PREPARAÇÃO DE DADOS PARA A MODELAGEM DE OCORRÊNCIAS DA FERRUGEM ASIÁTICA DA SOJA

GUILHERME AUGUSTO SILVA MEGETO[1]
CARLOS ALBERTO ALVES MEIRA[2]
STANLEY ROBSON DE MEDEIROS OLIVEIRA[3]
SILVIA MARIA FONSECA SILVEIRA MASSRUHÁ[4]
CLAUDIA VIEIRA GODOY[5]

RESUMO: A ferrugem asiática da soja é uma das principais doenças da cultura. As perdas estimadas em grãos com a doença no Brasil são significativas. O objetivo deste trabalho foi realizar a preparação dos dados para a modelagem de ocorrências da ferrugem asiática, visando descobrir padrões que permitam predizer o risco do aparecimento de focos da doença. Os dados brutos se constituíram de registros de ocorrência da ferrugem em diferentes locais do Brasil, de 2004 a 2009, obtidos do Consórcio Antiferrugem, e de registros meteorológicos da base de dados do Agritempo. A integração dos dados meteorológicos com os registros de ocorrência foi feita com base no período latente do patógeno. Atributos novos foram criados, enquanto outros foram eliminados, por irrelevância ou por falta de dados. A tarefa de mineração de dados escolhida inicialmente foi a classificação, e os casos de ocorrência da ferrugem definiram a classe positiva. Os casos da classe negativa foram criados para uma data anterior à de ocorrência da ferrugem, supondo-se não ter havido ocorrência antes da identificação confirmada. A preparação dos dados resultou no conjunto de treinamento para a modelagem com 7076 exemplos, divididos igualmente em 3038 casos de ocorrência e de não ocorrência. O número de ocorrências diminuiu bastante em relação à quantidade original, por falta de dados meteorológicos para vários dos locais de identificação, mas não inviabiliza a iniciativa de descoberta de conhecimento em bases de dados proposta.

PALAVRAS-CHAVE: *Phakopsora pachyrhizi*, mineração de dados, classificação, descoberta de conhecimento em bases de dados

DATA PREPARATION FOR MODELING THE OCCURENCE OF ASIAN SOYBEAN RUST

ABSTRACT: The Asian soybean rust is one of the most important diseases of this culture. The estimated grain losses are significant in Brazil. The aim of this study was to deal with the data preparation for modeling the occurrence of Asian rust to discover patterns that allow for predicting the risk of disease outbreaks, based on weather and the soybean culture information. The raw data consisted of records regarding the occurrence of Asian rust in various places in Brazil, from 2004 to 2009, obtained from Consórcio Antiferrugem, and from weather records of some part of these locations, stored in Agritempo database. The integration of meteorological data with records of the disease occurrence was performed based on the latent period of its pathogen. New attributes were created, while others were eliminated due to

1 Graduando em Matemática Aplicada e Computacional, Estagiário da Embrapa Informática Agropecuária, e-mail: guilhermeasm@cnptia.embrapa.br
2 Doutor em Engenharia Agrícola, Pesquisador da Embrapa Informática Agropecuária, e-mail: carlos@cnptia.embrapa.br
3 Doutor em Ciência da Computação, Pesquisador da Embrapa Informática Agropecuária, e-mail: stanley@cnptia.embrapa.br
4 Doutora em Computação Aplicada, Pesquisadora da Embrapa Informática Agropecuária, e-mail: silvia@cnptia.embrapa.br
5 Doutora em Fitopatologia, Pesquisadora da Embrapa Soja, e-mail: godoy@cnpso.embrapa.br

their irrelevance or the presence of missing values. Classification was the originally chosen data mining task, and the occurrence cases of the disease defined the positive class. The cases of the negative class were created considering a date before the date of the rust occurrence, assuming that there was no occurrence before the confirmation of the disease identification. The data preparation resulted in a training set for the modeling with 7076 examples, equally divided into 3038 cases of occurrence and no occurrence. The number of occurrences in the training set decreased considerably in relation to the original raw data due to the lack of meteorological data in several points of the disease identification, but it does not invalidate the initiative of knowledge discovery in the proposed dataset.

KEYWORDS: *Phakopsora pachyrhizi*, data mining, classification, knowledge discovery in databases, KDD

1. INTRODUÇÃO

A soja (*Glycine max*), atualmente, no Brasil, tem como um dos principais patógenos o fungo *Phakopsora pachyrhizi* causador da ferrugem asiática da soja. Esta doença é bastante agressiva e resulta na desfolha precoce, ocasionando deficiências na formação e enchimento das vagens e no peso final dos grãos (HENNING e GODOY, 2006). As perdas estimadas em grãos com a ferrugem na safra 2007/2008 foram próximas de 418,5 mil toneladas (EMBRAPA, 2009). A principal forma de combate à doença é o defensivo químico (HENNING e GODOY, 2006).

A ferrugem, entre outras doenças de plantas, é bastante dependente do clima, sendo o molhamento foliar e a temperatura os principais fatores que influenciam a infecção e o desenvolvimento da doença (ALVES *et al,* 2006). Com o conhecimento das condições meteorológicas que favorecem, ou não, o desenvolvimento da ferrugem, técnicas de manejo da cultura podem ser otimizadas, como escolha da época de plantio, imposição do "vazio sanitário" (HENNING e GODOY, 2006), e, principalmente, uso consciente e sustentável de fungicidas, gerando economia para o produtor e proteção ao meio ambiente.

Uma das alternativas para abordar o problema da ferrugem asiática é a descoberta de conhecimento em bases de dados (*KDD*). O processo de KDD, também conhecido como mineração de dados (*data mining*), é responsável por auxiliar, por meio de algoritmos computacionais, a descoberta de padrões ocultos em repositórios de dados. Após a descoberta de padrões nos dados, aliada ao conhecimento do domínio e com o auxílio de especialistas, é possível tomar decisões com base nesses novos conhecimentos (WITTEN e FRANK, 2005).

Neste trabalho, foi realizada a fase de preparação dos dados de uma iniciativa de KDD, que envolve a aquisição de dados das bases de dados, limpeza de dados, estatística descritiva, integração de dados, seleção e criação de atributos, entre outras tarefas.

2. OBJETIVO

O objetivo do trabalho foi preparar dados para a modelagem de ocorrências de ferrugem asiática da soja. O intuito dessa modelagem é a descoberta de regras que permitam predizer o risco de aparecimento de focos da doença com base em condições meteorológicas e dados sobre a cultura.

3. MATERIAL E MÉTODOS

Como base de dados para o estudo, foram consultadas duas principais fontes: o Consórcio Antiferrugem (www.consorcioantiferrugem.net) e o Agritempo (www.agritempo.gov.br).

Os dados do Consórcio Antiferrugem abrangem cinco safras (2004/05 a 2008/09) e contêm

diversas informações sobre ocorrências de ferrugem asiática da soja em todo o país. Incluem a localização (UF e Município), o estádio fenológico da planta (Estádio), se a área plantada é comercial, experimental etc. (Tipo de área), a instituição que identificou e confirmou a ocorrência (Identificação e Confirmação), a data em que foi verificada a ocorrência (Data da ocorrência), o cultivar da soja (Cultivar), a época em que foi plantada (Época de plantio) e observações sobre o clima no momento da verificação da ocorrência (Condições Climáticas).
Os dados meteorológicos diários, obtidos do Agritempo, incluem os atributos temperatura mínima, temperatura máxima e precipitação, medidos em diversas estações meteorológicas distribuídas pelo país.
Para a preparação dos dados em geral, foram criadas rotinas na linguagem de programação *Perl* (*Perl 5.8.8*), e também foram utilizados *softwares* com suporte a arquivos com extensão *csv* (*Comma separated values*), planilha eletrônica, estatística básica e geração de gráficos.
Antes da preparação dos dados, foram elaborados gráficos de freqüência (p.ex. ocorrências por estado, ano, mês etc.), análises de máximos e mínimos valores, e testes de consistência (p.ex cidades com nomes iguais e estados diferentes ou datas fora do intervalo pesquisado). O intuito dessa análise foi obter maior conhecimento do domínio, verificar se algum atributo poderia ser excluído e analisar possíveis falhas, relações e inconsistências nos dados.
Alguns procedimentos para a criação de novos atributos foram desenvolvidos com base nas características da soja, do patógeno e das variáveis meteorológicas. Como o interesse desse estudo está voltado para descobrir padrões que influenciaram a infecção, se fez necessário descobrir quando foi a possível infecção, uma vez que a base de dados possui apenas a data de ocorrência, ou seja, quando foram visualizados sintomas da doença. Por isso utilizou-se a equação, desenvolvida por Alves *et al.* (2006), que estima o período latente:

$$Y = 0{,}11\,T^2 - 5{,}20\,T + 69{,}53 \qquad (1)$$

em que Y é o período latente em dias e T é a temperatura.
Para obter a possível data de infecção, subtraiu-se o período latente estimado (Y) da data de ocorrência e, a partir dessa nova data, foram estudadas as condições meteorológicas de três dias anteriores com o objetivo de encontrar padrões que favoreceram a infecção.
A princípio, a tarefa de mineração de dados escolhida foi a classificação, utilizando a técnica de árvore de decisão, que cria modelos para a predição de eventos de cada classe. Como a base de dados fornece apenas os casos de ocorrência (classe positiva), se fez necessário criar eventos para a classe negativa, de não ocorrência. A solução adotada foi regredir um período fixo, a partir da data de ocorrência, e supor que, naquela data, não houve ocorrência.

4. RESULTADOS E DISCUSSÃO

Com o resultado da análise dos dados de ocorrências, foram descartados os atributos: “Tipo de área”, por apresentar uniformidade de classe, ou seja, mais de 90% dos eventos possuíam o tipo de área Comercial; “Identificação” e “Confirmação” por serem irrelevantes para o estudo proposto; “Época de plantio” e “Condições Climáticas” por apresentarem muitos dados faltantes; e “Cultivar” por todos os cultivares serem considerados suscetíveis à doença.
Além de atributos excluídos, houve também eliminação de ocorrências, por diversos motivos, principalmente por inconsistência nos dados (p.ex. registros em que a data da ocorrência indicava 01/01/0001). No geral, os dados continham poucos erros ou inconsistências.
A Figura 1 apresenta o esquema utilizado na preparação dos dados meteorológicos, com o intuito de relacioná-los com as ocorrências da ferrugem asiática da soja. Utilizou-se a temperatura média nos dias anteriores (T_i) à ocorrência da doença (Oc) para estimar o período latente (PL), de acordo com a equação 1, e identificar o provável dia de infecção (Inf). Os atributos meteorológicos foram criados para um período em dias (m) anterior ao dia de infecção, incluindo este.

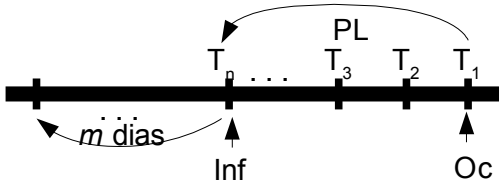

Figura 1 – Esquema da preparação dos dados meteorológicos. Oc. ocorrência da ferrugem; PL. período latente; T_1 a T_n. temperaturas médias nos dias do PL; Inf. provável dia de infecção; m dias. período considerado para derivar os atributos meteorológicos.

No presente trabalho, foram analisadas variáveis meteorológicas em três dias (m) anteriores à data de infecção. Assim, foram criados os atributos média da temperatura mínima nesses três dias (media_tmin_3), média da temperatura máxima (media_tmax_3), média da temperatura média (media_tmed_3), precipitação acumulada (precpt_acum_3), média de precipitação nos três dias (media_precpt_3) e o número de dias, dentre esses três dias, em que houve precipitação (precpt_dias_3), considerando valores maiores que zero.
Nessa fase da preparação dos dados, foi gerado um arquivo com as ocorrências que contém todos os atributos completos, tanto da cultura quanto meteorológicos. Inicialmente, o arquivo de ocorrências continha 9376 registros (linhas de ocorrências) em 764 cidades. Após o processo de tratamento de dados, o número de registros foi de 3038 em 139 cidades.
A Figura 2 ilustra a distribuição de ocorrências, entre os estados do Brasil, dos dados brutos e dos dados preparados. Com essa diferença de ocorrências entre os dados brutos e preparados, informação pode ter sido perdida, uma vez que quanto mais dados o conjunto de treinamento possui, melhor pode ser o resultado. A principal causa dessa diferença foi a falta de estações meteorológicas em algumas cidades com identificação de ocorrências da ferrugem asiática.

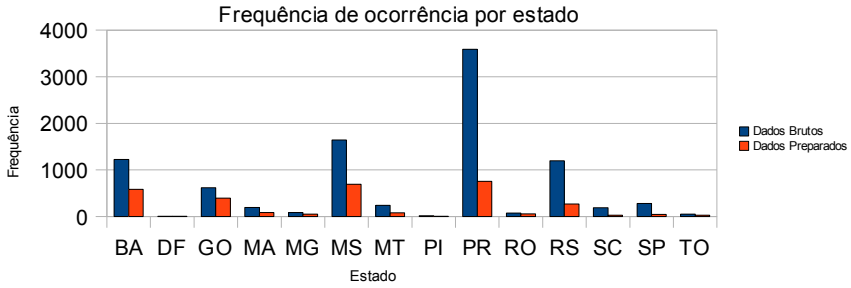

Figura 2 – Diferença de ocorrências entre os dados brutos e os dados preparados.

Como descrito na seção anterior, para se aplicar a tarefa de classificação, houve a necessidade de criação da classe negativa. Cada caso de ocorrência da ferrugem da soja teve um evento de não ocorrência associado, proporcionando um balanceamento exato entre as classes positiva e negativa.
A data de não ocorrência foi considerada dez dias antes da data de ocorrência observada, supondo-se que nessa data não havia sintoma da doença. Esse período fixo anterior à data de ocorrência foi denominado deslocamento de classe. Da mesma forma que foi estimado o período latente para os casos de ocorrência, também o foi para o evento de não ocorrência. A partir do período latente e a data da suposta não ocorrência, obteve-se a data denominada não infecção. Seguindo a mesma metodologia, foram analisadas as variáveis meteorológicas de três dias (m) anteriores à não infecção, com o objetivo de encontrar padrões nos dados para caracterizar as ocorrências e não ocorrências da ferrugem asiática da soja. O esquema de criação da classe negativa está apresentado na Figura 3.
O resultado da etapa de preparação dos dados foi o arquivo final contendo atributos

meteorológicos, informações sobre a cultura e a classe de predição. Com a criação da classe negativa, duplicou-se o número de registros, gerando assim 7076 linhas no conjunto de treinamento para realizar a tarefa de classificação. A partir desse arquivo, *softwares* específicos como o *Weka* (www.cs.waikato.ac.nz/ml/weka) e o *Rattle* (rattle.togaware.com) serão utilizados com algoritmos de árvore de decisão para tentar extrair os padrões desejados.

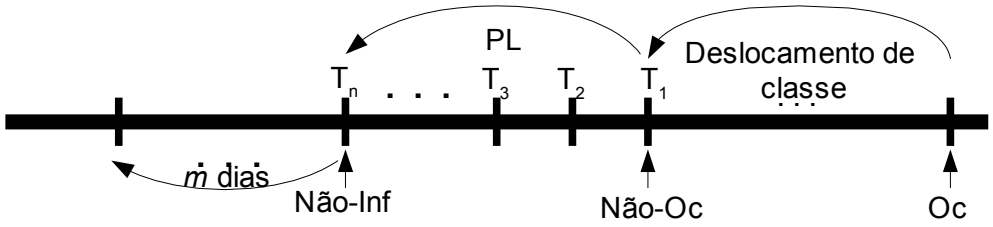

Figura 3 – Esquema de criação da classe negativa. Oc. data da ocorrência; Não-Oc. data da suposta não ocorrência, considerando um deslocamento de classe; PL. período latente; Não-Inf. data da suposta não infecção; m. dias anteriores à não infecção que foram analisados.

5. CONCLUSÕES E SUGESTÕES

A falta de dados meteorológicos em muitas das cidades para as quais existem registros de ocorrência da ferrugem asiática da soja reduziu bastante a quantidade de exemplos para a fase seguinte de modelagem. No entanto, não inviabiliza o prosseguimento do trabalho proposto.
Para contornar esse problema, há a possibilidade de uso de dados estimados pelo radar do satélite TRMM (*Tropical Rainfall Measuring Mission*) para precipitação e pelo Agritempo para as outras variáveis meteorológicas, como temperaturas máxima e mínima, com uma área de cobertura mais abrangente.
Outra abordagem para o "deslocamento de classe" seria considerar os estádios fenológicos da soja. Identificada a ocorrência em um determinado estádio, se suporia a não ocorrência no estádio anterior. O deslocamento seria, então, o tempo médio de transição entre um estádio fenológico e o outro.

6. REFERÊNCIAS

ALVES, S.A.M.; FURTADO, G.Q.; BERGAMIN FILHO, A. Influências das condições climáticas sobre a ferrugem da soja. In: ZAMBOLIM, L. (ed.). **Ferrugem asiática da soja**. Viçosa: Universidade Federal de Viçosa / Departamento de Fitopatologia. 2006. p.37-59.

EMBRAPA - EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Embrapa Soja divulga balanço sobre a ferrugem na safra 2007/08.** Disponível em: <http://www.cnpso.embrapa.br/noticia/ver_noticia.php?cod_noticia=469&desl=31>. Acesso em 18 de maio de 2009.

HENNING, A.A.; GODOY, C.V. Situação da ferrugem da soja no Brasil e no mundo. In: ZAMBOLIM, L. (ed.). **Ferrugem asiática da soja**. Viçosa: Universidade Federal de Viçosa / Departamento de Fitopatologia. 2006. p.1-14.

WITTEN, I.H.; FRANK, E.. **Data mining:** practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005. 525p.