Avaliação de Algoritmos de Seleção de Atributos para dados de Espectroscopia no Infravermelho

Caroline Heloíse de Oliveira – Universidade Estadual de Ponta Grossa – UEPG – carol42.helo@gmail.com
Gabriel Passos de Jesus – Universidade Estadual de Ponta Grossa – UEPG – gabrielpasje@gmail.com

Resumo: A tarefa de seleção de atributos é essencial para garantir a eficiência de um algoritmo de Aprendizado de Máquina. Portanto, objetivou-se avaliar o melhor método para dados de espectroscopia do infravermelho, levando em consideração o tempo de execução e a correlação entre os atributos selecionados e o atributo-meta.

Palavras-chave: Seleção de Atributos; Espectroscopia; Solos Tropicais

1. Introdução

Um dos principais desafios encontrados ao treinar algoritmos de Aprendizado de Máquina está a "maldição da dimensionalidade" (Bellman, 2003), que significa que, conforme a quantidade de atributos da base de dados utilizada aumenta, o desempenho do algoritmo diminui, bem como a complexidade computacional aumenta significativamente. Sendo assim, é imprescindível que se utilize técnicas para redução da dimensionalidade dos dados, mas sem perder a qualidade e mantendo a correlação com o atributo-meta.

2. Materiais e Métodos

Foram testados três algoritmos de seleção de atributos: Recursive Feature Elimination (RFE), Sequential Feature Selection (SFS) e Lasso Regression (LR).

A base utilizada para testar os algoritmos foi obtida a partir do equipamento FOSS NIRS DA1650, onde foram analisadas 266 amostras de solo e coletados os dados referentes à absorbância das amostras na faixa de comprimentos de onda de 1100 nm até 1648 nm. Os dados foram convertidos para reflectância, a fim de padronização para comparação com resultados da literatura. O atributo-meta utilizado foi o teor de carbono (TOC) das amostras, obtido através de métodos tradicionais laboratoriais (PAVAN et al., 1992; van RAIJ et al., 2001).

3. Resultados

O RFE finalizou a execução em 37m19s, retornando os atributos referentes aos comprimentos de onda 1342 nm, 1344 nm, 1346 nm, 1352 nm, 1354 nm, 1356 nm, 1358 nm e 1360 nm. O SFS finalizou em 16,038s, retornando 1100 nm, 1102 nm, 1106 nm, 1346 nm e 1522 nm. O LR finalizou em 0,004s, retornando 100 nm, 1522 nm, 1646 nm e 1648 nm. A partir de algoritmos de regressão, com os atributos selecionados por cada um dos métodos, foi possível chegar a um coeficiente de correlação de 0.99, 0.92 e 0.94, respectivamente.

4. Conclusão

Conclui-se que, em termos de custo computacional, os algoritmos SFS e LR são mais eficientes, enquanto o RFE pode chegar a maiores correlações, porém com maior custo.

Agradecimentos

Os autores agradecem à CAPES - Comissão de Aperfeiçoamento de Pessoal de Nível Superior - e à Fundação Araucária, pelo apoio financeiro.

Referências

BELLMAN, R. **Dynamic Programming**. Dover Publications, 2003.

PAVAN, M.A. et al. **Manual de análise química do solo e controle de qualidade**. Londrina, Instituto Agronômico do Paraná, 1992.

van RAIJ, B. et al.. Determinação de nitrogênio total em solo. **Análise química para avaliação da fertilidade de solos tropicais**, 2001.