Ajuste da generalização em Redes de Base Radial: uma abordagem multi-objetivo para a estimação de parâmetros

D. H. D. Carvalho[12], M. A. Costa[1] e A. P. Braga[1]
dhdc@ufmg.br, azevedo@est.ufmg.br e apbraga@cpdee.ufmg.br
[1]Universidade Federal de Minas Gerais
Caixa Postal 209, 30.161-970, Belo Horizonte, MG, Brasil
[2]Arte & Byte Sistemas Ltda
Rua Padre Marinho

Resumo

Este trabalho apresenta uma nova proposta para o treinamento de redes de função de base radial (RBFs) utilizando otimização multi-objetivo. Analisa-se a influência dos parâmetros livres das funções de base gaussianas na geração de um conjunto de soluções eficientes, propondo uma medida de complexidade para neurônios da camada escondida. O algoritmo proposto permite encontrar soluções de alta capacidade de generalização.

1. Introdução

As redes de funções de base radial ou RBFs [8] (*Radial Basis Function*), são modelos multi-camadas, semelhantes aos modelos multi-layer perceptron [11]. A principal diferença consiste na transformação, através de uso funções de base radial, dos padrões de entrada da rede RBF para a camada de saída linear. As redes MLPs utilizam funções de ativação não lineares do tipo sigmoidal.

Os métodos de ajuste dos parâmetros da camada de saída bem como a metodologia para a escolha do modelo final são semelhantes [4]. Entretanto, as RBF necessitam, na maioria dos algoritmos, de um ajuste a priori dos parâmetros das funções de base radial representados pelos respectivos centros e raios. O ajuste é realizado com base em um conjunto de observações sobre o qual deseja-se otimizar uma determinada função de custo, como a soma do erro quadrático.

Dessa forma, o ajuste dos parâmetros normalmente é definido em duas etapas: inicialmente procede-se com o ajuste dos centros e raios das funções de base radial e, em seguida, com o ajuste dos pesos da camada de saída.

A escolha do número de funções de base radial e seus parâmetros define a complexidade do modelo e, consequentemente, a sua capacidade de generalização ou a capacidade de aproximar a função representada pelo conjunto de dados. É desejável obter, ao final do ajuste dos parâmetros, uma solução de alta capacidade de generalização ou capaz de equilibrar os efeitos de polarização e variância [3] da rede.

Na abordagem multi-objetivo [13][2], o equilíbrio entre a polarização e variância é representado pela otimização dual da soma quadrática do erro de treinamento e da complexidade da rede neural, representada pela norma dos pesos. A medida da complexidade das redes multi-layer perceptron (MLP) pode ser aplicada aos parâmetros livres presentes na camada de saída das redes RBFs, uma vez que são semelhantes.

O algoritmo proposto neste trabalho tem como objetivo apresentar uma metodologia para ajuste simultâneo dos parâmetros de ambas as camadas presentes em uma rede do tipo RBF (*Radial Basis Function*), encontrando soluções de alta capacidade de generalização através de um treinamento supervisionado multi-objetivo [1]. Para isso, é necessário definir uma medida de complexidade, a ser otimizada, para os neurônios da camada escondida das RBFs.

2. Complexidade em redes RBFs

Os parâmetros livres presentes na camada escondida de uma rede RBF são definidos pelas posições dos centros (μ_m) e pelos valores dos raios (r_m). Para funções de base gaussiana, a saída da camada escondida é descrita pela Equação 1.

$$\phi_m(\mathbf{x_n}) = \mathbf{e}^{\frac{-(\|\mathbf{x_n}-\mu_\mathbf{m}\|)^\mathbf{2}}{\mathbf{2r_m^2}}} \quad (1)$$

A Equação 2 define a matriz de interpolação, $\mathbf{H}$. Seus elementos são determinados pelas c funções de base radial avaliadas para todos os p vetores de entrada, definidos no conjunto de treinamento.

$$\mathbf{H} = \begin{pmatrix} \phi_1(\mathbf{x_1}) & \dots & \phi_c(\mathbf{x_1}) \\ \vdots & & \vdots \\ \phi_1(\mathbf{x_p}) & \dots & \phi_c(\mathbf{x_p}) \end{pmatrix} \quad (2)$$

A saída de uma rede RBF é dada pelo sistema linear representado na Equação 3, onde a matriz $\mathbf{W}$ representa os valores de peso das conexões entre o(s) neurônio(s) da camada de saída e o(s) neurônio(s) da camada escondida.

$$\mathbf{Y} = \mathbf{HW} \quad (3)$$

Deseja-se explorar a situação em que o número de neurônios na camada escondida e as posições dos centros das funções de base gaussianas são pré-definidos, verificando-se a contribuição dos valores de raio para o equilíbrio de polarização e variância.

2.1. Influência do raio no dilema de polarização e variância

Seja a situação em que uma rede RBF possui um número de neurônios na camada escondida igual a quantidade de vetores de amostras de entradas. As suas funções de base possuem centros pré-determinados definidos pelos padrões de entrada do conjunto de treinamento e valores de raio muito inferiores à distância média entre estes centros.

A rede em questão possui um valor de erro de treinamento baixo, porém sua capacidade de generalização fica comprometida, pois existem regiões no espaço dos padrões de entrada que não serão satisfatoriamente mapeados para a camada de saída. O aumento do valor do raio torna a solução mais suave aumentando sua generalização. A Figura 1 ilustra este comportamento. Ela apresenta soluções de um problema de regressão não-linear da função *seno* para distintos valores de raios.

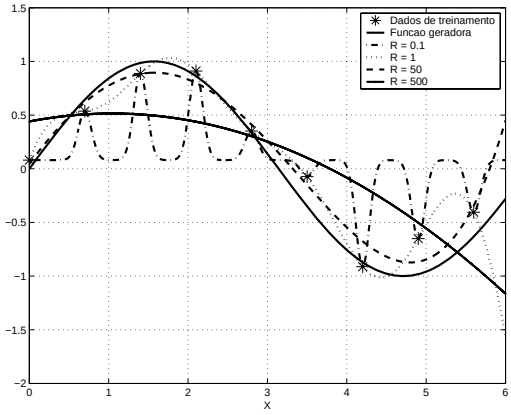

Figura 1: Soluções do problema de regressão da função seno para distintos valores de raios

A Tabela 2.1 demonstra o erro médio quadrático (MSE) e a norma dos pesos que caracterizam as soluções apresentadas.

Nota-se que a solução de melhor generalização, isto é, menor erro para conjunto validação, não possui o menor erro para o conjunto de amostras de treinamento. Neste caso, pode-se observar que existe um determinado valor de raio que garante uma melhor solução, caracterizada pelo equilibrio das contribuições do erro e complexidade do modelo (raio).

Tabela 1: Erro Médio Quadrático (MSE) e Norma do vetor de pesos para aproximação da função *seno*

Soluções	r = 0.1	r = 1	r = 50	r = 500
$\text{MSE}_{Trein.}$	4.2e-32	2.5e-30	0.0152	0.1511
$\text{MSE}_{Valid.}$	0.3414	0.0702	0.0323	0.1886
$\|\|\mathbf{W}\|\|$	1.844	13.173	3.03e+8	2.05e+4

Ao se utilizar valores mínimos de raio, as soluções obtidas com redes RBFs são polarizadas. Neste caso, a matriz de interpolação apresenta *rank* completo fornecendo uma única solução para o vetor de pesos.

Para valores elevados de raio, a matriz de interpolação apresenta *rank* incompleto e as soluções para o vetor de pesos possuem uma alta variância. Além disso, o aumento excessivo do raio contribui para um mal-condicionamento da matriz de interpolação, tornando-a próxima de uma matriz singular, dificultando o cálculo de sua inversa.

Geralmente, utiliza-se um único valor de raio para todas as funções de base. É possível, no entanto, utilizar valores diferentes de raio para cada função de base de uma mesma camada. Uma maneira de quantificar a magnitude deste parâmetro pode ser descrita pela norma de um vetor formado por todos os valores de raio ou simplesmente a soma algébrica dos valores, já que todos os raios são positivos.

Propõe-se apresentar uma medida que seja mais geral de modo a representar não somente os valores de raio mas também a influência das posições dos centros e do número de neurônios da camada escondida, possibilitando o desenvolvimento de algoritmos para o ajuste desses parâmetros.

Os elementos da matriz de interpolação são avaliados segundo os parâmetros das funções de base. A magnitude destes elementos está definida no intervalo real [0,1] segundo uma relação não-linear monotônica com o valor de raio e distância entre centros e padrões de entrada. Logo, o valor da norma da matriz de interpolação leva em consideração a contribuição de todas as funções de base em relação ao comportamento de raios, posição de centros e o número de funções de base para um determinado conjunto de padrões de entrada.

Neste trabalho será abordada somente a contribuição dos valores de raio para a caracterização da matriz de interpolação. O valor da norma da matriz de interpolação

será considerado uma medida para a complexidade da camada intermediária e a norma da matriz de pesos uma medida para a complexidade da camada de saída de redes RBF.

O algoritmo proposto procura obter soluções que equilibrem a relação de polarização e variância de soluções através da caracterização da norma da matriz de interpolação, da norma do vetor de pesos e o erro de treinamento.

2.2. Relação entre camadas

É possível estabelecer uma relação entre as grandezas presentes nas duas camadas de uma rede RBF. Partindo de um valor de raio mínimo e o aumentando gradativamente, os elementos de $\mathbf{H}$ começam a se tornar não-nulos e positivos. A matriz $\mathbf{H}$ modificada pode então ser uma soma algébrica da matriz original (de raios mínimos) e uma matriz de perturbação representada por $\delta\mathbf{H}$ (Equação 4).

$$\mathbf{H_2} = \mathbf{H} + \delta\mathbf{H} \tag{4}$$

Considerando a matriz de saídas desejadas, $\mathbf{Yd}$, constante durante o treinamento, a matriz de pesos também deverá ser modificada de modo a compensar a perturbação $\delta\mathbf{H}$ (Equação 5).

$$\mathbf{W_2} = \mathbf{W} + \delta\mathbf{W} \tag{5}$$

Pode-se então analisar a influência da perturbação $\delta\mathbf{H}$, causada pelo aumento do raio de suas funções radiais, na matriz $\mathbf{W}$ de pesos da solução de uma rede RBF. Considerando-se o mesmo conjunto de treinamento, tem-se:

$$\begin{aligned}
\mathbf{Yd} &= \mathbf{H_2}\mathbf{W_2} \\
\mathbf{Yd} &= (\mathbf{H}+\delta\mathbf{H})(\mathbf{W}+\delta\mathbf{W}) \\
\mathbf{W}+\delta\mathbf{W} &= (\mathbf{H}+\delta\mathbf{H})^{-1}\mathbf{Yd} \\
\delta\mathbf{W} &= -\mathbf{W} + (\mathbf{H}+\delta\mathbf{H})^{-1}\mathbf{Yd} \\
\delta\mathbf{W} &= [(\mathbf{H}+\delta\mathbf{H})^{-1} - \mathbf{H}^{-1}]\mathbf{Y} \\
\delta\mathbf{W} &= [(\mathbf{M})^{-1} - \mathbf{H}^{-1}]\mathbf{Y}
\end{aligned} \tag{6}$$

onde :

$$\begin{aligned}
\mathbf{M}^{-1} - \mathbf{H}^{-1} &= (\mathbf{H}^{-1}\mathbf{H})\mathbf{M}^{-1} - \mathbf{H}^{-1}(\mathbf{M}\mathbf{M}^{-1}) \\
\mathbf{M}^{-1} - \mathbf{H}^{-1} &= \mathbf{H}^{-1}(\mathbf{H}-\mathbf{M})\mathbf{M}^{-1}
\end{aligned} \tag{7}$$

substituindo 7 em 6,

$$\begin{aligned}
\delta\mathbf{W} &= [\mathbf{H}^{-1}(\mathbf{H}-\mathbf{M})\mathbf{M}^{-1}]\mathbf{Y} \\
\delta\mathbf{W} &= [\mathbf{H}^{-1}(\mathbf{H}-(\mathbf{H}+\delta\mathbf{H}))(\mathbf{H}+\delta\mathbf{H})^{-1}]\mathbf{Y} \\
\delta\mathbf{W} &= \mathbf{H}^{-1}(-\delta\mathbf{H})(\mathbf{H}+\delta\mathbf{H})^{-1}\mathbf{Y} \\
\delta\mathbf{W} &= \mathbf{H}^{-1}(-\delta\mathbf{H})(\mathbf{W}+\delta\mathbf{W})
\end{aligned}$$

e aplicando normas consistentes tem-se a seguinte inequação:

$$\| (\delta\mathbf{W}) \| \quad \leq \quad \| \mathbf{H^{-1}} \| \| -\delta H \| \| (\mathbf{W}+\delta\mathbf{W}) \|$$

Substituindo $cond(H) = \| \mathbf{H^{-1}} \| \| \mathbf{H} \|$ tem-se:

$$\frac{\| (\delta\mathbf{W}) \|}{\| (\mathbf{W}+\delta\mathbf{W}) \|} \quad \leq \quad cond(\mathbf{H}) \frac{\| \delta\mathbf{H} \|}{\| \mathbf{H} \|} \tag{8}$$

Analisando a Inequação 8 pode-se concluir que perturbações em $\mathbf{H}$, ocasionadas, por exemplo, pelo aumento dos raios possibilitam perturbações relativas na matriz de pesos. Se as perturbações na matriz de pesos são altas então a norma também será. Conclui-se que o aumento da norma de $\mathbf{H}$ permite o aumento no valor da norma de pesos. Desta forma, a norma de $\mathbf{H}$ e a norma de $\mathbf{W}$ são correlacionadas e representam medidas de complexidade para a rede RBF.

3. Treinamento MOBJ para redes RBF

A partir do estudo apresentado anteriormente, pretende-se avaliar a contribuição da norma de $\mathbf{H}$, representada pelo valor de raio, como medida de complexidade para o treinamento de redes RBFs. Sabe-se que existem valores de raios que possibilitam uma maior suavização da resposta. Existe então um comprometimento de seu valor de forma a possibilitar o equilíbrio dos efeitos de polarização e variância.

Dentro do contexto de treinamento multi-objetivo para redes neurais artificiais, espera-se que determinados valores de raios sejam capazes de gerar aproximações eficientes do conjunto pareto, ou seja, mais próximas do eixo das abcissas, onde se localizam as soluções de maior capacidade de generalização.

A Figura 2 apresenta as soluções obtidas para valores distintos de raios, ou seja, valores distintos de $||\mathbf{H}||$ para uma mesma arquitetura de rede RBF. Observa-se que o problema de se encontrar o conjunto pareto mais eficiente é representado por um problema de otimização cuja solução corresponde a um valor de raio que minimiza a distância média das soluções com relação ao ponto de origem do plano erro x norma de $\mathbf{W}$, conforme ilustra a Figura 3.

Conclui-se que existe um determinado valor de raio que permite alcançar as soluções mais eficientes, conforme ilustra a aproximação do pareto na Figura 2. No exemplo, o raio de valor unitário caracteriza tal situação.

Uma proposta inicial para se obter um subconjunto de soluções mais eficientes pode ser definida pela geração de várias aproximações para o conjunto pareto cada qual com um valor de raio pré-definido. Selecionando, como conjunto final, as soluções que estiverem mais próximas da solução utópica de erro e norma nulos.

Entretanto, pretende-se evitar a geração de várias aproximações do conjunto pareto a partir do uso de valores pré-definidos para os raios. Apresenta-se na seção seguinte uma metodologia para geração automática de soluções eficientes.

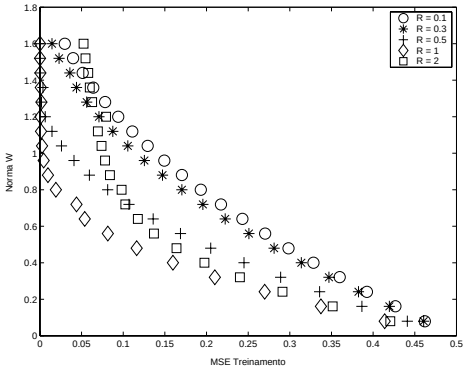

Figura 2: Aproximações do conjunto Pareto para distintos valores de raios

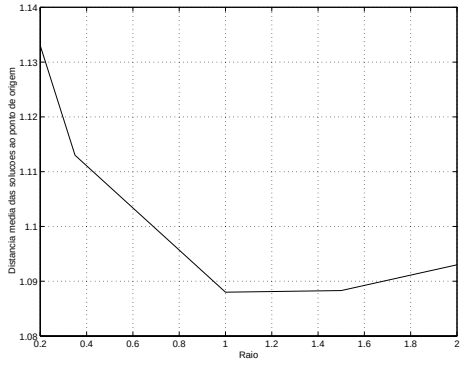

Figura 3: Distância média das soluções do Pareto em relação à origem do plano, para valores distintos de raios

3.1. Geração de soluções eficientes

Para a geração de cada aproximação do conjunto pareto, resolve-se o problema descrito pela Equação 9.

$$\mathbf{W}^* = arg_{\mathbf{W}} \ \ min \begin{cases} f_1(\mathbf{W}, \mathbf{r}) = e(\mathbf{W}, \mathbf{r}) \\ f_2(\mathbf{W}) = \|\mathbf{W}\| \end{cases} \tag{9}$$

Cada solução pode ser obtida, restringido o valor do funcional f_2 a valor pré-determinado e minimizando o funcional f_1.

Sob a hipótese de que o raio utilizado na resolução do problema apresentado em 9 não caracteriza a melhor resposta para a RBF, realiza-se então uma otimização do valor deste parâmetro segundo o problema mono-objetivo descrito em 10, visando a encontrar soluções mais eficientes.

$$\begin{aligned} &\mathbf{r}^* = arg_{\mathbf{r}} \ \ min \ \ f(\mathbf{W}, \mathbf{r}) = e(\mathbf{W}, \mathbf{r}) \\ &sujeito \ \ a : g(W) = \|\mathbf{W}\| < \|\mathbf{W}^*\| \end{aligned} \tag{10}$$

A solução do problema de otimização mono-objetivo (Equação 10) fornece um valor para o raio que minimiza o erro de treinamento para um determinado valor de $||\mathbf{W}||$. Isso significa que a partir de um valor de raio pré-definido, pode-se obter uma solução arbitrária no espaço de soluções e caminhar em direção a uma solução de erro reduzido obtendo uma solução mais eficiente, como ilustram as Figuras 4 e 5.

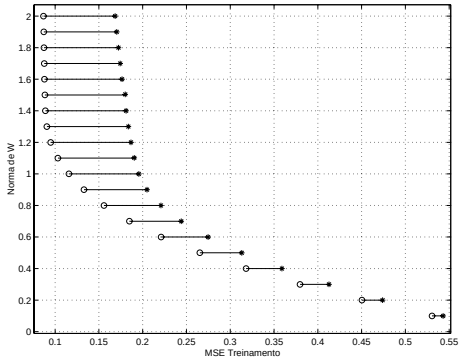

Figura 4: Deslocamento de soluções ocasionado pela otimização do valor de raio minimizando erro de treinamento

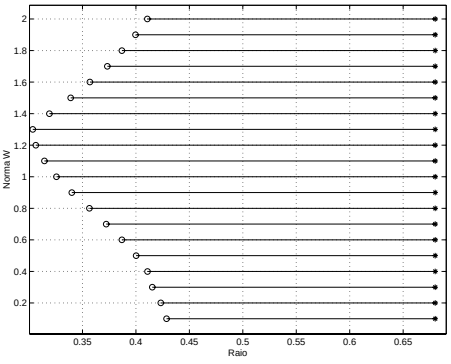

Figura 5: Deslocamento de soluções ocasionado pela otimização do valor de raio determinando valores distintos para cada restrição de norma de $\mathbf{W}$

Cada solução obtida possui um valor de raio distinto que minimiza o erro de treinamento mantendo constante o valor de norma. Desta forma realiza-se uma geração automática de soluções candidatas à solução de maior capacidade de generalização, definindo-se também os valores associados dos pesos das conexões da camada de saída e dos raios das funções de base da camada intermediária de redes RBF.

3.2. Decisor

A solução de maior capacidade de generalização está presente nas soluções constituíntes do conjunto Pareto [13]. O procedimento de seleção de modelo se restringe a um conjunto restrito de soluções ditas mais eficientes. Aplica-se um decisor para se escolher a melhor solução baseando-se no critério de menor erro para um conjunto de validação.

4. Resultados

Foram realizados testes comparativos entre o algoritmo proposto e outros algoritmos para controle de complexidade para redes RBF. São abordados dois problemas de regressão não-linear e um problema de predição de uma série temporal caótica.

As soluções encontradas pelo método multi-objetivo são comparadas com as soluções obtidas com o algoritmo de treinamento NEWRB [7] e métodos de regularização [9] utilizando critérios de seleção de modelos: *Unbiased estimate of variance* (UEV), *final prediction error* (FPE), *generalised cross-validation* (GCV) e *Bayesian information criterion* (BIC) [10].

Para a execução do algoritmo multi-objetivo, as posições dos centros foram selecionadas segundo o algoritmo K-means [5]. A geração do conjunto pareto foi realizada segundo o método ε-restrito [1] utilizando o algoritmo elipsoidal [12]. A topologia RBF utilizada apresenta uma entrada, 15 funções de base radial e uma saída (1-15-1) e foi padronizada para todas as simulações.

O primeiro problema de regressão não-linear é representado por um conjunto de 80 amostras obtidas a partir da função seno na qual foi adicionada um ruído gaussiano.A Tabela 2 apresenta uma comparação entre os erros obtidos para cada algoritmo. Observa-se que o treinamento multi-objetivo foi capaz de encontrar uma solução de alta capacidade de generalização, comprovado pelo o valor do erro de validação.

Tabela 2: Resultados de treinamento para problema de aproximação da função seno

Método	MSE Trein.	MSE Valid.
NEWRB	0.07322	0.00861
Ridge Regression GCV	0.081763	0.001302
Ridge Regression BIC	0.081934	0.001442
Ridge Regression MML	0.075834	0.004716
Ridge Regression FPE	0.081746	0.001305
Ridge Regression UEV	0.076115	0.003896
MOBJ-RBF	0.081209	0.001152

O segundo problema de regressão não-linear é representado por um conjunto de 80 amostras de uma função sinc na qual foi adicionada um ruído gaussiano. A Tabela 3 apresenta os resultados obtidos.

Tabela 3: Resultados de treinamento para problema de aproximação da função sinc

Método	MSE Trein.	MSE Valid.
NEWRB	0.074691	0.016119
Ridge Regression GCV	0.077487	0.021418
Ridge Regression BIC	0.073190	0.014418
Ridge Regression MML	0.105807	0.026284
Ridge Regression FPE	0.091833	0.021149
Ridge Regression UEV	0.079689	0.021055
MOBJ-RBF	0.084691	0.014633

A solução encontrada pelo método multi-objetivo é de qualidade próxima à encontrada pelo critério de informação Bayesiana (BIC).

O terceiro experimento constitui um problema de predição de valores para a série temporal caótica de Mackey-Glass. O vetor de dados de entrada (11) é formado por dados gerados a partir de uma solução obtida pelo método de Runge-Kutta de quarta ordem descrito em [6].

$$\mathbf{X} = \begin{pmatrix} x(t-18) & x(t-12) & x(t-6) & x(t) \end{pmatrix} \quad (11)$$

Deseja-se predizer o valor a seis passos a frente x(t+6). Foram utilizados 500 pontos para treinamento e 500 para validação. A topologia utilizada para uma rede RBF apresenta 4 entradas 15 funções de base radial e uma saída.

A Figura 6 apresenta as respostas para o conjunto de validação.

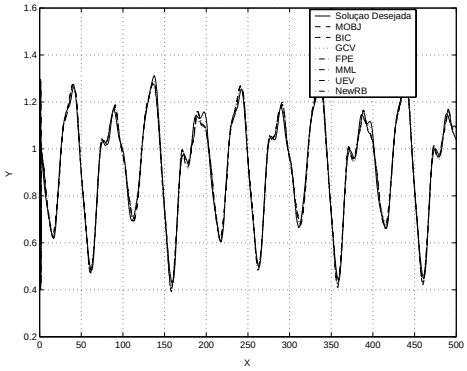

Figura 6: Série Caótica

Para o problema de predição de série, o método proposto obteve um resultado próximo aos demais métodos.

Tabela 4: Resultados de treinamento para problema de predição da série temporal caótica

Método	MSE Trein.	MSE Valid.
NEWRB	0.000271	0.000260
Ridge Regression GCV	0.000579	0.000550
Ridge Regression BIC	0.000657	0.000616
Ridge Regression MML	0.000353	0.000325
Ridge Regression FPE	0.000537	0.000513
Ridge Regression UEV	0.000324	0.000308
MOBJ-RBF	0.000424	0.000404

Os resultados demonstram que o treinamento multi-objetivo proposto é capaz de encontrar soluções de alta capacidade de generalização.

5. Discussões e conclusões

Apresentou-se neste trabalho um novo algoritmo de treinamento para redes RBFs. A metodologia apresentada é capaz de definir de maneira ótima o valor de raio e pesos da camada de saída, encontrando soluções de alta generalização.

Semelhante ao comportamento multi-objetivo para o treinamento de redes MLPs, os resultados obtidos indicam que a aproximação do conjunto pareto representa o conjunto de soluções eficientes do qual é possível extrair a solução de maior capacidade de generalização, no contexto das redes RBFs.

O trabalho apresentado contribui para uma extensão do treinamento multi-objetivo de redes MLP [13] aplicado ao treinamento de redes RBFs. O estudo pode ser expandido para avaliar a contribuição de outros parâmetros, bem como outros tipos de funções de base radiais.

Referências

[1] V. Chankong and Y. Y. Haimes. Multiobjetive decision making : Theory and methodology. 1983.

[2] MA Costa, AP Braga, BR de Menezes, GG Parma, and RA Teixeira. Control of generalization with a bi-objective sliding mode control algorithm. In *XVI Brazilian Symposium on Neural Networks*, november 2002.

[3] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, (4(1)):1–58, 1992.

[4] Simon Haykin. *Neural Networks : A Comprehensive Foundation*. Macmillan, NY, 2001.

[5] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, CA, 1967. University of California Press.

[6] MathWorks. *Fuzzy Logic Toolbox Users's Guide*. 2001.

[7] MathWorks. *Neural Networks Toolbox Users's Guide*. 2001.

[8] J. Moody and C. Darken. Learning with localised receptive fields. Research report, Yale University Department of Computer Science, 1988.

[9] M. J. L. Orr. Introdution to radial basis function networks. Technical report, University of Edinburgh, 1996.

[10] M. J. L. Orr. Matlab routines of subset selection and ridge regression in linear neural networks. Technical report, University of Edinburgh, 1997.

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representation by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 318–362. Cambridge, MA: MIT Press, 1986.

[12] N. Z. Shor. Cut-off method with space extension in convex programing problems. *Cybernetics*, pages 94–96, 1977.

[13] R. A. Teixeira, A. P. Braga, R. H. C. Takahashi, and R. R. Saldanha. Improving generalization of mlps with multi-objective optimization. *Neurocomputing*, 35(1–4):189–194, 2000.