MINERAÇÃO DE PADRÕES SEQÜENCIAIS CLIMÁTICOS PARA USO EM SISTEMA DE PREVISÃO DE DOENÇAS E DANOS NA CULTURA DO CAFÉ NO ESTADO DE SÃO PAULO

Carlos Alberto Alves Meira[1]
Luiz Henrique Antunes Rodrigues[2]

RESUMO
Este trabalho apresenta a proposta de mineração de padrões seqüenciais em bancos de dados agrometeorológicos visando descobrir regras de associação temporal que permitam prever, com confiança e antecipação aceitáveis, condições climáticas favoráveis ao surgimento e ao agravamento de doenças e danos na cultura do café no estado de São Paulo.

PALAVRAS-CHAVE: mineração de dados temporal, séries temporais, seqüências temporais, regras de associação temporal, descoberta de conhecimento em bancos de dados.

CLIMATE SEQUENTIAL PATTERN MINING TO USE IN A COFFEE DISEASE FORECASTING SYSTEM

ABSTRACT
This work presents a proposal of sequential pattern mining in climate databases to discover temporal association rules. These rules are intended to predict the conditions that favour the occurrence and severity of coffee diseases in São Paulo State, Brazil, with acceptable confidence and anticipation.

KEYWORDS: temporal data mining, time series, time sequences, temporal association rules, knowledge discovery in databases – KDD.

1. INTRODUÇÃO

Mineração de padrões seqüenciais é a descoberta de padrões freqüentes relacionados com o tempo ou outras seqüências. Um exemplo de padrão seqüencial: "Uma pessoa que adquiriu um PC nove meses atrás é um potencial comprador de uma nova CPU dentro de um mês" (Han & Kamber, 2000).

Descoberta de padrões seqüenciais é um importante problema de pesquisa de mineração de dados, com aplicações no mercado financeiro, comércio varejista, telecomunicações e medicina. De modo simplificado, dado um conjunto de seqüências de dados, o problema é descobrir subseqüências freqüentes tal que a porcentagem de seqüências que as contém atinge um mínimo especificado (Garofalakis et al., 2002).

A partir de padrões seqüenciais é possível obter regras de associação temporal, por exemplo, na forma $X \Rightarrow^{T} Y$, que indica que se X ocorrer então Y ocorrerá dentro do período de tempo T (Das et al., 1998). Regras de associação temporal são úteis para entender o comportamento de entidades de um domínio e para predizer o impacto da ocorrência de certos eventos na ocorrência de outro(s) evento(s).

No domínio agropecuário, são várias as situações em que eventos de interesse econômico, ambiental e/ou social dependem de fatores climáticos ao longo de um período de tempo. Exemplos são as geadas, os períodos de seca e as manifestações de doenças e danos em culturas agrícolas.

[1] Mestre em Ciências da Computação. 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.

2. OBJETIVOS

O objetivo do projeto de pesquisa apresentado neste trabalho é estabelecer, de maneira adequada, um processo e um conjunto de ferramentas específicos de mineração de dados em séries temporais agrometeorológicas para descoberta de padrões seqüenciais. Esses padrões seqüenciais podem desvendar ou reafirmar regras de associação temporal entre fatores climáticos e eventos de interesse econômico, ambiental e/ou social do setor agropecuário.

Tem-se como objetivo específico aplicar esse processo de mineração de padrões seqüenciais para descoberta de regras de associação temporal entre fatores climáticos e a ocorrência, bem como a severidade, de doenças e danos na cultura do café no estado de São Paulo. Serão considerados a ferrugem, o bicho mineiro e o abortamento de flor. As melhores regras, com confiança e antecipação aprovadas por especialistas e pelo público alvo, serão codificadas num sistema de monitoramento agroclimatológico e permitirão a geração de alertas aos produtores paulistas de café para que tomem medidas preventivas contra esses danos e doenças.

3. JUSTIFICATIVA

Existe uma demanda cada vez mais crescente da população mundial pela oferta de produtos agrícolas mais saudáveis, cultivados com menor aplicação de produtos químicos e com a preocupação de conservação do meio ambiente. Aliado a isso destaca-se o aumento da importância das barreiras fitossanitárias no comércio internacional, devido à globalização e conseqüente diminuição nas restrições econômicas. Fica mais evidente então a necessidade de combate às doenças de culturas agrícolas e, mais ainda, com ênfase no uso racional de agrotóxicos.

Os sistemas de previsão de doenças de plantas contribuem com esse objetivo (Reis et al., 2000). A aplicação usual de agrotóxicos se baseia num calendário fixo, considerando o período de proteção conferido pelo defensivo. A alternativa oferecida pelos sistemas de previsão é a aplicação de agrotóxico apenas quando forem verificadas as condições favoráveis de ocorrência de uma determinada doença, o que faz diminuir a média de aplicações durante um ciclo da cultura. Também, no caso de controle biológico, um sistema de previsão permite a definição mais precisa do melhor momento para a liberação dos inimigos naturais que irão combater as doenças.

Além dos impactos positivos sociais (saúde da população) e ambientais (conservação dos recursos naturais) conseqüentes da diminuição no uso de produtos químicos, tem-se como conseqüência um impacto econômico para os produtores, com redução nos custos de produção e menores perdas de produtividade decorrentes de doenças. Menor custo de produção pode significar também um produto com melhor preço para os consumidores.

Ocorre que o desenvolvimento convencional dos sistemas de previsão de doenças é caro, pelos equipamentos especializados que envolve, e complexo, pois é difícil descobrir as regras que interferem no surgimento e agravamento das doenças. Novas abordagens de resolução desse problema, como é o caso desta proposta, têm grande importância e podem causar fortes impactos positivos.

4. PROBLEMA DE PESQUISA

O processamento de dados com dependências seqüenciais, que se revelam normalmente como seqüências que dependem de uma variável, que pode ser o tempo ou outra grandeza física, é um problema de pesquisa em mineração de dados que desperta grande interesse na comunidade científica. As operações tradicionais, como classificação, associação e agrupamento (Han & Kamber, 2000), tendem a tratar esses dados como coleções não ordenadas de acontecimentos, ignorando a sua informação seqüencial ou temporal.

As seqüências que dependem do tempo podem ser de dois tipos: séries temporais (*time series*), em que os elementos são valores num domínio contínuo; ou seqüências temporais, em que os elementos pertencem a

um determinado alfabeto. Análises de dados em séries temporais geralmente buscam por modelos globais para o estudo de tendências, como o estudo das mudanças climáticas e do efeito estufa. As operações de mineração em seqüências temporais, por outro lado, geralmente visam descobrir padrões seqüenciais que se repetem numa mesma seqüência ou em várias seqüências (Das et al., 1998).

O interesse deste projeto de pesquisa está nesse tipo de operação. A estratégia é transformar séries temporais agrometeorológicas em seqüências temporais, convertendo os valores contínuos de baixo nível para símbolos de um alfabeto, e depois minerar as seqüências obtidas com o intuito de descobrir padrões seqüenciais e regras de associação temporal geradas a partir deles.

A extração nessa forma de regras a partir de séries temporais envolve dois problemas relacionados. O primeiro é a escolha da representação simbólica da série temporal, que pode ser obtida com o auxílio de especialistas no domínio e por meio de operações automáticas. O segundo problema é a indução das regras de associação temporal a partir das seqüências simbólicas. Naturalmente, existe uma dependência entre a qualidade da representação obtida com a transformação da série temporal e a qualidade das regras induzidas a partir dessa representação.

Na literatura, estão descritas várias formas de tornar discreta uma série temporal, que devem ser testadas e avaliadas, como, por exemplo, diferentes técnicas de agrupamento (Das et al., 1998). Os algoritmos de mineração de padrões seqüenciais relatados na literatura foram concebidos a partir de problemas reais bem caracterizados, como o interesse por padrões de consumo ao longo do tempo em transações de compra (Agrawal & Srikant, 1995).

Esses algoritmos se aplicam bem em problemas semelhantes. Aplicações em outros domínios, como é o caso desta proposta, requerem novos algoritmos. Smyth et al. (2002) salienta, de forma geral, que "mineração de dados é uma área direcionada por aplicação, onde questões de pesquisa tendem a ser motivadas por conjuntos de dados do mundo real".

As principais diferenças, no caso da descoberta de padrões seqüenciais, residem no formato dos dados, na forma de representação dos padrões e em como se considera a freqüência desses padrões. O novo algoritmo a ser concebido pode introduzir uma solução totalmente inovadora ou pode se basear nos esquemas e nas estruturas dos já existentes.

Outro problema relacionado com a descoberta de padrões seqüenciais é o grande número de regras gerado, como conseqüência do aumento do espaço de busca, isto é, das combinações possíveis, quando se incorpora a informação temporal no processo. Isso dificulta a análise pelo usuário das regras obtidas.

Duas linhas de pesquisa para mitigar esse problema, quando se trata de regras de classificação ou de associação na forma tradicional, também podem ser adequadas à mineração de padrões seqüenciais. A primeira investe em medidas de avaliação de regras, onde o objetivo é indicar as regras mais interessantes segundo um determinado aspecto (Freitas, 1999). A segunda trata da interação humano-computador no processo de descoberta de conhecimento (Smyth et al., 2002). Por exemplo, como podem o projetista do algoritmo e o especialista representarem conhecimento prévio no domínio de tal forma que não seja redescoberto o que já se conhecia? Ou então, como pode o especialista "entrar" e "guiar" o algoritmo de mineração para descobrir padrões de interesse?

São poucos os esforços encontrados na literatura nessas duas linhas de pesquisa quando se trata de mineração de padrões seqüenciais (Zaki, 2000; Garofalakis et al., 2002; Temporal Data Mining, 8., 2002.Antunes & Oliveira,). Então, também nessas direções, este projeto de pesquisa pretende dar sua contribuição, (1) fornecendo medidas de avaliação de regras próprias para a aplicação, (2) oferecendo mecanismos de especificação de restrições para diminuir o espaço de busca dos padrões e (3) propondo formas de representação de conhecimento prévio de agrometeorologia e das condições predisponentes de doenças e danos em culturas agrícolas.

5. CONSIDERAÇÕES FINAIS

O projeto de pesquisa apresentado neste trabalho está inserido num projeto amplo de monitoramento agroclimatológico para o estado de São Paulo. A cultura do café foi escolhida pela importância que possui nos setores agrícolas paulista e nacional, e as doenças e danos (ferrugem, bicho mineiro e abortamento de flor) pelas perdas e prejuízos que ocasionam. É objetivo do projeto maior que o trabalho seja estendido para outras culturas importantes para o estado e para o Brasil.

A relevância do estudo proposto pode ser vista também pelo grande potencial de aplicação da mineração de padrões seqüenciais em outras áreas do conhecimento. Dentre elas, cita-se a bioinformática e a agricultura de precisão, áreas também de destaque atual e potencial futuro para a pesquisa agropecuária.

6. REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R.; SRIKANT, R. **Mining sequential patterns**. In: Proceedings of the International Conference on Data Engineering, 11., 1995. pp 3-14.

ANTUNES, C. M.; OLIVEIRA, A. L. **Using context-free grammars to constrain apriori-based algorithms for mining temporal association rules**. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD-02) - Workshop on Temporal Data Mining, 8., 2002.

DAS G.; LIN K., MANNILA, H.; RENGANATHAN, G.; SMYTH, P. **Rule discovery from time series**. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD-98), 4., 1998. pp 16-22.

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

GAROFALAKIS M.; RASTOGI R., SHIM K. **Mining sequential patterns with regular expression constraints**. IEEE Transactions on Knowledge and Data Engineering, v. 14, n. 3, 2002. pp 530-552.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques**. Morgan Kaufmann Publishers, San Francisco, 2000.

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.

SMYTH, P.; PREGIBON, D.; FALOUTSOS, C. **Data-driven evolution of data mining algorithms**, Communications of the ACM, v. 45, n. 8, 2002. pp 33-37.

ZAKI, M. J. **Sequence mining in categorical domains: incorporating constraints**. In: International Conference on Information and Knowledge Management, 9., 2000. pp 422-429.