ISNN: 2358-6052

REGULARIZAÇÃO DE REDES NEURAIS ARTIFICIAIS DO TIPO MLP COM ANÁLISE DE ELEMENTOS DE BORDA DE SEPARAÇÃO E REAMOSTRAGEM

Brayan Rawlisom Castoril [1]; Servilio Sousa de Assis [2]; Bruno Alberto Soares Oliveira [3];

1 Engenheiro de Computação, IFMG Campus Bambuí, Bambuí - MG; brayanbrc@gmail.com
2 Programa de Pós-Graduação em Engenharia Elétrica, UFMG, Belo Horizonte – MG
3 Programa de Pós-Graduação em Engenharia Elétrica, UFMG, Belo Horizonte – MG; brunoalbertobambui@ufmg.br

RESUMO

Redes Neurais Artificiais podem ser consideradas como uma classe de algoritmos, que são modelados de forma que se comportam assim como o cérebro humano, sendo eles projetados para reconhecer padrões. Eles interpretam dados sensoriais por meio de um tipo de percepção de máquina, rotulagem ou agrupamento dado certas entrada de dados. Os padrões que essa família de algoritmos reconhece são numéricos, contidos em vetores, nos quais todos os dados do mundo real sejam imagens, sons, textos ou séries temporais, devem ser traduzidos para uma codificação própria. As Redes Neurais Artificiais podem ser aplicadas em problemas de regressão e classificação. É possível pensar nelas como uma camada de cluster e classificação sobre os dados armazenados e gerenciados. Elas ajudam a agrupar dados não rotulados de acordo com semelhanças entre as entradas de exemplo e classificam os dados quando eles têm um conjunto de dados rotulado para treinamento. As Redes Neurais Artificiais também podem extrair recursos que são alimentados a outros algoritmos para clustering e classificação, assim, é possível pensar em redes neurais profundas como componentes de aplicativos de aprendizado de máquina mais robustos, envolvendo algoritmos para aprendizado de reforço, classificação e regressão, por exemplo. Problemas de generalização de modelos de Redes Neurais Artificiais são frequentes. Diferentes técnicas são apresentadas pela literatura para contornar essa natureza de problemas. O presente trabalho propõe uma técnica de reamostragem para melhorar a generalização de um modelo de rede neural. A técnica consiste em analisar os elementos de uma dada amostra que estão em uma região de separação de classes e então adicionar ruídos pertencentes a cada classe presente no problema. Foram realizados experimentos com utilização da técnica proposta em derivação de modelos descritivos para bases de dados reais e sintéticas. Através dos experimentos realizados, foi observado que a técnica proposta ocasionou um ganho em acurácia em grande parte dos modelos treinados, além de um menor erro médio quadrático e maior área sob a curva ROC. Novos estudos são sugeridos para abordar um maior número de variáveis e parâmetros na técnica proposta visando aprimorar os resultados.

INTRODUÇÃO:

Redes Neurais Artificiais são compostas por unidades de processamento que armazenam conhecimento experimental e os tornam disponíveis para uso. Seu funcionamento se assemelha ao cérebro em relação ao conhecimento adquirido, que é baseado em um processo de aprendizagem e na existência dos pesos sinápticos, os quais são responsáveis por armazenar os conhecimentos adquiridos (HAYKIN, 2001).

A RNA é uma técnica computacional, inspirada no funcionamento de neurônios biológicos, que é utilizada para resolução de problemas computacionais de várias naturezas, como classificação, regressão e previsão.

No contexto de problemas de classificação, o aprendizado supervisionado é caracterizado pela utilização de um conjunto composto por pares de indivíduos amostrados de uma população e suas respectivas classificações. A capacidade efetiva de um modelo de rede neural depende das restrições impostas ao seu espaço de soluções. Estas restrições podem ser determinadas, por exemplo, pelo número de parâmetros do modelo.

A capacidade de generalização de uma rede MLP (*Multilayer Perceptron*) é um grande problema quando se utiliza o algoritmo de treinamento *backpropagation*, devido à sua grande capacidade de se ajustar aos dados utilizados no treinamento (CARUANA, LAWRENCE e GILES, 2001).

Visando contornar o problema da generalização, técnicas de regularização, introduzidas por Tikhonov (1963), têm sido utilizadas em Redes Neurais Artificiais (GIROSI, JONES E POGGO, 1995). As técnicas de regularização utilizam, geralmente, informações obtidas a priori sobre o problema. Dado esse contexto, o objetivo do presente trabalho é propor uma técnica de regularização para problemas de classificação, utilizando adição de ruídos nas bases de treinamento dos modelos.

ISNN: 2358-6052

METODOLOGIA:

Para o controle da capacidade efetiva das redes neurais artificiais aplicadas em problemas de classificação, quando treinadas com o algoritmo *backpropagation*, a técnica proposta consiste na reamostragem de dados em regiões de separação de classes. Tal técnica se divide em duas etapas, sendo elas a identificação de possíveis elementos da amostra presentes em uma região de separação de classes e posteriormente adição de ruídos em torno desses elementos para treinamento o modelo.

Na condução dos experimentos, foi utilizada uma base de dados sintética e sete bases de dados reais, onde suas características são apresentadas a seguir.

A. Base de Dados Sintética

Para o experimento com base de dados sintética, foram utilizadas duas amostras de dados de duas dimensões geradas a partir de distribuições gaussianas multivariadas. A Tabela 1 apresenta as médias e as matrizes de covariância de ambas as distribuições.

Distribuição 1	Distribuição 2
$\mu_1 = (1,1)$	$\mu_2 = (3,3)$
$\sum = \begin{matrix} 1 & 0 \\ 0 & 1/2 \end{matrix}$	$\sum = \begin{matrix} 1 & 0 \\ 0 & 1/2 \end{matrix}$

Tabela 1: Parâmetros de Distribuições de Gaussianas

Para cada distribuição, foi atribuída uma classe binária distinta. Assim, é dado um problema sintético de classificação. Dadas as distribuições descritas por densidades Gaussianas Multivariadas, com o conhecimento prévio acerca das funções geradoras das classes, é possível calcular a superfície de separação ideal (DUDA at. al, 2001).

B. Bases de Dados Reais

As sete bases de dados reais utilizadas na etapa experimental do trabalho são descritas nos tópicos seguintes.

YEAST: A base de dados contém informações sobre sítios de localização de proteínas em uma célula. Para cada registro, há observações de 9 variáveis e uma classe correspondente. A base possui, ao todo, 1484 registros (HORTON e NAKAI, 1996).

Mammographic Masses: A base de dados contêm informações sobre atributos de classificação BI-RADS e idade de pacientes relacionados com a severidade de um tumor na mama, podendo ser benigno ou maligno. A base apresenta 961 registros e apresenta 5 variáveis que podem ser utilizadas para predizer a severidade do tumor (ELTER, SCHULZ-WENDTLAND e WITTENBERG, 2007).

ILPD: Este conjunto de dados contém 416 registros de pacientes hepáticos e 167 registros de pacientes não hepáticos.

O conjunto de dados foi coletado do nordeste de Andhra Pradesh, na Índia. Na base, para cada registro, há valores de 10 variáveis independentes e uma respectiva classificação (DUA e TANISKIDOU, 2017).

PimaIndiansDiabetes: A base consiste em 8 variáveis preditoras médicas (independentes) e uma variável alvo (dependente), que define o diagnóstico da diabete como positivo ou negativo. Variáveis independentes incluem o número de gestações que cada paciente teve, seu IMC, nível de insulina, idade e assim por diante.

Breast Cancer Wisconsin: A base de dados possui nove variáveis contendo informações sobre tumores de pacientes com câncer. Para cada registro, o tumor é classificado como benigno ou maligno.

HouseVotes84: Este conjunto de dados inclui votos para cada um dos congressistas da câmara dos deputados dos EUA sobre as votações em 1984. A base é composta por um quadro de dados com 435 observações em 16 variáveis preditoras e uma de classificação.

Ionosphere: A base consiste em dados de transmissão de antenas de alta frequência. As antenas apresentavam elétrons livres na ionosfera como alvos. A classificação define, com base em dados coletados de um radar, se para cada configuração, há evidências de que os sinais alcançaram os elétrons ou passaram por eles. A base é composta por um quadro de dados com 351 observações em 34 variáveis independentes e uma última definindo a classe.

As bases de dados *YEAST*, *Mammographic Masses* e *ILPD* estão disponíveis no repositório UCI (NEWMAN et. al. 1998). As bases de dados *PimaIndiansDiabetes*, *BreastCancer*, *HouseVotes84* e *Ionosphere* estão disponíveis no pacote *Mlbench* do R (LEISCH e DIMITRIADOU, 2010).

C. Identificação de Possíveis Elementos em Regiões de Separação de Classes

Para a identificação de elementos pertinentes à regiões de separação de classes, foi observada a homogeneidade na vizinhança em torno de cada elemento. Dado um indivíduo P de classe C_P, com D dimensões, e valores $P_1, \ldots, P_D$ um ponto P' está em torno de P à uma distância $d(P, P')$, tal que:

$$d(P, P') = \sqrt{\sum_{i=1}^{D} (P_i - P_i')^2} \quad (1)$$

Através da Equação (1), é possível encontrar os elementos vizinhos de um dado elemento, respeitando-se um limite de distância R, onde P' está em torno de P se a restrição $d(P, P') \leq R$ é satisfeita. A distância R é um parâmetro a ser definido.

Uma vez dispondo dos pontos vizinhos do ponto analisado, é então observada a homogeneidade desses vizinhos, de acordo com as premissas a seguir:

- Se um elemento P de classe A possui muitos vizinhos de classe A e poucos vizinhos de classe diferente de A, então P é um improvável elemento próximo à uma região de separação de classes.
- Se um elemento P de classe A possui muitos vizinhos de classe diferente de A, então P é um improvável elemento próximo à uma região de separação de classes. Isso também pode ser um indício de que P seja um *outlier*.
- Se um elemento P de classe A apresenta vizinhos igualmente distribuídos, onde há elementos de classe A e elementos de classe diferente de A, então P é um provável elemento próximo à uma região de separação de classes.

Para quantificar os termos “muitos vizinhos” e “poucos vizinhos” de mesma classe de um elemento analisado, foram utilizados limites de percentuais de proporção. Dado um elemento P, a proporção α corresponde à quantidade de elementos vizinhos de P de mesma classe C_P em relação ao total de vizinhos de P.

Assim, definindo L_{min} como um limite inferior de proporção e L_{max} como um limite superior de proporção, é possível determinar se um elemento P é um provável elemento pertencente à uma região de separação de classes se a restrição $L_{min} \leq \alpha \leq L_{max}$ é satisfeita.

Os melhores valores para os parâmetros L_{min}, L_{max} e R podem variar com a natureza da base de dados trabalhada. A técnica pode apresentar bom potencial para detecção dos pontos de borda.

Na Figura 1, há duas classes de exemplo, representadas visualmente nos gráficos pelas cores vermelho e verde. Cada indivíduo apresenta valores em duas dimensões, o que torna possível a visualização dos indivíduos em um gráfico plano.

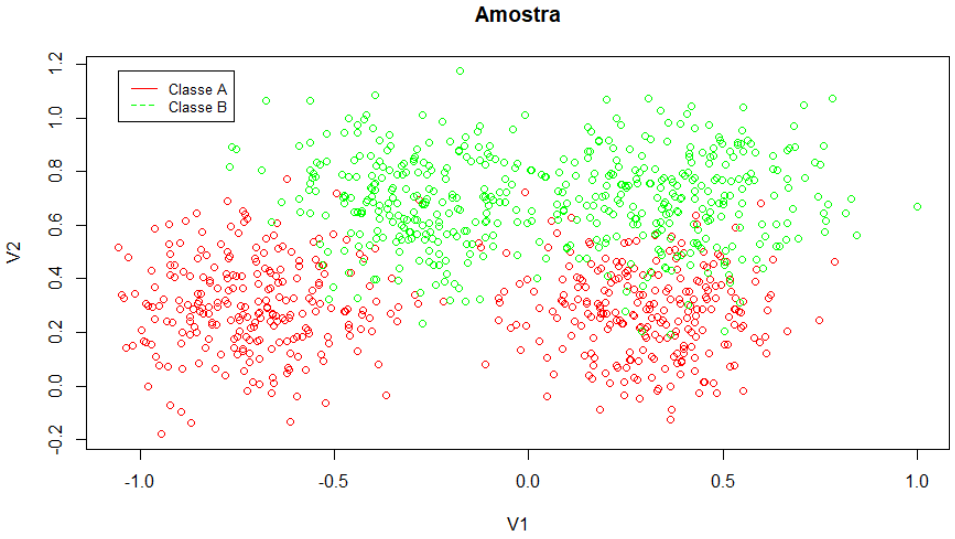

Figura 1: Representação gráfica de uma base de dados sintética

ISNN: 2358-6052

Aplicando a técnica de identificação de possíveis elementos pertencentes à regiões de separação de classes, é possível também visualizar graficamente os pontos encontrados como possíveis pontos de borda. A Figura 2 apresenta um gráfico contendo os prováveis pontos da base de dados que estão presentes na borda após a aplicação do método descrito.

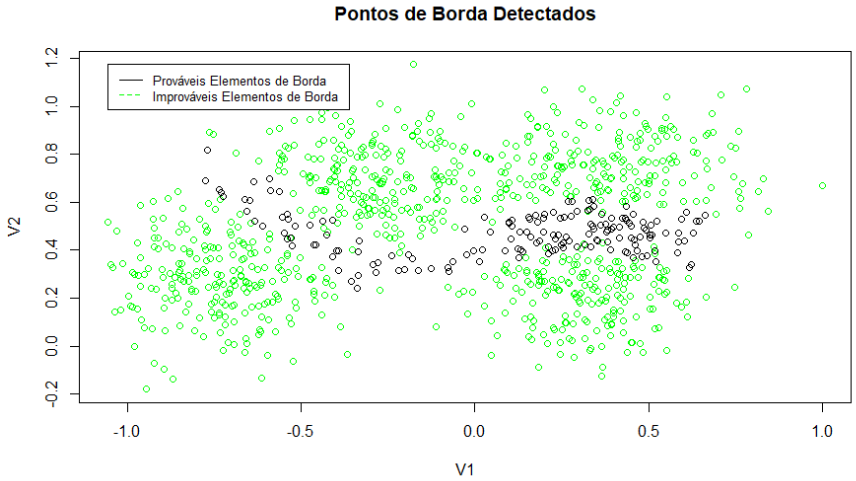

Figura 2: Pontos identificados na região de separação entre classes

D. Reamostragem

A reamostragem nos conjuntos de dados para o treinamento dos modelos foi feita da seguinte forma: uma vez com os elementos de borda detectados, foram adicionados quatro elementos de ruído, sendo dois para cada classe existente no problema. Cada um desses elementos foi posicionado aleatoriamente em torno do elemento de borda em questão.

Todos os experimentos foram feitos utilizando linguagem de programação R. A biblioteca RSNNS foi utilizada para criação dos modelos de redes neurais.

E. Condução de Experimentos

O trabalho se limitou a utilização de problemas com duas classes. Algumas bases de dados, como a YEAST, apresentam mais de duas classes. Assim, foram selecionadas duas classes mais predominantes para prosseguimento com os testes.

As bases de dados reais apresentam atributos dentro de intervalos distintos. Com isso, os dados foram mapeados para um intervalo entre 0 e 1.

Para a derivação dos modelos preditores, as bases foram divididas aleatoriamente em dois conjuntos, sendo um contendo 70% dos dados para treinamento e outro contendo 30% para teste.

Na etapa de identificação de elementos de borda, no conjunto de treinamento, os parâmetros de distância analisada e limites de proporção foram definidos baseados em testes.

Assim, com a aplicação da técnica, foram adicionados pontos de ruído dentro da base de treinamento para derivação dos modelos.

Para cada base de dados real, foi derivado um modelo preditor a partir dos dados utilizando uma Máquina de Vetores de Suporte (SVM) com kernel Gaussiano com os parâmetros C (regularização) e Γ (largura do kernel) otimizados via busca em *grid* com validação cruzada com 10 grupos. Para a base de dados sintética, foi gerada a superfície de separação ideal, com base nas equações geradoras das distribuições, que são conhecidas. Essas superfícies de separação foram utilizadas para comparações posteriores.

Dois modelos de RNA foram derivados para cada base, sendo um sem a utilização da técnica proposta e outro com a utilização da técnica. A arquitetura da RNA utilizada foi a MLP com uma camada escondida e função de aprendizado *backpropagation*. A Tabela 2 apresenta os parâmetros quantitativos utilizados nos modelos.

Número máximo de épocas	1000
ETA	0.05
Erro de critério de parada	0
Pesos Iniciais Aleatórios entre -0.3 e 0.3	

Tabela 2: Parâmetros da MLP

ISNN: 2358-6052

Assim, foi feito a análise das métricas de erro médio quadrático (MSE), área sob a curva ROC (AUC) e acurácia dos modelos. Através das métricas utilizadas, foram observados os ganhos desempenho em relação ao modelo convencional e à SVM.

RESULTADOS E DISCUSSÕES:

A análise de homogeneidade da vizinhança dos elementos para detecção de pontos próximos à regiões de separação entre as classes se mostrou eficiente, como pôde ser observado visualmente na base de dados de duas dimensões exemplificada na Figura 2. Diminuir o limite inferior de proporção e aumentar o limite superior de proporção implica em aumentar a quantidade de elementos de borda detectados. No mesmo sentido, aumentar o valor do raio de análise também faz encontrar mais elementos que possivelmente estão próximos à margem de separação. Os valores ótimos dependem das características dos dados trabalhados e podem ser alcançados com testes, alterando-os e verificando os resultados. Para cada base de dados, esses parâmetros foram definidos empiricamente.

No experimento com a base de dados sintética, foram geradas 50 amostras para cada classe. Assim, foi gerada a curva de separação e comparada com a curva obtida pela MLP sem reamostragem.

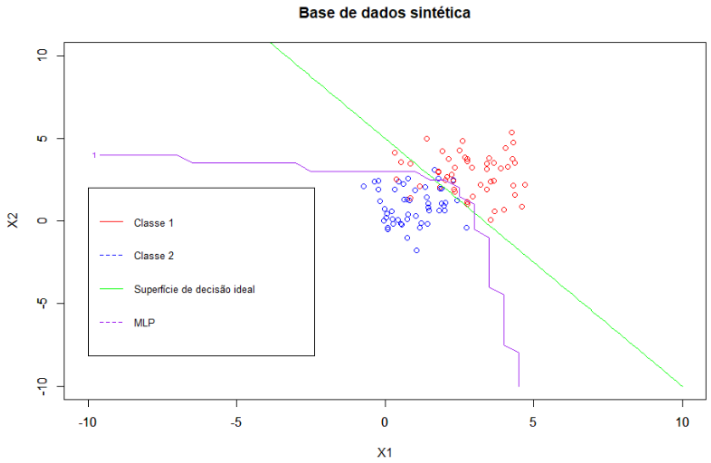

Figura 3: Comparação da curva de separação com MPL sem aplicação de ruídos

Como pode ser apreciada na Figura 3, a curva de separação obtida pelo modelo apresentou um deslocamento considerável em relação à curva de separação ideal. Apesar disso, a acurácia apresentou valor de 93.33%.

Em seguida, o mesmo procedimento foi feito para a MLP com a adição de ruídos na base de treinamento. A Figura 4 apresenta o gráfico de comparação entre as curvas de separação.

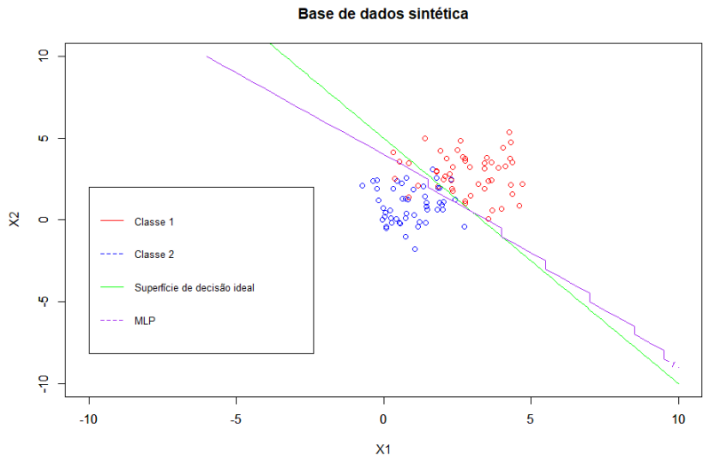

Figura 4: Comparação da curva de separação ótima com MLP com aplicação de ruídos

ISNN: 2358-6052

Com a adição de ruídos, a curva de separação obtida pelo modelo ficou mais próxima da curva de separação ideal. Além disso, a acurácia do modelo aumentou para 96.67%. Após as análises para a base sintética, foram feitas análises para os resultados dos experimentos com utilização de bases reais. Para cada base real, foi criado um modelo baseado em SVM e dois modelos baseados em MLP, sendo um com adição de ruídos no conjunto de treinamento e outro sem. A Tabela 3 apresenta as acurácias obtidas pelos modelos treinados para cada uma das bases reais trabalhadas.

	SVM	MLP	MLP*
YEAST	0.6528	0.6189	0.634
BreastCancer	0.9756	0.9707	0.9805
HouseVotes84	0.9565	0.942	0.9565
Ionosphere	0.9238	0.9048	0.9238
Mammographic Masses	0.8313	0.8554	0.8233
ILPD	0.7126	0.7011	0.7069
PimaIndiansDiabetes	0.7391	0.7435	0.7565
* Com utilização da técnica			

Tabela 3: Acurácia

Dos modelos descritivos derivados para as 7 bases reais trabalhadas, 6 apresentaram ganhos após a aplicação da técnica de regularização proposta. Apesar disso, os valores ficaram aproximados. A Tabela 4 apresenta os erros obtidos para a MLP com e sem aplicação da técnica.

	SVM	MLP	MLP*
YEAST	0.6526	0.6209	0.6328
BreastCancer	0.9694	0.9632	0.9759
HouseVotes84	0.9558	0.9448	0.9558
Ionosphere	0.9466	0.9350	0.9364
Mammographic Masses	0.8313	0.8580	0.8242
ILPD	0.6088	0.5597	0.6195
PimaIndiansDiabetes	0.7120	0.7170	0.7320
* Com utilização da técnica			

Tabela 4: Erro quadrático médio

A Tabela 5 apresenta a área sob a curva ROC apresentada para os modelos de MLP e SVM derivados para as bases de dados.

	MLP	MLP*
YEAST	0.3811	0.3660
BreastCancer	0.0292	0.0195
HouseVotes84	0.0579	0.0434
Ionosphere	0.0952	0.0761
Mammographic Masses	0.1445	0.1767
ILPD	0.2988	0.2931
PimaIndiansDiabetes	0.2565	0.2434
* Com utilização da técnica		

Tabela 5: Ária sobre curva ROC (AUC)

CONCLUSÕES:

Dada uma amostra representativa de uma população em um problema de classificação, elementos pertinentes à uma região de separação entre classes, são mais propensos a gerar erros em um modelo de generalização de classes. Devido ao seu alto poder de adaptação, o algoritmo *backpropagation* tende a super especializar nos dados de exemplo, o que pode ser um ponto negativo para o modelo, em termos de generalização.

ISNN: 2358-6052

Dentre as inúmeras abordagens existentes para contornar esse problema está a técnica proposta nesse trabalho.

Identificar elementos pertinentes à uma região de separação e adicionar ruídos em torno desses elementos é uma técnica plausível, pois tende a suavizar a curva de separação do modelo, garantindo melhor generalização. Através dos experimentos, foi possível observar uma melhor generalização nos modelos derivados com a adição de ruídos nas bases de treinamento.

Como trabalhos futuros, a técnica pode continuar sendo estudada, como exploração de maior número de parâmetros, como por exemplo heurísticas para definir o número de elementos de ruído adicionados e suas posições no espaço amostral, visando aprimorar os resultados da técnica.

REFERÊNCIAS BIBLIOGRÁFICAS:

BERGMEIR, Christoph Norbert et al. Neural networks in R using the Stuttgart neural network simulator: RSNNS.

BLAKE, Catherine. UCI repository of machine learning databases. **http://www. ics. uci. edu/~ mlearn/MLRepository. html**, 1998.

CARUANA, Rich; LAWRENCE, Steve; GILES, C. Lee. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In: **Advances in neural information processing systems**. 2001. p. 402-408.

DUDA, O. Richard et al, Pattern Classification. 2001.

ELTER, M.; SCHULZ-WENDTLAND, R.; WITTENBERG, T. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. **Medical physics**, v. 34, n. 11, p. 4164-4172, 2007.

FRANK, Andrew. UCI machine learning repository. **http://archive. ics. uci. edu/ml**, 2010.

GEMAN, Stuart; BIENENSTOCK, Elie; DOURSAT, René. Neural networks and the bias/variance dilemma. **Neural computation**, v. 4, n. 1, p. 1-58, 1992.

GIROSI, Federico; JONES, Michael; POGGIO, Tomaso. Regularization theory and neural networks architectures. **Neural computation**, v. 7, n. 2, p. 219-269, 1995.

HAYKIN, Simon. Redes neurais: Princípios e Prática. 2001. 2001.

HORTON, Paul; NAKAI, Kenta. A probabilistic classification system for predicting the cellular localization sites of proteins. In: **Ismb**. 1996. p. 109-115.

LEISCH, Friedrich; DIMITRIADOU, Evgenia. Machine Learning Benchmark Problems. 2010.

TIKHONOV, Andrei N. Solution of incorrectly formaulated problems and the regularization method. In: **Dokl. Akad. Nauk.** 1963. p. 1035-1038.