Análise e expansão de uma arquitetura neural capaz de calcular sua própria confiabilidade

Abner C. Rodrigues Neto[1], Mauro Roisenberg[1] Guenther Schwedersky Neto[2]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catatina (UFSC) – Florianópolis, SC – Brasil

[2]Petróleo Brasileiro S.A. - Petrobras CENPES/PDEP/TR
Rio de Janeiro, RJ – Brasil

{abnern, mauro}@inf.ufsc.br, guenther@petrobras.com.br

Abstract. *There are several ways to calculate a measure of confidence to the output of neural networks, but in general these approaches require some restrictions that are not always observed in real problems or even not provide a measure of performance that guarantees the desired level of confidence or which do not reflect the distribution of training data. This paper analyzes and extends a model of neural network that calculates the confidence of its outputs, the Validity Index Network, we remove it restrictions in the calculation of the density and improve the probability coverage of the prediction levels when the training data have variable density.*

Resumo. *Existem várias maneiras de calcular uma medida de confiança para a saída de redes neurais, mas em geral essas abordagens necessitam de algumas restrições que nem sempre são observadas em problemas reais ou mesmo não apresentam uma medida de desempenho que garanta o nível de confiança desejada ou ainda que não refletem a distribuição dos dados de treinamento. Este trabalho analisa e estende um modelo de rede neural capaz de calcular um intervalo de predição na saída, a Rede Índice de Validade, removendo-se restrições no cálculo da densidade por essa rede e melhorando a probabilidade de cobertura do intervalo de predição quando o dados de treinamento possuem densidade variável.*

1. Introdução

Devido a sua grande capacidade representacional, as redes neurais têm sido largamente utilizadas como aproximadores universais de funções na construção de modelos preditivos não-lineares a partir de dados extraídos do sistema real que se está desejando modelar. Porém, devido à sua natureza empírica, é difícil perceber quando a rede neural está extrapolando ou calculando a saída para uma região cujos dados de treinamento eram insuficientes para realizar uma boa aproximação. Medidas de desempenho global normalmente utilizadas para avaliar o desempenho da rede neural, tal como o Erro Médio Quadrática (EMQ), não são capazes de reconhecer regiões onde a resposta da rede possa estar contaminada por incertezas devido a fatores como, os erros do modelo devido ao ruído dos dados, ou a baixa densidade de dados de treinamentos nestas regiões. Para resolver esse problema, a solução proposta é calcular alguma forma de medida de confiabilidade do modelo, como por exemplo, os Intervalos de Predição (IP).

Em geral, a metodologia para cálculo do IP é específica da arquitetura de rede neural utilizada. Pode-se dizer que existem basicamente duas grandes abordagens: a abordagem baseada em aprendizagem local e estimação do IP através de técnicas de regressão linear e; a abordagem baseada em aprendizagem global que utiliza modelos de regressão não-linear para cálculo do IP.

Abordagens de aprendizagem local, como a modelo de Função de Base Radial ("*Radial Basis Function*" - RBF), possuem como principal característica o conceito de vizinhança. Um ponto, seja de treinamento ou teste, é considerado local a um ponto de teste, quando está espacialmente localizado em uma região limitada e bem definida em torno deste ponto. A rede Índice de Validade ("*Validity Index network*" - VInet), é uma extensão das redes RBF proposta por Leonard et al. [Leonard et al. 1992] que calcula o IP para sua saída, além de outras medidas de confiança como a função densidade de probabilidade dos dados e o *flag* de extrapolação [Chinman and Ding 1998].

Por outro lado, redes com aprendizagem global, como Perceptrons de Múltiplas Camadas ("*Multilayer Perceptrons*" - MLP), não possuem este conceito de vizinhança local e portanto não podem ser facilmente estendidas para incorporar o cálculo do IP. Algumas soluções para calcular o IP em redes MLP foram propostas na literatura [Chryssolouris et al. 1996], [Shao et al. 1997], [Veaux et al. 1998], [Hwang and Ding 1997]. Porém, esses algoritmos assumem fortes restrições que devem ser satisfeitas, tais como: o número de pontos de treinamento devem tender para o infinito [Hwang and Ding 1997]; os resíduos devem ser independentes e distribuídos de acordo com uma gaussiana com média zero [Chryssolouris et al. 1996] e; rede deve ser treinada até a convergência [Veaux et al. 1998].

Essas condições nem sempre são verdadeiras em problemas reais e não precisam ser satisfeitas na VInet, que é bem mais simples. Por outro lado, em estudos comparativos foi observado que o tamanho do intervalo de predição calculado na VInet nem sempre corresponde a distribuição dos dados de treinamento ou não atingem uma probabilidade de cobertura desejada [Shao et al. 1997], [Yang et al. 1991]. Além disso, a maneira como a densidade é calculada na VInet pode levar a resultados insatisfatórios, principalmente quando se utiliza dados multi-dimensionais [Wedding and Cios 1997].

O objetivo deste trabalho é estudar o modelo de rede VInet a fim de propor, desenvolver e testar técnicas capazes de transpor as limitações apresentadas. Deste modo, modifica-se a maneira de como a densidade dos dados é calculada pela VInet utilizando redes auto-organizáveis, uma abordagem que não depende da dimensionalidade dos dados. Em seguida incorpora-se essa densidade no valor final do intervalo de predição, com o objetivo de corrigir as deficiências da VInet em aproximar o IP à distribuição dos dados e obter um IP compatível com a probabilidade de cobertura esperada.

Este trabalho está organizado da seguinte forma: a próxima seção aborda as características da rede VInet. O método de cálculo da densidade utilizando redes auto-organizáveis é descrito na seção 3. Os testes efetuados e os resultados obtidos estão relatados na seção 4. Finalmente, a seção 5 apresenta as conclusões.

2. Rede Índice de Validade

Uma rede RBF é uma rede neural que adota o comportamento de certos neurônios biológicos chamado de resposta localmente sintonizada ("*locally tunned response*")

[Hassoun 1995]. Esse comportamento faz com que tais neurônios respondam somente a um intervalo finito do espaço de sinais de entrada. Essas redes são constituídas de 3 camadas com funções distintas. A primeira camada tem a função de conectar a rede ao ambiente e é formada por neurônios chamados de unidades sensoriais. Já a segunda camada (única camada oculta) constitui-se de neurônios com funções de ativação de base radial e não-linear (normalmente de base gaussiana), sendo que seu papel na arquitetura da rede é o de aplicar uma transformação não-linear do espaço de entrada para o espaço oculto que normalmente é de alta dimensionalidade. Estas funções de base radial representam alguma métrica de distância para o espaço de entrada ao determinar a ativação dos neurônios. A resposta dessas funções aumentam (ou diminuem) monotonicamente em relação a distância de um ponto central. Por fim a camada de saída é linear, fornecendo a resposta da rede para o estímulo recebido, como pode ser visto na Figura1.

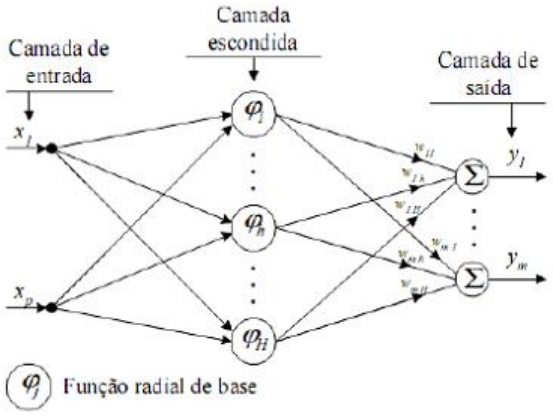

Figura 1. Diagrama da arquitetura das redes RBF.

A função típica escolhida para a RBF é a normal:

$$v_j(x) = exp\left(-\frac{(x-k_j)^2}{r_j^2}\right) \tag{1}$$

onde k é o centro da normal e r é seu raio, o índice j representa o j-*ésimo* neurônio da camada escondida.

O valor da saída do i-*ésimo* neurônio de saída y_i é dado por:

$$y_i = \sum_{j=1}^{m} w_{ij} v_j \tag{2}$$

O método de treinamento da RBF consiste em duas fases, primeiro se calcula os centros e os raios das funções de base e depois os pesos da camada de saída, determinados através de regressão linear [Moody and Darken 1989].

A VInet é uma extensão da rede RBF que calcula o IP e outras medidas de confiança na saída [Leonard et al. 1992]. A idéia básica é que o cálculo do IP da resposta da rede para uma entrada qualquer x é feito através da média dos limites de predição de

todos os neurônios de base radial ponderados pela contribuição de cada unidade escondida para a saída da entrada x. Assim, o cálculo do IP, para uma entrada *x* é:

$$IP(x) = \frac{\sum_{j=1}^{m} v_j(x) IP_j}{\sum_{j=1}^{m} v_j(x)} \tag{3}$$

onde m é o número de neurônios da camada escondida, $v_j(x)$ é a ativação do *m-ésimo* neurônio da camada escondida e IP_j é o IP local, calculado por:

$$IP_j = t_{n_j-1}^{\alpha/2} S_j \left(1 + \frac{1}{n_j}\right)^{1/2} \tag{4}$$

onde n_j é o número de pontos que caem no campo receptivo do neurônio j, dado por:

$$n_j = \sum_{i=1}^{n} v_j(x_i) \tag{5}$$

onde $v_j(x_i)$ é a ativação do neurônio j para o dado de treinamento x_i. S_j é o desvio padrão do erro para cada neurônio, calculado por:

$$S_j^2 = \frac{\sum_{i=1}^{n} v_j(x_i)(y_i - f(x_i; \hat{\theta})^2)}{n_j - 1} \tag{6}$$

onde $(y_i - f(x_i; \hat{\theta})^2$ é a diferença entre o valor desejado y_i e $f(x_i; \hat{\theta})$ é o valor obtido pela rede, para o dado de treinamento x_i.

Uma vantagem importante dessa abordagem de aprendizagem local em relação a abordagem de aprendizagem global para cálculo do IP na saída de redes neurais, é o fato do método não ficar limitado às severas restrições do modelo normalmente impostas na outra abordagem, de forma que o IP calculado é capaz de refletir a distribuição real do ruído. Em regiões do domínio que sejam naturalmente mais ruidosas, o IP fica maior do que em regiões menos ruidosas. Isso acontece na VInet, devido a sua natureza de aprendizagem local, onde cada neurônio da camada escondida vai ter um S_j e um n_j diferentes. Já nos métodos derivados da abordagem de aprendizagem global, as técnicas de regressão não-linear calculam um IP médio e global, que não reflete essa possível variação.

Por outro lado, como o método baseado em aprendizagem local usa basicamente uma composição de regressões lineares, o IP fornecido pelo método não incorpora no cálculo da confiabilidade a densidade de probabilidade dos pontos de treinamento nas diferentes regiões do domínio. Assim, a solução apresentada por Leonard et al. para esta limitação é apresentar além do IP calculado como medida de confiança, mais duas saídas: a densidade e um indicador de extrapolação [Leonard et al. 1992].

Na proposta de Leonard et al. [Leonard et al. 1992] a densidade de probabilidade para o dado de entrada x é calculada utilizando janelas de Parzen [Parzen 1962]. Nesta proposta a densidade para um ponto x é dada por:

$$p(x) = \frac{\sum_{j=1}^{m} v_j(x) p_j}{\sum_{j=1}^{m} v_j(x)} \tag{7}$$

onde p_j é determinado durante a fase de treinamento:

$$p_j = \frac{\sum_{i=1}^{n} v_j(x_i)}{n(\pi^{1/2}\sigma)^N} \tag{8}$$

onde N é o número de dimensões dos dados e i representa o índice de cada dado de treinamento.

Outra medida da VInet é o indicador de extrapolação, que é a maior ativação dos neurônios da camada escondida:

$$max - act = \max_j \{v_j(x)\} \tag{9}$$

Apesar de valores baixos no *max-act* indicarem extrapolação dos dados, também pode-se obter um valor baixo durante a interpolação [Leonard et al. 1992]. Por exemplo, pode acontecer de dois ou mais neurônios serem familiares com um dado, mas nenhum deles gerar uma ativação alta.

Em relação à densidade calculada com janela de Parzen, seu uso é semelhante ao do *max-act*. Um valor muito pequeno, abaixo de um limiar escolhido pelo projetista, significa uma possível resposta com erro elevado, pois esse dado é de uma região que teve poucos exemplos durante o treinamento da rede. A desvantagem da janela de Parzen, é que para dados com muitas dimensões, o valor da densidade fica muito pequeno, a ponto de resultar em problemas de arredondamento e tenderem a zero [Wedding and Cios 1997]. Outro problema é que a confiabilidade do resultado da janela de Parzen melhora com o aumento do número de neurônios na camada escondida, mas nem sempre a escolha ótima do número de neurônios para o cálculo da densidade é igual ao número ótimo para a aproximação da função [Bishop 1994], [Wedding and Cios 1997].

3. Cálculo da densidade usando Mapa Auto-Organizável

A distribuição dos pesos de um mapa auto-organizável ("*Self-Organizing Map*" - SOM) aproxima a distribuição dos dados de treinamento, regiões com mais dados tendem a ser representadas por mais neurônios que são ativados com maior frequência [Holmstrom and Hamalainen 1993]. Dessa forma, é possível usar a SOM para calcular a função distribuição de probabilidade ("*Probability Density Function*" - PDF) dos dados. Uma das maneiras de fazer isso é utilizando o conceito de *mistura de modelos*, que representa a distribuição por meio de uma combinação linear de algumas funções de *kernel* [Bishop 1995].

A densidade de um dado x é dada por:

$$p(x) = \sum_{j=1}^{m} P(j) p(x|j) \tag{10}$$

onde m é o número de neurônios da SOM, $P(j)$ são os pesos da mistura e podem ser interpretados como a probabilidade a *priori* do dado pertencer a célula do diagrama de Voronói correspondente ao neurônio j e $p(x|j)$ representa a densidade condicional do vetor desejado x para o *j-ésimo kernel*.

As probabilidades $P(j)$ devem satisfazer as condições

$$\sum_{j=1}^{m} P(j) = 1 \tag{11}$$

$$0 \leq P(j) \leq 1 \tag{12}$$

Similarmente as probabilidades $p(x|j)$ devem ser normalizadas para garantir que

$$\int p(x|j)dx = 1 \tag{13}$$

Para as funções de *kernel*, uma escolha comum é a gaussiana, onde o centro é o peso do neurônio e o raio é o campo de ativação desse neurônio.

A versão *batch* completa do algoritmo é:

- Treine a SOM e utilize os pesos de cada neurônio do mapa como os centros da funções de *kernel*.
- Calcule as probabilidades $P(j) = \frac{N_i}{N}$ onde N_i é o número de dados que caem na célula do diagrama de Voronói de cada neurônio j.
- Calcule os raios de cada função de *kernel*.

Para encontrar o valor dos raios, uma maneira é usar a distância euclidiana do centro de um neurônio j para os *k-ésimos* centros mais próximos [Terrell and Scott 1992].

A vantagem desse método sobre a janela de Parzen da VInet é não ser dependente da dimensão dos dados, evitando assim o problema de valores tendendo a zero quando a dimensionalidade dos dados de entrada for alta.

3.1. Combinando a densidade com intervalo de predição

Como nem sempre o IP reflete bem a densidade dos dados ou não atingem a Probabilidade de Cobertura do Intervalo de Predição (PCIP) prometida, uma sugestão dada por Shao et al. [Shao et al. 1997] é alterar a fórmula do IP para uma dada entrada x para:

$$IP_f(x) = \frac{2IP(x)}{1 + \frac{\hat{p}(x)}{p_{max}}} \tag{14}$$

Dessa forma, regiões com densidade próxima a zero vão ter o IP com o dobro do tamanho que um dado de uma região com densidade proxima a p_{max}.

4. Testes e Resultados

O objetivo dos testes é comparar a eficiência dos IPs calculado pela VInet com a metodologia de cálculo proposta neste trabalho, em situações onde a densidade dos dados de treinamento é variável ao longo do domínio. Essa comparação é feita treinando as redes e calculando o IP para 95% de confiança, depois são apresentados os dados de teste à rede e calculado a probabilidade de cobertura do intervalo de predição, ou seja, a percentagem dos dados do conjunto de testes estão localizados na região delimitada pelo IP.

Ao todo são instanciadas 100 RBFs para cada teste, treinadas com o mesmo conjunto de dados. Depois é calculado o PCIP para cada uma e finalmente um PCIP médio para todas as redes.

Para o primeiro teste, as redes possuiam 6 neurônios na camada escondida e foram treinadas com um conjunto de dados cuja densidade pode ser vista no histograma mostrado na Figura 2, ao todo foram utilizados 31 pontos de treinamento. A densidade dos dados de treinamento, calculado utilizando a SOM está na Figura 3.

A função utilizada para gerar os dados é dada por:

$$y(x) = 0.5 \sin(1.5\pi x + \pi/2) + 2.0 + v$$

onde x a variável independente, y a variável dependente e v o ruído gaussiano, com média zero e desvio padrão 0.1. Os conjunto de treinamento foi formado escolhendo-se uniformemente 1000 dados dentro do intervalo [-1, 1].

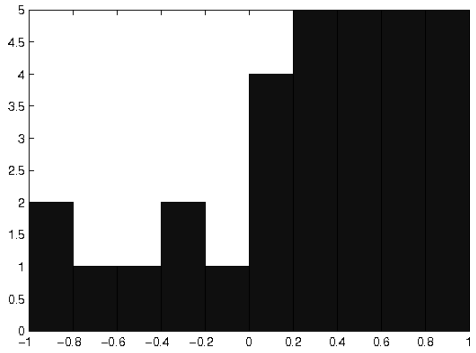

Figura 2. Histograma dos dados de treinamento.

Tabela 1. PCIP para as duas abordagens no primeiro teste.

IP-VInet	84.816
IP-densidade	95.192

A Tabela 1 mostra como ficou o PCIP para as duas abordagens para esse primeiro teste. No caso da VInet, ficou abaixo da cobertura de 95% desejada. A Figura 4 mostra como ficaram os dois IPs calculados pelos dois métodos. O IP calculado pelo método proposto por Leonard et al. praticamente não reflete a densidade dos dados e mantém um IP quase constante tanto para as regiões que receberam mais dados de treinamento,

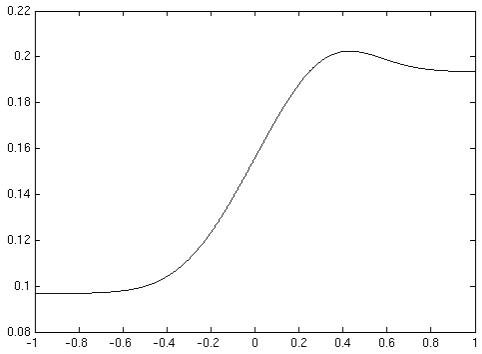

Figura 3. Densidade dos dados de treinamento calculados através da SOM.

quanto para as que receberam menos. O IP com a densidade incorporada foi capaz de refletir a distribuição dos dados de treinamento e obteve um desempenho muito próximo do desejado no PCIP.

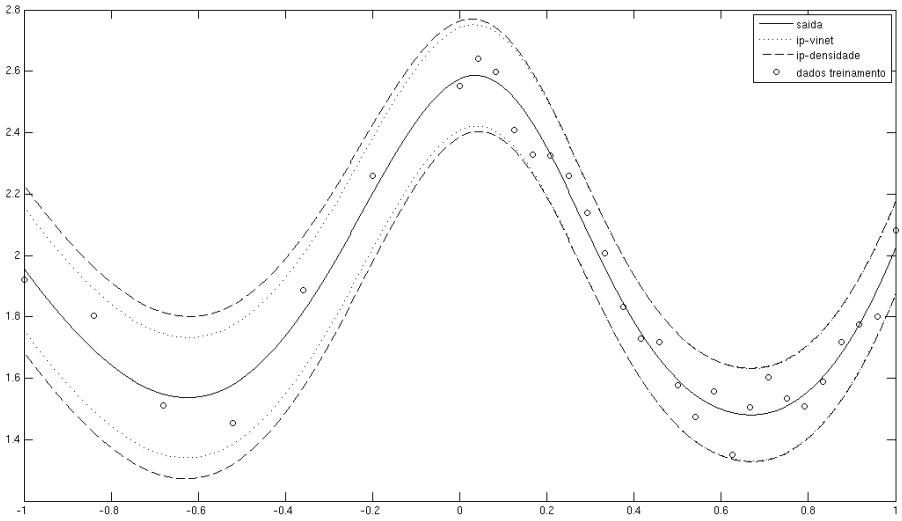

Figura 4. Dados de treinamento, saída da rede e IPs calculados pelas duas metodologias.

No segundo teste foi utilizada uma base de dados com atributos e preços de imóveis disponível em "http://www.inf.ufsc.br/˜ogliari/arquivos/Dados_rede_neural_Gazola.zip". Neste exemplo, os dados possuem 12 dimensões na entrada e uma dimensão na saída e descrevem atributos de imóveis. A entrada são as característica dos imóveis como: número de suítes, se possui dependência de empregada, o estado de conservação, idade, gasto médio de energia, vagas na garagem, dentre outros. A saída desejada é o preço total do imóvel. Ao todo são 397 dados, que foram divididos em 50% para teste e 50% para treinamento. As redes nesse teste foram criadas com 25 neurônios na camada escondida.

A Tabela2 mostra os resultados finais para esse exemplo multidimensional. Foi

Tabela 2. PCIP para as duas abordagens no segundo teste.

IP-VInet	84.8485
IP-densidade	93.90

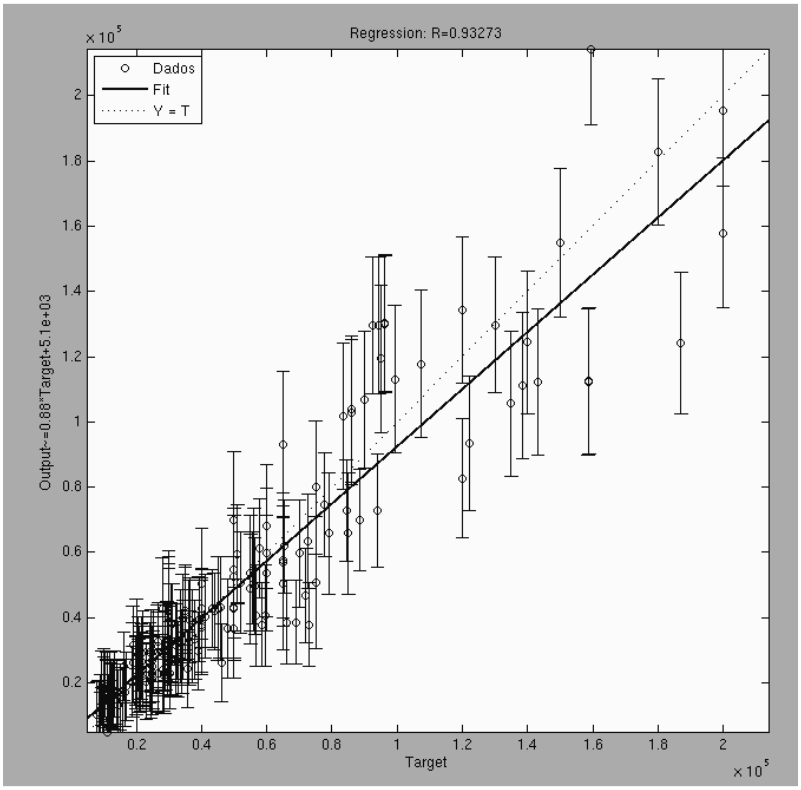

Figura 5. Figura com a regressão linear entre os valores reais versus valores obtidos pela rede, com seu respectivo IP.

semelhante ao resultado anterior e indica que a incorporação da densidade em locais de baixa densidade de treinamento torna o PCIP mais próximo do desejado. A Figura5 mostra o gráfico da saída desejada versus saída calculada por uma das redes instânciadas para o segundo teste. Pelo gráfico pode-se perceber que os imóveis mais caros, possuem menos exemplos parecidos, uma densidade menor e consequentemente apresentaram um IP mais largo que os outros imóveis mais baratos e que aparecem com maior frequência na base de dados.

5. Conclusões

Esse trabalho apresenta opções para corrigir algumas deficiências da VInet, uma maneira de calcular a densidade que não depende da dimensão dos dados, resolvendo assim o problema de se obter densidade com valor zero quando os dados possuem muitas dimensões. Além disso, foi demonstrado que em alguns casos o IP calculado pela VInet não corresponde ao nível de confiança que se deseja, e também não reflete a densidade dos dados de treinamento. Com a incorporação da densidade no IP, os pontos que caírem em regiões onde o treinamento não foi eficiente vão ter um IP mais largo, indicando uma menor confiança para as respostas dessas regiões. Estas modificações são bem vindas em aplicações industriais, onde é necessário ter a certeza que a rede vai continuar a

gerar dados confiáveis após o treinamento. Ao receber dados muito diferentes dos que foram apresentados no treinamento, seu IP será mais largo, indicando uma novidade para a rede e portanto ela deveria ser treinada novamente com exemplos desta região. Um exemplo de aplicação onde isto seria útil é na caracterização de reservatórios de petróleo; se em algum momento forem obtidos IPs mais largos que o desejado para uma região do reservátorio, então seria conveniente obter mais dados de treinamento naquele local, através da perfuração de novos poços. Com essas vantagens, os usuários da rede possuem maior controle para o treinamento e mais confiança nas aproximações.

Referências

Bishop, C. M. (1994). Novelty detection and neural network validation. In *IEE Proc.-Vis. Image Signal Process., Vol. 141*.

Bishop, C. M. (1995). *Neural Netowrks for Pattern Recognition*. Oxford Press.

Chinman, R. and Ding, J. (1998). Prediction limit estimation for neural network models. *Neural Networks, IEEE Transactions on*, 9(6):1515–1522.

Chryssolouris, G., Lee, M., and Ramsey, A. (1996). Confidence interval prediction for neural network models. In *IEEE Trans. Neural Networks 7*, pages 229–232.

Hassoun, M. H. (1995). *Fundamentals of Artificial Neural Networks*. MIT Press.

Holmstrom, L. and Hamalainen, A. (1993). The self-organizing reduced kernel density estimator. In *IEEE International Conference on Neural Networks*, pages 417–421.

Hwang, J. T. G. and Ding, A. A. (1997). Prediction intervals for artificial neural networks. In *J. American Statistical Association 92(438)*, pages 748–757.

Leonard, J. A., Kramer, M. A., and Ungar, L. H. (1992). A neural network architeture that compute its own reliability. In *Computers Chem. Engng.*

Moody, J. and Darken, C. J. (1989). Fast learning in networks of locally tuned processing units. In *Neural Computation 1*, pages 281–294.

Parzen, E. (1962). On estimation of a probability density function and mode ann. In *Math. Statist. 33*, pages 1065–1076.

Shao, R., Martin, E. B., Zhang, J., and Morris, A. J. (1997). Confidence bounds for neural network representations. In *Computers chem. Engng 21*, pages 1173–1178. Elsevier Science Ltd.

Terrell, G. R. and Scott, D. W. (1992). Variable kernel density estimation. In *The Annals of Statistics, Vol. 20, No. 3.*, pages 1236–1265. Institute of Mathematical Statistics.

Veaux, R. D., Schweinsberg, J., and Shellington, D. (1998). Prediction intervals for neural networks via nonlinear regression. In *Computers chem. Engng 21*, pages 1173–1178. Elsevier Science Ltd.

Wedding, D. K. and Cios, K. J. (1997). Cetainty factors versus pazen windows as reliability measures in rbf networks. In *Neurocomputing 19*, pages 151–165. Elsevier Ltd.

Yang, L., Kavli, T., Carlin, M., Clausen, S., and de Groot, P. (1991). An evaluation of confidence bound estimation methods for neural networks. In *ESIT 2000*.