Modelos de alerta para o controle da ferrugem-do-cafeeiro em lavouras com alta carga pendente

Carlos Alberto Alves Meira[1], Luiz Henrique Antunes Rodrigues[2] e Sérgio Almeida de Moraes[3]

[1]Embrapa Informática Agropecuária, Caixa Postal 6041, CEP 13083-970 Campinas, SP. E-mail: carlos@cnptia.embrapa.br [2]Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola, Caixa Postal 6011, CEP 13083-875 Campinas, SP. E-mail: lique@agr.unicamp.br [3]Instituto Agronômico, Centro de Pesquisa e Desenvolvimento de Fitossanidade, Caixa Postal 28, CEP 13012-970 Campinas, SP. E-mail: sergioam@iac.sp.gov.br

Resumo – O objetivo deste trabalho foi desenvolver árvores de decisão como modelos de alerta da ferrugem-do-cafeeiro em lavouras de café (*Coffea arabica* L.) com alta carga pendente de frutos. 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 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. A acurácia do modelo para o limite de 5 pontos percentuais foi de 81%, por validação cruzada, chegando até 89% segundo estimativa otimista. Esse modelo apresentou bons resultados para outras medidas de avaliação importantes, como sensitividade (80%), especificidade (83%) e confiabilidades positiva (79%) e negativa (84%). O modelo para o limite de 10 pontos percentuais teve acurácia de 79%, e não apresentou o mesmo equilíbrio entre as demais medidas. Em conjunto, esses modelos podem auxiliar na tomada de decisão referente ao controle da ferrugem-do-cafeeiro no campo. A indução de árvores de decisão é alternativa viável às técnicas convencionais de modelagem e facilita a compreensão dos modelos.

Termos para indexação: *Coffea arabica*, *Hemileia vastatrix*, árvores de decisão, doença de plantas, previsão.

Warning models for coffee rust control in growing areas with large fruit load

Abstract – The objective of this work was to develop decision trees as warning models of coffee (*Coffea arabica* L.) rust in growing areas with large fruit load. 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 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 accuracy obtained by cross-validating the model to the limit of 5 percentage points was 81%, reaching up to 89% according to an optimistic estimate. This model showed good results for other important evaluation measures, such as sensitivity (80%), specificity (83%), positive reliability (79%), and negative reliability (84%). The model for the limit of 10 percentage points had a 79% accuracy and did not show the same balance among the other evaluation measures. Together, these two models may support the decisions about coffee rust control in the field. The decision tree induction is a viable alternative to conventional modeling techniques, thus facilitating the comprehension of the models.

Index terms: *Coffea arabica*, *Hemileia vastatrix*, decision trees, plant disease, prediction.

Introdução

A ferrugem (*Hemileia vastatrix* Berk. & Br.) é a principal doença do cafeeiro (*Coffea arabica* L.) em todo o mundo. No Brasil, em regiões onde as condições climáticas são favoráveis à doença, os prejuízos na produção atingem cerca de 35%, podendo chegar a mais de 50% (Zambolim et al., 1997). Além da importância econômica, a ferrugem atende outros requisitos que justificam o desenvolvimento de modelos de previsão ou de alerta, como a variação na sua intensidade entre as estações de cultivo e a disponibilidade de medidas de controle economicamente viáveis. Existem vários exemplos de modelos empíricos de previsão de ocorrência de ferrugem-do-cafeeiro. O ajuste dos dados observados a equações de regressão é a forma mais comum de modelagem (Kushalappa & Eskes, 1989; Zambolim et al., 2002).

Kushalappa et al. (1983) propuseram um modelo para explicar o curso de ação biológica de *H. vastatrix*. Valores obtidos por este modelo a partir de dados observados no campo foram utilizados no desenvolvimento de equações de regressão para predizer a taxa de progresso da ferrugem-do-cafeeiro (Kushalappa et al., 1984). Pinto et al. (2002) avaliaram o potencial de redes neurais para descrever a epidemia da ferrugem-do-cafeeiro. As redes neurais elaboradas podem ser utilizadas como modelos de previsão da doença. Também de maneira empírica, mas com uma abordagem qualitativa, Garçon et al. (2004) propuseram um modelo de previsão com base em valores de severidade da ferrugem-do-cafeeiro.

A indução de árvores de decisão é uma técnica de modelagem alternativa. As árvores de decisão, com sua representação simbólica e interpretável, permitem a compreensão das fronteiras de decisão que existem nos dados e da lógica implícita neles (Apte & Weiss, 1997). Redes neurais, embora possam ter alta precisão, são difíceis de ser compreendidas quando comparadas com as árvores de decisão (Fayyad et al., 1996; Monard & Baranauskas, 2002b). Diferentemente das técnicas de regressão, a multicolinearidade entre as variáveis independentes não afeta o desempenho das árvores de decisão (Butt & Royle, 1990; Hand et al., 2001). Além disso, diversas variáveis, numéricas ou categóricas, podem ser analisadas ao mesmo tempo, uma vez que o próprio algoritmo de indução se encarrega de selecionar aquelas de maior importância.

A árvore de decisão é um modelo representado graficamente por nós e ramos (Monard & Baranauskas, 2002b; Witten & Frank, 2005). O nó-raiz, no topo da estrutura, e os nós internos são nós de decisão. Cada um contém um teste sobre uma variável independente e os resultados desse teste formam os ramos da árvore. Os nós-folhas, nas extremidades, representam valores de predição da variável dependente ou distribuições de probabilidade desses valores.

Paul & Munkvold (2004) usaram este tipo de modelagem para predizer categorias de severidade da cercosporiose do milho em estágio avançado do cultivo. Árvores de decisão também modelaram epidemias de giberela do trigo, procurando predizer se a severidade da doença seria maior ou igual a 10% (Molineros et al., 2005). Baker et al. (1993) desenvolveram uma árvore de decisão para predizer o risco (alto ou baixo) de mortalidade de pínus em decorrência de podridão das raízes causada por *Heterobasidion annosum*.

O propósito da indução de árvores de decisão é descobrir a estrutura preditiva do problema ou produzir modelos de predição precisos. Com relação à primeira opção, Meira et al. (2008) analisaram epidemias da ferrugem-do-cafeeiro usando árvore de decisão.

O objetivo deste trabalho foi desenvolver modelos de alerta da ferrugem-do-cafeeiro para lavouras com alta carga pendente de frutos a partir de dados meteorológicos e do espaçamento entre plantas.

Material e Métodos

Os dados utilizados foram coletados por Japiassú et al. (2007) e se referem ao acompanhamento mensal da incidência da ferrugem-do-cafeeiro na fazenda experimental da Fundação Procafé, em Varginha, MG (21°34'0"S, 45°24'22"W), de outubro de 1998 a outubro de 2006. Em cada ano, no mês de setembro, foram selecionadas oito lavouras de café adultas em produção, com idade entre 6 e 20 anos, quatro lavouras em espaçamento largo (aproximadamente 3,5 m entre linhas e 0,7 m entre plantas – densidade média de 4.000 plantas por hectare) e quatro adensadas (aproximadamente 2,5 m entre linhas e 0,5 m entre plantas – densidade média de 8.000 plantas por hectare). Para cada espaçamento, foram escolhidas duas lavouras com alta carga pendente de frutos (acima de 30 sacas beneficiadas por hectare) e duas com baixa carga (abaixo de 10 sacas beneficiadas por hectare). Dessas duas lavouras, uma foi da cultivar Catuaí e a outra da cultivar Mundo Novo. Não houve controle da doença durante o ano agrícola nos talhões escolhidos. O processo de amostragem, realizado no final de cada mês, foi aquele recomendado por Chalfoun (1997). Dados meteorológicos, como temperatura do ar (média, máxima e mínima), precipitação pluvial e umidade relativa do ar, foram registrados a cada 30 min por uma estação meteorológica automática (marca Davis, modelo Groweather Industrial) instalada próximo dos locais de avaliação da incidência da ferrugem.

O progresso da ferrugem entre uma avaliação e a outra foi definido como a variável de interesse. Foram calculadas taxas de infecção para cada mês, subtraindo-se da incidência da doença no mês a incidência no mês anterior. Os valores numéricos das taxas de infecção foram mapeados para duas categorias ou classes. A variável dependente foi definida como uma taxa de infecção binária, com uma classe a menos que a taxa

de infecção categórica usada na análise da epidemia da ferrugem-do-cafeeiro (Meira et al., 2008). O objetivo foi procurar obter modelos menos complexos, com menor quantidade de nós e com melhores valores das medidas de avaliação.

A primeira opção de taxa de infecção binária foi estabelecida com a criação da variável 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 (Tabela 1). Como opção, foi criada a variável TAXA_INF_M10, com valor 1 para taxas maiores ou iguais a 10 p.p. e valor 0 no caso contrário. A fronteira de decisão em 5 p.p. foi determinada com base no limite de 5% de incidência da ferrugem-do-cafeeiro recomendado por Zambolim et al. (1997) para o controle da doença por via foliar. A fronteira de decisão em 10 p.p. foi determinada com base em Kushalappa et al. (1984), que propuseram o limite de risco de 10% de incidência para recomendar aplicação de fungicida. Deve-se ressaltar que este valor está próximo do limite máximo de 12% de folhas doentes recomendado para a aplicação de fungicidas sistêmicos (Zambolim et al., 1997).

Outra alteração em relação à modelagem realizada por Meira et al. (2008) foi separar a geração dos modelos por carga pendente de frutos. Os principais motivos foram simplificar os modelos e particularizar seu uso de acordo com a característica bianual dos cafezais, que nos anos de alta carga pendente estão mais predispostos ao ataque da ferrugem que nos anos de baixa carga. Neste trabalho, são discutidos os modelos para lavouras com alta carga pendente. Todos os modelos de alerta gerados podem ser consultados em Meira (2008).

As variáveis independentes meteorológicas foram construídas a partir do nível horário (registros da estação), passando pelo nível diário até chegar a um nível que permitisse a análise de sua relação com a variável dependente. No nível diário, além de médias e somatórias das variáveis meteorológicas, foram calculados valores estimados de molhamento foliar prolongado (mínimo de 6 horas), porque a germinação dos uredósporos de *H. vastatrix* só ocorre se a folha estiver molhada. O tempo mínimo necessário para a ocorrência de infecção foi avaliado em 6 horas de água livre na superfície da folha (Kushalappa et al., 1983). O número de horas com alta umidade relativa do ar (acima de 95%) foi utilizado como medida indireta de molhamento foliar contínuo (Sutton et al., 1984). Em dias com períodos de molhamento disjuntos, foi considerado o maior período, com tolerância de até 1 hora entre eles, para uni-los em um único período. Os períodos de molhamento foliar foram analisados tanto na extensão total como na fração noturna (das 20h às 8h), já que a infecção ocorre preferencialmente com pouca ou nenhuma luminosidade (Montoya & Chaves, 1974). O intervalo considerado como um dia se estendeu das 12h de um dia até 12h do dia seguinte, pois os períodos de molhamento ocorrem geralmente entre um dia e outro. A temperatura média durante o período de molhamento foliar também foi calculada para cada dia, uma vez que é o fator principal que determina o percentual de germinação dos esporos e de penetração enquanto a superfície da folha está molhada (Kushalappa et al., 1983). Os dias com precipitação

Tabela 1. Relação das variáveis usadas na indução dos modelos de alerta da ferrugem-do-cafeeiro em lavouras com alta carga pendente de frutos.

Variáveis	Tipo	Medida	Significado
Variáveis dependentes			
TAXA_INF_M5	Binário	-	Taxa de infecção binária: 1 para taxas maiores ou iguais a 5 pontos percentuais; 0, caso contrário.
TAXA_INF_M10	Binário	-	Taxa de infecção binária: 1 para taxas maiores ou iguais a 10 pontos percentuais; 0, caso contrário.
Variáveis independentes			
DCHUV_PINF	Numérico	dias	Número de dias chuvosos (precipitação ≥1 mm) no período de infecção (PINF).
LAVOURA	Binário	-	Espaçamento: lavoura ADENSADA ou LARGA.
MED_PRECIP_PINF	Numérico	mm	Média das precipitações pluviais diárias no PINF.
NHNUR95_PINF	Numérico	h	Média diária do número de horas noturnas com umidade relativa do ar ≥95% no PINF.
NHUR95_PINF	Numérico	h	Média diária do número de horas com umidade relativa do ar ≥95% no PINF.
PRECIP_PINF	Numérico	mm	Precipitação pluvial acumulada no PINF.
THUR95_PINF	Numérico	ºC	Temperatura média diária durante as horas com umidade relativa do ar ≥95% no PINF.
TMAX_PINF	Numérico	ºC	Média das temperaturas máximas diárias no PINF.
TMAX_PI_PINF	Numérico	ºC	Média das temperaturas máximas diárias no período de incubação para os dias do PINF.
TMED_PINF	Numérico	ºC	Média das temperaturas médias diárias no PINF.
TMED_PI_PINF	Numérico	ºC	Média das temperaturas médias diárias no período de incubação para os dias do PINF.
TMIN_PINF	Numérico	ºC	Média das temperaturas mínimas diárias no PINF.
TMIN_PI_PINF	Numérico	ºC	Média das temperaturas mínimas diárias no período de incubação para os dias do PINF.
UR_PINF	Numérico	%	Umidade relativa do ar média diária no PINF.

maior ou igual a 1 mm foram considerados chuvosos segundo o mesmo critério usado por Kushalappa et al. (1983).

Na sequência da preparação dos dados meteorológicos, cada dia foi tratado como um eventual dia de infecção. Considerando um período de incubação estimado, de acordo com a equação proposta por Moraes et al. (1976), cada dia foi associado ao mês correspondente de avaliação da incidência da ferrugem e, consequentemente, à taxa de infecção para a qual possivelmente teve parcela de contribuição (Meira et al., 2008). O conjunto de dias associado a uma taxa de infecção foi denominado de período de infecção (PINF). As variáveis independentes meteorológicas usadas na modelagem foram derivadas para cada um desses períodos de infecção (Tabela 1). O espaçamento da lavoura (lavoura adensada ou larga) completou o conjunto das variáveis independentes.

A preparação dos dados foi feita por meio de programas de computador escritos na linguagem de programação Perl (ActivePerl versão 5.8.7, ActiveState Corp.). Alguns procedimentos finais foram realizados com o SAS Enterprise Miner, versão 4.3 (SAS Institute, 2004). 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.

As árvores de decisão foram geradas usando a ferramenta "Decision Tree" do SAS Enterprise Miner e depois foram visualizadas e analisadas usando a ferramenta SAS Enterprise Miner Tree Desktop Application, versão 9.1.32 (SAS Institute, 2007). O algoritmo de indução constrói uma árvore de decisão de forma recursiva, de cima para baixo. Inicia com o conjunto de treinamento, que é dividido de acordo com um teste sobre uma das variáveis independentes, formando subconjuntos mais homogêneos em relação à variável dependente. Esse procedimento é repetido até que se consiga conjuntos de exemplos bem homogêneos, para os quais seja possível atribuir uma única classe. O critério de escolha da variável que divide o conjunto de exemplos em cada repetição foi o ganho de informação (Witten & Frank, 2005), relacionado com a redução da entropia dos exemplos.

Cada árvore de decisão foi escolhida para ser binária, com dois ramos a partir dos nós internos. Para evitar que o modelo ficasse muito específico para o conjunto de treinamento ("overfitting"), o que comprometeria a sua generalização e o desempenho com novos exemplos, foram adotadas duas regras de parada do algoritmo de indução. A primeira regra limitou a profundidade da árvore, permitindo que ela tivesse no máximo seis níveis – o nó-raiz foi considerado o nível zero. A segunda regra limitou a fragmentação do conjunto de treinamento, requerendo um mínimo de 10 exemplos em cada nó para a busca de uma nova divisão e de pelo menos cinco exemplos em cada nó-folha. Além das regras de parada, denominadas de pré-poda, foi realizado um procedimento de pós-poda. Além da árvore de decisão completa, foram avaliadas todas as suas possíveis subárvores e foi escolhida a menor subárvore (menor complexidade), com a menor taxa de erro.

A acurácia e a taxa de erro são as medidas de avaliação mais comuns para modelos de classificação (Witten & Frank, 2005). São estimativas dos percentuais de acertos e de erros do modelo na predição da classe de novos exemplos. Essas medidas foram calculadas a partir da matriz de confusão, que também oferece meios efetivos para a avaliação de um classificador (Monard & Baranauskas, 2002a). Para um problema com duas classes, denominadas classe positiva e classe negativa, a matriz de confusão indica as quatro possibilidades de acertos e de erros do classificador: verdadeiros positivos (VP), quando os exemplos da classe positiva foram preditos corretamente; falsos negativos (FN), quando os exemplos da classe positiva foram preditos como da classe negativa; verdadeiros negativos (VN), quando os exemplos da classe negativa foram preditos corretamente; e falsos positivos (FP), quando os exemplos da classe negativa foram preditos como da classe positiva. Outras medidas de avaliação também foram derivadas da matriz de confusão (Monard & Baranauskas, 2002a): sensitividade ou precisão da classe positiva (1); especificidade ou precisão da classe negativa (0); confiabilidade positiva; e confiabilidade negativa. As equações de cálculo das medidas de avaliação são as seguintes:

$$\text{Acurácia} = (VP + VN)/n \quad (1)$$

$$\text{Taxa de erro} = (FP + FN)/n = 1 - \text{Acurácia} \quad (2)$$

$$\text{Sensitividade} = VP/(VP + FN) \quad (3)$$

$$\text{Especificidade} = VN/(VN + FP) \quad (4)$$

$$\text{Confiabilidade positiva} = VP/(VP + FP) \quad (5)$$

$$\text{Confiabilidade negativa} = VN/(VN + FN) \quad (6),$$

em que n é o número de exemplos no conjunto de treinamento.

As medidas de avaliação foram estimadas pelos métodos de ressubstituição e de validação cruzada.

A ressubstituição consiste em construir o modelo e avaliar o seu desempenho no mesmo conjunto de exemplos, ou seja, o conjunto de teste é idêntico ao conjunto de treinamento (Monard & Baranauskas, 2002a). Normalmente, resulta em estimativas otimistas devido à especialização do modelo em relação aos exemplos. A validação cruzada é uma forma de evitar esse viés otimista e é usada principalmente quando a quantidade de dados para dividir entre treinamento e teste é limitada (Witten & Frank, 2005). Na validação cruzada, os exemplos são aleatoriamente divididos em k partições mutuamente exclusivas, de tamanho aproximadamente igual. Uma das partições é reservada para teste, enquanto as demais, juntas, são usadas para treinamento. Este procedimento é executado k vezes, cada vez com uma partição diferente para teste. As medidas de avaliação são calculadas como a média das medidas obtidas em cada uma das partições de teste. A vantagem é usar cada um dos exemplos tanto para treinamento quanto para teste. A validação cruzada foi realizada com 10 partições (k = 10), pois testes extensivos em muitos e diferentes conjuntos de dados mostraram que 10 é um valor próximo do número exato de partições para se obter as melhores estimativas (Witten & Frank, 2005). Além de aleatórias, as partições foram estratificadas, e a distribuição dos exemplos de cada classe ficou parecida com a do conjunto de treinamento completo.

Resultados e Discussão

Os alertas da ferrugem-do-cafeeiro são emitidos quando a taxa de infecção da doença, prevista para o prazo de um mês, atinge ou ultrapassa 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.

O modelo de alerta considerando o limite de 5 p.p. (MA-TI5) é apresentado na Figura 1. O nó-raiz (nó 1) indica a distribuição percentual dos exemplos entre as duas classes no conjunto de treinamento: 46% (84 exemplos) da

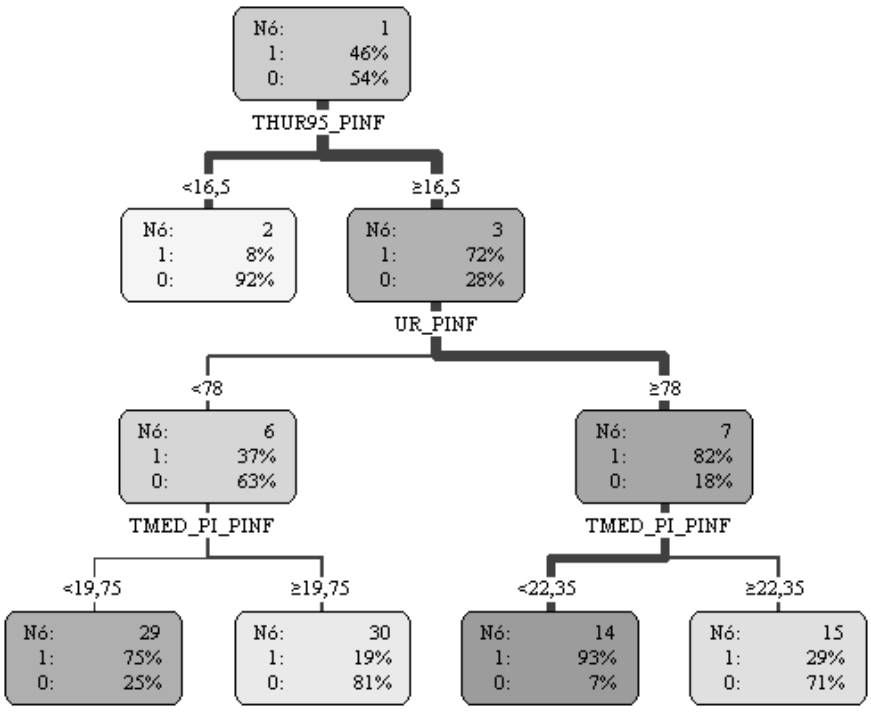

Figura 1. Árvore de decisão para alerta da ferrugem-do-cafeeiro em lavouras com alta carga pendente de frutos, considerando o limite de 5 pontos percentuais na taxa de infecção (MA-TI5). Os nós da árvore indicam o seu número identificador e a distribuição de probabilidade das duas classes. A classe de predição em um nó-folha é a que apresenta a maior probabilidade. Identificação das variáveis na Tabela 1.

classe 1 e 54% (98 exemplos) da classe 0. Os nós da árvore de decisão são coloridos em tons de cinza com base na proporção de exemplos da classe 1 – quanto mais escuro, maior a proporção. Cada ramo tem espessura proporcional à quantidade de exemplos classificados através do caminho de decisão que passa pelo ramo. A regra de divisão original do nó 6, com base na variável MED_PRECIP_PINF (divisão em 1,75 mm), teve regra concorrente com mesmo ganho de informação com base em TMED_PI_PINF (divisão em 19,75ºC). Esta última foi escolhida para o modelo final devido à melhor distribuição mensal dos exemplos classificados.

A árvore de decisão (Figura 1) estabelece que os alertas devem ser emitidos quando as condições das variáveis meteorológicas entre o nó-raiz e os nós-folhas 14 ou 29 forem satisfeitas. Caso contrário, é provável que a taxa de infecção da ferrugem fique abaixo de 5 p.p. O caminho de decisão entre o nó-raiz e um nó-folha pode ser traduzido para uma regra na forma 'SE <*condição*> ENTÃO <*decisão*>'. Por exemplo, o caminho até o nó 14 se traduz na regra 'SE (THUR95_PINF ≥ 16,5ºC) e (UR_PINF ≥ 78%) e (TMED_PI_PINF < 22,35ºC) ENTÃO TAXA_INF_M5 = 1'.

Segundo o modelo MA-TI5, as influências da temperatura e da umidade relativa do ar se revelaram mais importantes no progresso da ferrugem-do-cafeeiro. A participação da temperatura dividiu-se na sua influência no processo de germinação (THUR95_PINF) e durante o período de incubação (TMED_PI_PINF). Temperaturas médias mais baixas durante o molhamento foliar foram desfavoráveis às taxas de infecção maiores ou iguais a 5 p.p. (nó 2, Figura 1). Montoya & Chaves (1974) indicaram que o ponto mínimo de germinação de *H. vastatrix* seria encontrado em temperaturas inferiores a 18ºC e Kushalappa et al. (1983) consideraram 14ºC como limite mínimo de atividade do patógeno. Temperaturas médias diárias mais elevadas no período de incubação também exerceram efeito negativo nas taxas de infecção da ferrugem-do-cafeeiro (nós 15 e 30). Segundo Moraes et al. (1976), temperaturas médias máximas elevadas no período de incubação ocasionaram efeito depressivo sobre o desenvolvimento de *H. vastatrix*.

A umidade relativa média diária (UR_PINF) mais alta foi favorável às taxas de infecção maiores ou iguais a 5 p.p. (nó 7, Figura 1). A umidade relativa parece ter expressado melhor a importância das chuvas que as variáveis independentes relacionadas com a precipitação (Meira et al., 2008). Dependendo das condições para a infecção, menos favoráveis (nó 6) ou mais favoráveis (nó 7), a influência da temperatura no período de incubação foi diferenciada. Meira et al. (2008) observaram que infecções ocorridas em condições menos favoráveis foram mais sensíveis ao efeito da temperatura durante o período de incubação.

O modelo da Figura 1 é simples e compacto, com cinco regras e com base apenas em três variáveis de teste. O modelo foi avaliado com acurácia estimada em 81%, podendo alcançar 89% pela estimativa mais otimista (Tabela 2). A estimativa da acurácia é importante, pois dá a noção da proporção de acertos que o modelo pode ter caso venha a ser aplicado ao problema real. As outras medidas de avaliação 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, por sua vez, 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 MA-TI5 entre as medidas de avaliação (Tabela 2) é positivo. As matrizes de confusão desse modelo, obtidas pela ressubstituição e pela validação cruzada, estão apresentadas na Tabela 3.

As regras mais importantes do modelo MA-TI5 foram as correspondentes aos nós-folhas 2 ('SE THUR95_PINF < 16,5ºC ENTÃO TAXA_INF_M5 = 0') e 14 ['SE (THUR95_PINF ≥ 16,5ºC) e (UR_PINF ≥ 78%) e (TMED_PI_PINF < 22,35ºC) ENTÃO TAXA_INF_M5 = 1'], conforme apresentado na Figura 1. A primeira

Tabela 2. Avaliação dos modelos de alerta da ferrugem-do-cafeeiro em lavouras com alta carga pendente de frutos pelos métodos de validação cruzada e ressubstituição[(1)].

Medida de avaliação	Modelo MA-TI5		Modelo MA-TI10	
	Validação cruzada (%)	Ressubstituição (%)	Validação cruzada (%)	Ressubstituição (%)
Acurácia	81,3±4,8	89,0	79,2±3,0	89,6
Taxa de erro	18,7±4,8	11,0	20,8±3,0	10,4
Sensitividade	79,9±6,9	84,5	70,3±4,6	71,7
Especificidade	82,6±4,3	92,9	82,8±3,3	96,9
Confiabilidade positiva	79,4±6,1	91,0	65,3±4,1	90,5
Confiabilidade negativa	83,9±4,7	87,5	86,9±2,4	89,3

[(1)]Modelos MA-TI5 e MA-TI10, modelos de alerta considerando os limites de 5 e 10 pontos percentuais (p.p.), respectivamente, na taxa de infecção da ferrugem-do-cafeeiro.

cobriu corretamente a maior parte dos exemplos da classe 0 (VNs), enquanto a segunda cobriu corretamente a maioria dos exemplos da classe 1 (VPs). Vários exemplos dos meses de dezembro a abril, período crítico de evolução da ferrugem, foram cobertos corretamente pela regra referente ao nó 14 (48 VPs e apenas 4 FPs).

O modelo de alerta considerando o limite de 10 p.p. (MA-TI10) é apresentado na Figura 2. Com esse nível de taxa de infecção, a distribuição dos exemplos entre as duas classes no conjunto de treinamento foi mais desigual, conforme indica o nó-raiz (nó 1): 29% (53 exemplos) da classe 1 e 71% (129 exemplos) da classe 0.

Até o terceiro nível de profundidade do modelo MA-TI10, as variáveis de teste são idênticas e as fronteiras de decisão são praticamente as mesmas do modelo MA-TI5. A ramificação a partir do nó 12 (Figura 2) é a diferença principal entre os dois modelos. Essa ramificação identificou melhor os exemplos das duas classes e resultou na menor taxa de erro para o modelo MA-TI10. As combinações introduzidas de temperatura média diária (TMED_PINF), média diária de molhamento foliar (NHUR95_PINF) e média diária de precipitação (MED_PRECIP_PINF) estão coerentes com aspectos epidemiológicos conhecidos da ferrugem-do-cafeeiro.

A avaliação do modelo MA-TI10 foi satisfatória (Tabela 2). A acurácia de 79% (validação cruzada) é maior que a acurácia de 71% de um classificador que atribuísse a todos os exemplos a classe majoritária 0. A especificidade (83%) e a confiabilidade negativa (87%) foram melhores que a sensitividade (70%) e a confiabilidade positiva (65%), provavelmente em decorrência da distribuição desbalanceada dos exemplos entre as classes no conjunto de treinamento. As matrizes de confusão do modelo MA-TI10, segundo a ressubstituição e a validação cruzada, estão apresentadas na Tabela 3.

Como ocorreu para o modelo MA-TI5, a regra referente ao nó 2 (Figura 2) se destacou em relação aos VNs. Os VPs foram mais distribuídos no modelo MA-TI10, destacando-se as regras correspondentes aos nós 24 e 28, em que todos os exemplos foram cobertos corretamente, no período de janeiro a maio.

Segundo Zambolim et al. (2002), nos anos de alta carga pendente de frutos não são recomendadas atomizações tardias após a constatação de nível de incidência da ferrugem-do-cafeeiro maior que 5%. O modelo MA-TI5 seria, então, o mais indicado no suporte à decisão dos momentos oportunos para a adoção de medidas de controle da doença. O modelo MA-TI10 poderia servir como instrumento adicional, alertando que as medidas de controle deveriam ser urgentes e/ou mais eficazes, pois as condições estariam propícias a um desenvolvimento ainda mais acelerado da doença.

Em Kushalappa et al. (1984), a melhor equação de regressão explicou 76% da variação na taxa de infecção da ferrugem-do-cafeeiro (coeficiente de determinação igual a 0,76). As redes neurais desenvolvidas por Pinto et al. (2002) tiveram a incidência da ferrugem-do-cafeeiro como variável de saída. A rede neural com o melhor desempenho apresentou erro médio de previsão igual a 1,17% e quadrado médio dos desvios igual a 3,95. Tais medidas de avaliação, distintas das utilizadas no presente trabalho, são adequadas para problemas em que a variável dependente do modelo é numérica e contínua, diferente da abordagem adotada neste trabalho.

Comparando a outras aplicações de árvores de decisão como modelos de predição de doenças de plantas (Baker et al., 1993; Paul & Munkvold, 2004; Molineros et al., 2005), os resultados obtidos neste trabalho foram promissores, uma vez que os maiores valores relatados de acurácia não ultrapassaram 80%. A acurácia dos modelos apresentados, conforme se

Tabela 3. Matrizes de confusão dos modelos de alerta da ferrugem-do-cafeeiro em lavouras com alta carga pendente de frutos obtidas pelos métodos de validação cruzada e ressubstituição.

Método	Verdadeira[1]	Predita[2]	Ocorrência
		Modelo MA-TI5	
Validação cruzada	1	1	67
	1	0	17
	0	1	17
	0	0	81
Ressubstituição	1	1	71
	1	0	13
	0	1	7
	0	0	91
		Modelo MA-TI10	
Validação cruzada	1	1	37
	1	0	16
	0	1	22
	0	0	107
Ressubstituição	1	1	38
	1	0	15
	0	1	4
	0	0	125

[1]Classe verdadeira no conjunto de dados de treinamento. [2]Classe predita pelo modelo.

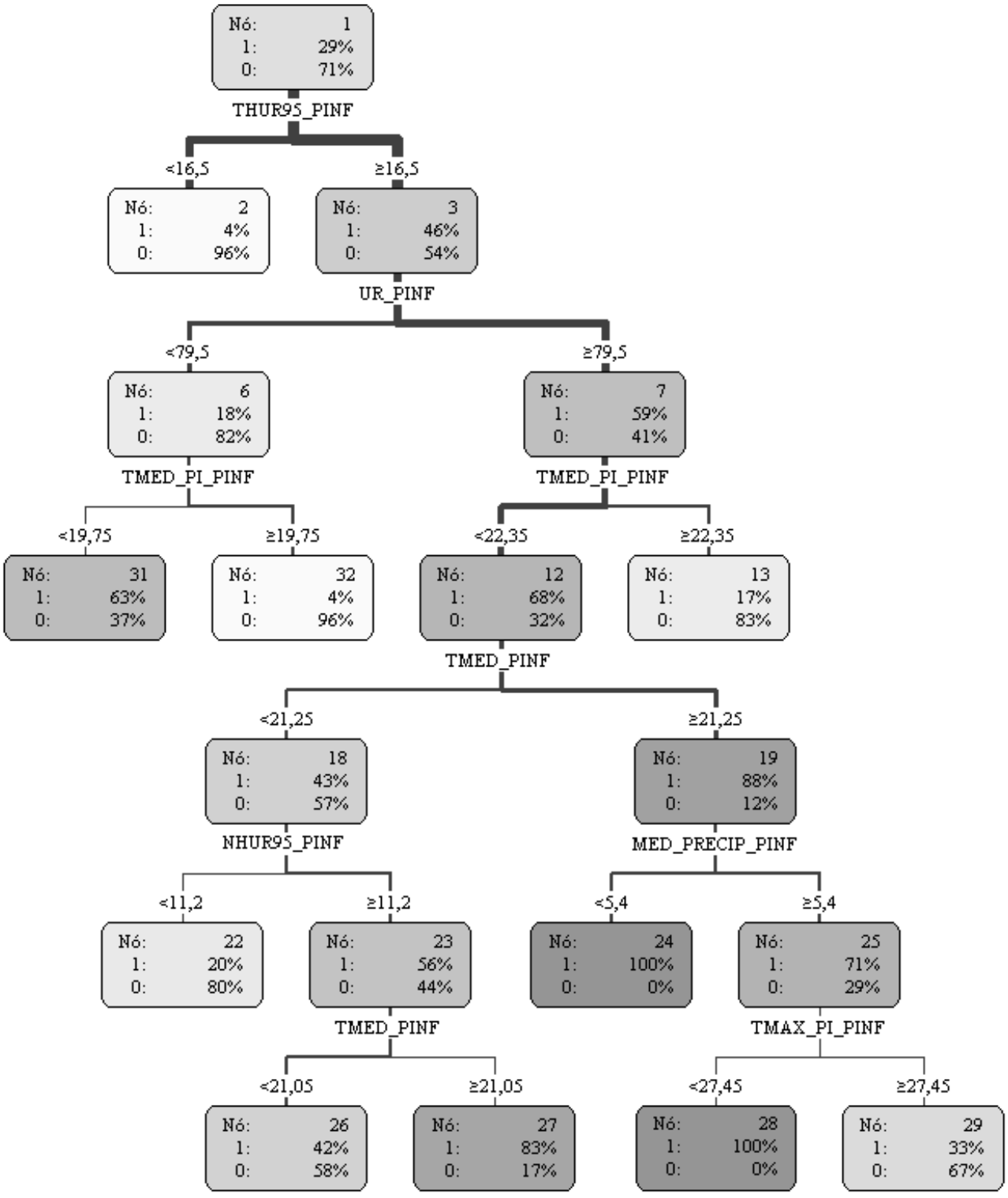

Figura 2. Árvore de decisão para alerta da ferrugem-do-cafeeiro em lavouras com alta carga pendente de frutos, considerando o limite de 10 pontos percentuais na taxa de infecção (MA-TI10). Os nós da árvore indicam o seu número identificador e a distribuição de probabilidade das duas classes. A classe de predição em um nó-folha é a que apresenta a maior probabilidade. Identificação das variáveis na Tabela 1.

esperava, foi superior à do modelo usado por Meira et al. (2008) na análise da epidemia da ferrugem-do-cafeeiro, que obteve 73% de acurácia na validação cruzada.

A principal limitação dos modelos de alerta desenvolvidos está relacionada à sua abrangência. O uso desses modelos deve ficar restrito à região onde os dados foram coletados ou a regiões com climas parecidos. Regiões com diferentes climas podem apresentar condições meteorológicas que não foram representadas nos dados analisados e que podem condicionar o progresso da ferrugem-do-cafeeiro de maneira diferente da capturada pelos modelos. Cabe ressaltar que, antes da adoção de qualquer dos modelos apresentados, é importante que seja realizada uma etapa de validação.

Conclusão

As árvores de decisão para alerta da ferrugem-do-cafeeiro em lavouras com alta carga pendente de frutos dão base para a decisão sobre as medidas a serem adotadas para o controle da doença e o melhor momento de implementá-las.

Agradecimentos

À Fundação Procafé, pela concessão dos dados relacionados com o monitoramento da incidência da ferrugem-do-cafeeiro; ao SAS Brasil, pela concessão da licença de uso do SAS Enterprise Miner por meio de seu Programa Acadêmico.

Referências

APTE, C.; WEISS, S. Data mining with decision trees and decision rules. **Future Generation Computer Systems**, v.13, p.197-210, 1997.

BAKER, F.A.; VERBYLA, D.L.; HODGES, C.S.; ROSS, E.W. Classification and regression tree analysis for assessing hazard of pine mortality caused by *Heterobasidion annosum*. **Plant Disease**, v.77, p.136-139, 1993.

BUTT, D.J.; ROYLE, D.J. Multiple regression analysis in the epidemiology of plant diseases. In: KRANZ, J. (Ed.). **Epidemics of plant diseases**: mathematical analysis and modeling. 2nd ed. Berlin: Springer-Verlag, 1990. p.143-180.

CHALFOUN, S.M. **Doenças do cafeeiro**: importância, identificação e métodos de controle. Lavras: Ufla/Faepe, 1997. 96p.

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

GARÇON, C.L.P.; ZAMBOLIM, L.; MIZUBUTI, E.S.G.; VALE, F.X.R. do; COSTA, H. Controle da ferrugem do cafeeiro com base no valor de severidade. **Fitopatologia Brasileira**, v.29, p.486-491, 2004.

HAND, D.J.; MANNILA, H.; SMYTH, P. **Principles of data mining**. Cambridge: MIT Press, 2001. 546p.

JAPIASSÚ, L.B.; GARCIA, A.W.R.; MIGUEL, A.E.; CARVALHO, C.H.S.; FERREIRA, R.A.; PADILHA, L.; MATIELLO, J.B. Influência da carga pendente, do espaçamento e de fatores climáticos no desenvolvimento da ferrugem do cafeeiro. In: SIMPÓSIO DE PESQUISA DOS CAFÉS DO BRASIL, 5., 2007, Águas de Lindóia. **Anais**. Brasília: Embrapa Café, 2007. 1 CD-ROM.

KUSHALAPPA, A.C.; AKUTSU, M.; LUDWIG, A. Application of survival ratio for monocyclic process of *Hemileia vastatrix* in predicting coffee rust infection rates. **Phytopathology**, v.73, p.96-103, 1983.

KUSHALAPPA, A.C.; AKUTSU, M.; OSEGUERA, S.H.; CHAVES, G.M.; MELLES, C. Equations for predicting the rate of coffee rust development based on net survival ratio for monocyclic process of *Hemileia vastatrix*. **Fitopatologia Brasileira**, v.9, p.255-271, 1984.

KUSHALAPPA, A.C.; ESKES, A.B. Advances in coffee rust research. **Annual Review of Phytopatology**, v.27, p.503-531, 1989.

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.

MEIRA, C.A.A.; RODRIGUES, L.H.A.; MORAES, S.A. Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. **Tropical Plant Pathology**, v.33, p.114-124, 2008.

MOLINEROS, J.; DE WOLF, E.; FRANCL, L.; MADDEN, L.; LIPPS, P. Modeling epidemics of fusarium head blight: trials and tribulations. **Phytopathology**, v.95, p.71, 2005.

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

MONARD, M.C.; BARANAUSKAS, J.A. Indução de regras e árvores de decisão. In: REZENDE, S.O. (Org.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Manole, 2002b. p.115-139.

MONTOYA, R.H.; CHAVES, G.M. Influência da temperatura e da luz na germinação, infectividade e período de geração de *Hemileia vastatrix* Berk. & Br. **Experientiae**, v.18, p.239-266, 1974.

MORAES, S.A.; SUGIMORI, M.H.; RIBEIRO, I.J.A.; ORTOLANI, A.A.; PEDRO JÚNIOR, 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. de; 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.

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

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, 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, 1997. v.1. p.83-139.

Recebido em 6 de novembro de 2008 e aprovado em 27 de fevereiro de 2009