ÁRVORES DE DECISÃO INDUZIDAS PELO WEKA PARA ALERTA DA FERRUGEM DO CAFEEIRO EM LAVOURAS COM ALTA CARGA PENDENTE

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

RESUMO: O objetivo deste trabalho foi desenvolver árvores de decisão com o software livre Weka para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente de frutos e comparar esses modelos com as árvores de decisão induzidas por um software proprietário. Dados de incidência mensal da doença no campo coletados durante oito anos foram transformados em valores binários considerando limites de 5 e 10 pontos percentuais (p.p.) na taxa de infecção. Foi gerado um modelo para cada taxa de infecção binária, a partir de dados meteorológicos e do espaçamento entre plantas. O alerta é indicado quando a taxa de infecção, prevista para o prazo de um mês, atingir ou ultrapassar o respectivo limite. As árvores de decisão induzidas pelo Weka tiveram desempenho semelhante às induzidas pelo software proprietário. Os modelos de alerta para o limite de 5 p.p. induzidos pelas duas ferramentas são praticamente iguais. O modelo considerando o limite de 10 p.p. induzido pelo Weka é mais simples e compacto que o induzido pelo software proprietário.

PALAVRAS-CHAVE: *Coffea arabica*, *Hemileia vastatrix*, doença de plantas, previsão, modelos, mineração de dados.

DECISION TREES INDUCED BY WEKA FOR COFFEE RUST WARNING IN GROWING AREAS WITH LARGE FRUIT LOAD

ABSTRACT: The objective of this work was to develop decision trees with the free software Weka for coffee rust warning in growing areas with large fruit load and compare these models with the decision trees induced by a proprietary software. Monthly data of disease incidence in the field collected during eight years were transformed into binary values considering limits of 5 and 10 percentage points (pp) in the infection rate. Models were generated from meteorological data and space between plants for each binary infection rate. The warning is indicated when the infection rate is expected to reach or exceed the respective limit in a month. The decision trees induced by Weka had performance similar to that induced by the proprietary software. The warning models for the limit of 5 pp induced by both tools are equal in pratice. The model for the limit of 10 pp induced by Weka is simpler and more compact than the one induced by the proprietary software.

KEYWORDS: *Coffea arabica*, *Hemileia vastatrix*, plant disease, prediction, models, data mining.

1. INTRODUÇÃO

A ferrugem do cafeeiro (*Hemileia vastatrix* Berk. & Br.) é a principal doença da cultura do café em todo o mundo. No Brasil, os prejuízos na produção atingem cerca de 35%, em média, podendo chegar a mais de 50%, nas regiões onde as condições climáticas são favoráveis à doença (ZAMBOLIM et al., 1997). A importância econômica e outros aspectos, como a variação na intensidade da doença em cada ano agrícola e a disponibilidade de medidas de

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

controle economicamente viáveis, justificam o desenvolvimento de modelos de alerta ou de previsão da ferrugem do cafeeiro.

A maioria dos modelos empíricos de previsão da ferrugem do cafeeiro utilizou o ajuste dos dados observados a equações de regressão (ZAMBOLIM et al., 2002). Redes neurais também foram utilizadas para modelar epidemias da doença (PINTO et al., 2002). A indução de árvores de decisão é outra alternativa, com exemplos de modelos para doenças de outras culturas agrícolas, como a cercosporiose do milho (PAUL e MUNKVOLD, 2004) e a giberela do trigo (MOLINEROS et al., 2005).

Meira et al. (2009) desenvolveram árvores de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente de frutos, em que a predisposição das plantas ao ataque da ferrugem é maior. Os modelos foram induzidos com o SAS® Enterprise Miner™ (www.sas.com/technologies/analytics/datamining/miner; SAS INSTITUTE INC., 2004), que é a solução SAS para iniciativas de mineração de dados.

O objetivo deste trabalho foi desenvolver árvores de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente, a partir dos mesmos dados utilizados por Meira et al. (2009), mas com o software livre Weka (www.cs.waikato.ac.nz/ml/weka; WITTEN e FRANK, 2005), e comparar esses modelos com os induzidos pelo software proprietário.

2. MATERIAL E MÉTODOS

Os dados analisados foram do acompanhamento mensal da incidência da ferrugem do cafeeiro na fazenda experimental da Fundação Procafé, em Varginha-MG, de outubro de 1998 a outubro de 2006. Em cada ano, foram selecionadas oito lavouras de café em produção, quatro em espaçamento largo e quatro adensadas. Dessas quatro, duas lavouras tinham alta carga pendente de frutos e duas tinham baixa carga pendente. Não houve controle da doença durante o ano agrícola nos talhões escolhidos. Dados meteorológicos foram registrados por uma estação meteorológica automática instalada próximo dos locais de avaliação da ferrugem.

A análise dos dados foi conduzida como um processo de descoberta de conhecimento em bases de dados (FAYYAD et al., 1996), de acordo com o modelo de processo de mineração de dados CRISP-DM (CHAPMAN et al., 2000). A instância do processo compreendeu as fases de compreensão do domínio, de entendimento dos dados, de preparação dos dados, de modelagem e de avaliação dos modelos obtidos (MEIRA, 2008).

A característica da epidemia de interesse foi o progresso da ferrugem do cafeeiro entre uma avaliação e a outra. Taxas de infecção foram calculadas pela diferença entre a incidência da doença em um mês e a incidência no mês anterior; em seguida, os valores numéricos foram mapeados para duas categorias ou classes. A primeira opção de taxa de infecção binária foi a variável dependente TAXA_INF_M5, com valor 1 para taxas de infecção maiores ou iguais a 5 pontos percentuais (p.p.) e valor 0 no caso contrário. Este limite de decisão foi determinado com base no limite de 5% de incidência recomendado por Zambolim et al. (1997) para o controle da doença por via foliar. A outra opção foi a variável TAXA_INF_M10, considerando o limite de decisão em 10 p.p., este determinado com base em Kushalappa et al. (1984), que propuseram 10% de incidência para recomendar a aplicação de fungicida.

As variáveis independentes meteorológicas foram construídas a partir do nível horário (registros da estação), até o nível que permitisse a análise de seu relacionamento com a variável dependente (MEIRA, 2008). O número de horas com alta umidade relativa do ar (≥95%) foi utilizado como medida indireta de molhamento foliar contínuo. Considerando um período de incubação estimado, de acordo com a equação proposta por Moraes et al. (1976), cada dia foi tratado como um eventual dia de infecção e foi associado ao mês correspondente da taxa de infecção para a qual possivelmente teve parcela de contribuição. O conjunto de dias associado a uma taxa de infecção foi denominado de período de infecção (PINF). As

variáveis meteorológicas usadas na modelagem foram derivadas para esses períodos de infecção. O espaçamento da lavoura completou o conjunto das variáveis independentes.
A modelagem foi específica por carga pendente de frutos, produzindo modelos de alerta próprios para cafeeiros com alta carga e com baixa carga. Procurou-se particularizar o uso dos modelos, de acordo com a característica bianual dos cafezais, que nos anos de alta carga estão mais predispostos ao ataque da ferrugem que nos anos de baixa carga (ZAMBOLIM et al., 2002). Neste trabalho, são discutidos os modelos para lavouras com alta carga pendente.
O conjunto de dados preparado totalizou 192 exemplos ou casos (8 anos x 12 meses x 2 espaçamentos). Dez exemplos foram eliminados em razão de períodos de falha no registro da estação meteorológica, encerrando o conjunto de dados para a modelagem, ou conjunto de treinamento, com 182 exemplos.
A indução das árvores de decisão foi feita com o software Weka, por meio do classificador J48 (WITTEN e FRANK, 2005). No J48, o número mínimo permitido de exemplos em cada nó folha foi alterado de dois para cinco exemplos, igual ao valor utilizado no Enterprise Miner por Meira et al. (2009), sendo este o único valor padrão de configuração alterado.
A acurácia e a taxa de erro (WITTEN e FRANK, 2005) foram avaliadas para cada árvore de decisão. Outras medidas de avaliação foram consideradas (MONARD e BARANAUSKAS, 2002): sensitividade, especificidade, confiabilidade positiva e confiabilidade negativa. Todas as medidas foram calculadas por validação cruzada (WITTEN e FRANK, 2005).

3. RESULTADOS E DISCUSSÃO

Os alertas da ferrugem do cafeeiro são considerados quando a taxa de infecção da doença, prevista para o prazo de um mês, atingir ou ultrapassar os limites de 5 p.p. e 10 p.p. Esses alertas dão base para a decisão sobre as medidas a serem adotadas para o controle da doença e o melhor momento de implementá-las (MEIRA et al., 2009).
A árvore de decisão gerada com o Weka para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente, considerando o limite de 5 p.p. (TAXA_INF_M5), é apresentada na Figura 1. Ela possui a mesma estrutura de árvore do modelo obtido com o Enterprise Miner (MEIRA et al., 2009). Os atributos de teste dos dois modelos são os mesmos e os valores para os limites de decisão são praticamente iguais.

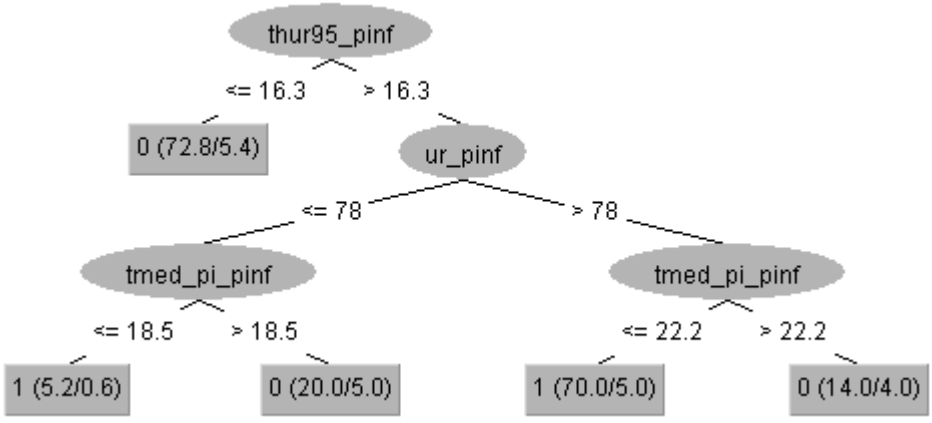

Figura 1. Árvore de decisão gerada com o Weka para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente, considerando o limite de 5 p.p. na taxa de infecção.

As influências da temperatura e da umidade relativa do ar se revelaram mais importantes no progresso da ferrugem do cafeeiro. Temperaturas médias mais baixas durante o molhamento foliar (THUR95_PINF ≤ 16,3°C) foram desfavoráveis às taxas de infecção maiores ou iguais a 5 p.p., assim como temperaturas médias diárias mais elevadas no período de incubação:

TMED_PI_PINF maior que 18,5°C ou 22,2°C, dependendo da condição de umidade (Figura 1). Umidade relativa média diária mais alta (UR_PINF ≥ 78%) foi mais favorável às taxas de infecção maiores ou iguais a 5 p.p. Esses efeitos estão de acordo com estudos epidemiológicos da doença, conforme discutido por Meira et al. (2009).

O modelo da Figura 1 é simples, compacto e foi bem avaliado. A sua avaliação foi semelhante à do modelo gerado com o Enterprise Miner, como mostra a Tabela 1. A acurácia dá a noção da proporção de acertos que o modelo pode ter caso venha a ser aplicado ao problema real. As outras medidas também são importantes: a sensitividade estima a capacidade que o modelo tem de acertar nas situações em que se deve emitir um alerta; a confiabilidade positiva estima a capacidade do modelo de emitir corretamente os alertas; a especificidade e a confiabilidade negativa são correspondentes às duas primeiras, mas para as situações em que o alerta não é necessário ou não é emitido. Portanto, o equilíbrio exibido pelo modelo entre as medidas de avaliação é positivo.

Tabela 1. Avaliação dos modelos de alerta da ferrugem do cafeeiro.

Medida de avaliação	TAXA_INF_M5		TAXA_INF_M10	
	Weka	EM	Weka	EM
Acurácia	81,9%	81,3%	79,2%	79,2%
Taxa de erro	18,1%	18,7%	20,8%	20,8%
Sensitividade	78,5%	79,9%	57,3%	70,3%
Especificidade	84,7%	82,6%	88,5%	82,8%
Confiabilidade positiva	82,7%	79,4%	69,8%	65,3%
Confiabilidade negativa	83,5%	83,9%	83,7%	86,9%

A Figura 2 apresenta a árvore de decisão induzida pelo Weka considerando o limite de 10 p.p. (TAXA_INF_M10). A temperatura média diária no período de infecção (TMED_PINF) foi considerada, sendo que temperaturas acima de 21,2°C foram mais favoráveis às taxas de infecção maiores ou iguais a 10 p.p. O modelo é uma subárvore da árvore obtida com o Enterprise Miner, resultado de dois tipos de poda realizados pelo classificador J48 (WITTEN e FRANK, 2005). O primeiro tipo de poda substituiu subárvores por nós folhas (*subtree replacement*) e o segundo tipo resultou na troca do atributo de teste do nó raiz, podando o teste em THUR95_PINF e elevando a subárvore a partir de UR_PINF (*subtree raising*).

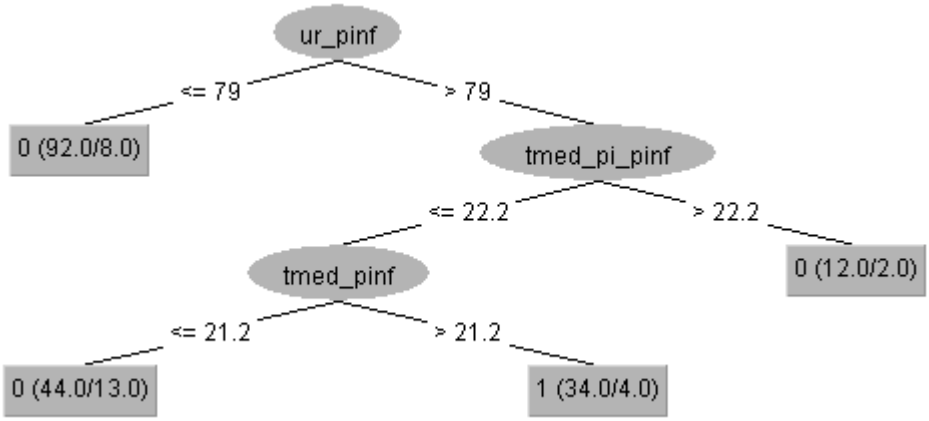

Figura 2. Árvore de decisão gerada com o Weka para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente, considerando o limite de 10 p.p. na taxa de infecção.

Com as podas, a árvore de decisão ficou mais simples e compacta do que a gerada com o Enterprise Miner. A estimativa de acurácia dos dois modelos, no entanto, foi igual (Tabela 1). A especificidade e a confiabilidade negativa foram melhores do que a sensitividade e a confiabilidade positiva, provavelmente em virtude da distribuição desbalanceada dos exemplos entre as classes no conjunto de treinamento (MEIRA et al., 2009).

4. CONCLUSÕES

As árvores de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente de frutos desenvolvidas com o software livre Weka tiveram desempenho semelhante às desenvolvidas com o SAS Enterprise Miner.

Os modelos de alerta induzidos pelas duas ferramentas para o limite de 5 p.p. na taxa de infecção da doença são praticamente iguais. O modelo induzido pelo Weka para o limite de 10 p.p. é mais simples e compacto que o gerado com o software proprietário.

A validação dos modelos daria maiores subsídios para a sua comparação e seria essencial no caso de se querer adotar tais modelos em uso real.

Agradecimentos: à Fundação Procafé pela concessão dos dados utilizados neste trabalho.

5. REFERÊNCIAS

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0**: step-by-step data mining guide. [Illinois]: SPSS, 2000. 78p.

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

MEIRA, C.A.A.; RODRIGUES, L.H.A.; MORAES, S.A. Modelos de alerta para o controle da ferrugem-do-cafeeiro em lavouras com alta carga pendente. **Pesquisa Agropecuária Brasileira**, v.44, n.3, p.233-242, 2009. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-204X2009000300003&lng=pt&nrm=iso>. Acesso em; 24 ago. 2009.

MEIRA, C.A.A. **Processo de descoberta de conhecimento em bases de dados para a análise e o alerta de doenças de culturas agrícolas e sua aplicação na ferrugem do cafeeiro**. 2008. 198p. Tese (Doutorado), Universidade Estadual de Campinas, Campinas.

MOLINEROS, J.E; DE WOLF, E.D.; FRANCL, L.; MADDEN, L.; LIPPS, P. Modeling epidemics of fusarium head blight: trials and tribulations. **Phytopathology**, v.95, p.S71. 2005.

MONARD, M.C.; BARANAUSKAS, J.A. Conceitos sobre aprendizado de máquina. In: REZENDE, S.O. (Org.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Manole, 2002. p.89-114.

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**, v.2, p.32-38, 1976.

PAUL, P.A.; MUNKVOLD, G.P. A model-based approach to preplanting risk assessment for gray leaf spot of maize. **Phytopathology**, v.94, p.1350-1357, 2004.

PINTO, A.C.S.; POZZA, E.A.; SOUZA, P.E.; POZZA, A.A.A.; TALAMINI, V.; BOLDINI, J.M.; SANTOS, F.S. Descrição da epidemia da ferrugem do cafeeiro com redes neuronais. **Fitopatologia Brasileira**, v.27, p.517-524, 2002.

SAS INSTITUTE INC. **Getting started with SAS® Enterprise Miner™ 4.3**. Cary, NC: SAS Institute Inc., 2004. 126p.

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

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, 2002. p.369-449.

ZAMBOLIM, L.; VALE, F.X.R.; PEREIRA, A.A.; CHAVES, G.M. Café (Coffea arabica L.): controle de doenças – doenças causadas por fungos, bactérias e vírus. In: VALE, F.X.R.; ZAMBOLIM, L. (Ed.). Controle de doenças de plantas: grandes culturas. Viçosa: UFV, v.1, 1997. p.83-139.