

# Redes Neurais Artificiais - Exercício 1

Victor São Paulo Ruela  
Programa de Pós-Graduação em Engenharia Elétrica  
Universidade Federal de Minas Gerais  
Belo Horizonte, Brasil  
Email: victorspruela@gmail.com

## Resumo

Neste exercício é feito um resumo e análise crítica do artigo “Improving generalization of MLPs with multi-objective optimization” [1]. Embora bastante curto, o trabalho foi um dos primeiros a abordar o problema de generalização no treinamento de redes neurais de forma multi-objetiva e consegue apresentar de forma simples e direta os principais conceitos envolvidos.

## I. RESUMO E ANÁLISE DO ARTIGO

Em geral, algoritmos de aprendizado supervisionado de redes neurais artificiais (RNA) possuem como objetivo minimizar o erro quadrático dos valores previstos pelo modelo em relação às saídas em estudo:

$$\sum_{i=1}^N [y_i - f(\mathbf{x}_i)]^2 \quad (1)$$

onde  $y_i$  é uma resposta desejada para uma entrada  $\mathbf{x}_i$ , e  $f$  é a função que aproxima a resposta desejada. Ou seja, estamos interessados em encontrar o conjunto de pesos  $\mathbf{w}$  da rede a partir dos pares de dados de entrada-saída  $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$  que melhor aproxima a função desconhecida  $f$ .

Entretanto, se os dados a serem modelados são ruidosos o uso deste único objetivo pode levar a um overfitting sobre o conjunto de dados de treinamento, de forma que este não consiga generalizar bem para novos valores observados. Estatisticamente, podemos definir a efetividade de  $f$  como um estimador de  $y$  como [2]:

$$E[(y - f(\mathbf{x}; \mathcal{D}))^2 | \mathbf{x}, \mathcal{D}] = E[(y - E[y|\mathbf{x}])^2 | \mathbf{x}, \mathcal{D}] + (f(\mathbf{x}; \mathcal{D}) - E[y|\mathbf{x}])^2 \quad (2)$$

É importante notar neste indicador que o primeiro termo representa a variância de  $y$  dado  $\mathbf{x}$ , não dependendo dos dados. Já o segundo termo mede a distância entre o estimador e a regressão. Logo, podemos definir o erro quadrático médio de  $f$  como um estimador da regressão  $E[y|\mathbf{x}]$  para um conjunto de dados  $\mathcal{D}$  como:

$$\begin{aligned} E_{\mathcal{D}}[(f(\mathbf{x}; \mathcal{D}) - E[y|\mathbf{x}])^2] = \\ (E_{\mathcal{D}}[f(\mathbf{x}; \mathcal{D})] - E[y|\mathbf{x}])^2 \quad \text{“viés”} \\ + E_{\mathcal{D}}[f(\mathbf{x}; \mathcal{D}) - E_{\mathcal{D}}[f(\mathbf{x}; \mathcal{D})]] \quad \text{“variância”} \end{aligned} \quad (3)$$

A derivação completa da relação acima pode ser encontrada em [2]. Logo é fácil notar que o aprendizado de RNAs é um problema multi-objetivo, no qual precisamos encontrar uma solução de compromisso entre o viés e a variância do modelo. Portanto, em um dos extremos teremos um conjunto de pesos que resultam em um viés máximo (*overfitting*) e no outro variância máxima (*underfitting*).

Para lidar com essa situação, o artigo escolhido para este exercício propõe uma nova técnica para melhorar a generalização de perceptrons de múltiplas camadas (MLP) através de uma abordagem por otimização multi-objetivo. Os autores iniciam com uma breve introdução e motivação do trabalho, descrevendo algumas abordagens da literatura já utilizadas: (i) algoritmos de poda de pesos da rede com o intuito de obter um melhor balanço entre flexibilidade e rigidez; (ii) o algoritmo SVM, que realiza uma regularização do problema de minimização de (1) através da inclusão da norma dos pesos da rede na função objetivo. A técnica proposta pelos autores consiste em buscar uma solução com boa generalização dentro do plano formado pelos dois objetivos.

Para a solução de um problema multi-objetivo, o primeiro passo consiste na obtenção do conjunto pareto ótimo, o qual conterá todas as soluções eficientes. Em seguida, é necessária uma estratégia de decisão para escolher a solução mais apropriada deste conjunto pareto ótimo. Para a estimativa deste conjunto, foi escolhida uma variação da abordagem  $\epsilon$ -restrito proposta por um dos autores em outro trabalho, a qual é mais eficiente pelo fato de evitar a geração de solução ineficazes. Neste ponto, vale a pena citar que o nome do artigo indicado nas referências não condiz com o presente na literatura, o que dificultou um pouco encontrá-lo para a leitura. Além disso, o leitor que deseja se aprofundar mais nesta abordagem pode se sentir um pouco confuso, já que a técnica desenvolvida é aplicada sobre um problema bem específico da área de controle.

A técnica utilizada aparenta ser bem eficiente, intuitiva e simples de entender, então sua escolha foi adequada para o problema em questão. De forma geral, ela contrói vetores dentro do cone de soluções viáveis a partir da combinação convexa

dos ótimos de cada objetivo em relação à solução utópica. Em seguida, o problema multi-objetivo é transformado em um mono-objetivo, para o qual minimizamos uma variável auxiliar que irá levar a uma solução bem balanceada na fronteira pareto ótima. Entretanto, algo que não ficou claro foi a estratégia de decisão para a escolha desta solução. Durante a descrição da técnica, os autores citam que o fator  $\gamma$  é variado no intervalo  $[0, 1]$ , porém após a definição do problema de otimização (Equações 7 e 8 de [1]) não fica muito explícito qual das soluções do conjunto pareto foi selecionada.

Em seguida, os autores apresentam uma comparação do algoritmo proposto em relação ao MPL treinado com *backpropagation* e o SVM, considerando tanto um problema de classificação quanto um de regressão. Os resultados mostraram que a técnica proposta é capaz de obter soluções suaves para ambos os problemas em relação ao *backpropagation*. Além disso, soluções com erros de predição similares foram obtidas se comparada ao SVM. Vale a pena notar que não foram feitos testes estatísticos para as comparações realizadas, de forma que, embora graficamente os resultados sugere a eficiência da técnica, não pode-se afirmar que um método é melhor em relação ao outro sem o desenho de um experimento com posterior realização de testes de hipótese.

Embora tenha se mostrado eficiente, a técnica proposta tem um grande custo computacional associado, uma vez que é necessário estimar muitos pontos do conjunto pareto ótimo. Apesar do problema de otimização mono-objetivo ter sido resolvido com o algoritmo elipsoidal, que é um método bem poderoso, se estivermos lidando com um volume de dados de treinamento muito grande, estimar este conjunto pareto pode ser tornar proibitivo computacionalmente. Infelizmente, o número de pontos amostrados e o esforço computacional dos estudos de caso não foram informados pelos autores. Isto constitui um fator importante ao se realizar uma comparação com demais métodos da literatura.

Conforme observado pelos autores, os resultados obtidos pelo SVM e a nova abordagem são bem similares, justamente devido ao uso da regularização na função objetivo. Se formos um pouco mais além, o SVM atribui um peso ajustável à norma dos pesos, o que no fundo consiste em uma técnica clássica de escalarização chamada  $\lambda$ -restrito. Logo, o ajuste deste hiper-parâmetro nada mais é que uma amostragem do conjunto pareto ótimo para este algoritmo. Ou seja, é algo válido pensar em estender a técnica multi-objetivo proposta para o ajuste deste hiper-parâmetro do SVM, por exemplo.

#### REFERÊNCIAS

- [1] de Albuquerque Teixeira, R., Braga, A. P., Takahashi, R. H., & Saldanha, R. R. (2000). Improving generalization of MLPs with multi-objective optimization. *Neurocomputing*, 35(1-4), 189-194.
- [2] Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1), 1-58.