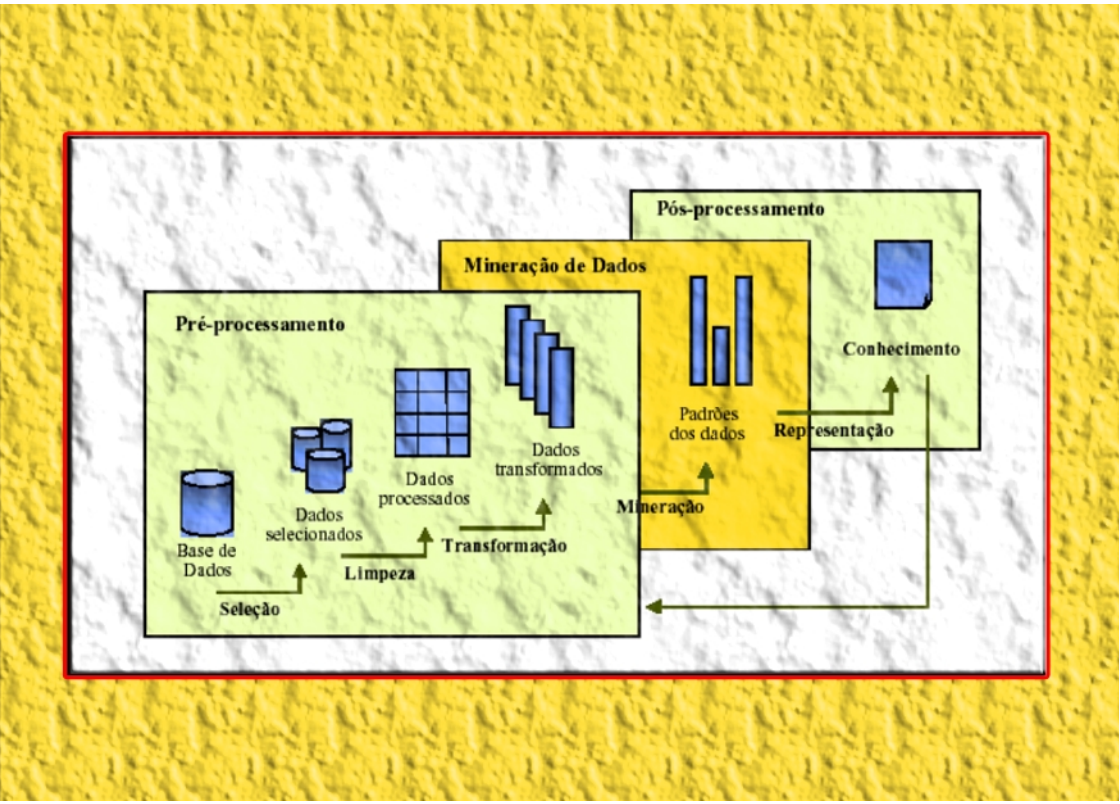
Ministério da Agricultura,
Pecuária e Abastecimento

Documentos

Novembro, 2002 20

ISSN 1677-9274

Mineração de Dados Climáticos para Previsão de Geada e Deficiência Hídrica para as Culturas do Café e da Cana-de-Açúcar para o Estado de São Paulo



Embrapa

ISSN 1677-9274
Novembro, 2002

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

Documentos 20

Mineração de Dados Climáticos para Previsão de Geada e Deficiência Hídrica para as Culturas do Café e da Cana-de-Açúcar para o Estado de São Paulo

Luciana Corpas Bucene
Luiz Henrique Antunes Rodrigues
Carlos Alberto Alves Meira

Campinas, SP
2002

Embrapa Informática Agropecuária
Área de Comunicação e Negócios (ACN)
Av. André Tosello, 209
Cidade Universitária "Zeferino Vaz" – Barão Geraldo
Caixa Postal 6041
13083-970 – Campinas, SP
Telefone (19) 3789-5743 - Fax (19) 3289-9594
URL: http://www.cnptia.embrapa.br
e-mail: sac@cnptia.embrapa.br

1ª. edição
on-line - 2002

Bucene, Luciana Corpas.
Mineração de dados climáticos para previsão de geada e deficiência hídrica para as culturas do café e da cana-de-açúcar para o Estado de São Paulo / Luciana Corpas Bucene, Luiz Henrique Antunes Rodrigues, Carlos Alberto Alves Meira. – Campinas : Embrapa Informática Agropecuária, 2002.
41 p. : il. – (Documentos / Embrapa Informática Agropecuária ; 20)

ISSN 1677-9274

1. Mineração de dados. 2. Previsão de geada. 3. Deficiência hídrica. 4. Café. 5. Cana-de-açúcar. I. Rodrigues, Luiz Henrique Antunes. II. Meira, Carlos Alberto Alves. III. Título. IV. Série.

CDD – 21st ed.
006.3

Autores

Luciana Corpas Bucene
Eng. Agrícola, M.Sc. em Geoprocessamento, Doutoranda da Faculdade de Engenharia Agrícola/Unicamp, Pesquisadora Colaboradora da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP.
Telefone (19) 3789-5791 – e-mail: lucianac@cnptia.embrapa.br

Luiz Henrique Antunes Rodrigues
Eng. Agrícola, Prof. Dr. da Faculdade de Engenharia Agrícola – Feagri/Unicamp, Cidade Universitária "Zeferino Vaz", Caixa Postal 6011 – 13083-970 – Campinas, SP.
Telefone (19) 3788-1000 – e-mail: lique@agr.unicamp.br

Carlos Alberto Alves Meira
M.Sc. em Ciência da Computação e Matemática Computacional, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP.
Telefone (19) 3789-5806 – e-mail: carlos@cnptia.embrapa.br

Apresentação

Este documento tem como objetivo identificar relações entre parâmetros climáticos, visando descobrir eventuais novos conhecimentos, através de técnicas de inteligência artificial, possibilitando a previsão de geada para a cultura de café e a previsão de déficit hídrico para as culturas de café e cana-de-açúcar no Estado de São Paulo, com maior grau de confiança e num intervalo de tempo satisfatório, com a intenção de auxiliar os produtores na tomada de decisões.

Este trabalho está vinculado ao projeto "Desenvolvimento e Evolução de um Sistema de Monitoramento Agroclimatológico para o Estado de São Paulo", desenvolvido pela Embrapa Informática Agropecuária, em conjunto com o Instituto Agronômico de Campinas (IAC) e Unicamp, ao qual se pretende adicionar novos produtos para incorporação no sistema de monitoramento agroclimatológico, além de propor uma nova infra-estrutura para o sistema já existente e evoluções nos modelos operacionais disponíveis.

Espera-se que esta publicação venha contribuir para o aprimoramento do sistema agroclimatológico do Estado de São Paulo e seja um instrumento útil para estudantes de graduação, pós-graduação e técnicos de áreas afins, suprindo a carência de bibliografia especializada no assunto, na língua portuguesa.

José Gilberto Jardine
Chefe-Geral

Sumário

Mineração de Dados Climáticos para Previsão de Geada e Deficiência Hídrica para as Culturas do Café e da Cana-de-Açúcar para o Estado de São Paulo

Luciana Corpas Bucene
Luiz Henrique Antunes Rodrigues
Carlos Alberto Alves Meira

Introdução

A descoberta automática de conhecimento a partir dos dados – usar os computadores para descobrir novas informações úteis – é um dos objetivos mais fascinantes da ciência da computação.

Cada vez mais, o volume de informações excede a capacidade de sua análise pelos métodos tradicionais (planilhas, consultas e gráficos). Esses métodos podem gerar relatórios a partir dos dados, mas não conseguem analisá-los sob o enfoque conhecimento. Para atender a essa necessidade foram pesquisadas e desenvolvidas novas técnicas e ferramentas, que permitem a extração de conhecimento a partir de grandes volumes de dados.

Mineração de dados ("data mining") é a descoberta de conhecimento interessante, mas escondido, em grandes bases de dados. Bases de dados corporativas freqüentemente contêm tendências desconhecidas, que são de importância estratégica para a organização. É uma tecnologia baseada numa nova geração de hardware e software que inclui análises estatísticas, exploração visual, árvores de decisão, redes neurais, entre

outras, para explorar grandes bases de dados e descobrir relações e padrões existentes nessas informações. Difere de técnicas estatísticas porque, ao invés de verificar padrões hipotéticos, utiliza os próprios dados para descobrir tais padrões.

As várias tarefas desenvolvidas em "data mining" têm como objetivo primário a predição e a descrição, ou seja, a produção de um modelo ou a produção de informação, respectivamente. A predição usa atributos para predizer o desconhecido ou os valores futuros de outras variáveis. "Data mining" utiliza técnicas estatísticas e de aprendizado de máquinas para construir modelos capazes de predizer o comportamento de determinado atributo. Como descrição, diversas técnicas podem trazer percepções diferentes das apresentadas em tabelas ou relatórios. A descrição procura por padrões que descrevem os dados e são interpretáveis facilmente pelos seres humanos.

O objetivo deste estudo é analisar bancos de dados climáticos disponíveis, buscando identificar novos conhecimentos, através da utilização de técnicas relacionadas à Mineração de Dados, possibilitando a previsão de ocorrência de geada e deficiência hídrica para as culturas de café e cana-de-açúcar no Estado de São Paulo.

Como qualquer iniciativa de mineração de dados, este trabalho parte da hipótese de que é possível descobrir conhecimento novo "escondido" no grande volume de dados climáticos e ainda, a partir do comportamento conhecido dos atributos climáticos, aumentar as chances de se descobrir padrões que podem explicar e ajudar a predizer o comportamento futuro dos mesmos. Estas hipóteses são reforçadas pela percepção dos especialistas em climatologia que possuem um claro sentimento de que é possível extrair informação nova e útil e predizer o comportamento futuro dos parâmetros climáticos. Para isso, serão aplicadas técnicas de mineração de dados nos grandes bancos de dados climáticos do Estado de São Paulo, possibilitando a previsão de geada para cultura do café e a previsão de deficiência hídrica no solo tanto para as culturas de café como de cana-de-açúcar no Estado de São Paulo, com maior grau de confiança e em um intervalo de tempo satisfatório, podendo, então, auxiliar os produtores na tomada de decisões visando a proteção contra essas ocorrências, reduzindo os impactos causados.

Mineração de Dados

A Descoberta de Conhecimento em Bancos de Dados (Knowledge Discovery of Database - KDD) é uma tecnologia que possui ferramentas poderosas para a descoberta eficiente de informações valorosas de uma grande coleção de dados, visando o auxílio no suporte a decisão. Mineração de dados ("data mining") é uma das ferramentas de KDD mais utilizadas, podendo ser considerada um passo dentro do processo de KDD. Há autores que utilizam os dois termos como sinônimos. A descoberta de conhecimento em base de dados é um processo não trivial de identificar padrões válidos, não conhecidos, potencialmente úteis e interpretáveis, consistindo, basicamente, em descobrir conhecimento útil nos dados armazenados, a partir da aplicação de técnicas de mineração de dados, da aplicação dos padrões obtidos e da interpretação dos resultados (Fayyad et al., 1996).

Esta área surgiu em 1981, devido a necessidade de métodos mais poderosos para a recuperação e utilização da informação, pois com o avanço da tecnologia, as bases de dados acumulam milhares de informações, aumentando expressivamente o volume de dados e a riqueza de suas informações. Como resultado desse aumento efetivo, o processamento dessas informações tornou-se cada vez mais complexo e difícil, e, normalmente, os dados ficam armazenados nas bases de dados sem que sejam utilizados de uma forma realmente eficiente (Halmenschlager, 2000).

Mineração de dados é entendida como o processo de exploração e análise de grandes quantidades de dados, com o objetivo de descobrir padrões ou regras que permitam uma melhor compreensão da informação contida nos mesmos. As ferramentas de mineração de dados podem prever futuras tendências e comportamentos, permitindo um novo processo de tomada de decisão, baseado principalmente no conhecimento acumulado e freqüentemente desprezado, contido em seus próprios bancos de dados (Fayyad et al., 1996). Enquanto as ferramentas tradicionais de banco de dados (transacionais) são capazes de mostrar "o que" está na base de dados, os softwares analíticos ajudam o usuário a descobrir o "porquê". Em um pacote estatístico, o usuário formula hipóteses com os prováveis porquês, para então testar suas validades. Mineração de dados estende a capacidade de gerar e validar hipóteses e por isso se diz que pode descobrir conhecimento novo (inesperado), útil e interessante (Munari, 2001).

Basicamente, mineração de dados se preocupa com a análise dos dados e com o uso de técnicas responsáveis por achar padrões e regularidades no conjunto de dados. É o computador que é o responsável por achar os padrões identificando as regras subjacentes e as características nos dados. A idéia é que é possível encontrar "ouro" em lugares inesperados, tal como os softwares de data mining extraem padrões não previamente encontrados ou tão óbvios que ninguém os notou antes (Monard et al., 2002).

O processo de descoberta de conhecimento envolve várias etapas complexas, que devem ser executadas corretamente, pois cada etapa é fundamental para que os objetivos estabelecidos e o sucesso completo da aplicação sejam alcançados. O processo é interativo, com muitas decisões a serem tomadas, e também iterativo, podendo possuir laços entre quaisquer das etapas, não existindo uma ordem ou seqüência única durante o andamento do processo.

Segundo Baranauskas & Monard (2000), no início do desenvolvimento, há a necessidade de preparação dos dados, fase considerada na literatura como a que consome mais tempo. Nesta fase há a necessidade do acompanhamento dos especialistas humanos visando auxiliar na identificação da relevância dos atributos.

Fayyad et al. (1996) representam um processo típico de mineração de dados, como mostra a Fig. 1.

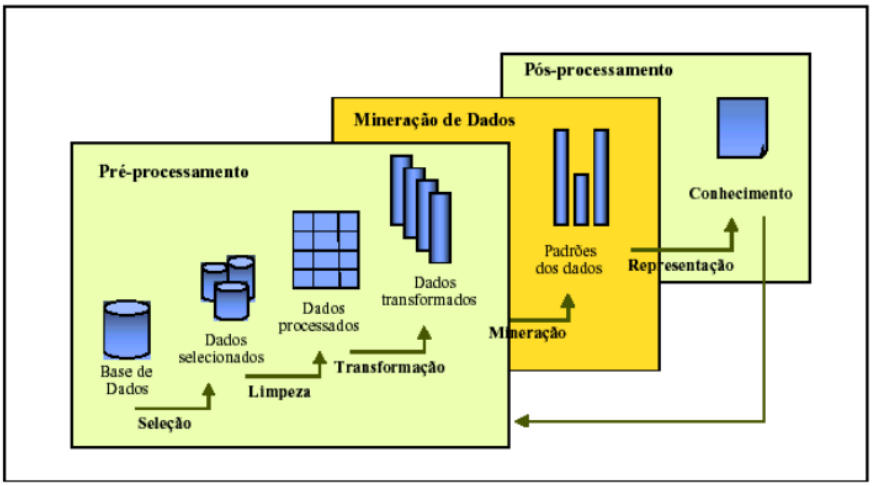


Fig. 1. Etapas do processo de descoberta do conhecimento.
Fonte: Fayyad et al. (1996).

Ele possui três passos. Inicialmente é preciso selecionar os tipos de dados que serão usados pelo algoritmo de mineração. Dados crus geralmente são variados, não estão organizados e nem todos são necessários para a mineração. Um grande esforço é necessário para se coletar uma boa quantidade de dados e transportá-los para um lugar onde se possa minerá--los. O primeiro passo é pré-processar os dados para aprontá-los para a análise. Usualmente os dados têm que ser formatados, amostrados, adaptados e, algumas vezes, transformados para que possam ser usados pelo algoritmo de mineração. Ocorre então, o desenvolvimento do entendimento do domínio da aplicação, avaliação do hardware e software disponíveis, seleção, limpeza e transformação dos dados. Após o pré-processamento, os dados estão prontos para serem minerados por um algoritmo. É definida a escolha da tarefa e das técnicas a serem utilizadas, identificação da ferramenta que satisfaça a essas condições e aplicação desta aos dados. Este passo pode envolver técnicas muito diversas e a informação descoberta é usada principalmente para construção de modelos, extração automática de padrões e exploração visual de dados. O último passo do processo de mineração de dados, o pós-processamento, é assimilar a informação minerada, chamado pós-processamento. É a interpretação dos resultados e incorporação do conhecimento adquirido. No caso da construção de modelos, este passo consiste em avaliar a robustez e efetividade dos modelos produzidos. No caso da extração de padrões e exploração visual de dados, este passo consiste em tentar interpretar a informação extraída.

A qualidade do conhecimento descoberto no final é dependente da qualidade do dado, do pré-processamento, do algoritmo de mineração e do processo de assimilação. Mais do que isso, a qualidade é altamente dependente de como o processo é montado como um todo.

Liu et al. (2001) utilizam técnicas de mineração de dados para predição de chuvas, baseado em uma série histórica de dados, alcançando ótimos resultados. McCullagh et al. (1999) desenvolveram um sistema inteligente, baseado em mineração de dados, utilizando técnicas de redes neurais artificiais, para estimar parâmetros meteorológicos, entre eles a precipitação. Os resultados mostram que o conhecimento adquirido a partir de mineração de dados, contribuiu para o sucesso do sistema desenvolvido. Howard & Rayward-Smith (1997) aplicaram técnicas de

descoberta de conhecimentos em uma base de dados meteorológica a fim de descobrir padrões climáticos. A construção de um modelo de classificação, baseado em fatores climáticos, de regiões com aptidão para o cultivo de uvas é indicada como aplicação potencial de mineração de dados em Witten et al. (1996).

Técnicas de Mineração de Dados

Na etapa de mineração de dados é definida a escolha da tarefa e da técnica a serem utilizadas, a identificação da ferramenta que satisfaça as condições exigidas e aplicação das ferramentas nos dados.

Mendonça Neto (2001) afirma que mineração de dados usa sistemas de aprendizado automatizados, que codificam informações de treinamento diretamente de repositórios de dados. Apesar da estrutura geral do processo de mineração de dados parecer similar ao de aprendizado de máquina, existem duas diferenças críticas. A primeira é que os dados crus do repositório onde se vai aplicar as técnicas de mineração foram derivados para outra finalidade. Provavelmente os dados não estão arrumados numa forma que irá facilitar a indução automática de conhecimento. Além disso, podem existir conjuntos de dados irrelevantes, incertos ou incompletos. A segunda diferença é que o produto da mineração não é necessariamente um modelo. Este produto, chamado de "informação minerada", terá ainda que passar pela interpretação de um perito no domínio de aplicação, para ser assimilada e transformada em conhecimento com valor real.

A automação dos processos de aprendizado tem sido estudada numa área da inteligência artificial chamada de aprendizado de máquina - "machine learning" (Baranauskas & Monard, 2000). O sistema típico de aprendizado de máquina não interage diretamente com o domínio (ambiente) externo. Ele usa informações codificadas ou de treinamento para aprender sobre este domínio. Ele amostra fatos do ambiente que se quer modelar e codifica estes fatos em conjuntos de informações de treinamento. Este conjunto de treinamento é usado para alimentar um mecanismo de aprendizado que irá produzir um modelo do ambiente observado. Este modelo pode ser usado para extrair informação útil e desconhecida sobre o domínio.

De acordo com Zaverucha et al. (2002), o aprendizado de máquina é uma área multidisciplinar de pesquisa, que compreende mecanismos pelos quais o conhecimento é adquirido através da experiência. Segundo Monard et al. (2002), a área de aprendizado de máquina estuda métodos computacionais apropriados para a aquisição de novos conhecimentos, novas habilidades e novas formas de organização do conhecimento já existente. Técnicas como Árvores de Decisão, Redes Neurais, entre outras, têm sido aplicada com sucesso. Paralelamente, avanços teóricos através de trabalhos de pesquisa na área de aprendizado de máquina têm definido limites para estes métodos, guiando a busca por novos modelos e aplicações, além de oferecerem embasamento teórico para os resultados experimentais obtidos. Por outro lado, a grande quantidade de informação armazenada em meio digital nas atuais organizações é atualmente um dos problemas mais graves trazidos com o advento da tecnologia. A maioria destas informações está armazenada em bases de dados, cujo tamanho cresce exponencialmente devido ao aparecimento de aplicações como meteorologia, Sistemas de Informações Geográficas, e outras, cujo volume de dados gerado é grande. Neste sentido, a extração de conhecimento em bases de dados, em cujo processo estão inseridas técnicas de mineração de dados, é uma forte tendência, e vem se estabelecendo como uma das áreas de pesquisa mais interessantes dos últimos tempos. Recentemente, a interligação das áreas de aprendizado de máquina e de extração de conhecimentos em bases de dados tem se tornado cada vez mais importante, na medida em que a manipulação e análise manuais do grande volume de dados armazenados pelas aplicações têm se tornado inviáveis (Zaverucha et al., 2002).

A escolha de quais técnicas de mineração de dados usar depende das metas do perito no domínio e das tarefas para atingir estas metas. As tarefas de mineração podem ser classificadas em seis principais categorias (Mendonça Neto, 2001):

1. **Estimação e predição -** estimação consiste em examinar atributos de um conjunto de entidades e, baseado nos valores destes atributos, assinalar valores e atributos de uma nova entidade que se quer caracterizar. A predição usa atributos para predizer o desconhecido ou os valores futuros de outras variáveis.
2. **Classificação -** consiste em examinar os atributos de uma determinada entidade para, baseada nestes atributos, assinalar esta entidade a uma determinada classe ou categoria.

3. **Descoberta de Associações -** consiste em identificar quais atributos estão associados com outros em um dado ambiente.
4. **Descoberta de Aglomerações (*Clustering*) -** consiste em segmentar uma população heterogênea em subgrupos homogêneos de entidades, com base na auto-similaridade entre registros. Esta técnica agrupa informações homogêneas de grupos heterogêneos entre os demais e aponta o item que melhor representa cada grupo, permitindo, desta forma, perceber as características de cada grupo.
5. **Visualização de Dados -** é a tarefa de descrever informações complexas através de apresentações visuais, como por exemplo, gráficos, planilhas, diagramas, etc.
6. **Exploração Iterativa de Dados -** é o processo de inspecionar grandes volumes de dados através de controles iterativos, que permitem rapidamente explorar novos cenários e questões abertas durante a análise dos dados.

Predição e classificação, têm por objetivo construir modelos explícitos que estão prontos para serem empregados por uma organização. Descoberta de associações e aglomerações têm por objetivo identificar padrões potencialmente úteis nos dados explorados. Estes padrões têm que ser interpretados por um perito no domínio para que ele perceba nestes padrões fatos de real valor. Visualização e exploração iterativa de dados, objetivam auxiliar os peritos no domínio a encontrar por eles próprios padrões interessantes nos dados explorados.

As técnicas para a execução dessas tarefas são variadas (Mendonça Neto, 2001), entre elas: árvores de decisão, redes neurais artificiais, regras de indução, mineração visual de dados e clusterização. A adequabilidade do tipo da função de mineração de dados ao tipo de problema que se está querendo solucionar, juntamente com a quantidade e qualidade dos dados são os fatores fundamentais para definir a técnica mais adequada de execução. Normalmente, os produtos para mineração de dados combinam as diversas técnicas, para se construir um produto mais preciso e mais rápido.

Árvores de Decisão

Árvores de decisão ou de classificação são técnicas de indução usadas para descobrir regras de classificação para um atributo a partir da subdivisão sistemática dos dados contidos no repositório que está sendo analisado.

As árvores de decisão consistem de nodos que representam os atributos, de arcos, provenientes destes nodos e que recebem os valores possíveis para estes atributos, e de nodos folha, que representam as diferentes classes de um conjunto de treinamento, como ilustra a Fig. 2. Uma árvore de decisão tem a função de particionar recursivamente um conjunto de treinamento, até que cada subconjunto obtido deste particionamento contenha casos de uma única classe. Para atingir esta meta, a técnica de árvores de decisão examina e compara a distribuição de classes durante a construção da árvore. O resultado obtido, após a construção de uma árvore de decisão, são dados organizados de maneira compacta, que são utilizados para classificar novos casos (Garcia & Alvares, 2002).

A partir de uma árvore de decisão é possível derivar regras. As regras são escritas considerando o trajeto do nodo raiz até uma folha da árvore. As regras e a árvore de decisão são geralmente utilizadas em conjunto. Devido ao fato das árvores de decisão tenderem a crescer muito, de acordo com algumas aplicações, elas são muitas vezes substituídas pelas regras. Isto acontece em virtude das regras poderem ser facilmente modularizadas. Uma regra pode ser compreendida sem que haja a necessidade de se referenciar outras regras (Ingargiola, 2002).

Nesta técnica escolhe-se a variável que se quer avaliar e o software procura as mais correlacionadas e monta a árvore com várias ramificações. As árvores de decisão são meios de representar resultados de mineração de dados na forma de árvore, e que lembram um gráfico organizacional horizontal. A partir de um grupo de dados com numerosas colunas e linhas, uma ferramenta de árvore de decisão pede ao usuário para escolher uma das colunas como objeto de saída, e aí mostra o único e mais importante fator correlacionado com aquele objeto de saída como o primeiro ramo (nó) da árvore de decisão. Isso significa que se pode rapidamente ver qual o fator que mais direciona o objeto de saída, e o pode entender porque o fator foi escolhido. Uma boa ferramenta de árvore de decisão vai, também, permitir que se explore a árvore de acordo com a sua vontade, do mesmo modo que poderá encontrar grupos alvos que lhe interessem mais, e aí ampliar o dado exato associado ao seu grupo alvo. Os usuários podem, também, selecionar os dados fundamentais em qualquer nó da árvore, movendo-o para dentro de uma planilha ou outra ferramenta para análise posterior. Nesta técnica consegue-se saber os itens que mais influenciam uma determinada variável (Gimenes & Seixas, 2000). Segundo Brazdil (2002), muitos são os algoritmos de classificação que elaboram árvores de decisão. Não há uma forma de determinar qual é o melhor algoritmo, um

pode ter melhor desempenho em determinada situação e outro algoritmo pode ser mais eficiente em outros tipos de situações. O algoritmo ID3 foi um dos primeiros algoritmos de árvore de decisão, tendo sua elaboração baseada em sistemas de inferência e em conceitos de sistemas de aprendizagem. Logo após foram elaborados diversos algoritmos, sendo os mais conhecidos: C4.5, CART (Classification and Regression Trees), CHAID (ChiSquare Automatic Interaction Detection), entre outros. Os algoritmos que constróem árvores de decisão buscam encontrar aqueles atributos e valores que provêem máxima segregação dos registros de dados, com respeito ao atributo que se quer classificar, a cada nível da árvore.

Após a construção de uma árvore de decisão é importante avaliá-la. Esta avaliação é realizada através da utilização de dados que não tenham sido usados no treinamento. Esta estratégia permite estimar como a árvore generaliza os dados e se adapta a novas situações, podendo, também, se estimar a proporção de erros e acertos ocorridos na construção da árvore (Brazdil, 2002).

A Fig. 2, mostra um exemplo de uma árvore de decisão, envolvendo um problema de condições meteorológicas, analisando-se o caso sair ou não sair de casa, de acordo com o tempo.

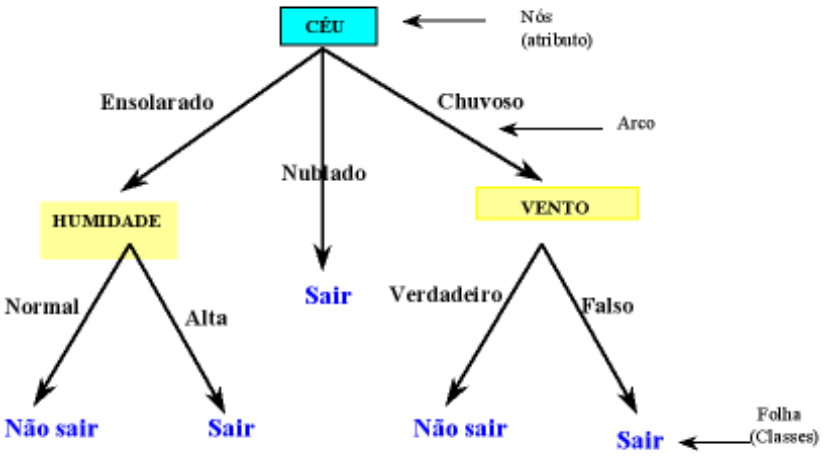


Fig. 2. Árvore de decisão para condições meteorológicas.
Fonte: Cechin & Osório (2002).

Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são modelos computacionais inspirados no sistema nervoso biológico, cujo funcionamento é semelhante a alguns procedimentos humanos, ou seja, aprendem pela experiência, generalizam exemplos através de outros e abstraem características (Wasserman, citado por Venturieri & Santos, 1998). De maneira geral, pode-se definir uma RNA como um sistema constituído por elementos de processamento interconectados, chamados de neurônios, os quais estão dispostos em camadas, uma camada de entrada, uma ou mais intermediárias e uma de saída. A Fig. 3 apresenta um modelo de rede neural com uma camada intermediária. São responsáveis pela não-linearidade da rede, através do processo interno de certas funções matemáticas. Essas RNAs possuem alguma forma de regra de aprendizagem que é responsável pela modificação dos pesos sinápticos a cada ciclo de iteração, de acordo com os exemplos que lhe são apresentados. Assim, pode-se dizer que as RNAs aprendem a partir de exemplos (Galvão & Valença, 1999).

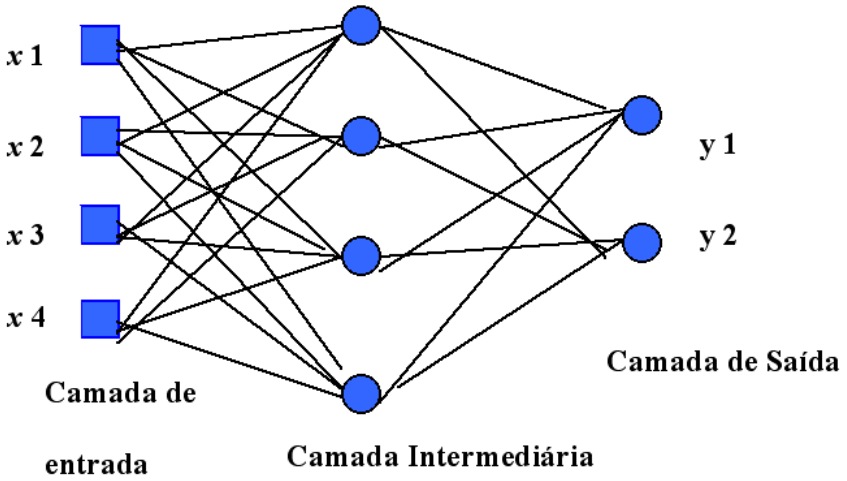


Fig. 3. Ilustração de uma rede multicamada com uma camada intermediária.

Segundo Galvão & Valença (1999), dentre as regras de aprendizado, o algoritmo de aprendizado Perceptron sugerido por Widrow e Hoff, que é também conhecido como regra delta, constitui-se num dos mais simples. Nesta técnica de treinamento fornecem-se, além dos dados de entrada, as respostas desejadas (treinamento supervisionado), de tal forma que o processamento ocorre de maneira bastante simples, ou seja, inicialmente atribui-se aos pesos valores aleatórios e, com eles, calcula-se a resposta da rede e então se compara os valores calculados com aqueles desejados. Caso o erro não seja aceitável, faz-se o ajuste dos pesos proporcionalmente ao erro. Neste caso, uma vez que duas classes se misturam e não possam ser separadas por uma linha reta, os exemplos não são linearmente separáveis (Galvão & Valença, 1999). Braga et al. (2000) afirmam que as redes de uma só camada resolvem apenas problemas linearmente separáveis. A solução de problemas não linearmente separáveis passa pelo uso de redes com uma ou mais camadas intermediárias, denominadas multicamadas.

Redes multicamadas apresentam um poder computacional muito maior do que aquele apresentado pelas redes sem camadas intermediárias. Ao contrário destas redes, as multicamadas podem tratar com dados que não são linearmente separáveis. Teoricamente, redes com duas camadas intermediárias podem implementar qualquer função, seja ela linearmente separável ou não.

Os trabalhos de Minsky & Paper, citados por Azevedo et al. (2000), provaram que redes diretas necessitam de camadas intermediárias para solucionar problemas não linearmente separáveis. Posteriormente, ficou provado que tudo que uma rede pode aprender com n camadas intermediárias pode ser aprendido por uma rede de única camada intermediária. O número de neurônios nas camadas de entrada e saída, normalmente, é função do problema em questão. O problema reside, então, no número de neurônios na camada intermediária: se for um número grande, a rede pode se especializar e perder a capacidade de generalização; se for um número pequeno, a rede pode não aprender.

Os algoritmos Perceptrons de Múltiplas Camadas (MLP) têm sido aplicados com sucesso para resolver diversos problemas difíceis, através do seu treinamento de forma supervisionada com um algoritmo muito popular conhecido como algoritmo de retropropagação de erro (*error*

back-propagation). Este algoritmo é baseado na regra de aprendizagem por correção de erro (Haykin, 2001). Segundo Sclünzen et al., citado por Venturieri & Santos (1998), é necessário um agente externo que direcione o sistema. Esse direcionamento é realizado através do algoritmo de treinamento denominado *backpropagation* (retropropagação de erro), que consiste na apresentação de um conjunto de amostras, o qual é comparado com uma imagem de saída desejada.

O algoritmo *backpropagation* (retropropagação do erro) é baseado na regra delta proposta por Widrow e Hoff, sendo por isto também chamada de regra delta generalizada. Este algoritmo propõe uma forma de definir o erro dos nodos das camadas intermediárias, possibilitando o ajuste de seus pesos (Braga et al., 2000). Segundo Azevedo et al. (2000), à medida que a rede aprende, o valor do erro converge para um valor estável, normalmente irredutível. O processo de aprendizagem prossegue até que algum critério seja estabelecido, como por exemplo, um valor mínimo de erro global, ou uma diferença sucessiva mínima entre erros calculados para cada iteração.

Allard & Fuchs, citados por Câmara & Medeiros (1998), comentaram que sistemas baseados em regras do conhecimento e redes neurais têm sido largamente utilizados na solução de problema complexos, onde os algoritmos e técnicas tradicionais são inadequados.

Regras de Indução

Indução, em oposição a dedução, é o processo de se obter uma hipótese a partir dos dados e fatos já existentes. Indução pode ser explicada como sendo a conclusão de informações provenientes de dados e aprendizagem indutiva é o processo de construção de um modelo onde o ambiente, isto é, o banco de dados, é analisado através de uma visão para identificar padrões. Objetos semelhantes são agrupados em classes e regras formuladas por meio das quais é possível predizer a classe de objetos não vistos. Este processo de classificação identifica grupos nos quais cada qual tem um padrão único de valores que constitue a descrição da classe. A natureza do ambiente é dinâmica e por isso o modelo deve ser adaptável, isto é, deve ser capaz de aprender (Unesp, 2002).

A aprendizagem indutiva, onde o sistema deduz o conhecimento pela observação do seu ambiente tem duas estratégias principais:

- aprendizagem supervisionada - é a aprendizagem por meio de exemplos onde um professor ajuda o sistema a construir um modelo definindo classes e exemplos abastecedores de cada classe. O sistema tem que achar uma descrição de cada classe, isto é, as propriedades comuns nos exemplos. Uma vez que foi formulada a descrição e a forma da classe, uma regra de classificação pode ser usada para predizer a classe de objetos previamente não vistos; e
- aprendizagem não supervisionada - é a aprendizagem por meio de observações e descobertas. O sistema de dados é provido com objetos mas nenhuma classe é definida. Assim tem-se que observar os exemplos e reconhecer padrões (isto é, descrição de classe) por si só. Este sistema resulta em um conjunto de descrições de classe, um para cada classe descoberta no ambiente.

Logo, indução é a extração de padrões. A qualidade do modelo produzido por métodos de aprendizagem indutiva é tal que o modelo poderia ser usado para predizer o resultado de situações futuras. Em outras palavras, poderia ser usado não somente para estados encontrados mas também para estados não vistos que pudessem acontecer.

Segundo Goulart Júnior et al. (2002), regra indutiva é o processo de olhar uma série de dados e, a partir dela, gerar padrões. Pelo fato de explorar automaticamente a série de dados, como mostra a Fig. 4, o sistema indutivo cria hipóteses que conduzem a padrões.

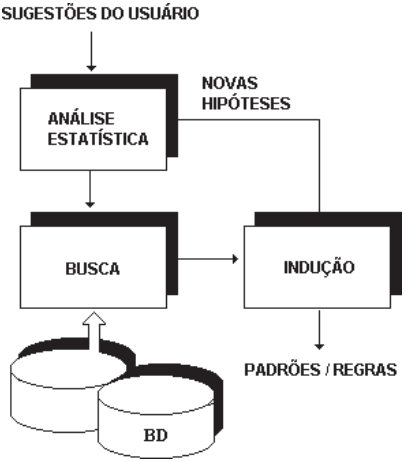


Fig. 4. Esquema da exploração de regras de indução.
Fonte: Goulart Júnior et al. (2002).

O processo é em sua essência semelhante àquilo que um analista humano faria em uma análise exploratória. A indução de regras pode descobrir regras muito gerais, as quais lidam tanto com dados numéricos quanto não numéricos.

Mineração Visual de Dados

Pode-se pensar a visualização de dados como técnicas que mapeiam volumes de dados multidimensionais para a tela bidimensional de um computador. Visualização é uma ferramenta importante para mineração de dados porque seres humanos são muito bons em processar informação visual e muito ruins em processar informação numérica e/ou tabular. Mineração visual de dados engloba técnicas que combinam visualização e exploração interativa de dados. Estas técnicas descrevem conjuntos complexos de dados através de gráficos envolvendo múltiplas variáveis simultaneamente. Elas normalmente permitem a exploração inteligente destes dados através de controle dos gráficos e seleção interativa da informação a ser analisada. Neste caso, o perito necessita interagir diretamente com a ferramenta para que possa extrair informações úteis dos dados explorados (Mendonça Neto et al. 2000).

Técnicas de Aglomeração (Clusterização)

Muitas vezes a clusterização é uma das primeiras etapas dentro de um processo de mineração de dados, já que identifica grupos de registros correlatos, que serão usados como ponto de partida para futuras explorações (Ikematu, 2002).

A descoberta por clusterização procura separar automaticamente elementos em classes que serão identificadas durante o processo (não há classes pré-definidas). A clusterização é diferente da classificação, pois a primeira visa criar as classes através da organização dos elementos, enquanto a segunda procura alocar elementos em classes já pré-definidas. A clusterização auxilia o processo de descoberta de conhecimento, facilitando a identificação de padrões nas classes. Geralmente, a técnica de clusterização vem associada com alguma técnica de descrição de conceitos, para identificar os atributos de cada classe. Esta posterior identificação das classes através de suas características é chamada de *cluster analysis* e gera uma nova abordagem de descoberta (Willet, 1988).

Aglomerações podem ser usadas para: produzir uma visão de alto nível do que acontece com os dados; automaticamente identificar pontos fora da curva; ou classificar ou predizer valores de novos registros usando as características dos aglomerados mais próximos a este registro (Mendonça Neto et al., 2000).

O Estado de São Paulo e o Agronegócio Café e Cana-de-Açúcar

O Estado de São Paulo possui mais de 190 mil quilômetros quadrados plantados, entre culturas, pastagens e florestas destinadas ao aproveitamento econômico. Líder em agronegócios, o Estado é responsável por um terço do PIB agroindustrial do Brasil. Contribuindo para isso, destacam-se as culturas de café e cana-de-açúcar (São Paulo, 2002).

O Estado de São Paulo tem destacada importância para o agronegócio do café, principalmente, por sua infra-estrutura portuária, fundamental para o escoamento da produção de outras regiões produtoras e por possuir o maior parque industrial de café do país. O cultivo do produto encontra-se hoje concentrado nas regiões de Campinas, Franca e Marília, correspondendo a 57,8% do total produzido no estado (Embrapa, 2002). O Estado de São Paulo é considerado o quarto maior produtor mundial e o terceiro nacional de café produzindo 3,5 milhões de sacas de 60 quilos (Caser et al., 2002).

O Estado é também o segundo maior produtor mundial de cana-de-açúcar (São Paulo, 2002). A distribuição da área e produção com cana-de-açúcar em São Paulo mostra que ela é cultivada em todo o Estado, preponderando na DIRA (Divisão Regional Agrícola) de Campinas e de Ribeirão Preto, que juntas são responsáveis por 60% da área total e da produção total (Veiga Filho, 2002).

Apesar desses números, as perdas na agricultura são muito grandes, devido à ocorrência de sinistros na plantação por eventos climáticos. Tanto o café, como a cana-de-açúcar, que são plantas perenes e semi-perenes, respectivamente, sofrem, portanto as influências do clima em todo o curso

do ano, ao contrário das culturas anuais que sofrem as influências em determinados períodos. Os problemas ligados às adversidades climáticas são os mais variados e se relacionam a diferentes elementos como: geadas; vento frio persistente; veranicos freqüentes; deficiências hídricas prolongadas; má distribuição do regime pluvial ao longo do ano, etc. Segundo Rossetti (2002), as culturas de verão apresentam 60% de perdas por seca e 32% por chuvas fortes. Nas culturas de inverno, as perdas por seca alcançam até 30%, por chuvas fortes 32% e por geada 30%. Tanto a produção de café e a de cana-de-açúcar no Estado de São Paulo sofrem alternâncias motivadas por eventos climáticos adversos e em especial, as geadas e as secas, que reduzem drasticamente as produções.

Neste sentido, justifica-se, uma proposta para prever a ocorrência de geada para a cultura do café e deficiência hídrica no solo para a cana-de-açúcar e para o café, podendo, auxiliar os produtores na tomada de decisões que permita eliminar e/ou reduzir os prejuízos causados pelos fatores climáticos. Para isso, utilizar-se-á da técnica de mineração de dados para extrair conhecimento do grande volume de dados climáticos, possibilitando a previsão de geada e de deficiência hídrica.

Geada x Café

O fenômeno das geadas ocorre com certa freqüência nas principais regiões cafeeiras brasileiras, causando prejuízos às lavouras.

De acordo com Tubelis & Nascimento (1992), geada é a ocorrência de temperatura do ar abaixo de 0°C, podendo ou não dar origem à formação de gelo sobre as superfícies expostas. Segundo Caramori et al. (2001), sob o aspecto agronômico, considera-se geada qualquer abaixamento de temperatura que acarrete na planta efeitos prejudiciais ao seu crescimento ou desenvolvimento. Portanto, deve-se destacar que nem sempre a presença de gelo sobre as superfícies expostas significa que ocorreu geada do ponto de vista agronômico, pois a temperatura que provoca danos às plantas pode não ter sido atingida.

Diversos estudos mostram que temperaturas entre -3 °C e -4 °C são letais para o tecido foliar do cafeeiro (Ferraz, 1968). Constatou-se também que quanto maior for a queda de temperatura e quanto mais tempo a planta

permanecer exposta a temperaturas inferiores à crítica, mais graves e mais extensos são os danos.

Segundo Assad & Pinto (2001), a condição climática considerada apta para a cafeicultura no Estado de São Paulo é de temperatura média anual entre 18°C e 22°C e não ocorrência de temperaturas máximas superiores a 34°C nos meses de outubro e novembro.

O cafeeiro arábica, quando em áreas com temperaturas médias anuais elevadas, acima de 23°C e para 34°C nos meses de outubro e novembro, devido ser o período de florescimento (toda a cafeicultura comercial do Brasil apresenta o florescimento na primavera, a frutificação no verão, a maturação no outono e a colheita no inverno), freqüentemente apresenta problemas de abortamento das flores e formação de estrelinhas, ocorrendo a frutificação e a maturação demasiadamente precoces, podendo esse fato, acarretar perdas na qualidade final do produto, pois as fases da colheita e secagem podem coincidir com períodos quentes e chuvosos.

Por outro lado, temperaturas médias anuais baixas (inferiores a 18°C) provocam o período de dormência das gemas florais retardado e o desenvolvimento dos frutos mais lentos, o que faz com que o período de maturação seja coincidente com novo florescimento, dificultando a colheita, ou seja, provocam aumento no período de frutificação, podendo ocorrer a maturação, se sobrepondo ao florescimento no ano seguinte, prejudicando a vegetação e produção final (Camargo & Pereira, citados por Caramori et al., 2001).

As taxas de sinistralidade agrícola devido às geadas são muito grandes. A produção de café pelos Estados do Sudeste brasileiro sofrem alternâncias motivadas por eventos climáticos adversos (geadas e secas), que reduzem drasticamente as produções. No Estado de São Paulo, a produção de café ocupa aproximadamente 240 mil hectares, contra 710 mil hectares em 1975, ano de inflexão da produção de café, decorrente da pior geada da história da cafeicultura nacional. A maior parte dos cafezais do Estado de São Paulo foram dizimados, iniciando o processo de perda relativa da participação da produção desse estado no total produzido no Brasil (Alfonsi, 2002).

Segundo Rebello & Neves (1987), o fenômeno de geadas nas Regiões Sul e Sudeste do Brasil é um assunto de muito interesse para meteorologistas ligados ao setor agropecuário, que procuram disseminar avisos alertando os agricultores sobre a aproximação de massas polares, causadoras de temperaturas mínimas extremas e ocorrência de geadas.

No Estado do Paraná, o IAPAR criou o sistema alerta geadas para elaborar previsões detalhadas, com o intuito de amenizar o problema que as geadas causam ao café. O potencial de retorno deste Sistema é de 50 a 60 milhões por ano em economia de novos plantios. A margem de acerto das previsões tem sido de 100%, dando total segurança ao produtor. No inverno de 2000, todas as geadas ocorridas foram previstas, possibilitando que muitos agricultores evitassem prejuízos em viveiros e plantios recentes (Caramori et al., 2001).

Tarifa et al. (1977) descreveram a situação dos danos causados pela geada de 1975 à cafeicultura no Estado de São Paulo. O principal resultado descrito é o grau de resfriamento na superfície. No Estado de São Paulo a pressão mínima foi 1028 hPa e a máxima foi 1030 hPa.

Fortune (1981) estudou o episódio de geada que ocorreu em 1979, buscando sinais no Oceano Pacífico, que pudessem dar indicações para uma previsão de geadas. Os resultados mostraram que uma onda longa do Pacífico amplifica-se, fornecendo um sinal da provável ocorrência de geadas no sul do Brasil com 3 a 4 dias de antecedência.

Fortune et al. (1982) analisaram os aspectos mais importantes encontrados para duas geadas, a de 1979 e a de 1981, e destacam importantes precursores: uma configuração de ondas longas, observada em altos níveis deslocando-se lentamente no Pacífico, amplificou-se entre 4 e 5 dias antes das geadas no Brasil.

Molion et al. (1981) discutiram as evidências sobre a ocorrência do fenômeno de geada, que podem ser detectadas com até 3 dias de antecedência. Isso seria possível analisando a intensidade da massa de ar polar que estivesse penetrando no sul da América do Sul. Um modelo estatístico seria utilizado para gerar a distribuição da temperatura.

Satyamurty et al. (1990) fizeram um estudo de caso, em que foi observada a ocorrência de duas ondas de ar frio que afetaram a região Sul, São

Paulo, Mato Grosso do Sul, sul de Minas Gerais e sul de Goiás. As massas de ar frio, com pressões centrais da ordem de 1030 hPa, levaram cerca de 72 horas para se deslocar da costa oeste do continente até o litoral da região Sudeste do Brasil.

Algarve & Cavalcanti (1994) mostraram padrões típicos para ocorrência de geadas no sul do Brasil, através de uma análise com dados de 10 anos.

Deficiência Hídrica x Café x Cana-de-Açúcar

Segundo Tubelis (1992), o balanço hídrico é um método de se calcular a disponibilidade de água no solo. Indica a contabilização da água do solo, representando o balanço entre o que entrou e o que saiu de água. Os valores de balanço hídrico positivos indicam excedentes hídricos e os negativos, deficiência hídrica ou falta de água. Ocorre excedente de água sempre que a precipitação for superior à quantidade necessária para alimentar a evapotranspiração potencial[1] e completar o armazenamento de água no solo. A deficiência aparece sempre que o solo não conseguir suplementar a precipitação no atendimento da evapotranspiração potencial. O balanço hídrico contabiliza a precipitação perante a evapotranspiração potencial, levando em consideração a capacidade de armazenamento de água no solo. Esta é a máxima quantidade de água, utilizável pelas plantas, que pode ser armazenada na sua zona radicular. O valor de armazenamento para as culturas de cana-de-açúcar e café já foi determinado, estipulado em 125mm, pelo método de Thornthwaite-Mather.

O cálculo do balanço hídrico pode ser feito pelo método Thornthwaite-Mather. É preciso conhecer os valores mensais e anual da precipitação e da evapotranspiração potencial, podendo ser representado num gráfico, indicando excedente ou deficiência hídrica no solo.

---

[1] Thornthwaite e Wilm introduziram o termo Evapotranspiração potencial (Etp), em 1944, que representa a perda natural de água do solo vegetado para a atmosfera através da ação conjunta da evaporação e da transpiração. A Etp é processo oposto à precipitação, representa a água que retorna forçosamente para a atmosfera, em estado gasoso, e depende da energia solar disponível na superfície do terreno para vaporizá-la (Camargo & Camargo, 2000).

Ao ponderar sobre as melhores condições hídricas para o cultivo da cana, tem-se que levar em conta que não é o total de precipitação anual o mais importante, mas sim a sua distribuição ao longo do ano ou, ainda melhor, a disponibilidade de água no solo à disposição da planta, durante o ciclo vegetativo. Para se estimar esta disponibilidade, pode-se utilizar o balanço hídrico, segundo Thornthwaite & Mather (1955), baseado num sistema de contabilização da água no solo, que nos indica os déficits e excedentes hídricos no curso do ano (Minas Gerais, 1980).

Um regime hídrico em que não ocorrem déficits hídricos é contra indicado para a cultura da cana-de-açúcar, por ser necessária a existência de um período seco, que favoreça a maturação em detrimento do crescimento. Por outro lado, quando a deficiência hídrica anual ultrapassa determinados limites, o desenvolvimento da planta poderá ficar seriamente reduzido. A deficiência hídrica anual menor que 200 mm é considerada ótima para o cultivo da cana por apresentar condições hídricas satisfatórias. Deficiência hídrica anual entre 200 a 400 mm indica deficiências hídricas sazonais pronunciadas, tornando-se recomendável o emprego de irrigação suplementar. Deficiência hídrica anual igual a 400 mm corresponde ao limite acima do qual torna-se imprescindível a irrigação. E excedente hídrico anual igual a 800 mm, apresenta o limite acima do qual ocorre excesso de umidade na estação vegetativa (Minas Gerais, 1980).

Zink, citado em Minas Gerais (1980), considera que a precipitação de 1.200 mm anuais é suficiente para o bom desenvolvimento da cana, necessidade esta de mais água nos primeiros meses de crescimento, concentrados na primavera e verão, e posteriormente de um período seco para a maturação, com inverno seco e/ou frio bem característico, sem geadas freqüentes.

Em relação à cultura do café, Carvajal (1972) constata que ao se avaliar o ótimo de precipitação para o cafeeiro, deve-se considerar algumas variáveis importantes: precipitação anual média, distribuição da precipitação durante o ano (número de meses secos), desvio da precipitação anual em relação a média (anos secos e úmidos) e condições do solo (características físicas). Coste (1968) também menciona a importância de se considerar o total das precipitações durante o ano e sua repartição mensal, quinzenal, decendial ou pentadial. As exigências das chuvas são da ordem de 1500 a 2000 milímetros anuais (Alfonsi, 2002).

A cultura do café apresenta quatro fases fenológicas distintas ao longo do ano, quais sejam: granação e abotoação, maturação e abotoamento, dormência e floração (Camargo, 1987). O mesmo autor observou, para as condições de Campinas, que a deficiência hídrica mostra-se bastante crítica para o cafeeiro nas fases de chumbinho (outubro a dezembro), granação (janeiro a março) e maturação/abotoamento (abril a junho). No período de julho a setembro, pode até ser benéfica, já que esta fase corresponde a dormência da planta. Ortolani (1991) comenta que a curva de demanda hídrica do cafeeiro, normalmente, é sazonal, com menores valores para o período de dormência (junho a setembro no Estado de São Paulo) e com elevação da evapotranspiração desde a antese, vegetação e granação.

O cafeeiro, para vegetar e frutificar normalmente, necessita encontrar umidade facilmente disponível no solo durante todo o período de vegetação e frutificação que vai de setembro a maio (Camargo, 1977). Para definir satisfatoriamente as disponibilidades hídricas climáticas, Assad & Pinto (2001) estabeleceram os seguintes limites para definir as áreas propícias para o cultivo do café no Estado de São Paulo. Déficit hídrico menor que 100 mm apresenta condições apta para o cultivo do café, déficit hídrico entre 100 e 150 mm indicam condições marginais para a cafeicultura, recomendando-se irrigação e, déficit hídrico maior que 150 mm representa condições inaptas para o cultivo do café.

Resultados de experimento realizado no Estado de São Paulo evidenciam a importância da precipitação pluviométrica e temperatura nas fases de abotoamento, florescimento, máxima vegetação e granação (Weill, 1990).

Previsão Meteorológica

Previsão meteorológica é uma estimativa do comportamento médio da atmosfera com algum tempo de antecedência. Atualmente, para se fazer esse tipo de previsão, os meteorologistas utilizam dois métodos, o estatístico e o dinâmico. O método estatístico, com equações matemáticas e conceitos de estatística, através de uma correlação entre duas ou mais variáveis, estima o prognóstico de uma delas. Já o método dinâmico, com equações matemáticas e conceitos físicos, através de equações

físicas, simula os movimentos atmosféricos para prever os acontecimentos futuros (Inpe, 2002a).

O comportamento da atmosfera é governado por leis físicas que podem ser expressas por equações matemáticas. Tais equações, entretanto, são muito complexas e não possuem soluções exatas para os valores futuros. Por esta razão, técnicas de modelagem numérica são utilizadas, dando origem aos "Modelos Numéricos de Previsão". Os modelos de previsão numérica meteorológica podem ser globais ou de área limitada (regionais). Quando as equações que governam a atmosfera são resolvidas sobre todo o globo, temos os modelos globais. Estes fazem previsões até 10 dias à frente. O modelo global tem um índice de acerto de até 60% para previsão até 7 dias. Para a previsão de 1 ou 2 dias, este acerto está acima de 90%. A atmosfera é previsível até um certo limite, a partir daí, não se consegue mais fazer previsão desse nível. Os modelos globais consideram todos os fenômenos atmosféricos que ocorrem no globo terrestre sem, entretanto, ater-se às peculiaridades de cada região. Vários centros de previsão do tempo no mundo executam rotineiramente modelos globais (também denominados de modelos de previsão de médio prazo), dentre eles podemos citar: NCEP (National Centers for Environmental Prediction), ECMWF (European Centre for Medium Wheather Forecasting), JMA (Japan Meteorological Agency) e CPTEC (Centro de Previsão de Tempo e Estudos Climáticos). Os modelos de área limitada resolvem as equações sobre uma área mais localizada, por exemplo, a América do Sul ou parte dela. Esses modelos podem fornecer uma previsão mais detalhada, mas eles normalmente fazem previsões de algumas horas até 2 a 3 dias à frente. Como exemplo pode-se destacar o modelo regional ETA, utilizado pelo CPTEC. No caso do ETA, as previsões se estendem até 48h e cobrem a maior parte da América do Sul (Ipmet, 2002).

O uso de modelos numéricos de previsão de tempo e clima, permite a elaboração de previsões com maior precisão, qualidade e antecedência. A previsão numérica depende muito das condições do plano de tempo (o campo que dá entrada para o modelo), porque se não houver precisão, a previsão é ruim também (Porto, 2002).

Desde 1995 o CPTEC/INPE é o único Centro Meteorológico na América Latina que operacionalmente produz previsões numéricas de tempo e

clima para o Brasil e para o globo. Essas previsões são de grande importância para a tomada de decisões em relação à agricultura. O CPTEC vem experimentando a previsão de longo prazo, de um a três meses, empregando o seu modelo dinâmico, com resultados promissores (Inpe, 2002a).

A previsibilidade numérica da atmosfera de forma determinística, baseada em modelos dinâmicos, tem sido amplamente discutida desde que foi observada que a solução de sistemas de equações semelhantes às que governam os movimentos atmosféricos apresentam dependência sensível em relação às condições iniciais fornecidas no início da integração. Notou-se que partindo de condições ligeiramente perturbadas, após algum tempo de integração, as soluções podem ser completamente diferentes. Tais fenômenos vieram a ser chamados "caóticos" devido ao comportamento irregular que apresentavam. Sabe-se que os modelos numéricos não conseguem reproduzir a enorme diversidade de fenômenos que influenciam a evolução das condições atmosféricas, o que seria suficiente para limitar o prazo de previsão, entretanto, mesmo que os modelos fossem perfeitos, os erros inerentes às observações, utilizadas no momento de geração da condição inicial, poderiam levar a uma previsão que não seria verificada depois de alguns dias (Inpe, 2002b).

Metodologia

O trabalho já está sendo desenvolvido nas dependências da Embrapa Informática Agropecuária, juntamente com a equipe do projeto "Desenvolvimento e Evolução de um Sistema de Monitoramento Agroclimatológico para o Estado de São Paulo" (Embrapa, 2001).

Os dados em estudo compreendem dados climáticos históricos do Estado de São Paulo, como temperatura máxima, temperatura mínima, precipitação diária e balanço hídrico, para um longo período de tempo, disponibilizados pelo Instituto Agronômico de Campinas (IAC). Contém dados de 136 estações climatológicas, coletados num total de 12 anos, no período de 1991 a 2002.

A extração de conhecimento é o principal objetivo da mineração de dados, permitindo que sejam descobertas informações de grande valor e que

não tenham relações óbvias a serem identificadas. Através do uso de algoritmos específicos, procura-se descobrir padrões e tendências nos dados e inferir regras para descrevê-los.

O processo de mineração de dados é um conjunto de atividades contínuas, descritas por etapas, que compartilham o conhecimento descoberto a partir de bases de dados. No conjunto dos dados presente, as seguintes etapas serão aplicadas:

a) pré-processamento: desenvolvimento do entendimento do domínio da aplicação, avaliação do hardware e software disponíveis, seleção, limpeza e transformação dos dados;

b) mineração de dados: escolha da tarefa e da técnica a serem utilizadas, identificação da ferramenta que satisfaça a essas condições e sua aplicação aos dados nesta ferramenta;

c) pós-processamento: interpretação dos resultados enumerados e incorporação do conhecimento adquirido.

Ao se trabalhar com mineração de dados, percebe-se que se utiliza uma grande variedade de técnicas. Porém, ao selecionar um algoritmo devem ser considerados vários aspectos decisivos para um bom desempenho da ferramenta de descoberta de conhecimento, pois algumas técnicas são mais adequadas para trabalhar com determinados tipos e volumes de dados do que outras.

Com a utilização de técnicas de mineração de dados procurar-se-á identificar padrões entre os parâmetros analisados, como por exemplo, identificar relações existentes entre os atributos presentes (temperatura máxima, temperatura mínima) e as ocorrências de chuvas. Quando determinados padrões de comportamento começam a se repetir com freqüência, as ferramentas de "data mining" indicam a presença de oportunidades e "*insights*" em relação àquele determinado atributo, descobrindo-se padrões e tendências nos dados e gerando-se regras.

A partir de análises e avaliações das regras geradas poderão sugerir novos conhecimentos através de evidências não detectadas anteriormente, indicando-se a necessidade de estudos aprofundados acerca dessas relações.

No presente estudo de pesquisa serão utilizados dados climatológicos históricos do Estado de São Paulo, e com a aplicação da metodologia apresentada é provável que sejam realizadas algumas previsões climáticas. Os dados estão disponíveis para um longo período de tempo, o que aumenta as chances de se descobrir padrões que podem explicar e ajudar a prever ocorrências de geada e deficiência hídrica beneficiando a cultura do café e cana-de-açúcar do Estado de São Paulo.

Os conhecimentos adquiridos com a utilização de técnicas de mineração de dados deverão ser verificados junto a especialistas humanos na área climatológica, ao longo de todo o desenvolvimento do projeto. Ao longo da execução das técnicas de mineração de dados, com a identificação de evidências que podem significar eventuais novos conhecimentos, também há a necessidade de acompanhamento de especialistas para avaliação da pertinência das eventuais relações encontradas.

A validação do modelo gerado deve ser realizada através da consulta a outros especialistas que não tenham participado do seu desenvolvimento. Devem ser apresentados aos especialistas diversos cenários e comparadas as suas previsões com as obtidas com o modelo gerado.

"Funções de interessantíssimo" também serão utilizadas para quantificar quanto uma regra poderá ser interessante para um perito. Especialistas podem então olhar aquelas regras consideradas mais "interessantes" pela "função de interessantíssimo" e tentar derivar conhecimento a partir delas (Mendonça Neto, 2001). O grau de interesse é uma maneira de selecionar regras tentando capturar o quanto o conhecimento é interessante (ou inesperado) segundo critérios de utilidade e potencialidade de uso. Num processo de análise quantitativa de regras, as medidas objetivas do grau de interesse podem ser usadas como uma espécie de filtro para selecionar regras potencialmente interessantes e, posteriormente, submeter essas regras a uma avaliação subjetiva, determinando assim qual o conhecimento é realmente interessante (Gomes, 2002).

O conhecimento descoberto e validado neste trabalho estarão incluídos no esquema de disseminação e transferência de informação do projeto.

Resultados Esperados

Espera-se no final desse projeto, partindo da mineração de dados, identificar novos conhecimentos, entre os parâmetros climáticos (temperatura máxima, temperatura mínima, precipitação diária, entre outros), permitindo a previsão de geada para a cultura do café e previsão de déficit hídrico para as culturas do café e da cana-de-açúcar, visando a prevenção contra déficit hídrico e geada para o Estado de São Paulo.

Referências Bibliográficas

ALFONSI, R. R. Histórico climatológico da cafeicultura brasileira. In: SAMAPIO ASSESSORIA DE COMUNICAÇÃO. **Coffee break – o portal do agronegócio café.** Disponível em: <http://www.coffeebreak.com.br/ocafezal.asp?SE=8&ID=67>. Acesso em: maio, 2002.

ALGARVE, V. R.; CAVALCANTI, I.F. A. Características da circulação atmosférica associadas à ocorrência de geadas no sul do Brasil. In: CONGRESSO BRASILEIRO DE METEOROLOGIA, 8., 1994, Belo Horizonte. **Anais.** Rio de Janeiro: Sociedade Brasileira de Meteorologia, 1994. v. 2

ASSAD, E. D.; PINTO, H. S. **Zoneamento climático do café para os estados de São Paulo, Paraná, Minas Gerais, Goiás e Sudoeste da Bahia.** Brasília, DF: Ministério da Agricultura e Abastecimento - Coordenação Nacional do Zoneamento Agrícola: Embrapa: Funcafé, 2001.

AZEVEDO, F. M.; BRASIL, L. M.; OLIVEIRA, R. C. L. **Redes neurais com aplicações em controle e em sistemas especialistas.** Florianópolis: Bookstore, 2000. 401 p.

BARANAUSKAS, J. A.; MONARD, M. C. **Reviewing some machine learning concepts and methods.** São Carlos, SP: ICMC-USP, 2000. (Relatório Técnico).

BRAGA, A. P.; LUDERMIR, T. B.; CARVALHO, A. C. P. L. F. **Redes neurais artificiais:** teoria e aplicações. Rio de Janeiro: LTC, 2000. 262 p.

BRAZDIL, P. B. **Construção de modelos de decisão a partir de dados.** Disponível em: <http://www.niaad.liacc.up.pt/~pbrazdil/Ensino/ML/ModDecis.html>. Acesso em: maio, 2002.

CÂMARA,G.; MEDEIROS, J. S. Tendências de evolução do geoprocessamento. In: ASSAD, E. D., SANO, E. E. **Sistema de informações geográficas:** aplicações na agricultura. 2. ed. rev. ampl. Brasília, DF: Embrapa–SPI: Embrapa–CPAC, 1998. p. 411-424.

CAMARGO, A. P. Balanço hídrico, florescimento e necessidade de água para o cafeeiro. In: SIMPÓSIO SOBRE O MANEJO DA ÁGUA NA AGRICULTURA, 1987, Campinas. **[Anais...].** Campinas: Fundação Cargill, 1987. 226 p. (Fundação Cargill, 127).

CAMARGO, A. P. Zoneamento de aptidão climática para a cafeicultura de arábica e robusta no Brasil. In: IBGE. **Recursos, meio ambiente e poluição.** Rio de Janeiro, 1977. p. 68-76.

CAMARGO, A. P.; CAMARGO, M. B. P. Uma revisão analítica da evapotranspiração potencial. **Bragantia**, Campinas, v. 59, n. 2, p. 125-137, 2000.

CARAMORI, P. H.; CAVIGLIONE, J. H.; WREGE, M. S.; GONÇALVES, S. L.; ANDROCIOLI FILHO, A.; SERA, T.; CHAVES, J. C.; LEAL, A. C.; MORAIS, H.; KOGUISHI, M .S. **Zoneamento de riscos climáticos para a cultura do café (*Coffea arabica* L.) no Paraná.** Londrina: IAPAR, 2001.

CARVAJAL, J. F. **Cafeto – cultivo y fertilización.** Berna: Instituto Internacional de La Potasa, 1972. 141 p.

CASER, D. V.; CAMARGO, A. M. M P.; FRANCISCO, V. L. F. S.; GHOBRIL, C. N. **Previsão de safra:** previsões e estimativas das safras agrícolas do estado de São Paulo, fevereiro de 2002. Disponível em: <http://www.iea.sp.gov.br/ps-0202-3l-t.htm>. Acesso em: jun. 2002.

CECHIN, A.; OSÓRIO, F. **KDD - o conhecimento: representação do conhecimento.** Disponível em: <http://www.inf.unisinos.br/~cechine~osorio>. Acesso em: jul. 2002.

COSTE, R. **Le caféier.** París: Techniques Agricoles et Productions Tropicales, 1968. 310 p.

EMBRAPA. **Desenvolvimento e evolução de um sistema de monitoramento agroclimático para o estado de São Paulo.** Campinas: Embrapa Informática Agropecuária, 2001. 15 p.

EMBRAPA. Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café. **Economia cafeeira.** Disponível em: <http://www.embrapa.br/cafe/consorcio/home_4.htm>. Acesso em: jun. 2002.

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

FERRAZ, E. C. **Estudo sobre o momento em que a geada danifica as folhas do cafeeiro.** 1968. 59 p. Tese (Doutorado) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba.

FORTUNE, M. A. **Cenário sinótico da invasão de ar frio na geada de maio de 1979 e mapeamento de geadas para prever áreas de risco.** São José dos Campos: Inpe, 1981. (Inpe-2166-RPE/383).

FORTUNE, M. A.; KOUSKY, V. E.; FERREIRA, N. J. **Duas geadas críticas no Brasil:** percursores no Oceano Pacífico e evolução na América do Sul. São José dos Campos: Inpe, 1982. (Inpe-E-2587-PRE/235).

GALVÃO, C. O.; VALENÇA, M. J. S. **Sistemas inteligentes:** aplicações a recursos hídricos e sistemas ambientais. Porto Alegre: Ed. Universidade: UFRGS: ABRH, 1999. 246 p.

GARCIA, S. C.; ALVARES, L. O. **O uso de árvores de decisão na descoberta de conhecimento na área da saúde.** Disponível em: <http://www.inf.ufrgs.br/pos/SemanaAcademica/Semana2000/SimoneGarcia/>. Acesso em: 05 jun. 2002.

GIMENES, E.; SEIXAS, J. A. **"Data mining – data warehouse" a importância da mineração de dados em tomadas de decisões.** 2000. 51 p. Monografia - Faculdade de Tecnologia de Taquaritinga - Centro Estadual de Educação Tecnológica "Paula Souza", Taquaritinga.

GOMES, A. K. **Análise do conhecimento extraído de classificadores simbólicos utilizando medidas de avaliação e de interessabilidade.** 2002. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

GOULART JÚNIOR, F. S.; FIDALGO, R. N.; SALGADO, A. C.; FONSECA, F. **Data mining.** Disponível em: <http://www.di.ufpe.br/~compint/aulas-IAS/agentes/taci1-981/DataMining.ppt>. Acesso em: maio, 2002. Disciplina – Banco de dados – Universidade Federal de Pernambuco.

HALMENSCHLAGER, C. **Utilização de agentes na descoberta de conhecimento.** Porto Alegre: UFRGS-PPGC, 2000. 55 f.

HAYKIN, S. **Redes neurais:** princípios e prática. 2. ed. Porto Alegre: Bookman, 2001. 900 p.

HOWARD, C. M.; RAYWARD-SMITH, V. J. **Streamling a meteorological database for knowledge discovery.** Los Alamitos: IEEE, 1997. 5 p.

IKEMATU, R. S. **Tutorial DBForum´98 – data mining:** ferramentas e técnicas. Disponível em: <http://www.pr.gov.br/celepar/celepar/batebyte/edicoes/1998/bb76/admin.htm>. 1998. Acesso em: maio, 2002.

INGARGIOLA, G. **Building classification models:** ID3 and C4.5. Disponível em: <http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html>. Acesso em: jun. 2002.

INPE. Centro de Previsão de Tempo e Estudos Climáticos. **Meio ambiente e ciências atmosféricas - a utilização de multimídia e da rede Internet no ensino público de nível médio.** Disponível em: <http://tucupi.cptec.inpe.br/~ensinop/aulas.html>. Acesso em: jun. 2002a.

INPE. Centro de Previsão de Tempo e Estudos Climáticos. **Previsões numéricas - o sistema de previsão de tempo global por Ensemble do CPTEC.** Disponível em: <http://www.cptec.inpe.br/prevnum/exp_ensemble.shtml>. Acesso em: jun. 2002b.

IPMET. **Previsão numérica.** Disponível em <http://www.ipmet.unesp.br/modelos/expl1.html>. Acesso em: jun. 2002.

LIU, J. N. K.; LI, B. N. L.; DILLON, T. S. An improved naïve bayesian classifier technique coupled with a novel input solution method. **IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews**, v. 31, n. 2, May, p. 249-256, 2001.

McCULLAGH, J.; BLUFF, K.; HENDTLASS, T. **Envolving expert neural networks for meteorological rainfall estimations.** Los Alamitos, IEEE, 1999. p. 585-590.

MENDONÇA NETO, M. G. de. Mineração de dados. In: ESCOLA REGIONAL DE INFORMÁTICA DA SBC REGIONAL DE SÃO PAULO, 6., 2001, São Carlos. **Minicursos:** coletânea de textos: anais. São Carlos, SP: ICMC-USP, 2001. p. 189-218.

MENDONÇA NETO, M. G. de ; NOGUEIRA, L. A.; PONTES, L. A. M.; TEIXEIRA, L. S. G.; GUIMARÃES, P. R. B. **Aplicação de técnicas de mineração visual de dados na regulação da indústria de energia: um estudo de casos.** Trabalho publicado nos Anais do 1. Congresso Brasileiro de Regulação de Serviços Públicos Concedidos, Salvador, BA, jun. 2000. Disponível em: <http://www.nuperc.unifacs.br/RT-NUPERC-2000-1p.pdf>. Acesso em: maio, 2002.

MINAS GERAIS. Secretaria de Estado da Agricultura. **Cultura da cana-de-açúcar.** Belo Horizonte, 1980. Disponível em: <http://www.agridata.mg.gov.br/mapgeo/html/cana.html>. Acesso em: maio, 2002.

MOLION, L. C. B.; FERREIRA, N. J.; MEIRA FILHO, L. G. **O uso de satélites ambientais para monitoramento de geadas.** São José dos Campos: Inpe, 1981. (INPE-2128-RPE/352).

MONARD, M. C.; BATISTA, G. E. A. P.; KAWAMOTO, S.; PUGLIESI, J. B. **Uma introdução ao aprendizado simbólico de máquina por exemplos.** Disponível em: <http://labic.icmsc.sc.usp.br/portugues/courses.htm>. Acesso em: 29 maio 2002.

MUNARI, A. C. B. **Uso de técnicas de classificação automática na análise ambiental:** um estudo de caso. 2001. 139 f. Dissertação (Mestrado) - Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas.

ORTOLANI, A. A. Relação clima-cafeicultura na região de Marília. In: ENCONTRO REGIONAL DE CAFÉ DE MARÍLIA, 1991, Marília. [Marília: s.n., 1991]. 27 p.

PORTO, M. **Modelagem matemática:** o contido e o residual – modelagem matemática na previsão do tempo e do clima. Disponível em: <http://www.comciencia.br/reportagens/modelagem/mod06.htm>. Acesso em: jun.2002.

REBELLO, E. R. G.; NEVES, E. K. Aspectos sinóticos da ocorrência de geadas severas nas regiões sul, sudeste e centro-oeste do Brasil. In: CONGRESSO BRASILEIRO DE AGROMETEOROLOGIA, 5., 1987, Belém. **Coletânea de trabalhos.** Belém: SBA, 1987. p. 313.

ROSSETTI, L. A. A seguridade e o zoneamento agrícola no Brasil – novos rumos. In: SEMINÁRIO BRASILEIRO DE ZONEAMENTO AGRÍCOLA 1., 2000, São Paulo. **Anais do Seminário.** Disponível em: <http://masrv54.agricultura.gov.br/anais/seminario.htm>. Acesso em: jun. 2002.

SÃO PAULO. Governo do Estado. **Agricultura.** Disponível em <http://www.saopaulo.sp.gov.br/invista/numeros/agric.htm#>. Acesso em: jun. 2002.

SATYAMURTY, P.; ETCHICHURY, P. C.; STUDZINSKI, C.; CALBETE, N. O.; LOPES, R. R.; GLAMMELSBACHER, I. A. V.; GLAMMELSBACHER, E. A. A. A primeira friagem de 1990: uma descrição sinótica. **Climanálise:** Boletim de Monitoramento e Análise Climática, v. 5, n. 5, p. 43-51, 1990.

TARIFA, J. R.; PINTO, H. S.; ALFONSI, R. R.; PEDRO JUNIOR, M. J. A gênese dos episódios meteorologicos de julho de 1975 e a variação espacial dos danos causados pelas geadas à cafeicultura no Estado de São Paulo. **Ciência e Cultura**, v. 29, n. 12, p. 1363-1374, dez. 1977.

THORNTHWAITE, C. W; MATHER, J. R. **The water balance.** New Jersey: Laboratory of Climatology, 1955. (Publication in Climatology, v. 8, n. 1)

TUBELIS, A.; NASCIMENTO, F. J. L. **Meteorologia descritiva:** fundamentos e aplicações brasileiras. São Paulo: Nobel, 1992. 374 p.

UNESP. Campus de Rio Claro. **Data mining.** Disponível em: <http://black.rc.unesp.br/IA/cintiab/datamine/teoria.html>. Acesso em: maio, 2002.

VEIGA FILHO, A. de A. **Estudo do processo de mecanização do corte na cana-de-açúcar:** o caso do estado de São Paulo, Brasil. Disponível em: <http://www.nuca.ie.ufrj.br/infosucro/biblioteca/agricultura/filho_estudo.rtf>. Acesso em: jun. 2002.

VENTURIERI, A.; SANTOS, J. R. Técnicas de classificação de imagens para análise de cobertura vegetal. In: ASSAD, E. D.: SANO, E. E. **Sistema de informações geográficas:** aplicações na agricultura. 2. ed. rev. ampl. Brasília: Embrapa–SPI: Embrapa–CPAC, 1998. p. 351-371.

WEILL, M. A. M. **Avaliação de fatores edafoclimáticos do manejo na produção de cafeeiros (*Coffea arabica* L.) na região de Marília e Garça, SP.** 1990. 182 p. Tese (Mestrado) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

WILLET, P. Recent trends in hierarchic document clustering: a critical review. **Information Processing & Management**, v. 24, n. 5, p. 577-597, 1988.

WITTEN, I. H.; CUNNINGHAM, S. J.; HOLMES, G.; MCQUEEN, R. J.; SMITH, L.A. Practical machine learning and its potential application to problems in agriculture. In: NEW ZEALAND COMPUTER CONFERENCE, 1996, Auckland. **Proceedings...** Auckland: [s. n.], 1996. v. 1, p. 308-325.

ZAVERUCHA, G.; BARBOSA, V. C.; DUTRA, I. C.; BAIÃO, F. A.; HALLACK, N.; BASILIO, R.; MENEZES, R. **I Escola Brasileira de Aprendizado de Máquina e Extração de Conhecimentos em Bases de Dados.** Disponível em: <http://www.cos.ufrj.br/~mlkdd/index_port.html>. Acesso em: jun. 2002.


Embrapa
Informática Agropecuária