TÉCNICAS DE MINERAÇÃO DE DADOS APLICADAS À PREVISÃO DE PRODUTIVIDADE DE CANA-DE-AÇÚCAR BASEADA EM PARÂMETROS DE FERTILIDADE DE SOLO

PAULO RODRIGUES PELOIA[1]
LUIZ HENRIQUE ANTUNES RODRIGUES[2]
JOSÉ PAULO MOLIN[3]

RESUMO: O objetivo deste trabalho foi elaborar um modelo de previsão da produtividade da cana-de-açúcar usando variáveis edáficas (fatores relacionados às características físicas e químicas do solo), por meio da aplicação de técnicas de mineração de dados. A base de dados possui 2.450 registros, coletados em pontos georreferenciados em regiões produtoras de duas usinas na Colômbia, sendo composta por 15 atributos de fertilidade do solo. As amostras de produtividade (t/ha), atributo meta, foram coletadas por toda a área em estudo. A categorização do atributo meta foi realizada pelo algoritmo EM (*Expectation-Maximization)*, resultando em três grupos (baixa, média e alta). O balanceamento de classes foi feito pelo método de amostragem com reposição (Resample), com igual probabilidade de selecionar registros de qualquer uma das três classes. O método de sub-amostragem selecionado foi a validação cruzada, com 10 subconjuntos. O software utilizado para aplicação das técnicas de mineração de dados foi o Weka versão 3.6.3. A técnica empregada foi a classificação com uso dos algoritmos J48 (árvore de decisão), Random Forest (árvore de decisão), Multilayer Perceptron (redes neurais), SMO (Support Vector Machine) e IBK (k vizinho mais próximo, com k=5). Cada um dos algoritmos foi avaliado com relação à sua capacidade de prever corretamente as classes de produtividade. Foi feita a seleção de atributos pelos seguintes métodos: χ^2, CFS, PCA, Info Gain, Gain Ratio e Wrapper. Os resultados mostram que o algoritmo Random Forest teve a maior precisão (83,5%). Os atributos de maior importância foram P, pH, Ca, Ca/Mg, K/(Ca + Mg)½. O de menor importância foi percentual de areia. Os resultados obtidos mostraram que algoritmos de mineração de dados possuem potencial na previsão de produtividade.

PALAVRAS-CHAVE: agricultura de precisão, árvore de decisão, redes neurais, support vector machine.

DATA MINING TECHNIQUES APPLIED TO PREDICTING YIELD OF SUGARCANE BASED ON PARAMETERS OF SOIL FERTILITY

ABSTRACT: The aim of this study was developing a model for predicting sugarcane yield using soil variables, by applying data mining techniques. The database has 2,450 records, collected from georeferenced points in producing areas of two sugarcane mills in Colombia, consisting in 15 attributes of soil fertility. The samples of yield (t/ha), target attribute, were collected on throughout the study area. The categorization of the target attribute was performed by the EM algorithm (Expectation-Maximization*)*, resulting in three groups (low, medium and high). The classes balancing have been done by sampling with replacement (Resample), with equal probability to select records from any of the three classes. The method of sub-sampling cross-validation was selected with 10 subsets. The software used for

[1] Engenheiro Agrônomo, FEAGRI - UNICAMP, E-mail: paulo.peloia@feagri.unicamp.br
[2] Engenheiro Agrícola, FEAGRI - UNICAMP, E-mail: lique@feagri.unicamp.br
[3] Engenheiro Agrícola, ESALQ - USP, E-mail: jpmolin@usp.br

applying the techniques of data mining was the Weka version 3.6.3. The technique selected was classification with use of the algorithms: J48 (decision tree), Random Forest (decision tree), Multilayer Perceptron (neural networks), SMO (Support Vector Machine) and IBK (k nearest neighbor). Each algorithm was evaluated by ability to correctly predict the productivity classes. It was made the selection of attributes by the following methods: $\chi 2$, CFS, PCA, Information Gain, Gain Ratio and Wrapper. The results show that the algorithm Random Forest had the highest accuracy (83.5%). The attributes of greatest importance were P, pH, Ca, Ca/Mg, K/(Ca + Mg) ½. The minor was the percentage of sand. Data mining algorithms have shown potential in predicting productivity.

KEYWORDS: precision agriculture, decision trees, neural networks, support vector machine.

1. INTRODUÇÃO

O conhecimento e correto manejo da capacidade produtiva dos solos é um fator importante para o sucesso da atividade agrícola. Na cultura da cana-de-açúcar são diversos os trabalhos que relacionam, por meio de métodos estatísticos clássicos, produtividade com fatores edáficos, isto é, aqueles relacionados às características físicas e químicas do solo (DIAS et al., 1999; TERAMOTO, 2003).

A técnica regressão linear múltipla foi utilizada por Beauclair (1991) para relacionar produtividade e teores de nutrientes disponíveis no solo e chegou a resultados que explicam menos de 40% da variação na produtividade. Como constatado por Teramoto (2003), técnicas estatísticas convencionais aplicadas de forma isolada, podem não ser a melhor metodologia na elaboração de modelos ou identificação de fatores de influência na produtividade.

Esta mesma dificuldade de interpretação de informações e transformação em ferramentas para auxílio no correto tratamento da variabilidade espacial da lavoura, vem sendo enfrentada por aqueles que adotam técnicas do conceito de agricultura de precisão, com um agravante, um volume de dados sensivelmente maior, que torna esta atividade ainda mais complexa (SOUZA et al., 2010). Neste sentido, o emprego de técnicas de mineração de dados, em função de sua capacidade de análise de bases de dados extensas, complexas e não triviais, na elaboração de um modelo de previsão de produtividade, pode trazer resultados superiores aos encontrados anteriormente (YANG et al., 2002).

2. OBJETIVO

O objetivo deste trabalho foi elaborar um modelo de previsão da produtividade da cana-de-açúcar usando variáveis edáficas, por meio da aplicação de técnicas de mineração de dados.

3. MATERIAL E MÉTODOS

A base de dados utilizada possui 2.450 registros, sendo composta por atributos de fertilidade do solo amostrados em pontos georreferenciados e pela produtividade de cana-de-açúcar associada a cada ponto. Os dados foram coletados em áreas produtoras pertencentes a duas usinas na Colômbia, representando uma área de aproximadamente 4.900 ha. As amostras de solo foram tiradas a uma profundidade de 0,00 a 0,25 m em malha regular com uma amostra a cada dois hectares. Para cada ponto de amostra de solo um valor de produtividade foi associado, obtido pela média das amostras de produtividades num raio de 15 m.

Os 15 atributos de caracterização da fertilidade do solo (variáveis independentes) e a produtividade de cana (atributo meta) utilizados na elaboração do estudo são apresentados na Tabela 1, juntamente com a estatística descritiva. A participação total de valores faltantes é de 0,7% (razão entre os 250 valores faltantes pelo produto de 2.450 registros por 15 variáveis); em função deste baixo percentual, optou-se por não realizar o seu preenchimento.

Tabela 1. Estatística descritiva dos atributos do solo e produtividade

Atributo	Unidade	Min.	P_{25}	P_{50}	P_{75}	Máx.	Missing
pH (água)		4,4	6,3	6,7	7,1	9,4	4 (0,2%)
Cond. Elét. Sol. Solo	dS/m	0,0	0,0	0,0	0,2	6,4	47 (2,0%)
Matéria Orgânica	%	0,1	1,9	2,4	2,9	10,3	1 (0,0%)
Fósforo	PPM	0,7	8,1	16,5	38,2	776,5	100 (4,3%)
Cálcio	$cmol_c/dm^3$	0,0	1,2	1,8	2,2	5,0	4 (0,2%)
Magnésio	$cmol_c/dm^3$	0,0	0,5	0,9	1,5	3,1	0 (0,0%)
Potássio	$cmol_c/dm^3$	0,05	0,20	0,29	0,44	4,74	25 (1,0%)
Ca/Mg		0,4	1,4	1,9	2,6	8,4	4 (0,2%)
Ca/K		0,3	4,0	5,4	7,4	40,6	28 (1,2%)
Mg/K		0,1	1,9	2,9	4,1	12,0	25 (1,0%)
$K/(Ca + Mg)^{½}$		0,039	0,143	0,185	0,244	2,616	28 (1,2%)
Soma de Bases	$cmol_c/dm^3$	0,7	18,5	28,4	38,6	77,9	68 (2,9%)
Areia	%	0,0	14,0	26,0	39,0	82,0	3 (0,1%)
Silte	%	4,0	26,0	32,0	38,0	84,0	3 (0,1%)
Argila	%	4,0	29,0	37,0	50,5	86,0	3 (0,1%)
Produtividade	t/há	60,4	105,4	119,0	131,0	187,0	0 (0,0%)

A categorização do atributo meta foi realizada pelo algoritmo EM, *Expectation-Maximization*, (DEMPSTER et al., 1977) otimização do algoritmo original *k-means*. O resultado desta etapa foram três grupos (Tabela 2), sendo denominados Baixa, Média e Alta com base em seus respectivos valores médios. O mesmo número de grupos de produtividade de cana foi encontrado por Ferraro et al. (2009) e Souza, et al. (2010), pelas respectivas metodologias *k-means* e média ± desvio-padrão, sendo considerado um valor adequado para a tomada de decisão.

Tabela 2. Estatística descritiva dos grupos de produtividade

Grupo	N. Registros	Min.	P_{25}	P_{50}	P_{75}	Máx.	Média	Desvio-padrão
Baixa	613	60,4	86,7	95,0a	100,7	105,3	92,6	9,95
Média	1.582	105,4	115,1	122,4b	130,4	141,6	122,7	9,49
Alta	255	141,7	144,5	149,1c	156,4	187,0	151,4	8,88

Letras diferentes indicam diferença significativa de medianas submetidas ao Teste de Kruskal-Wallis ($p<0,01$)

O balanceamento de classes foi feito pelo método de amostragem com reposição (Resample), com igual probabilidade de selecionar registros de qualquer uma das três classes. Após o balanceamento, o total de registros se manteve em 2.450, com 869 (35,5%) na classe baixa, 800 (32,7%) média e 781 (31,9%) alta. O método de sub-amostragem selecionado foi a validação cruzada, com 10 subconjuntos.

O software utilizado para aplicação das técnicas de mineração de dados foi o Weka versão 3.6.3. A técnica empregada foi a classificação, com uso dos algoritmos J48 (árvore de decisão), Random Forest (árvore de decisão), Multilayer Perceptron (redes neurais), SMO (Support Vector Machine) e IBK (k vizinho mais próximo, com k=5). Cada um dos algoritmos foi avaliado com relação à sua capacidade de prever corretamente as classes de produtividade. Como forma de melhorar o desempenho dos algoritmos de classificação, foi feita a seleção de atributos pelos seguintes métodos: χ^2 (5% de confiança), CFS, PCA (análise de componentes principais), Info Gain (mérito mínimo 0,01), Gain Ratio (mérito mínimo 0,01) e Wrapper.

4. RESULTADOS E DISCUSSÃO

O desempenho dos diferentes algoritmos de classificação (Tabela 3) mostrou uma diferença na precisão de 96,4% se comparado o modelo de menor desempenho (SMO), que identificou 42,5% corretamente, com relação ao de maior, Random Forest, com 83,5% de taxa de acerto. Os algoritmos SMO e Naive Bayes apresentaram desempenhos semelhantes, seguidos por Multilayer Perceptron, IBK, J48 e Random Forest, todos diferindo estatisticamente entre si.

Tabela 3. Precisão dos algoritmos de classificação para diferentes métodos de seleção de atributos

	J48	Random Forest	Naive Bayes	Multilayer Perceptron	IBK (5)	SMO	Média
Sem Seleção	0,767	0,839	0,448	0,508	0,557	0,433	**0,592B**
X^2	0,767	0,839	0,448	0,508	0,557	0,433	**0,592B**
CFS	0,751	0,837	0,427	0,457	0,569	0,411	**0,575A**
PCA	0,749	0,831	0,424	0,453	0,560	0,396	**0,569A**
Info Gain	0,757	0,822	0,428	0,463	0,551	0,432	**0,576A**
Gain Ratio	0,760	0,838	0,441	0,480	0,558	0,434	**0,585AB**
Wrapper	0,773	0,838	0,442	0,498	0,586	0,438	**0,596B**
Média	**0,760d**	**0,835e**	**0,437a**	**0,481b**	**0,562c**	**0,425a**	

Dados analisados por meio de ANOVA de fator duplo a 5%. Letras minúsculas (para algoritmos) e maiúsculas (para técnicas de seleção de atributos) diferentes indicam médias com diferença estatisticamente significativa, submetidas ao teste t (DMS) a 1%.

Com relação às diferentes técnicas de seleção de atributos, o ganho obtido pelo método de maior desempenho (Wrapper, com precisão média de 59,6%) com relação ao de menor (PCA, taxa de acerto de 56,9%) foi de 4,7%, evidenciando que para esta base de dados, o maior ganho em precisão é oriundo do algoritmo de classificação e não da seleção de atributos. O método χ^2 não excluiu nenhum atributo, mantendo a base de dados igual a ausência de seleção. Os métodos CFS, PCA e Info Gain foram, em média, os que apresentaram menor precisão, sendo inferiores inclusive à ausência de seleção. O Gain Ratio não diferiu estatisticamente de nenhum dos métodos testados, enquanto Wrapper apresentou resultado semelhante à ausência de seleção.

Os atributos selecionados por cada método de seleção para o algoritmo de classificação Random Forest são apresentados em ordem de importância decrescente na Tabela 4.

Tabela 4. Ranking em ordem decrescente dos atributos selecionados pelas diferentes técnicas para o classificador Random Forest

Atributos	X^2	CFS*	PCA*	Info Gain	Gain Ratio	Wrapper*	Média
Fósforo	15	S	S	15	15	S	**15,0**
pH (água)	14	S	S	14	11	S	**14,0**
Cálcio	13	S	S	13	10	S	**13,5**
Ca/Mg	7	S	S	7	5	S	**10,7**
K/(Ca + Mg)½	10	S	N	10	13	S	**10,5**
Silte	11	S	N	11	8	S	**10,0**
Magnésio	9	N	S	9	14	N	**7,8**
Mg/K	6	N	N	6	12	S	**6,5**
Cond. Elét. Sol. Solo	3	N	S	0	3	S	**6,0**
Matéria Orgânica	2	N	S	0	0	S	**5,3**
Soma de Bases	12	N	N	12	7	N	**5,2**
Potássio	5	N	S	5	4	N	**4,8**
Argila	4	N	N	0	9	S	**4,7**
Ca/K	8	N	N	8	6	N	**3,7**
Areia	1	N	N	0	0	N	**0,2**
N. atributos	15	6	8	11	13	10	

** métodos que apenas selecionam ou descartam atributos (S = 15 pontos; N = 0 pontos)*

Apesar do método de seleção de atributos CFS, na média dos algoritmos de classificação, ter apresentado um desempenho inferior, quando este tem seu desempenho analisado dentro do

algoritmo Random Forest, não difere estatisticamente dos demais métodos em precisão, porém gera o modelo mais simples, pois seleciona apenas seis atributos, que coincidem com os de maior importância no ranking.

O ranking dos atributos mostra que P, pH, Ca estão num primeiro grupo de importância, seguidos pelas relações Ca/Mg e K/(Ca + Mg)½ num segundo grupo. Os teores de P e pH também foram encontrados por Beauclair (1991) como sendo de importância para a previsão da produtividade, sendo o mesmo relatado por Dias et al. (1999) para o teor de Ca. As relações Ca/Mg e K/(Ca + Mg)½, segundo Orlando Filho et al. (1996), tem valores médios próximos a 2,0 e 0,19, respectivamente, sendo estes próximos dos valores medianos na base de dados (Tabela 1), o que contribui para a previsão da produtividade.

5. CONCLUSÕES

O algoritmo de classificação Random Forest apresentou a maior taxa de acerto para previsão de produtividade da cana-de-açúcar. O método de seleção de atributos CFS foi o que tornou o modelo menos complexo. As variáveis P, pH e Ca são as de maior importância para o modelo. Algoritmos de mineração de dados mostraram potencial na previsão de produtividade da cana-de-açúcar com base em atributos de fertilidade do solo.

6. AGRADECIMENTOS

Agradecimentos às Usinas Castilla e Rio Paila, que gentilmente disponibilizaram os dados.

7. REFERÊNCIAS

BEAUCLAIR, E.G.F. de. **Relações entre algumas propriedades químicas do solo e a produtividade da cana-de-açúcar (Saccharum spp) através de regressão linear múltipla.** 1991. 96 p. Dissertação (Mestrado) - Universidade de São Paulo, Piracicaba. 1991

DEMPSTER, A.P.; LAIRD, N.M.; RUBIN, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal of the Royal Statistical Society,** v. 39, n. 1, p. 1-38, 1977.

DIAS, F.L.F.; MAZZA, J.A.; MATSUOKA, S.; PERECIN, D.; MAULE, R.F. Produtividade da cana-de-açúcar em relação ao clima e solos da região noroeste do Estado de São Paulo. **Revista Brasileira de Ciência do Solo**, v. 23, n.3, p. 627-634, 1999.

FERRARO, D. O.; RIVERO, D.E.; GHERSA, C.M.. An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. **Field Crops Research**, n. 112, p. 149-157, 2009.

ORLANDO FILHO, J.; BITTENCOURT, V.C.; CARMELLO, Q.A.C.; BEAUCLAIR, E.G.F. Relações K, Ca e Mg de solo areia quartzoza e produtividade da cana-de-açúcar. **Stab: açúcar, álcool e subprodutos**, v.14, n. 5, p. 13-17, 1996.

SOUZA, Z. M. DE; CERRI, D.G.P.; COLET, M.J.; RODRIGUES, L.H.A.; MAGALHÃES, P.S.G.; MANDONI, R.J.A. Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística e árvore de decisão. **Ciência Rural**, v.40, n.4, p. 840-847, abril, 2010.

TERAMOTO, E.R. **Avaliação e aplicação de modelos de estimativa de produção de cana-de-açúcar (Sacharum spp) baseados em parâmetros do solo e do clima**. 2003. 96 p. Tese (Doutorado) - Universidade de São Paulo, Piracicaba. 2003

YANG, C.; PRACHER, S. O.; WHALEN, J.; GOEL, P. K. Use of hyperspectral imagery for identification of different fertilization with decision-tree technology. **Biosystems Engineering.** v. 83, n. 3, p.291-298, 2002.