidade das amostras do banco de dados, conforme Tabela 2, para barras nervuradas.

Tabela 2 – Valor-p de Shapiro-Wilk para barras nervuradas

Amostra	valor da estatística	Valor-p de Shapiro	p valor
N1	0,91	0,47	0,05
N2	0,79	0,08	
N3	0,84	0,18	
N4	0,98	0,94	
N5	0,90	0,20	
N6	0,83	0,03	
N6SW	0,83	0,05	

Na Tabela 2 as amostras N1, N2, N3, N4 e N5 possuem valor-p de Shapiro maior que o valor de referência (valor-p), concluindo que as amostras seguem distribuição normal.

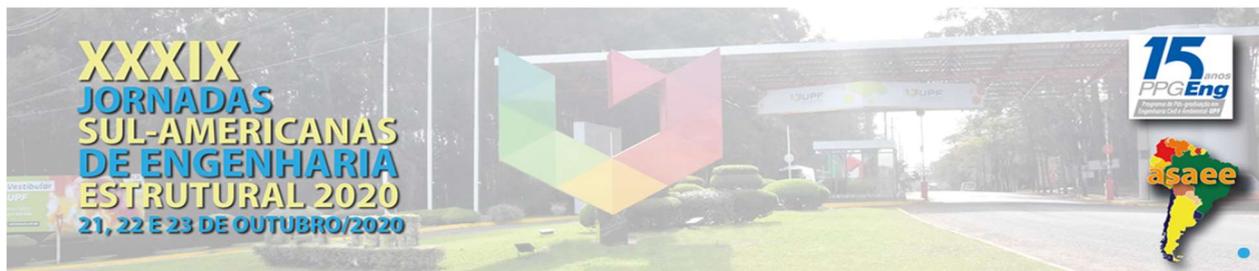
Da amostra N6 foram retirados os valores espúrios via tratamento preliminar, sendo renomeadas de N6SW, e obtendo valor-p de Shapiro igual ao valor de referência concluindo que estas seguem distribuição normal.

3.2 Tratamento Preliminar e Teste de Grubbs

O tratamento preliminar, que utiliza a média e o desvio padrão da amostra, e o teste de Grubbs, que faz a comparação entre coeficientes calculados e valores tabelados, propuseram a retirada dos *outliers*. As amostras apresentadas na Tabela 3 e na Tabela 4 mostram o número de elementos da amostra com e sem a retirada dos valores espúrios para cada tratamento.

Tabela 3 - Número de elementos das amostras com a retirada de *outliers* via tratamento preliminar

TRATAMENTO PRELIMINAR

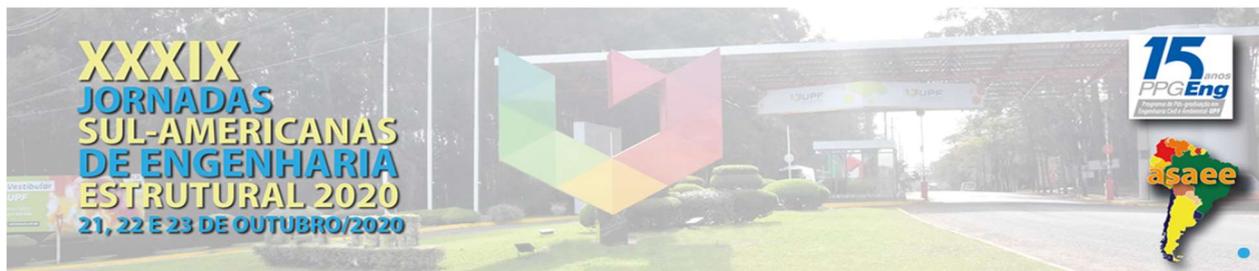


TIPO	AMOSTRA	NÚMERO DE ELEMENTOS DA AMOSTRA	NÚMERO DE ELEMENTOS SEM OUTLIERS
NERVURADAS	N1	5	5
	N2	5	5
	N3	5	5
	N4	4	3
	N5	11	9
	N6	10	9
ENTALHADAS	E1	12	11
	E2	11	9
	E3	10	9
	E4	5	4
	E5	4	2
	E6	6	4
	E7	5	4
	E8	5	4
	TOTAL	98	83

Do número total de elementos da amostra bruta de 98 elementos com o tratamento preliminar foram retirados 15 elementos de amostras, restando somente 83 elementos.

Tabela 4 - Número de elementos das amostras com a retirada de outliers via teste de Grubbs

TESTE DE GRUBBS			
TIPO	AMOSTRA	NÚMERO DE ELEMENTOS DA AMOSTRA	NÚMERO DE ELEMENTOS SEM OUTLIERS (GRUBBS)
NERVURADAS	N1	5	5
	N2	5	5
	N3	5	5
	N4	4	4
	N5	11	11
	N6	10	9
ENTALHADAS	E1	12	12
	E2	11	11
	E3	10	10
	E4	5	5
	E5	4	4
	E6	6	4
	E7	5	4



	E8	5	5
	TOTAL	98	94

Do número total de elementos da amostra bruta de 98 elementos com o tratamento preliminar foram retirados 4 elementos de amostras, restando somente 94 elementos.

3.4 Rede Neural Artificial

A rede neural proposta inclui as barras nervuradas e entalhadas conjuntamente. Na Tabela 5 são apresentados os resultados referentes a implementação das redes neurais artificiais desenvolvidas nesta pesquisa.

Tabela 5 - resultados da RNA

	DADOS BRUTOS	GRUBBS	TRATAMENTO PRELIMINAR
R^2	0,899	0,908	0,948
RMSE (N)	1545,604	1457,680	1358,704

Por meio dos resultados mostrados na Tabela 5, verificou-se que a retirada dos *outliers* melhorou o desempenho das métricas de performance. O coeficiente de determinação obteve um resultado de 0,899 para os dados brutos, seguido do valor de 0,908 para retirada de outliers usando o teste de Grubbs e, por fim, o valor de 0,948 por meio do tratamento preliminar (neste tratamento foi retirado o maior número total de elementos amostras). Com relação ao erro médio quadrático, pode-se notar que os resultados obtiveram o mesmo padrão e a seguinte ordem foi respeitada: o valor de menor performance foi com os dados brutos, seguido pelo teste de Grubbs e, por fim, o tratamento preliminar.

4 CONCLUSÕES

O presente trabalho teve como objetivo geral testar a influência da retirada de *outliers*, via tratamento preliminar e pelo teste de Grubbs, e seu impacto nas métricas de performance em uma Rede Neural Artificial.

Por meio da realização deste trabalho, foi possível concluir que as métricas de desempenho utilizadas na Rede Neural Artificial obtiveram melhora com o uso de técnicas estatísticas aplicadas para a base de dados de Carvalho et al. (2017). A RNA que utilizou o método estatístico denominado tratamento preliminar foi a que obteve melhor resultado apresentado nas métricas, seguido pelo teste de Grubbs e, por fim, a base de dados brutos. Desta forma, o uso de técnicas estatísticas para “limpeza” de dados deve ser encorajadas, uma vez que indicam melhoria no desempenho de algoritmos inteligentes.

REFERÊNCIAS

Carvalho EP, Ferreira EG, da Cunha JC, Rodrigues C de S, Maia N da S. Experimental investigation of steel-concrete bond for thin reinforcing bars. Lat Am J Solids Struct. 2017;14(11):1932–51.



Grubbs, Frank E. Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, p. 27-58, 1950.

Hagan, Martin T.; DEMUTH, Howard B.; BEALE, Mark. *Neural network design*. PWS Publishing Co., 1997.

Haykin S. "*Neural Networks – A Comprehensive Foundation*", Macmillan College Publishing Inc., 1994.

Silva, Leandro Nunes de Castro et al. *Análise e síntese de estratégias de aprendizado para redes neurais artificiais*. Universidade Estadual de Campinas. Campinas, p. 250, 1998.