Aprendizado de Máquina Aplicado ao Estudo de Marcadores Moleculares para Produção de Carne Bovina

Silvia H. M. G. da Silva[1], Ana C. Lorena[1], André C. P. L. F. de Carvalho[1], Danielle D. Tambasco[2] and Luciana C. A. Regitano[2]

[1] Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação,
Cx. Postal 668, CEP 13560-970,
São Carlos - São Paulo - Brasil
{silviah, aclorena, andre}@icmc.sc.usp.br
[2] Embrapa - Centro de Pesquisas de Pecuária do Sudeste,
Cx. Postal 339, CEP 13560-970,
São Carlos - São Paulo - Brasil
{danielle, luciana}@cppse.embrapa.br

Resumo Existem diversos fatores relacionados à produção de carne bovina. Identificá-los contribui de maneira direta para a escolha de cruzamentos genéticos mais promissores. Este trabalho investiga o uso de Redes Neurais Artificiais e Máquinas de Vetores Suporte no estudo da influência de marcadores moleculares no ganho de peso diário do animal, do nascimento à desmama. As taxas de erros obtidas se mostraram baixas, evidenciando que as variáveis envolvidas estão relacionadas com o ganho de peso nesta fase.

1 Introdução

A geração de tecnologias e conhecimento que contribuam para o aumento da disponibilidade de produtos de origem animal de melhor qualidade e menor custo é de fundamental importância para o progresso da pecuária. O uso de marcadores moleculares, principalmente de DNA, permite que o potencial genético de um animal seja determinado com maior precisão [4].

Essas pesquisas geram uma enorme quantidade de dados que necessitam de ferramentas computacionais que facilitem sua compreensão. Técnicas de Aprendizado de Máquina (AM), por serem capazes de aprender por si a partir de um conjunto de dados, representam uma alternativa atraente para lidar com este tipo de problema. Este trabalho busca investigar o uso duas técnicas inteligentes, Redes Neurais Artificiais (RNAs) e Máquinas de Vetores Suporte *(Support Vector Machines - SVMs)*, para interpretação de dados sobre marcadores moleculares.

As RNAs podem ser definidas como sistemas paralelos distribuídos compostos de unidades de processamento simples que computam determinadas funções matemáticas, simulando os neurônios [2]. As SVMs utilizam funções denominadas *Kernels* para mapear os dados para um espaço de grande dimensão. Nesse

espaço, a utilização de uma função linear é suficiente para aproximação da distribuição dos dados [7].

Este artigo encontra-se organizado como segue: a Seção 2 apresenta os materiais e métodos utilizados nos experimentos conduzidos. A Seção 3 lista e discute os resultados obtidos e finaliza este artigo.

2 Materiais e métodos

A base de dados utilizada neste trabalho inclui dados obtidos através do projeto "Marcadores moleculares aplicados à produção de carne bovina", desenvolvido no CPPSE - Embrapa. Foram obtidos dados dos marcadores moleculares LGB (encontrado no gene β-lactoglobulina) e GH (gene de hormônio de crescimento).

Os dados são de 189 animais resultantes do cruzamento de fêmeas Nelore com touros Aberdeen Angus, Canchim e Simental. Os atributos utilizados como entrada para o treinamento das técnicas de aprendizado foram o sexo do animal, grupo genético, tratamento (com ou sem ração), pai, idade da mãe ao parto e combinações dos marcadores (GH e LGB, GH, LGB e sem marcador). A saída, por sua vez, foi o ganho de peso médio.

O trabalho investigou a influência dos marcadores no ganho de peso diário dos animais, no período entre o nascimento e a desmama. Os resultados foram obtidos utilizando *10-fold cross validation* [6]. Foram utilizadas RNAs Perceptron multicamadas, treinadas com o algoritmo *backpropagation* [2]. Foram testadas redes com uma camada intermediária, variando o número de neurônios. A taxa de aprendizado e o termo momentum adotados foram ambos iguais a 0.1. Para implementação das RNAs, utilizou-se a ferramenta SNNS [8]. Para as SVMs, empregou-se diversas funções Kernel (Polinomiais, Gaussianas e Sigmoidal). As SVMs foram geradas com o auxílio da ferramenta SVMTorch II [3].

3 Resultados, Discussões e Conclusão

A Tabela 1 apresenta as taxas de erro obtidos nos testes das técnicas de aprendizado consideradas. A RNA com 1 neurônio na camada intermediária foi a melhor topologia de rede em todos experimentos. As SVMs com Kernel Gaussiano e desvio padrão de 50 apresentaram os melhores resultados nos experimentos com ambos os marcadores e somente com o marcador GH. Para os experimentos com LGB e sem marcadores, o melhor Kernel foi o Gaussiano com desvio padrão igual a 100.

Comparando-se o desempenho das técnicas utilizadas, pode-se observar que as SVMs foram mais precisas, embora não se possa afirmar a um nível de confiança de 95% [5].

Verifica-se que a menor taxa de erro foi obtida no experimento realizado utilizando somente o marcador GH. Do ponto de vista fisiológico, pode-se dizer que isto se deve ao fato do marcador GH estar relacionado ao gene que codifica o hormônio de crescimento.

Tabela 1. Erro quadrático médio obtido nos experimentos

	GH + LGB	GH	LGB	sem marcador
RNAs	11.67 ± 4.02	11.08 ± 3.39	11.69 ± 4.21	11.26 ± 4.07
SVMs	9.95 ± 3.16	9.83 ± 3.13	10.15 ± 3.13	10.07 ± 3.28

O experimento com LGB apresentou a maior taxa de erro, evidenciando que este marcador não influi no ganho de peso diário. Quando associado ao GH, essa taxa diminui. Porém, o erro obtido é maior que o da análise isolada do GH, sugerindo um efeito epistásico da β-lactoglobulina.

Na ausência de marcadores, observa-se uma taxa de erro semelhante às anteriores. Isto indica que existem outras variáveis influenciando fortemente o ganho de peso diário, sugerindo a realização de outros testes experimentais, bem como o uso de outros marcadores moleculares. Uma outra abordagem é analisar quais grupos genéticos estão mais relacionados com as características de produção.

Agradecimentos

Os autores agradecem ao CNPq e à Fapesp pelo apoio financeiro concedido para a realização deste projeto.

Referências

1. Baldi, P., Brunak, S.: Bioinformatics - The Machine Learning Approach. The MIT Press (1998)
2. Braga, A.P., Carvalho, A. C. P. L. F., Ludermir T. B.: Redes Neurais Artificiais: Teoria e Aplicações. Livro Técnico e Científico, Rio de Janeiro (2000)
3. Collobert, R., Bengio, S.: SVMTorch: Support vector machines for large scale regression problems. Journal of Machine Learning Research, Vol. 1 (2001) 143–160
4. Coutinho, L.L., Regitano, L.C.A.: Uso de marcadores moleculares na indústria animal. Biologia Molecular aplicada à produção animal, Embrapa Informação Tecnológica (2001) 215 p.
5. Johnson, R. A.: Miller and Freund's Probability and Statistics for engineers. Prentice Hall (2000)
6. Mitchell, T.: Machine Learning. McGraw Hill (1997)
7. Smola, A. and Schölkopf, B.: A tutorial on support vector regression. NeuroCOLT2 Technical Report NC2-TR-1998-030 (1998)
8. Zell, A., Mamier, G., Vogt, M., Mache, N., Hübner, R., Döring, S., Herrmann, K., Soyez, T., Schmalzl, M., Sommer, T., Hatzigeorgiou, A., Posselt, D., Schreiner, T., Kett, B., Clemente, G., Wieland, J.: SNNS - Stuttgart Neural Network Simulator - User Manual, Version 4.1. Technical Report 6/95, Institute for Parallel and Distributed High Performance Systems (IPVR), University of Stuttgart (1995)