XXXVIII CONGRESSO BRASILEIRO DE ENGENHARIA AGRÍCOLA
2 a 6 de agosto 2009
Juazeiro (BA)/Petrolina (PE)

MINERAÇÃO DE DADOS NA ESTIMATIVA DA PRODUTIVIDADE DE CANA-DE-AÇÚCAR USANDO DADOS AGROCLIMÁTICOS

WANDER J. PALLONE FILHO[1], JURANDIR ZULLO JUNIOR[2], STANLEY R. M. OLIVEIRA[3]

[1] Eng. Agrícola, Doutorando da Faculdade de Engenharia Agrícola da UNICAMP, Campinas-SP, Fone: (019)34298267, wander@ctc.com.br
[2] Eng. Agrícola, Pesquisador Dr. do Centro de Pesquisas Meteorológicas e Climáticas Aplicadas a Agricultura da UNICAMP, Campinas-SP
[3] Doutor em Mineração de Dados, Pesquisador da Embrapa Informática Agropecuária, Campinas-SP

Apresentado no
XXXVIII Congresso Brasileiro de Engenharia Agrícola – CONBEA 2009
2 a 6 de agosto de 2009 - Juazeiro-BA/Petrolina-PE

RESUMO: Este trabalho teve o objetivo de desenvolver modelos de estimativa de produtividade agrícola de cana planta (ano e meio), a partir de dados de produção e dados de precipitação e temperatura, da usina São Luiz SA, localizada no município de Ourinhos-SP. Essas estimativas são de grande interesse e importância às usinas de açúcar e álcool, para o planejamento de sua produção e comercialização. As usinas dispõem de grande volume de dados, gerados em suas atividades, pouco explorados com detalhes, pela dificuldade de extração de informação útil. Para abordar esse desafio, buscou-se o desenvolvimento de modelos, com base em técnicas de descoberta de conhecimento em banco de dados. A seleção de atributos, aliada ao balanceamento de classes, diminuiu a acurácia do classificador, mas aumentou a precisão da classificação da classe minoritária (alta produtividade).

PALAVRAS-CHAVE: temperatura, precipitação, classificação.

DATAMINIG FOR SUGAR CANE YIELD PREDICTION USING AGROCLIMATIC DATA

ABSTRACT: This work aimed at developing models to estimate sugarcane yield (first cut - one year and a half), from production, precipitation and temperature data collected from the Sugar Mill called Sao Luiz SA, located in Ourinhos, Sao Paulo State. These estimatives are relevant and important to support production and commercialization activities of sugar and alcohol mills. Several sugar mills have a large volume of data generated in their activities, but they have been little explored in detail, due to the complexity in extracting useful information. To address this challenge, some models were developed based on knowledge discovery in databases. The attribute selection in addition to class balancing reduced the classification accuracy, but improved the precision of the minority class (high productivity).

KEYWORDS: temperature, precipitation, classification.

INTRODUÇÃO: A cultura da cana-de-açúcar está entre as principais atividades agrícolas do Brasil, com 6,1 milhões de hectares (IBGE, 2007). A previsão de suas safras é de fundamental importância às usinas e destilarias para fins de planejamento agrícola (plantio, colheita e reforma, por exemplo) e de estratégias de negócios (estoques, preços e comercialização). A previsão de safras é baseada, principalmente, em levantamentos de área plantada e em estimativas de produtividade, tendo o clima forte influência sobre a última. Nessas estimativas, há subjetividade envolvida, especialmente, na produtividade. O clima é o fator que mais influencia a produtividade da cana-de-açúcar. A influência do clima na produtividade da cana-de-açúcar, especialmente envolvendo temperatura e disponibilidade de água, foi abordada em vários trabalhos citados por DELGADO ROJAS (1998), TERAMOTO (2003). As usinas canavieiras têm acumulado um grande volume de dados produzidos nas suas atividades, pouco explorados em nível de detalhe, pela dificuldade de extração de informação útil. As técnicas de mineração de dados tratam da busca de conhecimento não explícito em banco de dados,

como parte do processo de descoberta de conhecimento em banco de dados (HAN e KAMBER, 2006; FAYYAD et al., 1996). Dentre essas técnicas, a classificação permite obter modelos de predição, capazes de classificar um conjunto de dados em diferentes classes, a partir de atributos comuns a esse conjunto (CHEN, 1996). A classificação baseada em árvores de decisão é adequada às aplicações, onde a compreensão dos critérios de classificação é parte importante da análise (GANTI et al., 1999), com potencial para gerar conhecimento. O objetivo do trabalho foi desenvolver modelos de estimativa de produtividade, com foco na sua aplicabilidade, usando técnicas de descoberta de conhecimento em banco de dados: preparação de dados, seleção de atributos, geração e interpretação de modelos.

MATERIAL E MÉTODOS: Foram utilizados dados de produção e meteorológicos, da Usina São Luiz SA, localizada em Ourinhos-SP, relativos às safras de 2002 a 2005. Dados de produção: Safra, Variedade, Fazenda, Talhão, Estágio de Corte, Tipo de Solo, Distância, Data Plantio, Data Primeiro Corte, Data Corte Anterior, Data Corte Atual, Área, Fibra, Pureza, Produção Total, Produtividade Agrícola. Dados meteorológicos (diários): precipitação, temperatura máxima e temperatura mínima. A partir desses dados foram calculados os balanços hídricos (ciclo) da cana-de-açúcar (THORNTHWAITE e MATHER, 1955; DOORENBOS e KASSAM, 1994), e obtidos, os déficits hídricos (DH) e soma de graus de temperatura, acima de 19°C (SG), acumulados, por mês, do primeiro ao vigésimo mês após o plantio. A exploração inicial dos dados envolveu a análise de sua distribuição em tabelas e gráficos. Os dados de produção foram integrados às DH e SG, pelos meses de plantio e de colheita por safra (início e fim do balanço hídrico, respectivamente). A seleção de atributos baseou-se nos algoritmos: teste do Qui-quadrado, Ganho de Informação e Taxa de Ganho de Informação (HAN e KAMBER, 2006), visando eliminar atributos que não agregam informações úteis aos modelos. Conforme é necessário ao uso das árvores de decisão, os valores do atributo classe produtividade agrícola, foram transformados para nominais, formando três classes: A (alta, produtividade agrícola maior que 10% da média do bloco – safra, fazenda e variedade), B (baixa, produtividade agrícola menor que 10% da média do bloco) e N (normal, caso contrário). Os atributos "data plantio" e "data corte" foram transformados para o formato nominal, como meses de plantio e de corte, respectivamente. As árvores de decisão foram geradas usando algoritmo C4.5 (QUINLAN, 1993). Para o balanceamento das classes A (168 exemplos), N (667) e B (178), foram utilizados três métodos: *"Resample"* (amostragem com reposição), *NCL* e *Smotesampling* (BATISTA et al., 2004), os dois últimos são baseados no algoritmo de classificação k-vizinhos mais próximos. Os conjuntos de testes foram criados por amostragens estratificadas, antes do balanceamento de classes. Diferentes modelos foram gerados com conjuntos de dados balanceados (três métodos) e não balanceados, com e sem seleção de atributos. Os algoritmos utilizados estão disponíveis no software Weka (software gratuito, disponível em: http://www.cs.waikato.ac.nz/~ml/weka/index.html), que contém uma coleção de algoritmos de aprendizado de máquina para uso em aplicações de mineração de dados.

RESULTADOS E DISCUSSÃO: Verificou-se uma variação (ruído/inconsistência) importante dos valores de produtividade (atributo classe), entre talhões de mesmo bloco. A causa provável é a dificuldade de controle durante a colheita, podendo ocorrer lançamento da produção de um talhão no outro, inclusive entre talhões com variedades e/ou estágios diferentes. Verificou-se que talhões com cana planta de ano e meio, foram menos susceptíveis a esse problema. Observa-se na Tabela 1, que nas safras de 2002 a 2005, a cana de ano e meio (corte 1) representou 12,4% da área total plantada e 17% da produção total obtida. Os atributos data plantio, data corte, safra, fazenda, talhão, produção total, estágio corte, data primeiro corte e data corte anterior foram descartados, antes da seleção de atributos, porque não agregariam informação útil ao modelo. Na seleção de atributos, adotou-se o critério de descartar os últimos 12 atributos ordenados, por coincidirem nos três métodos utilizados. Atributos descartados: SG acumulados do 16° ao 20° mês após o plantio, e DH acumulada do

1° ao 8° mês após o plantio. Na Fig. 1, pode ser observado que o período de plantio da cana de ano e meio concentra-se nos meses de janeiro a março, principalmente em fevereiro. Já o período de colheita, distribui-se por vários meses entre maio e setembro, visando obter as melhores produtividades, considerando as características varietais. Os maiores picos de colheita são observados nos meses de maio e junho (plantio fevereiro). Pode ser observado na Tabela 2, que a seleção de atributos aumentou ligeiramente a acurácia dos modelos, para dados balanceados e não balanceados. A combinação da seleção de atributos com os métodos *NCL* e *Smotesampling* aumentou a precisão da classe minoritária (A - alta). Por outro lado, a seleção de atributos no caso dos dados não balanceados e balanceados pelo método *Resample,* diminuiu a precisão da classe minoritária (A – alta). O balanceamento de classes combinado com a seleção de atributos foi positivo, tendo sempre elevado a precisão da classe minoritária (A - alta), apesar de ter reduzido a acurácia dos modelos. Por outro lado, o balanceamento de classes foi negativo no caso dos dados sem seleção de atributos, pois nesse caso reduziu tanto a acurácia dos modelos quanto a precisão da classe minoritária (A - alta). A combinação que produziu o modelo com maior acurácia foi a seleção de atributos com os dados não balanceados, porém, esse modelo apresentou a precisão da classe minoritária (A - alta) nula.

Tabela 1 – Distribuição de área e produtividade, por estágio de corte, nas safras de 2002 a 2005.

Corte	Area (mil ha)	Área (%)	Produção (mil toneladas)	Área (%)
1	9,6	12,4	1.086,1	17,0
2	17,8	23,0	1.589,1	24,9
3	15,7	20,2	1.208,5	19,0
4	12,7	16,4	917,8	14,4
5	9,9	12,8	704,6	11,1
> 6	11,7	15,2	866,8	13,6
Total	77,5		6.372,8	

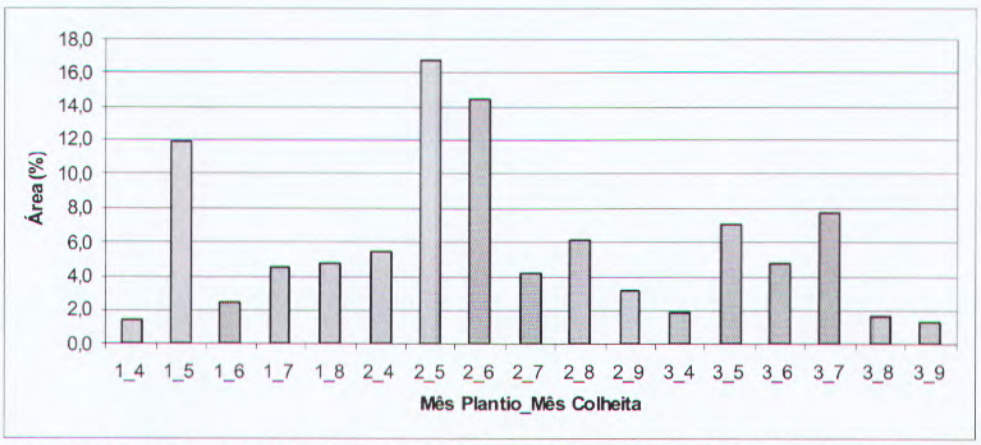

Figura 1 – Gráfico da distribuição de área ocupada, por época de plantio e de colheita, nas safras de 2002 a 2005.

Tabela 2 – Acurácia e precisão da classe minoritária (A) das diferentes árvores geradas, com dados balanceados e não-balanceados, com e sem seleção de atributos.

Seleção de Atributos	Não balanceado	*Resample*	*NCL*	*Smotesampling*
Sem	62,9 / 33,3	39,2 / 22,6	57,8 / 22,7	51,9 / 15,8
Com	64,6 / 0,0	41,2 / 18,8	58,8 / 25,0	53,9 / 20,0

Acurácia do modelo / Precisão da classe minoritária A.

CONCLUSÕES: As técnicas de mineração de dados propiciaram a geração de modelos, com base em dados disponíveis nas usinas, portanto, com potencial para aplicação prática. A seleção de atributos teve efeito positivo, no geral, na acurácia dos modelos, e irregular na precisão da classe minoritária (A - alta). O balanceamento de classes combinado com a seleção de atributos aumentou expressivamente a precisão da classe minoritária (A - alta). A combinação que proporcionou a obtenção dos melhores resultados, foi a com dados sem seleção de atributos e sem balanceamento de classes. O uso de informações agrometeorológicas com freqüência quinzenal, decendial ou semanal, além do tratamento dos ruídos e inconsistência presentes nos dados comerciais, podem contribuir para o aumento da precisão dos modelos.

REFERÊNCIAS

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. *A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data*. SIGKDD Explorations, v.6, n.1, p.20-29, 2004.

CHEN, Ming-Syan; HAN, Jiawei; YU, Philip S. Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering, v. 8, n. 6, p. 866-883, dec. 1996.

DELGADO ROJAS, J. S. *Modelo agrometeorológico para estimativa dos efeitos de deficiência hídrica na produtividade agro-industrial da cana-de-açúcar*. Dissertação (Mestrado), 74 p. Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, 1998.

DOORENBOS, J.; KASSAM, A.M. Efeito da água no rendimento das culturas. Campina Grande, UFPB, 1994.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, v. 17, n. 3, Fall 1996, p.37-54.

GANTI, V.; GEHRKE, J.; RAMAKRISHNAN, R. Mining Very Large Databases. *Computer*, v. 32, n. 8, p. 38-45, aug. 1999.

HAN, J. ; KAMBER, M. Data mining: concepts and techniques, 2nd edition. Morgan Kaufmann Publishers, 2006.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Disponível em: <http://www.ibge.gov.br/home/estatistica/indicadores/agropecuaria/lspa/defaulttab.shtm>. Acesso em: 10 ago. 2007.

QUINLAN, J. R. C4.5: programs for machine Learning. Morgan Kaufman, San Mateo, CA, 1993.

SOUSA, M. S. R. *Mineração de Dados: Uma implementação fortemente acoplada a um sistema gerenciador de banco de dados paralelo*, Rio de Janeiro, 1998. 67p. Dissertação (Mestrado). Universidade Federal do Rio de Janeiro.

TERAMOTO, E. R. *Avaliação e aplicação de modelos de estimativa de produção de cana-de-açúcar (Saccharum spp) baseados em parâmetros do solo e do clima*, 2003. 96p. Tese (Doutorado) Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

THORNTHWAITE, C.W.; MATHER, J.R. The water balance. Centerton, NJ: Drexel Institute of Technology - Laboratory of Climatology, 1955. 104p. (Publications in *Climatology*, vol. VIII, n.1)