Licenciatura em Engenharia Informática

Projeto em Engenharia Informática 2016/2017

Análise de Metodologias de Extração de Caraterísticas em Conjuntos de Dados

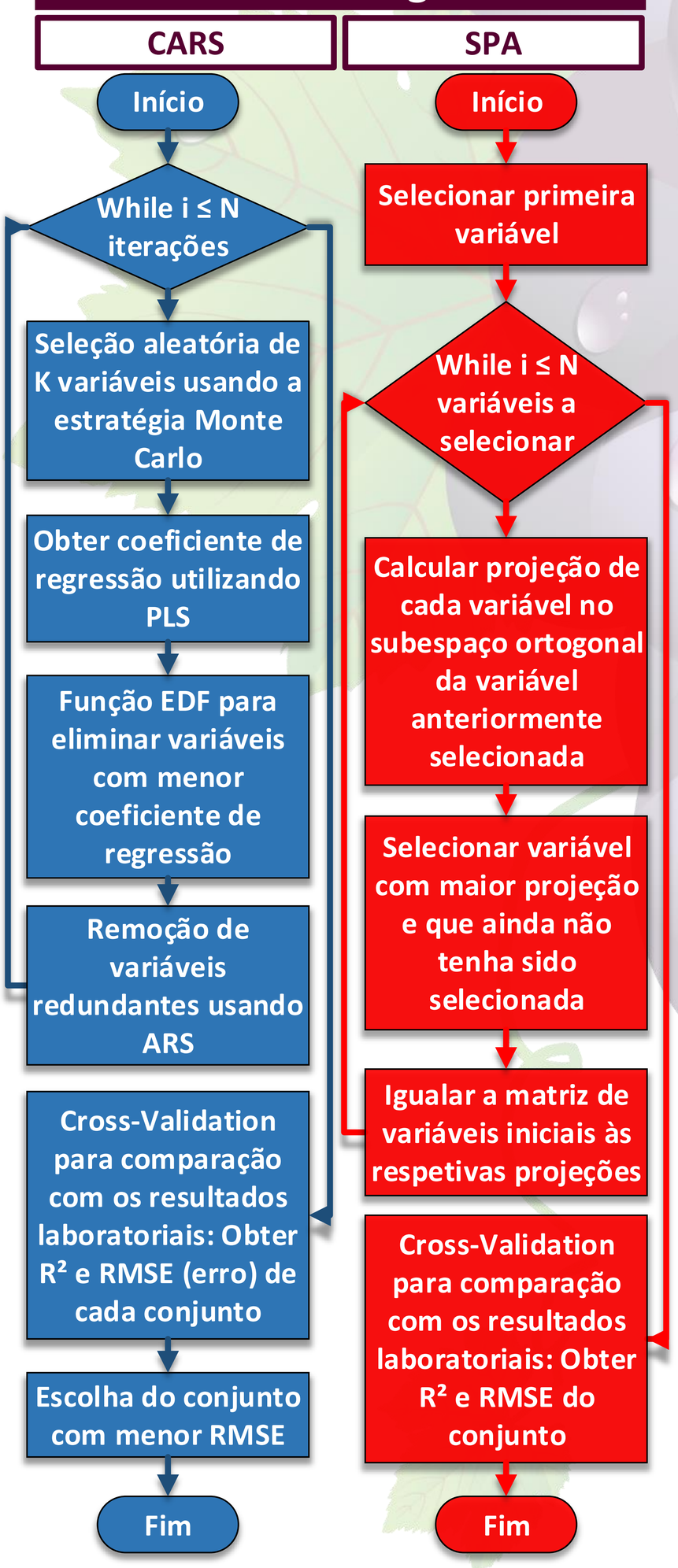
Marcelo Queirós Pinto (nº 60102)

Introdução [1-5]

Na vinicultura, a precisão e automação de processos é cada vez mais importante e, por isso, as tecnologias para apoio à avaliação do estado de amadurecimento da uva representam uma mais valia no mercado competitivo da produção de vinho.

A imagem hiperespectral permite a avaliação de forma não intrusiva de parâmetros enológicos relacionados com o amadurecimento da uva, disponibilizando conjuntos de dados complexos e de grande dimensão que permitirão estimar, com precisão, os referidos parâmetros. No entanto, a abundância de dados gerados obriga a métodos mais robustos para o processamento de processos complexos em tempo real. A solução passa pelo desenvolvimento de metodologias de extração de caraterísticas, permitindo a diminuição da carga computacional, identificando bandas espectrais relevantes para o problema em questão e eliminando as restantes, mantendo a quase totalidade da informação envolvida. Para este fim, foram desenvolvidas duas metodologias de extração de caraterísticas: Competitive Adaptive Reweighted Sampling (CARS) e Successive Projections Algorithm (SPA).

Metodologia



Resultados e Discussão

Foram utilizadas **1040 variáveis** correspondentes aos comprimentos de onda da refletância das amostras de uvas para posterior estimação do Brix. Na Figura 1, apresenta-se o melhor Coeficiente de Pearson (R^2) para cada N calculado pelas metodologias CARS e SPA. Foram feitos 1000 testes do algoritmo CARS e 102960 testes do algoritmo SPA.

O R^2 e o consequente RMSE (root mean square error) obtêm-se a partir da comparação dos valores preditos por um modelo e os valores realmente observados. Os valores reais foram observados a partir de análises laboratoriais de 240 amostras de uvas. O R^2 dá a correlação entre os valores preditos e os valores observados, mas não permite a análise mais fina dos erros envolvidos que é efetuada pelo RMSE. Assim, ambas as metodologias utilizam RMSE como função objetivo do processo de aprendizagem.

O algoritmo CARS selecionou um mínimo de 44 variáveis (N) na totalidade dos seus testes correspondendo a um R^2 de 0,9606. No entanto, **o R^2 máximo foi de 0.9694, correspondendo a N=77 e RMSE=0.5820.** O algoritmo SPA permite a seleção de qualquer N pré-definido, tendo sido testado para N=2 até N=100, **obtendo-se um R^2 máximo de 0.8922, correspondendo a N=17 e RMSE=1.0920.**

Comparando ambos os algoritmos, o CARS permitiu seleção de variáveis mais representativas da amostra, uma vez que todos os testes originam um maior R^2 do que os testes realizados pelo algoritmo SPA. No entanto, o algoritmo SPA consegue selecionar menos variáveis, possibilitando uma maior redução da carga computacional, contendo como melhor R^2 apenas 17 variáveis.

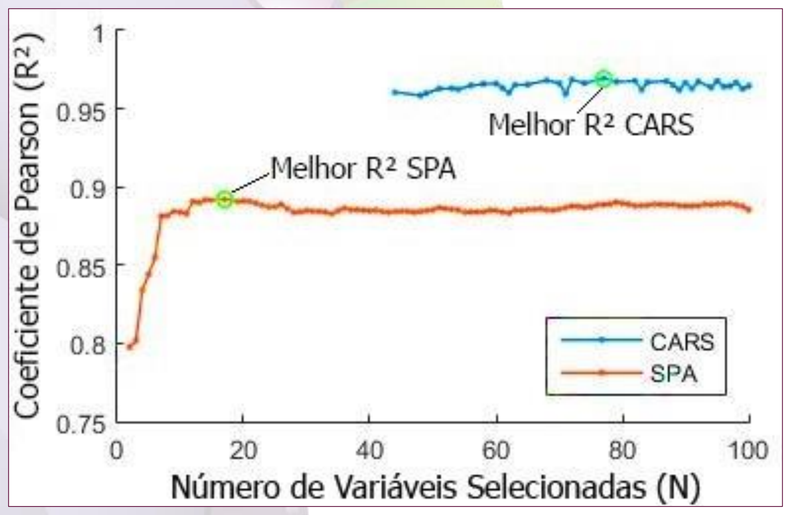


Figura 1: *Melhor R^2 para cada número de variáveis selecionadas (N) pelas metodologias CARS e SPA, correspondente ao parâmetro enológico Brix.*

Conclusão e Trabalho Futuro

Após todos os testes realizados, dos quais os principais resultados se apresentam neste poster, conclui-se que nos casos em que a metodologia CARS não obtenha uma redução dimensional suficiente, a metodologia SPA é uma boa opção. Nos restantes, a metodologia CARS é mais benéfica porque seleciona dados mais relevantes, como mostra o R^2.

Encontra-se em progresso um artigo onde serão analisadas em pormenor ambas as metodologias apresentadas neste poster.

Referências

[1] Tang, G., Huang, Y., Tian, K., Song, X., Yan, H., Hu, J., Min, S. (2014). A new spectral variable selection pattern using competitive adaptive reweighted sampling combined with successive projections algorithm. The Analyst, 139(19), 4894. doi:10.1039/c4an00837e

[2] Araújo, M. C., Saldanha, T. C., Galvão, R. K., Yoneyama, T., Chame, H. C., & Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics and Intelligent Laboratory Systems, 57(2), 65-73. doi:10.1016/s0169-7439(01)00119-8

[3] Goudarzi, N., & Goodarzi, M. (2010). Application of successive projections algorithm (SPA) as a variable selection in a QSPR study to predict the octanol/water partition coefficients (Kow) of some halogenated organic compounds. Analytical Methods, 2(6), 758. doi:10.1039/b9ay00170k

[4] Wu, D., He, Y., Shi, J., & Feng, S. (2009). Exploring Near and Midinfrared Spectroscopy to Predict Trace Iron and Zinc Contents in Powdered Milk. Journal of Agricultural and Food Chemistry, 57(5), 1697-1704. doi:10.1021/jf8030343

[5] Dai, Q., Cheng, J., Sun, D., & Zeng, X. (2014). Advances in Feature Selection Methods for Hyperspectral Image Processing in Food Industry Applications: A Review. Critical Reviews in Food Science and Nutrition, 55(10), 1368-1382. doi:10.1080/10408398.2013.871692

Artigo em Progresso

Equipa de Orientação:
Pedro José de Melo Teixeira Pinto
Véronique Medeiros Gomes

Escola de Ciências e Tecnologia
Universidade de Trás os Montes e Alto Douro