MINERAÇÃO DE DADOS NO DESENVOLVIMENTO DE SISTEMAS DE ALERTA CONTRA DOENÇAS DE CULTURAS AGRÍCOLAS

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

RESUMO

Este artigo apresenta um projeto de pesquisa em execução com a finalidade de avaliar a aplicação de tarefas de mineração de dados no desenvolvimento de sistemas de alerta contra doenças de culturas agrícolas. Através de processos de descoberta de conhecimento em bancos de dados, espera-se obter modelos de previsão para a ferrugem do cafeeiro (*Hemileia vastatrix*) e para a mancha preta do amendoim (*Cercosporidium personatum*), tal que possam vir a fazer parte futuramente de um sistema de monitoramento agrometeorológico.

PALAVRAS-CHAVE: descoberta de conhecimento em bancos de dados; sistema de previsão; modelo de previsão; *Hemileia vastatrix*; *Cercosporidium personatum*.

DATA MINING ON THE DEVELOPMENT OF PLANT DISEASE WARNING SYSTEMS

ABSTRACT

This article presents a research project which objective is to evaluate data mining tasks on the development of plant disease warning systems. Knowledge discovery processes will take place to discover predictive models for the coffee rust (*Hemileia vastatrix*) and the late leaf spot of peanut (*Cercosporidium personatum*) that could be incorporated in a meteorological monitoring system.

KEYWORDS: knowledge discovery in databases – KDD; prediction system; predictive model; *Hemileia vastatrix*; *Cercosporidium personatum*.

1. INTRODUÇÃO

A racionalidade no uso de agrotóxicos contribui, de um lado, para o bolso do produtor, reduzindo gastos com defensivos e com mão de obra, e de outro lado, para diminuir os riscos de contaminação da água e dos alimentos cultivados, em época que a população mundial se

[1] Mestre em Ciências da Computação e Doutorando em Engenharia Agrícola. Embrapa Informática Agropecuária. Caixa Postal 6041 – CEP 13083-970 – Campinas, SP. E-mail: carlos@cnptia.embrapa.br.
[2] Mestre em Engenharia Elétrica e Doutor em Sistemas de Suporte à Decisão. FEAGRI-UNICAMP. Caixa Postal 6011 – CEP 13083-970 – Campinas, SP. E-mail: lique@agr.unicamp.br.

preocupa, cada vez mais, com o consumo de produtos saudáveis e com a proteção do meio ambiente.

Um dos meios de promover o uso racional de agrotóxicos são os sistemas de previsão ou de alerta para o controle de doenças de culturas agrícolas, principalmente daquelas causadas por fungos (REIS, 2004; REIS et al., 2000). O aviso antecipado das condições predisponentes de uma doença permite diminuir o número de aplicações de defensivos agrícolas, em comparação com os esquemas convencionais de controle baseados num calendário fixo (PEDRO JR. e MORAES, 1992).

Entretanto, os sistemas de alerta são pouco empregados na prática pelos produtores. As dificuldades passam pela complexidade do modelo de previsão – normalmente, quanto maior a acurácia do modelo maior o número de variáveis exigidas –, pela necessidade de investimentos em equipamento para a obtenção de parâmetros meteorológicos e pelo emprego de mão de obra de acompanhamento do sistema e de manutenção do equipamento.

O avanço tecnológico dos últimos anos, com a instalação pela administração pública de inúmeras estações meteorológicas automáticas, com a organização de bancos de dados integrados e a disponibilidade de sistemas de monitoramento agrometeorológico via internet, e com a proliferação de técnicas avançadas de análise de dados, permite se pensar num sistema de alerta para o controle de doenças de plantas de alcance público, gratuito e simples de usar.

A análise de dados meteorológicos junto com dados de incidência de doenças em culturas agrícolas causadas por fungos, como a ferrugem do cafeeiro (*Hemileia vastatrix*) e a mancha preta do amendoim (*Cercosporidium personatum*), caracterizada como um processo de descoberta de conhecimento em bancos de dados (FAYYAD et al., 1996), indicará a viabilidade de uso dos padrões/modelos descobertos – em termos de confiança, de antecipação e de outras medidas cabíveis – na emissão de alertas contra essas doenças, como produto integrante futuro de um sistema de monitoramento agrometeorológico.

O objetivo é avaliar tarefas de mineração de dados quando aplicadas no desenvolvimento de modelos de previsão de doenças em culturas agrícolas, buscando estabelecer procedimentos de emissão de alertas, baseados em monitoramento agrometeorológico, quanto à incidência da ferrugem do cafeeiro e da mancha preta do amendoim. Pretende-se também caracterizar o processo de descoberta de conhecimento

realizado, tal que permita sua reprodução e adaptação com vistas a aplicá-lo em problemas similares de outras culturas agrícolas e dessas mesmas culturas, para outras doenças ou pragas.

Além da disponibilidade de dados confiáveis para análise, a escolha da ferrugem do cafeeiro e da mancha preta do amendoim se justifica pela importância dessas doenças (VALE e ZAMBOLIM, 1997; ZAMBOLIM, 2002). A ferrugem é a principal doença do cafeeiro em todo o mundo e pode ser encontrada em todas as lavouras de café cultivadas no Brasil. As manchas foliares do amendoim são consideradas as mais importantes em todas as regiões produtoras. No Estado de São Paulo, maior produtor nacional, a mancha preta tem se mostrado predominante e a mais severa.

2. MATERIAL E MÉTODOS

A fonte de dados para análise da ferrugem do cafeeiro é a Fazenda Experimental de Varginha, MG, da Fundação Procafé, ligada ao Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Refere-se ao acompanhamento mensal do índice de infecção, entre os anos agrícolas de 1998/1999 e 2004/2005, obtido a partir de folhas coletadas de talhões sem controle de ferrugem, em lavouras com espaçamento adensado e largo, sendo que para os dois espaçamentos foram utilizadas lavouras com carga pendente alta e baixa. Os dados agrometeorológicos correspondem a registros de uma estação meteorológica automática localizada na fazenda, que, em intervalos de 30 minutos, registra dados como temperaturas máxima e mínima, precipitação pluviométrica, radiação solar, fluxo e direção do vento, umidade relativa do ar e molhamento foliar.

Os dados relacionados com a mancha preta do amendoim foram obtidos junto ao Instituto Agronômico de Campinas (IAC). Referem-se a experimentos de avaliação de diferentes tratamentos para o controle das manchas foliares no cultivar Tatu, realizados nas estações experimentais localizadas em Pindorama, Ribeirão Preto e Adamantina, SP, entre os anos agrícolas de 1995/1996 e 2002/2003. O conjunto de dados é composto por duas planilhas para cada experimento: uma delas com registros diários de temperatura mínima, de períodos em horas com umidade relativa do ar acima de 90% e de chuva, associados com o tempo em dias após o plantio; a outra planilha com registros semanais da evolução do índice de infecção, da correspondente área sob a curva de progresso da doença (ASCPD) e do impacto

final na produtividade da cultura. Estão sendo considerados para análise os dados das áreas testemunhas sem controle com pulverização.

Inclui-se ainda como fontes de dados auxiliares das estações experimentais do IAC: o sistema CIIAGRO (IAC, 2005), que reúne dados diários das estações meteorológicas desses locais; arquivos com registros diários das estações automáticas, a partir da época que foram implantadas; e diagramas impressos de termohigrógrafos, com registros contínuos de temperatura e de umidade relativa do ar. Essas fontes de dados podem ser úteis, por exemplo, como base de consistência dos dados registrados nas planilhas ou para obtenção de valores ausentes, e abrem possibilidade de se analisar os dados num nível de granularidade menor.

O planejamento, a execução e o acompanhamento deste projeto de descoberta de conhecimento em banco de dados estão baseados no modelo de processo de mineração de dados CRISP-DM (*CRoss Industry Standard Process for Data Mining*), que divide o ciclo de vida de projeto em seis fases (CROSS..., 2004): compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição. Vale ressaltar que não existe uma seqüência rigorosa entre essas fases, sendo quase sempre necessário voltar e seguir em frente entre diferentes fases.

Técnicas e ferramentas de preparação dos dados serão selecionadas e empregadas (PYLE, 1999), e o *software* SAS (www.sas.com) será utilizado para análises estatísticas das variáveis e durante a fase de modelagem. Outra ferramenta adotada é o *software* WEKA (www.cs.waikato.ac.nz/~ml/weka), de domínio público e isento de pagamento de licença. É uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados (WITTEN e FRANK, 1999). Fonte adicional de informação e de programas para mineração de dados é o site KDnuggets™ (KNUGGETS, 2005).

A seleção das melhores regras será feita com o auxílio das ferramentas de mineração, por meio de suas informações de avaliação, e com o auxílio de medidas de desempenho e de qualidade de regras que têm sido largamente pesquisadas. Dentre as medidas de desempenho, encontram-se a precisão, o erro, a confiança negativa, a sensitividade, a especificidade, a cobertura, o suporte e a satisfação (RESENDE, 2002). Destacam-se como medidas de qualidade, a compreensibilidade e a interessabilidade (FREITAS, 1999).

3. RESULTADOS E DISCUSSÃO

Os principais resultados esperados são: (1) um conjunto de regras para a ferrugem do cafeeiro e (2) um conjunto de regras para a mancha preta do amendoim, escolhidos como os melhores dentre os demais descobertos, segundo os critérios de avaliação adotados, que permitam identificar os períodos críticos de incidência de cada doença e que permitiriam reduzir o número de pulverizações em relação ao calendário fixo convencional; e (3) um inventário do processo de descoberta de conhecimento, ou seja, descrição, discussão e caracterização detalhadas das atividades realizadas, das dificuldades encontradas, dos acertos e erros cometidos, especialmente das fases de preparação dos dados, de modelagem e de avaliação, que possa ser útil em projetos futuros similares.

Até o presente momento, foram realizadas atividades relativas às fases de compreensão do domínio e de entendimento dos dados. A seguir é feita uma breve descrição das fases de um projeto de mineração de dados, segundo a metodologia CRISP-DM (CROSS..., 2004), incluindo discussão referente a este projeto de pesquisa em particular.

Compreensão do domínio: fase inicial para entender os objetivos e requisitos do projeto pela perspectiva do domínio de aplicação, e depois converter esse conhecimento na definição do problema e no plano detalhado para atingir os objetivos.

A compreensão de como as interações entre a planta hospedeira, o patógeno e o meio ambiente determinam o desenvolvimento de doenças de plantas é fundamental, em especial da ferrugem do cafeeiro e da mancha preta do amendoim (MORAES, 1983; VALE e ZAMBOLIM, 1997; ZAMBOLIM, 2002).

Entendimento dos dados: começa com uma coleção inicial de dados e prossegue com atividades para se familiarizar com eles, identificar seus problemas de qualidade e se ter as primeiras compreensões (*insights*) do problema.

Esse entendimento dos dados envolve: descrição dos dados que foram obtidos, incluindo o formato dos dados, a quantidade de dados, o significado dos atributos e outras características dos dados; exploração dos dados, que inclui a distribuição dos valores de atributos chave, a relação entre pares ou pequenos grupos de atributos e análises estatísticas simples; e verificação da qualidade dos dados, examinando questões como: Os dados estão completos? Eles estão corretos ou contêm erros? Existem dados ausentes? Onde isso ocorre e com que freqüência?

Preparação dos dados: cobre todas as atividades para construir o conjunto de dados final – dados que alimentarão as ferramentas –, a partir dos dados iniciais brutos. Tarefas de preparação podem ser repetidas várias vezes, sem ordem predeterminada. Incluem seleção de tabelas, de registros e de atributos, bem como transformações e limpeza dos dados para as ferramentas de mineração.

A preparação dos dados é uma fase importante dentro do processo, que consome grande parte do tempo e, em muitos projetos, não recebe a devida atenção. O desafio é preparar os dados de forma que a informação contida neles seja exposta da melhor maneira para as ferramentas de mineração (PYLE, 1999), momento em que as atividades de compreensão do domínio fazem diferença.

Modelagem: nesta fase, várias técnicas de mineração são selecionadas e aplicadas, e seus parâmetros são calibrados para valores ótimos. Tipicamente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas têm requisitos específicos quanto ao formato dos dados. Portanto, voltar para a fase de preparação é freqüentemente necessário.

Pretende-se utilizar diferentes ferramentas e algoritmos, que implementam as tarefas de classificação e análise seqüencial para predição. Dentre as técnicas pertinentes, pode-se citar a indução de árvores de decisão, a indução de regras de classificação e a descoberta de padrões seqüenciais (HAN e KAMBER, 2001; p. 3-14, 1995.AGRAWAL e SRIKANT,).

Avaliação: neste estágio do projeto têm-se padrões (expressões ou modelos) que parecem ter boa qualidade. Antes de prosseguir, é importante avaliá-los completamente, com o auxílio de especialistas, e rever os passos executados.

Como colaboradores, participam especialistas da Embrapa, do IAC e do Centro de Ensino e Pesquisa em Agricultura da Unicamp (Cepagri), que auxiliarão nas etapas de preparação dos dados e de avaliação das regras obtidas na fase de mineração de dados.

Distribuição: descoberta de padrões geralmente não é o fim de um projeto de mineração. O conhecimento adquirido deve ser organizado e apresentado de modo que o usuário possa aproveitá-lo.

Nesta fase, pretende-se organizar e apresentar o conhecimento adquirido, de forma que os especialistas possam compreendê-lo, e que permita, com um esforço de programação em

linguagem de computador, ser incorporado num sistema de monitoramento agrometeorológico, para que os alertas aos produtores de café e de amendoim sejam emitidos.

4. AGRADECIMENTOS

À Fundação Procafé/MAPA por ceder os dados relacionados com o monitoramento de doenças da cultura do café, em específico da ferrugem do cafeeiro; ao Instituto Agronômico de Campinas, em especial ao pesquisador Dr. Sérgio Almeida de Moraes, por ceder os dados de seus experimentos com as manchas foliares do amendoim.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R.; SRIKANT, R. Mining sequential patterns. In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 11, 1995. **Proceedings…** p. 3-14, 1995.

CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING. **CRISP-DM – home**. Disponível em: <http://www.crisp-dm.org/index.htm>, Acesso em: 29 jun. 2004.

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

FREITAS, A. A. On rule interestingness measures. **Knowledge-Based Systems**, 12(5-6): 309-315, 1999.

HAN, J.; KAMBER, M. **Data mining**: concepts and techniques. San Francisco: Morgan Kaufmann Publishers, 2001. 550 p.

IAC. **Centro Integrado de Informações Agrometeorológicas**. Disponível em: <www.iac.sp.gov.br/ciiagro>. Acesso em: 25 jul. 2005.

KDNUGGETS. **KDnuggets**: data mining, web mining, and knowledge discovery guide. Disponível em: <www.kdnuggets.com>. Acesso em: 25 jul. 2005.

MORAES, S. A. **A ferrugem do cafeeiro**: importância, condições predisponentes, evolução e situação no Brasil. Campinas: Instituto Agronômico, 1983. 50p. (IAC. Circular, 119).

PEDRO JR., M. J.; MORAES, S. A. Racionalização pela previsão de ocorrência de doenças. **Summa Phytopathologica**, Jaboticabal, v. 18, n. 1, p. 62, 1992.

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

REIS, E. M.; FORCELINI, C. A.; BRESOLIN, A. C. R. **A informática na previsão de epidemias de doenças de plantas**. In: INFOAGRO 2000 CONGRESSO E MOSTRA DE AGROINFORMÁTICA, 2000, Ponta Grossa. PontaGrossa: UEPG, 2000. n.p. Palestra.

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

RESENDE, S. O. (Coord.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Editora Manole. 2002. 525 p.

VALE, F. X. R. do; ZAMBOLIM, L. (Ed.). **Controle de doenças de plantas**: grandes culturas. Viçosa: UFV, v. 1, 1997.

WITTEN, I. H.; FRANK, E. **Data mining**: practical machine learning tools and techniques with java implementations. San Francisco: Morgan Kaufmann, 1999. 416 p.

ZAMBOLIM, L. (Ed.). **O estado da arte de tecnologias na produção de café**. Viçosa: UFV, 2002. 568 p.