Anais do 6º Congresso Brasileiro de Agroinformática – SBIAgro 2007, São Pedro, Brasil, 8-11 outubro 2007, Embrapa Informática Agropecuária, p.381-385.

PREPARAÇÃO DE DADOS PARA OBTENÇÃO DE MODELOS DE ALERTA DA FERRUGEM DO CAFEEIRO

CARLOS ALBERTO ALVES MEIRA[1]
LUIZ HENRIQUE ANTUNES RODRIGUES[2]

RESUMO: Este trabalho descreve a fase de preparação de dados de um processo de descoberta de conhecimento em bases de dados para obtenção de modelos de alerta da ferrugem do cafeeiro. Parte dos atributos foi preparada segundo elementos epidemiológicos conhecidos da doença e parte como estatísticas descritivas comuns. Métodos de seleção de atributos indicaram que os atributos preparados de acordo com condições que favorecem ou inibem o desenvolvimento da ferrugem possuem maior poder preditivo em relação à taxa de infecção da doença do que os demais atributos preparados.

PALAVRAS-CHAVE: descoberta de conhecimento em bases de dados, mineração de dados, seleção de atributos, sistemas de alerta de doenças de plantas.

DATA PREPARATION TO OBTAIN COFFEE RUST WARNING MODELS

ABSTRACT: This paper describes the data preparation step of a process of knowledge discovery in databases which aims to generate coffee rust warning models. One part of the attributes was prepared according to epidemiology aspects of the disease while the other part was prepared as common descriptive statistics. Attribute selection methods pointed that the attributes prepared according to conditions that favor or inhibit rust development have more predictive power in relation to the disease infection rate than the others attributes.

KEYWORDS: knowledge discovery in databases - KDD, data mining, attribute selection, plant disease warning systems.

1. INTRODUÇÃO

Sistemas de alerta de doenças de plantas dão suporte à tomada de decisão ao indicar as condições que favorecem uma doença, permitindo agir somente quando necessário e diminuir o uso de agrotóxicos (CAMPBELL e MADDEN, 1990). Restrições que têm limitado o seu uso incluem (REIS, 2004): indisponibilidade dos dados requeridos por certos modelos e custo de implementação e manutenção para o agricultor. Um projeto em desenvolvimento (MEIRA e RODRIGUES, 2005) tem como hipótese que uma análise de dados meteorológicos junto com registros de intensidade de doenças causadas por fungos em culturas agrícolas, caracterizada como um processo de descoberta de conhecimento em bases de dados, conhecido como processo de KDD (FAYYAD et al., 1996), indicará a viabilidade de uso dos modelos obtidos na emissão de alertas, como produto integrante de um sistema de monitoramento agrometeorológico de alcance público, gratuito e simples de usar. Os objetivos gerais desse projeto são avaliar tarefas e técnicas de mineração de dados no desenvolvimento de modelos de alerta da ferrugem do cafeeiro e caracterizar o processo de KDD para utilizá-lo em problemas similares do domínio de aplicação.

Segundo Zambolim et al. (2002), a ferrugem do cafeeiro, causada pelo fungo *Hemileia vastatrix* Berk. & Br., é considerada a principal doença da cultura do café, proporcionando decréscimos de produção que variam de 35 a 50%. Esses autores, com relação à influência das condições do tempo sobre a ferrugem, compilaram alguns resultados de pesquisa: a

[1] Matemático, Embrapa Informática Agropecuária, E-mail: carlos@cnptia.embrapa.br.
[2] Engenheiro Agrícola, Feagri – Unicamp, E-mail: lique@agr.unicamp.br.

temperatura ótima de germinação do fungo é estimada em 23,7 °C; temperaturas superiores a 30 ºC e inferiores a 14 ºC são limitantes à infecção; seis horas de água livre na superfície da folha é o tempo mínimo para ocorrer infecção; o período noturno é mais favorável à infecção, devido à ausência de luz, que inibe a germinação; o período de incubação, dependendo da temperatura, pode variar de 28 a 65 dias; orvalho e chuva leve seriam as melhores condições para a germinação; vento e chuva são os principais agentes de disseminação do fungo.

A preparação dos dados para a modelagem é fundamental e consome a maior parte do tempo em projetos de mineração de dados. O desafio é preparar os dados de forma que a informação contida neles seja exposta da melhor maneira para as ferramentas de mineração (PYLE, 1999). Um aspecto importante da preparação é procurar incorporar conhecimento prévio do domínio da área de aplicação nos dados preparados. Outro aspecto importante, para o tipo de problema apresentado, são as séries temporais, isto é, atributos que são medidos ao longo do tempo em intervalos fixos, como a temperatura do ar, a umidade relativa e a precipitação pluvial. São necessárias transformações nos dados e derivação de novos atributos para que a dimensão temporal seja incorporada no formato de dados usual reconhecido pelos algoritmos tradicionais de mineração (WEISS e INDURKHYA, 1998).

Este trabalho descreve, na próxima seção, como foi planejada e executada a fase de preparação dos dados do projeto mencionado, de acordo com elementos epidemiológicos da ferrugem do cafeeiro e levando-se em consideração as séries temporais meteorológicas disponíveis para a análise. Em seguida, os resultados dessa fase são apresentados e discutidos em termos do poder preditivo dos atributos preparados em relação à taxa de infecção da doença. Ao final, são apresentadas as conclusões deste trabalho.

2. MATERIAL E MÉTODOS

Os dados disponíveis se referem ao acompanhamento mensal da incidência da ferrugem (percentual de folhas atacadas de uma amostra), desde o ano agrícola 1998/1999 até 2005/2006, na fazenda experimental da Fundação Procafé em Varginha, MG. Foram selecionadas, a cada ano, oito áreas em produção, sendo quatro em espaçamento largo e as demais adensadas. Foram coletadas folhas de talhões sem controle da doença, sendo que para os dois espaçamentos foram utilizadas lavouras com carga pendente (produção) alta e baixa. Uma estação meteorológica automática registrou diversos atributos a cada 30 minutos, como temperatura, precipitação, radiação solar, fluxo e direção do vento e umidade relativa do ar.

A Figura 1 ilustra um hipotético dia-a-dia de infecção e do conseqüente aparecimento dos sintomas da ferrugem. No final desse período, representa-se a data em que é feita uma das avaliações (A_i) mensais da incidência da doença. A manifestação da ferrugem nessa data corresponde à evolução do aparecimento dos sintomas desde a última avaliação (A_{i-1}). Os novos sintomas que surgem (D_s – dia de sintoma) são o resultado de infecções (D_i – dia de infecção) ocorridas anteriormente, que se desenvolvem durante o período de incubação (PI) até se expressarem como sintomas visíveis. Os períodos de infecção (P_{INF}) correspondentes a cada avaliação mensal da doença são variáveis, em conseqüência das variações nos períodos de incubação.

A preparação dos dados levou em consideração as condições que favorecem o desenvolvimento da ferrugem nos períodos P_{INF} e PI. As temperaturas médias mínimas e máximas diárias representam bem o efeito desta variável no período de incubação (MORAES et al., 1976). No processo de infecção, a temperatura média durante o período de molhamento foliar é que deve ser considerada, pois só há infecção na presença de água líquida sobre a superfície foliar, principalmente no período noturno. O número de horas com alta umidade relativa do ar (p. ex. $\geq$ 90%) foi utilizado como medida indireta de molhamento foliar (SUTTON et al., 1984).

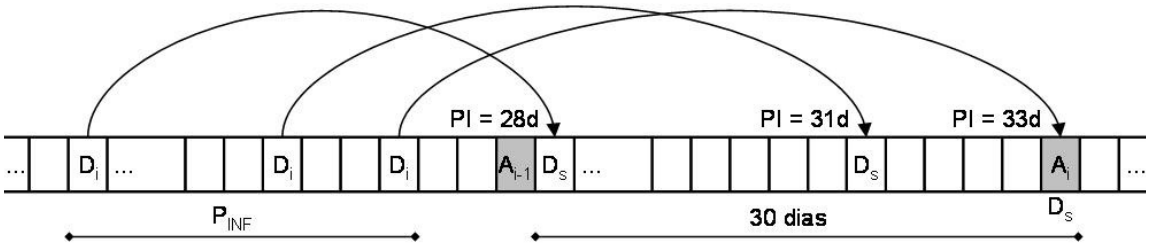

Figura 1. Dia-a-dia de infecção e aparecimento dos sintomas da ferrugem.

Os atributos meteorológicos para análise foram construídos a partir do nível horário (registros da estação), passando pelo nível diário, até o nível de granularidade mensal, para possibilitar a integração com os registros de incidência da ferrugem. Do nível horário para o diário, por exemplo, foi criado o atributo CDINF, representando a condição diária de infecção (desfavorável, favorável ou muito favorável), de acordo com condições de molhamento foliar e de temperatura e luminosidade durante o período de molhamento. Do nível diário para o mensal, derivam de CDINF três atributos, representando o número de dias desfavoráveis, favoráveis e muito favoráveis no período de infecção, respectivamente DDI_PINF, DFI_PINF e DMFI_PINF. Foram geradas também estatísticas descritivas (médias e somatórios) dos atributos meteorológicos durante cada período de infecção.

O atributo dependente ou classe (aquele que se pretende predizer) foi definido como a taxa de infecção da ferrugem (A_i - A_{i-1}). Os valores numéricos foram depois transformados em intervalos discretos, definindo três níveis categóricos de taxa de infecção: tx1(<0), tx2(>=0<=5) e tx3(>5).

Foram utilizados quatro métodos de seleção de atributos, com a finalidade de produzir um *ranking* de acordo com o poder preditivo de cada um em relação à taxa de infecção. Eles fazem parte do *software* de mineração de dados Weka (WITTEN e FRANK, 2005). Um dos métodos avalia subconjuntos de atributos baseado em correlação (CfsSubsetEval) e os demais avaliam os atributos individualmente com respeito à classe, um baseado no teste do qui-quadrado (ChiSquaredAttributeEval), outro no ganho de informação (InfoGainAttributeEval) e o terceiro na razão de ganho (GainRatioAttributeEval).

3. RESULTADOS E DISCUSSÃO

A Tabela 1 apresenta o resultado da avaliação dos atributos quanto à importância na predição da taxa de infecção da ferrugem (atributo dependente). Ao todo foram avaliados 33 atributos independentes ou preditores. São apresentados os 16 primeiros relacionados por cada método de seleção de atributos utilizado, admitindo-se como uma possível meta a redução em torno da metade da quantidade de atributos independentes para a fase seguinte de modelagem. Em síntese, a maioria dos atributos apresentados aparece simultaneamente nas quatro relações.

Dias desfavoráveis e muito favoráveis à infecção no período de infecção (DDI_PINF e DMFI_PINF, respectivamente) aparecem nas primeiras posições em todas as relações, enquanto DFI_PINF (dias favoráveis) aparece uma vez (CFS, posição 11). Também, DIAS_PINF (DDI_PINF+DFI_PINF+DMFI_PINF) está bem posicionado nas quatro relações. Além desses, outros atributos preparados segundo elementos epidemiológicos que favorecem o desenvolvimento da ferrugem no período de infecção aparecem bem posicionados: NHUR90_PINF/THUR90_PINF (média diária do número de horas com umidade relativa ≥ 90% e temperatura média durante esse período) e NHNUR90_PINF/THNUR90_PINF (idem, mas no período noturno). As temperaturas média e média mínima no período de incubação (TMED_PI_PINF e TMIN_PI_PINF) aparecem na tabela em mais de uma relação. Além disso, indiretamente, tanto a média mínima quanto a

média máxima de temperatura no período de incubação são consideradas na determinação do período de infecção.

Tabela 1: *Ranking* dos atributos independentes[3] quanto ao mérito preditivo em relação ao atributo dependente (taxa de infecção da ferrugem do cafeeiro).

Pos.	CFS	Qui-quadrado	Ganho de informação	Razão de ganho
1	30 ddi_pinf	30 ddi_pinf	30 ddi_pinf	29 dmfi_pinf
2	29 dmfi_pinf	29 dmfi_pinf	29 dmfi_pinf	24 thur90_pinf
3	34 dias_pinf	27 thnur90_pinf	7 tmin_pinf	30 ddi_pinf
4	9 nhluz_pinf	7 tmin_pinf	27 thnur90_pinf	27 thnur90_pinf
5	24 thur90_pinf	24 thur90_pinf	24 thur90_pinf	5 tmed_pinf
6	21 ur_pinf	11 bar_pinf	34 dias_pinf	7 tmin_pinf
7	11 bar_pinf	34 dias_pinf	11 bar_pinf	25 nhnur90_pinf
8	7 tmin_pinf	21 ur_pinf	21 ur_pinf	9 nhluz_pinf
9	5 tmed_pinf	5 tmed_pinf	5 tmed_pinf	34 dias_pinf
10	2 carga	9 nhluz_pinf	9 nhluz_pinf	22 nhur90_pinf
11	28 dfi_pinf	25 nhnur90_pinf	25 nhnur90_pinf	17 med_precip_pinf
12	27 thnur90_pinf	17 med_precip_pinf	17 med_precip_pinf	6 tmax_pinf
13	25 nhnur90_pinf	22 nhur90_pinf	22 nhur90_pinf	21 ur_pinf
14	12 vvento_pinf	6 tmax_pinf	6 tmax_pinf	33 tmin_pi_pinf
15	3 enfolhamento	31 tmed_pi_pinf	33 tmin_pi_pinf	10 esolar_pinf
16	22 nhur90_pinf	33 tmin_pi_pinf	31 tmed_pi_pinf	20 med_indpluv_pinf

As estatísticas descritivas de destaque (aparecem pelo menos três vezes na tabela; todas são médias para o período de infecção) correspondem aos seguintes atributos meteorológicos: temperatura (TMED_PINF, TMIN_PINF e TMAX_PINF), umidade relativa (UR_PINF), pressão barométrica (BAR_PINF) e precipitação (MED_PRECIP_PINF). Outro atributo que se destaca é NHLUZ_PINF (média diária do número de horas com luminosidade no período de infecção).

O cafeeiro é uma planta bianual, que de dois em dois anos apresenta alta produção. Nesses anos, identificados como de alta carga pendente, a ferrugem atinge maior intensidade do que nos anos de baixa carga. Outro fator da planta que influencia o desenvolvimento da doença é a densidade de plantio. Lavouras adensadas influenciam as condições microclimáticas dentro do cafezal, tornando o ambiente mais propício à incidência da ferrugem. No entanto, o atributo CARGA aparece apenas uma vez (CFS, posição 10); já o atributo LAVOURA (adensada ou larga) não figura na Tabela 1.

4. CONCLUSÕES

Os resultados apresentados indicam que o planejamento e a execução da preparação dos dados, segundo elementos epidemiológicos conhecidos da ferrugem do cafeeiro, devem permitir a obtenção de modelos de alerta da doença de melhor desempenho, em comparação caso a preparação não tivesse levado em conta tais fatores. No geral, os atributos preparados de acordo com condições que favorecem ou inibem o desenvolvimento da ferrugem tiveram melhores resultados na avaliação do que aqueles definidos como estatísticas descritivas, que são a forma mais comum de se preparar dados para esse tipo de análise.

[3] O número que antecede o nome do atributo em cada célula da tabela corresponde à posição em que o atributo aparece no conjunto de dados avaliado.

Estes resultados precisam ainda passar por uma análise e discussão junto com especialistas para verificar no que convergem ou divergem do que é senso comum no domínio de conhecimento da área de aplicação.

Este trabalho é importante para as fases seguintes do processo de descoberta de conhecimento em bases de dados. Um subconjunto dos atributos preparados deve ser selecionado para a fase de modelagem, o que vai permitir a obtenção de modelos de menor complexidade, sem comprometimento no desempenho. Para o projeto de pesquisa no qual está inserido este trabalho, cuja proposta é aplicar a técnica de árvores de decisão, essa menor complexidade se traduz em árvores de menor tamanho, com menor quantidade de nós e ramos, e, consequentemente, maior poder de interpretação das regras geradas.

5. AGRADECIMENTOS

À Fundação Procafé por ceder os dados relacionados com o monitoramento da ferrugem do cafeeiro, em especial ao Eng° Agrônomo Leonardo Bíscaro Japiassú.

6. REFERÊNCIAS BIBLIOGRÁFICAS

CAMPBELL, C. L.; MADDEN, L. V. **Introduction to plant disease epidemiology**. New York: John Wiley & Sons, 1990. 532 p.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996.

MEIRA, C. A. A.; RODRIGUES, L. H. A. Mineração de dados no desenvolvimento de sistemas de alerta contra doenças de culturas agrícolas. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 5., 2005, Londrina. **Agronegócio, tecnologia e inovação:** anais. Londrina: FAPEAGRO/SBI-AGRO, 2005. 1 CD-ROM.

MORAES, S. A.; SUGIMORI, M. H.; RIBEIRO, I. J. A.; ORTOLANI, A. A.; PEDRO JR., M. J. Período de incubação de *Hemileia vastatrix* Berk. et Br. em três regiões do Estado de São Paulo. **Summa Phytopathologica**, Piracicaba, v. 2, n. 1, p. 32-38, 1976.

PYLE, D. **Data preparation for data mining**. San Francisco: Morgan Kaufmann, 1999. 540 p.

REIS, E. M. (Ed.) **Previsão de doenças de plantas**. Passo Fundo: UPF, 2004. 316 p.

SUTTON, J. C.; GILLESPIE, T. J.; HILDEBRAND, P. D. Monitoring weather factors in relation to plant disease. **Plant Disease**, v. 68, n. 1, p. 78-84, 1984.

WEISS, S. M.; INDURKHYA, N. **Predictive data mining**: a practical guide. San Francisco: Morgan Kaufmann, 1998. 228 p.

WITTEN, I. H.; FRANK, E. **Data mining**: practical machine learning tools and techniques. 2 ed. San Francisco: Morgan Kaufmann, 2005. 525 p.

ZAMBOLIM, L.; VALE, F. X. R.; COSTA, H.; PEREIRA, A. A.; CHAVES, G. M. Epidemiologia e controle integrado da ferrugem-do-cafeeiro. In: ZAMBOLIM, L. (Ed.). **O estado da arte de tecnologias na produção de café**. Viçosa: Suprema Gráfica e Editora, p. 369-449, 2002.