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

A INCORPORAÇÃO DO CONHECIMENTO PRÉVIO NA ESTRATÉGIA DE DECISÃO DO APRENDIZADO MULTI-OBJETIVO

Talles H. Medeiros*, Antônio P. Braga†, Ricardo H. C. Takahashi†

* *Universidade Federal de Ouro Preto*
Departamento de Ciências Exatas e Aplicadas
João Monlevade, Minas Gerais, Brasil

† *Universidade Federal de Minas Gerais*
Centro de Pesquisa e Desenvolvimento em Engenharia Elétrica
Belo Horizonte, Minas Gerais, Brasil

Emails: talles.medeiros@decea.ufop.br, apbraga@cpdee.ufmg.br, taka@mat.ufmg.br

Abstract— This article presents the results of applying a new approach to decision making on the multi-objective classification. The objective of the strategy-making is to select, within the Pareto-optimal solutions, the learning model to minimize the structural risk (the generalization error). In this new approach has been incorporated a prior knowledge that, when available, allows the choice of a most efficient model. The prior knowledge used in decision making reduces the set of solutions in the space of objectives indicating a region of the Pareto composed of appropriate models to solve the problem. The new strategy was assessed in a classification problem with synthetic data, allowing us to conduct an analysis of efficiency in the decision-making in multi-objective learning.

Keywords— machine learning, multi-objective optimization, decision making, classification.

Resumo— Este artigo apresenta os resultados da aplicação de uma nova abordagem para tomada de decisão em problemas de classificação multi-objetivo. O objetivo da estratégia de decisão é selecionar, dentro do conjunto Pareto-ótimo de soluções, o modelo de aprendizado que minimize o risco estrutural(o erro de generalização). Nessa nova abordagem foi incorporado um conhecimento prévio que, quando disponível, permite a escolha de um modelo mais eficiente. O conhecimento prévio usado no decisor reduz o conjunto de soluções no espaço dos objetivos indicando uma região do Pareto composta por modelos adequados à solução do problema. A nova estratégia foi avaliada em um problema de classificação com dados sintéticos, permitindo-nos realizar uma análise da eficiência no contexto da tomada de decisão no aprendizado multi-objetivo.

Keywords— aprendizado de máquina, otimização multi-objetivo, tomada de decisão, classificação.

1 Introdução

Em um problema de aprendizado de máquina, o objetivo principal é encontrar a função desejada em um vasto conjunto de funções, sendo que esta busca é realizada tendo como base um número limitado de exemplos. Em uma das abordagens mais comuns para realizar essa busca, a estatística clássica definiu o problema como o dilema da polarização e da variância (Geman et al., 1992). Os modelos podem ficar demasiadamente complexos para o conjunto de exemplos, sendo, portanto, chamados de super-ajustados. Mas também podem ficar muito simples para modelar a função, sendo, portanto, chamados de sub-ajustados. O problema em questão é obter o modelo que se ajuste adequadamente ao conjunto de exemplos disponível.

O aprendizado de máquina multi-objetivo é uma abordagem para tentar encontrar um compromisso entre a polarização e a variância por meio da minimização de duas funções objetivo, geralmente conflitantes (Jin and Sendhoff, 2008). Normalmente, não é possível minimizar todos os objetivos simultaneamente, porque o ótimo de um dos objetivos raramente é o ótimo dos outros. Então, não existe um ótimo único, mas um conjunto deles, chamado conjunto Pareto-ótimo (PO) (Sawaragi et al., 1985). O conjunto com todas as soluções Pareto-ótimas corresponde aos melhores compromissos entre a medida do erro e a da complexidade do modelo.

Então, a resolução de um problema de aprendizagem sob o ponto de vista multi-objetivo não termina na construção do conjunto PO. É necessário que um decisor escolha um dos modelos desse conjunto. O modelo escolhido deverá mostrar que a função desejada foi melhor aproximada pelo modelo escolhido pelo decisor.

Normalmente, a seleção do modelo é feita por meio da avaliação de uma função de erro para novos exemplos (Teixeira et al., 2000). Porém, em problemas de aprendizado de máquina,geralmente, temos um número limitado de exemplos. Este artigo apresenta como principal contribuição, uma estratégia de decisão para problemas de classificação que incorpore um conhecimento prévio do problema. Uma vez incorporado, esse conhecimento fará com que a decisão não necessite mais de novos exemplos.

Para mostrar a eficiência da nova abordagem de decisão, os resultados da aplicação desse decisor

serão apresentados juntamente com os resultados do aprendizado estatístico de uma Máquina de Vetores Suporte (SVM) e de uma rede neural multi-objetivo com um decisor baseado em erro de validação.

O restante deste artigo está organizado da seguinte forma: na seção 2 é descrito o método multi-objetivo de aprendizagem de máquina, na seção 3 é descrito o processo de decisão no aprendizado multi-objetivo, na seção 4 descreve-se o princípio do decisor com conhecimento prévio, em seguida, na seção 5 são descritos os resultados da aplicação do decisor com conhecimento prévio e, por fim, na seção 6 são descritas as conclusões finais desse trabalho.

2 O Aprendizado Multi-objetivo

Para iniciar esta seção é importante apontar diversos trabalhos que já foram destacados na literatura ao se aplicar a teoria da otimização multi-objetivo no aprendizado de máquina (Jin, 2006). Nesse aspecto, o aprendizado multi-objetivo destaca-se como um método eficaz de minimização do risco estrutural (SRM - *Structural Risk Minimization*) (Vapnik, 1998). Onde o princípio SRM é conhecido por definir um compromisso entre o risco empírico (definido em termos de erro de treinamento) e a complexidade (que pode ser definida pela norma dos parâmetros livres). Portanto, pode-se entender o método de minimização do risco estrutural (erro da generalização) como um problema bi-objetivo que tenta encontrar uma solução de compromisso entre os dois objetivos conflitantes: o erro de treinamento e a norma dos parâmetros.

Nesse aspecto, destaca-se o trabalho de desenvolvimento do método multi-objetivo de treinamento de redes neurais (método MOBJ), proposto por (Teixeira et al., 2000), capaz de gerar um conjunto de modelos Pareto-ótimos. A equação 1 apresenta a formulação bi-objetivo do treinamento da rede neural.

$$\text{Minimizar} \begin{cases} f_1(w) = \sum (y_i - \hat{f}(x_i, w))^2 \\ f_2(w) = ||w|| \end{cases} \quad (1)$$

onde $f_1(.)$ é o funcional de erro, $f_2(.)$ é o funcional de complexidade, w é o vetor de pesos sinápticos, y_i e $\hat{f}(x_i, w)$ são, respectivamente, a saída esperada e a saída obtida correspondente ao i-ésimo exemplo. Estas duas funções objetivo são conflitantes em uma determinada região, que constitui o conjunto PO. Nesse conjunto há uma solução que minimiza o erro de generalização. A etapa de decisão é responsável por realizar a busca por essa solução.

3 O Método de Decisão Multi-objetivo

O problema da regra geral de decisão multi-objetivo deve obedecer o esquema dado pela Equação (2),

$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathcal{W}^*}{\arg \min} \; f_U \quad (2)$$

onde f_U é uma função de utilidade capaz de classificar as soluções Pareto-ótimas segundo uma métrica específica.

A função f_U deverá representar o erro entre a função aproximada do modelo, $f(\mathbf{x}; \mathbf{w})$, obtida com o treinamento multi-objetivo e a função geradora desconhecida, $f_g(\mathbf{x})$, descrita pela Equação (3):

$$\mathcal{E}[f(\mathbf{x}; \mathbf{w})] = E\left[[f_g(\mathbf{x}) - f(\mathbf{x}; \mathbf{w})]^2\right], \quad (3)$$

onde $f(\mathbf{x}; \mathbf{w})$ é a solução do Pareto que minimiza a expectativa de erro.

O primeiro critério de decisão implementado no método MOBJ foi o decisor por mínimo erro de validação (Teixeira et al., 2000). A descrição dessa estratégia de decisão é detalhada na subseção 3.1. Além disso, na subseção 3.2 é apresentado o conceito geral da *Curva L* para seleção do modelo com menor erro de generalização.

3.1 O Critério do Erro de Validação

Esta estratégia de decisão multi-objetivo é baseada em um processo de amostragem de novos exemplos do problema de aprendizado. Este decisor escolhe o modelo correspondente ao de mínino erro de validação exibido na curva discreta de validação. Conforme destacado em (Teixeira et al., 2000), o decisor baseado no critério mínimo erro de validação é capaz de assegurar o princípio SRM. A Figura 1 mostra que a curva de validação apresenta uma região de mínimo erro para uma norma dos pesos em particular. Essa região corresponde à localização da solução PO escolhida.

3.2 O Critério da Curva L

A *Curva L* é um gráfico, em escala logarítmica, da complexidade do modelo *versus* o erro de treinamento correspondente. O uso prático de tal gráfico foi sugerido pela primeira vez em (Lawson and Hanson, 1974). É fato que vários pesquisadores têm utilizado essa abordagem quando procuram tratar problemas inversos mal colocados por meio

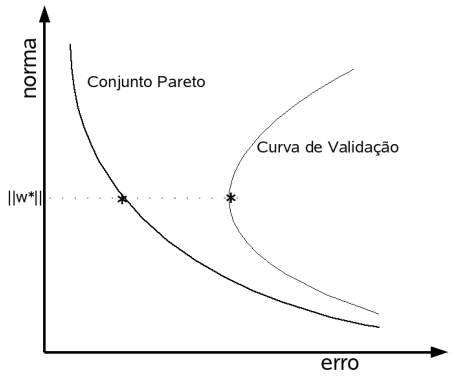

Figura 1: A Curva de Pareto e a Curva de Validação, ambas discretas.

da teoria de regularização (Tikhonov and Arsenin, 1977). Com a *Curva-L* é possível estimar o valor "ótimo" do parâmetro de regularização, tal como outros métodos, como a GCV (*Generalized Cross Validation*) (Golub et al., 1979) e a discrepância de Morozov (Ramlau and Ramlau, 2001).

Em muitos problemas de diferentes aplicações o objetivo da regularização com o critério da *Curva L* é encontrar uma curva, contínua ou discreta (Figura 2), em forma de "L", aproximadamente. Com esta técnica, a solução com o parâmetro de regularização mais adequado e, consequentemente, a de melhor ajuste, é a solução correspondente ao ponto mais próximo à "quina" da curva em forma de "L".

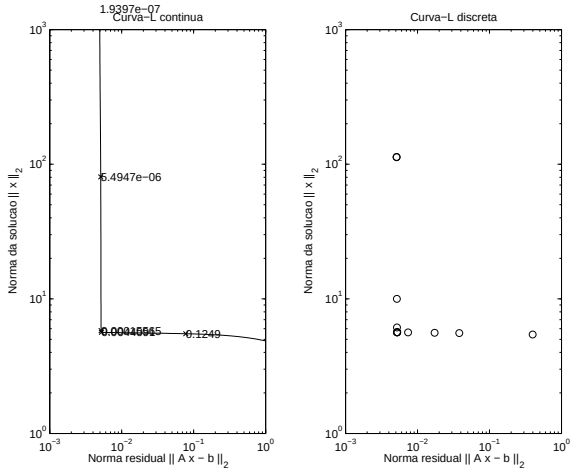

Figura 2: As *Curvas-L* para problemas contínuos e discretos, respectivamente.

A região de maior curvatura da *Curva L* corresponde a região das soluções que seguem o princípio SRM. A *Curva L* apresenta uma semelhança bem evidente da curva Pareto-ótima, no qual ambas representam o dilema entre o erro de treinamento e a complexidade do modelo. Toda via, o critério da *Curva L* foi proposto somente para a formulação ponderada do problema bi-objetivo de aprendizado. Isso implica que, no aspecto multi-ojetivo, a *Curva L* torna-se aplicável somente quando os funcionais envolvidos são convexos.

4 Decisor com Conhecimento Prévio

O decisor por mínimo erro de validação é capaz de obter respostas satisfatórias quando temos exemplos suficientes para decisão. Porém, esta é uma consideração que nem sempre pode ser garantida e pode levar a decisões por modelos inadequados.

A idéia principal deste novo decisor para classificação é um teste estatístico que quantifique a probabilidade de um modelo de classificação ser o melhor, comparado com os outros do conjunto Pareto-ótimo, com base em algum conhecimento prévio sobre a distribuição do erro.

O exemplo mais simples desta idéia é o caso em que existe uma probabilidade fixa do erro de classificação, não importando a posição do exemplo no espaço de características. Supondo que existem duas classes, A e B, e supondo que é conhecido que o gerador de dados produz um erro descrito por:

- Um exemplo que pertence à classe A possui uma probabilidade p_1 de ser rotulado como pertencente à classe B, e um exemplo que pertence a classe B possui uma probabilidade p_2 de ser rotulado como pertencente a classe A.

Neste caso, o número de classificacões errôneas em cada classe segue uma *distribuição binomial*.

Portanto, o erro presente nos exemplos pode ser visto como uma variável aleatória que segue a distribuição binomial. Assim, usa-se a função binomial para estimar a probabilidade dos modelos PO em cada classe, conforme mostra a equação 4:

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \cdots, n, \tag{4}$$

onde n é o número de modelos do conjunto PO, k é o número de exemplos de uma classe que o modelo classificou incorretamente no treinamento e p o a informaçao prévia da probabilidade de classificações incorretas nos exemplos fornecidos para o aprendizado.

Assim, o decisor gera uma distribuição binomial da probabilidade de cada modelo do conjunto PO representar os exemplos da classe em questão. A partir disso, é gerada uma distribuição para cada classe, onde o modelo mais provável é escolhido.

Em algumas situações, um modelo único pode ser escolhido, mas em outros, mais modelos pode ser considerados similares.

A seguir, um problema de classificação irá exemplificar o princípio desse proceso de decisão e sua eficiência no uso racional do conjunto de exemplos disponível.

5 Resultados

Para apresentar e discutir os resultados da simulação realizada com o método MOBJ com conhecimento prévio foi escolhido um problema de classificação de padrões não linearmente separáveis e com sobreposição entre as classes para avaliar a eficiência da estratégia de decisão.

5.1 Problema das Duas Luas

Neste problema de classificação, é utilizado o método MOBJ em uma rede neural para mostrar o resultado do processo de decisão baseado na incorporação do conhecimento prévio dos erros inseridos nos rótulos das classes.

O conjunto de dados possui 400 exemplos divididos, igualmente, em duas classes. Os resultados obtidos são confrontados com uma SVM (com kernel polinomial) treinada com mesmo conjunto de dados. Além disso, a solução do método MOBJ com validação (70% dos exemplos para treinamento e 30% dos exemplos para validação) também é apresentada. O conjunto PO de soluções foi composto por 31 redes *Multilayer Perceptron* com 10 neurônios na camada oculta. A função de transferência usada nos neurônios da camada oculta e de saída é a tangente hiperbólica. A diferença de norma entre os modelos gerados é de $\Delta||w|| = 0.5$. O algoritmo de treinamento é o Levemberg-Marquardt com restrição de norma dos pesos (Costa et al., 2002). O decisor baseado no conhecimento prévio usa o valor $p = 0.1$ para informar que existem, cerca de, 10% de exemplos rotulados incorretamente em cada classe. Por meio dessa informação, o decisor descarta o uso de exemplos para validação e permite que o treinamento use todos os exemplos.

A Figura 3 mostra as curvas de separação que definem a região limite entre as classes. As soluções encontradas pelo método MOBJ para ambos os decisores ficaram muito próximas entre si, enquanto a solução da SVM é ligeiramente mais complexa. Dada a não-linearidade das classes e, para complicar, a sobreposição gerada pelo ruído, todas as abordagens garantiram o controle da complexidade do modelo.

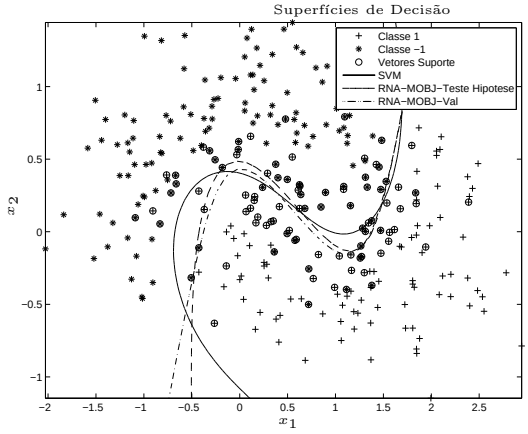

Figura 3: As curvas de separação obtidas pelo método MOBJ com validação, com o conhecimento prévio $p = 0.1$ e da SVM.

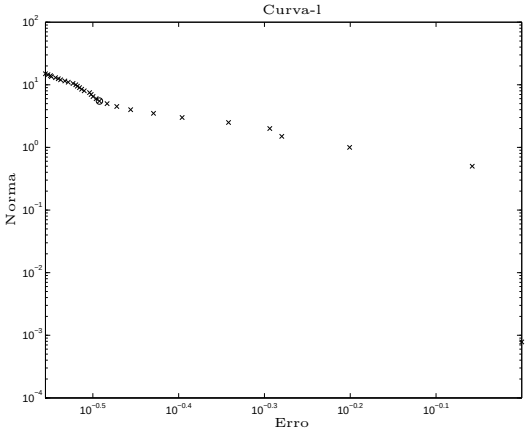

Figura 4: A *Curva L* gerada com dados do treinamento sobrepostos e com decisor por teste de hipótese com $p = 0.1$.

A Figura 4 exibe o princípio da *Curva L* para as soluções PO obtidas e destaca, com um círculo (o), a solução escolhida pelo critério do decisor com conhecimento prévio. Conforme esperado, a *Curva L* não destaca, claramente, a região de máxima curvatura. O critério da *Curva L* é usado para definir o parâmetro de regularização na formulação ponderada de um problema multi-objetivo. Porém, a convexidade dos funcionais objetivos é uma condição necessária nesse critério, caso contrário, a *Curva L* pode não ter a região de máxima curvatura bem evidente, conforme a Figura 4. A solução do método MOBJ com validação não foi indicada na Figura 4 porque seu treinamento foi realizado com um conjunto de exemplos diferente (70% dos exemplos), constituindo assim, um outro conjunto PO. A solução da SVM também não foi indicada porque sua medida de complexidade difere da medida usada na rede neural.

As Figuras 5 e 6 exibem a distribuições binomiais das classes do problema das Duas Luas. Em

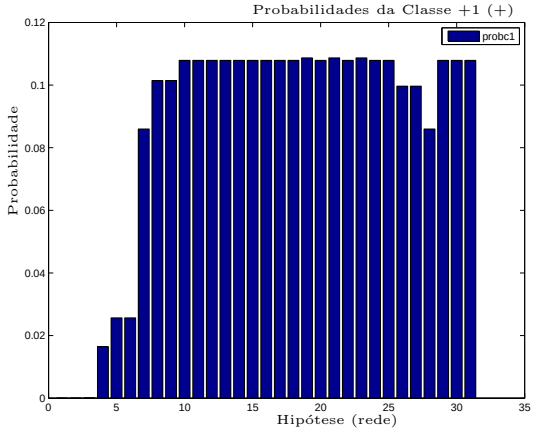

Figura 5: Distribuição da classe (+), com $p = 0.1$.

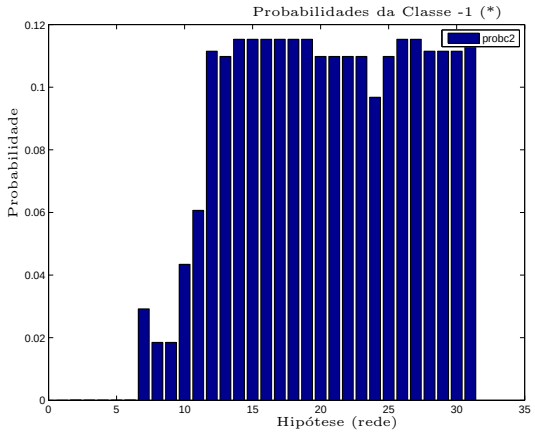

Figura 6: Distribuição da classe (*), com $p = 0.1$.

cada figura tem-se a probabilidade do modelo PO representar o conjunto de exemplos de acordo com o conhecimento prévio inserido. A Figura 5 indica a solução 19 como a mais provável para a classe +, enquanto a Figura 6 indica a solução 14 como a mais provável para a classe *. Portanto, as soluções intermediárias do intervalo [14, 19] constituem um conjunto de soluções PO aceitáveis como solução final para decisão. A curva de separação indicada na Figura 3 corresponde à solução 16 do Pareto, que está localizada no interior desse intervalo.

6 Conclusões

Esses resultados destacam a eficiência do uso do conhecimento prévio na tomada de decisão em um conjunto PO de modelos de máquinas de aprendizagem. Tal eficiência torna-se mais evidente à medida que os exemplos não precisam ser reservados para a tomada de decisão, permitindo, assim, que o algoritmo de treinamento use o conjunto completo de exemplos disponível.

A construção da distribuição de probabilidade binomial permite caracterizar não somente uma solução mas, uma região de soluções equivalentes de acordo com o valor p inserido pelo especialista que detém o conhecimento prévio sobre o problema e seus exemplos disponíveis.

Essa forma de abordagem na etapa de decisão do aprendizado multi-objetivo traz benefícios importantes quando há uma informação prévia do problema que pode, então, ser incorporada no decisor e garantir que o modelo escolhido segue princípio da minimização do risco estrutural.

Referências

Costa, M. A., Braga, A. P. and Menezes, B. R. (2002). Improved generalization learning with sliding mode control and the levenberg-marquadt algorithm, *VII Brazilian Symposium on Neural Networks*.

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

Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics* **21**: 215–223.

Jin, Y. (ed.) (2006). *Multi-objetive Machine Learning*, 1 edn, Springer.

Jin, Y. and Sendhoff, B. (2008). Pareto-based multiobjective machine learning: An overview and case studies, *IEEE Transactions on Systems, Man, and Cybernetics, Part C* **38**(3): 397–415.

Lawson, C. L. and Hanson, J. R. (1974). *Solving Least Square Problems*, Prentice Hall, Englewoods Clifs, NJ.

Ramlau, R. and Ramlau, R. (2001). Morozov's discrepancy principle for tikhonov regularization of nonlinear operators, *Numer. Funct. Anal. and Optimiz* **23**: 2002.

Sawaragi, Y., Nakayama, H. and Tanino, T. (1985). *Theory of Multiobjective Optimization*, Academic Press.

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

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill Posed Problems*, Vh Winston.

Vapnik, V. N. (1998). *Statistical Learning Theory*, John Wiley & Sons, Inc., New York, NY. 736 pages.