UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS AGRONÔMICAS
CAMPUS DE BOTUCATU

APLICAÇÃO DE COMPUTAÇÃO EVOLUCIONÁRIA NA MINERAÇÃO DE DADOS FÍSICO-QUÍMICOS DA ÁGUA E DO SOLO

ALAINE MARGARETE GUIMARÃES

Tese apresentada à Faculdade de Ciências Agronômicas da UNESP - Campus de Botucatu, para obtenção do título de Doutor em Agronomia (Energia na Agricultura).

BOTUCATU-SP
Dezembro - 2005

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS AGRONÔMICAS
CAMPUS DE BOTUCATU

APLICAÇÃO DE COMPUTAÇÃO EVOLUCIONÁRIA NA MINERAÇÃO DE DADOS FÍSICO-QUÍMICOS DA ÁGUA E DO SOLO

ALAINE MARGARETE GUIMARÃES

Orientador: Prof. Dr. Angelo Cataneo
Co-orientador: Prof. Dr. Fedro S. Zazueta

Tese apresentada à Faculdade de Ciências Agronômicas da UNESP - Campus de Botucatu, para obtenção do título de Doutor em Agronomia (Energia na Agricultura).

BOTUCATU-SP
Dezembro - 2005

DEDICATÓRIA

Dedico esta tese ao meu avô Alfredo Moletta (*In Memorium*), a quem sempre chamei de Tate, que tanto me ensinou por meio de seus exemplos, de sua simplicidade e de sua imensa sabedoria e que partiu há poucos dias deixando uma enorme saudade em meu coração. Jamais esquecerei nossas longas conversas junto ao fogão à lenha... tomando chimarrão...

A ele presto esta minha última homenagem.

AGRADECIMENTOS

Agradeço imensamente a Deus por ter-me provido com saúde, coragem e capacidade para desenvolver esse trabalho. Nos momentos difíceis durante essa jornada foi a Ele que sempre me dirigi primeiramente pedindo por forças para continuar e sabedoria para seguir o melhor caminho.

Quando penso nos motivos que me conduziram a realizar esse doutorado, lembro do meu marido Alexandre Lunardon insistindo muito para que eu o fizesse. Talvez eu nunca tenha lhe dito pessoalmente, por isso aproveito para deixar registrado aqui o meu profundo agradecimento por essa insistência e por suas palavras que sempre me motivaram a vencer desafios e seguir em frente.

Sem minha mãe Arlete Moletta eu nada seria. Seu esforço imensurável para proporcionar-me condições de chegar até aqui está eternizado em minha memória. Seu gosto pela boa música e boa leitura, sua coragem e desprendimento, assim como seu interesse em sempre adquirir novos conhecimentos, são para mim maravilhosos exemplos de conduta e comportamento, os quais eu espero seguir sempre. Muito obrigada minha querida mãe por, mesmo sentindo minha falta, ter me incentivado a conduzir esse doutorado, principalmente durante o tempo em que estive fora do país. Jamais esquecerei dessa sua demonstração de amor maior.

A minha irmã Alana Guimarães, que é o meu referencial de vida, eu agradeço imensamente por ter sempre, invariavelmente, estado ao meu lado encorajando-me, e motivando-me a seguir em frente. Agradeço também, e muito, por ter cuidado de nossa família durante minha ausência.

Aos meus sogros Lolia e Aristeu Lunardon agradeço muito pela constante preocupação comigo e pelas palavras sempre carinhosas.

Ao meu orientador, Dr. Angelo Cataneo, que sempre acreditou em mim e nas minhas idéias e que nunca mediu esforços para ajudar-me a desenvolver um bom trabalho. Agradeço-lhe profundamente por ter me dado a oportunidade de realizar esse doutorado sob sua orientação, por tudo que me ensinou, por sua amizade e por seu exemplo de caráter.

Ao meu co-orientador, Dr. Fedro Zazueta, pela oportunidade de realizar o estágio na Universidade da Flórida, pela forma atenciosa com que sempre me tratou, pelas sugestões que contribuíram em muito para o aprimoramento dessa tese e por ter me provido com toda a infraestrutura necessária para que eu pudesse realizar meus estudos de forma tão produtiva e agradável naquela Universidade.

A Universidade Estadual de Ponta Grossa (UEPG) pela liberação para a realização desse doutorado. Aos funcionários da UEPG que de alguma forma contribuíram, em especial a Marcinha da PROPESP. Aos colegas do DEINFO que sinceramente compartilham comigo dessa conquista, meus agradecimentos.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de pós-graduação e da bolsa para realização do estágio de doutoramento no exterior.

A Universidade Estadual Paulista (UNESP) e em especial a Faculdade de Ciências Agronômicas de Botucatu (FCA) pela oportunidade. Agradeço também aos funcionários que de alguma forma me auxiliaram na condução desse doutoramento, em especial a Marlene da Pós-Graduação por sua forma gentil e atenciosa e ao Marquinhos do Departamento de Gestão e Tecnologia Agroindustrial a quem considero um grande amigo.

A Universidade da Flórida – EUA por ter-me aceitado para a realização do doutorado “sandwich”. A todos os professores, pesquisadores e funcionários do Escritório de Tecnologia Acadêmica dessa Universidade por terem compartilhado comigo seus conhecimentos e por terem acolhido-me de forma tão carinhosa. Em especial Pat, Betty, Cathie e Ronald.

A Leila Vriesmann, ex-orientada e que hoje, para minha alegria, é minha colega na UEPG e mais do que isso, minha grande amiga, agradeço muito pelo excelente trabalho desenvolvido sob minha orientação o qual é parte fundamental dessa tese e por toda sua dedicação e boa-vontade atendendo sempre a todas às minhas solicitações.

Ao pesquisador Dr. Petraq Papajorgji, a quem hoje tenho como um grande amigo, pelas sugestões e por toda a ajuda durante o tempo em que morei nos Estados Unidos.

Ao Dr. José Paulo Molin, por ter acreditado no meu trabalho e por ter fornecido a base de dados brasileira que utilizei nessa tese.

Ao Dr. Alex Freitas, agradeço pelo apoio e por sua importantíssima contribuição técnica.

Aos amigos Dr. Jorim Sousa e Dr. Marcelo Canteri agradeço por me estimularem a realizar o doutorado na área de informática aplicada à agricultura e por terem me apresentado para o meu orientador e co-orientador, respectivamente.

Ao amigo MSc. Ivo Mathias pelas palavras de incentivo e por ter prontamente ajudado-me no desenvolvimento de um dos módulos do sistema.

Finalmente, agradeço de forma muito especial às amigas Cláudia Corseuil, Siumara Daer e Cristina Piazera pelas palavras carinhosas em nossas inúmeras conversas via MSN, algumas vezes sérias e outras divertidas, e que indubitavelmente contribuíram para que eu mantivesse a calma, a perseverança e o bom humor, principalmente durante o tempo em que estive sozinha fora do país. Vocês ocupam um lugar muito especial em meu coração.

LISTA DE ABREVIATURAS

ACO	Otimização de Colônia de Formigas
AD	Árvore de Decisão
AG	Algoritmo Genético
AM	Aprendizado de Máquina
AP	Agricultura de Precisão
CE	Computação Evolucionária
CSV	Campos Separados por Vírgula
DFD	Diagrama de Fluxo de Dados
DM	Data Mining (Mineração de Dados)
EPA	Agência de Proteção Ambiental dos Estados Unidos
GPS	Sistema de Posicionamento Global
IC	Inteligência Computacional
KDD	Descoberta de Conhecimento em Bases de Dados
MD	Mineração de Dados
PC	Computador Pessoal
RAM	Memória de Acesso Randômico
RNA	Redes Neurais Artificiais
SGBD	Sistema Gerenciador de Banco de Dados
SIG	Sistema de Informações Geográficas
UF	Universidade da Flórida
UM	Unidade Monetária
WEB	Rede na Internet

LISTA DE FIGURAS

SUMÁRIO

RESUMO

Essa tese apresenta o desenvolvimento de um sistema de mineração de dados baseado na técnica de computação evolucionária denominada Algoritmos Genéticos. O sistema resultante, de nome MinAG, realiza a tarefa de classificação de dados contínuos e destina-se a minerar dados físico-químicos do solo e da água. Os padrões de comportamento dos atributos minerados são apresentados no formato SE-ENTÃO, facilitando a compreensão da informação descoberta. Foram definidos alguns requerimentos e restrições para o uso desse sistema relacionados às características do arquivo de dados possível de ser minerado. O MinAG adota o conceito de computação em *grid*, o que propicia para que mais e melhores resultados sejam obtidos. Os testes realizados permitiram concluir que o sistema executou as tarefas definidas para o mesmo e gerou resultados corretos ao minerar as bases de dados a que se propôs, atingindo, portanto os objetivos dessa tese. Foram realizados dois estudos de casos. No primeiro foi utilizada uma base de dados brasileira sobre dados físico-químicos do solo obtidos por equipamentos de agricultura de precisão na região de Campos Novos Paulista – SP. No segundo estudo de caso usou-se uma base de dados de qualidade de água do estado da Flórida – EUA. Em ambos os casos o sistema foi capaz de atingir seu objetivo encontrando padrões de comportamento nos dados. Pode-se concluir que o sistema MinAG apresenta-se como uma nova maneira de analisar a correlação entre os elementos físico-químicos do solo e da água. Esse sistema não deve ser entendido como um substituto de métodos de análise tradicionais, como a estatística. Sua função é servir como uma ferramenta adicional na geração de informações para auxílio à compreensão do comportamento existente nos dados.

Palavras-chave: Agricultura de Precisão, Algoritmos Genéticos, Inteligência Computacional, Mineração de Dados, propriedades físico-químicas.

EVOLUTIONARY COMPUTATION APPLIED TO WATER AND SOIL PHYSICO-CHEMICAL DATA MINING. Botucatu, 2005. 127p. Tese (Doutorado em Agronomia/Energia na Agricultura) - Faculdade de Ciências Agronômicas, Universidade Estadual Paulista.

Author: ALAINE MARGARETE GUIMARÃES

Adviser: ANGELO CATANEO

Co-Adviser: FEDRO S. ZAZUETA

SUMMARY

This thesis presents the data mining system development based on an evolutionary computation technique named Genetic Algorithms. The MinAG system performs the continuous data classification task and mines water and soil physico-chemical datasets. The patterns discovered by mining the attributes are presented using the IF-THEN rule format. It makes it easier to understand the information discovered. Some requirements and restrictions related to the dataset features were defined in order to use the system. MinAG adopts the grid computing concept in order to produce more and better results. By the evaluation system, it was possible to conclude that it is able to perform the proposed tasks and produces correct results when mining the datasets. Therefore, the system reached the thesis goals. Two case studies were performed. In the first one, a Brazilian dataset related to soil physico-chemical properties was used. The data was obtained in Campos Novos Paulista – SP by Precision Agriculture equipment. In the second case study, a Florida – USA water quality dataset was utilized. The system discovered behavior patterns achieving the goals in both cases. The MinAG system presents a new way to analyse the correlation between the water and soil physico-chemical attributes. This system is not a substitute for traditional methods such as statistics. In fact, it is an auxiliary tool to generate information in order to help understand the behavior between data.

Keywords: Precision Agriculture, Genetic Algorithms, Computational Intelligence, Data Mining, physico-chemical properties.

1 INTRODUÇÃO

A última década trouxe avanços nos processos de obtenção e armazenamento de dados agrícolas e ambientais, resultantes do uso de sensores remotos, satélites espaciais, introdução de códigos de barras, automação de processos, além da sofisticação dos sistemas computacionais de gerenciamento de banco de dados. Com isso surgiu a necessidade de explorar a gama de dados disponíveis com o objetivo de obter novas informações que apresentem tendências, regularidades, enfim, padrões que possam contribuir de forma estratégica na tomada de decisões em domínios específicos.

Tal necessidade mostra-se fortemente presente no tratamento de dados associados a fatores físico-químicos do solo, visto que atualmente é possível se obter inúmeros diferentes tipos de dados do solo devido às tecnologias de agricultura de precisão. A compreensão da inter-relação entre tais dados é definitivamente complexa e de extrema importância na busca de maior produtividade e menor agressão ao meio ambiente. Essa complexidade se deve ao fato de existir uma grande quantidade de fatores físico-químicos envolvidos, além de outros fatores interferentes no desenvolvimento das plantas e no comportamento do solo, como temperatura e precipitação, os quais são incontroláveis e de difícil predição. Embora seja muito difícil compreender a interação entre os tais fatores, essa compreensão é interessante para que possam ser tomadas melhores decisões relacionadas ao uso e tratamento do solo.

Da mesma forma, a compreensão do comportamento e variabilidade dos fatores físico-químicos presentes na água é importante na definição do melhor aproveitamento

da água bem como no processo de controle da qualidade da mesma. O monitoramento e análise de dados relativos à água são onerosos, assim como o tratamento para redução de poluição das águas também o é. Os avanços tecnológicos permitem que estações de monitoramento da água sejam controladas e dados a serem analisados sejam periodicamente coletados, em diferentes intervalos de tempo. Assim como no solo, outras variáveis de difícil controle também interferem nos fatores físico-químicos da água tornando ainda mais complexa a compreensão do comportamento desses fatores.

Normalmente as análises dos dados físico-químicos, tanto do solo como da água, são feitas utilizando-se métodos estatísticos, tais como: estatística descritiva, geoestatística e inferência estatística. Tais métodos são úteis e indispensáveis, porém muitas vezes apresentam o resultado em uma forma não muito clara ou de difícil entendimento. Além disso, normalmente o executor da análise precisa pré-definir os conhecimentos/resultados esperados da análise, o que induz a uma dependência do conhecimento prévio do executor, o que de certa forma limita a possibilidade de obtenção de novos e inesperados conhecimentos sobre os dados.

Além do citado acima, é sabido que quanto mais próximo da linguagem humana um conhecimento é representado, mais fácil torna-se o seu entendimento e consequentemente sua aplicação. Em função disso, novas metodologias de análises têm sido propostas, utilizando os mais modernos conceitos da computação como ferramentas para agilizar o processo de análise dos dados. Entre essas metodologias destaca-se aquela denominada Descoberta de Conhecimento em Bases de Dados, conhecida pela sigla KDD do inglês *Knowledge Discovery in Databases.*

Inserido no contexto KDD está a Mineração de Dados (MD), mais conhecida como Data Mining (DM), a qual consiste no processo de extrair informação válida, previamente desconhecida e de máxima abrangência a partir de grandes bases de dados, usando-as para efetuar tomadas de decisões. DM é considerada uma das ferramentas de KDD mais utilizadas, tanto no meio comercial quanto no meio científico, para a descoberta eficiente de informações valiosas em uma grande coleção de dados, visando o auxílio no suporte a decisão (Canuto & Gottgtroy, 1997). Existem diferentes tipos de tarefas em mineração de dados, sendo uma delas a classificação, técnica que consiste na aplicação de um conjunto de

exemplos pré-classificados, para desenvolver um modelo capaz de classificar uma população maior de registros.

Associadas a essas tarefas estão diversas técnicas de mineração, que variam em função da complexidade, tempo de processamento e tarefa mais adequada a executar.

Uma dessas técnicas chama-se Algoritmos Genéticos (AGs). Esta técnica tem sido utilizada na otimização de solução de vários problemas complexos, porém ainda é pouco usada na área das ciências agrárias e ambientais.

1.1 Objetivos

O objetivo dessa tese consistiu em definir uma nova metodologia para análise de dados físico-químicos da água e do solo, por meio do desenvolvimento de um sistema de mineração de dados empregando a técnica de computação evolucionária denominada Algoritmos Genéticos. O propósito desse sistema foi realizar a classificação de dados, comportando-se como uma ferramenta auxiliar na análise e compreensão das correlações existentes entre os mesmos.

Para a validação da metodologia dois diferentes estudos de caso são apresentados. Ambos comportam as condições necessárias para serem submetidos ao sistema de mineração de dados proposto. O primeiro deles refere-se a uma base de dados obtidos na região de São Paulo – Brasil por processos de agricultura de precisão. O segundo refere-se a uma base de dados de qualidade da água do estado da Flórida - EUA.

O objetivo do primeiro estudo de caso consistiu em demonstrar o uso do sistema para a mineração de dados referentes às propriedades físico-químicas do solo associadas ou não à produtividade de grãos.

O objetivo do segundo estudo de caso foi demonstrar a aplicação sistema na análise do comportamento dos fatores físico-químicos da água de acordo com os critérios de uso para vida aquática.

É importante ressaltar que o objetivo desse trabalho consiste no desenvolvimento de um novo sistema de análise de dados, não visando resolver um problema

específico, mas sim demonstrar exemplos de casos em que tal sistema pode ser aplicado, bem como indicar suas limitações e requerimentos.

1.2 Organização da tese

A tese é apresentada em nove capítulos. O primeiro deles apresenta essa introdução. O capítulo 2 apresenta uma revisão da literatura referente à metodologia de mineração de dados a ser desenvolvida sendo abordada sua importância na análise de dados do solo e da água, justificando o emprego da mesma. O capítulo 3 trata da importância da análise dos fatores físico-químicos do solo e da água. O capítulo 4, denominado Material e Métodos: Sistema Desenvolvido, apresenta a metodologia utilizada para o desenvolvimento do sistema MinAG. Em seguida, o sistema resultante é apresentado no capítulo 5, o capítulo destinado a resultados e discussão, onde dois estudos de caso são abordados: o primeiro deles, referente à aplicação da mineração em dados obtidos por processo de Agricultura de Precisão e o segundo referente a dados de qualidade da água. As considerações finais são apresentadas no capítulo 6, as conclusões sobre o trabalho desenvolvido são expostas no capítulo 7 e as referências bibliográficas são citadas no capítulo 8.

2 REVISÃO DA LITERATURA

2.1 Mineração de dados e KDD

A busca por informações estratégicas ocorre nos mais diversos segmentos. Geneticistas procuram padrões em genomas de seres vivos, empresas tentam descobrir o perfil de seus clientes para realizar marketing direto, economistas fazem predições de mercado com base em dados históricos, pesquisadores da área agronômica têm como meta descobrir características em dados obtidos com recursos de sistemas de posicionamento global, entre outros inúmeros exemplos tanto da área científica como comercial.

Para atender a essas necessidades surgiu a área de pesquisa centrada na descoberta de conhecimento em bases de dados, conhecida como KDD, do inglês Knowledge Discovery in Databases.

Segundo Fayyad (1996) KDD consiste em um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, em conjuntos de dados. Os termos desta caracterização devem ser entendidos em mais detalhes:

- **Dados:** correspondem ao conjunto de fatos, por exemplo, casos em um banco de dados.

- **Padrão:** é uma expressão descrevendo fatos que ocorrem com determinada freqüência, demonstrando um padrão de comportamento nos dados.

- **Processo:** Usualmente em KDD um processo é composto por vários passos, que envolvem preparação de dados, busca de padrões, avaliação do conhecimento e refinamento envolvendo iteração depois de modificação. O processo é assumido como não trivial por ter algum nível de busca autônoma. Por exemplo, o cálculo da idade média em um conjunto de dados não pode ser qualificado como um processo de KDD por ser uma tarefa trivial.

- **Validade:** A descoberta de padrões deve ser válida sobre novos dados com determinado grau de certeza. Uma medida de certeza é uma função que mapeia as expressões (padrões) para um espaço de medida ordenado totalmente ou parcialmente.

- **Novo:** Os padrões são novos (pelo menos para o sistema). A novidade pode ser medida com respeito a:

  – **Mudanças nos dados:** comparando valores correntes com prévios ou já esperados.

  – **Conhecimento:** como um conhecimento novo está relacionado com um antigo.

- **Potencialmente útil:** Os padrões devem levar a uma ação útil, sendo medida por alguma função de utilidade.

- **Compreensível:** Um objetivo do KDD é gerar conhecimento compreensível por seres humanos para facilitar a compreensão dos dados subjacentes. Se o conhecimento não é compreensível pelo usuário ele não é capaz de interpretá-lo ou validá-lo. Neste caso o usuário não confiará o suficiente nesse conhecimento para usá-lo em tomada de decisão. Como é difícil de medir a compreensiblidade de uma regra, muitas vezes ela é substituída pela medida de simplicidade.

Várias medidas de simplicidade existem e elas variam de medidas puramente sintáticas (por exemplo, o tamanho de um padrão em bits) para a semântica (por exemplo, a facilidade para humanos entenderem em alguma colocação).

A combinação das medidas de validade, novidade, utilidade e simplicidade, resulta em uma medida de valor que expressa o grau de interesse de uma regra. Alguns sistemas KDD têm uma função explícita de grau de interesse. Outros sistemas definem o grau de interesse indiretamente ordenando os padrões descobertos.

Considere o exemplo de um padrão hipoteticamente descoberto:

SE Estado (grávida) **ENTÃO** Sexo (feminino).

Muito embora o padrão acima corresponda a uma informação correta e válida, não pode ser considerado novo nem potencialmente útil, por ser declaradamente óbvio. Com isso o grau de interesse dessa regra certamente resultaria em uma medida extremamente baixa, por não apresentar novo conhecimento.

A determinação de quando uma informação pode ser vista como novo conhecimento é subjetiva, estando fortemente relacionada ao usuário, no sentido em que dependerá de suas opções em relação a forma de mapeamento, as funções e limites adotados. Isso significa que uma informação sendo entendida como conhecimento por um determinado usuário pode não ser considerada da mesma forma para outro, em função dos parâmetros adotados em cada análise.

A Figura 1 apresenta uma visão geral sobre os passos que compreendem o processo de KDD, no qual a fase de mineração de dados ocorre sobre dados que foram transformados em passos anteriores. Para melhor compreensão vejamos a seguir uma breve descrição dos sete passos do processo KDD (Fayyad, 1996).

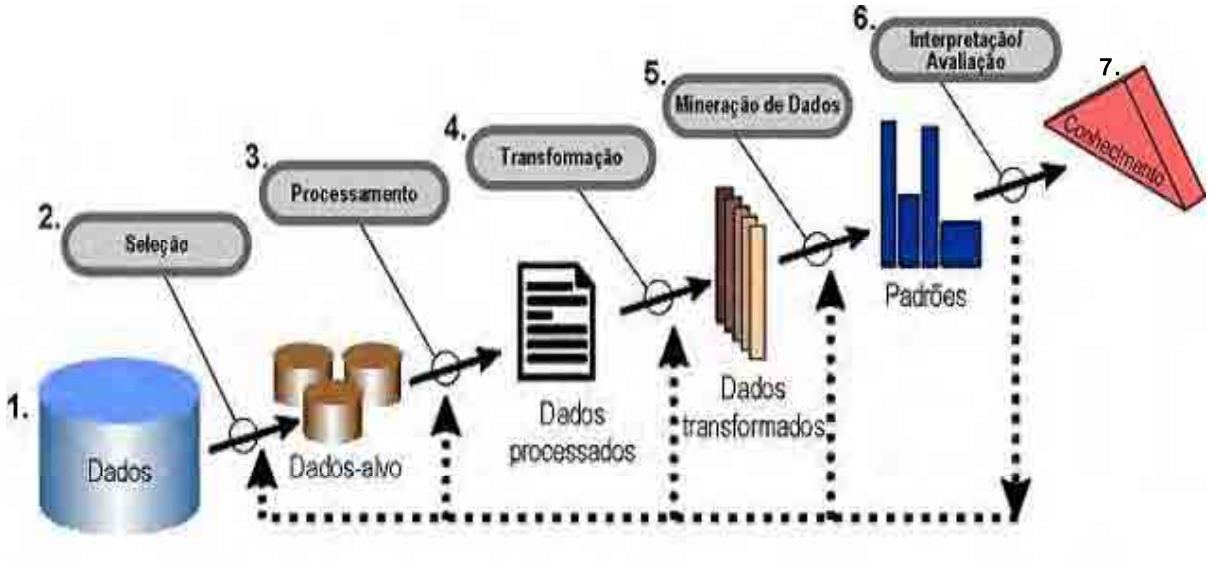


Figura 1. Visão geral dos passos do processo de KDD, adaptado de Fayyad (1996)

1. O primeiro passo para realizar o processo de KDD consiste em obter a compreensão do domínio da aplicação e as metas do usuário final. É preciso observar quais são as bases de dados disponíveis e o volume de dados associados de alguma forma com a meta estabelecida.

2. Criar um conjunto de dados meta selecionando um conjunto de dados, ou estabelecendo um subconjunto de variáveis ou exemplos de dados, sobre os quais a descoberta será executada.

3. Alguns fatores colaboram com os erros de reconhecimento de padrões. O ambiente de extração das características freqüentemente apresenta ruídos, distorções, etc. Para diminuí-los dever-se realizar operações básicas, tais como remover ruídos ou valores incoerentes, coletar as informações necessárias para o modelo e decidir sobre estratégias para controlar campos de dados perdidos.

4. Efetuar a redução e projeção dos dados, encontrando características úteis para representar-los dependendo do objetivo da tarefa. Deve-se também usar a redução de

dimensionalidade ou métodos de transformação para reduzir o número efetivo de variáveis em consideração.

5. Na fase de mineração de dados, primeiramente deve-se definir a tarefa de mineração que será feita, ou seja, classificação, regressão, agrupamento, etc. As várias tarefas possíveis de DM serão vistas em mais detalhes na seção 2.3. Uma vez definida a tarefa será eleito o algoritmo a ser usado para encontrar os padrões nos dados. Isso inclui decidir que modelos e parâmetros podem ser apropriados (por exemplo, modelos para dados categóricos são diferentes de modelos para dados numéricos reais). A mineração dos dados é executada então, por meio do algoritmo que, para apresentar os padrões encontrados, adotará uma forma representacional particular como regras de classificação, árvores, gráficos de agrupamento, etc. O usuário pode auxiliar significativamente o método de DM executando corretamente os passos precedentes.

6. Nesse passo é importante interpretar os padrões minerados, possivelmente retornando a qualquer um dos passos anteriores para novas iterações, caso necessário.

7. Finalmente, é necessário consolidar o conhecimento descoberto incorporando-o ao sistema global, ou simplesmente documentando-o e relatando-o para as partes interessadas. Isso também inclui checar e resolver possíveis conflitos com conhecimentos previamente extraídos ou conhecidos.

As fases do KDD estão diretamente relacionadas a tradicional pirâmide da informação (Figura 2). Assim como um organismo vivo, as empresas recebem informação do meio ambiente e também atuam sobre ele. Durante essas atividades, é necessário distinguir vários níveis de informação. Ao observar a Figura 2 pode-se notar o natural aumento de abstração conforme subimos de nível.

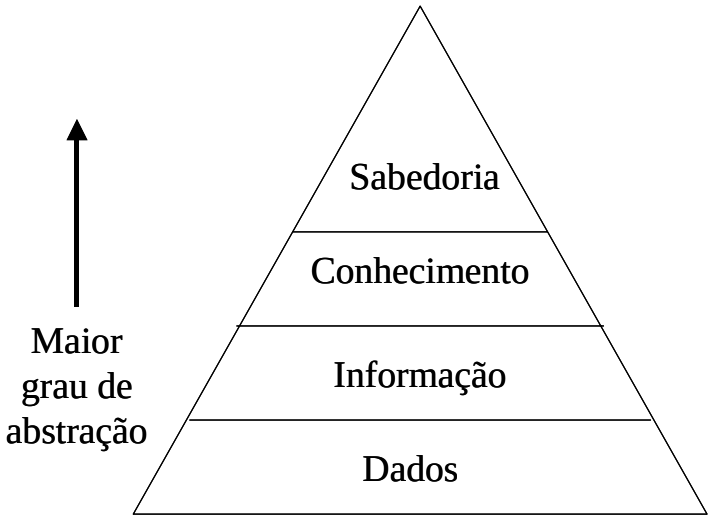


Figura 2. Pirâmide da informação (Navega, 2002)

Segundo Navega (2002) o diagrama apresentado na Figura 2, quando atualizado, fica como apresentado na Figura 3. O fundamental a se perceber neste novo diagrama é a sensível redução de volume que ocorre cada vez que subimos de nível. Essa redução é uma conseqüência natural do processo de abstração.

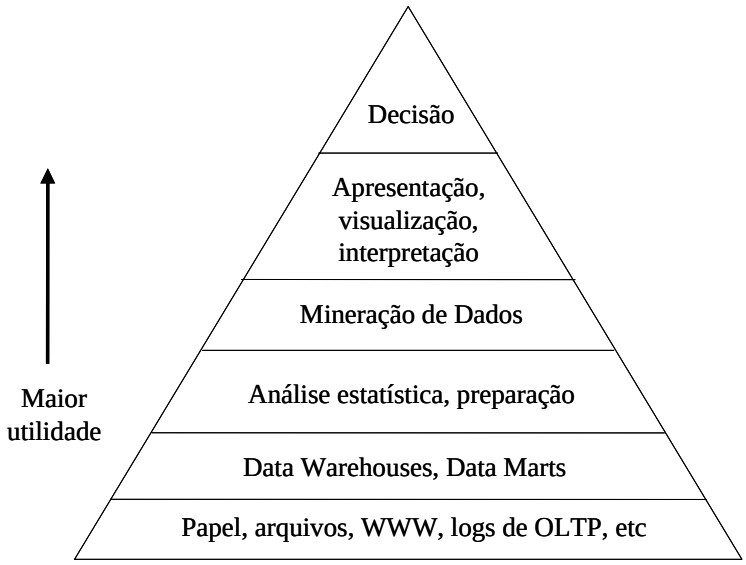


Figura 3. Pirâmide da informação atualizada (Navega, 2002)

Abstrair, no sentido aqui usado consiste em representar uma informação através de correspondentes simbólicos e genéricos. Este ponto é importante: como acabamos de ver, para ser genérico, é necessário "perder" um pouco dos dados, para só conservar a essência da informação. A fase de mineração de dados no processo de KDD localiza padrões através da aplicação de processos de generalização, algo que é conhecido como indução. O processo de KDD pode envolver iteração significativa. Todos os passos do KDD são importantes para o sucesso de sua utilização, porém abordaremos a seguir as questões pertinentes à fase de mineração de dados. Maiores detalhes sobre os demais passos podem ser obtidos em Fayyad (1996).

2.2 Definições e características do processo de mineração de dados

Fayyad (1996) define Mineração de Dados como um passo no processo de KDD consistindo de algoritmos de mineração que, sob algumas limitações aceitáveis de eficiência computacional, produzem uma enumeração de padrões sobre os fatos observados em uma base de dados.

Uma definição mais genérica considera a Mineração de Dados como sendo o uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertos a olho nu pelo ser humano (Carvalho, 2001).

A mineração de dados consiste em mais do que uma simples consulta a um banco de dados, no sentido em que permite a exploração e inferência de informação útil a partir dos dados, descobrindo relacionamentos escondidos.

Ainda que as técnicas de MD sejam antigas, apenas nos últimos anos passaram a ser usadas como exploração de dados por vários motivos:

- **O volume de dados disponível atualmente é enorme:** Mineração de dados é uma técnica que só se aplica a grandes massas de dados, pois necessita disso para calibrar seus algoritmos e extrair dos dados conclusões confiáveis. Empresas dos mais diversos segmentos, inclusive do meio agronômico, geram a cada dia uma

grande quantidade de dados sobre seus serviços e produtos. Esses dados são passíveis de análise por mineração.

- **Os dados estão sendo organizados:** Com a tecnologia de banco de dados os mesmos estão sendo organizados e padronizados de forma a possibilitar sua utilização dirigida para o auxílio à decisão. As técnicas de mineração de dados necessitam de bancos de dados limpos, padronizados e organizados.

- **Os recursos computacionais são potentes:** A MD necessita de muitos recursos computacionais para operar seus algoritmos sobre grandes quantidades de dados. O avanço da área de banco de dados também auxiliou em muito a mineração de dados.

- **A competição empresarial exige técnicas mais modernas de decisão:** As empresas da área de finanças, telecomunicações e agronegócios experimentam a cada dia mais e mais competição e necessidade de reduzir riscos e custos dos erros. Isso levou as empresas a buscarem, em seus dados, informações estratégicas que contribuam para as tornarem mais competitivas e para evitarem erros.

- **Programas comerciais de Data Mining já podem ser adquiridos:** Alguns padrões já podem ser encontrados contendo as técnicas mais conhecidas e bem definidas. As técnicas mais recentes, no entanto, ainda se encontram no campo acadêmico.

A mineração de dados é um campo interdisciplinar que envolve a Estatística, a Inteligência Computacional e o Aprendizado de Máquina (Andreatto, 2002).

- **Estatística**

A mais antiga linhagem de descendência do DM é a Estatística Clássica, sem a qual não seria possível termos o DM, visto que a mesma é a base da maioria das tecnologias a partir das quais é construído.

A Estatística Clássica envolve conceitos como distribuição normal, variância, análise de regressão, desvio simples, análise de conjuntos, análises de discriminantes e intervalos de confiança, todos usados para estudar dados e os relacionamentos entre eles. Esses são as pedras fundamentais onde as mais avançadas análises estatísticas se apóiam. E sem dúvida, no coração das atuais ferramentas e técnicas de mineração de dados, a análise estatística clássica desempenha um papel fundamental.

- **Inteligência Computacional**

A Inteligência Artificial, ou IA, é construída a partir dos fundamentos da heurística, em oposto à estatística, e tenta imitar a maneira como o homem pensa na resolução dos problemas estatísticos. Em muitos casos a credibilidade da aplicação efetiva de soluções baseadas em IA ficou comprometida em função das inúmeras discussões teóricas sobre a significância e magnitude do termo IA, envolvendo inclusive aspectos filosóficos, no sentido de até onde pode-se criar uma inteligência artificial.

Surgiu então o termo Inteligência Computacional (IC), como sendo uma área da computação concentrada na implementação de solução colaborativas e não competitivas, no sentido em que visa implementar em sistemas complexos do mundo real, soluções computacionais com algumas características de comportamento inteligente para contribuir na busca por resultados melhores e mais compreensíveis.

Existe ainda uma grande divergência na definição do termo Inteligência Computacional. Poole et al. (1998) a definem como o estudo do projeto de agentes inteligentes. Considerando-se que agente é algo que age em um ambiente e que para ser inteligente deve ser dotado de características como crenças, desejos e intenções (Guimarães, 2000), deve ser flexível para alterar o ambiente e suas metas (Mathias, 2000), essa definição acaba por ser restritiva no contexto das implementações possíveis na área de IC, uma vez que

muitas das soluções computacionais que envolvem a modelagem da inteligência não apresentam necessariamente as características básicas do que pode ser considerado um agente.

Segundo Palazzo (2003) a área denominada Inteligência Computacional (IC) é formada pelo estudo de sistemas Fuzzy, Redes Neurais e Computação Evolutiva (CE) que compreende uma ramificação da ciência da computação. A IC, por sua vez, em conjunto com outras áreas, tais como Vida Artificial, Geometria Fractal, Teoria do Caos, Sistemas Complexos, etc., delimitam um campo conhecido como Computação Natural (CN).

O importante a notar é que a principal diferença entre os termos IA e IC é que enquanto a IA ocupa-se de aspectos teóricos, conceituais e qualitativos da representação de modelos de inteligência e de conhecimento, a IC ocupa-se, de modo mais prático, da implementação computacional de modelos na busca de solução e otimização de problemas complexos.

- **Aprendizado de Máquina**

É a terceira e última linhagem do DM podendo ser descrita como a associação entre a Estatística e a IC, por combinar a heurística e análise estatística.

O Aprendizado de Máquina (AM) é o estudo de algoritmos que melhoram com a experiência, ou seja, cujo desempenho em determinada tarefa melhora com a experiência (Mitchell, 1997).

O Aprendizado de Máquina tenta fazer com que os programas de computador “aprendam” com os dados que eles estudam, de forma que esses programas tomem decisões diferentes baseadas nas características dos dados estudados, usando a estatística para os conceitos fundamentais, e adicionando heurística avançada da IC aos algoritmos para alcançar os seus objetivos.

Segundo Mitchell (1997) pode-se desenvolver algoritmos de aprendizado baseados em árvores de decisão, redes neurais artificiais, redes Bayesianas, métodos baseados em instância, modelos baseados na evolução biológica como algoritmos e programação genética, entre outros.

Algoritmos de AM são úteis em uma grande variedade de aplicações, especialmente na mineração de dados. Para se aplicar o AM é necessário que se tenha bem definida a tarefa a ser realizada, uma medida de performance e uma fonte de experiência para

treinamento. Projetar uma abordagem de AM envolve um número de escolhas como: tipo de experiência de treinamento, a função objetivo a ser aprendida, uma representação desta função objetivo, e um algoritmo para aprender a função objetivo a partir dos exemplos de treinamento.

O aprendizado envolve busca, ou seja, deve-se buscar no espaço das possíveis hipóteses aquela que melhor se ajusta aos exemplos de treinamento disponíveis e outras restrições e conhecimento prévios.

2.3 Tarefas de mineração de dados

Existem diversas tarefas de mineração de dados incluindo classificação, regressão, agrupamento, modelagem de dependência, etc. (Fayyad et al., 1996). Cada tarefa pode ser vista como um problema a ser resolvido por um algoritmo de mineração de dados. Portanto, o primeiro passo na hora de projetar um algoritmo de DM é definir a qual tarefa o algoritmo estará direcionado. Vejamos a seguir, as principais tarefas: associação, classificação e agrupamento.

2.3.1 Associação

A tarefa de associação consiste em identificar relações presentes entre diferentes atributos em um registro. Muito usada na área comercial, essa tarefa permite, por exemplo, identificar quais produtos que, normalmente, são comprados juntos por um mesmo consumidor.

Também conhecida como análise de afinidade (Carvalho, 2001), a tarefa de associação busca determinar que fatos ocorrem simultaneamente com probabilidade razoável (co-ocorrência) ou que itens estão presentes juntos com uma certa chance (correlação). Dos números obtidos desta análise, pode-se extrair “regras de associação”.

O algoritmo Apriori (Agrawal & Srikant, 1994) desempenha com muita propriedade esse tipo de tarefa.Vejamos o emprego desse algoritmo para o seguinte problema proposto (Guimarães, 2004):

Uma companhia de cartões de crédito deseja realizar uma campanha para aumentar o volume do uso de cartões em estabelecimentos de pequeno porte do ramo alimentício, como lanchonetes e panificadoras. Para tanto deseja identificar em quais estabelecimentos um cliente usa o cartão em determinado mês, para saber onde pode encontrar clientes que podem vir a utilizar seus cartões em pequenos estabelecimentos alimentícios também. O Quadro 1 apresenta as transações feitas por clientes em determinado período. Cada registro corresponde a uma transação de um cliente, com itens assumindo valores binários (sim/não), indicando se o cliente comprou ou não no respectivo estabelecimento.

Quadro 1. Transações feitas com cartão de crédito

Nº Trn	**Farmácia**	**Perfuma-ria**	**Panifica-dora**	**Restau-rante**	**Lancho-nete**	**Livraria**	**Super-mercado**
1	Sim	Não	Não	Sim	Sim	Não	Sim
2	Sim	Não	Sim	Sim	Sim	Não	Não
3	Não	Não	Não	Sim	Sim	Não	Sim
4	Não	Não	Não	Sim	Sim	Não	Sim
5	Não	Não	Sim	Não	Não	Não	Não
6	Não	Não	Não	Não	Sim	Não	Não
7	Não	Não	Não	Sim	Não	Não	Não
8	Não	Sim	Não	Não	Não	Não	Não
9	Não	Sim	Não	Não	Não	Sim	Não
10	Não	Não	Não	Não	Não	Sim	Não

Sabendo-se que uma regra de associação é um relacionamento, do tipo:

SE (X) ENTÃO (Y),

onde X e Y são conjuntos de itens, com $X \cap Y = \varnothing$, e que a cada regra são atribuídos dois fatores:

$$\textbf{Suporte (Sup.)} = \frac{\text{Nº de registros com X e Y}}{\text{Nº total de registros}}$$

a tarefa para o problema proposto consiste em descobrir todas as regras de associação com um suporte mínimo de 0,3 e confiança mínima de 0,8, por exemplo.

O algoritmo Apriori tem duas fases, divididas em vários passos de acordo com o número de itens existentes na base de dados. Basicamente os passos em cada

$$\textbf{Confiança (Conf.)} = \frac{\text{Nº de registros com X e Y}}{\text{Nº de registros com X}}$$

fase são:

Fase I:

- Descobrir conjuntos de itens freqüentes.
- Descobrir todos os conjuntos de itens com suporte maior ou igual ao mínimo suporte especificado pelo usuário.

Fase II:

- Descobrir regras com alto fator de confiança.
- A partir dos conjuntos de itens freqüentes, descobrir regras de associação com fator de confiança maior ou igual ao especificado pelo usuário.

Para realizar os passos do algoritmo sobre os dados propostos no Quadro 1, vamos considerar cada estabelecimento comercial como sendo um item a ser associado.

Passo 1: Calcular suporte de conjuntos com 1 item.

Item farmácia: Sup = 0,2

Item supermercado: Sup = 0,3

Item panificadora: Sup = 0,2

Item restaurante: Sup = 0,5

Item lanchonete: Sup = 0,5

Item livraria: Sup = 0,2

Item perfumaria: Sup = 0,2

Itens freqüentes (Sup >= 0,3): supermercado, restaurante, lanchonete.

Passo 2: Calcular suporte de conjuntos com 2 itens.

Otimização: Se um item I não é freqüente, um conjunto com dois itens, um dos quais é o item I, não pode ser freqüente. Logo, conjuntos contendo o item I podem ser ignorados visto que sua freqüência estará abaixo do mínimo aceitável.

Conjunto de itens: supermercado, restaurante. Sup =0,3.

Conjunto de itens: supermercado, lanchonete. Sup = 0,3.

Conjunto de itens: lanchonete, restaurante. Sup = 0,4.

Conjuntos de itens frequentes (Sup > = 0,3): {supermercado, restaurante}, {supermercado, lanchonete}, {lanchonete, restaurante}.

Passo 3: Calcular suporte de conjuntos com 3 itens.

Otimização: Se o conjunto de itens {I, J} não é freqüente, um conjunto com três itens, incluindo os itens {I, J} não pode ser freqüente. Logo, conjuntos contendo itens {I, J} podem ser ignorados.

Conjunto de itens: supermercado, restaurante, lanchonete. Sup = 0,3.

Conjunto de itens freqüentes (Sup >= 0,3): {supermercado, restaurante, lanchonete}.

Uma vez que foram realizadas, para o nível de suporte estabelecido, todas as associações possíveis entre os itens freqüentes, o próximo passo consiste em calcular o fator de confiança das regras, sabendo-se que:

$$\textbf{Conf. da regra “SE X ENTÃO Y”} = \frac{\text{Quantidade de transações contendo X e Y}}{\text{Quantidade de transações contendo X}}$$

Conjunto de itens: {supermercado, restaurante}.

SE supermercado ENTÃO restaurante. Conf = 1,0.

SE restaurante ENTÃO supermercado. Conf = 0,6.

Conjunto de itens: {supermercado, lanchonete}.

SE supermercado ENTÃO lanchonete. Conf = 1,0.

SE lanchonete ENTÃO supermercado. Conf = 0,6.

Conjunto de itens: {lanchonete, restaurante}.

SE lanchonete ENTÃO restaurante. Conf = 0,8.

SE restaurante ENTÃO lanchonete. Conf = 0,8.

Conjunto de itens: {supermercado, lanchonete, restaurante}.

SE supermercado, restaurante ENTÃO lanchonete. Conf = 1,0.

SE supermercado, lanchonete ENTÃO restaurante. Conf = 1,0.

SE lanchonete, restaurante ENTÃO supermercado. Conf = 0,75.

SE supermercado ENTÃO restaurante, lanchonete. Conf = 1,0.

SE restaurante ENTÃO supermercado, lanchonete. Conf = 0,6.

SE lanchonete ENTÃO supermercado, restaurante. Conf = 0,6.

Finalmente, seleciona-se as regras com confiança (Conf) maior ou igual ao valor mínimo especificado pelo usuário (para o problema proposto, 0,8).

SE lanchonete ENTÃO restaurante. Conf = 0,8.

SE restaurante ENTÃO lanchonete. Conf = 0,8.

SE supermercado, restaurante ENTÃO lanchonete. Conf = 1,0.

SE supermercado, lanchonete ENTÃO restaurante. Conf = 1,0.

SE supermercado ENTÃO restaurante, lanchonete. Conf = 1,0.

Observe que, em relação ao objetivo do problema proposto, nas regras encontradas pode-se concluir sobre estabelecimentos alimentícios de pequeno porte do tipo lanchonete. Informações sobre relações com panificadoras não puderam ser descobertas, dentro de um suporte e confiança mínimos aceitáveis.

2.3.2 Classificação

A classificação é uma das tarefas mais utilizadas para minerar dados, por ser a atividade cognitiva humana mais efetuada no auxílio à compreensão do ambiente. A todo momento estamos classificando e criando relações em função das situações que se

apresentam. A forma mais natural de raciocínio humano é caracterizar, ou seja, classificar elementos com base em condições relacionadas ao mesmo.

Nessa tarefa o objetivo é estabelecer para cada caso (registro, objeto ou instância) uma classe com base nos valores de alguns atributos (chamados atributos preditores, ou de predição). No contexto da tarefa de classificação o conhecimento descoberto também é expresso na forma de regras do tipo SE-ENTÃO.

SE <condição> ENTÃO <classe>

A parte antecedente da regra normalmente contém um conjunto de conjunções ligadas por operadores lógicos de conjunção, havendo algoritmos que implementam também o operador de disjunção.

Cada condição de uma regra pode ser entendida como um termo, sendo que o antecedente da regra é uma conjunção de termos na forma SE termo1 E termo2 E...

Cada termo é uma tripla <atributo, operador, valor> tal como:

<idade = 25>

A parte conseqüente da regra (parte ENTÃO) especifica a classe predita por meio do atributo de classe, também denominado atributo meta, para os casos cujos atributos preditores satisfazem todos os termos especificados na parte antecedente da regra.

Do ponto de vista de DM, este tipo de representação de conhecimento tem a vantagem de ser intuitivamente compreensível para o usuário, contanto que o número de regras descobertas e o número de condições no antecedente da regra não sejam muito grandes.

Para compreensão da tarefa de classificação vejamos o seguinte exemplo obtido de Guimarães (2004).

Uma companhia de cartões de crédito deseja identificar o perfil dos clientes que costumam pagar as faturas de forma parcelada. Para tanto dispõe de uma base de dados contendo o sexo, a renda em uma unidade monetária, a idade e o fato de o cliente ter ou não realizado ao menos um pagamento parcelado no último ano (Quadro 2).

Quadro 2. Base de dados de clientes para o problema proposto

Sexo	Idade	Renda (UM)	Pagamento Parcelado
M	25	até 1000	Sim
M	21	1001 a 3000	Sim
F	23	até 1000	Sim
F	34	1001 a 3000	Sim
F	30	até 1000	Não
M	21	acima de 3000	Não
M	20	acima de 3000	Não
F	18	acima de 3000	Não
F	34	até 1000	Não
M	55	até 1000	Não

A tarefa para o problema proposto consiste em classificar os clientes em duas categorias: aqueles que pagam sua fatura de forma parcelada, e aqueles que pagam sem parcelamento.

Primeiramente, deve-se caracterizar os atributos. Nesse caso:

Atributos de predição: Sexo, Renda (UM) e Idade
Atributo meta (ou categórico): Pagamento Parcelado

Um classificador é uma função que, dada uma coleção de dados D, composta por n objetos, cada um deles descrito por um conjunto de atributos A, produz para cada elemento de D um valor de classe C, com base nos valores internos dos atributos.

Os programas que realizam a tarefa de classificação são chamados indutores, e necessitam dividir a base de dados em dois conjuntos: de treinamento e de teste. A Figura 4 apresenta um esquema de programas indutores.

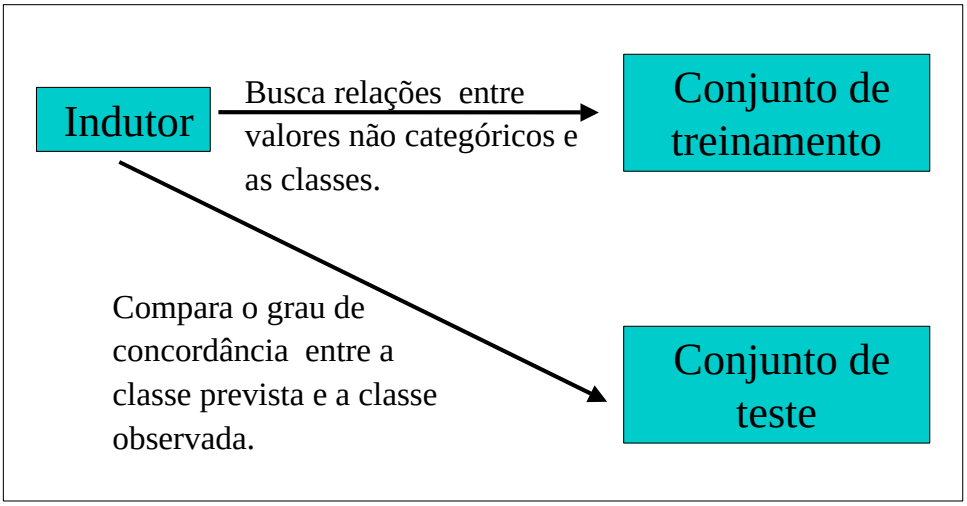


Figura 4. Esquema de programas indutores

As regras de classificação descobertas por um indutor, a partir dos dados de entrada são:

SE (Renda = “acima de 3000”) ENTÃO (Pagamento Parcelado = “não”)

SE (Renda = “1001 a 3000”) ENTÃO (Pagamento Parcelado = “sim”)

SE (Renda = “até 1000” E Idade ≤ 25) ENTÃO (Pagamento Parcelado = “sim”)

SE (Renda = “até 1000” E Idade > 25) ENTÃO (Pagamento Parcelado = “não”)

Com base nas regras obtidas a companhia de cartões verifica o perfil de seus clientes descobrindo, por exemplo, que para clientes com renda baixa, a decisão de parcelar ou não suas faturas é diferente, em função de suas idades.

2.3.3 Agrupamento

A tarefa de agrupamento, também conhecida como clusterização, tem como objetivo estabelecer para uma amostra de indivíduos, com base em algum critério pré-

determinado, grupos mutuamente exclusivos, em função de similaridades entre tais indivíduos. Os grupos resultantes devem exibir uma alta homegeneidade interna (dentro do grupo) e alta heterogeneidade externa (entre os grupos). Quando representados graficamente os objetos pertencentes a cada grupo estarão todos juntos.

O agrupamento é uma tarefa descritiva onde busca-se identificar um conjunto finito de categorias para descrever os dados (Titterington et al., 1985).

Segundo Carvalho (2001) o problema que a análise de agrupamento pretende resolver é: dada uma amostra de *n* objetos (ou indivíduos), cada um deles medido segundo *p* variáveis, procurar um esquema de classificação que agrupe os objetos em *g* grupos em função de suas similaridades, devendo ser determinados também o número e as características desses grupos.

A Figura 5 apresenta um gráfico que sugere um processo de agrupamento resultando em quatro grupos gerados. Pode-se notar que em cada grupo os elementos não são todos iguais, mas sim, semelhantes.

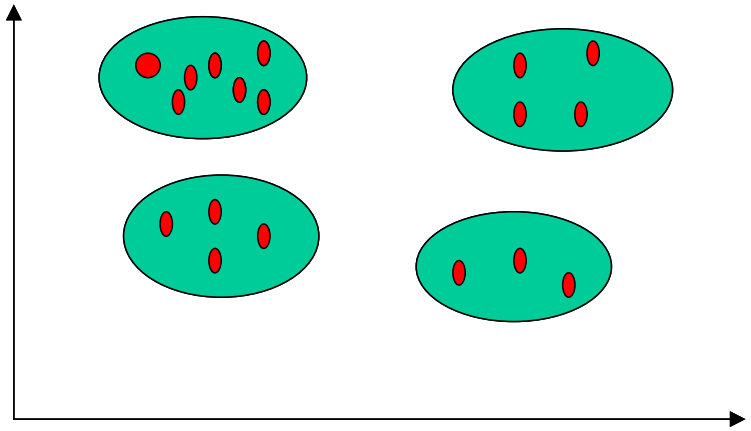


Figura 5. Agrupamento de indivíduos em função de alguma semelhança

Exemplos de aplicações de agrupamento no contexto da mineração de dados incluem: descoberta de sub-populações homogêneas para clientes em uma base de dados de marketing, identificação de regiões homogêneas em um estado em função das características de solo e de clima para definição de culturas adequadas a serem plantadas em cada grupo identificado.

A dificuldade do agrupamento reside no fato de que pode não haver classes, ou seja, os dados distribuem-se de forma eqüitativa em todo o espaço possível, não caracterizando nenhuma categoria. É fácil perceber isso se considerarmos que dadas as classes animal, vegetal e mineral, é trivial dizermos à qual delas um objeto pertence. Porém, de posse de uma massa de dados sobre o consumo no Brasil, determinar quantas classes ou padrões de comportamento consumista existem é algo bem diferente e mais complexo. Por exemplo, agrupar sintomas pode gerar classes que não representem nenhuma doença explicitamente, uma vez que doenças diferentes podem possuir os mesmos sintomas.

Após o agrupamento pode-se aplicar métodos de classificação e sumarização para descobrir regras de classificação (que discriminem registros de diferentes classes) e regras de sumarização (que produzam descrições características de cada classe).

2.4 Seleção da tarefa de mineração de dados

Uma das grandes dificuldades do processo de KDD é identificar qual tarefa de mineração é mais indicada em função dos tipos de dados e do objetivo a ser atingido com base nos padrões descobertos. É de fundamental importância que no momento de seleção da tarefa de DM se tenha um conhecimento suficiente tanto da teoria da mineração de dados, como do domínio da aplicação, para que se possa escolher a tarefa mais adequada, ou seja, aquela que possa vir a gerar o tipo de informação que o usuário espera.

Freitas (2000) estabeleceu uma comparação entre a tarefa de classificação e as de associação e agrupamento, a qual está detalhada a seguir.

2.4.1 Associação x Classificação x Agrupamento

A associação aplica-se a problemas simétricos (todos os itens podem aparecer no antecedente ou no conseqüente de uma regra). A qualidade de uma regra é avaliada por fatores de confiança e suporte definidos pelo usuário, e a definição do problema é determinística, uma vez que o sistema precisa achar todas as regras com o suporte e confiança maior ou igual a esses limiares pré-definidos. O desafio mais citado na literatura para a tarefa de associação é projetar algoritmos eficientes.

A tarefa de classificação é adequada para um problema assimétrico (um único atributo meta a ser previsto, dados demais atributos). As regras são avaliadas em dados de teste não vistos durante treinamento (prever o futuro), e a qualidade de uma regra é muito mais difícil de avaliar. Logo, não é muito claro quais regras deveriam ser descobertas pelo sistema. Em conseqüência do processo de indução envolvido nesse tipo de tarefa, o problema é considerado não-determinístico. Em relação à pesquisas, a eficiência ainda é importante, mas o desafio principal é projetar algoritmos eficazes.

Na classificação há um único atributo meta, e os demais atributos são previsores, sendo que do problema consiste justamente em determinar automaticamente a importância desses atributos previsores. Há medidas objetivas para medir a qualidade da classificação (ex. taxa de acerto), sendo uma tarefa usada principalmente para previsão.

No agrupamento não há um atributo especial, sendo que a importância de cada atributo é geralmente considerada equivalente à dos demais. Em relação à qualidade, é difícil de medi-la e, finalmente, é uma tarefa usada, principalmente, para exploração e sumarização de dados.

2.5 Inteligência Computacional em mineração de dados

Existem inúmeras tecnologias de Inteligência Computacional (IC) aplicadas em mineração de dados, sendo algumas delas: Redes Neurais, Indução de Regras, Árvores de Decisão, Algoritmos Genéticos e Colônia de Formigas. Cada tipo de tecnologia

tem suas próprias vantagens e desvantagens, do mesmo modo que nenhuma ferramenta consegue atender a todas as necessidades em todas as aplicações. Abordaremos a seguir algumas delas.

2.5.1 Árvores de Decisão

As Árvores de Decisão (AD) são uma evolução das técnicas que apareceram durante o desenvolvimento das disciplinas de Aprendizado de Máquina. Segundo Ye (2003) as Árvores de Decisão correspondem a um dos mais populares modelos de mineração de dados e têm sido aplicadas em uma grande variedade de problemas. Elas evoluíram a partir de uma abordagem de análise denominada Detecção de Interação Automática, desenvolvida na Universidade de Michigan. Essa análise trabalha testando automaticamente todos os valores do dado para identificar aqueles que são fortemente associados com os itens de saída selecionados para exame. Os valores que são encontrados com forte associação são os prognósticos chave ou fatores explicativos, usualmente chamados de regras sobre o dado.

Muitos pesquisadores de MD consideram Ross Quinlan, da Universidade de Sydney, Austrália, como sendo o “pai das Árvores de Decisão”. Quinlan (1993) apresentou um novo algoritmo chamado ID3 e suas evoluções (ID4, ID6, C 4.5, See 5), os quais são bem adaptados para usar em conjunto com as AD, na medida em que produzem regras ordenadas pela importância. Essas regras são, então, usadas para produzir um modelo de Árvore de Decisão dos fatos que afetam os itens de saída.

Segundo Andreatto (2002) novas fórmulas em Árvores de Decisão como a Gini, um índice computacional inventado por Ron Bryman, mostram-se adequadas e oferecem uma crescente velocidade de processamento assim como mais habilidades para processar números e textos concorrentemente.

Para realizar a tarefa de classificação, uma ferramenta de Árvore de Decisão pede ao usuário para definir na sua base de dados, qual será o atributo meta, o qual se apresenta nas folhas da árvore gerada, e mostra o único e mais importante atributo preditor correlacionado com o atributo meta como o primeiro ramo (nó) da AD. Os outros atributos

preditores são subseqüentemente classificados como nós do(os) nó(s) anterior(es). Isso significa que o usuário pode rapidamente ver qual o atributo preditor que mais direciona o seu atributo meta. A Figura 6 apresenta a estrutura de uma AD.

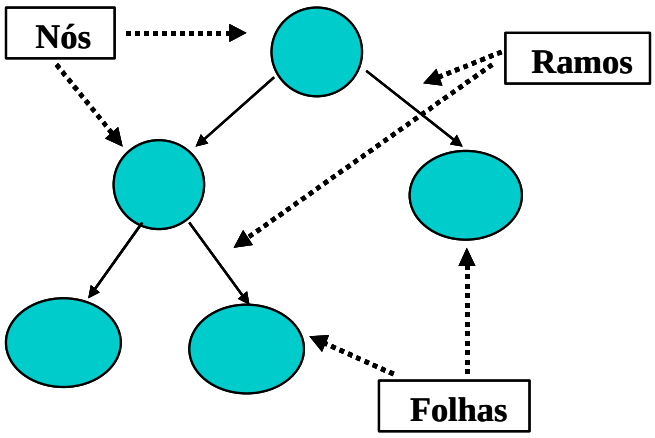


Figura 6. Estrutura de uma Árvore de Decisão

Uma boa ferramenta de AD deve permitir que o usuário explore a árvore de acordo com a sua vontade, o que pode ser feito por meio de parâmetros de gerenciamento do algoritmo, que devem ser solicitados ao usuário, uma vez que para cada combinação de parâmetros para um mesmo conjunto de dados, a árvore de decisão gerada será diferente.

A figura 7 apresenta um exemplo de árvore de decisão gerada para o problema de classificação proposto na seção 2.3.2.

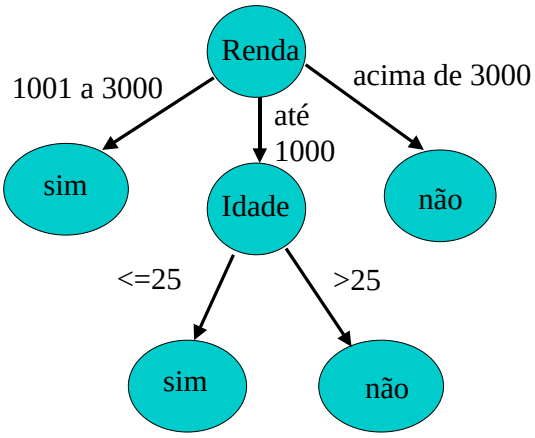


Figura 7. Árvore de Decisão gerada para um problema de classificação

As Árvores de Decisão são, quase sempre, usadas em conjunto com a tecnologia de Indução de Regras, mas são únicas no sentido de apresentar os resultados da Indução de Regras num formato com priorização. Então, a regra mais importante é apresentada na árvore como o primeiro nó, e as regras menos relevantes são mostradas nos nós subseqüentes. As principais vantagens das AD são a facilidade de manipulação e de compreensão por parte do usuário.

As regras no formato SE-ENTÃO são obtidas percorrendo todos os caminhos possíveis na árvore, sendo que cada regra corresponde a um caminho do nó raiz até uma determinada folha. As regras geradas, a partir da Árvore de Decisão apresentada na Figura 7, são:

SE (Renda = “1001 a 3000”) ENTÃO (Pagamento Parcelado = “sim”)
SE (Renda = “até 1000” E Idade ≤ 25) ENTÃO (Pagamento Parcelado = “sim”)
SE (Renda = “até 1000” E Idade > 25) ENTÃO (Pagamento Parcelado = “não”)
SE (Renda = “acima de 3000”) ENTÃO (Pagamento Parcelado = “não”)

Kantardzic (2003) afirma que embora os modelos de Árvores de Decisão sejam relativamente simples, compreensíveis e rápidos para serem gerados, tornando-os muito atrativos, existem algumas desvantagens e limitações dessa abordagem. A principal desvantagem está no fato de que as AD dividem o espaço de exemplos em regiões, sendo que cada região é associada a uma classe correspondente. As Árvores de Decisão são então construídas por refinamentos sucessivos. Dessa forma bases de dados com grande quantidade de atributos e com muita variabilidade nos valores dos mesmos tendem a gerar árvores extremamente complexas com grande número de regras e ainda com grande margem de erros.

2.5.2. Redes Neurais Artificiais

Redes Neurais Artificiais (RNA) consistem em uma técnica de IC, no contexto de Aprendizado de Máquina, sendo consideradas um modelo baseado na organização

dos neurônios no cérebro humano. O cérebro humano consiste de um grande número de neurônios, conectados entre si por meio de sinapses. Tipicamente uma rede neural consiste de um conjunto de nodos que recebem sinal de entrada, outro conjunto de nodos que geram sinal de saída, e vários níveis intermediários que contém nodos intermediários (Adriaans & Zantigue, 1996).

Um histórico resumido sobre Redes Neurais Artificiais deve começar por três das mais importantes publicações iniciais, desenvolvidas por: McCulloch e Pitts (1943), Hebb (1949), e Rosenblatt (1958). Estas publicações introduziram o primeiro modelo de redes neurais simulando "máquinas", o modelo básico de rede de auto-organização, e o modelo Perceptron de aprendizado supervisionado, respectivamente. A Figura 8 apresenta o neurônio artificial proposto por McCulloch & Pitts (1943).

Alguns históricos sobre a área costumam "pular" os anos 60 e 70 e apontar um reinício da área com a publicação dos trabalhos de Hopfield (1982) relatando a utilização de redes simétricas para otimização e de Rumelhart, Hinton e Williams que introduziram o poderoso método Backpropagation.

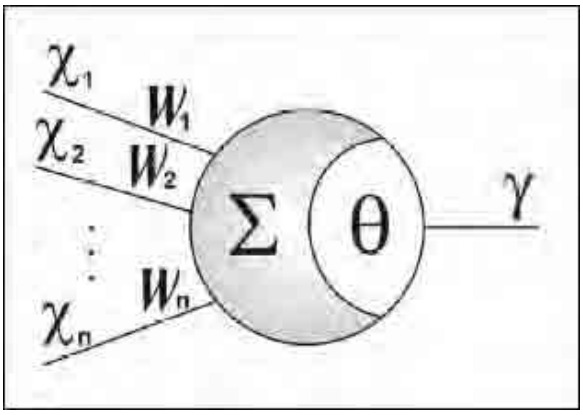


Figura 8. Neurônio artificial proposto por McCulloch & Pitts (1943)

O interesse dos pesquisadores em uma forma de computação que se inspirasse no cérebro (Figura 9) veio do fato deste último possuir características altamente desejáveis em qualquer sistema artificial, tais como: robustez e tolerância à falhas; flexibilidade; capacidade para lidar com informações probabilísticas, contendo ruídos, ou

inconsistentes; processamento paralelo; arquitetura compacta e com pouca dissipação de energia. Além destas características, a capacidade de aprendizado, generalização, e ainda associação, motivou o interesse por este tipo de computação alternativa.

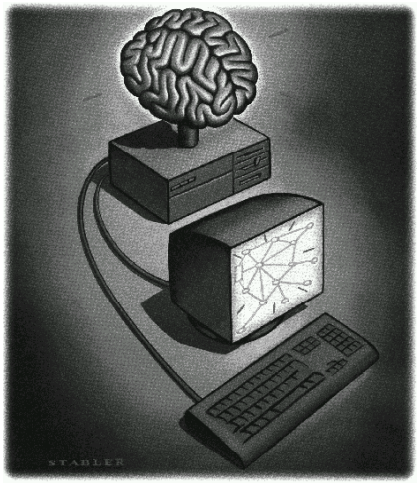


Figura 9. A inspiração do cérebro humano na computação

As Redes Neurais Artificiais possuem a capacidade de aprender por meio de exemplos e fazer interpolações e extrapolações do que aprenderam. Um conjunto de procedimentos bem-definidos para adaptar os parâmetros de uma rede neural para que a mesma possa aprender uma determinada função é chamado de algoritmo de aprendizado (Braga et al., 2000).

Essa tecnologia oferece boa capacidade de mineração, mas é também muito difícil de entender. As Redes Neurais tentam construir representações internas de modelos ou padrões encontrados nos dados, mas essas representações não são apresentadas para o usuário. Com elas, o processo de descoberta de padrões é tratado pelos programas de mineração de dados dentro de um processo “caixa-preta”, tornando difícil o entendimento de como os resultados foram obtidos.

A função básica de cada neurônio é:

- Avaliar valores de entrada
- Calcular o total para valores de entrada combinados
- Comparar o total com um valor limiar

- Determinar qual será a saída.

Enquanto a operação de cada neurônio é razoavelmente simples, procedimentos complexos podem ser criados pela conexão de um conjunto de neurônios. Tipicamente, as entradas dos neurônios são ligadas a uma camada intermediária (ou várias camadas intermediárias) que é então conectada com a camada de saída.

Para construir um modelo neural, primeiramente "adestra-se" a rede em um *dataset* (conjunto de dados) de treinamento e então usa-se a rede já treinada para fazer predições. Pode-se, às vezes, monitorar o *dataset* durante a fase de treinamento para checar seu progresso.

Cada neurônio geralmente tem um conjunto de pesos que determina como o neurônio avalia a combinação dos sinais de entrada. A entrada para um neurônio pode ser positiva ou negativa. O aprendizado se faz pela modificação dos pesos usados pelo neurônio em acordo com a classificação de erros que foi feita pela rede como um todo.

As redes neurais artificiais têm se mostrado uma poderosa ferramenta na previsão de séries temporais. Sua habilidade em extrair complicadas relações não lineares a partir dos dados de entrada com ruídos tem produzido resultados interessantes, muitas vezes melhores que os obtidos por procedimentos estatísticos convencionais.

2.5.3. Algoritmos Genéticos

Os Algoritmos Genéticos estão inseridos na área da Computação Evolucionária (Bäck et al., 2000), que é constituída também pela Programação Evolutiva (PE), Estratégias de Evolução (EE), Sistemas de Classificadores (SC) e Programação Genética (PG).

Os AGs realizam a tarefa de classificação de dados baseados na analogia com os processos de seleção natural e genética evolucionária (Goldberg, 1989). A essência do método consiste em manter uma população de indivíduos (cromossomos), os quais representam possíveis soluções para um problema. A melhor solução é obtida através de um processo de seleção competitiva (Herrera et al., 1998).

Os AGs foram idealizados por Holland (1975) tendo sido fundamentados nos princípios da genética populacional. Segundo esse princípio, a variabilidade entre indivíduos, em uma população de organismos que se reproduzem sexualmente, é produzida pela mutação e pela recombinação genética.

A base para o desenvolvimento dos AGs baseiam-se na Teoria da Seleção Natural descrita no livro “On the Origin of Species by Means of Natural Selection” publicado em 1859 pelos naturalistas ingleses Charles Darwin, o qual contém também contribuições de Lamark e Malthus sobre o processo de classificação biológica (Darwin, 1859).

Darwin (1859) apresentou sua teoria de evolução das espécies através de seleção natural, afirmando que variações (mutações) estão presentes em todas as espécies. Segundo ele, a evolução ocorre devido a uma força chamada de seleção natural que "escolhe" os indivíduos melhores adaptados ao ambiente. Em um ambiente mutável alguns indivíduos serão melhores que os originais e serão preservados, e novas espécies serão criadas desta maneira.

Segundo Bäck et al. (2000) os Algoritmos Genéticos compõem uma estratégia de busca e otimização que tem se mostrado extremamente útil na solução de problemas complexos que possuam as seguintes características:

a) Existe a possibilidade da definição de uma função real que mensura a adequação de cada possível solução, sendo esta função chamada de função de *fitness*;
b) A estratégia de solução pode ser visualizada como uma questão de otimização da função de *fitness*;
c) O espaço de busca é multimodal, isto é, existem vários máximos locais;
d) Existe um conjunto de restrições que deve ser observado pelas soluções;
e) Soluções aproximadas são aceitáveis.

Dessa forma um problema de mineração de dados que possa ser definido como a maximização ou minimização de alguma função pode ser, em princípio, resolvido com um Algoritmo Genético. Trata-se de uma heurística já antiga, mas cujo estudo se faz mais intenso há poucos anos, prometendo ser um processo de muita utilização e comum no futuro (Carvalho, 2001).

Em Algoritmos Genéticos, qualquer possível solução é representada

por meio de uma seqüência de símbolos de um alfabeto finito. Inicialmente esta representação era feita com base em um alfabeto binário {0,1}; atualmente outros alfabetos podem ser utilizados. A cada uma dessas seqüências, chamadas indivíduos, é atribuído um valor ou medida de *fitness* (Goldberg, 1989).

A medida de *fitness* denotada por F pode ser computada de diferentes formas. Quando usado para realizar a tarefa de classificação em mineração de dados, o AG normalmente apresenta uma função de *fitness* associada à sensibilidade e/ou especificidade da regra, com variações, sendo:

$$\textbf{sensibilidade} = \frac{VP}{VP + FN}$$

$$\textbf{especificidade} = \frac{VN}{FP + VN}$$

Onde:

VP (verdadeiro positivo): é o número de casos cobertos pela regra que tem a classe predita pela regra.

FP (falso positivo): é o número de casos cobertos pela regra que tem uma classe diferente da classe predita pela regra.

FN (falso negativo): é o número de casos que não estão cobertos pela regra, mas tem a classe predita pela regra.

VN (verdadeiro negativo): é o número de casos que não estão cobertos pela regra e não tem a classe predita pela regra.

Os valores de F estão na faixa $0 \leq F \leq 1$ e, quanto maior o valor de F, melhor a qualidade da regra.

Definida a forma de representação, o processo de solução segue da seguinte forma: uma população inicial de indivíduos (possíveis soluções) é gerada de maneira aleatória e a medida de *fitness* de cada indivíduo é calculada. Em seguida, usando um processo evolucionário, uma nova população, com o mesmo número de indivíduos, é gerada a partir da

antecessora. Esse processo é repetido um determinado número de vezes ou até a satisfação de um critério de parada. Os operadores que podem ser aplicados aos indivíduos para a geração das novas populações são: seleção, mutação, e *crossover* (recombinação).

O uso de Algoritmos Genéticos tem causado impacto (Guimarães et al., 2003), sendo que suas principais vantagens sobre outras técnicas no processo de descoberta de conhecimento são: grande capacidade de trabalhar com dados imprecisos (Freitas & Lavington, 1998), a tratabilidade em termos de custo computacional (Langley, 1996), o ajuste fino de parâmetros de acordo com o domínio (Mitchell, 1997), e a possibilidade de processamento em paralelo e distribuição de carga de processamento (Freitas & Lavington, 1998).

2.5.4 Otimização de Colônia de Formigas

Segundo Parpinelli et al. (2002) o desenvolvimento de algoritmos de otimização de colônias de formigas, também conhecidos como ACO (Ant Colony Optimization), em mineração de dados, consiste em uma área de pesquisa promissora, devido ao fato de que algoritmos de ACO envolvem agentes simples (formigas) que cooperam entre si para alcançar um comportamento emergente, unificado para o sistema como um todo, produzindo um sistema robusto capaz de achar soluções de alta qualidade para problemas com um espaço de busca grande.

Um algoritmo ACO pode ser entendido essencialmente como um sistema baseado em agentes que simulam o comportamento natural de formigas, inclusive mecanismos de cooperação e adaptação. Em Dorigo & Di Caro (1999) foi proposto o uso deste tipo de sistema para resolver problemas de otimização combinatorial, tendo se mostrado robusto e versátil quando aplicado em diferentes problemas.

No contexto de descoberta de regras, um algoritmo de ACO tem a habilidade de executar uma procura flexível e robusta para uma boa combinação de condições lógicas envolvendo valores do atributo preditor.

O uso de algoritmos ACO para descoberta de regras de classificação, no contexto de mineração de dados, ainda é uma área de pesquisa pouco explorada. Um algoritmo de colônia de formigas conhecido, desenvolvido para mineração de dados, é um algoritmo para a tarefa de agrupamento (Monmarché, 1999). Além desse, Parpinelli et al. (2002) desenvolveram o algoritmo ACO, denominado Ant-Miner, para realizar a tarefa de classificação, o qual será abordado em mais detalhes ainda nessa seção.

Algoritmos de ACO estão baseados nas seguintes idéias:

- Cada caminho seguido por uma formiga é associado com uma solução candidata para um determinado problema.
- Quando uma formiga seguir um caminho, a quantidade de feromônio depositada naquele caminho é proporcional à qualidade da solução candidata correspondente para o problema projetado.
- Quando uma formiga tem que escolher entre dois ou mais caminhos, existe uma maior probabilidade de aquele com uma quantidade maior de feromônio ser escolhido pela formiga.

Como resultado, as formigas convergem eventualmente para um caminho curto, que se espera seja o ótimo ou uma solução próxima à ótima para o problema designado.

Em essência, o projeto de um algoritmo de ACO envolve a especificação de:

- Uma representação apropriada do problema, que permite às formigas a construção e modificação de soluções pelo uso de uma regra de transição probabilística, baseado na quantidade de feromônio na trilha.
- Um método para obrigar a construção de soluções válidas, ou seja, soluções que são válidas no mundo real associado à definição de problema.
- Uma função heurística dependente do problema (h).

- Uma regra para atualizar o feromônio, que especifica como modificá-lo.
- Uma regra de transição probabilística baseada no valor da função heurística (h) e nos conteúdos do feromônio na trilha (t).

Formigas artificiais têm várias características semelhantes às formigas reais:

- Formigas artificiais têm uma preferência probabilística para caminhos com uma quantia maior de feromônio.
- Caminhos mais curtos tendem a ter taxas maiores de crescimento baseadas na proporção do feromônio no caminho.
- As formigas usam um sistema de comunicação indireto baseado na quantia de feromônio depositada em cada caminho.

Da mesma forma que no AG, a qualidade de uma regra em ACO é denotada pela associação da sensibilidade e da especificidade da regra (ver seção 2.5.3), sendo computada pela seguinte fórmula:

$$\mathbf{Q = sensibilidade * especificidade}$$

A atualização do feromônio para um $termo_{ij}$ particular, para todo os termos *ij*, é executada de acordo com a equação

$$\mathbf{\tau_{ij}\,(t+1) = \tau_{ij}\,(t) + \tau_{ij}\,(t) * Q,\ \ \forall\ ij \in R}$$

Onde:

R: é o conjunto de condições que acontecem na regra construída pela formiga na iteração *t*.

Portanto, para todo $termo_{ij}$ que ocorre na regra encontrada pela formiga atual, o valor de feromônio é acrescido por uma fração da quantia atual de feromônio, e esta fração é determinada por Q.

Uma abordagem interessante sobre o uso de ACO em mineração de dados é apresentada no algoritmo Ant-Miner, desenvolvido por Parpinelli et al. (2002) e citado anteriormente. A seguir apresentaremos algumas características desse algoritmo.

No Ant-Miner o problema projetado é a descoberta de regras de classificação, sendo que cada regra tem a forma:

SE <termo1> E <termo2> E...> ENTÃO <classe>.

Cada termo é uma tripla <atributo, operador, valor>, onde "valor" é um valor que pertence ao domínio de atributo. O elemento operador na tripla é um operador relacional. A versão do algoritmo Ant-Miner, publicada em Parpinelli et al. (2002), comporta somente atributos categóricos, de forma que o operador na tripla é sempre o "=".

Uma descrição de alto-nível de Ant-Miner indica que o ACO busca descobrir uma lista de regras de classificação que cobrem todos, ou quase todos, os casos de treinamento. No princípio, a lista de regras descobertas está vazia e o conjunto de treinamento consiste em todos os casos de treinamento.

A cada iteração do algoritmo, correspondendo a várias execuções da estrutura REPITA-ATÉ, uma regra de classificação é descoberta. Esta regra é acrescentada à lista de regras descobertas, e os casos de treinamento que são corretamente cobertos por esta regra (i.e., casos satisfazendo o antecedente de regra e tendo a classe predita pela regra conseqüente) são retirados do conjunto de treinamento.

Este processo é executado iterativamente enquanto o número de casos de treinamento descobertos é maior que um limiar especificado pelo usuário.

2.5.5 Sistemas Híbridos

Um problema comum em KDD é a presença de ruídos nos dados que são minerados. Redes Neurais são robustas e têm uma boa tolerância a ruídos tornando-se satisfatórias na mineração de dados com essas características. Porém, elas têm a bem conhecida desvantagem de não descobrir nenhuma regra de alto nível que possa ser usada como um suporte para a tomada de decisão humana, por representar seu conhecimento na forma de pesos numéricos e interconexões, que não são compreensíveis pelo usuário.

Santos et al. (2002) apresentaram um método híbrido para extrair regras precisas e compreensíveis de Redes Neurais. O método proposto usa um Algoritmo Genético para achar uma boa topologia de Rede Neural. Esta topologia é então passada a um algoritmo de extração de regras, e a qualidade das regras extraídas é retornada para o AG.

O AG para evoluir a topologia da Rede Neural foi implementado utilizando a ferramenta ENZO (Braun & Ragg, 1995), à qual foi acoplado um módulo de extração de regras baseado no algoritmo RX (Lu & Setiono, 1995). No AG cada indivíduo da população corresponde a uma topologia de rede treinada pelo algoritmo RPROP (Zell et al., 1995). O sistema proposto foi avaliado em três bases de dados de domínio público e os resultados mostraram que a abordagem é válida, tendo gerado regras com um nível de acurácia maior do que o obtido quando empregado o algoritmo C4.5 (Quinlan, 1993) sobre os mesmos dados.

Outro exemplo de sistema híbrido está associado à mineração de dados de uso da WEB, abordado a seguir.

O rápido crescimento de e-comércio tem colocado a comunidade empresarial e os clientes frente a uma nova situação. Devido à intensa competição e pela possibilidade do cliente escolher entre várias alternativas, a comunidade empresarial percebeu a necessidade de estratégias inteligentes de marketing e de gerenciamento de relacionamentos.

A mineração de dados de uso da WEB tenta descobrir conhecimento útil dos dados secundários obtidos das interações dos usuários com a WEB. Esse tipo de aplicação de mineração de dados tem sido muito útil, entre outros, para:

- Gerenciamento efetivo de sites WEB.

- Criação de sites WEB adaptativos.
- Negócios e serviços de suporte.
- Personalização.
- Análise de fluxo de tráfego na rede.

O estudo do comportamento de colônias de formigas e suas capacidades de auto-organização são de interesse para a recuperação/gerenciamento do conhecimento e nos sistemas de suporte a decisão, por fornecer modelos de organização adaptativa distribuída. Esses modelos são úteis para resolver problemas difíceis em otimização, classificação, e controle distribuído.

Com base no exposto acima, Abraham & Ramos (2003) propuseram um algoritmo de agrupamento de formigas para descobrir padrões de uso da WEB e uma abordagem de programação genética linear para analisar as tendências de visita a páginas da WEB.

Resultados empíricos mostraram que o agrupamento de colônia de formigas comporta-se bem quando comparado a um mapa auto-organizável (para agrupar padrões de uso da WEB), embora a precisão de desempenho não seja tão eficiente quando comparado a abordagem de agrupamento evolucionária-fuzzy (i-miner) (Abraham, 2003).

2.6 Barreiras da mineração de dados

As principais barreiras consideradas ao uso da mineração de dados são: alto custo das soluções, necessidade de grande volume de dados armazenados em servidores, pouca amigabilidade das ferramentas de DM para pessoas que não sejam altamente especializadas e alto custo das soluções. A seguir, algumas características dessas barreiras são abordadas, como fonte de motivação para possíveis temas de pesquisa na área.

- **Necessidade de grande volume de dados**

O maior obstáculo ao DM no passado foi a necessidade de armazenar e administrar milhares de dados e/ou servidores. Isso por si só já dificultava bastante o crescimento do mercado de mineração de dados.

Embora a maioria dos fornecedores dessa tecnologia continue insistindo no discurso de que o DM requer terabytes de dados e poderosos servidores, soluções mais acessíveis já têm aparecido no mercado e criado condições para um maior uso da tecnologia.

É aceita pelo mercado a afirmação de que 80% do valor de um determinado grupo de dados pode ser encontrado em 20% dos dados que o compõem. É evidente então que os fornecedores farão todo o possível para encontrar esses 20% e minerá-los.

As ferramentas que procuram tornar grupos de dados pessoalmente manejáveis fornecem aos usuários a possibilidade de minerar porções de dados em seu próprio computador, o que pode, efetivamente, contornar essa barreira. Embora não tenham a intenção de substituir aplicações que necessitam de grandes volumes de dados, essas ferramentas podem prover um acesso alternativo que pode também ser usado em conjunto com as ferramentas “pesadas”.

- **Complexidade das Ferramentas**

Mesmo com essa nova geração de ferramentas computacionais que permitem a mineração de dados, uma outra barreira ainda permanece: a grande maioria das ferramentas ainda continua incompreensível para os usuários comuns.

Considerando a complexidade existente no processo de mineração de dados na busca de novos conhecimentos, a primeira preocupação dos pesquisadores dessa área acaba sendo a lógica embutida e necessária para gerar bons resultados. Dessa forma, para a execução do sistema normalmente são requeridos inúmeros parâmetros para os quais muitas vezes não existe um valor *default* ou uma explicação detalhada dos valores mais adequados para esses parâmetros. Com isso torna-se realmente muito complexo para o usuário fazer uso da ferramenta.

Outro aspecto importante é o fato de que muitas ferramentas ainda utilizam a abordagem “caixa-preta”, não permitindo que se saiba como alcançaram os seus resultados. Isso significa que a mineração de dados ainda tem que ser feita por profissionais da área de sistemas de informação, a quem os usuários têm que submeter as suas solicitações, esperar por dias ou semanas enquanto os dados são processados, para então receberem e examinarem a saída sumarizada.

Felizmente, existem técnicas de mineração de dados compreensíveis e fáceis de serem empregadas, como as Árvores de Decisão. Essas técnicas permitem a um usuário com conhecimento básico sobre computadores que as utilize compreendendo o processo. Contudo, o poder dessas ferramentas não chega nem perto daquelas que utilizam técnicas mais sofisticadas. A medida em que as pesquisas avançam no uso de técnicas mais avançadas como Algoritmos Genéticos, Programação Genética e Sistemas Híbridos, a tendência é que se desenvolvam sistemas mais compreensíveis e mais fáceis de serem usados pelo usuário mesmo que a lógica embutida no mesmo seja complexa.

- **Altos Custos**

O alto custo da maioria das ferramentas faz com que fique difícil a disseminação das mesmas entre as corporações. Uma simples operação matemática mostra que um projeto, com o custo de R$ 75.000,00 (US$ 30.000) por usuário, pode atender não mais que um seleto grupo.

Certos fornecedores dessas ferramentas têm introduzido produtos com custo mais baixo, mas mesmo assim o preço continua limitando o uso em larga escala.

Evidentemente, os custos por usuário precisam ser reduzidos antes que os benefícios desta tecnologia possam atingir a massa.

2.7 Viabilidade do desenvolvimento de um sistema de mineração de dados

O estudo nesse capítulo procurou avaliar a viabilidade do uso das técnicas de mineração de dados nos mais diferentes segmentos. Analisando as barreiras do

DM é possível também compreender a importância do desenvolvimento de sistemas de mineração de dados que tornem mais fácil e acessível o seu uso. O sistema pode ser utilizado como uma ferramenta de apoio na busca de conhecimento novo e mais facilmente compreensível, visto que os resultados gerados pelo mesmo podem levar a uma nova interpretação do problema em estudo levando o usuário a repensar alguns conceitos previamente estabelecidos.

Existem diversos softwares de mineração de dados, tanto de uso comercial quanto acadêmico, porém os resultados gerados pelos mesmos nem sempre atendem às expectativas, visto que muitos dos softwares são desenvolvidos utilizando técnicas de mineração de dados bem conhecidas, como Redes Neurais e Árvores de Classificação. Embora essas técnicas sejam bem aceitas, o fato de a estrutura de busca ser local em Árvores de Classificação implica em resultados ineficientes em inúmeros casos, o que pode passar desapercebido para o usuário final do sistema. Já em relação às Redes Neurais o fato de a forma de obtenção dos resultados ser ainda considerada como uma “caixa preta” cria restrições quanto ao seu uso.

Detalhes sobre os softwares disponíveis para mineração de dados podem ser obtidos em www.kdnuggets.com. Não foi identificado até o momento nenhum software de mineração de dados contínuos referentes a fatores físico-químicos do solo e da água tendo sido desenvolvido usando a técnica de Algoritmos Genéticos, permitindo intensa parametrização, e que possa ser executado em ambiente paralelo.

Embora a proposta da mineração de dados seja a descoberta de conhecimentos novos, é importante salientar que muitos dos resultados consistem em informação já conhecida pelo usuário, apresentando baixo grau de novidade. Daí a necessidade de se desenvolver métodos de avaliação do grau de interesse das regras resultantes da mineração de dados, o que ocorre em fase posterior à mineração. Existem alguns estudos sobre a incorporação de medidas de grau de interesse de regras dentro do próprio processo de mineração, porém ainda não existe um consenso sobre a fase em que essa avaliação deve ser realizada, ou seja, durante ou após a mineração. Detalhes sobre essa discussão podem ser vistos em Freitas (2002).

Em relação à questão do grau de interesse é importante observar que o sistema de mineração de dados deve apresentar a facilidade de armazenar seus resultados em

um banco de dados de forma tal que facilite sua análise e manipulação posterior na busca das efetivamente novas e melhores regras. Essa facilidade não tem sido observada nos softwares de mineração de dados, nos quais as regras resultantes normalmente são armazenadas em forma de arquivos textuais.

3 IMPORTÂNCIA DA ANÁLISE DOS FATORES FÍSICO-QUÍMICOS DO SOLO E DA ÁGUA

Esse capítulo apresenta uma visão geral sobre a importância da análise dos fatores físico-químicos do solo e da água, verificando a possibilidade de se empregar a tecnologia de mineração de dados na análise desses fatores.

Considerando que a tecnologia de Data Mining é aplicável em dados volumosos, são consideradas aqui duas áreas de pesquisa que envolvem grandes bases de dados associados a fatores físico-químicos, sendo uma associada a solos e outra a água.

O estudo sobre os fatores físico-químicos do solo está focado em dados provenientes de processos de Agricultura de Precisão devido a essa tecnologia prover grande e interessante base de dados envolvendo inúmeras variáveis. Da mesma forma, optou-se por avaliar o potencial de uso de mineração de dados no contexto da análise da qualidade da água de rios, considerando o grande volume de dados físico-químicos gerados pelas estações de monitoramento.

3.1 Dados de solo obtidos por Agricultura de Precisão

Vários pesquisadores afirmam que é sabido há muito tempo que nos campos agrícolas existe variabilidade espacial nas propriedades do solo e na produtividade de grãos, e que pode-se notar uma melhora nessa produtividade por meio de um gerenciamento localizado (Drummond et al, 2003; Whitney et al, 1999).

A variabilidade no solo ocorre tanto vertical como horizontalmente devido à própria natureza dos fatores responsáveis pela sua formação. Este fato ocorre porque o próprio material de origem não é uniforme em toda a sua extensão, ou seja, o material de origem não sofre o processo de intemperização de forma homogênea e contínua. Nas camadas superficiais, os solos são mais intemperizados (Buol et al., 1997). No caso de uma área cultivada, existem outras fontes de variabilidade no solo devido ao manejo exercido pelo homem, como o cultivo em linhas e a conseqüente aplicação localizada de fertilizantes (Johnson et al., 1996; Souza et al., 1997).

A busca por um melhor aproveitamento da terra, firmada na convicção da existência de variabilidade espacial e temporal tanto nas características do solo quanto na produtividade de grãos, deu origem a criação da área de pesquisa denominada Agricultura de Precisão (AP), a qual recebe ainda diferentes nomes e definições.

Também conhecida como Gerenciamento Localizado, a Agricultura de Precisão consiste em um sistema de gerenciamento agrícola baseado em tecnologia e informação para identificar, analisar e gerenciar a variabilidade espacial e temporal da produtividade nos campos a fim de otimizar a rentabilidade, a sustentabilidade e a proteção ao ambiente (Pierce & Sadler, 1997)

McBratney et al. (2005) propõe uma definição genérica afirmando que AP é um tipo de agricultura que aumenta o número de decisões (corretas) por unidade de área da terra e unidade de tempo com benefícios de rede associados. Essa definição move um pouco o foco da simples resolução espacial para outro que envolve a importância das decisões no espaço e no tempo.

Segundo Fraisse (1998) a Agricultura de Precisão é uma técnica de gerenciamento sistêmico que se propõe a otimizar o processo agrícola permitindo a aplicação de insumos nos locais corretos e nas quantidades requeridas.

A diversidade de definições de AP é bem apresentada no site do Laboratory for Agricultural Machinery and Processing, Katholieke University,(http://www.agr.kuleuven.ac.be/aee/amc/research/precag/introduction/PAdefinitions.htm), citado por McBratney et al. (2005).

A Agricultura de Precisão tem hoje à sua disposição vários componentes tecnológicos. A impressão de alguns especialistas agrícolas de que ela seja muito complicada

não é necessariamente correta; é claro que a implementação de um sistema ideal, trabalhando no ponto ótimo para todas as informações imagináveis, é um grande desafio. Mas algumas operações podem ser selecionadas para trabalhar em um ponto satisfatório, tendo como orientação o sistema ideal com simplificações.

Os recursos mais avançados da eletrônica e da computação, como os Sistemas de Posicionamento Global, também conhecido por GPS (do inglês *Global Positioning System*), os Sistemas de Informação Geográfica (SIG), os sistemas de controle e aquisição de dados, sensores e atuadores, entre outros, fazem parte da Agricultura de Precisão.

Usando um GPS associado a um SIG, dados do campo podem ser coletados de forma precisa, gerando bases de dados volumosas.

O uso racional dessas tecnologias, utilizadas como ferramentas de acompanhamento, controle e análise, permitem determinar "qual, quando e onde" o insumo deve ser aplicado e "como" fazê-lo. Ou seja, com a Agricultura de Precisão define-se como aplicar no local correto e no momento adequado, as quantidades de insumos necessários à produção agrícola, para áreas cada vez menores e mais homogêneas.

Portanto, a quantificação da variabilidade espacial permitindo identificar sítios específicos com diferentes potenciais de produtividade, pode determinar ou não investimentos em insumos ou na correção de fatores limitantes à produção, desde que econômica e tecnicamente viáveis, visando à maximização da produtividade e minimização dos impactos ambientais.

A maior dificuldade ainda reside em trabalhar esse grande volume de informações, referentes à variabilidade espacial e temporal, e interpretá-las para tomada de decisão em campo. O sistema água-solo-planta-atmosfera e os processos físico-químico-biológicos presentes são complexos, o clima é um fator de grande peso e as incertezas das modelagens devem ser minimizadas para a tomada de decisão.

O grande desafio da Agricultura de Precisão está em considerar as variações espaciais e temporais dos diversos parâmetros envolvidos no processo de produção agrícola. No solo, o teor de nutrientes, o teor de matéria orgânica, o pH, a umidade, a profundidade de camadas compactadas, entre outros parâmetros, apresentam variações que podem atingir até uma ordem de grandeza de um local para outro ou de uma data para outra, na mesma área de produção. Toda prática agrícola convencional está baseada em tratar o

campo como homogêneo, ignorando tais variações. No manejo convencional, a informação para melhoria do processo de produção é obtida de umas poucas amostras dos parâmetros. A interpretação da informação assume um valor médio das amostragens. O uso da informação, ou seja, a aplicação de insumos (principalmente agroquímicos em geral) é uma constante baseada nessa média e independe da maior ou menor necessidade de cada ponto da aplicação (Jorge, 2002).

Na Agricultura de Precisão as informações obtidas sobre os diversos processos da produção agrícola consistem de dados manuais que envolvem análises laboratoriais, dados coletados automaticamente por sensores estáticos (instalados no campo) e sensores dinâmicos (instalados nos implementos) e também dados obtidos por sensoriamento remoto. Os dados de posicionamento são fornecidos por GPS Diferencial (DGPS – Differential GPS). A interpretação das informações é auxiliada por computador e integra sistemas SIG com técnicas de geoestatística, programas de modelagem, entre outros para estabelecer e gerar mapas de controle das operações de campo, como a aplicação de fertilizantes, pesticidas, plantio, irrigação e outras.

Espera-se com a Agricultura de Precisão, em princípio, resultados de ordem econômica, mas as vantagens em termos de impacto ambiental são conseqüências da sua adoção. A aplicação de agroquímicos, de acordo com necessidades específicas espacialmente determinadas, deve levar ao uso racional de tais insumos. Assim, espera-se, além da maior margem de lucro pela redução dos gastos com esses produtos, também o benefício ambiental em termos da redução de resíduos nas culturas e diminuição da contaminação do lençol freático por percolação. Desse ponto de vista, a Agricultura de Precisão é a grande proposta atual para resolver o equacionamento da máxima produtividade com mínimos danos ambientais.

De acordo com Jorge (2002) as três principais áreas de operação na Agricultura de Precisão são:

- **Aplicação de fertilizantes:** A operação de maior interesse da pesquisa é a aplicação de fertilizantes. Nos Estados Unidos são consumidas dezenas de milhões de toneladas anuais de fertilizantes. Estima-se que os fertilizantes representem entre 25 e 45 % do custo de produção do milho. A estreita margem de lucro da produção

agrícola e a maior preocupação com a poluição ambiental têm aumentado consideravelmente o interesse no uso eficiente dos produtos para fertilização. São relatados ganhos que vão de 10 a 80 dólares por acre em culturas de milho e trigo com a adoção da Agricultura de Precisão. O caso extremo ($80/acre) serve para encorajar o bom gerenciamento, pois é atribuída uma especial atenção aos processos de amostragem, testes e planejamento, baseados na variabilidade espacial.

- **Controle de aplicação de pesticidas:** A segunda área de maior interesse da Agricultura de Precisão é o controle da aplicação de pesticidas para o domínio de doenças e pragas. Pelo menos 45 tipos de pesticidas já foram detectados no lençol freático, o que tem tornado a legislação a esse respeito muito mais severa. O desenvolvimento da aplicação espacialmente variável de pesticidas requer qualidade e precisão dos aplicadores. Na prática, a minoria dos aplicadores (cerca de 25%) é capaz de manter a taxa de aplicação dentro de pelo menos 5% da taxa ajustada. Com isso, cerca de 1 bilhão de dólares podem estar sendo perdidos anualmente só nos Estados Unidos.

- **Controle de plantio.** A terceira área de interesse é o controle do plantio. Basicamente, são três tipos de controle: da população das sementes, variando-se o espaçamento entre elas; da profundidade de deposição das sementes, onde são consideradas a espessura da camada superficial, a umidade e a compactação do solo e, por fim, da variedade das sementes, alternando-se sua fonte. Aparentemente não há impossibilidades técnicas para o controle da população, da profundidade de plantio ou da alternância de variedade. O mercado oferece plantadoras onde a população de sementes e a profundidade podem ser controladas pelo operador e os fornecedores de sementes já dispõem de vários híbridos para combinar com condições localizadas.

Informações sobre a variabilidade espacial podem ajudar também no esquema de irrigação. A princípio, as topografias acidentadas seriam as mais beneficiadas.

Porém, a aplicação de água de acordo com a posição pode reduzir erros sistemáticos inerentes dos pivôs centrais mesmo em terrenos planos, como efeitos de bordas.

A adoção da Agricultura de Precisão tem um grande potencial para a racionalização do sistema de produção agrícola moderno devido a: a) diminuição da quantidade de agroquímicos aplicados nos solos e culturas; b) conseqüente redução dos custos de produção e da contaminação ambiental e c) melhoria da qualidade das safras. Assim, a Agricultura de Precisão vem de encontro às exigências de um mercado globalizado, que requer maior volume de produção, exige menores preços e repudia técnicas e tecnologias que possam contaminar o ambiente.

Hoje são poucos os locais em todo o mundo onde se realiza comercialmente a Agricultura de Precisão de forma completa por vários motivos, dentre eles o custo da tecnologia, a baixa correlação obtida entre produtividade e os fatores estudados e a falta de informações corretas sobre a tecnologia. Na maioria das vezes realiza-se parte do ciclo, como por exemplo, o mapeamento da produtividade ou a aplicação de insumos em taxa variada, que fornecem informações importantes para o manejo das culturas e redução de custos.

No Brasil, as tecnologias necessárias à plena prática da Agricultura de Precisão não estão completamente disponíveis à grande maioria dos agricultores. A Agricultura de Precisão é considerada ainda num estágio experimental, com poucos grupos de pesquisa e grandes grupos privados trabalhando de forma efetiva neste assunto.

A adoção da Agricultura de Precisão tem sido feita com critério, realizando primeiramente um levantamento da variabilidade presente nos campos produtivos, refletida pela variabilidade espacial da produção. Por este motivo é que o mapeamento da produtividade tem sido o primeiro passo adotado. A evolução da AP pode ser acompanhada em extensa lista de literatura, entre as quais ressaltam-se Schueller (1992, 1997), Auernhammer (1994), Molin (1997). No Brasil, os primeiros resultados na obtenção de mapas de produtividade foram conseguidos por Balastreire et al. (1998) em cultura de milho, e desde então diversos outros resultados vêm sendo apresentados.

As informações do mapa de produtividade e dos mapas gerados a partir das amostragens realizadas permitem a interpretação e a realização de correlações entre os fatores e a produtividade para verificar a importância de cada um deles, porém, ainda é

necessário uma melhor compreensão da correlação entre esses fatores, para que se tenha tomadas de decisão mais acertadas.

Uma vez identificados os responsáveis pela variabilidade da produção pode-se partir para a fase seguinte que é a interferência, corrigindo os fatores que podem ser manejados como, por exemplo: materiais genéticos, fertilizantes, defensivos e outros. Esta fase não tem sido realizada de forma plena em áreas produtivas, por questões de custos e pela dificuldade ainda existente em se compreender a correlação entre as variáveis e a produtividade.

Assim, por meio da melhor compreensão da relação entre as variáveis associadas ao processo agrícola, buscam-se estratégias para o aumento da produtividade em solos, uma vez que explorar a capacidade de produção máxima dos mesmos com o mínimo de recursos empregados é a ambição de quem pesquisa e/ou investe em qualquer tipo de cultura.

A análise de regressão e outras técnicas estatísticas correlacionando os vários fatores tem sido freqüentemente usadas. Drumond et al. (1995) apresentam um estudo onde usaram diferentes estratégias de análise de regressão nos valores interpolados em células espacializadas de 10 m. Clay et al. (1998) apresentaram a mesma idéia. Eles avaliaram o impacto da distância de *grid* na análise espacial e a rentabilidade, correlacionando produtividade e seus fatores limitantes para a fertilidade do solo. Molin et al. (2001) também desenvolveram vários estudos executando análises de regressão múltipla visando identificar causas possíveis da variabilidade da produtividade, relacionadas à fertilidade do solo, obtendo amostras em diferentes profundidades. Gupta et al. (1997) desenvolveram um trabalho sobre a variabilidade espacial e estratégias de amostragem para a determinação dos elementos NO3-N, P e K para locais específicos. Embora os resultados obtidos nesses trabalhos sejam interessantes nota-se a dificuldade em se obter um índice de correlação significativo. Pesquisadores e produtores da área buscam conhecer as relações ainda não encontradas.

O tamanho e complexidade dessas bases de dados podem ser considerados os motivos da dificuldade em se encontrar resultados satisfatórios. Assim, surge a necessidade do uso de técnicas mais avançadas como a mineração de dados.

A introdução de novas tecnologias integradas com recursos computacionais visando produtividade e qualidade tem sido considerada questão estratégica. Dada a forte vocação agrícola brasileira e a importância do agronegócio no panorama

econômico nacional, a pesquisa visando o desenvolvimento da sociedade da informação neste setor é por si só um projeto relevante. Não obstante, a informação é a chave para o sucesso de qualquer atividade a qual pode se aperfeiçoar cada vez mais se a informação sobre ela obedecer ao ciclo: obtenção de novas informações seguida da interpretação e utilização dessas novas informações para melhorar a atividade.

Associado ao suporte tecnológico que alavanca a Agricultura de Precisão, existem inúmeras oportunidades em instrumentação e automação nessa área, bem como em sistemas inteligentes para tomada de decisão, por meio da mineração de dados e fusão de sensores.

3.2 Dados de qualidade de água

A água é essencial para a vida e desempenha um papel vital no funcionamento do ecossistema da Terra.

A água representa um dos mais básicos elementos de suporte à vida e ao ambiente natural, um primeiro componente para a indústria, um item de consumo para humanos e animais, e um vetor para poluição industrial e doméstica (Quevauviller, 2002).

A análise da água revela a presença de gases, minerais suspensos ou dissolvidos e matéria orgânica e microorganismos.

Muitos componentes da água ocorrem naturalmente, originários de, por exemplo, rochas, solos e ar, ou de fontes animais e humanas. Para estes componentes, substâncias antropogênicas são adicionadas por forças humanas devido a atividades urbanas, industriais e agrícolas.

Técnicas de tratamento da água desperdiçada nas cidades e indústrias também levam à formação e subseqüente liberação de contaminantes nas águas processadas. A qualidade e a quantidade dos vários (natural e/ou antropogênico) constituintes, na verdade formam a base para a definição da qualidade da água, sobre a qual a adequação para vários usos será decidida (por exemplo, consumo por animal doméstico e humano, uso doméstico ou industrial, irrigação, vida aquática, etc.). Segundo Quevauviller (2002) a observação, análise e

interpretação dos componentes da água são de suma importância para a definição da qualidade da mesma.

Decisões relativas ao gerenciamento ambiental, incluindo a qualidade das águas de várias fontes, são essencialmente suportadas pelos dados providos pelas análises laboratoriais. Dados de qualidade pobre podem produzir decisões erradas com severas conseqüências econômicas ou sociais. Por exemplo, erros feitos em relação a programas de monitoramento podem levar à não detecção de substâncias tóxicas ou a identificação de contaminantes não existentes.

Portanto, é imprescindível que as análises sejam conduzidas da forma mais correta possível. Porém, a atividade de análise e controle da água é onerosa, sendo preciso extrair-se o melhor conjunto possível de informações da volumosa base de dados que se forma em conseqüência do monitoramento.

Pela análise e acompanhamento da base de dados é possível identificar regiões em que a qualidade de água está abaixo do esperado, é possível identificar em cada região a funcionalidade possível que pode ser dada à água, pode-se fazer um trabalho de prevenção em relação ao uso da água, pode-se também estimar a qualidade da água em determinado período pelo acompanhamento temporal dos dados. Ou seja, pela melhor compreensão dos parâmetros ambientais inúmeras ações podem ser tomadas.

Boyd (2000) afirma que os conceitos de qualidade de água e quantidade de água foram desenvolvidos simultaneamente durante a história da humanidade, mas até recentemente, pouco significado quantitativo de avaliação da qualidade da água estava disponível.

Segundo Quevauviller (2002) existe ainda um campo de pesquisa aberto para a compreensão da interação dos componentes presentes na água. As atividades de pesquisa são multidisciplinares, envolvendo ecotoxicologia, geoquímica, microbiologia, biologia, química analítica, computação, entre outras.

Dessa forma a busca por padrões e pela melhor compreensão do comportamento dos parâmetros físico-químicos relacionados com a água é importante e pode contribuir para melhores tomadas de decisão em relação aos critérios de uso da mesma. Considerando as propostas da técnica de mineração de dados e a escassez de trabalhos sobre

análise de água apoiado nessa técnica, torna-se válido o estudo da aplicação de MD para análise de bases de dados de qualidade de água.

3.2.1 Critério de Qualidade da Água

Critérios de qualidade da água são valores quantitativos ou qualitativos para faixas aceitáveis de características físicas, químicas, biológicas e de aparência da água para usos específicos.

Os padrões e critérios são baseados na experiência, nos aspectos técnicos e econômicos, testes, evidência de efeitos de saúde pública, modelos matemáticos e obrigações legais (Tchobanoglus & Schroeder, 1985). Uma experiência considerável tem sido acumulada em padrões de qualidade de água, e muitos manuais sobre esse assunto têm sido publicados, por exemplo, manuais para água potável, para proteção de ecossistemas aquáticos, para água para irrigação e ainda para o uso de águas em recreação.

Nos Estados Unidos a Agência de Proteção Ambiental dos Estados Unidos, a qual atende pela sigla U.S. EPA (do inglês United States Environmental Protection Agency) tem publicado critérios de qualidade de água para mais de 50 poluentes. Os critérios estaduais normalmente são baseados no critério federal (EPA, 1994).

Segundo Boyd (2000) é muito difícil estabelecer critérios químicos para qualidade de água porque a toxidade de alguns metais e químicos orgânicos varia muito com as condições da qualidade da água. Portanto existe considerável uso de limitações baseadas em toxidade que são estabelecidas pelo completo teste de toxidade do efluente.

As pessoas precisam usar a água para diferentes propósitos, e a qualidade da água deteriora como resultado. A demanda pela água está crescendo em função do rápido crescimento da população humana. A deterioração da qualidade da água tem sério problema em muitas nações. A menos que medidas sejam implementadas para conservar ambas a quantidade e a qualidade da água, o mundo enfrentará uma séria falta de água no futuro.

Algumas nações têm desenvolvido sistemas elaborados de regulamentação de qualidade de água que ajudam a manter ou melhorar a qualidade de suas

águas. Contudo, muitas outras nações têm feito muito pouco para proteger a qualidade da água e sérios problemas estão ocorrendo. É urgente para todas as nações desenvolver regulamentações e modelos da qualidade da água e fazê-las ser cumpridas. É igualmente importante educar as pessoas sobre a importância de proteger nosso frágil abastecimento de água para usos futuros (Boyd, 2000).

Segundo Palmer (2001) existem várias aplicações típicas de modelo de qualidade de água, como a seguir:

- Para mudança nos sistemas de tratamento de água; para mudanças em carregamento de massas; para mudanças em processos de plantas.
- No desenvolvimento da terra e no planejamento de uso da terra.
- Para resolver conflitos de uso da água.
- Problemas de crescimento de plantas aquáticas.
- Na alocação de recursos de água para diferentes usos como água potável, irrigação, pesca e facilidades recreativas.
- No gerenciamento de inundações.

Para definir tais modelos é importante buscar novas correlações entre as variáveis envolvidas no contexto. A mineração de dados pode contribuir na busca dessas correlações. Um exemplo de aplicação da mineração é a classificação para predição da temperatura da água. Segundo Palmer (2001), a temperatura é um importante fator nas reações químicas e atividades biológicas na água. A temperatura também é um importante parâmetro para o habitat de peixes onde a temperatura máxima da água e as faixas de temperatura precisam estar dentro de limites especificados para as diferentes estações do ano.

Pela classificação dos dados em função das faixas de temperatura aceitável ou não, é possível, por exemplo, tentar descobrir os valores dos demais parâmetros e identificar aqueles que normalmente estão fora do limite aceitável quando a água está sob determinada temperatura. Dessa forma, sabendo-se as épocas em que a água atinge tal temperatura, é possível tomar-se ação preventiva sobre aqueles elementos para evitar que fique fora dos padrões de normalidade em função da temperatura da água.

Existem inúmeros parâmetros ambientais presentes para todos os tipos de água, entre eles pH, alcalinidade, temperatura, oxigênio dissolvido, matéria orgânica,

Potássio, Cálcio, entre outros. Não é objetivo dessa tese detalhar nem discutir esses fatores. Maiores detalhes sobre esse assunto podem ser obtidos em Chapman (1996) e Vigil (2003).

3.2.2 Uso de técnicas de IC na análise de dados de qualidade de água

Inúmeros grupos de pesquisa distribuídos no mundo vêm apresentando trabalhos que demonstram a preocupação e tentativa em utilizar técnicas de inteligência computacional e mineração de dados na estimativa dos parâmetros que interferem na qualidade da água.

Considerando que a temperatura interfere incisivamente nos fatores físico-químicos da água e, conseqüentemente, na capacidade de vida aquática, Risley et al. (2003) utilizaram modelos de Redes Neurais Artificiais para estimar a temperatura da água em pequenos córregos no oeste do estado de Oregon - EUA. Para a realização do mesmo foram coletados dados sobre a temperatura da água em 148 locais entre os meses de junho a setembro de 1999, de hora em hora. Como resultado, os modelos de temperatura da água desenvolvidos podem estimar a temperatura horária aproximada da água natural em pequenos córregos na região estudada. Obviamente, essa estimativa está restrita à região e período analisado, mas o trabalho traz sua contribuição demonstrando a viabilidade do uso de redes neurais nesse tipo de análise.

Dzeroski et al. (2000) desenvolveram um estudo relacionado ao problema da inferência de parâmetros químicos da qualidade de água de rios a partir de parâmetros biológicos. Esta tarefa é importante contribuindo no monitoramento químico da qualidade da água dos rios. Foi aplicada a técnica de indução por árvores de regressão, uma técnica de aprendizado de máquina, para correlacionar dados químicos e biológicos sobre a qualidade de água dos rios da Eslovênia.

Os dados sobre os rios eslovenos foram obtidos do Instituto Hidrometeorológico da Eslovênia (abreviado como HMZ) que executa o monitoramento da qualidade da água para a maioria dos rios eslovenos e mantém uma banco de dados de amostras de qualidade da água. Os dados providos pelo HMZ referem-se a um período de seis anos, de 1990 a 1995. Exemplos biológicos foram obtidos duas vezes por ano, no verão e no

inverno, enquanto que as análises químicas e físicas foram executadas várias vezes por ano para cada local amostrado.

As árvores de regressão foram construídas para predizer valores de parâmetros químicos com base na presença de bioindicadores. Nesse trabalho os autores apresentam uma interessante comparação justificando a opção de usar a técnica de árvores de regressão ao invés de métodos mais tradicionais de predição, como a regressão linear e o método do vizinho mais próximo. Segundo os autores, as três abordagens mostram-se muito próximas, em termos de performance de predição, especialmente a predição por árvores de regressão e pelo método do vizinho mais próximo. Porém as árvores de regressão têm a vantagem de produzir generalizações dos dados de entrada, estruturadas em um tamanho razoável, que são entendíveis, conforme os comentários de um especialista, sendo considerado mais fácil interpretar árvore de regressão de um tamanho moderado do que dezenas ou centenas de coeficientes em uma equação linear. Por outro lado, o método do vizinho mais próximo, não produz generalização dos dados de entrada como um todo.

Outro trabalho aplicando técnicas de inteligência computacional no tratamento de dados de qualidade de rios foi o desenvolvido por Cianchi et al. (2000). Baseado no fato de que o monitoramento e controle de qualidade dos rios são problemas de controle distribuído, foi proposto um sistema utilizando arquitetura distribuída, em ambiente de internet, para operar como uma ferramenta de gerenciamento de qualidade de água, com tarefas que variam desde o gerenciamento do banco de dados até a modelagem do sistema e projeto de controle para prover um sistema completo de suporte a decisão. As funções desse sistema são executadas por agentes inteligentes os quais consistem em módulos de software especializados que podem migrar sobre a rede e ativar as tarefas para os quais são designados. Entre as funções do sistema estão: Data Warehousing (repositório de dados para aquisição de dados normalizados), Data Mining, utilitário de configuração e serviço de simulação. Embora o sistema apresente uma arquitetura bastante interessante, não foram especificadas as tarefas de mineração de dados que podem ser executadas pelo sistema, o que dificulta uma melhor análise do potencial do mesmo em termos de mineração.

3.3 Considerações

Os trabalhos relacionados à análise de dados físico-químicos têm aplicado diferentes técnicas de inteligência computacional, porém não foi identificado nenhum sistema de mineração de dados específico para esse domínio e que possa manipular adequadamente o tipo de dado envolvido nesse contexto, os quais são do tipo numérico contínuo.

Sendo assim, torna-se viável o desenvolvimento de um sistema de mineração de dados para esse domínio de aplicação, utilizando a técnica de Algoritmos Genéticos, visto que suas características e vantagens podem contribuir para um melhor desempenho do sistema na manipulação do tipo de dado a ser analisado.

4. MATERIAL E MÉTODOS: SISTEMA DESENVOLVIDO

O sistema de mineração de dados físico-químicos desenvolvido nessa tese denomina-se MinAG, destina-se a realizar a tarefa de classificação e utiliza a técnica de Algoritmos Genéticos.

A opção por AGs foi feita em função de suas vantagens sobre outras técnicas no processo de descoberta de conhecimento, sendo elas: grande capacidade de trabalhar com dados imprecisos (Freitas & Lavington, 1998), a tratabilidade em termos de custo computacional (Langley, 1996), o ajuste fino de parâmetros de acordo com o domínio (Mitchell, 1997), e a possibilidade de paralelização e distribuição de carga de processamento (Freitas & Lavington, 1998).

A próxima seção detalha a técnica de Algoritmos Genéticos, brevemente citada na seção 2.5.3.

4.1 Definição do Algoritmo Genético

O objetivo dessa seção consiste em demonstrar a composição e comportamento do Algoritmo Genético incorporado no sistema MinAG, desenvolvido nessa tese.

A construção do algoritmo foi baseada em Goldberg (1989) e a implementação foi feita utilizando a linguagem de programação Delphi 5®.

As etapas que antecedem a execução desse Algoritmo Genético são as seguintes:

- Representar geneticamente as soluções viáveis do problema.
- Determinar uma população inicial de cromossomos.
- Definir uma função de avaliação dos cromossomos.
- Definir operadores genéticos eficazes na representação (geração) de novos cromossomos.
- Definir parâmetros para o tamanho da população; critérios de parada; critérios de renovação de cromossomos; taxa de mutação, taxa de crossover; etc.

A seguir cada etapa é detalhada.

- **Representação genética**

Qualquer possível solução é representada por meio de uma seqüência de símbolos, denominada indivíduo, ou cromossomo, em analogia com a biologia.

A representação genética é associada a um cromossomo p representado na forma de um vetor com n posições $p=(x_1, x_2, x_3, \ldots, x_{n)}$, onde cada componente x_i representa um gene (ou uma variável da solução). Usualmente, esta representação é feita com base em um alfabeto binário em que cada gene pode assumir o valor 0 ou 1. Contudo, outras formas de representação são possíveis.

No caso do MinAG, a representação não é binária. Cada indivíduo (cromossomo) é composto por genes, sendo que cada um corresponde a um atributo preditor da base de dados. Cada gene possui três campos: peso, operador e valor (Figura 10).

Campo Peso: indica se o atributo pertence ou não a regra. Ele é um valor pertencente ao intervalo de 0 a 1. Se o seu valor for maior ou igual ao valor dado pelo usuário, significa que o atributo deve ser considerado na regra, caso contrário, não.

Campo Operador: indica qual operador relacional será usado no tratamento daquele atributo, sendo possível indicar um ou dois dos seguintes operadores: <, >=, = ou <>.

Campo Valor: Corresponde ao valor do atributo. Na geração inicial ele corresponde ao valor adquirido do registro e é único. Caso sejam utilizados em um mesmo gene os operadores < e >= , o algoritmo entenderá que existe um intervalo de valores para esse atributo, podendo conseqüentemente, existir dois valores para o atributo representado nesse gene.

Cromossomo 1

...	Gene[11] = K			...	Gene[15] = Silte			...
...	Peso	Operador	Valor	...	Peso	Operador	Valor	...
...	1	=	1,275	...	0,96	>=	3,2805	...

Figura 10. Representação de um cromossomo no AG do MinAG

- **Geração da População Inicial**

A geração da população inicial é feita por procedimentos aleatórios ou algoritmos heurísticos. A população inicial no MinAG é definida por procedimentos aleatórios com o auxílio do gerador de números pseudo-aleatórios do ambiente de programação Delphi.

- **Função de Avaliação de Cromossomos**

Essa função avalia o nível de aptidão (adaptação) de cada cromossomo gerado, sendo também conhecida como função de *fitness*, ou ainda função objetivo.

A medida de *fitness* adotada inicialmente no MinAG para avaliar um cromossomo foi:

$$\textbf{\textit{Fitness}} = tp/(tp+fp)$$

Onde:

tp = número de verdadeiros positivos

fp = número de falsos positivos

Verificou-se que essa medida não se mostrava muito consistente, visto que, caso houvesse apenas um caso verdadeiro positivo e nenhum falso positivo, a função de *fitness* seria 1, ou seja, excelente. Contudo, a regra não poderia ser considerada consistente, mesmo com *fitness* 1, porque poderiam existir casos do tipo falso negativo que estariam contradizendo tal regra. Adotou-se então a seguinte medida de *fitness*, levando em consideração os casos falso negativo:

$$\textbf{\textit{Fitness}} = tp/(tp+fp) * tp/(tp+fn)$$

Onde:

tp = número de verdadeiros positivos

fp = número de falsos positivos

fn = número de falsos negativos

- **Operadores Genéticos**

Os operadores genéticos que podem ser aplicados são: mutação, *crossover* e clonagem, sendo que foram adotados no MinAG os dois primeiros operadores.

Mutação: Esse operador realiza alterações nos valores de um ou mais genes de um cromossomo. Quando esse cromossomo é binário este operador consiste da inversão aleatória nos bits do genótipo, como pode ser visto na Figura 11.

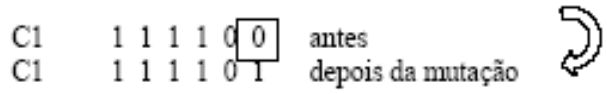


Figura 11. Aplicação de um operador de mutação em um cromossomo

No caso do MinAG, existe a possibilidade de mutação não somente dos valores dos atributos armazenados no cromossomo como também dos operadores relacionais usados na regra o que proporciona um grande diferencial para o algoritmo desenvolvido nessa tese, visto que por manipular dados contínuos é necessário tratar os dados em intervalos de valores usando os operadores >, >=, < e <=, além da igualdade (=) .

Crossover (Recombinação): O uso desse operador consiste em realizar uma recombinação por meio de cortes nos cromossomos pais, segundo uma regra pré-definida. Os cortes podem ser fixos para todos os cromossomos ou aleatórios. A Figura 12 apresenta o uso do crossover de um corte onde D1 e D2 são os filhos resultantes do cruzamento de G1 e G2.

Figura 12. Crossover de um corte

O tipo de recombinação em função do corte adotado no AG do MinAG foi o *crossover* de um corte.

- **Determinação dos Parâmetros**

Vários parâmetros precisam ser definidos para o funcionamento do AG. Para cada conjunto de parâmetros uma solução diferente é gerada. Cabe ao usuário, executar o algoritmo com diferentes combinações de parâmetros, até encontrar a solução mais satisfatória. Os parâmetros adotados no AG do MinAG são os seguintes:

Tamanho da População: afeta o desempenho global e a eficiência do AG. Populações maiores oferecem menor risco de convergência prematura com maior possibilidade de alcançar o ótimo global; populações menores exigem menos tempo e menos recursos computacionais, mas pode chegar a uma convergência sem atingir o ponto ótimo global. O tamanho *default* adotado para a população no MinAG é 50.

Taxa de Crossover (Recombinação): Quanto maior a taxa de recombinação, mais rapidamente novas estruturas serão introduzidas na população. Mas se esta for muito alta pode

ocorrer perda de estruturas de alta aptidão, e o algoritmo pode tornar-se muito lento. No intervalo de 0 a 1, o MinAG assume-se como *default* a taxa de *crossover* de 0,7.

Taxa de Mutação: Quando a taxa de mutação é alta, a busca torna-se essencialmente aleatória, e existe a possibilidade de perda de estruturas de alta aptidão. Porém, é importante que exista uma taxa razoável de mutação porque assim previne-se que uma dada posição fique estagnada em um valor, além de possibilitar que se chegue em qualquer ponto do espaço de busca. No caso do MinAG, observou-se que a taxa de mutação apropriada deveria ser baixa, como por exemplo, 0,1.

Intervalo de Gerações: Controla a porcentagem da população que será substituída durante a próxima geração. Com um intervalo de geração alto, a maior parte da população será substituída. Porém, com valores altos demais pode ocorrer perda de estruturas de alta aptidão, além de o algoritmo tornar-se muito lento. O valor *default* é 30%.

Critério de Parada: Existem basicamente dois critérios de parada. Um deles consiste em se estabelecer o número máximo de gerações que serão produzidas. Outro critério baseia-se na homogeneidade da população, ou seja, quando o coeficiente de variação de uma população é menor que um dado valor D, então o processo de geração de novas populações é terminado e a melhor solução é aquela dentre os indivíduos que mais se adaptam à função de avaliação.

O critério de parada adotado no MinAG foi o número de gerações definido pelo usuário, cujo valor *default* é 50..

Mecanismo de Seleção: Um ponto importante a ser comentado sobre as especificações do AG, consiste no mecanismo de seleção utilizado para identificar o indivíduo que participará da próxima geração. Existem vários mecanismos, sendo que abordaremos aqui a roleta e o elitismo.

Roleta: consiste em escolher aleatoriamente um indivíduo, sendo que a probabilidade de ele ser o selecionado aumenta de acordo com sua aptidão, ou seja, quanto maior sua função de *fitness*, maior será sua fração na roleta.

Elitismo: Nesse mecanismo os melhores indivíduos são perpetuados para a próxima geração. Na prática isto resulta numa busca mais agressiva, que é geralmente efetiva. No entanto existe o perigo de uma convergência prematura para mínimos locais. Cada indivíduo selecionado e cruzado com seu parceiro é colocado no lugar do pior indivíduo da população anterior. A aptidão é atribuída de acordo com um "ranking", ou seja, a aptidão de cada indivíduo assume valores discretos.

O MinAG adota o mecanismo da roleta com a possibilidade do usuário dividi-la em classes, para privilegiar a seleção de determinados tipos de elementos. Porém, observou-se que os melhores resultados foram obtidos com apenas uma classe, ou seja, sem privilegiar elementos.

Com todas as especificações necessárias estabelecidas, o funcionamento do AG pode ser expressado segundo os passos mostrados na Figura 13.

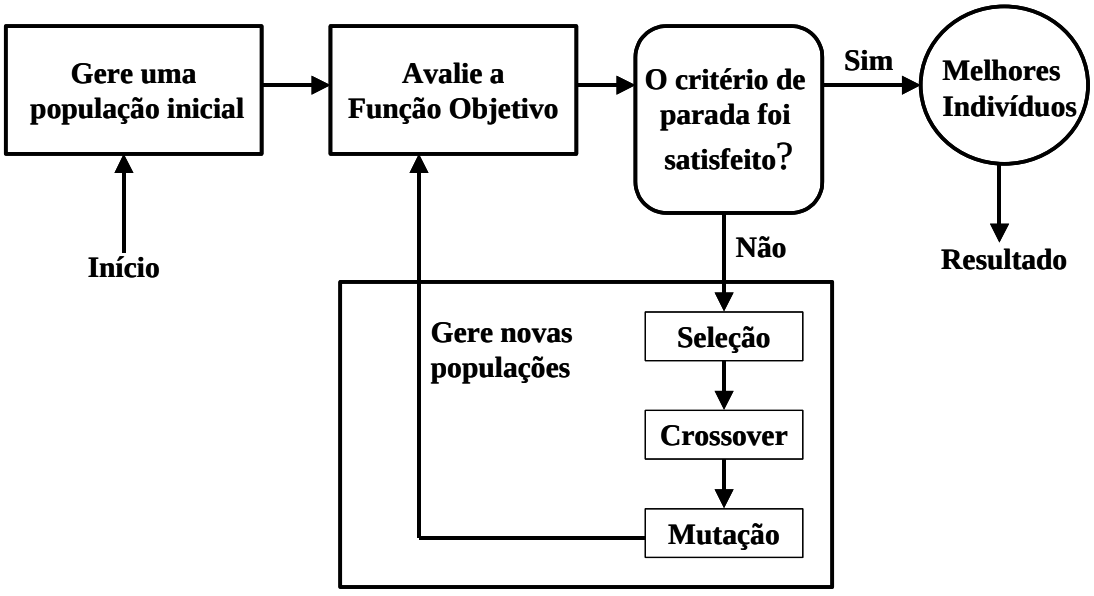


Figura 13. Esquema de funcionamento do AG implementado no MinAG

Como pode ser visto, existe toda uma terminologia utilizada em AGs, em função de sua associação com a biologia. O Quadro 3 apresenta os principais termos usados, associando sua origem com sua função no AG.

Quadro 3. Terminologia adotada em Algoritmos Genéticos (Freitas, 2003).

Natural	**Algoritmos Genéticos**
Indivíduo	Solução de um problema
População	Conjunto de soluções
Cromossomo	Representação de uma solução
Fitness	Qualidade de uma solução
Gene	Parte da representação de uma solução
Crossover e Mutação	Operadores genéticos
Seleção natural	Reutilizar as melhores soluções

4.2 Manipulação da base de dados pelo AG

Para a tarefa de classificação faz-se necessário dividir a base de dados em treinamento e teste, para ser possível avaliar as regras geradas em um domínio não explorado.

O sistema foi projetado para permitir que o usuário escolha o percentual de sua base que deseja utilizar para treinamento e para teste. O usuário pode ainda ter diferentes bases de treinamento e teste já definidas previamente à execução do sistema e apenas informar o nome das mesmas, porém normalmente o usuário tende a optar por dividir sua base original. Essa divisão é feita de maneira aleatória, ou seja, por meio de geração automática de números aleatórios são escolhidos registros de diferentes partes do arquivo original, de forma que, tanto a base de treinamento como a de teste contenham registros de todas as partes do arquivo original. A quantidade de registros selecionados para teste e

treinamento depende do percentual de divisão definido pelo usuário para cada uma das bases. Normalmente, a base de treinamento deve ser maior do que a de teste, tendo uma maior porcentagem dos registros da base original, por exemplo, 60% de treinamento e 40% de teste. A Figura 14 mostra a divisão da base de dados e seu uso no processo de descoberta de padrões pelo AG.

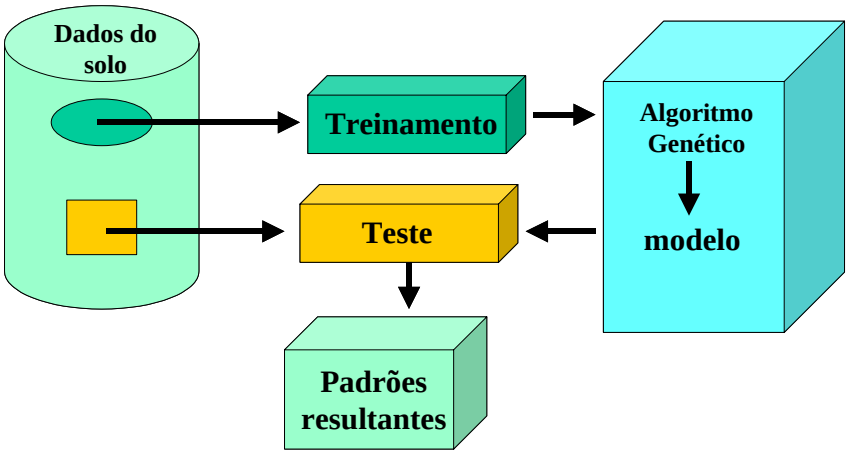


Figura 14. Divisão e uso de uma base de dados pelo AG para mineração

O formato padrão assumido para a base de dados foi o CSV (campos separados por vírgulas), devendo existir na primeira linha de cada coluna o nome do atributo cujos valores estarão expressos na referida coluna. Portanto, uma linha deve representar uma ocorrência para os valores dos atributos.

4.3 Funcionalidades do Sistema MinAG

Uma vez detalhada a composição do Algoritmo Genético responsável por efetivamente realizar o trabalho de minerar os dados, o sistema MinAG como um todo pode ser definido.

A modelagem funcional desse sistema está representada em um diagrama de fluxo de dados (DFD). Segundo Gustafson (2003), o DFD mostra o fluxo dos dados entre um conjunto de componentes. Os componentes podem ser tarefas, componentes de software ou mesmo abstrações das funcionalidades incluídas no sistema de software.

As regras e interpretações para o DFD aqui implementado foram extraídas de Gane (2003), sendo elas:

- Entidades externas (origens ou destinos de fluxo de dados para dentro e para fora do sistema) são representadas por um quadrado que recebe sombreamento em dois lados.
- A seta no DFD deve ser considerada como um caminho através do qual uma ou mais estruturas de dados poderão passar em tempo não especificado, representando, portanto, o fluxo dos dados. As setas devem ser nomeadas.
- Seta ramificada indica que o mesmo fluxo de dados está indo de uma fonte para dois lugares diferentes.
- O processo ou função, que transforma os dados de alguma maneira, é representado por um retângulo com cantos arredondados. A descrição da função do processo, contida no retângulo, começa por um verbo no infinitivo, seguida por uma cláusula do objeto, como “Gerar” (verbo) “regras” (cláusula do objeto).
- Os armazéns de dados (tabelas) são representados em um retângulo alongado, aberto em uma das laterais.

O diagrama de fluxo de dados do sistema desenvolvido está representado na Figura 15.

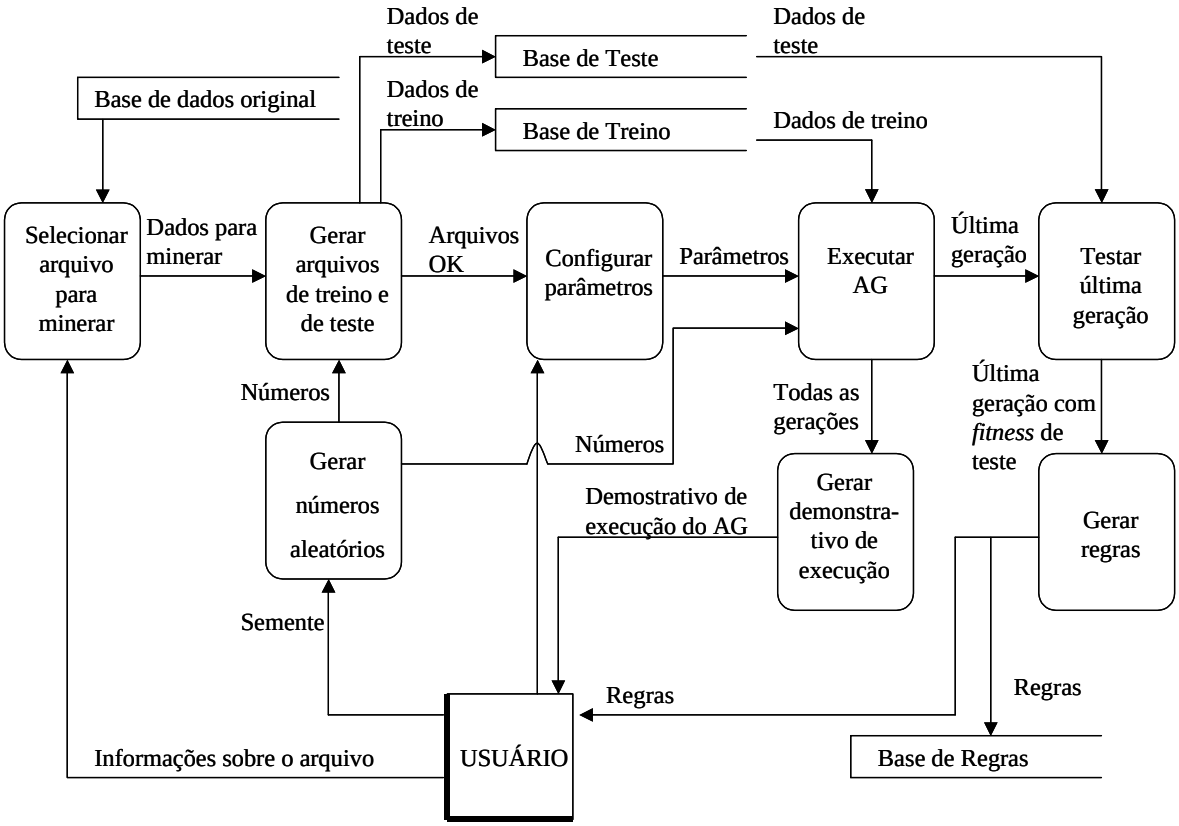


Figura 15. Diagrama de fluxo de dados do sistema MinAG

5 RESULTADS E DISCUSSÃO

Com base no diagrama de fluxo de dados o sistema MinAG foi desenvolvido resultando em um sistema de mineração de dados que atende aos requerimentos necessários para classificar dados físico-químicos do solo e da água, sob algumas restrições. Essas restrições estão relacionadas às características do arquivo de dados aceito pelo MinAG e à tarefa de mineração à qual o sistema se destina, que é a classificação.

O sistema resultante pode ser executado em modo convencional (um computador pessoal - PC) ou em paralelo (processamento em *grid*). O usuário poder ajustar os parâmetros conforme suas necessidades, assim como pode deixar que o sistema assuma os valores *default* para tais parâmetros, o que torna seu uso facilitado.

A seguir são descritas as funcionalidades, representadas nos processos do diagrama de fluxo de dados, e as formas de execução do sistema resultante.

5.1 Descrição das funcionalidades do MinAG

- **Selecionar arquivo para minerar**

O processo de mineração de dados físico-químicos do solo e da água para a tarefa de classificação é o objetivo principal do software. Antes, porém, deve-se

selecionar a base de dados (Figuras 16 e 17), a qual deverá estar em um formato reconhecido pelo software.

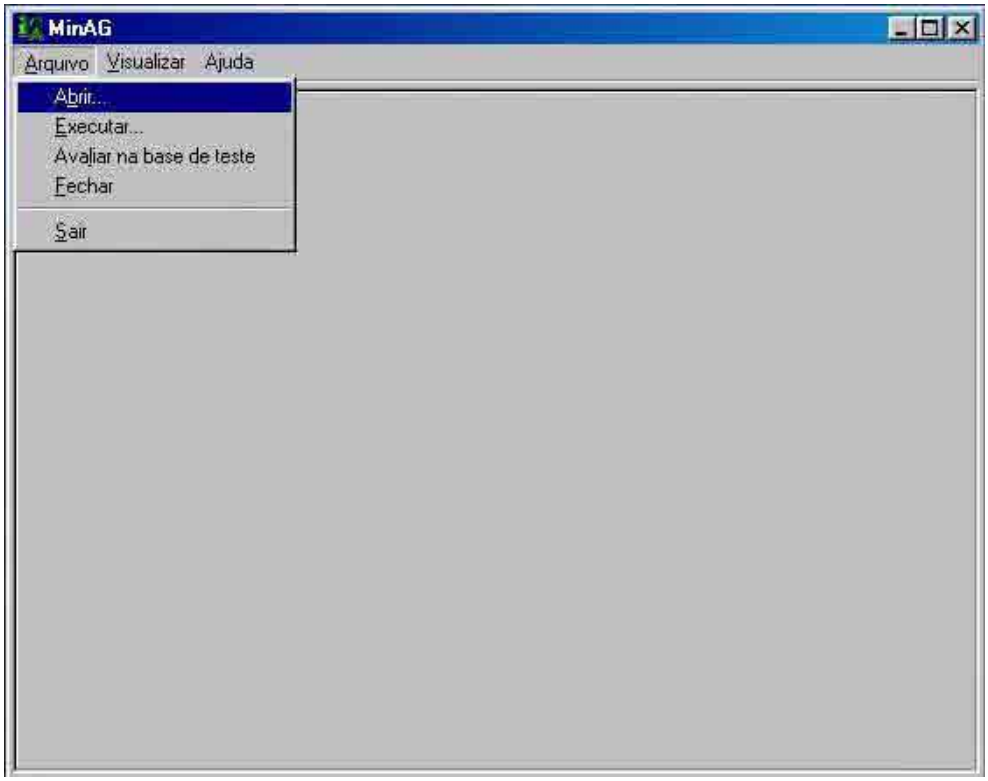


Figura 16. Caixa de diálogo inicial

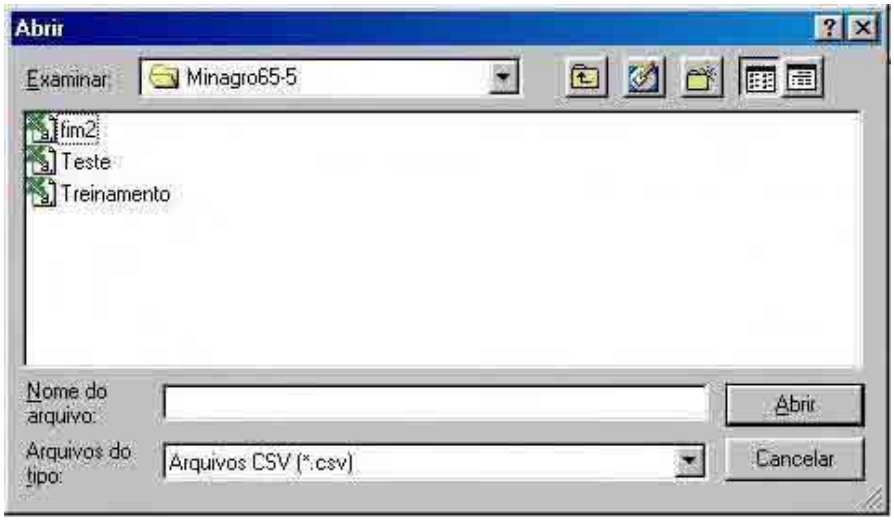


Figura 17. Caixa de diálogo para seleção de arquivo de dados contínuos

Em seguida o usuário deve informar alguns dados específicos da base de dados. Esses dados dizem respeito às colunas do atributo meta e do campo identificador de cada registro (Figura 18). Feito isso, o arquivo informado pelo usuário será aberto e apresentado na tela (Figura 19).

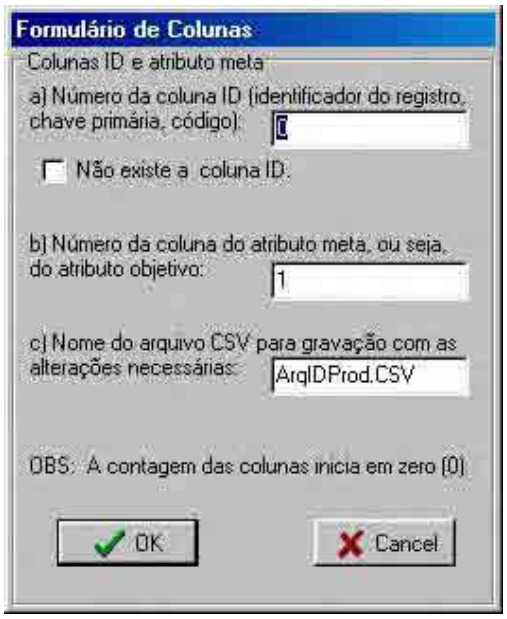
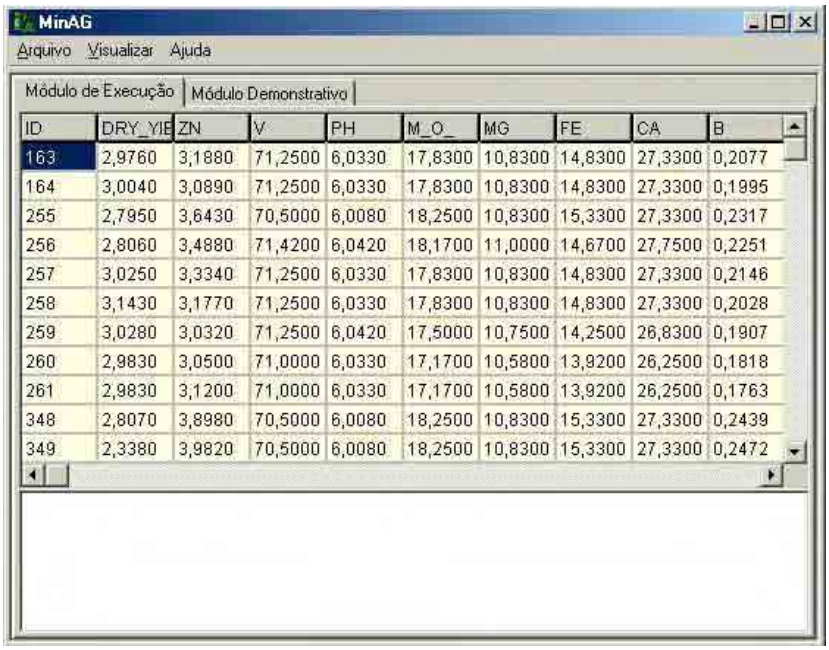


Figura 18. Formulário de colunas



ID	DRY_YIE	ZN	V	PH	M_O_	MG	FE	CA	B
163	2,9760	3,1880	71,2500	6,0330	17,8300	10,8300	14,8300	27,3300	0,2077
164	3,0040	3,0890	71,2500	6,0330	17,8300	10,8300	14,8300	27,3300	0,1995
255	2,7950	3,6430	70,5000	6,0080	18,2500	10,8300	15,3300	27,3300	0,2317
256	2,8060	3,4880	71,4200	6,0420	18,1700	11,0000	14,6700	27,7500	0,2251
257	3,0250	3,3340	71,2500	6,0330	17,8300	10,8300	14,8300	27,3300	0,2146
258	3,1430	3,1770	71,2500	6,0330	17,8300	10,8300	14,8300	27,3300	0,2028
259	3,0280	3,0320	71,2500	6,0420	17,5000	10,7500	14,2500	26,8300	0,1907
260	2,9830	3,0500	71,0000	6,0330	17,1700	10,5800	13,9200	26,2500	0,1818
261	2,9830	3,1200	71,0000	6,0330	17,1700	10,5800	13,9200	26,2500	0,1763
348	2,8070	3,8980	70,5000	6,0080	18,2500	10,8300	15,3300	27,3300	0,2439
349	2,3380	3,9820	70,5000	6,0080	18,2500	10,8300	15,3300	27,3300	0,2472

Figura 19. Formulário apresentando dados do arquivo aberto

- **Gerar números aleatórios**

O sistema MinAG utiliza números aleatórios para a geração dos arquivos de treino e de teste, na seleção dos registros que deverão pertencer a um ou outro arquivo. Além disso, durante toda a execução do AG, números aleatórios são necessários para definir a ação a ser executada, como por exemplo, em função do número aleatório gerado define-se se ocorrerá ou não uma mutação em determinado elemento.

Para a execução do gerador de números aleatórios é necessário que o usuário informe uma semente (número inicial). Caso o usuário desejar reproduzir uma mineração de dados repetindo os resultados ocorridos em uma execução prévia é necessário que todos os parâmetros de entrada sejam iguais, inclusive a semente do Gerador. O valor da semente deve ser um número real entre 0 e 1. O valor *default* para o sistema MinAG é 0,3333.

Figura 20. Caixa de diálogo para coleta da semente

- **Gerar arquivos de treino e de teste**

É o processo responsável pela divisão da base de dados em uma base de treinamento e uma base de teste (Figuras 21). Caso o usuário optar por dividir a base original (arquivo atual) em treinamento e teste uma caixa de diálogo será ativada para que seja informada a proporção para a divisão (Figura 22). Caso contrário, a base atual será usada integralmente para a realização do treinamento e no momento da execução do teste o nome do arquivo de teste será solicitado.

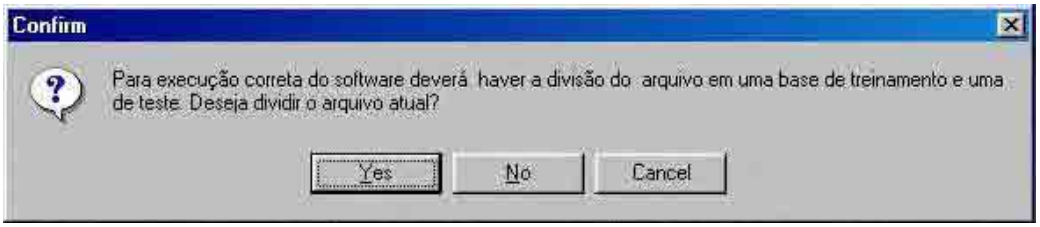


Figura 21. Caixa de diálogo para confirmação da divisão do arquivo

Figura 22. Formulário para coleta de dados sobre a divisão do arquivo

- **Configurar parâmetros**

Nesse processo é feita a identificação de parâmetros específicos necessários ao AG (Figuras 23 a 28). Os parâmetros referem-se à classe procurada, população e operadores genéticos. Uma vez configurados os parâmetros, é feita a chamada da execução do AG para realizar a mineração dos dados.

Figura 23. Tela para coleta de dados sobre a classe procurada

Figura 24. Tela para coleta de dados sobre a população

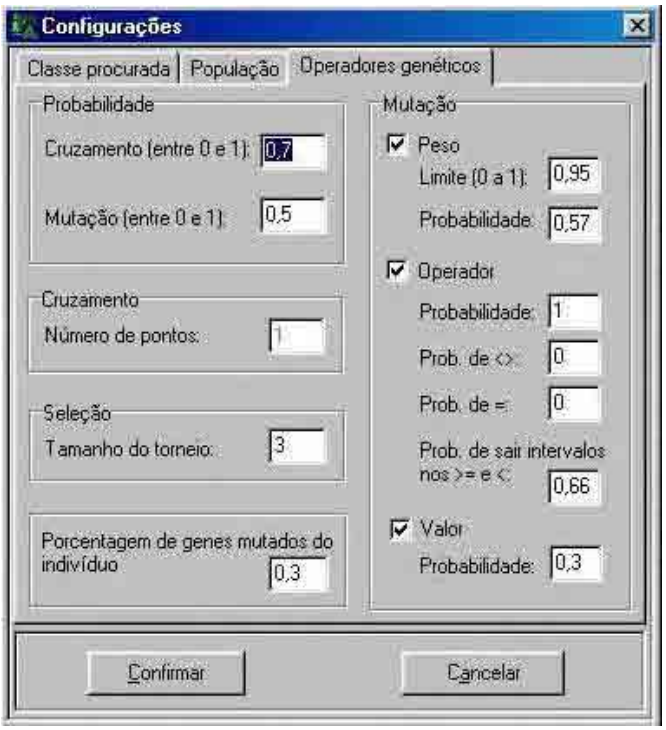


Figura 25. Tela para coleta de dados sobre os operadores genéticos

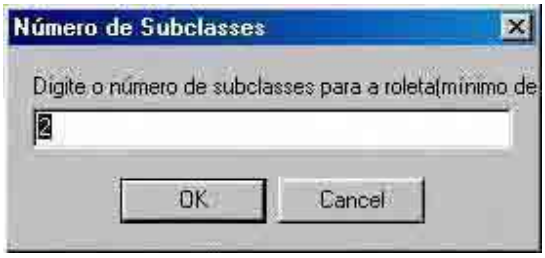


Figura 26. Caixa de diálogo para informação de número de subclasses para a roleta

Figura 27. Formulário para coleta de dados para a roleta

Figura 28. Caixa de diálogo para informar o nome do arquivo de resultados

- **Executar AG**

Esse processo realiza a mineração de dados propriamente dita, por meio da execução do AG. Com base nos parâmetros definidos a população inicial de soluções candidatas é definida e a quantidade de gerações estabelecida pelo usuário é executada, de forma a evoluir os elementos (cromossomos) da população buscando as melhores soluções (regras). A execução da mineração é feita utilizando-se a base de dados de treinamento.

- **Testar última geração**

Finalizada a mineração de dados na base de treinamento, faz-se necessário a comprovação do conhecimento descoberto, a qual deve ser realizada na base de dados de teste. Caso o usuário tenha dividido a base original em teste e treinamento a execução do teste é feita automaticamente após a ativação do ícone de teste. Se o usuário não dividiu a base original, nessa etapa será necessário informar o nome do arquivo de teste (Figura 29).

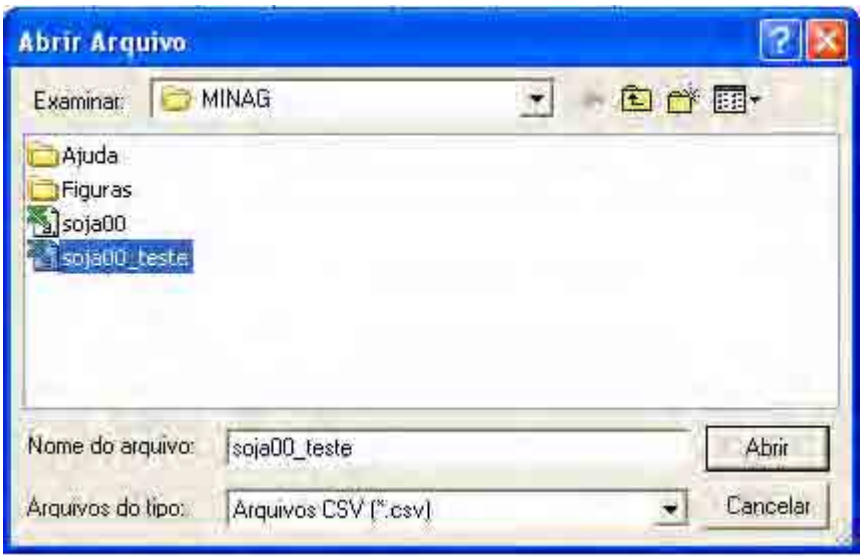


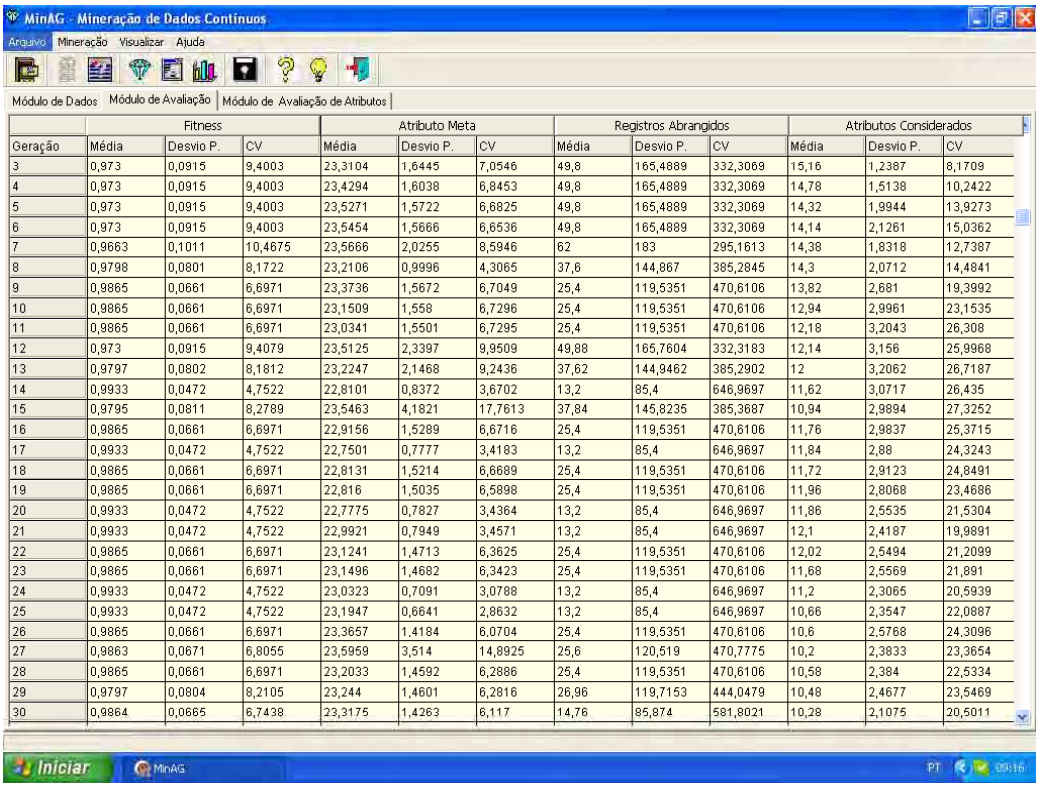
Figura 29. Caixa de diálogo para informar o nome do arquivo de teste

- **Gerar demonstrativo de execução**

Após a execução de um processo de mineração de dados muitos usuários se preocupam com a melhora dos resultados (regras) obtidos. Isso pode ser observado em Holmes et al. (2002), que relata o estado da arte de Sistemas Classificadores de Aprendizado, os quais objetivam aperfeiçoar regras anteriormente obtidas. Essa funcionalidade busca contribuir para identificar a geração com melhores indivíduos e conseqüentemente, melhores regras.

O AG utiliza busca global, impedindo que o algoritmo limite-se a um máximo local. No entanto, após uma execução de 100 gerações no algoritmo, como somente a última geração é codificada na forma SE-ENTÃO, pode acontecer das regras geradas serem de qualidade inferior as que seriam obtidas na geração 98. A finalidade do módulo demonstrativo (ou módulo avaliador) é apresentar alguns resultados do AG, de modo que o usuário perceba quais foram as gerações de indivíduos que poderiam originar as regras mais interessantes. Assim, o usuário pode escolher esta geração para codificar na forma SE-ENTÃO, devendo, para tanto, executar novamente a mineração informando como número de geração aquela escolhida.

O Módulo Demonstrativo (Figura 30) apresenta a média, o desvio padrão e o coeficiente de variação de determinados atributos em cada geração. Além disso, informa o número de vezes em que cada atributo preditor apareceu nas gerações determinando então quais os atributos que mais ocorreram e os que menos ocorreram na execução total do AG.



	Fitness			Atributo Meta			Registros Abrangidos			Atributos Considerados		
Geração	Média	Desvio P.	CV	Média	Desvio P.	CV	Média	Desvio P.	CV	Média	Desvio P.	CV
3	0,973	0,0915	9,4003	23,3104	1,6445	7,0546	49,8	165,4889	332,3069	15,16	1,2387	8,1709
4	0,973	0,0915	9,4003	23,4294	1,6038	6,8453	49,8	165,4889	332,3069	14,78	1,5138	10,2422
5	0,973	0,0915	9,4003	23,5271	1,5722	6,6825	49,8	165,4889	332,3069	14,32	1,9944	13,9273
6	0,973	0,0915	9,4003	23,5454	1,5666	6,6536	49,8	165,4889	332,3069	14,14	2,1261	15,0362
7	0,9663	0,1011	10,4675	23,5666	2,0255	8,5946	62	183	295,1613	14,38	1,8318	12,7387
8	0,9798	0,0801	8,1722	23,2106	0,9996	4,3065	37,6	144,867	385,2845	14,3	2,0712	14,4841
9	0,9865	0,0661	6,6971	23,3736	1,5672	6,7049	25,4	119,5351	470,6106	13,82	2,681	19,3992
10	0,9865	0,0661	6,6971	23,1509	1,558	6,7296	25,4	119,5351	470,6106	12,94	2,9961	23,1535
11	0,9865	0,0661	6,6971	23,0341	1,5501	6,7295	25,4	119,5351	470,6106	12,18	3,2043	26,308
12	0,973	0,0915	9,4079	23,5125	2,3397	9,9509	49,88	165,7604	332,3183	12,14	3,156	25,9968
13	0,9797	0,0802	8,1812	23,2247	2,1468	9,2436	37,62	144,9462	385,2902	12	3,2062	26,7187
14	0,9933	0,0472	4,7522	22,8101	0,8372	3,6702	13,2	85,4	646,9697	11,62	3,0717	26,435
15	0,9795	0,0811	8,2789	23,5463	4,1821	17,7613	37,84	145,8235	385,3687	10,94	2,9894	27,3252
16	0,9865	0,0661	6,6971	22,9156	1,5289	6,6716	25,4	119,5351	470,6106	11,76	2,9837	25,3715
17	0,9933	0,0472	4,7522	22,7501	0,7777	3,4183	13,2	85,4	646,9697	11,84	2,88	24,3243
18	0,9865	0,0661	6,6971	22,8131	1,5214	6,6689	25,4	119,5351	470,6106	11,72	2,9123	24,8491
19	0,9865	0,0661	6,6971	22,816	1,5035	6,5898	25,4	119,5351	470,6106	11,96	2,8068	23,4686
20	0,9933	0,0472	4,7522	22,7775	0,7827	3,4364	13,2	85,4	646,9697	11,86	2,5535	21,5304
21	0,9933	0,0472	4,7522	22,9921	0,7949	3,4571	13,2	85,4	646,9697	12,1	2,4187	19,9891
22	0,9865	0,0661	6,6971	23,1241	1,4713	6,3625	25,4	119,5351	470,6106	12,02	2,5494	21,2099
23	0,9865	0,0661	6,6971	23,1496	1,4682	6,3423	25,4	119,5351	470,6106	11,68	2,5569	21,891
24	0,9933	0,0472	4,7522	23,0323	0,7091	3,0788	13,2	85,4	646,9697	11,2	2,3065	20,5939
25	0,9933	0,0472	4,7522	23,1947	0,6641	2,8632	13,2	85,4	646,9697	10,66	2,3547	22,0887
26	0,9865	0,0661	6,6971	23,3657	1,4184	6,0704	25,4	119,5351	470,6106	10,6	2,5768	24,3096
27	0,9863	0,0671	6,8055	23,5959	3,514	14,8925	25,6	120,519	470,7775	10,2	2,3833	23,3654
28	0,9865	0,0661	6,6971	23,2033	1,4592	6,2886	25,4	119,5351	470,6106	10,58	2,384	22,5334
29	0,9797	0,0804	8,2105	23,244	1,4601	6,2816	26,96	119,7153	444,0479	10,48	2,4677	23,5469
30	0,9864	0,0665	6,7438	23,3175	1,4263	6,117	14,76	85,874	581,8021	10,28	2,1075	20,5011

Figura 30. Módulo demonstrativo do sistema MinAG

- **Gerar regras**

É o processo que possibilita ao usuário a visualização dos resultados da mineração dos dados (Figura 31), os quais são apresentados em regras de produção no formato SE-ENTÃO. O usuário pode optar por visualizar todas as regras ou fornecer um valor de *fitness* mínimo aceitável para visualização. O *fitness* é uma função de avaliação de um cromossomo, sendo, portanto, uma função que comprova a qualidade da regra gerada. Quanto maior o valor do *fitness*, melhor a qualidade da regra.

Cada regra possui um valor de *fitness* para a base de dados de treinamento e para a base de teste. Para a geração das regras de produção, o MinAG considera o *fitness* mínimo de ambas as bases.

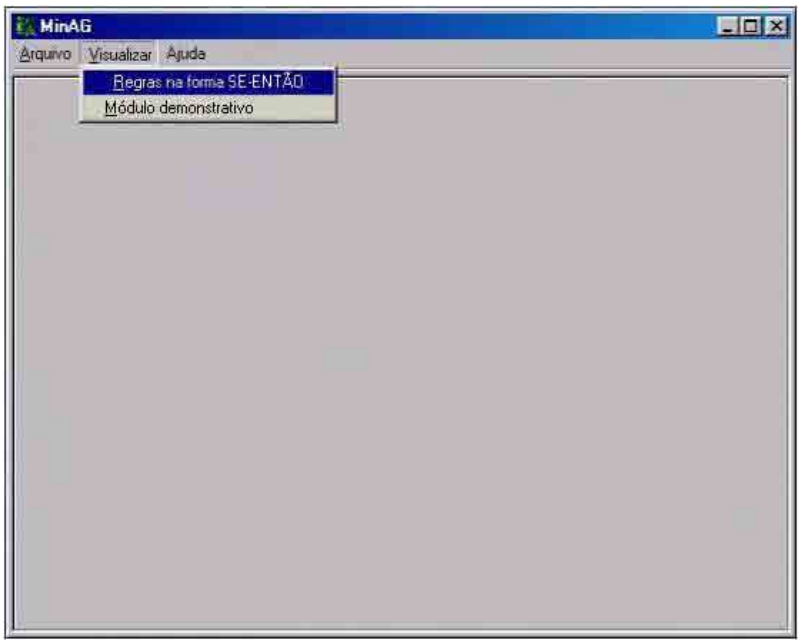
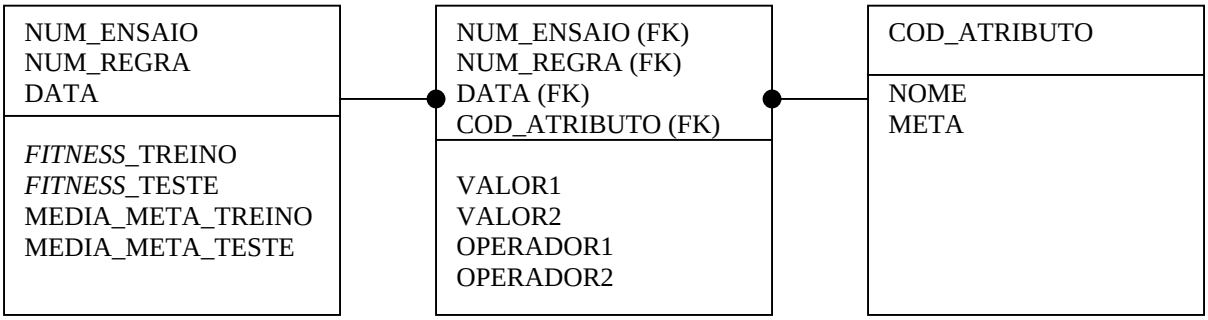


Figura 31. Menu Visualizar Regras

O usuário pode optar por gravar os resultados da mineração em um arquivo texto para posterior recuperação no próprio MinAG.

Visando facilitar futuras análises do grau de interesse e manipulações das regras resultantes o MinAG gera, além de um arquivo texto com o resultado da execução, um banco de dados das regras resultantes da mineração. O Sistema Gerenciador de Banco de Dados (SGBD) adotado foi o Interbase®, por ser uma ferramenta gratuita e largamente utilizada no ambiente acadêmico.

O banco de dados de regras é composto por três tabelas que se relacionam e contém a seguinte estrutura:



Com base nessa estrutura de banco de dados é possível armazenar e recuperar a qualquer tempo cada uma das regras geradas durante uma execução do sistema, sendo que todas as informações pertinentes às regras estão presentes no banco.

5.2 Execução do Sistema usando Paralelismo

Inicialmente o sistema foi projetado para ser executado de forma convencional, ou seja, utilizando apenas um computador do tipo PC (Personal Computer).

Durante o estágio realizado na Universidade da Flórida foi disponibilizado, para a otimização do sistema MinAG, o ambiente de paralelismo. Esse ambiente permitiu que o sistema fosse executado em 400 computadores da Universidade da Flórida em paralelo, ou seja, ao mesmo tempo. Esse tipo de execução melhorou em muito o desempenho e aplicação do sistema MinAG. A mineração dos dados referentes aos casos de

uso apresentados nessa tese foi executada em 84 horas em ambiente paralelo, tendo sido realizadas um total de 10880 minerações, ou seja, o sistema foi executado 10880 vezes. Estima-se que, para realizar essa quantidade de minerações, considerando-se um tempo médio de 7 minutos por execução e usando-se somente um computador, o tempo de processamento seria em torno de 1450 horas, o equivalente a 60 dias ininterruptos. Portanto, por meio do sistema de paralelismo é possível realizar em um tempo expressivamente menor várias execuções do sistema com diferentes bases de dados, ou com diferentes tabelas de uma mesma base de dados, e ainda com diferentes combinações de parâmetros. Isso é importante por três principais motivos:

- Existe dificuldade em se determinar os parâmetros ideais para o Algoritmo Genético, os quais dependem da base de dados em uso. Por meio das execuções do sistema é possível identificar os parâmetros que contribuíram para os melhores resultados, os quais podem então ser usados como referencial em outras execuções futuras.

- Visto que o Algoritmo Genético insere-se no contexto de aprendizado de máquina, torna-se necessário que sejam realizadas inúmeras iterações durante a execução do AG para que o aprendizado sobre os dados ocorra. Essa quantidade de iterações requer tempo de execução. Podendo-se realizar o processamento da mineração de dados em paralelo, pode-se testar então o comportamento do AG com diferente número de iterações.

- Para se extrair mais conhecimento implícito em uma base de dados, diferentes combinações dos atributos devem ser consideradas, ou seja, a cada execução pode-se considerar um grupo específico dentre todos os atributos contidos na base a ser minerada. Da mesma forma, pode-se classificar os dados tendo-se a cada tempo um atributo meta diferente. Isso requer um grande número de execuções do sistema de mineração para atender às diferentes combinações possíveis, o que torna-se viável quando a adoção do paralelismo é possível.

Para a execução do sistema usando paralelismo foram necessárias algumas adaptações no sistema MinAG de forma que o mesmo pudesse ser executado sem a intervenção do usuário. Para isso foi criado um arquivo de parâmetros, onde devem ser informados os dados que seriam fornecidos ao sistema em tempo de execução. Assim, na execução em paralelo o software de gerenciamento do paralelismo inicia a execução do sistema MinAG, direcionando sua execução para 400 diferentes computadores com um arquivo de parâmetros específico associado a cada execução. Dessa forma, garante-se que em cada máquina estará sendo executado o mesmo sistema, porém com parâmetros diferentes, gerando conseqüentemente resultados diferentes.

O arquivo de parâmetro tem o formato TXT, pode ser gerado individualmente, ou por meio do software de gerenciamento de paralelismo, disponível na Universidade da Flórida, o qual gera a partir de um exemplo e de critérios estabelecidos, vários arquivos de parâmetros com composições diferentes para serem utilizados durante a mineração em paralelo.

Cada linha do arquivo de parâmetros do MinAG corresponde a um parâmetro. A composição do arquivo é a seguinte:

Linha1: nome do arquivo no formato CSV que se deseja abrir (nesse caso o arquivo CSV deve estar no mesmo diretório do MinAG.exe).
Linha2: sim ou nao (existe a coluna ID?).
Linha 3: número da coluna ID (identificador do registro).
Linha 4: número da coluna contendo o atributo meta, atributo objetivo.
Linha 5: nome do arquivo CSV a ser gerado quando não existir a coluna ID ou quando as colunas ID e meta não estiverem, respectivamente, nas posições 0 e 1.
Linha 6: semente para o gerador de números (maior que 0,0 e menor que 1,0).
Linha 7: sim ou nao (deseja dividir o arquivo de dados em uma base de treinamento e uma de teste?).
Linha 8: proporção para divisão da base de dados de treinamento.
Linha 9: nome do arquivo de treinamento: se o arquivo de treinamento já existe, sobrescrever.
Linha 10: proporção para divisão da base de dados de teste.
Linha 11: nome do arquivo de teste: se o arquivo de teste já existe, será sobrescrito.

Linha 12: x (valor inicial do intervalo meta).

Linha 13: y (valor final do intervalo meta).

Linha 14: número de indivíduos na população.

Linha 15: número de gerações.

Linha 16: valor de *fitness* para o indivíduo passar de uma geração para outra.

Linha 17: probabilidade de cruzamento (entre 0 e 1).

Linha 18: probabilidade de mutação (entre 0 e 1).

Linha 19: tamanho do torneio para seleção.

Linha 20: porcentagem de genes mutados do indivíduo.

Linha 21: sim ou nao (deseja mutar o peso?).

Linha 22: limite de 0 a 1 do peso.

Linha 23: probabilidade do peso.

Linha 24: sim ou nao (deseja mutar o operador?).

Linha 25: probabilidade de mutar o operador.

Linha 26: probabilidade de <> no operador.

Linha 27: probabilidade de = no operador.

Linha 28: probabilidade de sair intervalos nos >= e <.

Linha 29: sim ou nao (deseja mutar o valor?).

Linha 30: probabilidade de mutar o valor.

Linha 31: número de subclasses para a roleta.

Linha 32: y1;prop1;x2;y2;prop2;x3;prop3 (x e y são os intervalos para a roleta e prop são as proporções.

Linha 33: nome do arquivo de resultados (result.mga).

Linha 34: sim ou nao (testar).

Linha 35: nome do arquivo de teste (caso não houve a divisão no próprio programa).

Linha 36: sim ou nao (gerar todas as regras de produção).

Linha 37: sim ou nao (gravar todas as regras de produção geradas).

Linha 38: sim ou nao (gerar módulo de avaliação).

Linha 39: sim ou nao (fechar o MinAG após execução).

Observando-se a composição do arquivo de parâmetros percebe-se que existem inúmeras diferentes combinações possíveis dos parâmetros, sendo que a variação de seus valores depende do objetivo da mineração e das características da base de dados sendo considerada.

Com as adequações feitas a versão atual do sistema MinAG pode ser executada tanto no modo convencional (usando apenas um computador) ou em paralelo (*grid*).

5.3 Requisitos e restrições do sistema MinAG

O sistema MinAG possui alguns requisitos em termos de equipamento para a execução do mesmo, bem como em relação às características da base de dados a ser minerada.

5.3.1 Equipamentos

A configuração mínima para que o sistema seja executado consiste em um computador pessoal (PC) com processador de 1.2 GHz, 256 Mb da RAM, Windows 98, XP ou 2000. Porém, dependendo do processamento realizado sobre os dados, do tamanho da base de dados e do número de iterações a serem realizadas na execução do sistema, tanto o tipo de processador como o tamanho da memória RAM precisam ser melhorados, caso contrário, a execução do sistema torna-se muito demorada, correndo-se ainda o risco de travamento do mesmo. Dessa forma sugere-se como configuração razoável um processador de 1.8 GHz ou superior com memória RAM de 512 MB, preferencialmente 1 GB ou superior.

Para a execução em paralelo deve-se distribuir o processamento de forma tal que as execuções do sistema que realizem mais iterações sejam executadas nos computadores que apresentem melhor configuração, ou seja, melhor performance.

5.3.2 Restrições da base de dados

O sistema MinAG foi desenvolvido para atender as especificidades de dos dados relacionados aos fatores físico-químicos do solo e da água, mas pode ser utilizado com outras bases de dados que atendam as seguintes especificações:

- Todos os dados devem ser do tipo numérico contínuo e contidos em uma mesma tabela.
- O arquivo de dados deve ser do formato CSV (Campos Separados por Vírgula).
- Os atributos (variáveis) devem estar dispostos nas colunas e cada linha corresponde a uma observação.
- Não são aceitos campos sem valores, sendo necessário, portanto, que na fase de pré-processamento tais campos sejam preenchidos com um valor que não tenha ocorrido na base de dados.
- As tabelas devem ter um mínimo de 2 atributos e de 200 observações, porém quanto mais observações, desde que com dados de boa qualidade e confiáveis, maior a tendência de se obter bons resultados com a mineração.
- A tarefa específica realizada pelo MinAG é a classificação, portanto, deve ser sempre especificado um atributo que será considerado como meta em cada execução.

5.4 Estudo de caso 1: Agricultura de Precisão

Para a avaliação do sistema desenvolvido, a fim de demonstrar a potencialidade do mesmo, foram definidos dois estudos de casos, inseridos em diferentes contextos. O primeiro estudo de caso refere-se a aplicação do sistema em uma base de dados brasileira gerada por procedimentos de Agricultura de Precisão.

5.4.1 Objetivos desse estudo de caso

O objetivo desse estudo de caso consistiu em demonstrar o uso da mineração de dados na análise das características físico-químicas do solo associadas ou não à produtividade de grãos.

5.4.2 Especificação da base de dados usada

Para este estudo de caso foi utilizada uma base de dados fornecida pelo Professor Doutor José Paulo Molin, pesquisador na área de Agricultura de Precisão na Escola Superior Agricultura Luis Queiroz da Universidade do Estado de São Paulo – ESALQ/ USP. Essa base de dados contém informações sobre uma região de 22 ha localizada em Campos Novos Paulista – SP na qual são desenvolvidas pesquisas relacionadas à Agricultura de Precisão. A base de dados possui as colheitas de 1998 a 2003 (safra de verão e safrinha de outono, com soja, milho, aveia e trigo), com 12 colheitas mapeadas (uma foi perdida por geada e uma por problemas de equipamento). Nessa mesma área existem três amostragens de solo com intervalo de dois anos e um conjunto de dados de compactação com intervalo de 5 cm no perfil do solo, feito em 2004.

O conjunto de dados mostrou-se interessante oferecendo a oportunidade de analisar os dados sob diferentes aspectos. As amostras em campo foram coletadas com o auxílio de aparelhos de GPS. Assim, cada ponto amostrado está associado à sua coordenada XY, que indica sua localização (latitude e longitude) na Terra.

Foram controlados 388 pontos de amostragem em campo, sendo que os métodos empregados para a coleta dos vários dados foram iguais em todas as células e no mesmo período, de forma que a amostra não fosse prejudicada por mudanças climáticas. A partir dessas 388 amostras foram estimados, pelo método de interpolação das áreas vizinhas, mais 2028 pontos com seus respectivos dados resultando então em um total de 2416 pontos controlados ao longo do período de 1998 a 2003. Os valores de produtividade foram obtidos por meio de mapas de produtividades gerados por equipamento de Agricultura de Precisão. Os

valores dos atributos físico-químicos do solo foram obtidos das células amostradas em campo e estimados para os demais pontos por processo de interpolação conforme acima citado.

Os dados estão distribuídos em seis diferentes tabelas, cada uma contendo 2416 registros, sendo que cada registro refere-se a um dos 2416 pontos amostrados controlados. Embora as tabelas refiram-se sempre aos mesmos pontos, diferentes variáveis (atributos) estão presentes em cada tabela, visto que em cada ano diferentes dados foram observados. A seguir as tabelas são detalhadas.

TABELA ANO 1998

A tabela referente ao ano de 1998 apresenta a produtividade do milho em 2416 pontos amostrados. Essa tabela é composta pelos seguintes atributos:

- **Id:** Corresponde ao código individual de identificação do ponto amostrado. Cada código corresponde a uma coordenada XY específica no campo. Por exemplo, o código 67 corresponde a coordenada X -49,985395, e coordenada Y -22,700661. Dessa forma, um mesmo código Id em todos os arquivos corresponde a uma mesma coordenada. Isso é importante por facilitar o agrupamento das tabelas em função desse código.
- **X:** Corresponde à coordenada X no espaço do ponto amostrado.
- **Y:** Corresponde à coordenada Y no espaço do ponto amostrado.
- **Milho:** Esse atributo possui a produtividade do milho no ano 1998 no ponto amostrado definido por X e Y.

TABELA ANO 1999

A tabela referente ao ano de 1999 apresenta a produtividade da soja e do milho em 2416 pontos amostrados, sendo composta pelos seguintes atributos:

- **Id:** Corresponde ao código individual de identificação do ponto amostrado.
- **X:** Corresponde à coordenada X no espaço do ponto amostrado.
- **Y:** Corresponde à coordenada Y no espaço do ponto amostrado.

- **Soja:** Esse atributo possui a produtividade da soja obtida no respectivo ponto amostrado na safra de verão no ano de 1999.
- **Milho Safrinha:** Esse atributo possui a produtividade do milho obtida no respectivo ponto amostrado na safrinha de outono no ano de 1999.
- ***P:*** *teor de* Fósforo no ponto amostrado.
- **K:** teor de Potássio no ponto amostrado.
- **Mg:** teor de Magnésio no ponto amostrado.
- **Ca:** teor de Cálcio no ponto amostrado.
- **Zn:** teor de Zinco.
- **Fe:** teor de Ferro.
- **B:** teor de Boro.
- **Cu:** teor de Cobre.
- **Mn:** teor de Manganês.
- **MO:** teor de Matéria Orgânica no ponto amostrado.
- **H_Al:** Hidrogênio+Alumínio.
- **pH:** acidez do solo observada no ponto.
- **Sb:** Soma de Bases.
- **Saturação_Al:** Saturação por Alumínio.
- **CTC:** capacidade de troca catiônica observada no ponto amostrado.
- **Silte:** teor de silte.
- **Argila:** teor de argila.
- **Areia_mf:** areia média-fina.
- **Areia_f:** areia_fina.
- **Areia_m:** areia média.
- **Areia_g:** areia grossa.
- **Areia_mg:** areia média-grossa.
- **Areia:** teor de areia.

TABELA ANO 2000

A tabela referente ao ano de 2000 apresenta a produtividade da soja em 2416 pontos amostrados, sendo composta pelos seguintes atributos:

- **Id:** Corresponde ao código individual de identificação do ponto amostrado.
- **X:** Corresponde à coordenada X no espaço do ponto amostrado.
- **Y:** Corresponde à coordenada Y no espaço do ponto amostrado.
- **Soja:** Esse atributo possui a produtividade da soja obtida no respectivo ponto amostrado na safra de verão no ano de 2000.

TABELA ANO 2001

A tabela referente ao ano de 2001 apresenta a produtividade da soja e da aveia em 2416 pontos amostrados, sendo composta pelos seguintes atributos:

- **Id:** Corresponde ao código individual de identificação do ponto amostrado.
- **X:** Corresponde à coordenada X no espaço do ponto amostrado.
- **Y:** Corresponde à coordenada Y no espaço do ponto amostrado.
- **Soja:** Esse atributo possui a produtividade da soja obtida no respectivo ponto amostrado na safra de verão no ano de 2001.
- Aveia: Esse atributo possui a produtividade da aveia obtida no respectivo ponto amostrado na safrinha de outono no ano de 2001.
- ***P:*** *teor de* Fósforo no ponto amostrado.
- **K:** teor de Potássio no ponto amostrado.
- **Mg:** teor de Magnésio no ponto amostrado.
- **Ca:** teor de Cálcio no ponto amostrado.
- **S:** teor de Enxofre.
- **Zn:** teor de Zinco.
- **Fe:** teor de Ferro.
- **B:** teor de Boro.
- **Cu:** teor de Cobre.
- **Mn:** teor de Manganês.
- **MO:** teor de Matéria Orgânica no ponto amostrado.

- **H_Al:** Hidrogênio+Alumínio.
- **pH:** acidez do solo observada no ponto.
- **Sb:** Soma de Bases.
- **V:** Percentual de Saturação.
- **CTC:** capacidade de troca catiônica observada no ponto amostrado.

TABELA ANO 2002

A tabela referente ao ano de 2002 apresenta a produtividade da soja e do milho em 2416 pontos amostrados, sendo composta pelos seguintes atributos:

- **Id:** Corresponde ao código individual de identificação do ponto amostrado.
- **X:** Corresponde à coordenada X no espaço do ponto amostrado.
- **Y:** Corresponde à coordenada Y no espaço do ponto amostrado.
- **Soja:** Esse atributo possui a produtividade da soja obtida no respectivo ponto amostrado na safra de verão no ano de 2002.
- **Milho Safrinha:** Esse atributo possui a produtividade do milho obtida no respectivo ponto amostrado na safrinha de outono no ano de 2002.

TABELA ANO 2003

A tabela referente ao ano de 2003 apresenta os fatores físico-químicos e a produtividade da soja e do milho nos 2416 pontos amostrados, sendo composta pelos atributos:

- **Id:** Corresponde ao código individual de identificação do ponto amostrado.
- **X:** Corresponde a coordenada X no espaço do ponto amostrado.
- **Y:** Corresponde a coordenada Y no espaço do ponto amostrado.
- **Soja:** Esse atributo possui a produtividade da soja obtida no respectivo ponto amostrado na safra de verão no ano de 2003.
- **Milho Safrinha:** Esse atributo possui a produtividade do milho obtida no respectivo ponto amostrado na safrinha de outono no ano de 2003.

- ***P:*** *teor de* Fósforo no ponto amostrado.
- **K:** teor de Potássio no ponto amostrado.
- **Mg:** teor de Magnésio no ponto amostrado.
- **Ca:** teor de Cálcio no ponto amostrado.
- **MO:** teor de Matéria Orgânica no ponto amostrado.
- **H_Al:** Hidrogênio+Alumínio.
- **pH:** acidez do solo observada no ponto.
- **Sb:** Soma de Bases.
- **Saturação_Al:** Saturação por Alumínio.
- **CTC:** capacidade de troca catiônica observada no ponto amostrado.
- **IC_0_5mp:** índice de cone em 0 a 5 metros de profundidade.
- **IC_5_10mp:** índice de cone em 5 a 10 metros de profundidade.
- **IC_10_15mp:** índice de cone em 10 a 15 metros de profundidade.
- **IC_15_20mp:** índice de cone em 15 a 20 metros de profundidade.
- **IC_20_25mp:** índice de cone em 20 a 25 metros de profundidade.
- **IC_25_30mp:** índice de cone em 25 a 30 metros de profundidade.
- **IC_30_35mp:** índice de cone em 30 a 35 metros de profundidade.
- **IC_35_40mp:** índice de cone em 35 a 40 metros de profundidade.
- **CE_Deep:** condutividade elétrica profunda.
- **CE_ Shallow:** condutividade elétrica rasa.

O Quadro 4 apresenta uma fração da base de dados do ano de 2003, utilizada nesse estudo de caso.

Quadro 4. Parte da base de dados utilizada.

Id	X	Y	Soja	Milho Safrinha	P	K	Mg	Ca
67	-49,98539534	-22,70066087	3149.30	3599.54	39,5478	2	13,6663	27,1835
68	-49,98529797	-22,70066087	3300.65	1786.06	39,717	2	13,6125	27,1633
69	-49,9852006	-22,70066087	1983.36	2805.07	40,002	2	13,5856	27,1575
70	-49,98510323	-22,70066087	2053.18	2700.52	40,4572	2	13,5896	27,1672
157	-49,98549271	-22,70057104	2256.62	6039.44	39,0154	2	13,6533	27,1199
158	-49,98539534	-22,70057104	2426.74	4283.92	39,0773	2	13,5617	27,0898
159	-49,98529797	-22,70057104	1436.97	2625.93	39,1642	2	13,4862	27,0725
160	-49,9852006	-22,70057104	239.97	3306.47	39,4018	2	13,4507	27,0677
161	-49,98510323	-22,70057104	1188.42	2539.98	39,9242	2	13,4646	27,0764
162	-49,98500586	-22,70057104	2579.51	2868.86	40,7145	2	13,5107	27,101
163	-49,98490849	-22,70057104	3229.57	2913.45	41,623	2	13,5674	27,1468
164	-49,98481112	-22,70057104	3525.10	2722.34	42,5114	2	13,6274	27,2222
247	-49,98559008	-22,70048122	1243.92	7560.76	38,5057	1,9	13,6818	27,0572
248	-49,98549271	-22,70048122	1817.70	6916.78	38,5682	2	13,5726	27,0201
249	-49,98539534	-22,70048122	2190.84	3675.32	38,5674	2	13,4509	27,0034
250	-49,98529797	-22,70048122	1825.39	2113.18	38,5012	2	13,3393	26,9986
251	-49,9852006	-22,70048122	842.79	642.51	38,6085	2	13,2885	26,9981
252	-49,98510323	-22,70048122	1381.14	3568.17	39,2247	2	13,3224	27,0001
253	-49,98500586	-22,70048122	2259.63	3152.00	40,3037	2	13,3978	27,0106
254	-49,98490849	-22,70048122	1377.49	3792.61	41,5069	2	13,4708	27,0445
255	-49,98481112	-22,70048122	2475.90	2766.19	42,5985	2	13,5388	27,1218
256	-49,98471375	-22,70048122	3245.71	2638.82	43,5209	2	13,6164	27,2583
257	-49,98461638	-22,70048122	2614.78	2532.43	44,3058	2	13,7113	27,455
337	-49,98568745	-22,70039139	1023.44	7783.41	37,9284	1,9	13,7278	26,9921
338	-49,98559008	-22,70039139	1917.17	7444.48	37,9932	1,9	13,6251	26,9308
339	-49,98549271	-22,70039139	2864.52	3677.22	38,0985	2	13,5043	26,9124
340	-49,98539534	-22,70039139	2648.74	5277.28	38,0866	2	13,3522	26,9294
341	-49,98529797	-22,70039139	1698.78	4930.46	37,817	2	13,1895	26,9605
342	-49,9852006	-22,70039139	2088.49	3180.84	37,6645	2	13,1133	26,9754
343	-49,98510323	-22,70039139	1377.76	2985.22	38,419	2	13,1821	26,9571
344	-49,98500586	-22,70039139	2592.69	7098.90	39,9512	2	13,2937	26,9266

5.4.3 Pré-processamento da Base de Dados

Uma vez definida a base de dados a ser usada a fase de pré-processamento é executada de forma a preparar a base para ser submetida ao algoritmo de mineração de dados. O pré-processamento foi feito utilizando o software Microsoft Excel® e consistiu em:

- **Verificar se o formato da tabela é adequado para a mineração**

As tabelas originais apresentavam-se de acordo com esse requerimento do sistema, visto que para realizar a mineração é necessário que cada coluna corresponda a um atributo diferente e cada linha represente um registro específico (uma observação).

- **Verificar a existência de caracteres conflitantes com a configuração do sistema**

Como pode ser observado no Quadro 4, o qual apresenta uma parte da base de dados original, alguns atributos apresentavam-se com o separador decimal sendo o ponto (.), o que não é aceito pelo sistema que foi configurado para aceitar como separador decimal a vírgula (,). Foi necessário então alterar os separadores para que todos estivessem dentro do padrão requerido pelo MinAG.

- **Verificar a existência de valores absurdos ou faltantes**

A base de dados fornecida apresentou-se com poucos erros no pré-processamento sendo que na verificação de valores de certa forma incoerentes a equipe responsável pelo fornecimento da base prontamente fez as correções necessárias nos valores.

- **Verificar se as unidades de medida utilizadas estão padronizadas**

Considerando que o mesmo tipo de informação aparece em diferentes tabelas, é indispensável verificar se um mesmo atributo está sendo expresso sempre em uma mesma unidade de medida.

Por meio dessa verificação identificou-se que a produtividade estava sendo apresentava em algumas tabelas em Kg/ha e em outras em Ton/ha. Nesse caso, foi necessário converter os valores de alguns atributos para realizar essa padronização, tendo sido adotado como padrão para a produtividade a unidade de medida Ton/ha.

- **Verificar a existência da primeira linha como sendo o cabeçalho**

Todas as tabelas já continham a primeira linha como sendo o cabeçalho, porém ocorreram casos em que um mesmo atributo recebeu diferentes nomes nas tabelas. Por exemplo, o atributo que representa o elemento Potássio recebeu o nome de “Potássio” na

tabela do ano de 1999 e o nome “K” na tabela de 2003. Muito embora seja óbvio que o nome “K” corresponde à representação do Potássio na Tabela Periódica, é adequado que sempre se adote um mesmo nome para facilitar a junção das tabelas, bem como a compreensão das regras geradas.

Solucionou-se essa questão realizando a padronização para todos os atributos. No caso citado acima padronizou-se para o elemento Potássio o nome de atributo “K”, tendo sido adotado para os elementos químicos os nomes conforme são representados na Tabela Periódica e para os demais atributos foram usadas siglas evitando assim nomes longos.

- **Eliminar atributos (colunas) desnecessários**

Considerando que o sistema MinAG tem como objetivo gerar regras sem o mapeamento da localização dos pontos observados, ou seja, o objetivo da mineração nesse estudo de caso é correlacionar os valores dos atributos desconsiderando seu posicionamento geográfico, as colunas X e Y foram eliminadas das tabelas após ter sido conferido que cada If (Código de identificação) correspondia efetivamente a uma mesma coordenada X,Y. A manutenção desses atributos se justificaria se no estudo de caso (na mineração) fosse de interesse que as regras apresentassem as posições de coordenadas fazendo parte das condições para se classificar um determinado atributo.

Vale ressaltar aqui que está previsto como trabalho futuro, a incorporação de um módulo de mapeamento das regras no MinAG, no qual o usuário poderá verificar no mapa da região a localização geográfica dos pontos em que uma determinada regra é satisfeita. Nesse caso, ou seja, a partir dessa incorporação os atributos referentes às coordenadas X e Y serão considerados no sistema.

- **Consolidar tabelas**

Após análise do conteúdo das tabelas e dos objetivos da mineração foram definidas algumas formas de consolidação das tabelas, o que se fez necessário devido ao fato de que as tabelas possuíam diferentes atributos, sendo que alguns deles estavam presentes em apenas uma tabela, porém deveriam ser analisados juntamente com atributos

existentes em outras tabelas. Por exemplo, a tabela do ano de 2000, após ter sido eliminados os atributos de coordenada X e Y, continha apenas os atributos Id (código de identificação) e Soja (produtividade da soja).

Portanto, para que fosse possível realizar a tarefa de classificação sobre os dados de produtividade da soja no ano 2000 em função dos fatores físico-químicos, seria necessário associar essa tabela a outra que contivesse dados físico-químicos do solo. Da mesma forma, a tabela de 1998 conteve somente os atributos Id (código de identificação) e Milho (produtividade do milho). Sendo assim, decidiu-se consolidar os dados das tabelas dos anos 1998, 1999 e 2000 em uma nova tabela denominada PA98_99_00. Isso só foi possível pelo fato de existir o atributo Id (Código de identificação) que permite associar um mesmo ponto em diferentes tabelas, visto que para cada ponto existe um código de identificação único.

Outra consolidação feita refere-se à junção das tabelas dos anos 2001 e 2002 em uma nova tabela denominada PA01_02. Essa consolidação se fez necessária para que as produtividades da soja e do milho no ano de 2002 pudessem ser correlacionadas com os fatores físico-químicos do solo.

Considerando que o objetivo desse estudo de caso foi identificar padrões de comportamento dos atributos do solo associados a produtividade e também dissociados dos atributos de produtividade, decidiu-se criar também uma tabela que consolidasse os valores dos atributos físico-químicos do solo, sem nenhum atributo referente à produtividade de forma a poder executar a mineração desses dados buscando identificar a interferências de determinadas características físicas do solo (condutividade elétrica, índice de cone, etc) na variabilidade dos demais elementos presentes no mesmo.

- **Gerar tabela em formato CSV**

Uma vez organizadas as tabelas o próximo passo foi gerar uma versão de cada tabela no formato CSV (Campos Separados por Vírgula) conforme é requerido pelo MinAG.

- **Gerar os arquivos de parâmetros para execução do sistema em paralelo**

Tendo-se as tabelas prontas para serem mineradas o sistema oferece duas opções: realizar cada mineração separadamente informando ao sistema o arquivo a ser usado bem como os parâmetros em tempo de execução, ou a realização da mineração dos dados usando o sistema de paralelismo. Em se adotando a primeira opção não seria necessário criar os arquivos de parâmetros. Porém, considerando os inúmeros benefícios do paralelismo e a disponibilidade de usar esse ambiente na Universidade da Flórida optou-se por executar a mineração em paralelo.

Conseqüentemente, fazendo-se uso das facilidades do ambiente de execução em *grid*, foi criado um arquivo de parâmetros para cada tabela da base de dados. Foram definidos, no total, 64 arquivos de dados e conseqüentemente 64 arquivos de parâmetros. Considerando que 58 arquivos de parâmetros tinham 60 combinações diferentes e 6 apresentaram 40 combinações, o total de combinações definidas para essa base de dados foi de 3720. Isso significa que o sistema MinAG foi executado 3720 vezes para realizar todas as minerações com as diferentes combinações estabelecidas.

5.4.4 Execução da mineração dos dados

As 3720 execuções do MinAG foram realizadas na Universidade da Flórida. O sistema foi executado em paralelo em 400 computadores do tipo PC (Personal Computer). A configuração desses equipamentos é basicamente Pentim III ou IV, com 256 a 512 Mb de memória RAM e processador de 1.2 a 1.8 GHz. Os arquivos de parâmetros que exigiam mais tempo de processamento e mais memória de máquina em função do número de iterações a serem feitas pelo sistema, foram alocados nas máquinas que apresentavam melhor performance (processador de 1.8 GHz e 512 Mb de memória RAM).

5.4.5 Regras Geradas

Ao final da mineração dos dados foram geradas inúmeras regras correlacionando e identificando os padrões entre os dados.

Os resultados das 3720 execuções foram divididos em três grupos: grupo de resultados com regras com *fitness* aceitável (>= 0,6), resultados com regras com *fitness* inaceitável (< 0,6) e o último grupo consistiu nos resultados em que não foram geradas regras. A qualidade da combinação dos valores adotados para os parâmetros só pôde ser identificada justamente após a execução do sistema. Assim, ao avaliar o conjunto de regras resultantes com *fitness* aceitável, foi possível identificar os valores de parâmetros mais adequados para os dados desse estudo de caso.

Para a interpretação das regras deve-se considerar que:

- A parte SE da regra apresenta a conjunção de valores dos atributos de predição considerados na regras.
- A parte ENTÃO apresenta a classe (faixa de valores do atributo meta) a qual a regra pertence.
- A média do atributo meta (treino) indica o valor médio obtido para o atributo meta, considerando todos os registros que satisfizeram a regra durante o treino.
- O *fitness* de treino indica o valor de *fitness* da regra obtido durante a fase de treinamento, ou seja, utilizando a base de dados de treinamento.
- A média do atributo meta (teste) indica o valor médio obtido para o atributo meta, considerando todos os registros que satisfizeram a regra durante o teste.
- O *fitness* de treino indica o valor de *fitness* da regra obtido durante a fase de teste, ou seja, utilizando a base de dados de teste.

Tomemos como exemplo a regra [0]:

Regra[0]: **SE 1,0583>=Potássio>=0,3465 E Soma_Bases=53,58 E CTC=72,24 ENTÃO 4,01<= Soja00<=4,68**

{Média do Atributo Meta (Treino): 4,1921} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 0} {*Fitness* (Teste): 0}

A interpretação da regra[0] pode ser feita da seguinte maneira: se o valor do potássio estiver entre 1,0583 e 0,3465 e a soma das bases for igual a 53,58 e a capacidade de troca catiônica (CTC) for igual a 72,24 então a produtividade da soja está entre 4,01 e 4,68. Durante o treinamento o valor médio encontrado para o atributo meta nos registros que satisfazem essa regra foi 4,1921, o que indica que, embora o valor do atributo meta esteja dentro da classe desejada (entre 4,01 e 4,68), em média seu valor foi 4,1921, um pouco abaixo do ponto médio da classe. O valor de *fitness* durante o treinamento foi 1, ou seja, excelente. Porém, durante os testes, que validam a regra na base de dados de teste, o resultado foi o oposto, ou seja, a média do atributo meta foi zero, o que indica que não foi encontrado nenhum registro que satisfizesse a regra, e conseqüentemente o *fitness* no teste foi zero também, indicando que a regra é fraca e não pode ser generalizada.

Algumas das combinações de parâmetros utilizadas na mineração podem ser observadas no Quadro 5 e algumas das regras geradas com esses arquivos de parâmetros podem ser vistas a seguir.

R_PA00SC58.MGA

Regra[1]: **SE Potássio=0,61 E 140,5>=Soma_Bases>=53,58 E CTC=72,24 ENTÃO 4,01<= Soja00<=4,68**

{Média do Atributo Meta (Treino): 4,1921} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 0} {*Fitness* (Teste): 0}

Regra[2]: SE Soma_Bases=53,58 E CTC=72,24 ENTÃO 4,01<= Soja00<=4,68 {Média do Atributo Meta (Treino): 4,1921} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 0} {*Fitness* (Teste): 0}

R_PA01AC56.MGA

Regra[0]: **SE 75,25>=V%>=66,3326 E Areia_m=31,7 ENTÃO 1,51<= Aveia01<=1,77** {Média do Atributo Meta (Treino): 1,61} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 1,54} {*Fitness* (Teste): 1}

Regra[1]: **SE 11,3624<Mg<13,54 E Areia_m=31,7 ENTÃO 1,51<= Aveia01<=1,77** {Média do Atributo Meta (Treino): 1,61} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 1,54} {*Fitness* (Teste): 1}

Regra[7]: **SE 5,554>=pH>=5,4093 E 35,11>=Argila>=19,9821 E 42,91>=Areia_f>=27,6024 E Areia_m=31,7 E 2,0447<Areia_g<3,95 ENTÃO 1,51<= Aveia01<=1,77** {Média do Atributo Meta (Treino): 1,52} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 1,54} {*Fitness* (Teste): 1}

Regra[11]: **SE 59,08>=P>=51,6416 E 5,554>=pH>=5,3162 E Areia_m=31,7 E 2,0447<Areia_g<3,95 ENTÃO 1,51<= Aveia01<=1,77** {Média do Atributo Meta (Treino): 1,52} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 1,54} {*Fitness* (Teste): 1}

Regra[12]: **SE 19,9821<Argila<27,6947 E Areia_m=31,7 ENTÃO 1,51<= Aveia01<=1,77** {Média do Atributo Meta (Treino): 1,61} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 1,54} {*Fitness* (Teste): 1}

Regra[36]: **SE 71,7331<CTC<81,21 E Areia_m=31,7 E 3,95>=Areia_g>=2,0447 ENTÃO 1,51<= Aveia01<=1,77** {Média do Atributo Meta (Treino): 1,61} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 1,54} {*Fitness* (Teste): 1}

R_PA03MsE36.MGA

Regra[0]: **SE Areia_g=0,86 E 0<Areia_mg<0,64 E 81,0395>=Areia>=76 ENTÃO 3,01<= Milho Safrinha<=6**

{Média do Atributo Meta (Treino): 5,125} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 0} {*Fitness* (Teste): 0}

Regra[1]: **SE 29,6169<Areia_f<42,91 ENTÃO 3,01<= Milho Safrinha<=6**

{Média do Atributo Meta (Treino): 3,8789} {*Fitness* (Treino): 0,8274} {Média do Atributo Meta (Teste): 3,8465} {*Fitness* (Teste): 0,8358}

Regra[4]: **SE 42,91>=Areia_f>=33,28 ENTÃO 3,01<= Milho Safrinha<=6**

{Média do Atributo Meta (Treino): 4,0756} {*Fitness* (Treino): 0,7273} {Média do Atributo Meta (Teste): 3,9875} {*Fitness* (Teste): 0,7769}

Regra[6]: **SE 3,95>=Areia_g>=0,86 E 76<Areia<91,58 ENTÃO 3,01<= Milho Safrinha<=6**

{Média do Atributo Meta (Treino): 3,8362} {*Fitness* (Treino): 0,831} {Média do Atributo Meta (Teste): 3,7949} {*Fitness* (Teste): 0,8354}

Regra[11]: **SE Areia_g=0,86 E 0,64>=Areia_mg>=0 ENTÃO 3,01<= Milho Safrinha<=6**

{Média do Atributo Meta (Treino): 5,074} {*Fitness* (Treino): 0,8} {Média do Atributo Meta (Teste): 3,984} {*Fitness* (Teste): 0,8}

Regra[16]: **SE 29,6169<Areia_f<42,91 E 0,64>=Areia_mg>=0 E 81,0395>=Areia>=76 ENTÃO 3,01<= Milho Safrinha<=6**

{Média do Atributo Meta (Treino): 3,7378} {*Fitness* (Treino): 0,8974} {Média do Atributo Meta (Teste): 3,7717} {*Fitness* (Teste): 0,9077}

Quadro 5. Composição de três dos arquivos de parâmetro utilizados.

nome do arquivo aabrir	PA03Ms_E.CSV	PA01A_C.CSV	PA00S_F.CSV
existe a coluna ID?	sim	sim	sim
número da coluna id	0	0	0
número da coluna do atributo meta	1	1	1
nome do arquivo temporario a gerar	NOME.CSV	NOME.CSV	NOME.CSV
semente para o gerador	0,3333	0,3333	0,3333
dividir o arquivo em treinamento e teste?	sim	sim	sim
proporção para a base de treinamento	50	50	50
nome do arquivo de treinamento	TREIN.CSV	TREIN.CSV	TREIN.CSV
proporção para a base de teste	50	50	50
nome do arquivo de teste	TEST.CSV	TEST.CSV	TEST.CSV
x (inicio intervalo meta)	3,01	1,51	0,78
y (fim intervalo meta)	6	1,77	2
número de indivíduos na população	20	38	38
número de gerações	200	200	50
valor de fitness	0,7	0,7	0,7
probabilidade de cruzamento	0,95	0,95	0,95
probabilidade de mutação	0,9	0,9	0,9
tamanho do torneio para seleção	3	3	3
porcentagem de genes mutados	0,4	0,4	0,4
deseja mutar o peso?	sim	sim	nao
limite de 0 a 1 do peso	0,95	0,95	1
probabilidade do peso	0,5	0,5	0
deseja mutar o operador?	sim	sim	sim
probabilidade de mutar o operador	0,95	0,95	0,95
probabilidade de <> no operador	0	0	0
probabilidade de = no operador	0	0	0
probabilidade de sair intervalos nos >= e <	1	1	1
deseja mutar o valor?)	sim	sim	sim
probabilidade de mutar o valor	0,0625	0,0625	0,0025
número de subclasses para a roleta	2	2	2
y1;prop1;x2;y2;prop2;x3;prop3	4,5;50;4,5;50	1,6;50;1,6;50	1;50;1;50
nome do arquivo mga	R_PA03MsE36.mga	R_PA01AC56.mga	R_PA00SF9.mga
testar?	sim	sim	sim
nome do arquivo de teste	TESTE.csv	TESTE.csv	TESTE.csv
gerar todas as regras de produção	sim	sim	sim
gravar todas as regras de produção geradas	sim	sim	sim
gerar módulo de avaliação	sim	sim	sim
nome do arquivo aabrir	sim	sim	sim

5.4.6 Discussão dos resultados

Conforme pode ser observado cada regra resultante da mineração é apresentada acompanhada do valor médio do atributo meta tanto na fase de treinamento quanto na fase de testes. Esse valor corresponde à média do valor do atributo meta, considerando os valores do mesmo presente na base de dados em cada ocorrência que confirmou a regra gerada. Essa informação é importante visto que, uma vez estabeleceu-se um intervalo de interesse para o atributo meta (0,78 a 2 por exemplo), é interessante saber qual o valor médio observado para tal atributo nas ocorrências da tabela que foram consideradas na criação (treinamento) e avaliação (teste) da regra.

Os valores de *fitness* de treino e de teste também são apresentados em cada regra para que o usuário possa identificar qual a qualidade da regra em termos de generalização, ou seja, o quanto significativa é a regra em relação à base de dados tanto de treinamento como de teste.

Os resultados contidos em R_PA00SC58.MGA não foram bons visto que a função de *fitness* apresentou valor zero na base de teste para várias regras. Analisando tais regras verificou-se que embora na base de treinamento a regra foi considerada boa (*fitness* 1), na base de teste a regra não se confirmou (*fitness* 0), o que significa que essa regra não pode ser generalizada. Nesse caso pode-se considerar que a combinação de parâmetros adotada não foi eficiente

Os demais arquivos de resultados considerados aqui (R_PA01AC56.MGA e R_PA03MsE36.MGA) apresentaram regras melhores no sentido em que foram identificados regras com elevado *fitness* (por exemplo , >0,8) tanto na base de teste como de treinamento.

5.5 Estudo de caso 2: Qualidade da Água

O segundo estudo de caso refere-se a uma base de dados selecionada durante o estágio de doutoramento realizado pela autora dessa tese na Universidade da Flórida, nos Estados Unidos.

5.5.1 Objetivo desse estudo de caso

O objetivo desse estudo de caso foi demonstrar a utilização do sistema MinAG para identificar padrões de comportamento nos dados das propriedades físico-químicas da água no Estado da Flórida – EUA, considerando os critérios adotados nos Estados Unidos sobre a qualidade da água para a vida aquática.

5.5.2 Especificação da base de dados usada

A base de dados das propriedades físico-químicas da água no Estado da Flórida – EUA foi disponibilizada pelo Centro de Pesquisa de Informações e Planejamento de Geo-Facilidades – GEOPLAN (do inglês Geo-Facilities Planning and Information Research Center) setor da Universidade da Flórida responsável pelo armazenamento e manutenção dos dados geográficos do estado da Flórida, em parceria com a Agência de Proteção Ambiental dos Estados Unidos, a qual atende pela sigla U.S. EPA (do inglês United States Environmental Protection Agency) e consiste em uma agência do governo americano responsável por desenvolver programas para proteger a saúde humana e o ambiente (EPA, 2005).

A U.S. EPA desenvolveu um vasto banco de dados que provê resumo estatístico do monitoramento da qualidade da água para 47 parâmetros físicos e químicos relacionados, tendo sido os dados coletados em todo o país. Foram selecionadas estações de monitoramento caracterizadas como lago, reservatório, canal, estuário ou oceano (EPA, 2005).

Essa base foi preparada para dar suporte ao U.S. EPA BASINS (Better Assessment Science Integrating Point and Nonpoint Sources System), que consiste em um sistema norte-americano de análise ambiental com multi-propósitos que integra um sistema de informação geográfica (SIG), dados nacionais importantes e ferramentas de modelagem e análise ambiental. O sistema BASINS está atualmente na versão 3.1 e maiores informações do mesmo podem ser obtidas em http://www.epa.gov/waterscience/basins/index.html. A base de dados desenvolvida pela U.S. EPA possui dados que resultam de uma contribuição de várias

organizações incluindo agências federais, estaduais e interestaduais, universidades e laboratórios de água. A Universidade da Flórida tem acesso aos dados referentes ao estado da Flórida, os quais foram disponibilizados para serem usados nesse estudo de caso.

Considerando que o objetivo desse estudo de caso consiste somente na demonstração da aplicação do sistema nessa base de dados, não é de interesse para esse trabalho discutir a maneira de obtenção e de análise das amostras de água. Maiores detalhes sobre esse assunto podem ser obtidos em http://geoplan.ufl.edu.

O conteúdo da base de dados aqui usada refere-se a resumos estatísticos coletados pelas estações de monitoramento para intervalos de cinco anos desde 1970 a 1994 e um intervalo de três anos de 1995 a 1997. As estatísticas incluem o número de observações, desvio padrão, média, e os percentis 15, 25, 50, 75 e 85. Além disso, um identificador (ID) foi assinalado para cada estação para facilitar ligações entre as tabelas existentes na base de dados.

A base de dados era composta originalmente por seis arquivos em formato DBF (Data Base File), sendo que todos apresentaram tabelas relacionadas a observações de qualidade da água contendo informação estimada de elementos físico-químicos e toxinas. Os arquivos trabalhados foram:

WQ70_74: tabela referente ao período de 1970 a 1974, contendo 51741 registros.
WQ75_79: tabela referente ao período de 1975 a 1979, contendo 60838 registros.
WQ80_84: tabela referente ao período de 1980 a 1984, contendo 45726 registros.
WQ85_89: tabela referente ao período de 1985 a 1989, contendo 49178 registros.
WQ90_94: tabela referente ao período de 1990 a 1994, contendo 55041 registros.
WQ95_97: tabela referente ao período de 1995 a 1997, contendo 32650 registros.

Cada tabela era composta pelos seguintes atributos:

- **ID:** Número único para cada estação definido pelo sistema BASINS.
- **AGENCY:** Código da agência.
- **STATION:** Código da estação.
- **BWQID:** Número da estação definida pelo sistema BASINS.

- **PARM CODE:** Código do parâmetro físico-químico definido pela U.S. EPA.
- **NO OBS:** Número de observações.
- **MEAN:** Valor médio das observações.
- **A15TH_P:** valor do 15º percentil.
- **A25TH_P:** valor do 25º percentil.
- **A50TH_P**: valor do 50º percentil.
- **A75TH_P:** valor do 75º percentil.
- **A85TH_P:** valor do 85º percentil.
- **STD:** Desvio padrão em relação ao valor médio das observações.

O Quadro 6 apresenta uma parte da tabela original do período de 1990 a 1994.

Quadro 6. Parte da tabela original de dados do período de 1990 a 1994.

ID	AGENCY	STATION	BWQID	PARM_CODE	NO_OBS	MEAN	A15TH_P	A25TH_P	A50TH_P	A75TH_P
01-01+21FLPDEM	21FLPDEM	01-01	32083	00630	32	0,0046800	0,0000000	0,0000000	0,0000000	0,0000000
01-01+21FLPDEM	21FLPDEM	01-01	32083	32211	32	3,3502400	1,1990000	1,4250000	2,2750000	4,9850000
01-01+21FLPDEM	21FLPDEM	01-01	32083	00665	24	0,0874900	0,0000000	0,0600000	0,0799900	0,1225000
01-01+21FLPDEM	21FLPDEM	01-01	32083	00625	23	0,6186900	0,3300000	0,4100000	0,5400000	0,8700000
01-01+21FLPDEM	21FLPDEM	01-01	32083	00612	23	0,0009500	0,0000000	0,0003600	0,0006700	0,0014000
01-01+21FLPDEM	21FLPDEM	01-01	32083	00610	23	0,0373900	0,0000000	0,0200000	0,0300000	0,0600000
01-01+21FLPDEM	21FLPDEM	01-01	32083	00400	75	7,7168600	7,3820000	7,5300000	7,7399900	7,9499900
01-01+21FLPDEM	21FLPDEM	01-01	32083	00310	30	1,1666600	0,0000000	0,7500000	1,2500000	1,7250000
01-01+21FLPDEM	21FLPDEM	01-01	32083	00010	75	23,6735000	17,1680000	19,0100000	24,2500000	28,4300000
01-01+21FLPDEM	21FLPDEM	01-01	32083	00671	33	0,0278700	0,0000000	0,0000000	0,0000000	0,0600000
01-02+21FLPDEM	21FLPDEM	01-02	32084	00310	14	1,7000000	1,1250000	1,3500000	1,5500000	2,0000000
01-02+21FLPDEM	21FLPDEM	01-02	32084	00625	10	0,5260000	0,3130000	0,3425000	0,5100000	0,6675000
01-02+21FLPDEM	21FLPDEM	01-02	32084	32211	14	3,9779200	1,3700000	1,9625000	2,8500000	5,7000000
01-02+21FLPDEM	21FLPDEM	01-02	32084	00671	14	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000
01-02+21FLPDEM	21FLPDEM	01-02	32084	00665	10	0,0480000	0,0000000	0,0000000	0,0250000	0,1050000
01-02+21FLPDEM	21FLPDEM	01-02	32084	00630	14	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000
01-02+21FLPDEM	21FLPDEM	01-02	32084	00612	10	0,0008700	0,0000000	0,0000000	0,0007000	0,0010100
01-02+21FLPDEM	21FLPDEM	01-02	32084	00400	27	7,9822100	7,8340000	7,8800000	7,9700000	8,0700000
01-02+21FLPDEM	21FLPDEM	01-02	32084	00010	27	23,8288000	16,0760000	20,1100000	24,1200000	28,6600000
01-02+21FLPDEM	21FLPDEM	01-02	32084	00610	10	0,0190000	0,0000000	0,0000000	0,0200000	0,0300000
01-03+21FLPDEM	21FLPDEM	01-03	32085	00625	12	0,7458300	0,4440000	0,6050000	0,7150000	0,9824900
01-03+21FLPDEM	21FLPDEM	01-03	32085	00310	16	1,1187500	0,0000000	1,0000000	1,1500000	1,4000000
01-03+21FLPDEM	21FLPDEM	01-03	32085	32211	16	4,5025000	2,1500000	2,8475000	4,0500000	6,1825000
01-03+21FLPDEM	21FLPDEM	01-03	32085	00671	16	0,0768700	0,0000000	0,0125000	0,0600000	0,0974900
01-03+21FLPDEM	21FLPDEM	01-03	32085	00665	12	0,1741600	0,0569900	0,0699900	0,1050000	0,1500000
01-03+21FLPDEM	21FLPDEM	01-03	32085	00630	16	0,0256200	0,0000000	0,0000000	0,0000000	0,0375000
01-03+21FLPDEM	21FLPDEM	01-03	32085	00400	22	7,2240800	6,9935000	7,0400000	7,2000000	7,3525000
01-03+21FLPDEM	21FLPDEM	01-03	32085	00010	22	24,2381000	18,4455000	20,9700000	24,6800000	27,7400000
01-03+21FLPDEM	21FLPDEM	01-03	32085	00610	12	0,0258300	0,0000000	0,0100000	0,0200000	0,0400000
01-03+21FLPDEM	21FLPDEM	01-03	32085	00612	12	0,0003700	0,0000000	0,0001000	0,0001300	0,0004400
01-04+21FLPDEM	21FLPDEM	01-04	32086	00612	12	0,0011000	0,0000900	0,0005000	0,0008200	0,0018500
01-04+21FLPDEM	21FLPDEM	01-04	32086	32211	16	12,8581000	3,2750000	4,7325000	12,2000000	14,9750000

Além das tabelas de dados acima citadas, a base de dados era composta também pelas tabelas:

TABELA WQPARM

Essa tabela contém a descrição dos parâmetros adotados para os critérios de vida aquática. Contém 30 registros, sendo composta pelos seguintes atributos:

- **PAR_CODE:** Código do parâmetro, definido pela U.S. EPA para relacionar essa tabela com as tabelas de observações.
- **PARM_NAME:** Nome do parâmetro, o qual vem a ser um elemento físico-químico ou uma toxina.
- **UNITS:** Unidade de medida empregada.
- **SAMPLE_TYPE:** tipo de amostra.
- **UP_REF_LVL:** Maior nível de referência aceitável.
- **LW_REF_LVL:** Menor nível de referência aceitável.
- **UNKNOWN:** tipo de uso da água para o intervalo de referência contido em UP_REF_LVL e LW_REF_LVL.
- **REF_LVLSRC:** Fonte de informação dos intervalos de referência adotados.

O Quadro 7 apresenta os parâmetros adotados segundo os critérios para vida aquática. Para se executar a mineração em função de outros tipos de uso da água, como para consumo, por exemplo, bastaria adotar os critérios para o tipo de uso especificado.

Quadro 7. Parâmetros para os critérios de uso para vida aquática.

PARM_CODE	PARM_NAME	UNITS	SAMPLE_TYP	UP_REF_LVL	LW_REF_LVL
00010	TEMPERATURE, WATER	C		32,20	0,00
00095	SPECIFIC CONDUCTANCE	UMHOS/CM AT 25C			
00300	OXYGEN, DISSOLVED	MG/L	DISSOLVED	0,00	5,00
00310	BOD, 5 DAY, 20 DEG C	MG/L	TOTAL	7,00	0,00
00400	PH	SU		9,00	6,50
00410	ALKALINITY, TOTAL (AS CACO3)	MG/L AS CACO3	TOTAL	400,00	20,00
00515	RESIDUE, TOTAL FILTRABLE DRIED AT 105C (TDS)	MG/L	DISSOLVED		
00530	RESIDUE, TOTAL NONFILTRABLE (TSS)	MG/L	TOTAL	500,00	0,00
00610	NITROGEN, AMMONIA, TOTAL	MG/L AS N	TOTAL	15,70	0,00
00612	AMMONIA, UNIONZED	MG/L AS N	TOTAL	93,00	0,00
00620	NITRATE NITROGEN, TOTAL	MG/L AS N	TOTAL		
00625	NITROGEN, KJELDAHL, TOTAL	MG/L AS N	TOTAL	0,00	0,00
00630	NITRITE PLUS NITRATE, TOTAL 1 DET.	MG/L AS N	TOTAL	0,00	0,00
00631	NITRITE PLUS NITRATE, DISS. 1 DET.	MG/L AS N	DISSOLVED		
00665	PHOSPHORUS, TOTAL	MG/L AS P	TOTAL	1,00	0,00
00671	PHOSPHORUS, DISSOLVED ORTHOPHOSPHATE	MG/L AS P	DISSOLVED	0,00	0,00
00900	HARDNESS, TOTAL	MG/L AS CACO3	TOTAL	200,00	0,00
00940	CHLORIDE,TOTAL IN WATER	MG/L	TOTAL	860,00	0,00
00945	SULFATE, TOTAL	MG/L AS SO4	TOTAL	250,00	0,00
01000	ARSENIC, DISSOLVED	UG/L	DISSOLVED		
01005	BARIUM, DISSOLVED	UG/L	DISSOLVED		
01025	CADMIUM, DISSOLVED	UG/L	DISSOLVED	3,90	0,00
01040	COPPER, DISSOLVED	UG/L	DISSOLVED	18,00	0,00
01046	IRON, DISSOLVED	UG/L	DISSOLVED		
01049	LEAD, DISSOLVED	UG/L	DISSOLVED	82,00	0,00
01065	NICKEL, DISSOLVED	UG/L	DISSOLVED		
01090	ZINC, DISSOLVED	UG/L	DISSOLVED		
01106	ALUMINUM, DISSOLVED	UG/L	DISSOLVED		
32730	PHENOLICS, TOTAL, RECOVERABLE	UG/L	TOTAL	10200,00	0,00
71900	MERCURY, TOTAL	UG/L	TOTAL	2,40	0,00

TABELA WQOBS

Essa tabela contém a descrição das agências e estações provedoras dos dados observados. A tabela WQOBS é composta pelos seguintes atributos:

- **ID:** Número único para cada estação definido pelo sistema BASINS.
- **AGENCY:** Código da agência.
- **AGENCY_COD:** Código complementar da agência.
- **STATION:** Código da estação.
- **ST_DEPTH:** profundidade em que a amostra foi coletada.
- **STATE:** Código do Estado – Nesse caso é sempre 42 – Flórida.
- **LAT:** Latitude da estação.
- **LONG:** Longitude da estação.
- **TYPE:** Tipo de estação (pequeno rio, lago, etc).
- **LOCATION:** descrição da localização da estação.

O Quadro 8 apresenta uma parte da tabela WQOBS, cujas linhas referem-se às estações de monitoramento localizadas no estado da Flórida, o que justifica o campo STATE ser sempre igual a 42.

Quadro 8. Parte da tabela WQOBS.

ID	AGENCY	AGENCY_COD	STATION	ST_DEPTH	STATE	LAT	LONG	TYPE
03374	112WRD	1	03039925	0	42	40,26612	-79,01695	/TYPA/AMBNT/STREAM
03371	11COEHUN	0	4CONW0105	10	42	40,44973	-79,28834	/TYPA/AMBNT/LAKE
03372	11COEHUN	0	4CONW0106	10	42	40,41625	-79,28292	/TYPA/AMBNT/STREAM
03373	11COEHUN	0	4CON20201	20	42	40,46139	-79,36806	/TYPA/AMBNT/LAKE
03375	21PA	0	WQN0810	0	42	40,45445	-79,39112	/TYPA/AMBNT/STREAM/BIO
03376	21PA	0	WQN0814	0	42	40,47334	-79,18362	/TYPA/AMBNT/STREAM/BIO
03377	21PA	0	WQN0816	0	42	40,33000	-78,90723	/TYPA/AMBNT/STREAM
03378	21PA	0	WQN0817	0	42	40,29362	-78,91889	/TYPA/AMBNT/STREAM
03379	21PA	0	WQN0864	0	42	40,67473	-78,94445	/TYPA/AMBNT/STREAM/BIO

5.5.3 Pré-processamento da Base de Dados

Uma vez definida a base de dados a ser usada a fase de pré-processamento foi executada de forma a preparar a base para ser submetida ao algoritmo de mineração de dados.

O pré-processamento realizado para essa base de dados foi complexo, requerendo inclusive o desenvolvimento de um programa específico para manipular as tabelas. Vejamos a seguir os passos seguidos em cada fase do pré-processamento.

- **Verificar a adequação do formato da tabela para a mineração**

As tabelas originais não se apresentavam de acordo com o formato requerido pelo sistema porque cada linha da tabela de dados representava um parâmetro diferente. Segundo os requerimentos do sistema é necessário que os parâmetros a serem minerados estejam dispostos em colunas e que cada linha da tabela corresponda a uma observação de todo o conjunto de parâmetros. Sendo assim em função do grande volume de registros, foi necessário desenvolver um programa específico para realizar essa etapa do pré-processamento, convertendo as tabelas em um formato adequado.

- **Verificar a existência de caracteres conflitantes com a configuração do sistema**

O caractere usado como separador da parte decimal nas tabelas era a vírgula, ou seja, estava de acordo com os requerimentos do MinAG.

- **Verificar a existência de valores absurdos ou faltantes**

As tabelas continham muitos campos em branco, o que não é aceito pelo sistema. Os campos em branco indicavam que aquele parâmetro não havia sido observado no momento em que outros o foram. Para resolver esse problema os campos em branco precisaram ser substituídos por outro valor. Esse valor não poderia ser o número zero porque existiam inúmeras observações em que o valor encontrado para um elemento observado foi igual a zero, o que é diferente do valor nulo por falta de observação. Foi então verificado um valor que não ocorreu em nenhuma das bases para ser adotado como padrão de substituição de dados em branco. O número escolhido foi 999. Assim, em todas as tabelas foi feita a substituição dos campos em branco pelo número 999.

No programa de conversão das tabelas foi incorporado um controle para que campos com conteúdo 999 não fossem tratados como um valor referente ao parâmetro, mas sim que fosse considerado como informação nula.

- **Verificar a padronização das unidades de medida utilizadas**

As tabelas existentes na base apresentavam-se padronizadas em termos de unidades de medidas, sendo que todas as observações referentes a cada parâmetros estavam em uma mesma unidade, específica para aquele parâmetro, conforme indicado na tabela WQPARM.

- **Verificar a existência da primeira linha como sendo o cabeçalho**

Embora as tabelas originais dispusessem de cabeçalho foi necessário que o programa de conversão desenvolvido criasse uma nova linha de cabeçalho para as tabelas convertidas, de acordo com os campos estabelecidos nas tabelas resultantes. Foi necessário também eliminar os espaços contidos nos nomes das variáveis.

- **Eliminar atributos (colunas) desnecessárias**

Pelas características de geração das tabelas originais que visavam a integração entre todas as tabelas sempre que possível, existiam vários códigos de identificação que, no contexto específico desse estudo de caso, mostraram-se desnecessários. Assim, associando as tabelas de dados da água com a tabela de dados das estações de monitoramento, foi mantido apenas um atributo de identificação para individualização dos pontos de coleta dos dados das estações de monitoramento.

Outros atributos que foram eliminados das tabelas são aqueles referentes aos percentis e o desvio padrão, visto que não seria coerente buscar padrões de comportamento entre os dados tomando por base os percentis e/ou desvio padrão do mesmo, visto que o sistema MinAG não estaria fazendo uma análise puramente estatística sobre esses dados. Vale mencionar também que a preservação desses atributos nas tabelas iria interferir nas regras resultantes da mineração, as quais certamente se mostrariam confusas e com pouca significância.

- **Consolidar tabelas**

Para cada tabela original foram criadas três novas versões em formato adequado para ser minerado. As novas tabelas contiveram a mesma quantidade de observações, porém cada uma tinha um conjunto diferente de parâmetros da água para serem minerados. Isso foi feito porque alguns parâmetros apresentavam um número de observações muito reduzido em relação aos demais. Assim em uma versão da tabela os parâmetros com baixa ocorrência eram considerados e em outras não. Considerando que o número de observações (linhas) presentes nas tabelas finais variou de 4863 a 7761, foram considerados como parâmetros possíveis de mineração aqueles que apresentaram um mínimo de 1000 observações, ou seja, parâmetros que estavam presentes em mais de 20% das observações realizadas, tomando por base a menor tabela.

Dependendo dos objetivos da mineração a consolidação das tabelas se faria de diferentes formas. Para fins desse estudo de caso optou-se por manter nas tabelas resultantes dados referentes ao mesmo período das tabelas originais. Assim as tabelas

wq90_94C, wq90_94D e wq90_94E contém todas o mesmo número de observações dentro do mesmo período, sendo que todos os dados foram extraídos da tabela wq90_94. As tabelas com final A e B (wq90_94A e wq90_94B, por exemplo) embora tenham sido criadas não foram consideradas para a mineração porque consistiram em tabelas intermediárias no processo de conversão não sendo destinadas a mineração.

As tabelas convertidas foram compostas pelo código de observação da estação (um valor único para cada estação) e pelo atributo referente ao valor médio de cada parâmetro observado que constava na tabela original.

O Quadro 9 apresenta parte da tabela WQ90_94 resultante e pronta para ser submetida ao processo de mineração. Comparando-se esse Quadro com aquela fração da tabela original apresentada no Quadro 6, pode-se notar que enquanto no Quadro 6 cada parâmetro era apresentado em uma linha diferente, na tabela resultante (Quadro 9) cada parâmetro está disposto agora em colunas em uma mesma linha para cada observação realizada.

Pode-se notar ainda no Quadro 9 que o parâmetro P630_MEAN, que corresponde ao parâmetro de código 630 da água (nitrito+nitrato) tem apenas valor 999, o que significa que ele não foi coletado naquela estação naquele determinado período.

Quadro 9. Parte da tabela WQ90_94 resultante para ser minerada.

ESTACAO	P10_MEAN	P400_MEAN	P610_MEAN	P625_MEAN	P630_MEAN	P665_MEAN
898	21,36	6,295	0,07499	0,32	999	0,035
899	25,8633	7,81	0,09499	1,095	999	0,045
900	20,755	6,855	0,055	0,15	999	0,055
901	22,71	7,145	0,07499	0,47	999	0,05
902	25,15	7,5	0,07499	0,475	999	0,04
903	25,5933	8,03666	0,105	0,98	999	0,06999
904	24,14	7,07666	0,145	0,275	999	0,04
905	23,6233	8,16	0,09499	0,51	999	0,1
906	24,95	6,88666	0,11	0,44	999	0,03
907	24,76	7,10999	0,06	0,53	999	0,035
908	25,425	7,505	0,16	1,465	999	0,1
909	24,73	6,885	0,055	0,5	999	0,035
910	24,04	6,91	0,06999	0,425	999	0,045
911	25,105	6,795	0,01	1,71	999	0,055
912	24,39	7,4725	0,17	1,0175	999	0,9025
913	24,19	7,425	0,43	1,09	999	0,2375
914	24,7375	7,10749	0,355	1,0675	999	0,2975
915	25,1575	7,21	0,175	1,01	999	0,19
916	23,9525	7,37999	0,23	0,97666	999	0,15
917	24,77	7,6075	0,16333	0,83333	999	0,14333
918	26,03	7,35399	0,45	0,87799	999	0,07199
919	24,8381	7,13818	0,21909	1,09182	999	0,10545
920	25,0975	7,3675	0,415	1,225	999	0,57
921	24,645	7,3925	0,3	0,615	999	0,9925
922	25,0675	7,475	0,265	1,065	999	0,3625
923	25,15	7,34	0,10666	1,17333	999	0,35
924	24,5525	7,3825	0,29	1,6625	999	0,225
925	25,4633	7,28	0,15666	0,89	999	0,02
926	26,8841	7,07666	0,16083	1,15917	999	0,3075
927	25,2375	7,555	0,2925	0,75	999	0,07249
928	24,9725	7,14249	0,2725	1,2375	999	0,1725
929	25,0725	7,425	0,4975	1,08	999	0,205

- **Gerar tabela em formato CSV**

Uma vez organizadas as tabelas o próximo passo foi gerar uma versão de cada tabela no formato CSV (Campos Separados por Vírgula) conforme é requerido pelo MinAG, o que foi feito utilizando-se o aplicativo Excel.

- **Gerar os arquivos de parâmetros para execução do sistema em *Grid* (em paralelo)**

Tendo-se as tabelas prontas para serem mineradas o próximo passo consistiu em gerar os arquivos de parâmetros para execução do sistema MinAG em paralelo.

Vale ressaltar que esse passo é necessário somente para o uso do paralelismo e que caso se optasse por executar a mineração em apenas um computador não seria necessário gerar arquivos de parâmetros visto que tais informações seriam prestadas ao sistema a medida em que ele fosse executado.

Foram criados então um total de 18 arquivos de dados e conseqüentemente 18 arquivos de parâmetros. Cada arquivo tinha uma quantidade de combinações variando de 160 a 620, o que gerou um total de 6960 combinações definidas para essa base de dados. Isso significa que o sistema MinAG foi executado 6960 vezes para realizar todas as minerações com as diferentes combinações estabelecidas.

5.5.4 Execução da mineração de dados

As 6960 execuções do MinAG foram realizadas na Universidade da Flórida. O sistema foi executado em paralelo em 400 computadores do tipo PC (Personal Computer). A configuração desses equipamentos é basicamente Pentim III ou IV, com 256 a 512 Mb de memória RAM e processador de 1.2 a 1.8 GHz. Os arquivos de parâmetros que exigiam mais tempo de processamento e mais memória de máquina em função do número de iterações a serem feitas pelo sistema foram alocados nas máquinas que apresentavam melhor performance (processador de 1.8 GHz e 512 Mb de memória RAM).

5.5.5 Regras Geradas

Ao final da mineração dos dados foram geradas inúmeras regras correlacionando e identificando os padrões entre os dados. As regras foram geradas em função dos parâmetros definidos em cada execução. Os resultados foram divididos em três

grupos: um em que os resultados apresentaram regras com *fitness* igual ou maior a 0,6, outro em que foram geradas regras com *fitness* inferior a 0,6 e o outro em que os resultados foram ruins, não tendo gerado regras.

Para a interpretação das regras considere-se como exemplo a regra[7]:

Regra[7]: **SE 22242,2511<P95_MEAN<51500 E P612_MEAN>859,7314 ENTÃO 0<= P10_MEAN<=32,2**
{Média do Atributo Meta (Treino): 23,2927} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 23,4361} {*Fitness* (Teste): 1}

A interpretação da regra[7] pode ser feita da seguinte maneira: se o valor do parâmetro 95 estiver entre 22242,2522 e 51500 e o parâmetro 612 for maior que 859,7314 então o parâmetro 10, que corresponde ao atributo meta, está entre 0 e 32,2. Durante o treinamento o valor médio encontrado para o atributo meta nos registros que satisfazem essa regra foi 23,2927, o que indica que, embora o valor do atributo meta esteja dentro da classe desejada (entre 0 e 32,2), em média seu valor foi 23,2927. O valor de *fitness* durante o treinamento foi 1, ou seja, excelente. Durante os testes, a regra se confirmou obtendo um *fitness* de teste igual a 1 também. O valor médio do atributo meta durante a fase de teste foi 23,4361, próximo do valor médio obtido no treinamento.

O Quadro 10 apresenta o conteúdo de três arquivos de parâmetros utilizados no processamento, os quais geraram regras com *fitness* maior ou igual a 0,6.

Quadro 10. Alguns arquivos de parâmetros utilizados para a mineração dos dados.

nome do arquivo aabrir	WQ95_97E.CSV	WQ90_94C.CSV	WQ90_94C.CSV
existe a coluna ID?	sim	sim	sim
número da coluna id	0	0	0
número da coluna do atributo meta	6	1	1
nome do arquivo temporario a gerar	NOME.CSV	NOME.CSV	NOME.CSV
semente para o gerador	0,3333	0,333	0,333
dividir o arquivo em treinamento e teste?	sim	sim	sim
proporção para a base de treinamento	50	50	50
nome do arquivo de treinamento	TREIN.CSV	TREIN.CSV	TREIN.CSV
proporção para a base de teste	50	50	50
nome do arquivo de teste	TEST.CSV	TEST.CSV	TEST.CSV
x (inicio intervalo meta)	0	0	0
y (fim intervalo meta)	1	32,2	32,2
número de indivíduos na população	38	38	38
número de gerações	50	100	100
valor de fitness	0,7	0,65	0,65
probabilidade de cruzamento	0,95	0,7	0,7
probabilidade de mutação	0,9	0,5	0,5
tamanho do torneio para seleção	6	3	3
porcentagem de genes mutados	0,8	0,8	0,3
deseja mutar o peso?	sim	sim	sim
limite de 0 a 1 do peso	0,95	0,95	0,95
probabilidade do peso	0,57	0,57	0,57
deseja mutar o operador?	sim	sim	sim
probabilidade de mutar o operador	1	1	1
probabilidade de <> no operador	0	0	0
probabilidade de = no operador	0	0	0
probabilidade de sair intervalos nos >= e <	0,66	0,66	0,95
deseja mutar o valor?)	sim	sim	sim
probabilidade de mutar o valor	0,3	0,5	0,5
número de subclasses para a roleta	2	2	2
y1;prop1;x2;y2;prop2;x3;prop3	0,5;50;0,5;50	1;50;1;50	1;50;1;50
nome do arquivo mga	R_WQ9597E419.mga	R_WQ9094C20.mga	R_WQ9094C21.mga
testar?	sim	sim	sim
nome do arquivo de teste	TESTE.csv	TESTE.csv	TESTE.csv
gerar todas as regras de produção	sim	sim	sim
gravar todas as regras de produção geradas	sim	sim	sim
gerar módulo de avaliação	sim	sim	sim
nome do arquivo aabrir	WQ95_97E.CSV	WQ90_94C.CSV	WQ90_94C.CSV

A seguir é apresentada uma fração do conjunto de regras geradas nas execuções em paralelo dos arquivos de parâmetros contidos no Quadro 6.

R_WQ9597E419.MGA

Regra[4]: **SE P10_MEAN=29,88 ENTÃO 0<= P665_MEAN<=1**
{Média do Atributo Meta (Treino): 0,11} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 0,06} {*Fitness* (Teste): 1}

Regra[13]: **SE P10_MEAN=30,37 ENTÃO 0<= P665_MEAN<=1**
{Média do Atributo Meta (Treino): 0,11} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 0,11} {*Fitness* (Teste): 1}

Regra[29]: **SE P10_MEAN=24,0833 ENTÃO 0<= P665_MEAN<=1**
{Média do Atributo Meta (Treino): 0,0317} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 0,33} {*Fitness* (Teste): 0,6667}

R_WQ 9094C20.MGA
Regra[0]: **SE 22242,2511<P95_MEAN<51500 ENTÃO 0<= P10_MEAN<=32,2**
{Média do Atributo Meta (Treino): 24,7176} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 24,6034} {*Fitness* (Teste): 1}

Regra[1]: **SE P95_MEAN>22242,2511 ENTÃO 0<= P10_MEAN<=32,2**
{Média do Atributo Meta (Treino): 23,7869} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 23,6431} {*Fitness* (Teste): 1}

Regra[5]: **SE P400_MEAN=5,6 ENTÃO 0<= P10_MEAN<=32,2**
{Média do Atributo Meta (Treino): 26,25} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 25,8} {*Fitness* (Teste): 1}

Regra[7]: **SE 22242,2511<P95_MEAN<51500 E P612_MEAN>859,7314 ENTÃO 0<= P10_MEAN<=32,2**
{Média do Atributo Meta (Treino): 23,2927} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 23,4361} {*Fitness* (Teste): 1}

Regra[10]: **SE P310_MEAN<=7181,2769 E P400_MEAN=5,6 E P665_MEAN<=214,4262 ENTÃO 0<= P10_MEAN<=32,2**
{Média do Atributo Meta (Treino): 27,5} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 25,8} {*Fitness* (Teste): 1}

R_WQ9094C21.MGA
Regra[1]: **SE 92,6884<P625_MEAN<713,1748 E 667,0692>=P665_MEAN>=0,4848 ENTÃO 0<= P10_MEAN<=32,2**
{Média do Atributo Meta (Treino): 23,65} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 17,625} {*Fitness* (Teste): 1}

Regra[8]: **SE 5077,2275>=P310_MEAN>=1,4349 E 270,6399<P625_MEAN<840,7924 E P665_MEAN<=988,0422 E ENTÃO 0<= P10_MEAN<=32,2**
{Média do Atributo Meta (Treino): 27,8} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 19,25} {*Fitness* (Teste): 1}

Regra[20]: **SE 92,6884<P625_MEAN<713,1748 E P665_MEAN<=988,0422 ENTÃO 0<= P10_MEAN<=32,2**
{Média do Atributo Meta (Treino): 23,65} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 17,625} {*Fitness* (Teste): 1}

Regra[34]: **SE 1,4349<P310_MEAN<8888 E 270,6399<P625_MEAN<840,7924 E P665_MEAN<=988,0422 E 980>=P32211_MEAN>=744,9656 ENTÃO 0<= P10_MEAN<=32,2**
{Média do Atributo Meta (Treino): 27,8} {*Fitness* (Treino): 1} {Média do Atributo Meta (Teste): 19,25} {*Fitness* (Teste): 1}

5.5.6 Discussão dos resultados

Da mesma forma que no primeiro estudo de caso cada regra resultante da mineração é apresentada acompanhada do valor médio do atributo meta, bem como dos valores de *fitness* de treino e de teste para auxiliar na análise das regras.

Um ponto importante a ser discutido aqui trata do nome dos atributos. Cada parâmetro físico-químico da água possui um código específico nas tabelas. Por exemplo, a temperatura média é referenciada pelo código P10_MEAN.

Na fase de pré-processamento foi mantido na tabela de mineração o código original sem substituir pelo nome do parâmetro. Como conseqüência pode-se observar aqui que as regras tornaram-se mais difíceis de serem compreendidas do que no primeiro estudo de caso. Isso se deve ao fato de que é necessário primeiramente identificar a qual parâmetro cada código corresponde para poder-se melhor entender a regra, a menos que o usuário esteja tão familiarizado com os códigos, que lhe seja natural a leitura da regra utilizando os mesmos.

Essa dificuldade seria facilmente sanada substituindo-se na tabela de dados para mineração o código pelo nome do parâmetro durante a fase de pré-processamento.

Por exemplo, a regra 20 resultaria na regra 20a (mais simples de se compreender), a seguir:

Regra[20]: **SE 92,6884 < P625_MEAN < 713,1748 E P665_MEAN <= 988,0422 ENTÃO 0 <= P10_MEAN <= 32,2**

Regra[20a]: **SE 92,6884 < Nitrogênio < 713,1748 E Fósforo <= 988,0422 ENTÃO 0 <= Temperatura <=32,2**

Fica evidente aqui a importância do pré-processamento e da atenção que deve ser dada aos detalhes nessa fase para que os resultados sejam melhores e mais compreensíveis.

6 CONSIDERAÇÕES

Considerando que a tarefa de mineração de dados é um processo de aprendizado de máquina em que o número de iterações e as diferentes combinações dos parâmetros levam a diferentes resultados, a capacidade de computação em *grid* incrementada ao sistema proporcionou vantagens no seu uso, uma vez que o tempo de processamento dos dados foi otimizado permitindo gerar mais e melhores resultados.

O grande conjunto de regras resultantes do processamento em *grid* realizado nos estudos de caso contribuiu em muito para a definição de valores adequados para os parâmetros do MinAG, os quais foram adotados como *default* no sistema. Isso é muito importante visto que o usuário do sistema, na maioria das vezes, não tem o conhecimento teórico da técnica de Algoritmos Genéticos e conseqüentemente seria difícil ao mesmo a escolha de parâmetros adequados. Dessa forma o sistema pode ser utilizado mantendo-se os parâmetros *default* para os tipos de dados abordados nessa tese.

Com base nos resultados obtidos e no comportamento atual do sistema, os seguintes trabalhos futuros foram definidos: incorporação de um módulo de análise do grau de interesse de regras, o qual auxiliará o usuário na identificação das regras mais interessantes; implementação de um módulo de mapeamento de

regras, o qual permitirá ao usuário identificar os pontos geográficos em que determinada regra foi observada, desde que a base de dados seja georreferenciada; adoção de módulo de seleção de atributos meta e de predição, o qual facilitará o processo de executar a mineração em *grid*, visto que no próprio sistema poderá se definir a cada execução os atributos da tabela a serem considerados como preditores bem como o atributo meta. Atualmente, todos os atributos presentes na tabela são considerados no processo de mineração, levando a necessidade de se construir diferentes tabelas a partir de uma mesma original para poder-se minerar os dados considerando diferentes atributos a cada execução.

O MinAG não deve ser entendido como um substituto de métodos de análise tradicionais como a estatística. Sua função é sim servir como uma ferramenta adicional na geração de informações para auxílio à compreensão da variabilidade existente nos dados.

7 CONCLUSÕES

Por meio dos testes e estudos de caso realizados pode-se concluir que esse trabalho atingiu os objetivos propostos e que o sistema desenvolvido apresenta-se como uma nova maneira de analisar a correlação entre os elementos físico-químicos do solo e da água.

8 REFERÊNCIAS BIBLIOGRÁFICAS

ABRAHAM, A. i-Miner: A Web Usage Mining Framework Using Hierarchical Intelligent Systems, In: The IEEE International Conference on Fuzzy Systems, FUZZ-IEEE'03. **Anais…**, 2003. p. 1129-1134.

ABRAHAM, A.; RAMOS, V. Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming. In: CEC´03 - Congress on Evolutionary Computation, IEEE Press, Canberra, Australia. **Anais...**, 8-12 Dec. 2003. p.1384-1391.

ADRIAANS, P.; ZANTIGUE, D. **Data Mining.** New York: Addison-Wesley, 1996.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining associations rules. In: **20th INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES. Anais…** San Francisco: Morgan Kaufmann, 1994. p. 487-499.

ANDREATTO, R. Construindo um data warehouse e analisando suas informações com Data Mining e OLAP. **DW Brasil Online**. Disponível em <http://www.dwbrasil.com.br/html/dmining.html>. Acesso em: 6 ago. 2002.

AUERNHAMMER, H. Special Issue: Global Positioning Systems in Agriculture**. Comp. & Elec. in Ag.** , v. 11, n. 1, 1994.

BÄCK, T.; FOGUEL, D. B.; MICHALEWICZ, T. **Evolutionary Computation 1 – Basic Algorithms ans Operators.** Philadelphia, USA : Institute of Physics Publishing , 2000.

BÄCK, T.; HAMMEL, U.; SCHWEFEL, H. Evolutionary Computation – Comments on the History and Current State. **IEEE Transactions on Evolutionary Computation,** v. 1, n. 1, abr. 1997.

BALASTREIRE, L.A.; ELIAS, A.I.; AMARAL, J.R. Agricultura de Precisão: mapeamento da produtividade da cultura de milho. **Revista de Engenharia Rural**, v. 8, n. 1, p. 97-111, 1998.

BITTNER, K. **Use case modeling.** Boston, MA: Addison Wesley, 2003.

BOYD, C. E. **Water quality: an introduction.** Boston, MA: Kluwer Academic Publishers, 2000.

BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDEMIR, T. B. **Redes Neurais Artificiais: Teoria e aplicações**. Rio de Janeiro: Livros Técnicos e Científicos (LTC), 2000.

BRAUN, H.; RAGG, T.; **ENZO – User Manual and Implementation Guide**, Version 1.0, University of Karlsruhe; 1995.

BUOL, S.W. et al. **Soil Genesis and classification**. Iowa: Iowa State University, 1997.

CANUTO, A. M. P.; GOTTGTROY, M. P. B. Data Mining: Geração de dados com qualidade para sistemas agropecuários. **Agrosoft. 1997**. Disponível em: <http://www.agrosoft.com/eventos/agrosoft97/trabalhos.html>. Acesso em: 1 mar. 2001.

CARVALHO, L. A. **Data Mining: a mineração de dados no marketing, medicina, economia, engenharia e administração.** São Paulo: Érica, 2001.

CHAPMAN, D. **Water quality assessments. A guide to use of biota, sediments and water in environmental monitoring.** New York: E & FN Spon, 1996.

CIANCHI, P. et al. INTEGRATED RIVER QUALITY MANAGEMENT USING INTERNET TECHNOLOGIES**.** In: WATERMATEX 2000, Gent (B). **Anais…**, 2000, p. 18 - 20.

CLAY, D. E.et al. Systemic evaluation of precision farming soil sampling requirements. In: ROBERT, P. C.; RUST, R. H.; LARSON, W. E. (Eds.). **Precision Agriculture.** Minneapolis, MN, USA, 1998, p. 253-265.

DARWIN, C. **On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life microform.** London: J. Murray, 1859.

DORIGO, M.; DI CARO, G. The ant colony optimization meta-heuristic**.** In: CORNE, D.; DORIGO, M.; GLOVER, F. (Eds.). **New Ideas in Optimization**, London, UK: McGraw Hill, 1999, p. 11-32.

DRUMMOND, S. T.; SUDDUTH, K. A..; JOSHI, A.; BIRREL, S. J.; KITCHEN, N. R. Statistical and Neural Methods for site-specific yield prediction. **Transactions of the ASAE**. v. 46, n. 1, p. 5-14, 2003.

DRUMMOND, S.T.; SUDDUTH, K.A .; BIRREL, S.J. Analysis and correlation methods from spatial data, **ASAE Paper 95-1335**, ASAE, St. Joseph, MI, USA, 1995.

DZEROSKI, S.; DEMSAR, D.; GRBOVIC, J. Predicting Chemical Parameters of RiverWater Quality from Bioindicator Data. **Applied Intelligence,** Kluwer Academic Publishers, Manufactured in The Netherlands, v. 13, p. 7–17, 2000.

EPA. **US Environmental Protection Agency.** Disponível em: <http://www.epa.gov>. Acesso em: 13 jul 2005.

EPA. US Environmental Protection Agency. **Water Quality Standards.** Washington: Office of Science and Technology. 1994.

FAYYAD, U. M. Data mining and knowledge discovery: making sense out of data. **IEEE Expert**, 1996.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: an overview. In: FAYYAD, U. M. et al. (Eds.). **Advances in Knowledge Discovery & Data Mining.** Cambridge, MA: AAAI/MIT, 1996, p. 1-34.

FRAISSE, C.W. **Agricultura de Precisão – A tecnologia de GIS/GPS chega às fazendas.** 1998. Disponível em: <http://www.fatorgis.com.br>. Acesso em: 14 nov. 2001.

FREITAS, A. Undestanding the crucial differences between classification and discovery of association rules. A postion paper. **ACM SIGKDD Explorations**, v. 2, n. 1, p. 65-69, 2000.

FREITAS, A. **Data mining and knowledge discovery with evolutionary algorithms.** Berlin; New York: Springer, 2002.

FREITAS, A .; LAVINGTON, S. H. **Mining very large databases with parallel processing.** Boston: Kluwer Academic, 1998.

GANE, C. **Desenvolvimento rápido de sistemas.** Rio de Janeiro: Livros Técnicos e Científicos (LTC), 2003.

GOLDBERG, D. E. **Genetic algorithms in search, optimization, and machine learning.** New York: Addison-Wesley, 1989.

GUIMARÃES, A. M. **Agente-M: Um Matriculador Inteligente.** Curitiba, PR: Universidade Federal do Paraná, 2000.

GUIMARÃES, A. et al. O impacto da Computação Evolucionária na tarefa de classificação de dados agronômicos. In: IV CONGRESSO BRASILEIRO DA SBIAGRO, 2003, Porto Seguro - BA. **Anais...** v. 2, 2003, p. 436-438.

GUIMARÃES, A. M. Inteligência Computacional Aplicada à Data Mining. In: UNICENTRO, Sociedade Brasileira de Computação E. (Org.). **XII Escola Regional de Informática.** Guarapuava, v. 1, 2004, cap. 3, p. 90-132.

GUPTA, R. K. et al. Spatial variability and sampling strategies for NO3-N, P and K determinations for site specific farming. **Transactions of the ASAE**, v. 40, n. 2, p. 337-342, 1997.

GUSTAFSON, D. A. **Teoria e problemas de engenharia de software.** Porto Alegre: Bookman, 2003.

HEBB D. **The organization of behavior – a neurophysiological theory.** New York: Wiley, 1949.

HERRERA, F.; LOZANO, M.; VERDEGAY, J. L. Tackling real-coded genetic algorithms: operators and tools for behavioral analysis. **Artificial Intelligence Review**, v. 12, p. 265-319, 1998.

HOLLAND, J. H. **Adaptation in Natural and Arti_cial Systems**. Ann Arbor: The University of Michigan Press, 1975.

HOLMES, G. et al. **A** logistic boosting approach to inducing multiclass alternating decision trees. **Working Paper 1/02**, Department of Computer Science, The University of Waikato, Hamilton. 2002.

HOPFIELD, J. *J* Neural Networks and Physical Systems with Emergent Computational Abilities. In: **Proceedings of the National Academy of Sciences**, Washington, USA, v.79, April. 1982 p. 2554-2558.

JOHNSON, G. et al. Spatial and temporal analysis of weed populations using geoestatistics. **Weed Science**, Champaign, v.44, n.3, p.704-710, 1996.

JORGE, L. A .C. Agricultura de Precisão. **Workshop Agrosoft 2002.** Disponível em: <http://www.agrosoft.com.br/ag2002/workshop>. Acesso em: 1 jun. 2002.

KANTARDZIC, M. **Data Mining. Concepts, Models, Methods, and Algorithms.** New Jersey: IEEE Computer Society. Wiley-Interscience. 2003

LANGLEY, P. **Elements of machine learning.** San Francisco, CA: Morgan Kaufmann, 1996.

LU, H.; SETIONO, H.; NeuroRule: a connectionist approach to data mining. In: **Proceedings of the 21st Conf. onVery Large Databases.** Zurich, 1995.

MCCULLOCH, W.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, v. 7, p. 115-133, 1943.

MATHIAS, I. M. **SISMAT - Sistema de Matrícula Inteligente.** Curitiba, PR: Universidade Federal do Paraná, 2000.

McBRATNEY, A.; WHELAN, B.; ANCEV, T. Future Directions of Precision Agriculture. **Precision Agriculture**, v. 6, p. 7-23, 2005.

MITCHELL, M. **An introduction to genetic algorithms**. Cambridge: Mit Press, 1997.

MOLIN, J.P. Agicultura de precisão, parte I: O que é e estado da arte em sensoriamento. **Eng. Agrícola,** Jaboticabal, v. 17, n. 2, p. 97-107, dez. 1997.

MOLIN, J.P et al. Regression and correlation analysis of *grid* soil data versus cell spatial data. In: THIRD EUROPEAN CONFERENCE ON PRECISION AGRICULTURE**,** Montpellier. **Proceedings…**Agro Montpellier, 2001. p. 449-453.

MONMARCHÉ, N. On data clustering with artificial ants. In: FREITAS, A. A. (Ed.). Data Mining with Evolutionary Algorithms, Research Directions –AAAI Workshop, Menlo Park, CA: AAAI Press, 1999. p. 23-26.

NAVEGA, S. Princípios Essenciais do Data Mining. In: Infoimagem 2002, Cenadem, **Anais...**, Nov. 2002.

PALAZZO, L. Algoritmos para Computação Evolutiva. **Universidade Católica de Pelotas - Escola de Informática.** Grupo de Pesquisa em Inteligência Artificial. Disponível em: <www.ucpel.tche.br> Internet. Acesso em: 2 set. 2003.

PALMER, M.D. **Water quality modeling: a guide to effective practice**. Washington, D.C.: World Bank, 2001.

PARPINELLI, R. S.; Lopes, H. S; FREITAS, A. A. Data Mining with an Ant Colony Optimization Algorithm. **IEEE Trans on Evolutionary Computation, special issue on Ant Colony Algorithms**, v. 6, n. 4, August 2002.

PARPINELLI, R.S.; LOPES, HS; FREITAS, AA. **An Ant Colony Algorithm for Classification Rule Discovery**. In: ABBASS, H.; SARKER, R.; NEWTON, C. (Eds.). **Data Mining: a Heuristic Approach**,. London: Idea Group Publishing, 2002. p. 191-208.

PIERCE, F.J.; SADLER, E.J. **The State of Site-Specific Management for Agriculture.** Madison, WI: ASA-CSSA-SSSA, 1997.

POOLE, D.; MACKWORTH, A.; GOEBEL, R. **Computational Intelligence. A Logical Aprroach.** New York: Oxford University Press, 1998.

QUEVAUVILLER, P. **Quality assurance for water analysis.** New York: John Wiley & Sons, 2002.

QUINLAN, J.R. **C4.5: Programs for Machine Learning**, San Francisco, CA: Morgan Kaufmann, 1993.

RISLEY, J. C.; ROEHL JR., E. A.; CONRADS, P. A.Estimating Water Temperatures in Small Streams in Western Oregon Using Neural Network Models. U.S. Geological Survey. **Water-Resources Investigations Report 02-4218**. Portland, Oregon. 2003.

ROSENBLAT, F. The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain. **Psychological Review**, v 65. 1958{Reprinted in (Anderson & Rosenfeld 1988)}.

SANTOS, R. T.; NIEVOLA, J. C.; FREITAS, A. **Extracting Comprehensible Rules from Neural Networks via Genetic Algorithms**. CEFET-PR/CPGEI, PUC-PR/PPGIA, PUC-PR/PPGIA, 2002.

SCHUELLER, J.K. A review and integrating analysis of spatially-variable control of crop production**. Fertilizer Research.** v. 31, p. 1-34, 1992.

SCHUELLER, J.K. Thechnology for precision agriculture. European Conference on Precision Agriculture. **Anais…,** v. 1, Set., 1997. p. 33-44.

SOUZA, L.S. et al. Variabilidade de propriedades físicas e químicas do solo em um pomar cítrico. **Revista Brasileira de Ciência do Solo**, Viçosa, v.21, n.3, p. 367-372, 1997.

TCHOBANOGLUS, G.; SCHROEDER, E. D. **Water Quality: Characteristics, Modeling, Modification**. Massachusetts, USA: Adison-Wesley Publishing Company, 1985.

TITTERINGTON, D. M.; SMITH, A. F. M.; MAKOV, U. E. **Statistical Analysis of Finite Mixture Distributions.** Chichester, U.K.: John Wiley and Sons. 1985.

VIGIL, K. M. **Clean water: an introduction to water quality and water pollution control.** Corvallis: Oregon State University Press, 2003.

WHITNEY, J. D. et al. PRECISION FARMING APPLICATIONS IN FLORIDA CITRUS. **Applied Engineering in Agriculture**. v. 15, n. 5, p. 399-403. 1999

YE, N. **The handbook of data mining.** New Jersey: Lawrence Erlbaum Associates (LEA), 2003.

ZELL, A. et al., **SNNS – Stuttgart Neural Network Simulator – User Manual.** Version 4.1, University of Stuttgart; 1995.

GLOSSÁRIO

Agricultura de Precisão	Conjunto de técnicas, equipamentos e procedimentos para realizar aplicação localizada de insumos no campo
Algoritmo Genético	Algoritmo de aprendizado baseado na teoria da evolução das espécies
Aprendizado de Máquina	Técnica implementada em algoritmos para que o mesmo apresente melhores resultados por meio de auto-ajuste (aprendizado)
Atributo de predição	Atributo contido em uma tabela utilizado para predizer a condição para que um outro atributo esteja dentro de uma faixa esperada de valores
Atributo meta	Atributo que se deseja classificar. Na mineração de dados define-se a classe de interesse do atributo meta
Base de teste	Arquivo que contém o conjunto de dados que serão utilizados na mineração de dados durante a fase de teste
Base de treinamento	Arquivo que contém o conjunto de dados que serão utilizados na mineração de dados durante a fase de treinamento
Computação em *grid*	Tipo de processamento que permite distribuir em vários computadores o processamento das execuções da mineração de dados, para que as mesmas sejam executadas em paralelo
Computação Evolucionária	Área de pesquisa que desenvolve técnicas baseadas no conceito da evolução das espécies
Critério de água	Conjunto de critérios que determinam a finalidade de uso da água
Cromossomo	Forma de representação do conjunto de atributos de um registro a ser manipulado pelo Algoritmo Genético

Fitness	Função que indica a qualidade de uma regra
Gene	Representação de cada atributo dentro do cromossomo
Grau de interesse	Função que indica o quanto uma regra é interessante
Inteligência Computacional	Área da computação que ocupa-se em desenvolver aplicações que usem técnicas inteligentes, representando o comportamento humano na resolução de problemas
Intervalo de Gerações	Número de vezes que o Algoritmo Genético será executado durante o processo de mineração de dados. A cada execução existe a possibilidade de um elemento (cromossomo) da população ser modificado (evoluir)
KDD	Metodologia que ocupa-se do desenvolvimento de processos de descoberta de conhecimento em bases de dados
Mineração de Dados	Conjunto de algoritmos que identificam padrões de comportamento em bases de dados
Operadores Genéticos	Conjunto de procedimentos que produzem mudança em dados durante a execução do Algoritmo Genético
Padrão	Fato freqüente em uma base de dados
Pré-processamento	Conjunto de procedimentos que antecedem a mineração de dados
Roleta	Método usado para selecionar registros de uma tabela (base de dados) para serem utilizados como elemento da população do Algoritmo Genético