ENTENDIMENTO E PREPARAÇÃO DE DADOS NO PROCESSO DE DESCOBERTA DE CONHECIMENTO APLICADO A SISTEMA DE ALERTA DA FERRUGEM DO CAFEEIRO

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

[1] Mestre em Ciências de Computação, Pesquisador EMBRAPA, Doutorando Engenharia Agrícola, FEAGRI/UNICAMP, Campinas – SP, (19) 3789-5839, carlos.meira@agr.unicamp.br.
[2] Engº Agrícola, Prof. Doutor, FEAGRI, UNICAMP, Campinas - SP.

Escrito para apresentação no
XXXV Congresso Brasileiro de Engenharia Agrícola
31 de julho a 04 de agosto de 2006 - João Pessoa - PB

RESUMO: Descoberta de Conhecimento em Bases de Dados é a extração de informação implícita, previamente desconhecida e potencialmente útil a partir de dados. É um processo que estabelece um pré-processamento inicial nos dados a fim de ajudar na compreensão do problema e de expor os dados da melhor maneira para as fases posteriores. Este trabalho discute e apresenta os resultados do entendimento e da preparação dos dados de uma instância do processo de descoberta de conhecimento aplicado a sistema de alerta da ferrugem do cafeeiro, um dos meios de promover o uso racional de agrotóxicos nas lavouras de café. Diferentes procedimentos de investigação e de manipulação dos dados foram adotados, e os resultados obtidos, na forma de descrição dos dados, de problemas identificados na sua qualidade e de conjuntos de dados preparados, confirmam a importância desse tipo de atividade para o êxito do projeto.

PALAVRAS-CHAVE: DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS, MINERAÇÃO DE DADOS, *HEMILEIA VASTATRIX*.

DATA UNDERSTANDING AND DATA PREPARATION IN THE PROCESS OF KNOWLEDGE DISCOVERY APPLIED TO A COFFEE RUST FORECASTING SYSTEM

ABSTRACT: Knowledge Discovery in Databases is the extraction of implicit, previously unknown and potentially useful information from data. It is a process that establishes an initial data pre-processing which aims to help understand the problem and to better expose the data for the subsequent phases. This work discusses and presents the results of data understanding and data preparation of a knowledge discovery process instance applied to a coffee rust forecasting system. Different investigation and manipulation procedures has been adopted, and the results, in the form of data descriptions, identified data quality problems and prepared data sets, confirm the importance of this type of activity on the success of the project.

KEYWORDS: KNOWLEDGE DISCOVERY IN DATABASES - KDD, DATA MINING, *HEMILEIA VASTATRIX*.

INTRODUÇÃO: Sistemas de alerta de doenças de plantas dão suporte à tomada de decisão ao indicar as condições que favorecem uma doença, permitindo agir somente quando necessário e podendo trazer diminuição no uso de agrotóxicos (REIS, 2004). São pouco utilizados na prática pela dificuldade de obtenção dos dados necessários e pelo custo de implementação e de manutenção para o agricultor. O desenvolvimento tecnológico pode ajudar a melhorar essa situação, pela disponibilidade de estações meteorológicas automatizadas, de bancos de dados meteorológicos e de técnicas avançadas de análise de dados. A análise de dados meteorológicos junto com registros de intensidade de doenças causadas por fungos em culturas agrícolas, como a ferrugem do cafeeiro, caracterizada como um processo de descoberta de conhecimento em bases de dados (FAYYAD et al., 1996), indicará a viabilidade de uso

dos modelos descobertos na emissão de alertas, como produto integrante futuro de um sistema de monitoramento agrometeorológico de alcance público. O objetivo é avaliar tarefas e técnicas de mineração de dados no desenvolvimento de modelos de previsão de doenças, e caracterizar o processo de descoberta de conhecimento para utilizá-lo em problemas similares. Pretende-se também obter um modelo de previsão útil e confiável para a ferrugem do cafeeiro. A ferrugem, causada pelo fungo *Hemileia vastatrix* Berk et Br., é considerada a principal doença da cultura, proporcionando decréscimos de produção que variam de 35 a 50% (ZAMBOLIM et al., 1997). Além da importância econômica, atende outros requisitos que justificam um sistema de alerta (COAKLEY, 1988): a doença varia entre as estações de cultivo e existem medidas de controle economicamente viáveis. O presente trabalho tem como objetivo discutir e apresentar os resultados de duas fases iniciais do processo, o entendimento e a preparação dos dados.

MATERIAL E MÉTODOS: O planejamento, a execução e o acompanhamento do projeto 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 em seis fases (CHAPMAN et al., 2000): compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição. Não existe uma seqüência rigorosa entre elas, sendo quase sempre necessário voltar e seguir em frente entre as diferentes fases. O entendimento dos dados começa com uma coleção inicial e prossegue com atividades para se familiarizar com os dados, para identificar problemas de qualidade nesses dados e para buscar as primeiras compreensões (insights) a partir deles. Esta fase envolve obter os dados, descrever os dados, explorar os dados e verificar a sua qualidade. A preparação dos dados reúne as atividades de construção do(s) conjunto(s) de dados para a fase de modelagem, a partir dos dados iniciais brutos. É uma fase muito importante, que consome grande parte do tempo. 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). Um aspecto importante dos dados são as séries temporais, ou seja, atributos que são medidos ao longo do tempo em intervalos fixos, como a temperatura, a umidade e a precipitação. São necessárias transformações nos dados e derivação de novos atributos tal que a dimensão temporal seja incorporada no formato de dados usual "linhas e colunas" reconhecido pelos algoritmos tradicionais de mineração. Também, devido à escolha pelas técnicas de indução de árvore de decisão e de regras de classificação (MONARD e BARANAUSKAS, 2002), por permitirem a extração de padrões compreensíveis, é preciso fazer com que o atributo meta ou classe, que representa a intensidade da doença, assuma valores discretos em vez de valores contínuos. É necessário definir valores categóricos, correspondentes a intervalos de valores contínuos, para representar diferentes níveis de intensidade (p. ex. alta, média e baixa). Este procedimento pode ser realizado automaticamente, mas é importante o apoio de especialistas no domínio, pois é por meio desses níveis de intensidade que futuramente será definido o limite crítico para a emissão dos alertas. O suporte ferramental para essas atividades é dado por um editor de planilhas eletrônicas e pela linguagem de programação Perl, considerada uma linguagem poderosa de manipulação de dados. Os dados relacionados com a ferrugem do cafeeiro são da Fundação Procafé, ligada ao Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Ela mantém uma estação de avisos fitossanitários, instalada na Fazenda Experimental de Varginha – MG (altitude 1.010m, latitude 21° 34' 00'' e longitude 45° 24' 22''), que emite mensalmente boletins aos produtores, técnicos e órgãos ligados ao setor. Os dados se referem ao acompanhamento entre os anos agrícolas de 1998/1999 a 2004/2005. Para o monitoramento da infecção de ferrugem foram selecionadas, a cada ano, oito áreas em lavouras de café em produção, sendo quatro áreas de lavouras em espaçamento largo e as demais adensadas. Foram coletadas folhas de talhões sem controle de ferrugem, sendo que para os dois espaçamentos foram utilizadas lavouras com carga pendente alta e baixa. As amostras foram obtidas coletando-se 100 folhas do terço médio das plantas, entre o terceiro e quarto par de folhas em cada talhão, e contando-se o número de folhas com lesões de ferrugem. Na fazenda, existe uma estação meteorológica automatizada, marca Davis modelo Groweather industrial, que registra dados meteorológicos a cada 30 minutos, tais como temperaturas máxima e mínima, precipitação, radiação solar incidente, fluxo e direção do vento, umidade relativa e molhamento foliar. O conjunto de dados recebido é composto por três tipos de arquivo, para cada mês do ano: um arquivo texto (.txt) com os valores dos atributos registrados pela estação meteorológica; uma planilha (.xls) com valores diários

consolidados de atributos selecionados do registro da estação; e um documento (.doc) referente ao boletim de avisos mensal.

RESULTADOS E DISCUSSÃO: Todo o conjunto de dados foi recebido da Fundação Procafé por meio de correio eletrônico. Os dados foram examinados e descritos detalhadamente na forma de relatório para cada um dos três tipos de arquivo mencionados. Com o intuito de verificar a qualidade dos dados, gerou-se automaticamente arquivos com valores diários consolidados, a partir dos arquivos da estação meteorológica, para comparação com as planilhas eletrônicas correspondentes recebidas. Alguns problemas foram constatados, como falhas no registro da estação não identificadas pelo pessoal de Varginha e erros comuns de digitação e de cálculo devido à participação humana na elaboração das planilhas. Também para verificar a qualidade dos dados, elaborou-se gráficos de evolução da incidência de ferrugem a partir dos boletins mensais. Foi possível constatar comportamentos estranhos da doença, que precisam ser analisados com mais detalhe e com o auxílio de especialistas. A primeira atividade de preparação de dados foi montar o que se está chamando de conjunto de dados primário. Os dados em estado bruto foram selecionados, descartando aqueles sem significado para o problema, e foram reestruturados, colocando-os em um formato próximo do que é usado na modelagem. A partir desse conjunto primário é que são construídos os conjuntos de dados para cada iteração da fase de modelagem. Para isso estão sendo realizadas transformações nos dados, derivação de novos atributos, integração dos dados meteorológicos com os registros de incidência da ferrugem e conversão dos dados para os formatos das ferramentas de mineração (Figura 1). Informações pesquisadas em relação à epidemiologia da ferrugem do cafeeiro são utilizadas na determinação dessas operações de transformação e derivação de dados: a temperatura ótima para germinação e penetração do fungo varia de 22 a 24º C; temperaturas superiores a 30º C e inferiores a 14º C são limitantes para a infecção; um período de seis horas de água livre na superfície da folha é o tempo mínimo necessário para ocorrer infecção; o período de incubação está em média entre 25 e 30 dias (ZAMBOLIM et al., 1997; ZAMBOLIM et al., 2002). Além da medida de molhamento foliar feita pela estação meteorológica, serão analisadas também algumas medidas indiretas de molhamento que consideram períodos de alta umidade relativa do ar (p.ex. NHUR90 – número de horas com umidade relativa maior ou igual a 90%) ou a intensidade das chuvas (p.ex. precipitação maior ou igual a 2,5mm). Como é comum a repetição dessas atividades de preparação, vem sendo desenvolvidos programas em Perl com o intuito de automatizar ao máximo os procedimentos. As fases seguintes do processo, a modelagem e a avaliação, ainda não executadas, é que vão permitir a obtenção dos modelos de previsão da ferrugem e a avaliação dos resultados obtidos. Cada conjunto de dados preparado será submetido à ferramenta de mineração, que produzirá, dependendo do algoritmo de indução utilizado, uma árvore de decisão ou um conjunto de regras de classificação. Essas próprias técnicas serão as responsáveis por escolher os melhores atributos meteorológicos, quer sejam originais medidos pela estação ou derivados na preparação, para predizer o nível de intensidade da doença. Os modelos escolhidos, com a participação de especialistas do domínio, devem permitir predizer a intensidade da ferrugem do cafeeiro para cada uma das combinações de espaçamento (lavoura larga ou adensada) e de produção (carga pendente alta ou baixa) da cultura. Dadas as condições meteorológicas requeridas, por meio da árvore de decisão ou do conjunto de regras de classificação, dependendo do modelo usado, vai ser possível identificar o nível previsto de incidência da ferrugem, com o grau de precisão fornecido pelo modelo, e decidir sobre o alerta, de acordo com o limiar de ação estabelecido.

CONCLUSÕES: Na etapa em que se encontra a execução do projeto, foi possível verificar que as atividades de entendimento e de preparação de dados ajudam bastante na compreensão do problema e na garantia de qualidade dos dados a serem utilizados nas fases seguintes. Isto confirma o que se encontra na literatura, mas é sabido também que muitos projetos não dão a devida atenção ao pré-processamento. Pode-se concluir também que investir esforço na automatização das atividades de preparação de dados não significa perda de tempo. É despendido um tempo maior no início, mas esse tempo é recuperado no decorrer do projeto à medida que a preparação de dados vai se repetindo em cada iteração do processo. Além disso, e muito importante, com a automatização é possível evitar

certos erros comuns originados por intervenção manual, que poderiam vir a interferir nos resultados pretendidos.

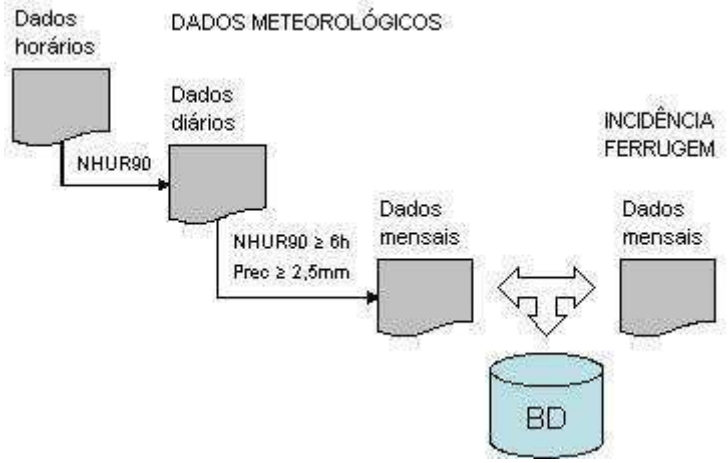

Figura 1: Esquema de preparação dos dados para a fase de modelagem.

AGRADECIMENTOS: À Fundação Procafé/MAPA por ceder os dados relacionados com o monitoramento da ferrugem do cafeeiro, em especial ao Eng° Agrônomo Leonardo Bíscaro Japiassú.

REFERÊNCIAS BIBLIOGRÁFICAS

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

COAKLEY, S. M. Variation in climate and prediction of disease in plants. **Annual Review of Phytopathology**, v. 26, p. 163-181, 1988.

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.

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

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

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

ZAMBOLIM, L.; VALE, F. X. R.; COSTA, H.; PEREIRA, A. A.; CHAVES, G. M. Epidemiologia e controle integrado da ferrugem-do-cafeeiro. In: ZAMBOLIM, L. (Ed.). **O estado da arte de tecnologias na produção de café**. Viçosa: Suprema Gráfica e Editora, p. 369-449, 2002.

ZAMBOLIM, L.; VALE, F. X. R. do; 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. do; ZAMBOLIM, L. (Ed.). **Controle de doenças de plantas**: grandes culturas. Viçosa: UFV, v. 1, p. 83-139, 1997.