Documentos

ISSN 1677-9274
Dezembro, 2012
123

Processo especializado de descoberta de conhecimento em bases de dados para a modelagem de doenças de plantas - versão 1.0

Embrapa

ISSN 1677-9274
Dezembro, 2012

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 123

Processo especializado de descoberta de conhecimento em bases de dados para a modelagem de doenças de plantas - versão 1.0

Carlos Alberto Alves Meira

Embrapa Informática Agropecuária
Campinas, SP
2012

Embrapa Informática Agropecuária
Av. André Tosello, 209 - Barão Geraldo
Caixa Postal 6041 - 13083-886 - Campinas, SP
Fone: (19) 3211-5700 - Fax: (19) 3211-5754
www.cnptia.embrapa.br
sac@cnptia.embrapa.br

1ª edição on-line 2012

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Informática Agropecuária

Meira, Carlos Alberto Alves.

Processo especializado de descoberta de conhecimento em Bases de Dados para modelagem de doenças de plantas - versão 1.0 / Carlos Alberto Alves Meira. - Campinas : Embrapa Informática Agropecuária, 2012.

45 p. : il. - (Documentos / Embrapa Informática Agropecuária , ISSN 1677-9274 ; 123).

1. Mineração de dados. 2. CRISP-DM. 3. Epidemiologia. 4. Alerta doença de planta. I. Embrapa Informática Agropecuária. II. Título. III. Série.

CDD (21. ed.) 006.3

Autor

Carlos Alberto Alves Meira

Doutor em Engenharia Agrícola
Pesquisador da Embrapa Informática Agropecuária
Av. André Tosello, 209, Barão Geraldo
Caixa Postal 6041 - 13083-970 - Campinas, SP
Telefone: (19) 3211-5748
e-mail: carlos.meira@embrapa.br

Apresentação

A metodologia apresentada neste documento descreve o processo de descoberta de conhecimento em bases de dados, ou mineração de dados, com algumas tarefas especializadas para a modelagem de doenças de plantas.

O objetivo em produzir este processo especializado foi organizar e registrar a experiência adquirida em um projeto real para orientar e facilitar sua adoção em iniciativas de modelagem de doenças de plantas em contextos de mineração de dados semelhantes ao do projeto original.

Para embasar a especialização conduzida, são apresentados, inicialmente, de forma geral, conceitos sobre descoberta de conhecimento em bases de dados, a metodologia padrão utilizada e o seu modelo de referência do processo de mineração de dados.

Espera-se que esta versão inicial da metodologia seja atualizada e incrementada com o passar do tempo, em virtude de novas experiências coletadas em projetos futuros de descoberta de conhecimento em bases de dados, o que justifica a identificação da versão 1.0 no título deste documento.

Kleber Xavier Sampaio de Souza

Chefe-Geral
Embrapa Informática Agropecuária

Sumário

Processo especializado de descoberta de conhecimento em bases de dados para a modelagem de doenças de plantas - versão 1.0

Carlos Alberto Alves Meira

1 Introdução

Este documento apresenta um processo especializado de descoberta de conhecimento em bases de dados, ou mineração de dados, com tarefas específicas para aplicação na modelagem de doenças de plantas. Ele é resultante da Tese de Doutorado intitulada “Processo de descoberta de conhecimento em bases de dados para a análise e o alerta de doenças de culturas agrícolas e sua aplicação na ferrugem do cafeeiro” (MEIRA, 2008).

O objetivo desta publicação foi produzir um documento autocontido e completo, no formato de uma metodologia, para referência e uso futuro. Parte do seu conteúdo pode ser encontrada no documento original citado, mas a organização e a redação deste documento procuram facilitar a consulta, a leitura e a compreensão pelo leitor.

Além disso, a visão geral da metodologia base utilizada para a proposição do processo especializado, conhecida pela sigla CRISP-DM (CHAPMAN et al., 2000), do inglês *CRoss-Industry Standard Process for Data Mining*, foi acrescida de uma descrição um pouco mais elaborada das fases e tarefas genéricas do seu modelo do processo de mineração de dados.

O documento está assim organizado: a próxima seção apresenta conceitos básicos de descoberta de conhecimento em bases de dados e mineração de dados; a seção 3 é a visão geral da metodologia CRISP-DM, incluindo o seu modelo de referência para o processo de mineração de dados; a seção 4 apresenta o processo especializado de descoberta de conhecimento em bases de dados para a modelagem de doenças de plantas; e na seção 5 são feitas as considerações finais.

2 Descoberta de conhecimento em bases de dados

Nas últimas décadas, a capacidade de gerar e armazenar dados aumentou rapidamente. Esse crescimento explosivo na quantidade de dados armazenados gerou a necessidade por novas técnicas e ferramentas automatizadas que pudessem auxiliar na transformação dos dados em informação útil e conhecimento. A abundância de dados, em conjunto com a necessidade por ferramentas de análise, ficou conhecida como a situação “rica em dados, mas pobre em informação” (HAN et al., 2011).

Ao mesmo tempo em que se percebia uma desproporção entre a geração de dados e a sua compreensão, havia uma expectativa crescente de que os dados, analisados e apresentados de maneira inteligente, seriam um recurso valioso a ser usado como vantagem competitiva (FRAWLEY et al., 1992).

Nesse contexto, surgiu a Descoberta de Conhecimento em Bases de Dados ou KDD, do termo em inglês *Knowledge Discovery in Databases* (KDD). KDD foi definida inicialmente como a extração de informação implícita, previamente desconhecida e potencialmente útil a partir de dados (FRAWLEY et al., 1992). Essa definição foi posteriormente revisada:

> “Descoberta de conhecimento em bases de dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em repositórios de dados.” (FAYYAD et al., 1996a).

O termo “mineração de dados” é bastante comum ser encontrado como sinônimo de KDD, como é o caso neste documento. Alternativamente, de

acordo com Fayyad et al. (1996b), KDD refere-se ao processo global de descoberta de conhecimento a partir de dados, enquanto a mineração de dados é uma fase desse processo. Por essa visão, a mineração refere-se à aplicação de algoritmos específicos para extrair os padrões dos dados. Também, a mineração de dados está muitas vezes associada à noção de extração de conhecimento a partir de grandes volumes de dados. Mas, independente disso, o processo de KDD pode ser realizado, em todas as suas fases, com qualquer quantidade de dados suficiente para que os padrões sejam extraídos.

A Figura 1 apresenta uma visão geral das fases do processo de KDD (FAYYAD et al., 1996a): primeiro, antes de se começar a mexer com os dados, é preciso compreender o domínio de aplicação e identificar a meta da descoberta de conhecimento pelo ponto de vista do usuário; em seguida, os dados de interesse são selecionados, é feito um pré-processamento nos dados (p.ex. eliminação de ruídos e tratamento de dados ausentes), os dados sofrem transformações (p. ex. conversão de dados e derivação de novos atributos) e é realizada a mineração de dados; ao final, é feita a interpretação e a avaliação dos resultados obtidos, e o conhecimento descoberto é distribuído conforme se tenha definido no planejamento. Esse processo pode envolver várias iterações e quase sempre é necessário o retorno para fases anteriores.

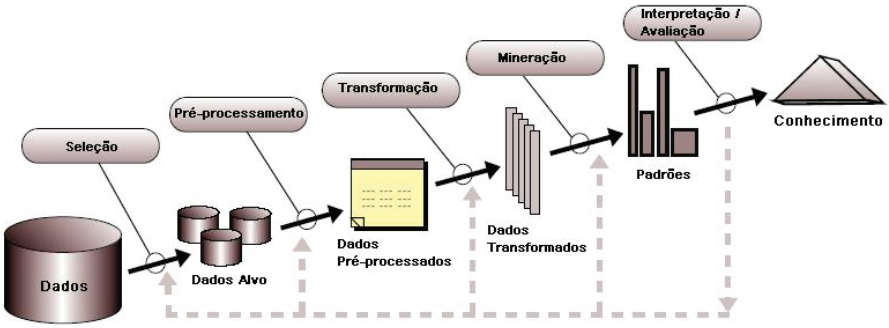

Figura 1. Visão geral das fases do processo de KDD.
Fonte: Fayyad et al. (1996a).

A mineração de dados é centrada na cooperação entre os seus diversos atores, e o seu sucesso depende, em parte, dessa cooperação. Os atores do processo podem ser divididos em três classes (REZENDE et al., 2002):

1) Especialista do domínio: pessoa que deve possuir amplo conhecimento do domínio de aplicação e deve fornecer apoio para a execução do processo.
2) Analista de dados: pessoa responsável pela execução do processo de KDD. Este usuário deve conhecer a fundo as etapas que compõem o processo.
3) Usuário final: representa a classe de usuários que vai utilizar o conhecimento extraído, por exemplo, em auxílio para a tomada de decisão.

A área de pesquisa evoluiu, e continua a evoluir, da intersecção de áreas como aprendizado de máquina, reconhecimento de padrões, banco de dados, estatística e visualização de dados (FAYYAD et al., 1996b). KDD se baseia fortemente em técnicas conhecidas de aprendizado de máquina, de reconhecimento de padrões e de estatística para encontrar os padrões nos dados. A estatística oferece também métodos de quantificação da incerteza inerente quando se procura inferir padrões gerais a partir de amostras de uma população. As técnicas de visualização de dados estimulam naturalmente a percepção e a inteligência humana, aumentando a capacidade de entendimento e de associação de novos padrões (REZENDE et al., 2002).

3 A metodologia CRISP-DM

3.1 Histórico e motivação

Na época em que se despertou o crescente interesse pela mineração de dados, por volta de 1996, não havia uma abordagem metodológica amplamente aceita e adotada. Surgia, assim, a clara necessidade por um processo que padronizasse e auxiliasse as organizações e suas equipes a planejarem e executarem os seus projetos de mineração de dados, de forma a incentivar boas práticas e oferecer a estrutura necessária para o alcance de melhores resultados com maior rapidez (SHEARER, 2000).

Por outro lado, havia a motivação comercial por um processo com tais características, como forma de demonstrar aos potenciais clientes que a

mineração de dados estaria suficientemente madura para ser adotada em seus negócios (CHAPMAN et al., 2000).

Nesse contexto, foi concebida a metodologia CRISP-DM, com o propósito de oferecer um modelo de processo bem documentado, sem direitos de propriedade e de livre acesso. O processo deveria também ser considerado neutro, isto é, estar organizado independente, ou sem influência, de empresa, de ferramenta de software e de domínio de aplicação.

A metodologia CRISP-DM não foi idealizada na academia, de maneira teórica e fundamentada em conceitos, nem por comitês elitizados a portas fechadas. Ao contrário, a sua criação foi coordenada por um consórcio de empresas líderes da indústria, formado em 1997. O modelo de processo foi desenvolvido e refinado, durante cerca de três anos, por um grupo especial *Special Interest Group* (SIG) com mais de 200 membros, entre praticantes e fornecedores de ferramentas e serviços. O seu sucesso está intimamente relacionado e fortemente baseado em experiências práticas e reais de como conduzir projetos de mineração de dados.

A primeira versão da metodologia foi lançada como CRISP-DM 1.0 em 2000 (CHAPMAN et al., 2000). Em 2006, o consórcio anunciou que iria iniciar os trabalhos para a segunda versão. Em setembro daquele ano, o grupo SIG reuniu-se para discutir possíveis melhorias a ser incluídas na versão 2.0. Entretanto, os esforços não foram adiante. De qualquer forma, pesquisas conduzidas em 2002, 2004 e 2007 mostram que CRISP-DM manteve-se como a metodologia líder utilizada por pessoas que conduziram projetos de mineração de dados (KDNUGGETS, 2012a; 2012b; 2012c).

3.2 Organização hierárquica

A metodologia CRISP-DM é descrita em termos de um modelo de processo hierárquico com quatro níveis de abstração, do geral para o específico (CHAPMAN et al., 2000): fases, tarefas genéricas, tarefas especializadas e instâncias do processo (Figura 2).

No nível mais alto, o processo de mineração de dados é organizado em fases. Cada fase consiste de diversas tarefas genéricas no segundo nível, o qual se destina a ser suficientemente geral para cobrir todas as possíveis situações de mineração de dados. As tarefas genéricas procuram ser as

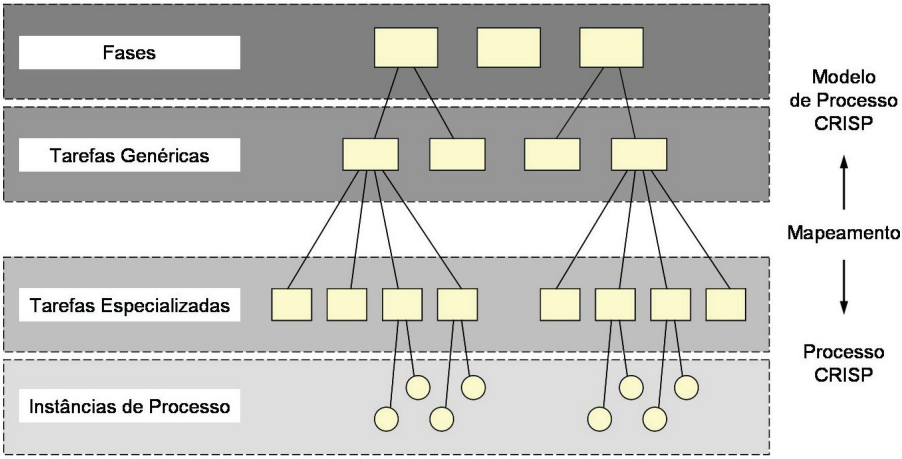

Figura 2. Visão hierárquica da metodologia CRISP-DM.

mais completas e estáveis, o que significa, respectivamente, que abrangem o processo completo de mineração de dados e todas as prováveis aplicações, e que o processo deve permanecer válido mesmo para desenvolvimentos futuros, como novas técnicas de modelagem.

O terceiro nível, das tarefas especializadas, descreve como as ações, dentro das tarefas genéricas, devem ser conduzidas em certas situações específicas. Por exemplo, no segundo nível, existe a tarefa de limpeza os dados. O terceiro nível descreve como esta tarefa difere em situações distintas, tais como limpeza de valores numéricos ou limpeza de valores categóricos, ou caso o tipo de problema seja agrupamento ou modelagem preditiva.

O quarto e último nível, das instâncias do processo, é um registro das ações, decisões e resultados de uma iniciativa real de mineração de dados. Uma instância do processo é organizada de acordo com a definição das tarefas nos níveis superiores, mas representa o que realmente aconteceu em um projeto posto em prática, ao invés do que acontece em geral.

Além da decomposição hierárquica vertical, a metodologia CRISP-DM distingue na horizontal entre o modelo de referência e o guia do usuário. O modelo de referência apresenta uma visão geral das fases e tarefas do processo (ver próxima seção), descrevendo “o que fazer” em um projeto de mineração de dados. O guia do usuário descreve “o como fazer”, forne-

cendo dicas mais detalhadas para cada fase e para cada tarefa dentro das fases (CHAPMAN et al., 2000).

3.3 O modelo de referência

O modelo de referência para mineração de dados fornece uma visão geral do processo, das suas fases e respectivas tarefas, e dos relacionamentos entre as tarefas. O ciclo de vida de um projeto de mineração de dados, representado na Figura 3, é composto de seis fases (CHAPMAN et al., 2000): compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição dos resultados obtidos.

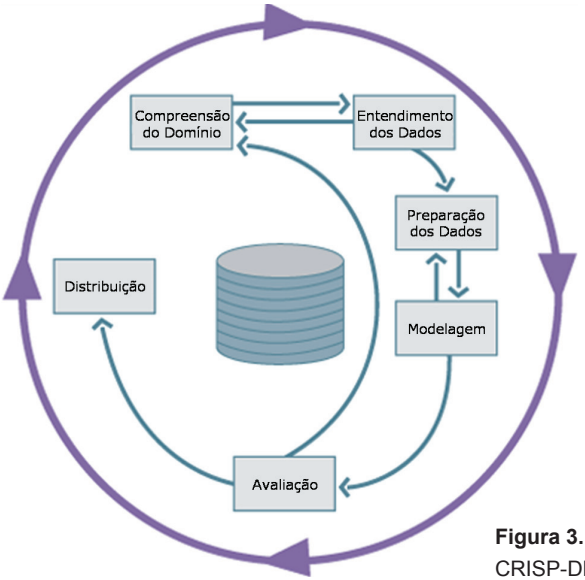

Figura 3. Fases do modelo de processo CRISP-DM.

A sequência dessas fases não é rígida, sendo comum, e quase sempre necessário, voltar e avançar entre elas – as setas internas na Figura 3 indicam as dependências entre fases mais importantes e mais frequentes. Qual fase, ou tarefa específica de uma fase, deve ser executada na sequência depende do resultado da fase anterior.

O círculo externo da Figura 3 representa o aspecto cíclico de um projeto típico de mineração de dados. Uma vez encontrada uma solução para o

problema, o projeto não é necessariamente finalizado. As lições aprendidas durante o processo, e a partir da solução encontrada, podem desencadear novos questionamentos, geralmente mais focados. Processos de mineração de dados subsequentes beneficiam-se das experiências adquiridas nas instâncias anteriores.

3.4 Mapeando os modelos genéricos para modelos especializados

O mapeamento entre os níveis genérico e especializado da metodologia CRISP-DM é direcionado por um contexto de mineração de dados. Quatro diferentes dimensões podem estar discriminadas em um contexto de mineração de dados (CHAPMAN et al., 2000):

1) Domínio de aplicação: é a área específica em que o projeto de mineração de dados se enquadra.
2) Tipo de problema de mineração de dados: indica a(s) classe(s) de objetivos específicos que o projeto de mineração de dados trata. Por exemplo: classificação, regressão, associação, etc.
3) Aspecto técnico: cobre questões específicas relativas a diferentes desafios técnicos que geralmente ocorrem durante a mineração de dados. Por exemplo: valores ausentes, *outliers*, etc.
4) Ferramenta e técnica: especifica as ferramentas e/ou técnicas aplicadas no projeto de mineração de dados. Exemplos de técnicas: indução de árvores de decisão, redes neurais, regressão, algoritmos genéticos e outras; com relação às ferramentas, existem opções de software livre e proprietárias (ver seção 4.4).

Um contexto de mineração de dados específico são valores concretos para uma ou mais dessas dimensões. Por exemplo, um projeto tratando de um problema de classificação em predição de migração de clientes constitui um contexto de mineração de dados específico. Quanto mais valores são determinados para as diferentes dimensões, mais concreto é o contexto de mineração de dados.

Destacam-se dois tipos diferentes de mapeamento entre o nível genérico e o nível especializado:

1) Mapeamento para o presente: quando se aplica o modelo de processo genérico para executar um projeto de mineração de dados e se procura mapear as tarefas genéricas e suas descrições para este projeto específico, conforme necessário. Neste caso, trata-se de um mapeamento único para, provavelmente, apenas um uso.
2) Mapeamento para o futuro: quando se especializa sistematicamente o modelo de processo genérico segundo um contexto de mineração de dados pré-definido ou, de maneira similar, quando se analisa e se consolida sistematicamente experiências de um projeto de mineração de dados em particular. Neste caso, trata-se de elaborar explicitamente um modelo de processo especializado em termos da metodologia CRISP-DM.

A escolha do tipo de mapeamento apropriado depende do contexto de mineração de dados específico e das necessidades da organização. A estratégia básica para mapear o modelo de processo genérico para o nível especializado, no entanto, é idêntica para ambos os tipos de mapeamentos:

1) Analisar o contexto específico.
2) Eliminar detalhes que não se aplicam ao contexto.
3) Adicionar detalhes específicos do contexto.
4) Especializar conteúdos genéricos de acordo com as características concretas do contexto.
5) Renomear conteúdos genéricos, possivelmente, provendo significados mais explícitos do contexto, por uma questão de clareza.

4 Processo especializado de descoberta de conhecimento em bases de dados para a modelagem de doenças de plantas

A metodologia CRISP-DM foi utilizada em um projeto de pesquisa de mineração de dados no domínio agropecuário, mais especificamente na área de epidemiologia de doenças de plantas. Uma instância do processo de descoberta de conhecimento em bases de dados - KDD foi realizada para

avaliar a aplicação de classificação e de árvores de decisão na análise e no alerta da ferrugem do cafeeiro (MEIRA, 2008; MEIRA et al., 2008; MEIRA et al., 2009).

A ferrugem, causada pelo fungo *Hemileia vastatrix*, é a principal doença do cafeeiro em todo o mundo, ocasionando decréscimos significativos na produção de café (ZAMBOLIM et al., 1997). O conhecimento dos fatores que determinam as epidemias de ferrugem, portanto, é fundamental. O desenvolvimento de modelos de alerta para esta doença se justifica não apenas pelos seus prejuízos econômicos, mas pela variação na sua intensidade entre os anos agrícolas e pela disponibilidade de medidas de controle economicamente viáveis.

A técnica de indução de árvores de decisão foi escolhida por permitir se descobrir a estrutura preditiva de um problema e/ou produzir modelos precisos (APTE; WEISS, 1997). Além disso, árvores de decisão são de interesse particular para a descoberta de conhecimento em bases de dados devido à sua representação simbólica e interpretável (FAYYAD et al., 1996a).

Foram utilizadas duas ferramentas ou software de modelagem: o SAS® Enterprise Miner™ (versão 4.3, SAS Institute) e o Weka (versão 3.4.11, Universidade de Waikato, Nova Zelândia). O Enterprise Miner™ é a solução SAS para o processo de mineração de dados (SAS INSTITUTE, 2004). O Weka é um software livre com uma coleção de algoritmos de aprendizado de máquina e de ferramentas relacionadas, que também oferece suporte ao processo completo de mineração de dados (HALL et al., 2009).

Além do mapeamento entre os níveis genérico e especializado do modelo do processo CRISP-DM, para atender aos objetivos específicos do projeto, foi feito um esforço para caracterizar o processo de descoberta de conhecimento em bases de dados realizado, com o intuito de permitir a sua reprodução e adaptação em problemas similares de outras culturas agrícolas ou mesmo da cultura do café, para outras doenças e pragas.

A caracterização ou a especialização do processo, segundo o termo usado na metodologia CRISP-DM, foi feita de acordo com o mapeamento para o futuro entre os níveis genérico e especializado do modelo do processo, com base na instância do processo para a obtenção dos modelos de análise e de alerta da ferrugem do cafeeiro, ou seja, baseada em uma análise sistemática das características do projeto e do registro das ações,

decisões, resultados e dificuldades durante a sua execução. O contexto de mineração de dados usado na especialização do modelo do processo é apresentado na Tabela 1. Este contexto conduziu o mapeamento entre os níveis genérico e especializado da metodologia CRISP-DM.

Tabela 1. Contexto de mineração de dados usado na especialização do modelo do processo.

Domínio de aplicação	Tipo de problema	Aspecto técnico	Técnica e ferramenta
Modelagem de doenças de plantas	Classificação	Definição de classes Derivação de atributos Indução interativa	Árvore de decisão SAS® Enterprise Miner™ Weka (J48)

A seguir, são descritas as fases e tarefas dos níveis genéricos do modelo de processo de mineração de dados da metodologia CRISP-DM (CHAPMAN et al., 2000; SHEARER, 2000). Dentro de algumas das tarefas genéricas, são apresentadas tarefas especializadas do processo de KDD para a modelagem de doenças de plantas, de acordo com o contexto de mineração de dados definido. São apresentados também relatos da experiência na instância do processo para a obtenção dos modelos de análise e de alerta da ferrugem do cafeeiro (MEIRA, 2008; MEIRA et al., 2008; MEIRA et al., 2009).

4.1 Compreensão do domínio

A fase inicial, talvez a mais importante de qualquer iniciativa de mineração de dados, foca em compreender os objetivos do projeto pela perspectiva do domínio, convertendo este entendimento em uma definição do problema de mineração de dados e, em seguida, desenvolvendo um plano preliminar para atingir os objetivos definidos. É crucial que os profissionais de mineração de dados compreendam plenamente o problema para o qual estão buscando uma solução, a fim de compreender quais e como os dados devem posteriormente ser analisados.

A fase de compreensão do domínio envolve várias tarefas, incluindo determinar os objetivos, avaliar a situação, determinar as metas de mineração de dados e elaborar o plano do projeto.

Determinar os objetivos

Compreender corretamente os objetivos do projeto é fundamental para se determinar os fatores importantes envolvidos e para garantir que a iniciativa não acabe encontrando as respostas certas para as perguntas erradas. Para isso, o analista de dados deve identificar o objetivo principal do projeto, bem como as questões importantes que devem ser abordadas.

O analista de dados deve, também, estabelecer a medida de sucesso do projeto. Sucesso pode ser medido, por exemplo, por redução de perdas, por incremento de ganhos ou, simplesmente, por um melhor entendimento dos dados. Deve-se evitar estabelecer metas inalcançáveis e garantir que cada critério de sucesso diz respeito a pelo menos um dos objetivos específicos.

Tarefa especializada: Verificar pré-requisitos da doença

Desenvolver um modelo de alerta requer que a doença atenda quatro requisitos (COAKLEY, 1988): (1) a doença ocasiona perdas economicamente significativas na qualidade ou na quantidade da produção; (2) a doença varia entre cada estação de cultivo (p.ex. severidade na colheita e taxa de aumento); (3) medidas de controle da doença estão disponíveis e são economicamente viáveis; e (4) informação sobre a natureza da dependência da doença em relação às condições meteorológicas é suficientemente conhecida.

Relato da experiência na instância do processo:

A ferrugem do cafeeiro causa decréscimos significativos na produção e é considerada a principal doença da cultura. Além da importância econômica, os demais requisitos da doença também foram atendidos: ela varia em intensidade a cada ano agrícola; existem várias opções de medidas de controle economicamente viáveis; e diversos estudos foram encontrados na literatura sobre a influência das condições meteorológicas no seu desenvolvimento.

Tarefa especializada: Identificar o esquema de análise de interesse

Esquemas de análise são adotados de acordo com o interesse em uma ou mais características de uma epidemia (BUTT; ROYLE, 1990): o interesse pode estar focado no progresso da epidemia no decorrer do tempo; pode estar concentrado na taxa de aumento da doença; pode estar relacionado

com os fatores que determinam o nível de severidade da doença em um determinado momento (p.ex. na época da colheita); pode estar centrado em eventos do ciclo da doença; e pode estar vinculado com perda de produção, que é uma consequência e não parte de uma epidemia.

Relato da experiência na instância do processo:

No caso da ferrugem do cafeeiro, como a doença é policíclica, o interesse estava em relacionar mudanças nas condições meteorológicas com aumentos na intensidade da doença.

O interesse foi pela taxa de aumento da doença, calculada como a taxa de infecção, subtraindo-se a incidência no mês em questão com a incidência no mês anterior.

Chegou-se a cogitar o uso de uma taxa que indicasse, ao invés do aumento absoluto, o aumento percentual no nível de incidência da ferrugem. No entanto, percebeu-se que haveria dois tipos de problema: aumentos a partir do nível zero de incidência (não há como fazer o cálculo do aumento percentual); e aumentos significativos a partir de níveis baixos de incidência (p.ex. um aumento na incidência de 0,5% para 13% corresponderia a um aumento percentual muito alto de 2500%; nesse caso, a taxa de infecção seria de 12,5 p.p.). Esses problemas dificultariam a definição das classes de taxa de infecção.

Avaliar a situação

Nesta tarefa, o analista de dados descreve os recursos, desde pessoal até software, que estão disponíveis para a execução do projeto de mineração de dados. Particularmente, é importante identificar quais dados estão disponíveis para atender ao objetivo principal. O analista de dados deve também: relacionar as premissas feitas no projeto; identificar os riscos do projeto e listar possíveis soluções para esses riscos; criar um glossário de termos, do problema e de mineração de dados; e fazer uma análise de custo-benefício para o projeto.

Tarefa especializada: Identificar as fontes de dados e garantir acesso a elas

Mencionar a importância disso parece óbvio, mas é essencial que se identifique todas as fontes de dados e se garanta o acesso a elas desde o início da concepção do projeto. É enganoso acreditar que o importante

está na proposta do projeto e que garantia prévia de acesso a uma parte dos dados é suficiente. Sem os dados completos, depois, principalmente relativos à doença, não há o que fazer.

Relato da experiência na instância do processo:

À época da concepção do projeto, considerou-se que já se tinha uma grande quantidade de dados meteorológicos e "só faltavam" os dados da ferrugem do cafeeiro. Havia a percepção de que aqueles dados eram importantes, mas preocupou-se, naquele momento, apenas com a proposta do projeto.

Com o passar do tempo, percebeu-se que o mais difícil estava em conseguir fontes de dados sobre doenças e pragas do cafeeiro. No estado de São Paulo, para onde se planejava originalmente desenvolver os modelos, não foram encontradas tais fonte de dados.

A Fundação Procafé foi o único lugar em que se encontrou um monitoramento sistemático e duradouro de doenças e pragas da cultura do café, dentre elas a ferrugem do cafeeiro (JAPIASSÚ et al., 2007).

Tarefa especializada: Conferir a disponibilidade de especialistas do domínio

A participação no projeto de especialistas do domínio é fundamental. Fitopatologistas, atuantes na epidemiologia de doenças de plantas e, de preferência, que tenham experiência e interesse na doença, são os mais indicados. A participação de agrometeorologistas, com experiência no desenvolvimento de modelos de previsão de doenças de plantas, também é bem-vinda.

Contudo, a participação de um especialista não desobriga o analista de dados de se inteirar sobre a epidemiologia da doença em questão. O nível de conhecimento do analista deve estar estreitamente relacionado com o nível de participação do especialista no projeto. De qualquer forma, quanto mais o analista de dados conhecer do domínio de aplicação, melhor.

Relato da experiência na instância do processo:

O especialista do domínio, durante o projeto, foi o Dr. Sérgio Almeida de Moraes, fitopatologista e pesquisador do Instituto Agronômico de Campinas (IAC). O Dr. Moraes desenvolveu estudos importantes sobre a ferrugem do cafeeiro (MORAES, 1983; MORAES et al., 1976).

A sua participação foi como um consultor. Diversas reuniões foram realizadas, com a sua presença, desde a elaboração do plano de pesquisa até a fase final de execução do projeto.

Determinar as metas de mineração de dados

Uma meta de mineração de dados especifica ou quantifica os objetivos do projeto em termos da aplicação - por exemplo, “Predizer quantos itens um cliente poderá comprar a partir de suas compras nos últimos três anos, de informações demográficas (idade, salário, cidade, etc.) e do preço do item”. As metas de mineração de dados exprimem os objetivos do projeto em termos técnicos, descrevendo os resultados esperados do projeto que vão permitir o alcance dos objetivos.

É preciso também definir critérios de sucesso da mineração de dados - por exemplo, êxito poderia ser definido como obter um determinado nível de acurácia de previsão. Como nos critérios de sucesso para o domínio de aplicação, pode ser necessário descrevê-los em termos subjetivos, no caso em que a pessoa, ou grupo de pessoas, que fará o julgamento deverá ser identificada.

Se a meta do projeto não puder ser efetivamente traduzida para uma meta de mineração de dados, é recomendável considerar a redefinição do problema neste ponto. Além de transformar as questões do domínio em metas técnicas, outra atividade consiste em especificar o tipo de problema de mineração de dados.

Relato da experiência na instância do processo:

A hipótese do trabalho realizado na instância do processo foi que uma análise de dados meteorológicos, junto com registros de intensidade de doenças de culturas agrícolas causadas por fungos, caracterizada como um processo de descoberta de conhecimento em bases de dados, indicaria a viabilidade de uso dos modelos obtidos, em termos de acurácia de predição e de outras medidas cabíveis, na emissão de alertas dessas doenças.

O objetivo específico foi produzir modelos confiáveis de alerta da ferrugem do cafeeiro, a partir de dados meteorológicos, da carga pendente de frutos do cafeeiro e do espaçamento entre plantas, por meio de classificação e de indução de árvores de decisão.

Não foi estabelecido como meta um nível de acurácia desejado para os modelos. A pesquisa era inédita e o que se podia esperar era alcançar níveis de acurácia parecidos com os de outras aplicações de árvores de decisão como modelos de alerta de doenças de plantas, cujos maiores valores estimados não ultrapassaram 80% (MEIRA, 2008).

O propósito básico da indução de uma árvore de decisão é produzir um modelo de predição preciso ou descobrir a estrutura preditiva do problema. No último caso, procura-se compreender quais variáveis e interações dessas variáveis conduzem o fenômeno estudado. Esses dois objetivos podem aparecer juntos em um mesmo estudo (BREIMAN et al., 1984).

Sendo assim, outro objetivo específico foi aplicar e avaliar o potencial da indução de árvores de decisão na análise da epidemia da ferrugem do cafeeiro. A meta era obter uma árvore de decisão capaz de auxiliar na compreensão de como as condições do ambiente, a carga pendente de frutos do cafeeiro e o espaçamento entre as plantas na lavoura condicionaram a taxa de infecção da doença, identificando os fatores mais importantes no progresso da ferrugem do cafeeiro no campo.

Elaborar o plano do projeto

O plano do projeto descreve o planejamento para se alcançar as metas de mineração de dados, incluindo a estratégia de ação, com o esboço de passos específicos e de um cronograma proposto, uma avaliação dos riscos potenciais e uma avaliação inicial das ferramentas e técnicas necessárias para apoiar o projeto. Padrões de tempo e esforço normalmente aceitos são (CHAPMAN et al., 2000): 50 a 70% do tempo e esforço em um projeto de mineração de dados envolvem a fase de preparação dos dados; 20 a 30% envolve a fase de entendimento dos dados; 10 a 20% é dispendido em cada uma das fases de modelagem, avaliação e compreensão do domínio; e 5 a 10% é gasto no planejamento da fase de distribuição.

4.2 Entendimento dos dados

A fase de entendimento dos dados começa com uma coleção de dados inicial. O analista, em seguida, inicia atividades para adquirir familiaridade com os dados, identificar problemas de qualidade nos dados, obter os

primeiros *insights* a partir dos dados e detectar subconjuntos interessantes para formular hipóteses acerca das informações escondidas. A fase de entendimento dos dados envolve quatro passos, incluindo a coleção de dados iniciais, a descrição dos dados, a exploração dos dados e a verificação da qualidade dos dados.

Coletar os dados iniciais

O analista de dados deve adquirir os dados necessários, incluindo o carregamento e a integração desses dados, se necessário. Deve certificar-se de relatar os problemas enfrentados e as soluções encontradas, para auxiliar em possíveis replicações do projeto. Por exemplo, dados podem ser coletados de várias fontes diferentes e algumas dessas fontes podem ter um tempo de espera longo. Saber disso antecipadamente é útil, para se evitar possíveis atrasos.

Descrever os dados

Nesta tarefa, o analista de dados examina as características dos dados adquiridos e documenta os resultados, examinando questões como o formato dos dados, a quantidade de dados, o número de registros e atributos em cada tabela, as identificações dos atributos e qualquer outra característica superficial dos dados. A questão chave a ser respondida: os dados obtidos satisfazem os requisitos relevantes? Por exemplo, se idade é um atributo importante, mas os dados não representam o intervalo possível de idades, pode ser recomendável coletar um conjunto de dados diferente. Esta tarefa também fornece uma compreensão básica dos dados que serão utilizados nas tarefas subsequentes.

Explorar os dados

Esta tarefa aborda as questões de mineração de dados que podem ser acomodadas por meio de consultas, visualização e relatórios. Considerando casos hipotéticos, um analista pode consultar os dados para descobrir os tipos de produtos que consumidores de determinado nível econômico normalmente compram; ou um analista pode executar uma análise de visualização para descobrir possíveis padrões de fraude. O ana-

lista de dados deve criar um relatório de exploração dos dados, descrevendo as primeiras descobertas, ou uma hipótese inicial, e o potencial impacto sobre o restante do projeto.

Tarefa especializada: Explorar os dados da doença

Recomenda-se a elaboração de gráficos de evolução da doença, como os da Figura 4, para se conhecer e conferir a sua periodicidade estacional e para se identificar possíveis erros no processo de avaliação da intensidade da doença (p.ex. erros na amostragem).

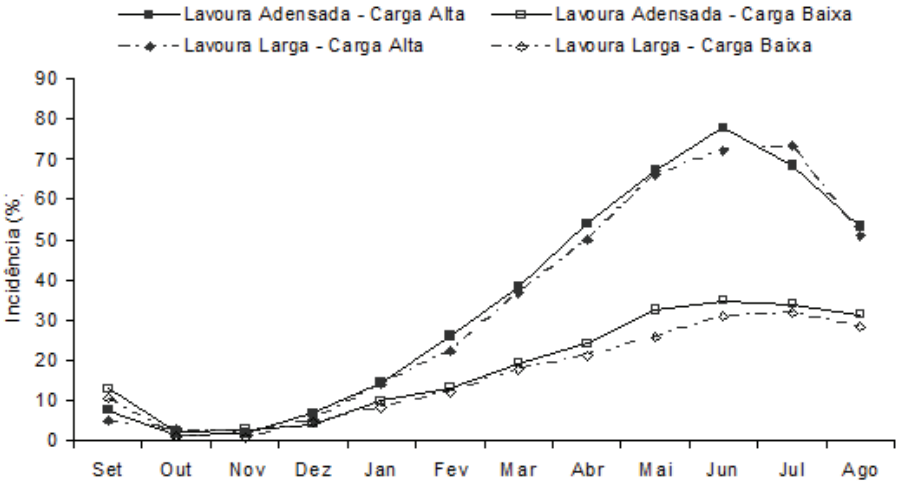

Figura 4. Evolução média mensal da incidência da ferrugem do cafeeiro em lavouras com diferentes espaçamentos e cargas pendentes de frutos.

Relato da experiência na instância do processo:

Com esse tipo de gráfico foi possível identificar claramente um problema na amostragem que determinou um dos valores de incidência da ferrugem do cafeeiro (Figura 5).

Tarefa especializada: Explorar os dados meteorológicos

A exploração visual dos dados meteorológicos também é importante. Diagramas de dispersão (Figura 6), gráficos do tipo *box plot* (Figura 7) e histogramas (Figura 8) permitem um melhor entendimento dos dados e podem fornecer subsídios para as fases posteriores do processo.

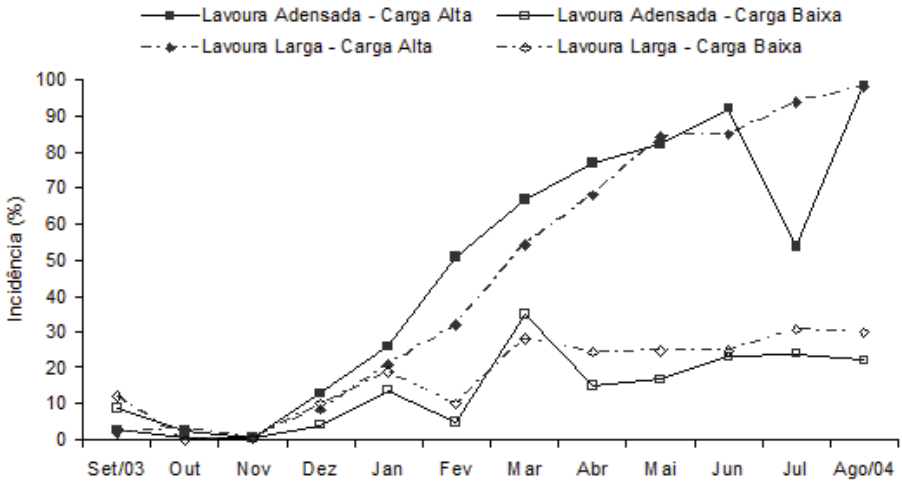

Figura 5. Evolução mensal da incidência da ferrugem do cafeeiro no ano agrícola 2003/2004 em lavouras com diferentes espaçamentos e cargas pendentes.

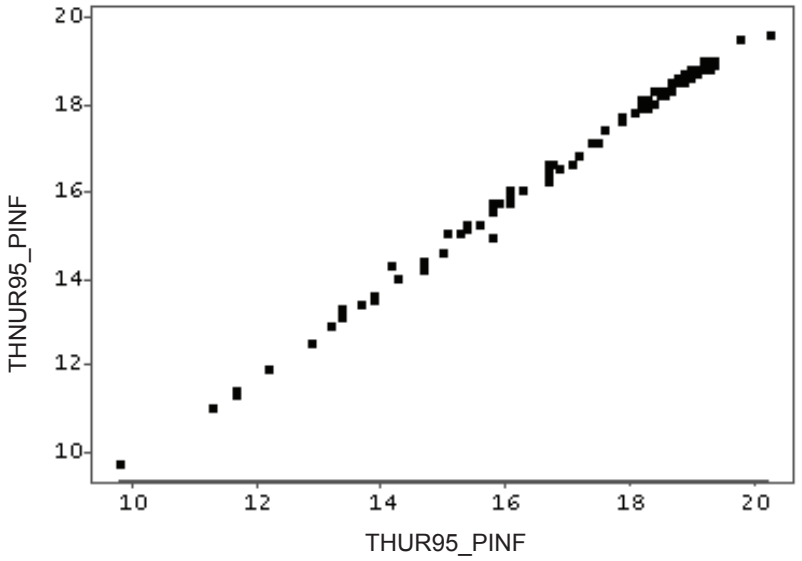

Figura 6. Relação entre a temperatura média durante o período de molhamento foliar (THUR95_PINF) e a temperatura média durante o período noturno de molhamento foliar (THNUR95_PINF).

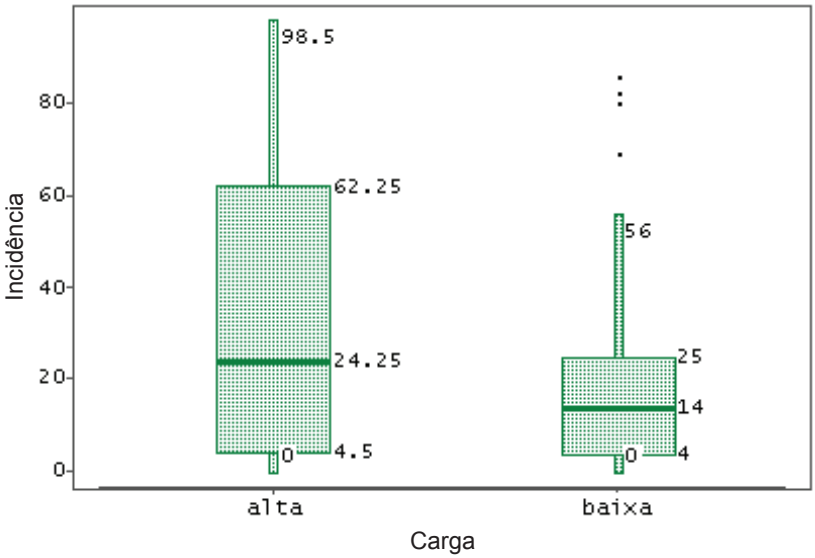

Figura 7. Distribuição dos valores de incidência da ferrugem do cafeeiro de acordo com a carga pendente de frutos da lavoura.

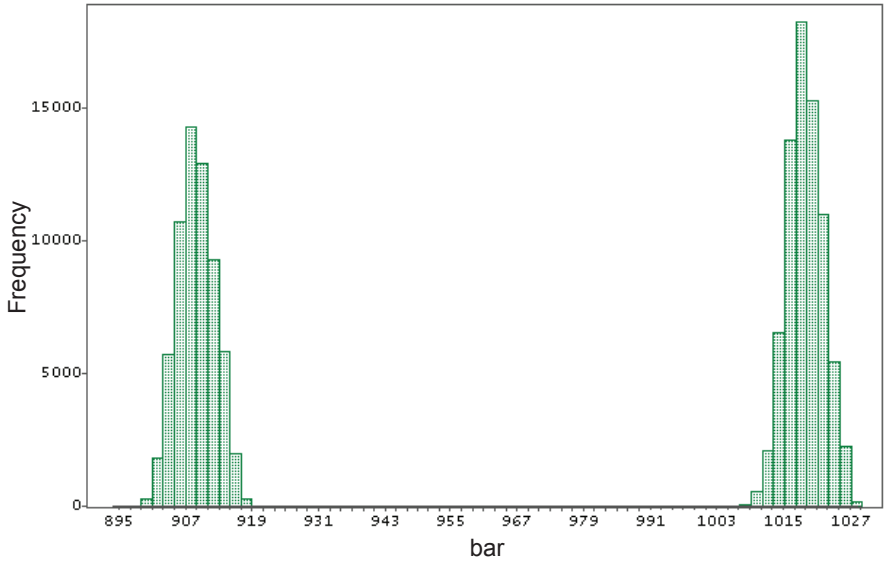

Figura 8. Histograma com distribuição de valores de pressão barométrica.

Estatísticas descritivas dos dados meteorológicos (valores máximos e mínimos, quantidade de registros e de valores nulos para cada atributo etc.) também são úteis, tanto no entendimento dos dados como no auxílio à verificação da qualidade desses dados.

Relato da experiência na instância do processo:

Para exemplificar a capacidade representativa de um gráfico de exploração de dados, considerar o histograma da Figura 8. Percebe-se, claramente, pela visualização da distribuição dos valores, que houve algum problema. Esse problema, esclarecido posteriormente, foi um período em que o sensor da estação meteorológica esteve descalibrado (faixa de distribuição à direita). O pessoal da Fundação Procafé nunca havia detectado esse problema.

Verificar a qualidade dos dados

Neste ponto, o analista examina a qualidade dos dados, verificando, por exemplo, se os dados estão completos. Valores faltantes ocorrem com frequência, especialmente se os dados foram coletados durante longos períodos de tempo. Alguns itens comuns a checar incluem: atributos faltantes e campos em branco; se os valores possíveis estão representados; se os valores fazem sentido; a grafia dos valores; e se atributos com valores diferentes têm significados semelhantes. O analista de dados também deve rever qualquer atributo ou registro que esteja em conflito com o senso comum – por exemplo, adolescentes com rendimentos bastante elevados.

Tarefa especializada: Inspecionar os dados meteorológicos

Além de elaborar e conferir testes de consistência dos dados meteorológicos, como os da Tabela 2, é importante inspecionar os arquivos gerados por uma estação meteorológica. Pode parecer uma tarefa desnecessária, capaz até de causar certo receio ao analista de dados, dado ao volume de dados registrado, mas uma simples inspeção visual pode revelar bastante sobre a qualidade dos dados.

Relato da experiência na instância do processo:

A inspeção visual nos arquivos gerados pela estação meteorológica da Fundação Procafé permitiu identificar problemas do tipo: registros ausentes; registros repetidos; registros não pertencentes ao mês correspondente

Tabela 2. Testes de consistência para atributos registrados por estação meteorológica.

Atributo DATA	Teste:	DATA = DD/MM/AAAA
Atributo HORA	Teste:	HORA = HH:[00|30]
Atributo TEMP	Teste:	TEMP > 0
Atributo TMAX	Testes:	TMAX > 0 TMAX ≥ TEMP
Atributo TMIN	Testes:	TMIN > 0 TMIN ≤ TEMP
Atributo VVENTO	Testes:	VVENTO ≥ 0 VVENTO ≤ 150
Atributo PRECIP	Testes:	PRECIP ≥ 0 PRECIP múltiplo de 0,2[1]
Atributo UR	Testes:	1. UR ≥ 0 2. UR ≥ 20

[1] Caso o coletor da estação meteorológica registre a precipitação em incrementos de 0,2 mm.

do arquivo da estação; registros para um dia que não existe (29 de fevereiro de um ano que não foi bissexto); e valores inconsistentes de determinados atributos.

4.3 Preparação dos dados

A fase de preparação dos dados abrange todas as atividades para construir, a partir dos dados brutos iniciais, o conjunto de dados final, ou seja, os dados que servirão de entrada para as ferramentas de modelagem. As tarefas incluem seleção de dados, limpeza dos dados, construção de dados, integração dos dados e formatação dos dados.

Selecionar os dados

A decisão de quais dados serão utilizados para a análise se baseia em vários critérios, incluindo a sua relevância para as metas de mineração de dados, bem como restrições técnicas e de qualidade, como limites em

volume ou tipos de dados. Parte do processo de seleção dos dados deve envolver a explicação do porquê certos dados foram incluídos ou excluídos. Também, é uma boa ideia decidir se um ou mais atributos são mais importantes do que outros.

A melhor forma de selecionar atributos relevantes é manualmente, baseada em conhecimento do significado dos atributos e do problema de aprendizado. Entretanto, métodos automáticos também são úteis, principalmente quando a quantidade de atributos é grande. Reduzir a dimensionalidade dos dados melhora o desempenho dos algoritmos e ajuda a produzir modelos mais compactos e mais adequados para interpretação, quando for o caso, permitindo aos usuários focarem sua atenção nas variáveis mais importantes (WITTEN et al., 2011).

Existem duas abordagens diferentes para a seleção de um bom subconjunto de atributos: uma delas é fazer uma avaliação independente baseada em características gerais dos dados; a outra é avaliar o subconjunto usando o mesmo algoritmo de aprendizado a ser aplicado na modelagem. A primeira abordagem é chamada de filtro, pois o conjunto de atributos é filtrado para selecionar aqueles que se mostram mais promissores. Dentre diversas técnicas e algoritmos disponíveis, pode-se citar *Correlation Feature Selection* (CFS), *InfoGain*, *GainRatio* e *Chi-square* (WITTEN et al., 2011). A segunda abordagem é chamada de *wrapper*, pois o próprio algoritmo de aprendizado é envolvido no procedimento de seleção (WITTEN et al., 2011).

Limpar os dados

Sem limpar os dados, os resultados de uma análise de mineração de dados são duvidosos. Assim, neste ponto, o analista de dados deve selecionar subconjuntos limpos de dados ou utilizar técnicas mais elaboradas, como estimar dados faltantes por meio de análises de modelagem. Por exemplo, um atributo com valor numérico ausente pode ser estimado de modo simples, atribuindo-se o valor médio do atributo no conjunto de treinamento, ou pode-se usar técnicas para estimar o valor ausente, como uma regressão linear a partir dos valores de outro atributo ou uma análise de agrupamento. O analista de dados deve se certificar de que descreveu como foi tratado cada problema de qualidade relatado na tarefa anterior “verificar a qualidade dos dados”.

Construir os dados

Após os dados estarem limpos, o analista de dados deve realizar operações de construção de dados, tais como criar registros novos ou produzir atributos derivados. Um exemplo de um novo registro seria a criação de um registro de compras vazio para clientes que não fizeram compras durante o ano anterior. Atributos derivados, ao contrário, são novos atributos que são construídos a partir de atributos existentes, tais como Área = Comprimento x Largura. Estes atributos derivados só devem ser adicionados se eles facilitam o processo ou o algoritmo de modelagem, não apenas para reduzir o número de atributos de entrada. Por exemplo, talvez "rendimento per capita" seja um atributo melhor/mais fácil para usar do que "rendimento por domicílio". Outro tipo de atributo deriva de transformações em um único atributo, geralmente executadas para atender às necessidades das ferramentas de modelagem. Estas transformações podem ser necessárias para transformar intervalos numéricos em campos categóricos (por exemplo, de idades para faixas de idade) ou vice-versa.

Tarefa especializada: Definir as classes do atributo meta

Em problemas de classificação, o atributo meta deve ser categórico. O ideal é quando já se tem esse atributo com as classes definidas. Em algumas situações, no entanto, é preciso derivar o atributo meta a partir de um atributo numérico contínuo. A divisão do intervalo contínuo em intervalos delimitados, que formam as classes, pode ser feita com o apoio de ferramentas computacionais ou com o auxílio de especialistas e/ou de literatura específica.

Relato da experiência na instância do processo:

Embora as ferramentas de modelagem possuíssem a funcionalidade de transformar atributos numéricos em categóricos, procurou-se, na literatura, resultados que permitissem a definição das classes da taxa de infecção da ferrugem do cafeeiro, por recomendação do especialista do domínio (MEIRA, 2008).

Tarefa especializada: Derivar os atributos preditivos meteorológicos

É comum se ter a impressão de que atributos com derivação mais elaborada podem dotar os modelos de maior acurácia. Contudo, pode ser que não funcione dessa forma.

Uma sugestão é começar pela derivação de atributos mais diretos e, portanto, mais simples; em seguida, proceder a uma rodada de modelagem e de avaliação dos modelos; depois, conforme a necessidade, ou mesmo se já estiver planejado dessa forma, partir para a derivação de atributos mais elaborados, com vistas a melhorar o desempenho dos modelos.

Esta sugestão permite que os resultados, mesmo que parciais, sejam obtidos de maneira mais rápida do que realizar toda a fase de preparação dos dados pretendida antes de se encaminhar para a fase de modelagem.

Relato da experiência na instância do processo:

A sugestão dada decorre da própria experiência, com a preparação dos dados, para a obtenção dos modelos de alerta da ferrugem do cafeeiro. Procurou-se pensar e realizar a preparação dos dados de maneira completa, sempre com o auxílio e o conhecimento do especialista do domínio; os programas foram implementados para derivar todos os atributos imaginados e considerados importantes, desde os mais simples até os mais elaborados.

Isso se justifica em um projeto de pesquisa com mais tempo, por exemplo um doutorado, mas pode não ser adequado para projetos em que seja importante apresentar resultados em um período de tempo mais curto.

Além disso, os programas de preparação dos dados foram implementados para proporcionar flexibilidade, permitindo a atribuição de alguns parâmetros (limite de umidade para considerar molhamento foliar; período de incubação fixo ou dependente de equação, qual equação considerar etc.). Quase a totalidade desses parâmetros foi mantida fixa desde a determinação dos valores iniciais, pela adequação presumida destes, com o aval do especialista do domínio, mas também pela impossibilidade de se testar todas as possíveis alternativas.

- Considerar o período de incubação na derivação

  No caso de doenças fúngicas, é preciso considerar o período de incubação[2] (PI) do fungo (referenciado por alguns autores como período latente[2]) na derivação dos atributos preditivos meteorológicos.

[2] Período de incubação é o intervalo de tempo entre a inoculação e o surgimento dos sintomas. Período latente é o tempo decorrido desde a penetração do patógeno até a sua esporulação em pústulas ou lesões.

O PI pode ser considerado fixo (KUSHALAPPA et al., 1983) ou variável, segundo alguma equação já desenvolvida. No caso do PI fixo, a estimativa dos períodos de infecção é mais grosseira e podem ocorrer distorções/contaminações mais significativas nos dados preparados.

Relato da experiência na instância do processo:

No projeto, o período de incubação foi considerado variável, dependente das médias das temperaturas máximas e mínimas durante o período, segundo a equação desenvolvida por Moraes et al. (1976).

- Estimar o período de molhamento foliar

  Caso se queira considerar o molhamento foliar na análise dos dados, e não se tenha ou não se queira usar registros de sensores de molhamento, é possível se fazer uma estimativa da duração do molhamento foliar diário.

  A forma mais comum utilizada para estimar o período de molhamento é por meio da duração da umidade relativa do ar acima de um limite específico, como 90% ou 95% (JENSEN; BOYLE, 1966; SUTTON et al.,1984). Neste caso, é recomendado se considerar os registros da estação meteorológica no decorrer de um “dia epidemiológico” - de 12h de um dia até 12h do dia seguinte -, pois os períodos de molhamento ocorrem geralmente entre um dia e o outro.

  Existem outras maneiras de se estimar os períodos de molhamento foliar (HUBER; GILLESPIE, 1992). Inclusive, a indução de árvores de decisão já foi utilizada para obtenção de um modelo de estimativa da duração de períodos de molhamento (GLEASON et al., 1994).

  Os períodos de molhamento foliar podem ser analisados tanto na sua extensão total como na sua fração noturna (das 20h às 8h), já que a infecção ocorre preferencialmente na ausência de ou com pouca luminosidade (MONTOYA; CHAVES, 1974).

  DICA: quando se tem que calcular o número de horas do dia em que ocorreu, ou se manteve, determinada situação, como é o caso do número de horas de molhamento foliar, é preciso estar atento para considerar um registro às 24 horas de cada dia. Por exemplo, em uma estação que registra dados de zero hora às 23h30 de um dia, é preciso considerar o registro à 0:00 h do dia seguinte como o registro às 24 horas daquele dia; caso contrário, o dia não vai ter 24 horas.

Relato da experiência na instância do processo:

De início, o molhamento foliar foi considerado quando a umidade relativa do ar era maior ou igual a 90%. Chegou-se a gerar alguns modelos com a taxa de infecção categórica com três classes como o atributo meta.

Posteriormente, um modelo experimental mostrou que 95% seria um limite de decisão mais adequado para o conjunto de dados disponível. Esse modelo foi uma árvore de decisão com a umidade relativa sendo o único atributo preditivo e o atributo meta indicando a existência ou não de molhamento foliar ('1' = molhado e '0' = seco) – esse atributo meta binário foi construído a partir dos registros do sensor de molhamento foliar da estação meteorológica.

Integrar os dados

A integração dos dados envolve combinar informações de múltiplas tabelas ou registros para criar novos registros ou valores. Por exemplo, um analista de dados pode juntar duas ou mais tabelas que possuem diferentes informações sobre os mesmos objetos. Integração também abrange agregações, que se referem a operações onde novos valores são calculados pela sumarização de dados de vários registros e/ou tabelas.

Formatar os dados

Em alguns casos, é preciso alterar o formato dos dados. Essas alterações podem ser simples – por exemplo, removendo caracteres inválidos de cadeias de caracteres – ou podem ser mais complexas, como aquelas que envolvem uma reorganização completa dos dados. Mudanças podem ser necessárias para tornar os dados adequados para uma ferramenta específica de modelagem.

4.4 Modelagem

Nesta fase, diversas técnicas de modelagem podem ser selecionadas e aplicadas, e seus parâmetros calibrados para valores ótimos. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas têm requisitos específicos com

respeito ao formato dos dados. Portanto, o retorno para a fase de preparação dos dados pode ser necessário. As tarefas de modelagem incluem a seleção da técnica de modelagem, a elaboração do projeto de teste, a criação de modelos e a avaliação dos modelos.

Selecionar a técnica de modelagem

Esta tarefa se refere a escolher uma ou mais técnicas de modelagem específicas, tais como árvores de decisão ou redes neurais. Se existem premissas associadas à técnica de modelagem, estas devem ser registadas.

Relato da experiência na instância do processo:

No caso da modelagem da ferrugem do cafeeiro, um dos principais objetivos era procurar entender as relações entre os atributos preditivos e o atributo meta representado pela taxa de infecção da doença, ou seja, entender como as variações nas condições meteorológicas conduziram as epidemias da ferrugem nas lavouras de café. Por este motivo e conforme já discutido anteriormente na tarefa “determinar as metas de mineração de dados”, foi escolhida a técnica de indução de árvores de decisão.

Elaborar o projeto de teste

Após construir um modelo, o analista de dados deve testar a qualidade e a validade do modelo, executando testes empíricos para determinar a força do modelo. Em tarefas supervisionadas de mineração de dados, como a classificação, é comum usar taxas de erro como medidas de qualidade para modelos. Uma estratégia é separar o conjunto de dados em conjuntos de treinamento e de teste. Se constrói o modelo sobre o conjunto de treinamento e se estima a sua qualidade sobre o conjunto de teste. Ou seja, o analista de dados desenvolve o modelo com base em um conjunto de dados e testa sua validade usando um conjunto de dados separado. É conveniente planejar o procedimento de teste antes de criar os modelos.

Construir modelos

Após planejar como serão feitos os testes, o analista de dados executa a(s) ferramenta(s) de modelagem sobre o conjunto de dados preparado

para criar um ou mais modelos. Existem muitas ferramentas que podem ser usadas para mineração de dados, desde soluções proprietárias até, e principalmente, soluções livres. Para escolher dentre elas, uma boa ideia é consultar as pesquisas divulgadas no site KDnuggets™. A pesquisa mais atual, até a data desta publicação, indicou RapidMiner e R, ambas software livre, como as ferramentas mais populares, com SAS sendo a primeira entre as ferramentas comerciais (KDNUGGETS, 2012d). O Weka e o SAS Enterprise Miner™, que é uma camada de interface específica para mineração de dados sobre o pacote SAS, apareceram na 7ª e 13ª posições da pesquisa, respectivamente.

Tarefa especializada: Construir modelos aplicados na epidemiologia de uma doença

A técnica de indução de árvores de decisão se mostrou bastante interessante quanto à sua aplicação na epidemiologia de doenças de plantas.

No caso dessa aplicação, considera-se como características desejáveis para a ferramenta de modelagem: (1) a possibilidade de indução interativa da árvore de decisão e (2) a visualização da distribuição de probabilidade entre as classes nos nós internos da árvore, além dos nós folhas.

A indução interativa proporciona flexibilidade e liberdade ao analista de dados, em conjunto com o especialista do domínio, para verificarem e analisarem diversas possibilidades de configuração, ou estruturação, da árvore de decisão.

É interessante que o processo de indução interativo permita se construir a árvore de decisão desde o nó raiz. Em cada nó que se queira ramificar, deve aparecer uma relação dos atributos mais importantes, de acordo com o critério de seleção de atributos. A partir daí, deve-se permitir escolher o mesmo atributo que seria escolhido no processo automático de indução ou escolher outro atributo de interesse com ganho de informação equivalente.

A indução interativa pode também partir da árvore de decisão gerada pelo processo automático. Neste caso, deve-se permitir a troca de atributos de teste nos nós internos e, também, a poda ou ramificação dos nós já criados.

A visualização da distribuição de probabilidade entre as classes nos nós internos da árvore de decisão permite melhorar a percepção e a compreensão dos relacionamentos entre os atributos preditivos/explicativos e a doença.

Essa visualização fica ainda melhor se a ferramenta de modelagem permitir a coloração dos nós da árvore de decisão conforme a proporção de distribuição de uma das classes, que pode ser a classe correspondente ao nível mais intenso de ataque da doença.

Relato da experiência na instância do processo:

A visualização da distribuição de probabilidade entre as classes, com coloração proporcional à classe de maior interesse, característica disponível na ferramenta SAS Enterprise Miner™, foi o que permitiu descobrir o potencial de aplicação da técnica de indução de árvores de decisão na epidemiologia da ferrugem do cafeeiro (MEIRA et al., 2008).

A indução interativa, também disponível no software Enterprise Miner™, apesar de não ter sido usada diretamente na indução da árvore de decisão para analisar a epidemia da ferrugem do cafeeiro, permitiu a verificação de todas as alternativas de escolha dos atributos de teste, em cada nó da árvore, auxiliando na discussão dos resultados.

Tarefa especializada: Construir modelos aplicados em alertas de uma doença

Quando o interesse está em modelos de alerta de uma doença, parece ser mais adequado definir o atributo meta como binário. Primeiro, porque o alerta fica melhor caracterizado com apenas duas classes: ‘1’ e ‘0’ ou ‘positivo’ e ‘negativo’. Segundo, porque os modelos tendem a ter um número menor de regras (menor complexidade) e a acertar mais (melhor acurácia).

Relato da experiência na instância do processo:

Os modelos desenvolvidos para a ferrugem do cafeeiro com as taxas de infecção binárias (MEIRA, 2008; MEIRA et al., 2009) são realmente mais indicados para a emissão de alertas, do que o modelo considerando a taxa de infecção categórica com três classes (MEIRA, 2008; MEIRA et al., 2008).

A separação entre modelos para lavouras com alta carga pendente e para lavouras com baixa carga pendente também contribuiu para um melhor desempenho dos modelos com as taxas de infecção binárias, especialmente aqueles para lavouras com alta carga pendente (MEIRA et al., 2009).

Avaliar os modelos

O analista de dados interpreta os modelos de acordo com o seu conhecimento do domínio, os critérios de sucesso de mineração de dados e o projeto de teste desejado. Ele avalia tecnicamente o sucesso da aplicação das técnicas de modelagem, mas deve também trabalhar com especialistas do domínio para interpretar os resultados da mineração de dados no contexto do problema. O analista de dados ainda pode optar por ter o especialista do domínio envolvido na criação dos modelos.

Neste ponto, o analista de mineração de dados pode também ordenar os modelos. Ele avalia os modelos de acordo com o critério de avaliação e leva em consideração os objetivos do projeto e os critérios de sucesso. Na maioria dos projetos de mineração de dados, o analista aplica uma única técnica mais de uma vez ou gera resultados de mineração de dados com diferentes técnicas alternativas. Nesta tarefa, o analista também compara todos os resultados de acordo com os critérios de avaliação.

Tarefa especializada: Avaliar os modelos por meio de validação cruzada

Epidemiologia e alertas de doenças de plantas é o tipo de aplicação com poucos dados para a modelagem, o que dificulta dividir esses dados entre um conjunto de treinamento e outro conjunto de teste. Então, para a avaliação dos modelos, a validação cruzada é uma das alternativas recomendáveis para a obtenção da acurácia e de outras medidas cabíveis (WITTEN et al., 2011).

É importante que se faça uma avaliação inicial da ferramenta de modelagem no momento do planejamento do projeto. Nessa avaliação, todos os aspectos do processo de KDD devem ser considerados, de forma geral, mas ressalta-se que os aspectos da avaliação dos modelos não podem ser esquecidos.

Relato da experiência na instância do processo:

O software Weka permite avaliar os modelos obtidos por meio de validação cruzada. Já o Enterprise Miner™, na sua versão 4.3, não possuía essa funcionalidade pronta. Foi preciso implementar a validação cruzada com programação específica na linguagem do SAS®, o que demandou esforço e tempo consideráveis.

4.5 Avaliação

Antes do analista de dados proceder com a distribuição final de algum modelo construído, é importante avaliar o modelo de forma mais completa e revisar a sua construção para se certificar de que ele atende os objetivos do negócio ou da pesquisa. É crucial determinar se alguma questão importante não foi suficientemente considerada. No final desta fase, o líder do projeto deve decidir como exatamente usar os resultados da mineração de dados. As tarefas chaves aqui são a avaliação dos resultados, a revisão do processo e a determinação dos próximos passos.

Avaliar os resultados

A avaliação na fase anterior de modelagem trata de fatores como a acurácia e a generalidade do modelo (WITTEN et al., 2011). Esta tarefa, por sua vez, avalia o grau com que o modelo atende os objetivos definidos e determina se existe alguma razão pela qual esse modelo pode ser considerado deficiente. Outra opção é testar o(s) modelo(s) em aplicações no mundo real, se as restrições de tempo e de orçamento permitirem. Além disso, a avaliação também procura desvendar desafios adicionais e informações ou sugestões para direcionamentos futuros.

Revisar o processo

Neste momento, é conveniente fazer uma revisão mais profunda do esforço de mineração de dados para determinar se existe algum fator importante ou tarefa que, de alguma forma, não tenha tido a devida atenção. Esta revisão também cobre questões de garantia da qualidade - por exemplo: Os modelos foram construídos corretamente? Foram usados apenas atributos permitidos, que estão disponíveis para implantação futura?

Determinar os próximos passos

Neste ponto, o líder do projeto deve decidir se finaliza o projeto e avança para a distribuição, se inicia novas iterações do processo ou se cria novos projetos de mineração de dados.

4.6 Distribuição

A criação de modelos geralmente não é o fim do projeto. O conhecimento adquirido deve ser organizado e apresentado de forma que se possa usá-lo. Dependendo dos requisitos, a fase de distribuição pode ser tão simples como gerar um relatório ou tão complexa como implementar um processo de mineração de dados que possa ser repetido por toda a organização.

Embora seja o cliente, não o analista de dados, que frequentemente realiza as etapas de distribuição, é importante para o cliente entender com antecedência quais ações devem ser tomadas, a fim de realmente utilizar os modelos criados. As tarefas chaves são planejar a distribuição, planejar o monitoramento e a manutenção, produzir o relatório final e revisar o projeto.

Planejar a distribuição

Esta tarefa considera os resultados da avaliação e desenvolve uma estratégia para a distribuição do(s) resultado(s) da mineração de dados.

Planejar o monitoramento e a manutenção

O monitoramento e a manutenção são questões importantes no caso de o resultado da mineração de dados for se tornar parte da atividade diária de uma organização e do seu ambiente. Uma estratégia de manutenção cuidadosamente preparada evita o uso incorreto dos resultados da mineração de dados.

Elaborar o relatório final

Ao final do projeto, o líder e sua equipe devem elaborar um relatório final. Dependendo do plano de distribuição, este relatório pode ser apenas um resumo do projeto e suas experiências ou pode ser uma apresentação final e global dos resultados da mineração de dados. Também, muitas vezes haverá uma reunião de conclusão do projeto, onde os resultados são verbalmente apresentados ao cliente.

Revisar o projeto

O analista de dados deve avaliar falhas e êxitos, bem como áreas potenciais de melhoria para usar em projetos futuros. Esta etapa deve incluir um resumo das experiências importantes durante o projeto e pode incluir entrevistas com os principais participantes do projeto. Este documento poderia incluir abordagens equivocadas ou dicas para selecionar as técnicas de mineração de dados mais adequadas em situações semelhantes. Em projetos ideais, a documentação da experiência abrange também quaisquer relatórios escritos por membros do projeto durante a realização das fases e tarefas.

5 Considerações finais

O intuito, na elaboração deste documento, não foi o de produzir, por completo, um processo de descoberta de conhecimento em bases de dados com todas as tarefas especializadas possíveis, de acordo com o contexto de mineração de dados definido, para todas as fases e tarefas genéricas do modelo do processo CRISP-DM.

O que se procurou fazer foi contemplar o que aconteceu de importante (as “lições aprendidas”) no transcorrer da instância do processo para a obtenção dos modelos da ferrugem do cafeeiro, que possa vir a ser útil em iniciativas futuras.

Espera-se que o registro desses acontecimentos, organizado de acordo com a estrutura genérica do modelo hierárquico da metodologia CRISP-DM, possa servir de orientação para novos projetos no mesmo domínio de aplicação e em contextos de mineração de dados similares.

Além disso, e tão importante quanto, espera-se que esta versão inicial da metodologia seja atualizada e incrementada com o passar do tempo, em virtude de novas experiências coletadas em projetos futuros de descoberta de conhecimento em bases de dados.

6 Referências

APTE, C.; WEISS, S. Data mining with decision trees and decision rules. **Future Generation Computer Systems**, v. 13, p. 197-210, 1997.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and regression trees**. Boca Raton: CRC Press, 1984. 358 p.

BUTT, D. J.; ROYLE, D. J. Multiple regression analysis in the epidemiology of plant diseases. In: KRANZ, J. (Ed.). **Epidemics of plant diseases**: mathematical analysis and modeling. 2nd ed. Berlin: Springer-Verlag, 1990. p. 143-180.

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. 76 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: an overview. In: FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Ed.). **Advances in knowledge discovery and data mining**. Menlo Park: AAAI Press, 1996a. p. 1-34.

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

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: an overview. **AI Magazine**, v. 14, n. 3, p. 57-70, 1992.

GLEASON, M. L; TAYLOR, S. E., LOUGHIN, T. M.; KOEHLER, K. J. Development and validation of an empirical model to estimate the duration of dew periods. **Plant Disease**, v. 78, n. 10, p. 1011-1016, 1994.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA data mining software: an update. **SIGKDD Explorations**, v. 11, n. 1. 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3ed. San Francisco: Morgan Kaufmann Publishers, 2011.

HUBER, L.; GILLESPIE, T. J. Modeling leaf wetness in relation to plant disease epidemiology. **Annual Review of Phytopathology**, Palo Alto, v. 30, p. 553-577, 1992.

JAPIASSÚ, L. B.; GARCIA, A. W. R.; MIGUEL, A. E.; CARVALHO, C. H. S.; FERREIRA, R. A.; PADILHA, L.; MATIELLO, J. B. Influência da carga pendente, do espaçamento e de fatores climáticos no desenvolvimento da ferrugem do cafeeiro. In: SIMPÓSIO DE PESQUISA DOS CAFÉS DO BRASIL, 5., 2007, Águas de Lindóia, SP. **Anais**... Brasília, DF: Embrapa Café, 2007. 1 CD-ROM.

JENSEN, R. E.; BOYLE, L. W. A technique for forecasting leafspot on peanuts. **Plant Disease Reporter**, v. 50, n. 11, p. 810-814, 1966.

KDNUGGETS. **Poll**: what main methodology are you using for data mining? 2012a. Disponível em: <www.kdnuggets.com/polls/2002/methodology.htm>. Acesso em: 03 out. 2012.

KDNUGGETS. **Poll**: data mining methodology. 2012b. Disponível em: <www.kdnuggets.com/polls/2004/data_mining_methodology.htm>. Acesso em: 03 out. 2012

KDNUGGETS. **Poll**: data mining methodology. 2012c. Disponível em: <www.kdnuggets.com/polls/2007/data_mining_methodology.htm>. Acesso em: 03 out. 2012.

KDNUGGETS. **Poll**: data mining/analytic tools used. 2012d. Disponível em: <http://www.kdnuggets.com/polls/2011/tools-analytics-data-mining.html>. Acesso em: 03 dez. 2012.

KUSHALAPPA, A. C.; AKUTSU, M.; LUDWIG, A. Application of survival ratio for monocyclic process of Hemileia vastatrix in predicting coffee rust infection rates. **Phytopathology**, St. Paul, v. 73, n. 1, p. 96-103, 1983.

MEIRA, C. A. A. **Processo de descoberta de conhecimento em bases de dados para a análise e o alerta de doenças de culturas agrícolas e sua aplicação na ferrugem do cafeeiro**. 2008. 198p. Tese (Doutorado), Universidade Estadual de Campinas, Campinas.

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, p. 114-124, 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, p. 233-242, 2009.

MONTOYA, R. H.; CHAVES, G. M. Influência da temperatura e da luz na germinação, infectividade e período de geração de *Hemileia vastatrix* Berk. & Br. **Experientiae**, v. 18, n. 11, p. 239-266, 1974.

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

MORAES, S. A.; SUGIMORI, M. H.; RIBEIRO, I. J. A.; ORTOLANI, A. A.; PEDRO JR., M. J. Período de incubação de *Hemileia vastatrix* Berk. et Br. em três regiões do Estado de São Paulo. **Summa Phytopathologica**, v. 2, n. 1, p. 32-38, 1976.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. F. de Mineração de dados. In: REZENDE, S. O. (Org.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Ed. Manole, 2002. p. 307-335.

SAS INSTITUTE. **Getting started with SAS® Enterprise Miner™ 4.3**. Cary, NC: SAS Institute Inc., 2004. 126 p.

SHEARER, C. The CRISP-DM model: the new blueprint for data mining. **Journal of Data Warehousing**, v. 5, n. 4, p. 13-22, 2000.

SUTTON, J. C.; GILLESPIE, T. J.; HILDEBRAND, P. D. Monitoring weather factors in relation to plant disease. **Plant Disease**, v. 68, n. 1, p. 78-84, 1984.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. 3. ed. San Francisco: Morgan Kaufmann, 2011. 629 p.

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

Ministério da
Agricultura, Pecuária
e Abastecimento

GOVERNO FEDERAL
BRASIL
PAÍS RICO É PAÍS SEM POBREZA

CGPE 10217