Anais do IX Congresso Brasileiro de Redes Neurais / Inteligência Computacional (IX CBRN)
Ouro Preto 25-28 de Outubro de 2009

SELEÇÃO DE MODELOS BASEADO NA APLICAÇÃO DE PCA À REDE MLP DUAL

Cláudio M. S. Medeiros*, Guilherme A. Barreto†

*Depto. Mecatrônica, Instituto Federal de Educação, Ciência e Tecnológica do Ceará
Av. 13 de Maio, 2081 - Benfica, Fortaleza, Ceará.

†Depto. Engenharia de Teleinformática, Universidade Federal do Ceará (UFC)
Av. Mister Hull, S/N - Campus do Pici, Centro de Tecnologia, Bloco 725
CP 6007, CEP. 60455-970, Fortaleza, Ceará.

Emails: claudiosa@cefetce.br, guilherme@deti.ufc.br

Abstract— In this paper we present a novel method for pruning redundant hidden neurons of a trained multilayer Perceptron (MLP). The proposed method is based on the principal components analysis (PCA) applied to the dual network, which is linear in parameters. Simulations using real-world data indicate that the proposed method presents equivalent or better performance than traditional pruning techniques.

Keywords— Model Selection, MLP, Dual Network, PCA.

Resumo— Neste artigo é introduzida uma metodologia de poda de neurônios ocultos redundantes em um perceptron multicamadas (MLP) previamente treinado. O método proposto é baseado na análise de componentes principais (PCA) aplicada à rede dual, a qual é linear nos parâmetros. Simulações computacionais usando dados reais indicam que o método proposto apresenta desempenho equivalente ou melhor que técnicas tradicionais de poda.

Keywords— Seleção de Modelos, MLP, Rede Dual, PCA.

1 Introdução

Um passo crucial no projeto de MLPs está relacionado com o problema de *seleção de modelos neurais*. Este problema, o qual ainda é um tema atual de pesquisas (Seghouane & Amari 2007, Curry & Morgan 2006, Nakamura et al. 2006, Xiang et al. 2005), pode ser entendido como achar a menor arquitetura que generalize bem, fazendo boas predições para dados desconhecidos.

Recentemente, os algoritmos evolucionários têm se apresentado como ferramentas atrativas para a determinação de topologias sub-ótimas. O projetista começa treinando uma rede com um pequeno número de neurônios ocultos e adiciona neurônios durante o processo de treinamento, com o objetivo de atingir uma estrutura neural que satisfaça as especificações de projeto. Esta é a abordagem usada pela arquitetura conhecida como *Cascade-Correlation* (Fahlman & Lebiere 1990). Entretando, em geral, estes algoritmos apresentam alguma dificuldade no início do processo de evolução, pois pequenas redes tendem a ser mais sensíveis a condições iniciais e parâmetros de treinamento, e são mais suscetíveis a cair em mínimos locais do que grandes redes, dificultando assim, a adição de novos neurônios.

Por outro lado, como alertado por Reed (1993), grandes redes geralmente aprendem relativamente rápido e apresentam menor sensibilidade à inicialização dos pesos. Entretanto, tais topologias apresentam incovenientes como grande demanda de memória, condição que pode ser crítica em sistemas embarcados, e tempos de execução inaceitáveis em aplicações em tempo real. Além disso, topologias com parâmetros em excesso são mais propensas à ocorrência de *overfitting*.

A aplicação de um método de poda de neurônios, entretanto, leva quase sempre a uma redução da estrutura ao eliminar neurônios desnecessários (ou redundantes), além de geralmente melhorar a capacidade de generalização da rede. Isto favorece o uso de métodos de poda em relação a algoritmos evolucionários. Nesta metodologia nós podemos treinar uma rede com um número relativamente grande de neurônios ocultos e então podamos os neurônios menos significantes ou redundantes. Esta é a abordagem sugerida neste artigo.

O restante deste artigo é organizado como se segue. Na Seção 2 nós abordamos brevemente a rede dual. Na Seção 3 o algoritmo proposto é apresentado. Os resultados de simulações são apresentados na Seção 4. Nós concluímos o artigo na Seção 5 com um sumário dos principais aspectos relacionados à aplicação do método proposto e sugestões para desenvolvimentos futuros.

2 A Rede Dual

A aplicação do algoritmo *backpropagation* para o treinamento de MLPs requer dois passos computacionais: o de sentido direto e o de sentido reverso. Durante o sentido direto, com a apresentação do t-ésimo padrão, são realizados os cálculos das ativações e das saídas dos neurônios

com a manutenção dos pesos sinápticos inalterados. A relação entrada-saída é não-linear, pois os neurônios são equipados com funções de ativação sigmoidais. Entretanto, após o cálculo do erro em cada neurônio de saída ($e_k^{(o)}(t)$), os pesos são atualizados, camada a camada, em função dos erros retropropagados a partir da camada de saída. Aqui, o fluxo de informação ocorre no sentido reverso.

A topologia da rede MLP durante a fase de retropropagação dos erros reduz-se a uma estrutura linear nos parâmetros, comumente chamada de rede MLP dual, ou simplesmente rede dual (Principe et al. 2000), como pode ser visto na Figura 1. O erro projetado em cada um dos Q neurônios ocultos é uma combinação linear dos gradientes locais dos neurônios de saída. Caso a função de ativação dos neurônios de saída seja linear, o erro projetado no i-ésimo neurônio oculto ($e_i^{(h)}(t)$) pode ser expresso simplesmente como a combinação linear dos erros na camada de saída, ou seja

$$e_i^{(h)}(t) = \sum_{k=1}^{M} m_{ki}(t) e_k^{(o)}(t), \qquad (1)$$

em que $m_{ki}(t)$ é o peso entre o i-ésimo neurônio oculto e o k-ésimo neurônio de saída e M é o número de neurônios de saída.

É importante ressaltar que os procedimentos tradicionais de seleção de modelos neurais não se aproveitam do fato da dinâmica da rede dual ser linear nos parâmetros. Esta propriedade permite que uma gama de métodos e técnicas clássicas oriundas da Álgebra Linear e da Estatística possam ser explorados com o objetivo de encontrar topologias ótimas para modelos neurais.

3 A Metodologia Proposta: PCA Aplicado à Rede Dual

A aplicação de PCA a MLPs no sentido direto da rede como uma ferramenta de seleção de modelos neurais é uma prática bastante investigada (Chen et al. 2001, Henrique et al. 2000). Entretanto, o uso desta técnica sobre a rede dual, a qual é linear nos parâmetros, se apresenta como uma novidade.

A aplicação de PCA à rede dual é usada como uma ferramenta para estabelecer o ordenamento dos neurônios ocultos em função de suas relevâncias para a solução do mapeamento entrada-saída da rede. O algoritmo elaborado para uma rede MLP com uma camada oculta apresenta os seguintes passos:

Passo 1: Apresentar os N padrões do conjunto de dados de treinamento à rede previamente treinada e calcular a matriz $\mathbf{E}_h$[1];

$$\mathbf{E}_h = \begin{bmatrix} e_0^{(h)}(1) & \cdots & e_Q^{(h)}(1) \\ e_0^{(h)}(2) & \cdots & e_Q^{(h)}(2) \\ \vdots & \vdots & \vdots \\ e_0^{(h)}(N) & \cdots & e_Q^{(h)}(N) \end{bmatrix}_{N\times(Q+1)} \qquad (2)$$

Passo 2: Calcular a matriz de covariância ($\mathbf{C}_h$) de $\mathbf{E}_h$;

Passo 3: Calcular autovalores e autovetores de $\mathbf{C}_h$;

Passo 4: Organizar em ordem decrescente os autovalores de $\mathbf{C}_h$ baseado em seus valores absolutos e construir uma matriz de transformação $\mathbf{M}$ a partir dos autovetores de $\mathbf{C}_h$. A primeira linha de $\mathbf{M}$ é composta pelo transposto do autovetor correspondente ao autovalor que possui o maior valor absoluto, a segunda linha de $\mathbf{M}$ é composta pelo transposto do autovetor correspondente ao autovalor que possui o segundo maior valor absoluto e etc.;

Passo 5: Aplicar uma transformação à matriz $\mathbf{E}_h$ dada por

$$\mathbf{E}_h^* = \mathbf{M}.\mathbf{E}_h^T, \qquad (3)$$

onde $\mathbf{E}_h^*$ é a matriz transformada dos erros projetados na camada escondida. A primeira linha de $\mathbf{E}_h^*$ supostamente contém as informações mais relevantes no espaço transformado;

Passo 6: Encontrar similaridade entre as linhas de $\mathbf{E}_h^*$ e as colunas de $\mathbf{E}_h$ através do produto escalar (correlação). O neurônio oculto, cujo vetor de erros projetados esteja mais correlacionado com a primeira linha de $\mathbf{E}_h^*$ (informação mais relevante no espaço transformado), é supostamente o neurônio mais relevante da camada escondida. Ao contrário, o neurônio oculto, cujo vetor de erros projetados esteja mais correlacionado com a última linha de $\mathbf{E}_h^*$ (informação menos relevante no espaço transformado), é supostamente o neurônio menos relevante da camada escondida.

Passo 7: Montar um vetor de índices de posição dos neurônios na camada oculta na ordem crescente de suas relevâncias. O primeiro elemento contém o índice de posição do neurônio oculto considerado menos relevante para a solução da rede MLP.

[1] As linhas da matriz $\mathbf{E}_h$ correspondem aos erros retropropagados associados aos neurônios da camada escondida. Em particular, a primeira coluna de $\mathbf{E}_h$ corresponde aos erros retropropagados associados aos limiares.

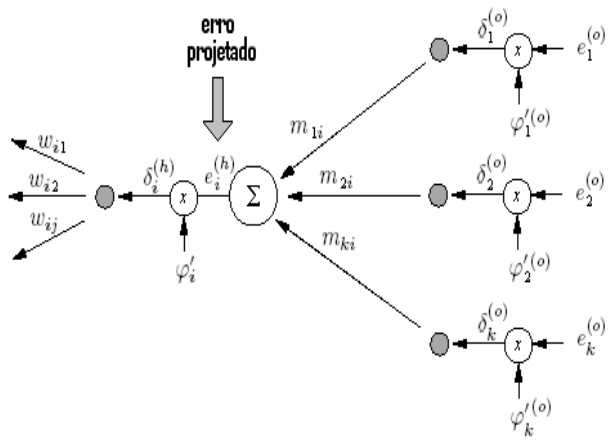

Figura 1: Rede Dual: topologia linear nos parâmetros.

O algoritmo que efetivamente realiza a poda dos neurônios ocultos de uma MLP com uma camada escondida usa o vetor de índices de posição $Ip = [Ip_1\ Ip_2\ \ldots\ Ip_q\ \ldots\ Ip_Q]$ como referência para o seqüenciamento da poda. O procedimento pode ser visto na Tabela 1, o qual deve ser aplicado sucessivamente até que não haja mais poda de neurônio. Na Tabela 1, J_{train} é a taxa de acerto no conjunto de treinamento e J_{tol} é a taxa de acerto mínima tolerável.

4 Simulações e Resultados

Os resultados da aplicação de PCA à rede MLP dual são apresentados na Tabela 2. Nela, foram utilizados alguns dos mais conhecidos conjuntos de dados disponíveis em *sites* relacionados a reconhecimento de padrões e, para efeito de comparação, também são apresentados resultados referentes à aplicação do método de poda denominado CAPE (Medeiros & Barreto 2007)[2] e de PCA à rede direta. Esta última metodologia se utiliza basicamente dos mesmos procedimentos do PCA aplicado à rede dual, mas é aplicada à rede direta. Aqui, a matriz $\mathbf{E}_h$, referente aos erros projetados na camada oculta, é substituída pela matriz $\mathbf{Y}_h$, de mesma dimensão, mas referente à saída dos neurônios ocultos mediante a apresentação de todo o conjunto de treinamento.

Os conjuntos de dados utilizados para a avaliação do método proposto foram *Iris*, *Wine*, *Dermatology*[3] e Coluna Vertebral. O conjunto *Iris* é composto de 150 padrões com 4 atributos distribuídos equitativamente entre 3 classes (Setosa, Versicolor e Virgínica). O conjunto *Wine* é composto por 178 padrões com 13 atributos também distribuídos em 3 classes de produtores de vinho com 59, 71 e 48 padrões, respectivamente. O conjunto *Dermatology* é composto por 366 padrões com 34 atributos distribuídos em 6 classes de doenças de pele: Psoríase(112), Dermatite seborréica(61), Líquen plano(72), Pitiríase rósea(49), Dermatite crônica(52) e Pitiríase rubra pilar(20). A informação entre parêntesis refere-se ao número de padrões por classe.

O conjunto de dados denominado como Coluna Vertebral contém dados biomédicos relacionados a patologias na coluna vertebral. A questão consiste em classificar pacientes pertencentes a uma de três categorias: Normal (100 pacientes), Hérnia de Disco (60 pacientes) ou Spondilolistese (150 pacientes). Cada paciente é representado na base de dados por seis atributos relacionados com a forma e orientação da pélvis e da coluna vertebral. Maiores detalhes sobre esses atributos bem como sua relação com as patologias na coluna vertebral podem ser encontrados em (Berthonnaud et al. 2005).

Nos conjuntos de dados *Wine* e *Dermatology* foram utilizados todos os padrões para treinamento. Para o conjunto de dados *Iris* foram reservados 35 padrões por classe para treinamento e o restante para teste. No caso do conjunto Coluna Vertebral, um total de 42 padrões por classe foram selecionados aleatoriamente para formarem o conjunto de treinamento. Este número refere-se 70% dos dados representativos da classe com menor número de amostras. Os 184 exemplos restantes foram usados para o propósito de testes.

Em todas as simulações apresentadas neste artigo o procedimento de normalização dos dados e de inicialização do processo de treinamento são os mesmos. Os pesos são inicializados randômicamente na faixa ente -0.5 to $+0.5$. Os neurônios usam funções

[2]O procedimento consiste em inicialmente treinar a rede MLP com um número relativamente grande de neurônios na camada escondida, e então descartar os pesos *desnecessários* ou *redundantes*, melhorando assim o desempenho da rede na generalização.

[3]Disponíveis em www.ics.uci.edu/∼mlearn.

Tabela 1: Procedimento de poda de neurônios ocultos baseado na aplicação de PCA à rede dual.

1.	Faça $q = 1$;	`// atribua 1 ao índice de contagem`
2.	**ENQUANTO** $q \leq Q$ **FAÇA**	`// comece o ciclo de poda`
	2.1. Faça $a_i = m_{i,Ip_q}$; $i = 1, \ldots, M$	`// salve os pesos e bias relacionados`
	$b_j = w_{Ip_q,j}$; $j = 1, \ldots, Q+1$	`// ao neurônio a ser podado`
	2.2. Faça $m_{i,Ip_q} = 0$; $i = 1, \ldots, M$	`// atribua zero aos pesos e bias`
	$w_{Ip_q,j} = 0$; $j = 1, \ldots, Q+1$	`// relacionados ao neurônio a ser podado`
	2.3. Calcule J_{train};	`// taxa de acerto no conjunto de treinamento`
	2.4. **SE** $J_{train} < J_{tol}$,	
	Faça $m_{i,Ip_q} = a_i$; $i = 1, \ldots, M$	`// recupere os pesos e bias`
	$w_{Ip_q,j} = b_j$; $j = 1, \ldots, Q+1$	`// relacionados ao neurônio`
	FIM DO SE	
	2.5. Faça $q = q + 1$;	`// continue a poda`
	FIM DO ENQUANTO	

de ativação do tipo tangente hiperbólica. Os vetores de rótulos de saída usam codificação binária 1-*out-of*-M. A taxa de aprendizado foi ajustada em $\eta = 0.001$. Todas os algoritmos foram implementados em MATLAB 7.0. Os dados foram pré-processados pela remoção da média e normalização pelo desvio-padrão.

A Tabela 2 apresenta resultados de poda de neurônios em duas redes MLP, com mesma estrutura ($Q = 9$), treinadas para a classificação do conjunto de dados Iris. A aplicação dos métodos de seleção de modelos a estas estruturas leva a resultados bastante diferentes. No primeiro caso, os três métodos apresentam resultados semelhantes, observando-se plena coincidência na poda dos 3 neurônios ocultos (coluna N_Q Podados) quando se refere ao PCA aplicado à rede direta (PCA) e à rede dual (PCAD). No caso da aplicação do CAPE, também obteve-se a poda de 3 neurônios ocultos, 2 dos quais coincidentes com os resultados anteriores. No segundo caso, observa-se que os neurônios 1, 2, 4, 6 e 8 são podados pelos três métodos, mas que os métodos baseados em PCA ainda sugerem a poda dos neurônios 3 e 5, e ainda mais, o PCAD sugere a eliminação dos pesos relacionados ao *bias*. É importante notar que a aplicação dos métodos às duas redes resultaram em desempenhos de podas bastante diferentes. Embora as taxas finais de acerto nos conjuntos de treinamento (CR_{tr}) e teste (CR_{ts}) sejam iguais, os números de neurônios ocultos (Q) e conexões (N_c) remanescentes são bem distintos, além dos erros nos conjuntos de treinamento (ε_{tr}) e teste (ε_{ts}). Esta discrepância deve-se ao fato de que cada processo de treinamento conduz o posicionamento dos hiperplanos de forma diferente, criando suas próprias redundâncias, mesmo que ao final, tenha-se taxas de classificação semelhantes.

Em seguida, os três métodos foram aplicados às redes treinadas com os conjuntos de dados *Wine* e *Dermatology*. Na rede treinada com o conjunto *Wine* o desempenho do PCAD foi superior aos demais, enquanto na rede treinada com o conjunto *Dermatology* o método CAPE apresentou clara vantagem. Digno de nota é o fato do PCAD sempre recomendar a poda de pelo menos um neurônio em comum com os outros dois métodos.

No caso da aplicação dos métodos à rede MLP treinada com o conjunto de dados referente a doenças relacionadas com a coluna vertebral, a poda com o PCAD resultou no descarte do maior número de neurônios ocultos. A coincidência na poda de neurônios entre os três métodos ocorre apenas em 2 neurônios (1 e 9).

Nos três exemplos apresentados na Tabela 2 observa-se a drástica redução do número de conexões sinápticas (N_c) alcançada pela aplicação do CAPE. Mesmo na aplicação dos métodos à rede MLP treinada com o conjunto de dados Coluna, quando o PCAD descartou um número maior de neurônios ocultos (7) do que o CAPE (5), o descarte de conexões sinápticas é maior com o CAPE. Entretanto, neste caso particular, a maior redução do número de neurônios ocultos obtida com o PCAD pode ter um maior impacto no custo computacional da rede podada, pois a poda de neurônios ocultos efetivamente reduz a dimensão das matrizes de pesos, reduzindo o número de operações matemáticas executadas na aplicação da rede MLP, enquanto o descarte de conexões sinápticas sem a eliminação de neurônios ocultos apenas acrescenta zeros às matrizes de pesos.

5 Conclusões

Este artigo introduziu a aplicação de PCA à rede MLP Dual, um procedimento eficiente e de fácil aplicação para poda de neurônios ocultos desnecessários de uma rede MLP previamente treinada. O método baseia-se na análise de componentes principais sobre a matriz de erros retroprojetados na camada oculta de MLPs com uma única camada escondida. O método pode ser facilmente

Tabela 2: Comparação PCA e CAPE.

Dado	Método	Q	N_c	CR_{tr}	ε_{tr}	CR_{ts}	ε_{ts}	N_Q Podados
	original	9	75	99,05	0,0203	93,33	0,0678	-
Iris	CAPE	6	20	99,05	0,2969	93,33	0,3144	4,5,9
	PCAD	6	51	99,05	0,0303	93,33	0,0751	4,5,7
	PCA	6	51	99,05	0,0303	93,33	0,0751	4,5,7
	original	9	75	99,05	0,0127	93,33	0,0858	-
Iris	CAPE	4	16	99,05	0.1765	93,33	0,2422	1,2,4,6,8
	PCAD	2	16	99,05	0,2161	93,33	0,2707	1,2,3,4,5,6,8,b
	PCA	2	19	99,05	0,0934	93,33	0,1406	1,2,3,4,5,6,8
	original	9	156	100	0,0005	100	0,0005	-
Wine	CAPE	7	37	100	0,2018	100	0,2018	1,9
	PCAD	5	88	100	0,0108	100	0,0108	1,5,6,9
	PCA	7	122	100	0,0133	100	0,0133	6,7
	original	10	416	99,73	0,0030	99,73	0,0030	-
Derm	CAPE	7	85	99,73	0,0193	99,73	0,0193	2,7,9
	PCAD	9	375	99,73	0,0037	99,73	0,0037	7
	PCA	9	375	99,73	0,0037	99,73	0,0037	7
	original	24	243	89,68	0,1132	87,50	0,1274	-
Col.	CAPE*	18	124	92,06	0,1662	86,96	0,1273	1,4,5,9,22
	PCAD	17	173	90,48	0,2267	87,50	0,1679	1,5,9,12,13,20,21
	PCA	20	203	89,68	0,1581	86,96	0,1365	1,4,7,9

adaptado para aplicação de poda em MLPs com mais de uma camada escondida e até mesmo para a seleção de atributos.

Simulações usando dados reais indicaram que, em termos de desempenho, o método de poda baseado em PCA aplicada à rede dual tem desempenho comparável ou melhor que técnicas de poda convencionais. Em termos de aplicações embarcadas, o PCAD pode ter um maior impacto no custo computacional da rede podada em relação aos métodos de poda de conexões sinápticas, pois a poda de neurônios ocultos efetivamente reduz a dimensão das matrizes de pesos, reduzindo o número de operações matemáticas executadas.

Agradecimentos: Os autores agradecem à CAPES/PRODOC pelo apoio financeiro.

Referências

Berthonnaud, E., Dimnet, J., Roussouly, P. & Labelle, H. (2005). Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters, *Journal of Spinal Disorders & Techniques* **18**(1): 40–47.

Chen, Y., Hu, S. & Chen, D. (2001). Fast pruning strategy for neural network size optimization and its applications, *Journal of Chemical Industry and Engineering* pp. 522–526.

Curry, B. & Morgan, P. H. (2006). Model selection in neural networks: Some dificulties, *European Journal of Operational Research* **170**(2): 567–577.

Fahlman, S. E. & Lebiere, C. (1990). The cascade-correlation learning architecture, *in* D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems*, Vol. 2, Morgan Kaufmann, San Mateo, pp. 524–532.

Henrique, H. M., Lima, E. L. & Seborg, D. E. (2000). Model structure determination in neural network models, *Chemical Engineering Science* **55**(22): 5457–5469.

Medeiros, C. M. & Barreto, G. A. (2007). An efficient method for pruning the multilayer perceptron based on the correlation of errors, *in* J. M. de Sá, L. A. Alexandre, W. Duch & D. Mandic (eds), *Artificial Neural Networks – ICANN 2007*, Vol. 4668 of *LNCS*, Springer, pp. 219–228.

Nakamura, T., Judd, K., Mees, A. I. & Small, M. (2006). A comparative study of information criteria for model selection, *International Journal of Bifurcation and Chaos* **16**(8): 2153–2175.

Principe, J. C., Euliano, N. R. & Lefebvre, W. C. (2000). *Neural and Adaptive Systems*, John Wiley & Sons.

Reed, R. (1993). Pruning algorithms - a survey, *IEEE Transactions on Neural Networks* **4**(5): 740–747.

Seghouane, A.-K. & Amari, S.-I. (2007). The AIC criterion and symmetrizing the kullback-leibler divergence, *IEEE Transactions on Neural Networks* **18**(1): 97–106.

Xiang, C., Ding, S. Q. & Lee, T. H. (2005). Geometric interpretation and architecture selection of the MLP, *IEEE Transactions on Neural Networks* **16**(1): 84–96.