MODELOS DE PREDIÇÃO DA FERRUGEM DO CAFEEIRO (*Hemileia vastatrix* Berkeley & Broome) POR TÉCNICAS DE MINERAÇÃO DE DADOS

Cesare Di Girolamo Neto[1], Luiz Henrique Antunes Rodrigues[2], Carlos Alberto Alves Meira[3]

(Recebido: 4 de outubro de 2013; aceito: 17 de dezembro de 2013)

RESUMO: A ferrugem é a principal doença do cafeeiro, podendo gerar perdas significativas na produção, caso medidas de controle não sejam adotadas. Modelos de alerta de doenças agrícolas são capazes de gerar informações para aplicações de defensivos somente quando necessário, reduzindo gastos por parte do produtor e impactos ambientais. Objetivou-se, neste trabalho, desenvolver, comparar e selecionar modelos de alerta baseados em técnicas de mineração de dados para a predição da ferrugem do cafeeiro, em anos de alta e baixa carga pendente de frutos. Foram utilizados dados obtidos em lavouras de café em produção, ao longo de 13 anos (1998-2011). Vinte e três atributos foram considerados como variáveis independentes (preditoras) e, como variável dependente, a taxa de progresso mensal da ferrugem do cafeeiro, obtida a partir de dados de incidência da doença. Os atributos mais importantes do conjunto de dados foram filtrados por métodos de seleção de atributos e a modelagem foi realizada por meio de quatro técnicas de mineração de dados: máquinas de vetores suporte, redes neurais artificiais, árvores de decisão e florestas aleatórias. Para anos de alta e baixa carga pendente de frutos, as melhores taxas de acerto foram 85,3% e 88,9%, respectivamente. Outras medidas de desempenho como sensitividade e especificidade também apresentaram valores altos e equilibrados. Os modelos desenvolvidos neste trabalho fornecem melhores subsídios para o monitoramento da doença, em anos de alta carga pendente de frutos do que outros modelos existentes, além de prover uma possibilidade de monitoramento, em anos de baixa carga pendente de frutos.

Termos para indexação: Alerta de doenças, florestas aleatórias, máquinas de vetores suporte, redes neurais artificiais, árvores de decisão.

WARNING MODELS FOR COFFEE RUST (*Hemileia vastatrix* Berkeley & Broome) BY DATA MINING TECHNIQUES

ABSTRACT: *Coffee rust can cause severe yield losses if control measures are not taken. Warning models are capable of generating useful information regarding to the application of fungicides, decreasing economic losses and environmental impacts. The aim of this study was to develop, compare and select warning models developed by data mining techniques in order to predict the coffee rust in years of high and low fruit load. For 13 years (1998-2011), data was collected from an automatic weather station. The independent variables were 23, obtained from the weather station, and the dependent variable was the monthly progress rate for the coffee rust, which was generated by the values of disease incidence. The most important features were refined by feature selection techniques, and the modeling was performed using four data mining techniques: support vector machines, artificial neural networks, decision trees and random forests. For high fruit load years the best accuracy was 85.3% and for low fruit load years it was 88.9%. Other performance measures like recall and specificity also had high and balanced values. The warning models developed on this study provide further information for monitoring the disease on high fruit load years than other models previously developed, and also provide a possibility for the monitoring on years of low fruit load.*

Index terms: *Predictive models, random forest, support vector machines, artificial neural networks, decision trees.*

1 INTRODUÇÃO

O café possui grande importância econômica para o Brasil, uma vez que o País é responsável por cerca de 37% da produção mundial e suas exportações representam um lucro de US$ 5,7 bilhões/ano (United States Department of Agriculture - USDA, 2013). A ferrugem do cafeeiro, causada pelo fungo *Hemileia vastatrix* Berk. & Br., tem alto potencial de dano e pode causar perdas de até 50% na produção, caso medidas de controle não sejam adotadas (ZAMBOLIM et al., 2002). Pela bienalidade do cafeeiro, a doença é mais agressiva em anos de alta carga pendente de frutos do que em anos com baixa carga pendente de frutos.

Estas observações evidenciam a importância de pesquisas capazes de gerar informações que possam ser utilizadas para fornecer suporte ao controle da doença, o que pode ser realizado por modelos de predição ou alerta. Um modelo dessa natureza procura antecipar quando uma doença pode atingir um nível crítico (HARDWICK, 2006).

[1]Universidade Estadual de Campinas/UNICAMP - Faculdade de Engenharia Agrícola (FEAGRI) - Cx. P. 6011 - Campinas SP 13083-875 - cesare.neto@gmail.com
[2]Universidade Estadual de Campinas/UNICAMP - Faculdade de Engenharia Agrícola (FEAGRI) - Cx. P. 6011 - Campinas - SP 13083-875 - lique@feagri.unicamp.br
[3]Embrapa Informática Agropecuária - Cx. P. 6041 - Campinas - SP - 13083-886 - carlos.meira@embrapa.br

Uma predição correta pode evitar aplicações desnecessárias de defensivos agrícolas, reduzindo gastos por parte do produtor na compra e na mão de obra para sua aplicação, além de diminuir impactos ambientais.

A partir da divulgação no meio científico da mineração de dados, por Fayyad, Piatetsky-Shapiro e Smyth (1996), tem-se notado um aumento no número de modelos de alerta desenvolvidos por meio dessa metodologia. Nesse sentido, Batchelor, Yang e Tschanz (1997) avaliaram a severidade da ferrugem asiática da soja, por meio de redes neurais artificiais; Paul e Munkvold (2004, 2005) utilizaram, respectivamente, árvores de decisão e redes neurais artificiais para predizer a severidade da cercosporiose do milho, em estágios avançados de cultivo; Molineiros et al. (2005) utilizaram árvores de decisão para modelar epidemias de giberela do trigo; Souza et al. (2013) utilizaram árvores de decisão para avaliar as condições de ocorrência da cercosporiose, em lavouras convencionais e orgânicas de café.

A epidemia da ferrugem do cafeeiro foi estudada por Pinto et al. (2002), com o uso de redes neurais artificiais, e por Meira, Rodrigues e Moraes (2008), por meio de árvores de decisão. Posteriormente, para determinar a taxa de progresso mensal da ferrugem do cafeeiro, para lavouras com alta carga pendente de frutos, Meira, Rodrigues e Moraes (2009) desenvolveram modelos de alerta em árvores de decisão e Cintra et al. (2011), modelos de alerta em árvores de decisão fuzzy. Outras técnicas também foram utilizadas para a modelagem da ferrugem do cafeeiro, como redes neurais artificiais fuzzy, desenvolvidas por Alves et al. (2010); máquinas de vetores suporte, induzidas por Luaces et al. (2011); florestas aleatórias, que também foram mencionadas no trabalho de Cintra et al. (2011); e redes bayesianas, geradas por Pérez-Ariza, Nicholson e Flores (2012).

A escolha pelo uso de uma ou outra técnica de modelagem requer a análise do problema em questão. Árvores de decisão possuem representação simbólica e interpretável, facilitando compreender quais as variáveis e suas interações que conduzem ao fenômeno estudado (HAN; KAMBER; PEI, 2011). A principal vantagem das redes neurais artificiais é resolver problemas que requerem solução complexa (HAYKIN, 2009), enquanto que as máquinas de vetores suporte podem gerar modelos com melhor taxa de acerto do que as redes neurais artificiais (LEE; TO, 2010).

As florestas aleatórias evitam sobreajuste (overfitting) e são pouco sensíveis a ruídos (BREIMAN, 2001).

Objetivou-se, neste trabalho, desenvolver, comparar e selecionar modelos baseados em técnicas distintas de mineração de dados, para a predição da taxa de progresso mensal da ferrugem do cafeeiro, em anos de alta e baixa carga pendente de frutos, buscando melhorar o desempenho em relação a modelos já existentes.

2 MATERIAL E MÉTODOS

Os dados utilizados neste trabalho foram coletados, referindo-se ao acompanhamento mensal da incidência da ferrugem do cafeeiro, em uma fazenda experimental da Fundação PROCAFÉ. Essa fazenda está localizada na cidade de Varginha/MG, latitude sul de 21º 34' 00'', longitude oeste de 45º 24' 22'' e altitude de 940 m. Foram coletados registros, entre outubro de 1998 até outubro de 2011, totalizando 13 safras agrícolas. As lavouras selecionadas tinham idade entre 6 e 20 anos, sendo quatro em espaçamento largo (3,5 m entrelinhas e 0,7 m entre plantas – densidade média de 4.000 plantas/ha) e quatro adensadas (2,5 m entrelinhas e 0,5 m entre plantas – densidade média de 8.000 plantas/ha). Para cada espaçamento, havia duas lavouras com alta carga pendente de frutos (acima de 30 sacas beneficiadas/ha) e duas com baixa carga (abaixo de 10 sacas beneficiadas/ha). Em cada par de lavouras, uma foi da cultivar Catuaí (Vermelho e Amarelo) e a outra da cultivar Mundo Novo. Não houve controle da doença durante o ano agrícola. O período de colheita foi entre junho e agosto. O processo de amostragem foi realizado ao final de cada mês, como recomendado por Chalfoun (1997).

Além dos dados referentes à doença, foram obtidos dados meteorológicos, como temperatura (média, máxima e mínima), precipitação pluviométrica, umidade relativa do ar e velocidade do vento. Esses dados foram registrados a cada 30 minutos por uma estação meteorológica automática, presente na fazenda experimental.

A taxa de progresso mensal da ferrugem do cafeeiro (TP) foi definida como a variável dependente, também chamada de atributo meta. Essa taxa consiste no aumento, diminuição ou manutenção do nível de incidência da doença entre dois meses subsequentes, calculada pela diferença entre a incidência em um mês e a incidência no mês anterior. Seus valores foram mapeados em um atributo de origem binária, sendo que a classe '1' indica TP maior ou igual a 5 p.p. (pontos percentuais), e a classe '0' indica TP inferior a 5 p.p. O valor de 5 p.p. foi baseado em Meira, Rodrigues e Moraes (2009).

As variáveis independentes, ou atributos preditivos, foram criadas a partir do nível horário, forma em que os dados meteorológicos foram coletados, e passaram por transformações, chegando a um nível que permitiu a integração desses atributos com o atributo meta. A construção dos atributos preditivos iniciou-se tratando cada dia como um dia eventual de infecção. Considerando o período de incubação do fungo, estimado pela equação de Moraes et al. (1976), cada dia foi associado ao mês de avaliação da incidência da ferrugem para o qual, possivelmente, contribuiu com o aparecimento de novos sintomas da doença (MEIRA; RODRIGUES; MORAES, 2008). O conjunto de dias associado a uma taxa de infecção foi denominado de período de infecção (PINF).

No nível diário, além de médias e somatórias das variáveis meteorológicas, foram calculados valores estimados de molhamento foliar prolongado (mínimo de 6 horas), uma vez que a germinação dos uredósporos de *H. vastatrix* só ocorre se a folha estiver molhada. O número de horas contínuas com alta umidade relativa do ar (maior ou igual a 90%) foi utilizado como medida indireta de molhamento foliar. O molhamento ocorre geralmente de um dia para o outro, então foi considerado o intervalo entre 12h de um dia até 12h do dia seguinte (dia epidemiológico). Os períodos de molhamento foliar foram analisados na extensão total e na fração noturna (das 20h às 8h). A temperatura média durante o período de molhamento também foi considerada, 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; AKUTSU; LUDWIG, 1983). Os dias com precipitação maior ou igual a 1 mm foram considerados chuvosos, seguindo o mesmo critério usado pelos autores.

Para completar o conjunto de dados foram criados atributos especiais. Tais atributos reuniram condições diárias de molhamento foliar, luminosidade e temperatura, durante o período de molhamento foliar, com relação ao processo de infecção. A partir das correlações da Tabela 1, cada dia foi classificado como desfavorável, favorável ou muito favorável para a infecção por *H. vastatrix*, associado a um índice numérico. Um atributo especial considerou o acúmulo dos índices numéricos, ao longo dos dias do PINF.

O conjunto de dados utilizado na modelagem totalizou 612 registros, sendo que alguns registros foram eliminados devido a falhas na estação meteorológica. Esse conjunto foi dividido em dois, um para anos de alta carga pendente de frutos e outro para anos de baixa carga pendente de frutos. O conjunto para baixa carga pendente de frutos apresentava cerca de 20% dos registros relativos à classe 1 (TP superior a 5 p.p.), mostrando a necessidade de utilização de um método de balanceamento de classes. Foi utilizado o método chamado "Smote+Tomek" (BATISTA; PRATI; MONARD, 2004), o qual deixou cada classe com cerca de 50% dos registros.

Modelos também foram induzidos com o conjunto de dados não balanceado.

O conjunto de alta carga pendente já estava originalmente balanceado.

TABELA 1 - Classificação de um dia como favorável ou não para a infecção por *Hemileia vastatrix,* a partir de condições de infecção e seus respectivos índices numéricos.

Temperatura (T) [ºC] Molhamento foliar* (MF) e luminosidade	Desfavorável (T < 15 ou T > 29)	Pouco favorável (15 ≤ T < 18 ou 27 < T ≤ 29)	Favorável (18 ≤ T < 21 ou 24 < T ≤ 27)	Muito favorável (21 ≤ T ≤ 24)
Desfavorável (MF noturno** < 4 ou MF < 6)	Desfavorável 0	Desfavorável 0	Desfavorável 0	Desfavorável 0
Pouco favorável (MF noturno ≥ 4 e MF ≥ 6)	Desfavorável 0	Desfavorável 0***	Favorável 1***	Favorável 2***
Favorável (MF noturno ≥ 8 e MF ≥ 12)	Desfavorável 0	Favorável 1	Favorável 2	Muito favorável 3
Muito favorável (MF noturno ≥ 8 e MF ≥ 18)	Desfavorável 0	Favorável 2	Muito favorável 3	Muito favorável 4

* Molhamento foliar foi medido pelo número de horas contínuas com umidade relativa maior ou igual a 90%; ** MF noturno foi considerado das 20:00 h às 8:00 h e MF total foi considerado no dia epidemiológico, entre 12:00 h de um dia e 12:00 h do dia seguinte; *** Caso MF noturno ≥ 8 ou MF ≥ 12, incrementa-se 0,5 no índice.

Antes da modelagem, foram ainda realizadas seleções de atributos de forma subjetiva e objetiva. A maneira subjetiva consistiu na seleção de atributos de acordo com a complexidade e dificuldade de obtenção dos mesmos. O primeiro conjunto (C1) reuniu todos os atributos. O segundo conjunto (C2) conteve atributos meteorológicos mais simples e de ampla disponibilidade, como temperatura, precipitação e umidade relativa, além dos atributos relativos ao molhamento foliar. O terceiro conjunto (C3) teve os mesmos atributos de C2, com exceção dos atributos relacionados ao molhamento foliar. A outra forma de seleção foi por meio de métodos objetivos, quando um algoritmo de seleção é utilizado para filtrar o conjunto de dados. Cinco métodos amplamente conhecidos na área de mineração de dados foram utilizados: CFS, InfoGain, GainRatio, Chi-quadrado (Chi2) e Wrapper (WITTEN; FRANK; HALL, 2011). Os métodos de seleção de atributos foram aplicados ao conjunto C1. As informações sobre os atributos e conjuntos de dados estão na Tabela 2.

O software utilizado na indução dos modelos foi o WEKA, versão 3.7.9 (HALL et al., 2009). Foram utilizadas as seguintes técnicas de modelagem: árvores de decisão, redes neurais artificiais, florestas aleatórias e máquinas de vetores suporte.

A indução por máquinas de vetores suporte foi feita com a utilização da biblioteca LIBSVM (CHANG; LIN, 2011).

A avaliação dos modelos ocorreu por meio de suas medidas de desempenho, derivadas de uma matriz de acertos e erros, chamada matriz de confusão. Medidas como taxa de acerto, taxa de erro, sensitividade e especificidade foram utilizadas para avaliação dos modelos. A sensitividade é a porcentagem de exemplos positivos que foram classificados corretamente, já a especificidade é a porcentagem de exemplos negativos que foram classificados corretamente. As medidas de desempenho foram geradas por meio de uma validação cruzada em 10 partes (WITTEN; FRANK; HALL, 2011).

Gráficos do tipo ROC (do inglês Receiver Operating Characteristic) também foram utilizados para avaliar e selecionar os modelos. Um modelo de classificação é representado por um ponto no espaço ROC (FAWCETT, 2006). Ao se analisar um grupo de modelos no espaço ROC, pode-se notar a presença de um "envelope externo convexo" (PRATI; BATISTA; MONARD, 2008). Os modelos que se encontram nos "vértices" desse envelope são modelos considerados ótimos, já os que não fazem parte do envelope têm desempenho inferior e podem ser descartados (FAWCETT, 2006).

TABELA 2 - Lista e significado de cada um dos atributos presentes nos respectivos conjuntos de dados utilizados na indução dos modelos de alerta.

Natureza	Nome	Medida	Significado	Conjunto
Espacial	LAVOURA	-	Lavoura ADENSADA ou LARGA.	C1,C2,C3
Meteorológico	TMAX_PINF	ºC	Média das temperaturas máximas diárias no período de infecção (PINF).	C1,C2,C3
Meteorológico	TMIN_PINF	ºC	Média das temperaturas mínimas diárias no PINF.	C1,C2,C3
Meteorológico	TMED_PINF	ºC	Média das temperaturas médias diárias no PINF.	C1,C2,C3
Meteorológico	UR_PINF	%	Umidade relativa do ar média diária no PINF.	C1,C2,C3
Meteorológico	MED_PRECIP_PINF	mm	Média das precipitações pluviais diárias no PINF.	C1,C2,C3
Meteorológico	PRECIP_PINF	mm	Precipitação pluvial acumulada no PINF.	C1,C2,C3
Meteorológico	DCHUV_PINF	dias	Número de dias chuvosos (precipitação ≥ 1 mm) no PINF.	C1,C2,C3
Meteorológico	TMAX_PI_PINF	ºC	Média das temperaturas máximas diárias no período de incubação (PI) para os dias do PINF.	C1,C2,C3
Meteorológico	TMIN_PI_PINF	ºC	Média das temperaturas mínimas diárias no PI para os dias do PINF.	C1,C2,C3

Continua ...

TABELA 2 Cont.

Meteorológico	TMED_PI_PINF	ºC	Média das temperaturas médias diárias no PI para os dias do PINF.	C1,C2,C3
Meteorológico	NHUR90_PINF	h	Média diária do número de horas com UR ≥ 90% no PINF.	C1,C2
Meteorológico	THUR90_PINF	ºC	Temperatura média diária durante as horas com UR ≥ 90% no PINF.	C1,C2
Meteorológico	NHNUR90_PINF	h	Média diária do número de horas noturnas com UR ≥ 90% no PINF.	C1,C2
Meteorológico	SMT_NHUR90_PINF	h	Somatório de NHUR90 no PINF.	C1
Meteorológico	SMT_NHNUR90_PINF	h	Somatório de NHNUR90 no PINF.	C1
Meteorológico	VVENTO_PINF	km/h	Velocidade média diária do vento no PINF.	C1
Meteorológico	SMT_VVENTO_PINF	km/h	Média do somatório da velocidade do vento de cada dia do PINF.	C1
Meteorológico	MED_INDPLUVMAX_PINF	mm/h	Média do índice pluviométrico* médio no PINF.	C1
Especial	ACDINF_PINF	-	Acumulado da condição diária de infecção no PINF, isto é, o somatório dos índices de condição diária de infecção no PINF.	C1
Especial	DMFI_PINF	dias	Número de dias muito favoráveis à infecção no PINF.	C1
Especial	DFMFI_PINF	dias	Número de dias favoráveis e muito favoráveis à infecção no PINF.	C1
Especial	DDI_PINF	dias	Número de dias desfavoráveis à infecção no PINF.	C1

* O índice pluviométrico é uma medida de intensidade das chuvas, calculada por meio do intervalo de tempo entre aumentos na precipitação.

3 RESULTADOS E DISCUSSÃO

Os modelos apresentados nesta seção são responsáveis por emitirem alertas sobre o aumento da incidência da ferrugem do cafeeiro, sendo que um alerta é emitido quando houver um aumento igual ou superior a 5 p.p. (pontos percentuais), na taxa de progresso mensal da ferrugem do cafeeiro (TP). Esse alerta é a informação que pode dar suporte à tomada de decisão, por parte do produtor, quanto à adoção de medidas de controle e o melhor momento para implementá-las.

A Figura 1 apresenta o gráfico ROC (Receiver Operating Characteristic) para modelos de alerta em anos de alta carga pendente de frutos. Os pontos numerados no gráfico representam o desempenho desses modelos. Dois modelos (22 e 28) estão presentes no envelope convexo e podem ser considerados modelos ótimos. As suas medidas de desempenho, as técnicas de modelagem e os métodos de seleção de atributos correspondentes estão apresentados na Tabela 3.

O que se busca em um modelo, em termos de medidas de desempenho, é uma alta taxa de acerto, aliada a altos e equilibrados valores de sensitividade e especificidade. Para a ferrugem do cafeeiro, uma predição correta de aumento da TP pode ser importante para indicar ou confirmar a necessidade de uma aplicação de fungicida, já um alarme negativo pode auxiliar na decisão de adiamento dessa aplicação. No caso de um esquema de controle com duas aplicações de fungicida, diminui-se o risco da ferrugem tardia quando a segunda aplicação puder ser adiada.

Dentre os modelos selecionados no envelope convexo para alta carga pendente de frutos, o que apresentou medidas de desempenho melhores, além de mais equilibradas, foi o modelo 28. O modelo 22 apresentou valores superiores de sensitividade em relação ao 28, entretanto, esse parâmetro não pode ser utilizado isoladamente para avaliar o desempenho de um modelo. O modelo 28, a princípio, desponta como o mais indicado para realizar a predição de aumento da TP.

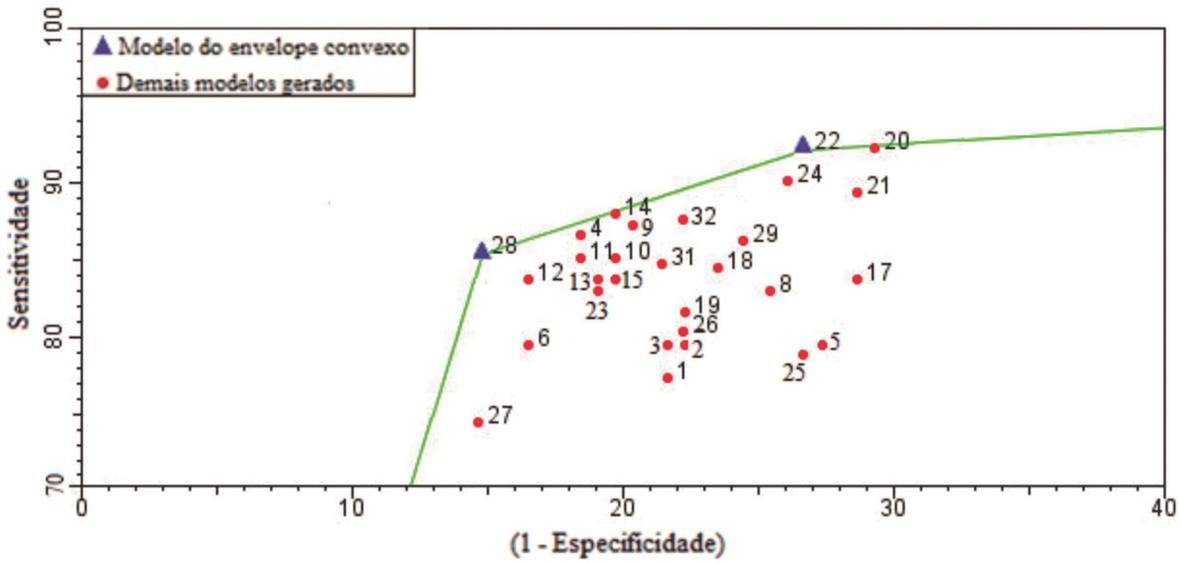

FIGURA 1 - Gráfico ROC com os pontos relativos ao desempenho dos 32 modelos desenvolvidos – numerados de acordo com a técnica de modelagem e método de seleção de atributos, destacando-se os selecionados no envelope convexo para anos de alta carga pendente de frutos.

TABELA 3 - Características construtivas e medidas de desempenho para modelos de alerta gerados para anos de alta e baixa carga pendente de frutos.

Modelos para alta carga pendente de frutos (Figura 1)					
Modelo	Técnica de modelagem	Seleção de atributos	Taxa de acerto (%)	Sensitividade (%)	Especificidade (%)
22	RNA	Chi^2	82,2	92,2	73,3
28	SVM	Wrapper	85,3	85,4	85,2
Modelos para baixa carga pendente de frutos (Figura 2).					
Modelo	Técnica de modelagem	Seleção de atributos	Taxa de acerto (%)	Sensitividade (%)	Especificidade (%)
25	RF	C1	88,9	86,3	89,9

RF = Florestas aleatórias (RF, do inglês Random Forests).
RNA = Redes neurais artificiais.
SVM = Máquinas de vetores suporte (SVM, do inglês Support Vector Machines).

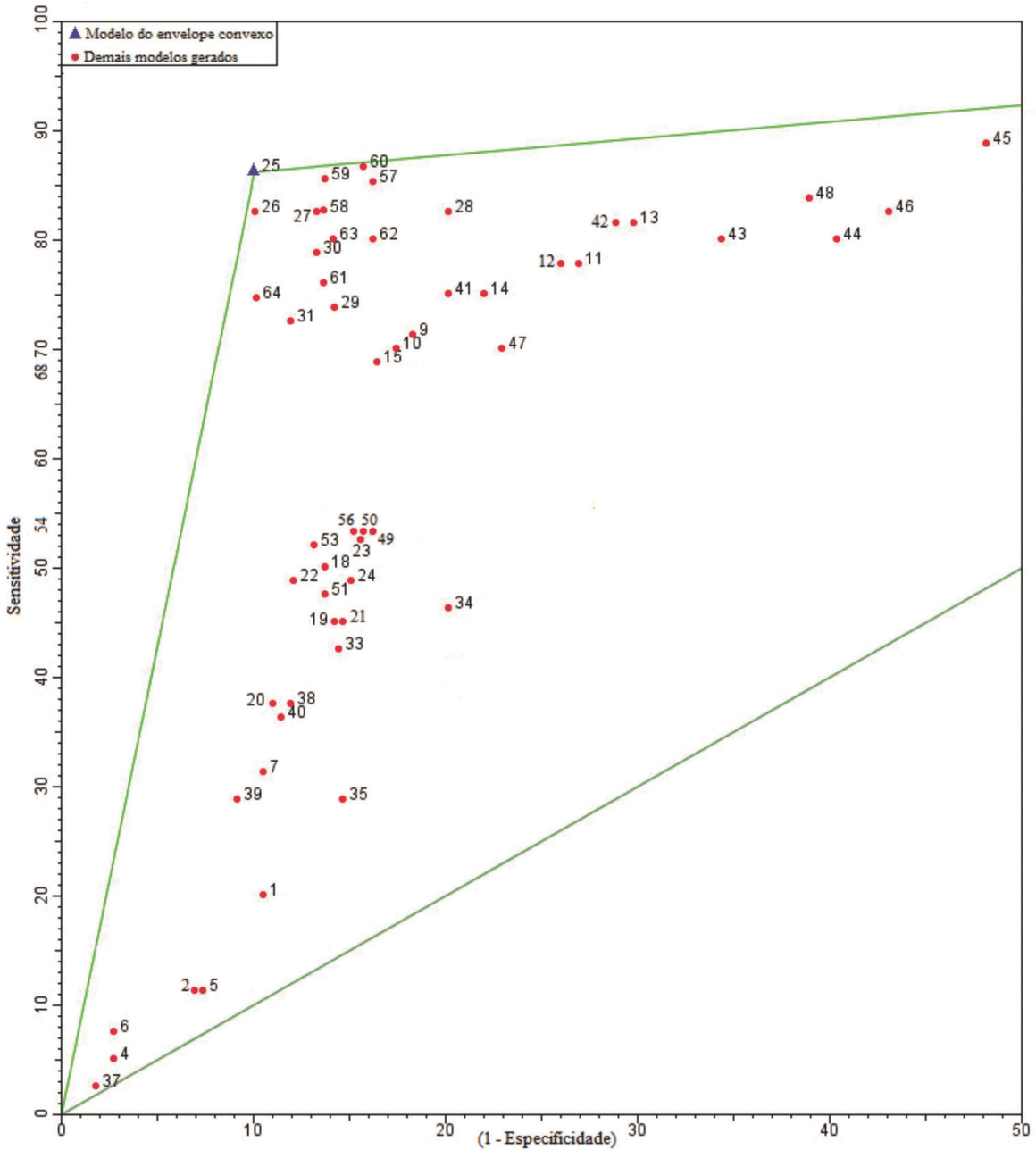

FIGURA 2 - Gráfico ROC com os pontos relativos ao desempenho dos 64 modelos desenvolvidos – numerados de acordo com a técnica de modelagem, método de seleção de atributos e balanceamento ou não de classes, destacando-se os selecionados no envelope convexo para anos de baixa carga pendente de frutos.

A Figura 2 apresenta o gráfico ROC, para modelos de alerta em anos de baixa carga pendente de frutos. Apenas um modelo (25) está presente no envelope convexo, sendo o modelo que pode ser considerado ótimo. As suas medidas de desempenho e informações construtivas estão apresentadas na Tabela 3.

Pela disposição da Figura 2, percebe-se uma grande discrepância de desempenho entre os modelos gerados por arquivos balanceados e não balanceados.Os modelos localizados abaixo da faixa de 54% de sensitividade foram gerados por conjuntos de dados não balanceados, o que indicou que eles classificaram, no máximo, pouco mais da metade dos exemplos positivos corretamente. Os demais modelos ficaram acima da faixa de 68% de sensitividade, desempenho de, no mínimo, 14 p.p. melhor, mostrando a superioridade dos modelos gerados por conjuntos balanceados.

Considerando os modelos gerados por arquivos balanceados, 14 deles ficaram mais próximos ao envelope convexo, ou fizeram parte dele. Esses modelos foram gerados pelas técnicas de máquinas de vetores suporte e florestas aleatórias, evidenciando o melhor desempenho médio dessas técnicas em relação às demais.

Para modelos de alerta com medidas de avaliação próximas, o conjunto de dados de treinamento pode ser usado na decisão de qual é o mais indicado para predizer a TP. Modelos com uma maior quantidade de atributos no seu conjunto de dados, além de atributos difíceis de serem calculados, têm sua aplicabilidade reduzida, justamente pela dificuldade de obtenção dos dados para gerá-los. Em contrapartida, modelos com poucos atributos podem ser considerados impróprios, pelo fato desses não apresentarem atributos relevantes às mais diversas condições de ocorrência e desenvolvimento da ferrugem do cafeeiro. Os atributos utilizados em cada um dos modelos estão na Tabela 4.

Ao menos um atributo especial foi utilizado por cada modelo da Tabela 4. A escolha deve estar relacionada à representatividade desses atributos com relação à doença, por reunirem diversas condições propícias ao seu desenvolvimento (luminosidade, temperatura e duração do período de molhamento foliar). Apesar de representativos, esses atributos são difíceis de serem calculados, o que pode dificultar a aplicabilidade dos modelos nos casos em que não haja disponibilidade suficiente de dados coletados para gerá-los.

Os modelos 22 e 25 (Tabela 4) utilizaram os atributos de temperatura mínima (TMIN_PINF) e média (TMED_PINF). O primeiro desses atributos apresentou valor médio de 15 ^{o}C, mas cerca de 35% de seus registros foram inferiores a 14 ^{o}C, valor considerado limitante para a infecção de *H. vastatrix* (KUSHALAPPA; AKUTSU; LUDWIG, 1983). Já o segundo atributo conteve cerca de 20% dos seus registros entre 22 e 24 ^{o}C, faixa considerada ótima para o desenvolvimento do fungo (ZAMBOLIM et al., 2002). O atributo de temperatura máxima (TMAX_PINF) foi utilizado apenas pelo modelo 25 e seu valor variou de 22 a 30 ^{o}C, não superando a faixa considerada limitante para o desenvolvimento de *H. vastatrix* (KUSHALAPPA; AKUTSU; LUDWIG, 1983). Ao contrário da temperatura mínima, o atributo de temperatura máxima desse local não foi responsável por condições que inibissem o desenvolvimento do fungo na lavoura. Esses três atributos reunem condições importantes para o desenvolvimento da ferrugem do cafeeiro e são simples de serem calculados, sendo que não reduzem a aplicabilidade de um modelo.

Atributos relacionados à precipitação (MED_PRECIP_PINF, PRECIP_PINF, DCHUV_PINF) estiveram presentes em dois modelos (25 e 28) da Tabela 4. Eles forneceram subsídios para explicar a distribuição e intensidade das chuvas. Muitos dias chuvosos, com média de chuva diária e total mensal baixos, indicam chuvas prolongadas e com baixa intensidade; em contrapartida, poucos dias chuvosos, aliados a uma média de chuva diária e total mensal altos, representam chuvas curtas e de alta intensidade. Chuvas leves e constantes aumentam a umidade relativa, deixando água livre na superfície das folhas por mais tempo e levando a períodos de molhamento foliar prolongados, favorecendo a ocorrência das condições mínimas de molhamento foliar (6 horas contínuas) para o desenvolvimento do fungo (KUSHALAPPA; AKUTSU; LUDWIG, 1983). Além disso, chuvas leves e seus respingos são apontados como um dos principais fatores de disseminação do fungo dentro da própria planta (ZAMBOLIM et al., 2002). Já chuvas intensas podem conduzir a maior parte dos esporos para o chão (KUSHALAPPA; ESKES, 1989), além de não promoverem períodos longos de molhamento foliar. Todos esses atributos contêm condições importantes para o desenvolvimento da ferrugem e são de baixa complexidade, ou seja, não reduzem a aplicabilidade de um modelo.

TABELA 4 - Atributos do conjunto de dados para os modelos de alerta selecionados no envelope convexo, para anos de alta e baixa carga pendente de frutos.

	Modelos para alta carga		Modelos para baixa carga
Atributos	22	28	25
LAVOURA			*
TMAX_PINF			*
TMIN_PINF	*		*
TMED_PINF	*		*
UR_PINF			*
MED_PRECIP_PINF		*	*
PRECIP_PINF			*
DCHUV_PINF			*
TMAX_PI_PINF			*
TMIN_PI_PINF			*
TMED_PI_PINF			*
NHUR90_PINF			*
THUR90_PINF		*	*
NHNUR90_PINF			*
SMT_NHUR90_PINF			*
SMT_NHNUR90_PINF			*
VVENTO_PINF			*
SMT_VVENTO_PINF		*	*
MED_INDPLUVMAX_PINF			*
ACDINF_PINF	*		*
DMFI_PINF	*		*
DFMFI_PINF	*	*	*
DDI_PINF	*		*

* Atributo presente no modelo.

A ocorrência de períodos de molhamento foliar pode ser ainda mais esclarecida pelo atributo de umidade relativa (UR_PINF), o qual só esteve presente no modelo 25. Quanto maior a umidade relativa média, maior a tendência para a ocorrência de períodos de molhamento foliar. Esse é um atributo de baixa complexidade e serve como base para o cálculo dos atributos de medida indireta de molhamento foliar.

O atributo de temperatura no molhamento foliar (THUR90_PINF) esteve presente em dois modelos (25 e 28) da Tabela 4. A presença desse atributo, juntamente com as informações sobre o molhamento foliar, podem representar a condição ótima de desenvolvimento do fungo (8 horas de molhamento foliar e temperatura entre 22 a 24 ^{o}C).

Os atributos relacionados à medição indireta de molhamento foliar (NHNUR90_PINF e NHUR90_PINF) foram utilizados apenas no modelo 25. Esses atributos têm complexidade mediana, menor que a dos atributos especiais, mas maior que a de atributos de temperatura e pluviosidade.

Após a avaliação dos atributos do conjunto de dados e das medidas de desempenho, foi possível selecionar os modelos recomendados para realizar as predições da taxa de progresso da ferrugem do cafeeiro.

Para anos de alta carga pendente de frutos, notou-se que o modelo 28 foi o que obteve a melhor taxa de acerto, além de valores de sensitividade e especificidade elevados e equilibrados.

Assim, o modelo 28 classifica corretamente uma maior porcentagem de casos do que o modelo 22, além de ter altas porcentagens de acerto para as classes positiva e negativa. Com relação ao conjunto de dados, os dois modelos contaram com atributos especiais, os quais diminuem sua aplicabilidade. Assim, não houve simplicidade maior de um modelo sobre o outro, tornando o modelo 28 o recomendado para realizar a predição da TP, em anos de alta carga pendente de frutos.

Para os anos de baixa carga pendente de frutos, apenas um modelo foi selecionado no envelope convexo, o modelo 25. Em termos de medidas de desempenho, esse modelo é o mais eficiente para realizar predições da TP da ferrugem. Entretanto, ao analisar os atributos utilizados para gerá-lo, notou-se que ele utilizou o conjunto C1, com todos os atributos especiais, o que reduz a sua aplicabilidade. No caso de uma indisponibilidade de dados necessários para utilizar esse modelo, uma alternativa viável seria o uso do modelo 26. Esse modelo também foi gerado por florestas aleatórias e obteve valores da taxa de acerto (87,9%), da sensitividade (82,5%) e da especificidade (89,9%) muito próximos ou iguais aos do modelo 25. Seu conjunto de dados foi o C2, o qual não conteve atributos especiais, deixando-o relativamente mais simples do que o modelo 25.

A aplicabilidade destes modelos ainda está relacionada com sua abrangência. O seu uso deve ser limitado à região onde os dados foram coletados ou em regiões com condições climáticas similares. Regiões com diferentes climas podem apresentar condições meteorológicas que não foram representadas nos dados analisados e que podem condicionar o desenvolvimento da ferrugem de maneira diferente da capturada pelos modelos (MEIRA; RODRIGUES; MORAES, 2008).

Considerando-se apenas o modelo escolhido para anos de alta carga pendente de frutos (28), suas medidas de desempenho como taxa de acerto (85,3%), sensitividade (85,4%) e especificidade (85,2%) foram superiores às obtidas pelas árvores de decisão geradas por Meira, Rodrigues e Moraes (2009), onde esses valores foram 81,3%, 79,9% e 82,6%, respectivamente. A taxa de acerto do modelo 28 também foi superior em relação às árvores de decisão *fuzzy* desenvolvidas por Cintra et al. (2011), onde a melhor obteve como resultado o valor de 84,7%.

4 CONCLUSÕES

Os melhores modelos foram gerados pelas técnicas de máquinas de vetores suporte e florestas aleatórias, sendo que o procedimento de balanceamento de classes ajudou na melhora da taxa de acerto. Os modelos de predição da taxa de progresso mensal da ferrugem do cafeeiro desenvolvidos neste trabalho fornecem melhores subsídios para o monitoramento da doença, em anos de alta carga pendente de frutos do que outros modelos existentes, além de prover uma possibilidade de monitoramento, em anos de baixa carga pendente de frutos.

5 AGRADECIMENTOS

Ao Consórcio Pesquisa Café e à CAPES pelo apoio financeiro. À fundação PROCAFÉ pelo fornecimento dos dados.

6 REFERÊNCIAS

ALVES, M. C. et al. A Soft computing approach for epidemiological studies of coffee and soybean rusts. International Journal of Digital Content Technology and its Applications, Sandy Bay, v. 4, n. 1, p. 149-154, Feb. 2010.

BATCHELOR, W. D.; YANG, X. B.; TSCHANZ, A. T. Development of a neural network for soybean rust epidemics. Transactions of the American Society of Agricultural Engineers, Saint Joseph, v. 40, n. 1, p. 247-252, 1997.

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, New York, v. 6, n. 1, p. 20-29, June 2004.

BREIMAN, L. Random forests. Machine Learning Journal, Hingham, v. 45, p. 5-32, Jan. 2001.

CHALFOUN, S. M. Doenças do cafeeiro: importância, identificação e métodos de controle. Lavras: UFLA/FAEPE, 1997.

CHANG, C. C.; LIN, C. J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, New York, v. 2, n. 3, p. 1-27, Apr. 2011.

CINTRA, M. E. et al. The use of fuzzy decision trees for coffee rust warning in Brazilian crops. In: INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS DESIGN AND APPLICATIONS, 11., 2011, Córdoba. **Proceedings...** Córdoba: IEEE, 2011. p. 1347-1352.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, New York, v. 27, n. 8, p. 861-874, June 2006.

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

HALL, M. A. et al. The WEKA data mining software: an update. **SIGKDD Explorations**, New York, v. 11, n. 1, p. 10-18, June 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data mining:** concepts and techniques. 3rd ed. San Francisco: M. Kaufmann, 2011. 703 p.

HARDWICK, N. V. Disease forecasting. In: COOKE, B. M.; JONES, D. G.; KAYE, B. (Ed.). **The epidemiology of plant diseases**. 2nd ed. Wageningen: Springer, 2006. p. 239-267.

HAYKIN, S. **Neural networks and learning machines**. 3rd ed. Englewood Cliffs: Prentice-Hall, 2009. 936 p.

KUSHALAPPA, A. C.; AKUTSU, M.; LUDWIG, A. Application of survival ratio for monocyclic process of *Hemileia vastatrix* in predicting coffee rust infection rates. **Phytopathology**, Saint Paul, v. 73, n. 1, p. 96-103, 1983.

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

LEE, M. C.; TO, C. Comparison of support vector machine and back propagation neural network in evaluating the enterprise financial distress. **International Journal of Artificial Intelligence & Applications**, Niskayuna, v. 1, p. 31-43, 2010.

LUACES, O. et al. Using nondeterministic learners to alert on coffee rust disease. **Expert Systems With Applications**, New York, v. 38, n. 11, p. 14276-14283, Jan. 2011.

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**, Brasília, v. 33, n. 2, p. 114-124, mar./abr. 2008.

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**, Brasília, v. 44, n. 3, p. 233-242, mar. 2009.

MOLINEROS, J. E. et al. Modeling epidemics of fusarium head blight: trials and tribulations. **Phytopathology**, Saint Paul, v. 95, n. 6, p. S71, 2005.

MORAES, S. A. et al. Período de incubação de *Hemileia vastatrix* Berk. e Br. em três regiões do Estado de SP. **Summa Phytopathologica**, Piracicaba, v. 2, n. 1, 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**, Saint Paul, v. 94, n. 12, p. 1350-1357, 2004.

PAUL, P. A.; MUNKVOLD, G. P. Regression and artificial neural network modeling for the prediction of gray leaf spot of maize. **Phytopathology**, Saint Paul, v. 95, n. 4, p. 388-396, 2005.

PÉREZ-ARIZA, C. B.; NICHOLSON, A. E.; FLORES, M. J. Prediction of coffee rust disease using bayesian networks. In: EUROPEAN WORKSHOP ON PROBABILISTIC GRAPHICAL MODELS, 6., 2012, Granada. **Proceedings...** Granada: PGM, 2012. p. 259-266.

PINTO, A. C. S. et al. Descrição da epidemia da ferrugem do cafeeiro com redes neuronais. **Fitopatologia Brasileira**, Brasília, v. 27, n. 5, p. 517-524, set./out. 2002.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Curvas ROC para avaliação de classificadores. **Revista IEEE América Latina**, São Paulo, v. 6, n. 2, p. 215-222, jun. 2008.

SOUZA, V. C. O. et al. Técnicas de extração de conhecimento aplicadas a modelagem de ocorrência da cercosporiose (*Cercospora coffeicola Berkeley & Cooke*) em cafeeiros na região sul de minas gerais. **Coffee Science**, Lavras, v. 8, n. 1, p. 91-100, jan./mar. 2013.

UNITED STATES DEPARTMENT OF AGRICULTURE. **Coffee:** world markets and trade. Disponível em: <http://www.fas.usda.gov/psdonline/circulars/coffee.pdf>. Acesso em: 15 fev. 2013.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining:** practical machine learning tools and techniques. 3rd ed. San Francisco: M. Kaufmann, 2011. 629 p.

ZAMBOLIM, L. et al. Epidemiologia e controle integrado da ferrugem-do-cafeeiro. In: ZAMBOLIM, L. (Ed.). **O estado da arte de tecnologias na produção de café**. Viçosa, MG: UFV, 2002. p. 369-449.