Potencial de técnicas de mineração de dados para modelos de alerta da ferrugem do cafeeiro

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

[1]Embrapa Informática Agropecuária – Caixa Postal 6041 - 13083-970 – Campinas – SP – Brasil.

[2]Faculdade de Engenharia Agrícola – Universidade Estadual de Campinas – Campinas – SP – Brasil.

cesare.neto@colaborador.embrapa.br, lique@feagri.unicamp.br, thiago.tamada@colaborador.embrapa.br, carlos.meira@embrapa.br

Abstract. *This paper aims to evaluate the potential of data mining techniques on developing warning models for coffee rust. Four techniques were evaluated: Neural Networks, Decision Trees, Support Vector Machines and Random Forests. The results show that models developed by the two last techniques obtained higher accuracy and also better values for recall and specificity. The Neural Networks were responsible for models with high values of recall, while the performance of Decision Tree models was the worst when compared to the others. The class balancing also proved to be an essential procedure to improve the accuracy of the models developed.*

Resumo. *Este trabalho procurou avaliar o potencial de técnicas de mineração de dados no desenvolvimento de modelos de alerta da ferrugem do cafeeiro. Foram avaliadas quatro técnicas: Redes Neurais Artificiais, Árvores de Decisão, Support Vector Machines e Random Forest. A avaliação dos modelos gerados mostrou que as duas últimas técnicas geram modelos com maior taxa de acerto e melhores medidas de sensitividade e especificidade. As Redes Neurais Artificiais geraram modelos com alto valor de sensitividade, enquanto que as Árvores de Decisão obtiveram desempenho inferior quando comparadas às demais técnicas. O balanceamento de classes se mostrou um procedimento fundamental na melhora da taxa de acerto dos modelos.*

1. Introdução

Causada pelo fungo *Hemileia vastatrix* Berk. e Br., a ferrugem é a principal doença do cafeeiro. As perdas de produção por esta doença podem chegar a 50%, caso nenhuma medida de controle seja adotada [Zambolim et al. 2002]. O controle da ferrugem pode ser feito com fungicidas, entretanto, métodos tradicionais de controle podem levar a aplicações desnecessárias, gerando gastos excessivos para o produtor, na compra e mão de obra para sua aplicação, além de causar impactos ambientais.

Ferramentas como modelos de predição, ou alerta, podem ser utilizadas para antecipar quando uma doença de planta ocorrerá, sendo que uma predição correta pode evitar aplicações desnecessárias de fungicidas. A partir da divulgação no meio científico da Mineração de Dados (DM) por Fayyad et al. (1996), tem-se notado um aumento no número de modelos gerados por meio de tal metodologia. Neste sentido, modelos de alerta para determinar a taxa de progresso da ferrugem do cafeeiro (TP) foram desenvolvidos por Meira et al. (2009) e Cintra et al. (2011), os quais desenvolveram

Árvores de Decisão (AD) e AD f*uzzy*, respectivamente. Outras técnicas também têm sido utilizadas para a modelagem da ferrugem do cafeeiro, como Redes Neurais Artificiais (RNA) *fuzzy*, desenvolvidas por Alves et al. (2010); Support Vector Machines (SVM), induzidas por Luaces et al. (2011) e Random Forest (RF), que também foram mencionadas no trabalho de Cintra et al. (2011).

A escolha pelo uso de uma ou outra técnica de modelagem requer a análise do problema em questão. Os modelos em AD são fáceis de serem interpretados [Witten et al. 2011], a medida que as RF, além de evitar *overfitting* (sobreajuste), também são pouco sensíveis a ruídos [Breiman 2001]. A principal vantagem das RNA é resolver problemas que apresentam uma solução muito difícil de ser encontrada [Haykin 2009], enquanto que as SVM também evitam sobreajuste e normalmente produzem classificadores muito precisos [Witten et al. 2011].

Sendo assim, a geração de novos modelos de alerta da ferrugem do cafeeiro por meio de diferentes técnicas de modelagem poderia gerar modelos com poder de predição maior do que os gerados anteriormente, além do fato que uma técnica pode ser responsável por uma propriedade específica nos modelos gerados. Logo, o objetivo deste trabalho foi avaliar o potencial das técnicas de Mineração de Dados para modelos de predição da TP da ferrugem, analisando a peculiaridade de cada uma delas.

2. Material e Métodos

Os dados utilizados neste trabalho referem-se ao acompanhamento mensal da incidência da ferrugem do cafeeiro em fazendas experimentais da Fundação PROCAFÉ. Estas fazendas estão em 3 municípios de Minas Gerais: Varginha, Carmo de Minas e Boa Esperança. Foram coletados dados para períodos entre outubro de 1998 e outubro de 2011. Havia lavouras com alta carga pendente de frutos (acima de 30 sacas beneficiadas/ha) e com baixa carga (abaixo de 10 sacas beneficiadas/ha), para os cultivares Catuaí (Vermelho e Amarelo) e Mundo Novo. Não houve controle da doença durante o ano agrícola. O período de colheita foi entre junho e agosto. 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. Estes dados foram registrados a cada 30 minutos por uma estação meteorológica automática presente em cada uma das fazendas experimentais.

A TP foi definida como a variável dependente (atributo meta), a qual consiste no aumento, diminuição ou manutenção da incidência da doença entre dois meses subsequentes. 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’ para TP inferior a 5 p.p. Uma opção de TP com 10 p.p. também foi utilizada para as lavouras em alta carga pendente de frutos. Os valores do atributo meta foram baseados em Meira et al. (2008). Os atributos preditivos, ou variáveis independentes, foram criados a partir dos dados meteorológicos e espaçamento da lavoura. Sua criação partiu da forma que foram coletados (nível horário) e transformados até chegar em um nível que permitiu a integração com o atributo meta.

O conjunto de dados utilizado na modelagem totalizou 738 registros, sendo que alguns registros foram eliminados devido a falhas na estação meteorológica. Este conjunto foi dividido de acordo com os locais e períodos em que os dados foram coletados. Para alta carga pendente de frutos foram utilizados os atributos meta 5 p.p e

10 p.p. e para baixa carga apenas o atributo meta 5 p.p. As divisões geraram diversos cenários de simulação, a fim de obter mais informações sobre o comportamento das técnicas de modelagem.

Para duas combinações (atributo meta 5 p.p. e carga baixa; atributo meta 10 p.p. e carga alta) foi necessário realizar o balanceamento de classes, pois a classe minoritária contava com 20% de registros. Utilizou-se o método "Smote+Tomek" [Batista et al. 2004], o qual deixou cada classe com cerca de 50% dos registros. O conjunto de dados não balanceado também foi utilizado na indução de modelos. Os modelos gerados com arquivos balanceados tiveram seu desempenho avaliado nos conjuntos originais.

A partir destes conjuntos foi realizada a seleção de atributos de duas formas: uma delas subjetiva, a qual consistiu na seleção de atributos de acordo com a complexidade e dificuldade de obtenção dos mesmos, gerando três conjuntos de dados; e a outra por meio de métodos objetivos, em que um algoritmo de seleção foi utilizado para filtrar o conjunto de dados. Cinco métodos amplamente conhecidos na área de MD foram utilizados: CFS, InfoGain, GainRatio, Chi-quadrado e Wrapper [Witten et al. 2011]. Os métodos de seleção foram aplicados ao conjunto contendo todos os atributos.

O software utilizado na indução dos modelos foi o WEKA, versão 3.7.9. [Hall et al. 2009]. Foram utilizadas quatro técnicas de modelagem para induzir os modelos: AD, RNA, RF e SVM. As AD foram geradas pelo classificador "J48". Uma opção de modelagem foi o número mínimo de objetos por folha igual a 5, gerando AD com, no mínimo, 5 registros por folha. Este número foi escolhido pois evitou um sobreajuste muito grande, como no caso de 1 ou 2 objetos por folha, e não deixou que o modelo perdesse um percentual grande de sua taxa de acerto (notou-se uma perda acentuada após 10 registros por folha). Para as RNA foi necessário determinar sua topologia (formato). Foram geradas RNA de uma, duas e três camadas intermediárias, onde cada camada teve seu número de neurônios avaliado de 1 até 10. Após a geração de todas as RNA, a que obteve melhor desempenho foi uma RNA com duas camadas intermediárias de dois neurônios cada.

As SVM foram geradas com a biblioteca LIBSVM [Chang e Lin 2011] e tiveram quatro parâmetros ajustados. O primeiro foi o seu "kernel" (produto interno que eleva os dados a uma dimensão maior, para posteriormente classificá-los). Foram testadas 4 opções de Kernel: linear, polinomial, RBF e Sigmóide, sendo que o escolhido foi o RBF. Dentro do Kernel o parâmetro gamma (γ) foi calibrado em $10^{(-1)}$. O coeficiente de custo (c) e o parâmetro epsilon (ε) também foram ajustados, chegando-se aos valores de 1 e $10^{(-3)}$, respectivamente. Para as RF dois parâmetros foram calibrados: a profundidade das árvores e o número de atributos aleatórios utilizados. O valor que apresentou melhor resultado para ambos os casos foi o valor 8. Assim, tem-se até 8 atributos aleatórios nas árvores da floresta, cada uma com uma profundidade de até 8 níveis. A quantidade de árvores geradas em cada floresta foi definida em 100 [Breiman 2001].

Medidas como taxa de acerto (acurácia), sensitividade e especificidade foram utilizadas para avaliação dos modelos gerados. As medidas de desempenho foram geradas por meio de validação cruzada em 10 partes. Gráficos do tipo ROC também foram utilizados para avaliar e selecionar os melhores modelos.

3. Resultados e Discussão

Foram desenvolvidos 640 modelos. Como exemplo, o gráfico ROC da Figura 1

representa o desempenho dos modelos desenvolvidos para um dos cenários de simulação. Neste cenário, foram selecionados 2 modelos no envelope convexo: 22 e 28. Estes modelos foram agrupados junto com os demais selecionados para os outros cenários de simulação e suas características foram estudadas. Um total de 43 modelos foram selecionados nos envelopes convexos, sendo 22 destes provenientes da seleção de atributos subjetiva e 21 da seleção de atributos objetiva.

Analisando cada técnica de modelagem em questão, verificou-se que modelos gerados pelas RNA apresentaram altos valores de sensitividade em alguns casos. Três cenários de simulação ilustraram este comportamento, onde grupos de modelos em RNA obtiveram altos valores de sensitividade, sempre dentro deste grupo havia um modelo no envelope convexo. A Figura 1 apresenta os modelos 20, 21, 22 e 24 (gerados por RNA) com altos valores de sensitividade (cerca de 92%). Estes resultados foram melhores do que outras técnicas, como, por exemplo, as RF, que chegaram a obter valores de 89%. O maior valor de sensitividade registrado em um modelo selecionado foi de 92,6%. A sensitividade indica como o modelo trabalha com exemplos de aumento da TP, altos valores indicam que um modelo classificou corretamente muitos desses exemplos. Entretanto, ocorreu que valores baixos de especificidade estavam atrelados aos altos valores de sensitividade, fazendo com que, em diversos cenários, os modelos gerados pelas RNA não fizessem parte do envelope convexo, apenas 3 modelos dos 43 foram RNA. Estes modelos isoladamente não são interessantes para predizer o aumento da TP, mas têm alto potencial de serem usados em sistemas que combinam modelos, podendo ser utilizados para confirmar o aumento da TP.

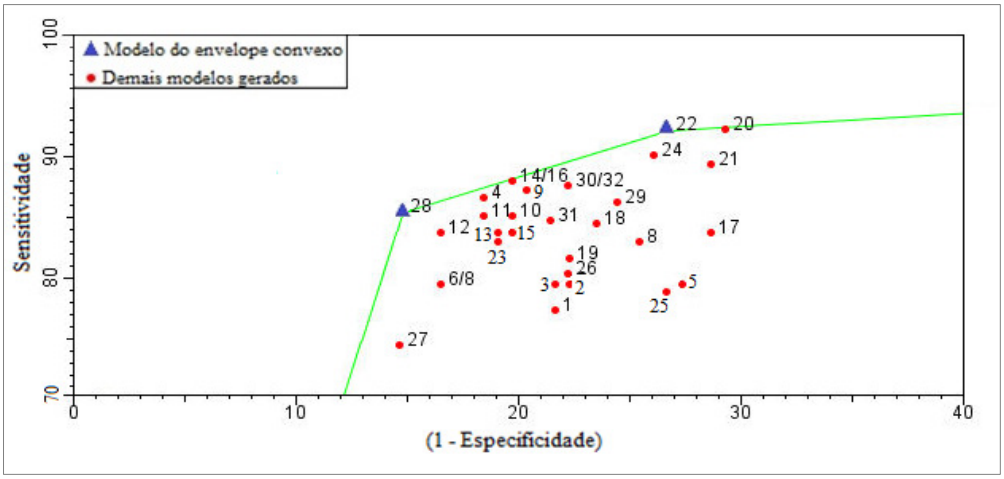

Figura 1: Gráfico ROC para um dos cenários de simulação.

Outra técnica de modelagem que apareceu pouco nos modelos selecionados nos envelopes convexos foram as AD, apenas em 4 dos 43 modelos. Isto ocorreu pois geralmente as AD obtiveram desempenho inferior às técnicas como RF e SVM, em termos de todas as medidas de desempenho. Como um exemplo, na Figura 1, os modelos 1, 2, 3 e 5 foram gerados por AD, e se encontram distantes do envelope convexo. Nenhuma propriedade específica foi notada nas AD, elas simplesmente obtiveram medidas de avaliação inferiores aos demais modelos desenvolvidos. Vale lembrar que AD são representativas e facilitam a visualização das soluções do problema. Sua utilização torna-se mais interessante em casos em que se deseja entender

os fatores que causam algum fenômeno, por exemplo, a descrição da epidemia da ferrugem do cafeeiro [Meira et al. 2008].

Por sua vez, as técnicas de SVM e RF formaram o restante dos modelos selecionados nos envelopes convexos (20 modelos em SVM e 16 em RF). O desempenho destas duas técnicas foi equivalente, onde ambas obtiveram ótimos resultados para diversas medidas de avaliação, com destaque para a taxa de acerto, que chegou próxima a casa dos 91%, no melhor caso. Também destacou-se o equilíbrio das medidas de sensitividade e especificidade, estando elas muitas vezes próximas dos 90% e com baixa diferença entre si, o que indica que o modelo classificou corretamente porcentagens próximas de exemplos de aumento e não aumento da TP, evitando problemas como o baixo valor de especificidade que ocorreu nas RNA.

Outro ponto a ser discutido é a diferença de desempenho de modelos gerados por arquivos com e sem o balanceamento de classes. Esta diferença começa com a presença de apenas 4 modelos provenientes de arquivos não balanceados nos envelopes convexos. Este baixo desempenho foi verificado com mais intensidade nos cenários de carga baixa e atributo meta 5 p.p., mas também foi notado nos cenários de carga alta e atributo meta 10 p.p.. A maior diferença de sensitividade entre estes dois tipos de modelos (entre o pior modelo gerado por arquivos balanceados e o melhor modelo gerado por arquivos não balanceados) foi de 27 p.p.

Sabe-se que a ferrugem apresenta evolução diferenciada nos anos alternados de carga pendente de frutos, sendo mais agressiva em anos de carga alta, trazendo mais ocorrências de aumento da TP. Além disso, quanto maior o limite da TP, há uma tendência de que ocorram menos registros com este aumento. A pequena quantidade de exemplos de uma classe deve ter sido a responsável pelo desempenho indesejado em alguns modelos desenvolvidos sem o balanceamento de classes.

De maneira geral, os modelos presentes nos envelopes convexos obtiveram desempenhos superiores aos desenvolvidos anteriormente. Como exemplo, o modelo 28 (Figura 1) apresentou medidas de taxa de acerto (85,3%), sensitividade (85,4%) especificidade (85,2%) superiores as AD geradas por Meira et al. (2009), onde estes valores foram 81,3%, 79,9% e 82,6%, respectivamente. A taxa de acerto do modelo 28 também foi superior às AD *fuzzy* desenvolvidas por Cintra et al. (2011), onde a melhor obteve como resultado o valor de 84,7%. Os trabalhos de Alves et al. (2010) e Luaces et al. (2011) utilizaram métricas de avaliação diferentes, impossibilitando uma comparação direta com os resultados deste trabalho.

Alguns modelos desenvolvidos neste trabalho estão sendo avaliados para safras agrícolas dos anos de 2011/2012 e 2012/2013. Após a avaliação, eles serão incorporados em um sistema de alerta, que estará disponível na internet para uso exclusivo pelos técnicos da fundação PROCAFÉ, a fim de auxiliar nas recomendações relacionadas ao controle da ferrugem do cafeeiro.

4. Conclusões

Técnicas de mineração de dados mostraram alto potencial para gerar modelos de alerta para a ferrugem do cafeeiro, em especial as Support Vector Machines e as Random Forest, quando se deseja uma alta taxa de acerto. O balanceamento de classes se mostrou imprescindível para melhorar a taxa de acerto destes modelos. Essas técnicas mostram-se promissoras na área de epidemiologia e podem ser utilizadas para outras

doenças do cafeeiro ou doenças de outras culturas.

5. Agradecimentos

À fundação PROCAFÉ por ceder os dados relacionados ao monitoramento de incidência da ferrugem do cafeeiro. Ao Consórcio Pesquisa Café pelo apoio financeiro.

6. Referências

Alves, M. C.; Carvalho, L. G.; Pozza, E. A.; Alves, L. S. A Soft Computing Approach For Epidemiological Studies of Coffee And Soybean Rusts. International Journal of Digital Content Technology and its Applications, v.4, n.1, p.149-154, fev., 2010.

Batista, G. E. A. P. A.; Prati, R. C.; Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations, v.6, n.1, p.20-29, jun., 2004.

Breiman, L. Random forests. Machine Learning Journal. Hingham, v.45, p.5–32, jan. 2001.

Chang, C-C.; Lin, C-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. v.2, n.3, Artigo 27 – 27p., abr, 2011.

Cintra, M. E.; Meira, C. A. A.; Monard, M. C.; Camargo, H. A.; Rodrigues, L. H. A. 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, ES, Proceedings... Córdoba: IEEE, p. 1347-1352, 2011.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. AI Magazine, v.17, n.3, p.37-54, jul., 1996.

Hall, M. A.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update; SIGKDD Explorations. New York, v.11, n.1, p. 10-18, jun., 2009.

Haykin, S. Neural Networks and Learning Machines. 3ed., Englewood Cliffs: Prentice-Hall. 2009.

Luaces, O.; Rodrigues, L. H. A.; Meira, C. A. A.; Bahamonde; A.. Using nondeterministic learners to alert on coffee rust disease. Expert systems with applications, 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. 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. v.44, n.3, p.233–242, mar., 2009.

Witten, I. H.; Frank, E.; Hall, M. A. Data mining: practical machine learning tools and techniques. 3ed. San Francisco: Morgan Kaufmann, 2011.

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. O estado da arte de tecnologias na produção de café.Viçosa: Suprema Gráfica Editora, 2002. p. 369-449.