MINERAÇÃO DE DADOS DE REFLETÂNCIA ESPECTRAL DA CULTURA DA SOJA UTILIZANDO O SOFTWARE SAS

Luiz Rodolfo Machado (PIBIC/Cnpq/UEPG), Pablo Henrique Ferreira, Alaine Margarete Guimarães (Orientadora), e-mail: alainemg@uepg.br

Universidade Estadual de Ponta Grossa/Departamento de Informática

Área do conhecimento: 1.03.00.00-7 Ciência da Computação
Subárea: 1.03.03.00-6 Metodologia e Técnicas da Computação

Palavras-chave: Computação Aplicada, Estimativa de teor de nutrientes, Redes Neurais Artificiais

Resumo
A mineração de dados de espectroscopia vem sendo utilizada como alternativa na estimativa dos teores de nutrientes na agricultura. O objetivo desse trabalho foi avaliar a ferramenta SAS On Demand For Academics: Enterprise Miner na mineração de dados de espectroscopia NIR e MIR para estimativa dos teores de nutrientes em grãos de soja. Para a realização do experimento, foram utilizadas 285 amostras de grãos de soja. Todas as amostras foram analisadas em um espectrofotômetro de infravermelho próximo (NIR). Três conjuntos diferentes de atributos foram utilizados. Foram executados todos os métodos disponíveis para a estimativa de teor de cada nutriente da base de dados. A comparação dos resultados foi feita por meio da raiz quadrada do erro quadrático médio (RMSE) normalizada (NRMSE). O algoritmo utilizado na mineração de dados que se destacou foi a rede neural, obtendo NRMSE de 12,49%. A redução da quantidade de atributos melhorou os resultados. A ferramenta SAS On Demand for Academics: Enterprise Miner se mostrou robusta, possuindo algoritmos eficientes.

Introdução

A estimativa do teor de nutrientes em grãos é de grande importância, pois a qualidade do grão é um importante parâmetro tanto para o cultivo quanto em seu valor econômico (FERREIRA et al, 2014).

A análise de dados de espectroscopia vem sendo utilizada como alternativa na estimativa do teor de nutrientes em diversas áreas, inclusive na agricultura. Essa técnica oferece vantagens em relação às metodologias correntes de análise, sendo a principal delas, a possibilidade de analisar amostras sólidas e líquidas sem tratamento prévio (BALABIN; SMIRNOV, 2011).

A mineração de dados consiste na descoberta automática de informação útil em grandes depósitos de dados, sendo uma etapa integral do processo de Descoberta de Conhecimento em Bases de Dados (KDD – do inglês, Knowledge Discovery in Databases). Essa etapa é antecedida pelo pré-processamento, realizado para transformar os dados em um formato apropriado para a mineração (TAN; STEINBACH; KUMAR, 2006).

Embora apresente vantagens, o processo de estimativa de nutrientes por mineração de dados de espectroscopia possui diversos fatores que podem variar e interferir nos resultados. Entre eles estão o espectro, o conjunto de atributos e os

algoritmos de mineração utilizados, os quais permitem que esse processo seja realizado de diversas maneiras diferentes.

A ferramenta SAS On Demand For Academics: Enterprise Miner é uma versão acadêmica gratuita do software SAS para mineração de dados, que possibilita a utilização de diferentes bases de dados e disponibiliza diferentes algoritmos de mineração de dados.

Com isso, o objetivo desse trabalho foi avaliar a ferramenta SAS On Demand For Academics: Enterprise Miner na mineração de dados de espectroscopia NIR e MIR para estimativa do teor de nutriente em grãos de soja.

Materiais e Métodos

Para a mineração, foram utilizados dados de espectroscopia de 285 amostras grãos de soja, em conjunto com os resultados de análises laboratoriais de determinação dos teores dos nutrientes N, P, K, Ca, Mg, S, Cu, Mn, Fe, Zn e B, além do valor da produtividade, fornecidos pelo Laboratório de Fertilidade de Solo da UEPG (Universidade Estadual de Ponta Grossa).

Grãos de Soja

O equipamento utilizado nas análises dentro do espectro de infravermelho próximo (NIR) foi o espectrofotômetro NIRS DA1650 da marca FOSS, com volume amostral de 40ml, sendo gerada uma base de dados que representa a resposta espectral de cada amotra.

Pré-processamento e Mineração de Dados

As etapas de pré-processamento e mineração de dados foram realizadas utilizando o software SAS On Demand For Academics: Enterprise Miner, que é uma versão acadêmica gratuita do módulo de mineração do SAS (Statistical Analysis System).

Para cada um dos nutrientes de cada base de dados gerada, foram executados todos os métodos disponíveis na ferramenta SAS para estimativa de nutrientes por mineração de dados, além da comparação de resultados por meio da raiz quadrada do erro quadrático médio (RMSE).

Foram também definidos diferentes conjuntos de atributos preditores na execução da mineração de dados em cada diagrama. Além da base completa de bandas espectrais NIR (Conjunto 1), foi aplicado o método de seleção de atributos padrão da ferramenta SAS (Conjunto 2) e a seleção de atributos do software SynthesisFS (Conjunto 3).

Foi utilizado para facilitar a comparação, o erro RMSE normalizado em relação à amplitude (NRMSE), que consiste no valor percentual que o erro representa em relação à amplitude dos valores de cada teor de nutriente sendo predito.

Resultados e Discussão

Para os métodos de mineração de dados utilizados, foi calculada a média do erro NRMSE para a base de dados. A Tabela 1 apresenta as médias de erro para cada um dos métodos de mineração de dados, onde observa-se que o menor erro foi obtido pela rede neural.

Tabela 1 – Médias de erro NRMSE para cada método de mineração aplicada na base de dados.

Base de Dados	AutoNeural	DMNeural	LARS	Neural	PLS	Reg
	NRMSE médio (%)					
Grãos – NIR	16,11	14,96	14,61	12,49	13,62	14,37

Tso e Yau (2007) utilizaram o software SAS Enterprise Miner para realizar a comparação de diferentes métodos de mineração de dados na predição de consumo de energia elétrica em Hong Kong, obtendo valores de RMSE de 44,142 para a rede neural "Neural Network", de 44,184 a 52,096 para os métodos de regressão e 44,397 para a árvore de decisão "Decision Tree", em uma das fases de seu trabalho, obtendo também seus melhores resultados com o algoritmo baseado na técnica de redes neurais.

Para analisar o conjunto de atributos utilizado, foi calculada a média do NRMSE de cada conjunto. A tabela 2 apresenta as médias de erro para cada conjunto de atributos.

Tabela 2 – Médias de erro NRMSE para cada conjunto de atributos em cada base de dados.

Base de Dados	Conjunto 1	Conjunto 2	Conjunto 3
	NRMSE médio (%)		
Grãos - NIR	13,97	13,56	15,54

Observa-se que a redução da quantidade de atributos em relação ao conjunto 1, que representa o espectro inteiro, apresentou resultados melhores nos três casos, com exceção do conjunto 3 na base de grãos e espectroscopia NIR. Os resultados são semelhantes aos de Balabin e Smirnov (2011), que fizeram a comparação de métodos de seleção de atributos para a predição de parâmetros de qualidade em biodiesel por meio de espectroscopia NIR, obtendo resultados melhores conforme o número de atributos diminuiu: para o teor de metanol, os resultados extremos foram de RMSE igual a 306 ppm, utilizando 316 atributos e 107 ppm utilizando 122 atributos.

Conclusões

Os resultados da mineração dos dados para a estimativa do teor de nutrientes de grãos de soja foram influenciados pela redução da quantidade de atributos e indicam que a ferramenta SAS On Demand for Academics: Enterprise Miner é robusta nessa estimativa, possuindo algoritmos eficientes, com destaque para as redes neurais artificiais.

Agradecimentos

Ao CNPq pelo apoio financeiro.

Referências

BALABIN, R. M.; SMIRNOV, S. V. Variable selection in near-infrared spectroscopy: Benchmarking of features selection methods on biodiesel data. **Analytica Chimica Acta**, v. 692, p. 63-72, 2011.

FERREIRA, D.S. et al. Comparison and application of near-infrared (NIR) and mid-infrared (MIR) spectroscopy for determination of quality parameters in soybean samples. **Food Control**, [S.l], v.35, p. 227-232, jan. 2014.

KIM, Y. S. Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. **Expert Systems with Applications**, v. 34, p. 1227-1234, 2008.

SANTOS, G. A.; SANTOS, A. P.; KORNDÖRFER, G. H. Sistema por Infravermelho próximo (NIR) para análise de nitrogênio foliar. **Biosci. J.**, Uberlândia, v. 28, n. 1, p. 83-90, mar. 2012.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Pearson Education, 2006.

TSO, G. K. F.; YAU, K. K. W. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. **Energy**, v. 32, n. 1, p. 1761-1768, 2007.