

Uso de técnicas de Mineração de Dados em Espectroscopia de Reflectância em Solos.

Sandro Teixeira*, Alaine Margarete Guimarães

*steixeira@uepg.br

Universidade Estadual de Ponta Grossa. Av. Gen. Carlos Cavalcanti, 4748 Uvaranas, Ponta Grossa - PR, 84030-900, Brasil.

Palavras Chave: *Mineração de Dados, solos, reflectância*

Introdução

Segundo dados do Ministério da Agricultura, Pecuária e Abastecimento (MAPA) o Brasil se consolidará como uma potência agrícola nos próximos dez anos e disputará a liderança na produção de alimentos com os Estados Unidos. De acordo com o estudo, produtos agrícolas de alto consumo interno e que já fazem parte da pauta de exportação brasileira, tendem a ter um aumento de produção, sobretudo devido ao avanço tecnológico. Neste contexto, é imprescindível que a área de pesquisa brasileira relacionada à produção continue avançando. E, para que isto aconteça é preciso que o uso de métodos de manejo, como por exemplo a Agricultura de Precisão (AP), sejam explorados a fundo.

Um dos objetivos básicos da AP é o manejo localizado do solo. A produtividade das culturas é influenciada quando os atributos físicos, químicos e biológicos do solo estão equilibrados e suficientemente disponíveis, de forma a fornecer condições para as plantas expressarem todo seu potencial produtivo. Por meio da análise de solo é possível conhecer sua fertilidade, avaliar o grau de deficiência de nutrientes e determinar as quantidades de corretivos e fertilizantes a serem recomendados com vista à produção.

Para a avaliação das características do solo relacionadas com o seu potencial, assim como a avaliação da fertilidade, as análises químicas e físicas de rotina são métodos convencionalmente usados. Porém, estas análises agredem a natureza pela utilização dos reagentes químicos. Nesse contexto o sensoriamento próximo tem despertado o interesse dos pesquisadores área de solos, já que possui algumas vantagens sobre as análises convencionais. Segundo SHEPHERD E WALSH (2007), é uma das técnicas analíticas mais eficientes e disponíveis do século 21. A utilização da espectroscopia de reflectância na região do visível e no infravermelho próximo (vis-NIRS) pode ser altamente viável na análise de solos, e como tal pode ajudar a identificar teores dos atributos do solo de maneira mais limpa e rápida.

É possível correlacionar os comprimentos de onda adquiridos da reflectância de componentes do solo por meio de análises estatísticas, bem como por meio de técnicas de Mineração de Dados, como no trabalho de PROENÇA (2012).

O objetivo deste trabalho foi comparar algoritmos de classificação de dados implementados no software WEKA¹⁰ (Waikato Environment for Knowledge Analysis) para estimar matéria orgânica (MO) e teor de argila do solo utilizando uma base de dados obtidos pelo método convencional de análise do solo e pelo método vis-NIRS.

Material e Métodos

Para a realização do trabalho foi utilizada uma base de dados coletados em 2011 pela Fundação ABC, empresa que atua no desenvolvimento de pesquisa aplicada à agricultura, localizada em Castro no estado do Paraná. Este levantamento de dados foi realizado na região dos Campos Gerais, utilizando uma área que está localizada no município de Pirai do Sul, no Centro-Sul do estado do Paraná, Brasil. Sua posição geográfica tem como coordenadas 24° 22' 30" S, 50° 04' 00" O. A gleba, com uma extensão de 110 hectares, é composta predominantemente por Latossolos de textura média a argilosa. Foram 111 amostras coletadas sendo uma por hectare na profundidade de 00-20 cm PROENÇA (2012).

A planilha original era constituída de 1064 atributos, correspondendo aos diferentes comprimentos de onda sendo lido pelo equipamento. A base foi dividida em quatro conjuntos de dados, conforme Tabela 1, para que o objetivo do trabalho fosse alcançado ela continha o atributo meta que está diretamente relacionado com o domínio do problema este atributo teria sua classe ou valor prevista nas regras geradas pelo algoritmo. Também continha os atributos precursores cuja função é determinar a classe do atributo meta.

Tabela 1 – Divisão da base de dados utilizada.

Conjunto	Nome do Conjunto	Atributo Meta	Comprimentos de Onda
----------	------------------	---------------	----------------------

1	MO VIS	MO	400 a 700 nm
2	MO NIRS	MO	702 a 2498 nm
3	ArgilaVIS	Teor de Argila	400 a 700 nm
4	Argila NIRS	Teor de Argila	702 a 2498 nm

O Conjunto 1 consistia além do atributo meta MO de mais 150 atributos relativos aos comprimentos de onda da região do espectro visível (VIS). Já o conjunto 2 era composto pelo atributo meta MO e 898 atributos que correspondiam aos comprimentos de onda do espectro referente ao infravermelho próximo (NIRS). Da mesma forma, os Conjuntos 3 e 4 continham além do atributo teor de argila, os comprimento de onda VIS e NIRS respectivamente.

Como auxílio na tarefa de mineração dos dados foi utilizado o WEKA que é um software que contempla uma série de algoritmos de preparação de dados, de aprendizagem de máquina e de validação de resultados.

Procurando reduzir a dimensionalidade dos dados foi aplicado nas quatro bases o algoritmo de filtro de seleção de atributos supervisionados disponível na sessão de pré-processamento do WEKA. Como resultado, foram selecionados determinados atributos de comprimentos de onda que tiveram uma maior correlação com as classes MO e teor de argila.

De posse dos atributos selecionados, para cada base de dados foi aplicada a tarefa de classificação utilizando-se dos algoritmos disponibilizados pela ferramenta. Todos os classificadores utilizados foram processados com seus valores default.

Resultados e Discussão

Depois de aplicado o filtro de seleção de atributos para cada uma das bases utilizadas no estudo, o quadro apresenta:

Tabela 02 – Resultado da aplicação do filtro nas bases utilizadas

Conjunto	Nome do Conjunto	Atributo Meta	Comprimento de Onda
1	MO VIS	MO	480
2	MO NIRS	MO	702,704,860,2270,2498
3	ArgilaVIS	Teor Argila	480
4	Argila NIRS	Teor Argila	702,862,2268,2270,2272,2496,2498

Levando-se em conta os três melhores resultados para o índice de correlação e o erro absoluto relativo para cada uma das bases de dados, as Tabelas 3,4,5 e 6 destacaram os algoritmos de classificação que ...

Tabela 03 – Melhores resultados para o Conjunto de dados Matéria Orgânica VIS

	Meta Bagging	Rules Decision Table	Lazy KStar
Coefficiente de correlação	0.8549	0.857	0.8779
Erro médio absoluto	2.5879	2.5021	2.360
Erro quadrático	3.5359	3.5071	3.250
Erro absoluto relativo	47.3278%	45.758%	43.1688%
Erro relativo de raiz quadrada	51.921%	51.4977%	47.7307%

Tabela 04 – Melhores resultados para o Conjunto de dados Matéria Orgânica NIRS

	Meta Random Subspace	Meta Bagging	Rules Decision Table
Coefficiente de correlação	0.792	0.7978	0.8068
Erro médio absoluto	2.9798	2.949	2.9333
Erro quadrático	4.1775	4.091	4.0148
Erro absoluto relativo	54.4947%	53.9307%	53.6444%
Erro relativo de raiz quadrada	61.343%	60.0721%	58.9528%

Tabela 05 – Melhores resultados para o Conjunto de dados Argila VIS

	Meta Bagging	Rules M5 rules	Lazy KStar
Coefficiente de correlação	0.9457	0.9467	0.982
Erro médio absoluto	24.0756	24.981	24.9271
Erro quadrático	36.3701	36.0666	36.4875
Erro absoluto relativo	26.7981%	27.8059%	27.746%

Erro relativo de raiz quadrada	32.4387%	32.168%	32.5434%
--------------------------------	----------	---------	----------

Tabela 06 – Melhores resultados para o Conjunto de dados Argila NIRS

	Rules M5 rules	MLP	Lazy KStar
Coeficiente de correlação	0.8798	0.8856	0.914
Erro médio absoluto	34.6106	35.6668	32.3151
Erro quadrático	53.2434	53.6988	45.5163
Erro absoluto relativo	38.5245%	39.7001%	35.9694%
Erro relativo de raiz quadrada	47.488 %	47.8942%	40.5962%

Os resultados demonstraram uma melhor resposta de classificação para o teor de argila em comparação com a quantidade de MO. O algoritmo de classificação *Lazy KStar* teve um coeficiente de correlação de 0.982 no Conjunto de dados Argila VIS com o comprimento de onda 480, assim como foi o melhor classificador para o Conjunto de dados Argila NIRS. Já os resultados adquiridos dos classificadores nas bases para MO foram satisfatórios tendo novamente o algoritmo *Lazy KStar* com melhor desempenho apresentando um índice de coeficiente de correlação de 0.8779. A classificação do Conjunto de dados que obteve o pior desempenho foi o MO NIRS tendo como o algoritmo *Decision Table* dentre os três descritos com coeficiente de correlação mais alto 0.8068.

Conclusões

Com relação aos resultados obtidos pode-se observar que para a predição do teor de argila foi obtido um resultado significativo como por exemplo índices de correlação maiores que 0.9. Já para a quantidade de MO os resultados foram satisfatórios. O trabalho demonstrou que o uso da espectroscopia é viável, trazendo vantagens por ser uma técnica rápida e não poluente.

Para confirmar eficiência da técnica aplicada neste trabalho, seria interessante obter outras bases de diferentes regiões e realizar os mesmos testes.

Agradecimentos

A Fundação ABC por ter disponibilizado a planilha de dados.

¹ Shepherd, K.D.; Walsh, M.G. Infrared spectroscopy - enabling an evidence based diagnostic surveillance approach to agricultural and environmental management in developing countries: Journal of Near Infrared Spectroscopy, Charlton, v.15, p.1-19, **2007**.

² Proença, C. A.; Redes Neurais Artificiais para predição dos teores de matéria orgânica e argila do solo na região dos Campos Gerais utilizando Espectroscopia de Reflectância Difusa. Dissertação de Mestrado, Ponta Grossa-PR, UEPG, **2012**.

³ Moreira, F. M. S.; Siqueira, J. O. Microbiologia e bioquímica do solo. 2. ed. Lavras: Editora UFLA. p.729, 2006.

⁴ Poppi R. J.; Sena M. Avaliação do uso de métodos quimiométricos em análise de solos. Departamento de Química Analítica - Instituto de Química - UNICAMP - CP 6154 - 13083-970 - Campinas – SP, **1999**.

⁵ Mcbratney, A.B.; Minasny, B.; Viscarra Rossel, R. Spectral soil analysis and inference systems: a powerful combination for solving the soil data crisis, Geoderma, Amsterdam, v.136. p.272-278, **2006**.

⁶ Souza, D. M., Beata, E. M., Guimarães, F. F. Aplicação de técnicas multivariadas e inteligência artificial na análise de espectros de infravermelho para determinação de matéria orgânica em amostras de solos. Química Nova, Vol. 35, No. 9, 1738-1745, **2012**

⁷ Tan, P.N.; Steinbach, M.; Kumar, V. Introdução ao Data Mining: Mineração de Dados. Rio de Janeiro, RJ, Editora Ciência Moderna, v.1, p.1, **2009**.

⁸ Carvalho, L. A. Data Mining: a mineração de dados no marketing, medicina, economia, engenharia e administração. São Paulo: Érica, **2001**.

⁹ Fayyad, U.M.; Piatetski-Shapiro, G.; Smyth, P.; Uthurusamy, R. Advances in Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, p. 11-34, **1996**.

¹⁰ Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Ian H. Witten; The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, **2009**.