Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka

Marcelino Pereira dos Santos Silva[1,2]

[1]Universidade do Estado do Rio Grande do Norte (UERN)
BR 110, Km 48, 59610-090, Mossoró, RN, Brasil

[2]Instituto Nacional de Pesquisas Espaciais (INPE)
C. Postal 515, 12201-097, São José dos Campos, SP, Brasil

mpss@dpi.inpe.br

Abstract. *Tools and techniques employed for automatic and smart analysis of huge data repositories of industries, governments, corporations and scientific institutes are the subjects dealt by the emerging field of Knowledge Discovery in Databases (KDD). Data mining is the KDD step where it's performed the method selection to search patterns in data, followed by the search for interesting patterns in a particular representation and the best parameter tuning of the chosen algorithms. This course will present the fundamentals of data mining, as well some research and application areas of this technology. In order to reach a practical and applied approach, data mining tasks will be performed using Weka, a collection of machine learning algorithms for real data mining tasks. The activities will help to fix concepts shown, allowing the perception of potentialities of this recent and challenging research area.*

Resumo. *As ferramentas e técnicas empregadas para análise automática e inteligente dos imensos repositórios de dados de indústrias, governos, corporações e institutos científicos são os objetos tratados pelo campo emergente da Descoberta de Conhecimento em Bancos de Dados (Knowledge Discovery in Databases - KDD). Mineração de dados é a etapa em KDD responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão. Neste minicurso, os fundamentos de mineração de dados serão apresentados, bem como diferentes áreas de pesquisa e aplicação desta tecnologia. Visando um enfoque prático e aplicado, atividades de mineração serão realizadas com o Weka, um pacote de algoritmos de aprendizagem de máquina para resolver problemas reais de mineração de dados. Estas atividades auxiliarão na fixação dos conceitos apresentados, bem como numa melhor percepção do potencial desta recente e desafiadora área de pesquisa.*

1. Introdução

As áreas governamentais, corporativas e científicas têm promovido um crescimento explosivo em seus bancos de dados, superando em muito a usual capacidade de

interpretar e examinar estes dados, gerando a necessidade de novas ferramentas e técnicas para análise automática e inteligente de bancos de dados [Fayyad et al. 1996].

Nos diferentes segmentos da sociedade, as instituições têm buscado na tecnologia recursos que agreguem valor aos seus negócios, seja agilizando operações, suportando ambientes ou viabilizando inovações. Diariamente, pessoas e instituições disponibilizam dados oriundos de tarefas cotidianas a estas plataformas tecnológicas através de simples atividades como compras no supermercado do bairro ou operações bancárias. Os sistemas de computação participam da vida das pessoas de forma cada vez mais próxima e constante. Não obstante, institutos científicos, indústrias, corporações e governos acumulam volumes gigantescos de dados, impulsionados também pela versatilidade e alcance proporcionados pela Internet.

Esta ampla disponibilidade de imensas bases de dados, aliada à eminente necessidade de transformar tais dados em informação e conhecimento úteis para o suporte à decisão, têm demandado investimentos consideráveis da comunidade científica e da indústria de software. A informação e o conhecimento obtidos podem ser utilizados para diversas aplicações, que vão do gerenciamento de negócios, controle de produção e análise de mercado ao projeto de engenharia e exploração científica [Han & Kamber 2001].

As ferramentas e técnicas empregadas para análise automática e inteligente destes imensos repositórios são os objetos tratados pelo campo emergente da descoberta de conhecimento em bancos de dados (DCBD), da expressão em inglês Knowledge Discovery in Databases (KDD). Mineração de dados é a etapa em KDD responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão.

1.1. O Processo de Descoberta de Conhecimento em Bancos de Dados (KDD)

Descoberta de conhecimento em bancos de dados, é o processo não trivial de identificar em dados padrões que sejam válidos, novos (previamente desconhecidos), potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão [Fayyad et al. 1996]. Examinando estes termos individualmente:

- Dados: conjunto de fatos F, como instâncias de um banco de dados. Por exemplo, uma coleção de n cadastros de pessoas físicas contendo idade, profissão, renda etc.
- Padrão: expressão E em uma linguagem L descrevendo fatos em um subconjunto F_E de F. E é dito um padrão se é mais simples do que a enumeração de todos os fatos em F_E. Por exemplo, o padrão: "Se renda < \$$r$ então a pessoa não recebe financiamento" seria aplicável para uma escolha apropriada de r.
- Processo: geralmente em KDD, processo é uma seqüência de vários passos que envolve preparação de dados, pesquisa de padrões, avaliação de conhecimento, refinação envolvendo iteração e modificação.
- Validade: os padrões descobertos devem ser válidos em novos dados com algum grau de certeza. Uma medida de certeza é uma função C mapeando expressões

em L para um espaço de medidas M_C . Por exemplo, se um limite de padrão de crédito é ampliado, então a medida de certeza diminuiria, uma vez que mais financiamentos seriam concedidos a um grupo até então restrito a esta operação.

- Novo: em geral, assume-se que "novidade" pode ser medida por uma função N(E,F), que pode ser uma função booleana ou uma medida que expresse grau de "novidade" ou "surpresa". Exemplo de um fato que não é novidade: sejam E = "usa tênis" e F = "alunos de colégio" então N(E,F) = 0 ou N(E,F) = *false*. Por outro lado: sejam E = "bom pagador" e F = "trabalhador da construção civil" então N(E,F) = 0,85 ou N(E,F) = *true*.
- Potencialmente útil: padrões devem potencialmente levar a alguma atitude prática, conforme medido por alguma função de utilidade. Por exemplo, regras obtidas no processo podem ser aplicadas para aumentar o retorno financeiro de uma instituição.
- Compreensível: um dos objetivos de KDD é tornar padrões compreensíveis para humanos, visando promover uma melhor compreensão dos próprios dados. Embora seja um tanto subjetivo medir compreensibilidade, um dos fatores freqüentes é a medida de simplicidade. O fator de compreensão dos dados está relacionado à intuitividade da representação destes, bem como da granularidade alta o suficiente para que estes sejam compreendidos. Por exemplo: o log de um servidor Web não é uma representação compreensível; já fatos estatísticos extraídos deste log, tais como totais de acesso ou classificação dos acessos realizados, fornecem informação num formato mais intuitivo e de granularidade humanamente compreensível.

1.2. Etapas do Processo de Descoberta de Conhecimento em Bancos de Dados

O processo de KDD é interativo, iterativo, cognitivo e exploratório, envolvendo vários passos (Figura 1) com muitas decisões sendo feitas pelo analista (que é um especialista do domínio dos dados, ou um especialista de análise dos dados), conforme descrito:

1. Definição do tipo de conhecimento a descobrir, o que pressupõe uma compreensão do domínio da aplicação bem como do tipo de decisão que tal conhecimento pode contribuir para melhorar.
2. Criação de um conjunto de dados alvo (Selection): selecionar um conjunto de dados, ou focar num subconjunto, onde a descoberta deve ser realizada.
3. Limpeza de dados e pré-processamento (Preprocessing): operações básicas tais como remoção de ruídos quando necessário, coleta da informação necessária para modelar ou estimar ruído, escolha de estratégias para manipular campos de dados ausentes, formatação de dados de forma a adequá-los à ferramenta de mineração.
4. Redução de dados e projeção (Transformation): localização de características úteis para representar os dados dependendo do objetivo da tarefa, visando a redução do número de variáveis e/ou instâncias a serem consideradas para o conjunto de dados, bem como o enriquecimento semântico das informações.
5. Mineração de dados (Data Mining): selecionar os métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de

interesse numa forma particular de representação ou conjunto de representações; busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão.

6. Interpretação dos padrões minerados (Interpretation/Evaluation), com um possível retorno aos passos 1-6 para posterior iteração.
7. Implantação do conhecimento descoberto (Knowledge): incorporar este conhecimento à performance do sistema, ou documentá-lo e reportá-lo às partes interessadas.

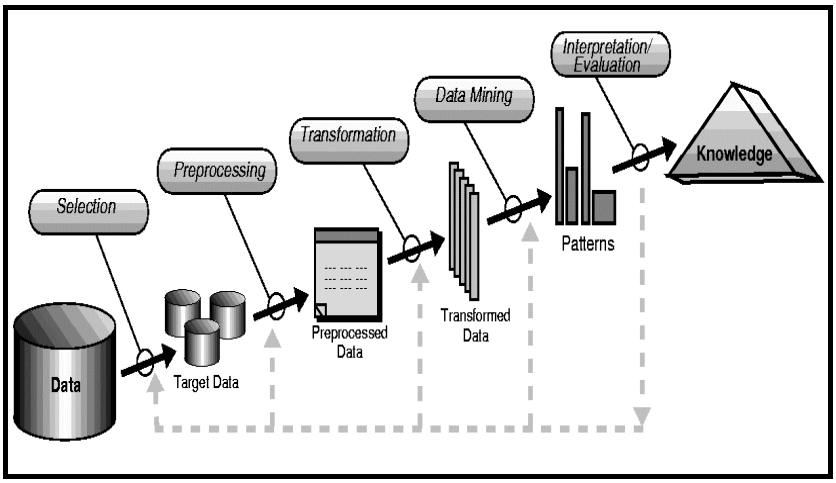

Figura 1. Etapas de KDD [Fayyad et al. 1996]

1.3. Aplicabilidade de Descoberta de Conhecimento em Bancos de Dados

Visando uma exemplificação da aplicabilidade de KDD, são apresentados a seguir casos onde a descoberta de conhecimento em bancos de dados pode desempenhar tarefas relevantes [Witten & Frank 2000]:

- Submissões a empréstimos demandam do proponente o fornecimento de dados pessoais e financeiros relevantes. Estas informações são utilizadas pelas instituições financeiras como base para a decisão de efetuar ou não o empréstimo. Tal decisão é comumente tomada em dois estágios. Primeiro, métodos estatísticos são utilizados para determinar situações bem definidas em relação à aceitação ou rejeição do pedido. Os casos remanescentes, ou seja, aqueles que estão no limite necessitam de análise humana. KDD pode ser aplicado neste problema da seguinte forma: suponha-se a disponibilidade de um banco de dados histórico sobre clientes da instituição, com aproximadamente 5000 cadastros contendo 20 diferentes atributos, tais como idade, tempo de serviço, vencimentos, bens, status atual de crédito etc. O tratamento dessas informações por métodos de KDD geraria automaticamente regras objetivas e claras sobre características fundamentais a bons e maus clientes, podendo estas regras serem aplicadas para aumentar a taxa de sucesso das operações de empréstimo.
- Diagnóstico é uma das principais aplicações de sistemas especialistas. A manutenção preventiva de dispositivos eletromecânicos pode evitar falhas que interrompam processos industriais. Técnicos regularmente inspecionam cada

dispositivo, medindo vibrações e outros fenômenos que indicam necessidade de manutenção. Instalações de indústrias químicas chegam a utilizar mais de mil diferentes dispositivos, que vão de pequenas bombas a grandes turbo-alternadores. A medição de vibrações e demais fenômenos é muito ruidosa, devido às limitações dos procedimentos de medição e registro. Estes dados, uma vez estudados por um especialista, conduzem a um diagnóstico. As limitações dos procedimentos técnicos, aliadas à subjetividade humana, oferecem uma margem de erro preocupante. Por outro lado, um universo de 600 falhas, cada uma devidamente registrada com seus conjuntos de medições representando 20 anos de experiência, pode ser utilizado para determinar o tipo de falha através de procedimentos de KDD, aperfeiçoando assim o processo de busca e correção de problemas eletromecânicos.

- Desde o princípio da tecnologia de satélites, cientistas ambientais tentam detectar manchas de óleo a partir de imagens de satélite, com o intuito de alertar e tomar providências rapidamente contra desastres ambientais. Estas imagens fornecem uma oportunidade para monitorar águas litorâneas dia e noite, independentemente de condições atmosféricas. Manchas de óleo aparecem como regiões escuras na imagem cujo tamanho e forma modifica-se dependendo do clima e condições marítimas. Entretanto, outras regiões negras semelhantes podem ser causadas por fatores climáticos, tais como ventos altos. Detecção de manchas de óleo é um processo manual de alto custo, que requer pessoal altamente treinado para avaliar cada região na imagem. Sistemas de detecção têm sido desenvolvidos para selecionar imagens para subseqüente processamento manual. Entretanto, é necessário ajustá-los detalhadamente para circunstâncias distintas. KDD permite que estes sistemas sejam treinados para fornecer padrões de manchas e de ausência delas, permitindo ainda ao usuário controlar compromissos entre manchas não detectadas e falsos alarmes.

2. Técnicas e Algoritmos

Bases de dados são altamente suscetíveis a dados ruidosos (erros e valores estranhos), incompletos (valores de atributos ausentes) e inconsistentes (discrepâncias semânticas) devido a seus típicos volumes. Técnicas de pré-processamento e transformação de dados são aplicadas para aumentar a qualidade e o poder de expressão dos dados a serem minerados. Estas fases tendem a consumir a maior parte do tempo dedicado ao processo de KDD (aproximadamente 70%). A etapa de mineração de dados é responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados, com efetiva busca por padrões de interesse numa forma particular de representação, além da busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão. Esta etapa pressupõe que os dados tenham uma boa qualidade (limpos, sem inconsistências, campos nulos etc.), além de uma boa representação e relevância semântica (dados devidamente tratados, transformados e enriquecidos).

2.1. Pré-processamento de Dados

Rotinas de limpeza de dados tentam suprir valores ausentes, reduzir discrepâncias de valores ruidosos e corrigir inconsistências. Para valores ausentes, algumas técnicas aplicáveis são [Han & Kamber 2001]:

1 - Ignorar a tupla

2 - Suprir valores ausentes

a) manualmente;

b) através de uma constante global;

c) utilizando a média do atributo;

d) utilizando a média do atributo para todas as instâncias da mesma classe;

e) com o valor mais provável (regressão, inferência etc.).

As técnicas 2b, 2c, 2d e 2e podem "viciar" os dados. A técnica 2e é uma estratégia interessante, pois em comparação com outros métodos utiliza um maior número de informações dos dados disponíveis.

Ruídos nos dados são erros aleatórios ou variâncias numa variável mensurada. A eliminação de ruídos pode ser realizada através de:

1 - Interpolação;

2 - Agrupamento;

3 - Inspeção humana e computacional combinadas;

4 – Regressão.

Alguns tipos de inconsistências podem ser corrigidos manualmente através de referências externas. Rotinas de consistência evitam a inserção de dados incorretos através da interface do banco de dados (infelizmente, a maioria dos softwares não são projetados e desenvolvidos levando em conta KDD). Ferramentas de engenharia do conhecimento podem detectar a violação de restrições de dados. Tanto redundâncias como discrepâncias podem ser combatidas através de dependências funcionais.

2.2. Transformação de Dados

O processo de mineração geralmente demanda a integração de dados (combinação de diferentes bases de dados) e a transformação destes (modificações de formato e enriquecimento semântico).

No caso da integração de dados, várias fontes podem ser utilizadas (diferentes bancos de dados, cubos de dados, flat files, arquivos XML etc.). Alguns tópicos relevantes neste processo são [Han & Kamber 2001]:

1 - Integração de esquemas - casamento de entidades relevantes do mundo real (utilização dos metadados);

2 - Redundância de atributos (análise de correlação - medida de quanto um atributo implica em outro);

3 - Identificação e resolução de valores de dados conflitantes (especialmente devido a diferenças na representação, escala ou codificação);

Uma integração de dados criteriosa pode reduzir e evitar redundâncias e inconsistências no conjunto de dados resultante, aumentando a precisão e velocidade do processo de mineração de dados.

No caso da transformação de dados, estes são modificados ou transformados em formatos apropriados à mineração:

1 - Agregação: geração de totalizadores levando em conta determinados atributos ou instâncias. Por exemplo, vendas mensais, anuais, sazonais etc.;

2 - Generalização: substituição de dados por conceitos de mais alto nível. Idades, por exemplo, podem ser representadas por faixas etárias, localidades por regiões etc.;

3 - Normalização: atributos são escalonados para uma faixa específica como -1.0 a 1.0, ou 0.0 a 1.0;

4 - Construção de atributos: novos atributos são construídos a partir de informações pré-existentes (ex.: classificação de crédito a partir de renda e histórico).

5 - Redução de dados

a) Agregações;

b) Redução dimensional: detecção e remoção de atributos irrelevantes;

c) Compressão de dados: utilização de mecanismos de codificação para reduzir o tamanho do conjunto de dados;

d) Redução numérica (instâncias): amostragem, por exemplo.

2.3. Mineração de Dados

Etapas de mineração de dados utilizam técnicas e algoritmos de diferentes áreas do conhecimento, principalmente inteligência artificial (especialmente aprendizagem de máquina), banco de dados (recursos para manipular grandes bases de dados) e estatística (comumente na avaliação e validação de resultados), conforme a Figura 2. Uma questão que freqüentemente surge é a seguinte: porque não utilizar tão somente os conhecidos procedimentos estatísticos para obter informações relevantes a partir de um conjunto de dados?

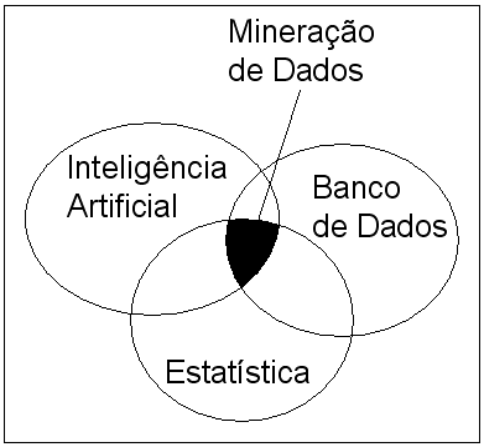

Figura 2. Mineração de dados utiliza recursos de diferentes áreas

Conforme mencionado, procedimentos estatísticos são utilizados nas etapas de KDD e mais especificamente na mineração de dados. Entretanto, o volume, complexidade e peculiaridades dos eventos e dos dados por eles originados impõem severas limitações a metodologias puramente estatísticas, dentre elas:

- Dados nem sempre possuem independência estatística entre eles, ou seja, muitos domínios possuem inter-relação entre seus objetos e respectivos atributos, comprometendo a aplicação de métodos estatísticos;
- A análise estatística demanda um grau de conhecimento e domínio desta área que apenas estatísticos e profissionais de áreas correlatas possuem, restringindo assim a atuação da grande maioria dos potenciais usuários de procedimentos analíticos;
- Métodos estatísticos manipulam muito bem dados numéricos, mas não manipulam bem valores simbólicos, incompletos ou inconclusivos;
- Estes métodos são computacionalmente caros quando se trata de grandes bases de dados.

Desta forma, percebe-se claramente que a mineração de dados possui grande relevância, contribuição e abrangência no que diz respeito a aplicações. Visando uma melhor compreensão das tarefas, será apresentado a seguir uma breve descrição dos principais métodos de mineração de dados utilizando aprendizagem de máquina.

A exemplificação de cada tópico toma por base recursos do Weka, uma ferramenta de KDD que contempla uma série de algoritmos de preparação de dados, de aprendizagem de máquina (mineração) e de validação de resultados. Na seção 4 outros pontos do software serão abordados.

2.3.1. Aprendizagem Supervisionada

Esta categoria de algoritmos possui esta denominação porque a aprendizagem do modelo é supervisionada, ou seja, é fornecida uma classe à qual cada amostra no treinamento pertence. Estes algoritmos são preditivos, pois suas tarefas de mineração desempenham inferências nos dados com o intuito de fornecer previsões ou tendências, obtendo informações não disponíveis a partir dos dados disponíveis:

- Classificação: através destes algoritmos supervisionados (com ênfase na precisão da regra) é possível determinar o valor de um atributo através dos valores de um subconjunto dos demais atributos da base de dados. Por exemplo, num conjunto de dados comerciais deseja-se descobrir qual o perfil dos clientes que consomem cosméticos importados. Com classificadores pode-se inferir (prever) que “clientes do sexo feminino, com renda superior a R$ 1.500,00 e com idade acima de 30 anos compram cosméticos importados. Neste caso, o atributo “compra cosmético importado” é denominado classe, pois é o atributo alvo da classificação (cujos possíveis valores, neste caso, são "sim" ou "não"). As formas mais comuns de representação de conhecimento dos algoritmos de classificação são regras e árvores. Os algoritmos Id3, C45, J48, ADTree, UserClassifier, PredictionNode, Splitter, ClassifierTree, M5Prime, por exemplo, geram como resultado árvores de classificação (Figura 3), enquanto que outros como Prism, Part, OneR geram regras de classificação. Outra opção seria a representação através de tabela de decisão implementada, por exemplo, pelo algoritmo DecisionTable. Modelos matemáticos, de regressão e redes neurais também representam resultados de algoritmos como SMO, LinearRegression, Neural, dentre outros.

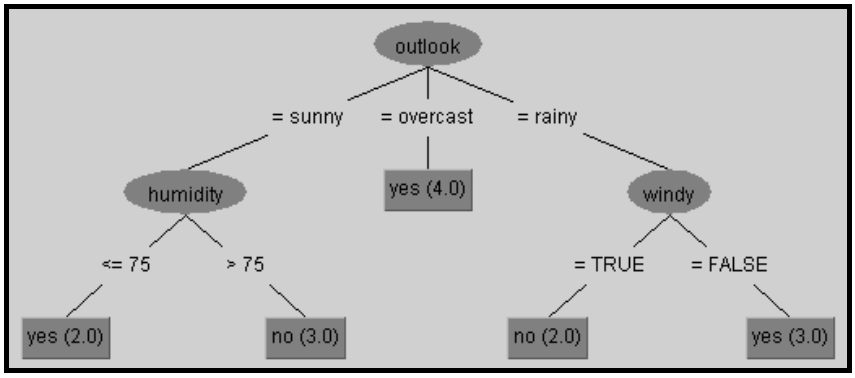

Figura 3. Exemplo de árvore de classificação no Weka [Waikato 2004]

- Seleção de atributos: em bases de dados encontram-se atributos que têm um peso maior ou até determinante nas tarefas de mineração de dados. Por exemplo, no caso do cliente, a sua renda com certeza é um atributo determinante nos seus hábitos de consumo. Com algoritmos de seleção de atributos é possível determinar os atributos de fato relevantes para a mineração dos dados, separando-os dos atributos irrelevantes, como por exemplo nome do cliente (que neste caso não influencia seus hábitos de consumo). O Weka disponibiliza vários algoritmos para esta categoria de mineração, dentre eles InformationGain, PrincipalComponents e ConsistencyEval.

2.3.2. Aprendizagem Não-Supervisionada

Nestes algoritmos o rótulo da classe de cada amostra do treinamento não é conhecida, e o número ou conjunto de classes a ser treinado pode não ser conhecido a priori, daí o fato de ser uma aprendizagem não-supervisionada. Além disso são também descritivos, pois descrevem de forma concisa os dados disponíveis, fornecendo características das propriedades gerais dos dados minerados:

- Associação: quando a classe de uma tarefa de mineração não é determinada como no caso da classificação, uma boa opção é o algoritmo de associação Apriori do Weka. Ele é capaz de gerar regras do tipo: clientes do sexo masculino, casados, com renda superior a R$ 1.800,00 têm o seguinte hábito de consumo: roupas de grife, perfumes nacionais, relógios importados. Esta regra teria a seguinte representação: sexo(X,[masc]) ^ est_civil(X,[casado]) ^ renda(X,[1800,∞]) ⇒ consome(X,[roupa_grife, perfume_nacional, relógio_importado]). Neste caso, o próprio algoritmo elege os atributos determinantes (lado esquerdo da regra) e os atributos resultantes (lado direito) na tarefa revelando associações entre valores dos atributos, tendo o algoritmo sua ênfase no compromisso entre precisão e cobertura (Figura 4).
- Clustering: em algumas situações, torna-se necessário verificar como as instâncias de uma determinada base de dados se agrupam devido a características intrínsecas de seus atributos, sem que seja definida uma classe para a tarefa. A partir da definição de uma métrica de similaridade para cada atributo e uma função de combinação destas métricas em uma métrica global, os objetos são agrupados com base no princípio da maximização da similaridade intraclasse e da minimização da similaridade interclasse. Weka possui os

algoritmos Cobweb, Simple Kmeans e Em para tarefas que demandam a descoberta de padrões de agrupamento nos dados. Como exemplo podemos utilizar algoritmos de clustering para identificar subgrupos homogêneos de clientes de uma determinada loja.

```
Best rules found:

 1. outlook=overcast 4 ==> play=yes 4    conf:(1)
 2. temperature=cool 4 ==> humidity=normal 4    conf:(1)
 3. humidity=normal windy=FALSE 4 ==> play=yes 4    conf:(1)
 4. outlook=sunny play=no 3 ==> humidity=high 3    conf:(1)
 5. outlook=sunny humidity=high 3 ==> play=no 3    conf:(1)
 6. outlook=rainy play=yes 3 ==> windy=FALSE 3    conf:(1)
 7. outlook=rainy windy=FALSE 3 ==> play=yes 3    conf:(1)
 8. temperature=cool play=yes 3 ==> humidity=normal 3    conf:(1)
 9. outlook=sunny temperature=hot 2 ==> humidity=high 2    conf:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2    conf:(1)
```

Figura 4. Exemplos de regras de associação no Weka [Waikato 2004])

2.3.3. Validação de Resultados

É muito importante que os resultados e modelos possam ser avaliados e comparados. Alguns elementos relevantes neste domínio: teste e validação, que fornecem parâmetros de validade e confiabilidade nos modelos gerados (cross validation, suplied test set, use training set, percentage split); indicadores estatísticos para auxiliar a análise dos resultados (matriz de confusão, índice de correção e incorreção de instâncias mineradas, estatística kappa, erro médio absoluto, erro relativo médio, precisão, F-measure, dentre outros).

2.3.4. Medidas de Interesse

Diferentes fatores retratam a qualidade dos resultados obtidos a partir de tarefas de mineração de dados. Neste ponto, abordaremos duas medidas de interesse muito relevantes na avaliação de regras:

Confidência:

Seja uma regra "A => B".

A confidência da regra é dada por:

$$\text{Confidência } (A \Rightarrow B) = \frac{\#\_\text{Tuplas\_Contendo\_Tanto\_A\_Como\_B}}{\#\_\text{Tuplas\_Contendo\_A}}$$

Exemplo: Uma confidência de 85% (0,85) da regra *compra (mulher, computadores) => compra (mulher, software)* significa que 85% das mulheres que compram computadores também compram software.

Suporte:

O suporte da regra é dado por:

$$\text{Suporte } (A \Rightarrow B) = \frac{\#\_\text{Tuplas\_Contendo\_Tanto\_A\_Como\_B}}{\#\_\text{Total\_De\_Tuplas}}$$

Ou seja, um suporte de 5% significa que de todas as transações comerciais realizadas, 5% são efetuadas por *mulheres que comprando computador também compram softwares*.

2.3.5. Critérios de comparação

Critérios para comparar métodos e resultados de mineração de dados permitem avaliar e optar pelo melhor custo/benefício a ser adotado para a tarefa em questão. Alguns critérios relevantes neste contexto são:

- Precisão avaliativa ou preditiva: habilidade do modelo para avaliar ou prever corretamente classes, agrupamentos, regras;
- Velocidade: refere-se ao custo computacional da geração e utilização do modelo;
- Robustez: habilidade do modelo para avaliar ou prever corretamente utilizando dados ruidosos ou com valores ausentes;
- Escalabilidade: capacidade de construir modelos eficientemente a partir de grandes volumes de dados;
- Interpretabilidade: nível de compreensão fornecido pelo modelo.

3. Aplicações

O número de pesquisadores e profissionais que utilizam técnicas de mineração de dados ainda é muito pequeno no Brasil, haja vista o potencial e demanda desta tecnologia. Tanto no campo acadêmico como no corporativo, os bancos de dados abarrotados de informações são geralmente utilizados para consultas triviais, e muitos dados preciosos fadados ao backup. O grande potencial do conhecimento intrínseco nestas montanhas de dados continua ignorado ou inacessível por muitas instituições. Entretanto, diferentes aplicações têm atestado a relevância e poder desta tecnologia.

3.1. Aplicações Acadêmicas

Na área acadêmica, a demanda por novas e poderosas abordagens de mineração de dados está presente em muitos segmentos de pesquisa, dentre eles:

- Mineração em datawarehouses: repositórios com dados de boa qualidade, integrados, estratégicos, históricos, disponibilidade de metadados, infraestrutura de processamento (inclusive ferramentas avançadas);
- Mineração em bancos de dados espaciais: aplicável sobre elementos geográficos, imagens de sensoriamento remoto, imagens médicas, layout de chips VLSI etc. No caso de dados geográficos (Figura 5), aplicações relevantes contemplam estudos ambientais, vigilância territorial, detecção de desmatamentos, planejamento urbano etc.;
- Mineração de dados multimídia: extração de padrões relevantes a partir de animações, áudio, vídeo, imagens e textos (busca por similaridades, análise multidimensional, classificatória, preditiva, dentre outros);

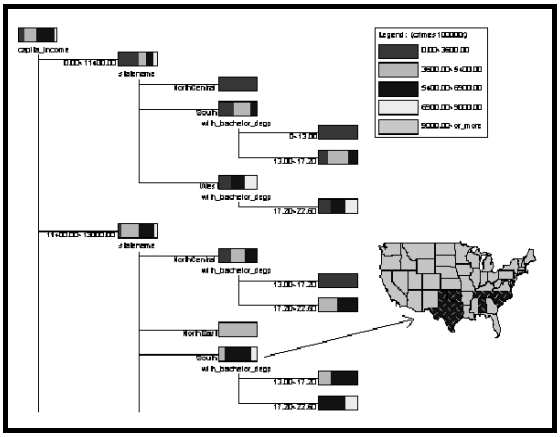

Figura 5. Mineração de dados geográficos (GeoMiner [Han et al. 1997])

- Mineração de dados (séries) temporais: mercado de ações, processos de produção, experimentos científicos, tratamentos médicos, análises de tendências, de históricos e de similaridades. Envolve ainda a avaliação de eventos cíclicos, sazonais e aleatórios;
- Mineração de textos: muitas informações estão disponíveis em documentos (artigos de jornais ou científicos, livros, e-mails, páginas Web etc.). Dentre as abordagens encontramos a recuperação de informações e a aplicação de técnicas de mineração em informação semi-estruturada;
- Minerando a Web: mineração de conteúdo, uso e estrutura. Configura-se num repositório imenso, distribuído e global que contem uma ampla e rica coleção de hiperlinks. Seu tamanho, complexidade e dinamismo oferecem grandes desafios científicos. Os engenhos de busca (Google, por exemplo) subsidiam abordagens interessantes.

3.2. Aplicações Corporativas

Atentas ao poder da informação que coletam a todo instante (e em volumes imensos), as empresas investem continuamente em novas tecnologias que agreguem valor ao negócio. O efetivo emprego de técnicas de mineração de dados atinge diferentes empreendimentos:

- Minerando o DNA: atividades de mapeamento de seqüências (padrões de seqüenciamento em larga escala), busca por similaridades e comparações, e identificação de seqüências genética co-ocorrentes têm encontrado na mineração de dados poderosos recursos computacionais;
- Minerando dados médicos: hospitais, clínicas e planos de saúde, cada vez mais interessados no aperfeiçoamento de procedimentos (diagnósticos e tratamentos), buscam em KDD meios para maximizar recursos e resultados, preferencialmente com redução de custos. Exemplos: mineração de imagens tomográficas, precisão na prescrição de exames e procedimentos;
- Dados estruturais (grafos): muitas bases de dados possuem características estruturais, cujos dados são compostos por segmentos e relacionamentos entre estes. A mineração de subestruturas busca subgrafos semelhantes num conjunto

de grafos. Subestruturas descobertas são utilizadas para comprimir os dados originais, permitindo abstrair estruturas detalhadas e representar conceitos estruturais nestes dados. Assim, uma subestrutura descoberta é usada para simplificar os dados, substituindo instâncias da subestrutura por um ponteiro para esta nova subestrutura descoberta, conforme exemplo da Figura 6.

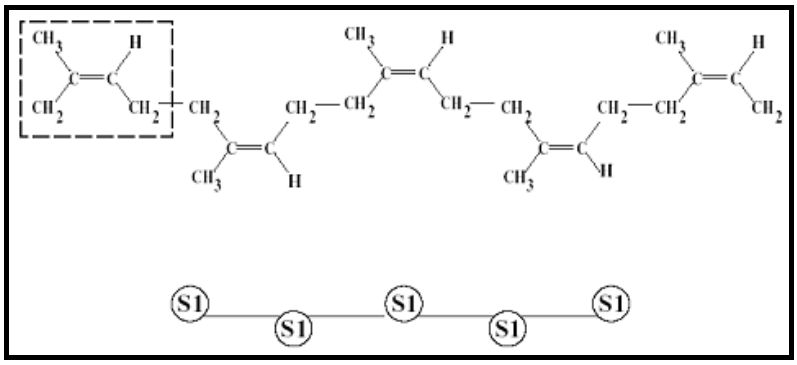

Figura 6. Mineração de grafos (estrutura atômica - borracha) [Holder et al. 2002]

- Dados financeiros: o volume de interesses e poder atrelados a ativos financeiros têm despertado a atenção de muitos para informações estratégicas deste domínio. Aplicações de minerações de dados vão da detecção de fraudes e lavagem de dinheiro à analise de mercados, tendências e fomento especulativo. Análise de crédito de consumidores e classificação de clientes para estratégias de marketing figuram dentre as aplicações mais comuns.
- Dados comerciais: empresas de varejo, especialmente as grandes redes, contemplam minas de ouro nos seus grandes bancos de dados. Análise de vendas, comportamento da clientela, giro de produtos, fenômenos sazonais e preferências regionais motivam grandes investimentos em mineração de dados. Além disso, é possível avaliar campanhas publicitárias, potencializar o comércio eletrônico, avaliar e incentivar fidelidade de clientes.
- Telecomunicações: grande demanda na detecção de invasões e comportamentos anômalos em sistemas, avaliação de uso e tráfego, análise de padrões de consumo.

3.3. Tendências, Desafios e Perspectivas

Observando as aplicações acadêmicas e corporativas acima, é possível citar algumas das fortes tendências da área, bem como os desafios correlatos e as perspectivas da mineração de dados:

- Os casos citados e outros possuem potenciais não explorados. O volume e complexidade dos dados, aliados à peculiaridade das respectivas operações, revelam em cada aplicação um grande conjunto de oportunidades para atividades de pesquisa que aperfeiçoem e inovem métodos e tecnologias [Sarawagi et al. 1998];
- Considerando o volume de dados (que em alguns casos mostra-se espantoso) e a demanda por alto desempenho, verifica-se que novos métodos escaláveis de

mineração muito contribuirão com o desenvolvimento da área [Agrawal & Srikant 1994];

- O alto nível de integração de plataformas e de bases de dados remotas demanda uma igual integração das ferramentas de mineração de dados com diferentes sistemas de bancos de dados, datawarehouses e Web [Silva & Robin 2002];
- Devido à iteratividade e interatividade das tarefas, linguagens que especifiquem consultas e processos são muito bem vindas em ambientes de KDD, uma vez que a utilização de diferentes ferramentas, o controle do fluxo do processo e o gerenciamento do conhecimento demandam esforço extra na ausência dos recursos providos por uma linguagem [Silva & Robin 2004];
- Visual data mining: concerne ao emprego de recursos de computação gráfica (CG) para evidenciar padrões em bases de dados. A evolução de ambas as áreas (KDD e CG) amplia as oportunidades de relevantes trabalhos neste domínio [DBMiner 2000] (Figura 7);

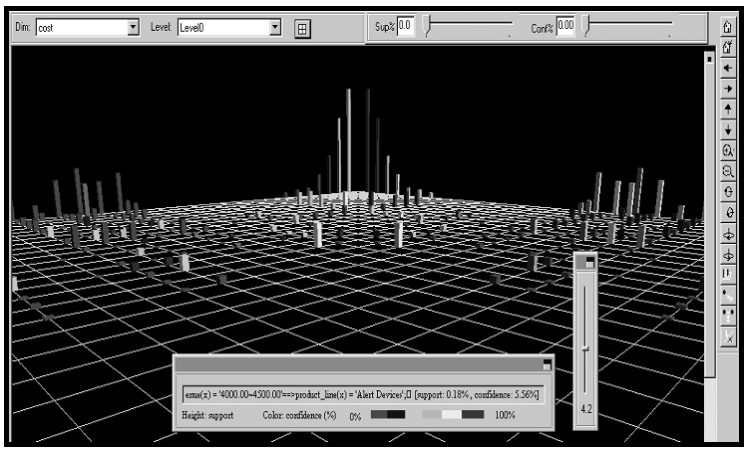

Figura 7. Exemplos de regras de associação [DBMiner 2000]

- Mineração de dados complexos e semi-estruturados: além das gigantescas bases de dados convencionais, repositórios de dados não convencionais (imagens, textos, grafos, Web, multimídia etc.) apresentam-se como grandes motivações para pesquisas e projetos inovadores [Simoff et al. 2002];
- Proteção de privacidade e segurança de dados: os freqüentes ataques a sistemas computacionais, especialmente através da Web, oferecem um excelente campo de aplicação para métodos de mineração de dados em tempo real como, por exemplo, avaliação de comportamento e padrões de uso.

4. Weka: Um Software GNU para Mineração de Dados

Waikato Environment for Knowledge Analysis – WEKA [Waikato 2004, Witten & Frank 2000] é uma ferramenta de KDD que contempla uma série de algoritmos de preparação de dados, de aprendizagem de máquina (mineração) e de validação de resultados. WEKA foi desenvolvido na Universidade de Waikato na Nova Zelândia, sendo escrito em Java e possuindo código aberto disponível na Web (a atual versão - 3.4.3 - demanda Java 1.4). A equipe de desenvolvimento tem lançado periodicamente correções e releases do software, além de manter uma lista de discussões acerca da

ferramenta. Grande parte de seus componentes de software são resultantes de teses e dissertações de grupos de pesquisa desta universidade. Inicialmente, o desenvolvimento do software visava a investigação de técnicas de aprendizagem de máquina, enquanto sua aplicação inicial foi direcionada para a agricultura, uma área chave na economia da Nova Zelândia.

O sistema possui uma interface gráfica amigável e seus algoritmos fornecem relatórios com dados analíticos e estatísticos do domínio minerado. Grande parte de seus recursos é acessível via sua GUI, sendo que os demais podem ser utilizados programaticamente através de API´s. Foi disponibilizada também uma abrangente documentação online do código fonte. Por ser escrito em Java, o código pode ser rodado em diferentes plataformas, conferindo uma boa portabilidade ao software. Uma limitação da ferramenta é a sua escalabilidade, uma vez que suas versões atuais limitam o volume de dados a ser manipulado à dimensão de memória principal. Mesmo assim, é possível minerar bases de dados relevantes, tornando o pacote atrativo para um grande número de aplicações (componentes do Weka têm sido utilizados em um considerável número de projetos). Algumas funcionalidades do software são introduzidas na seção 2.3.

4.1. Interface e Funcionalidades

A interface gráfica do Weka disponibiliza grande parte de suas funcionalidades. Embora seja intuitiva, para uma abordagem inicial faz-se necessário reconhecer alguns elementos estratégicos da GUI Explorer, cujo guia do usuário encontra-se em http://aleron.dl.sourceforge.net/sourceforge/weka/ExplorerGuide.pdf

Esta primeira tela (Figura 8) apresenta elementos de pré-processamento:

Figura 8. Interface do Weka (Preprocess)

- (A) Open File, Open URL, Open DB: através destes botões é possível selecionar, respectivamente, bases de dados a partir de flat files locais (formato .arff), bases remotas (Web), e diferentes bancos de dados (via JDBC). Para

acessar dados no MS Access, um roteiro de configuração está disponível em http://www.cs.waikato.ac.nz/~ml/weka/opening_windows_DBs.html . Uma breve descrição do formato .arff pode ser encontrada em http://www.cs.waikato.ac.nz/~ml/weka/arff.html ;

- (B) No botão filter é possível efetuar sucessivas filtragens de atributos e instâncias na base de dados previamente carregada (seleção, discretização, normalização, amostragem, dentre outros);
- (C) Navegando interativamente pelos atributos (quadro Attributes) é possível obter informações quantitativas e estatísticas sobre os mesmos (quadro Selected attribute);

Nesta interface, é possível desenvolver tarefas de classificação (Figura 9):

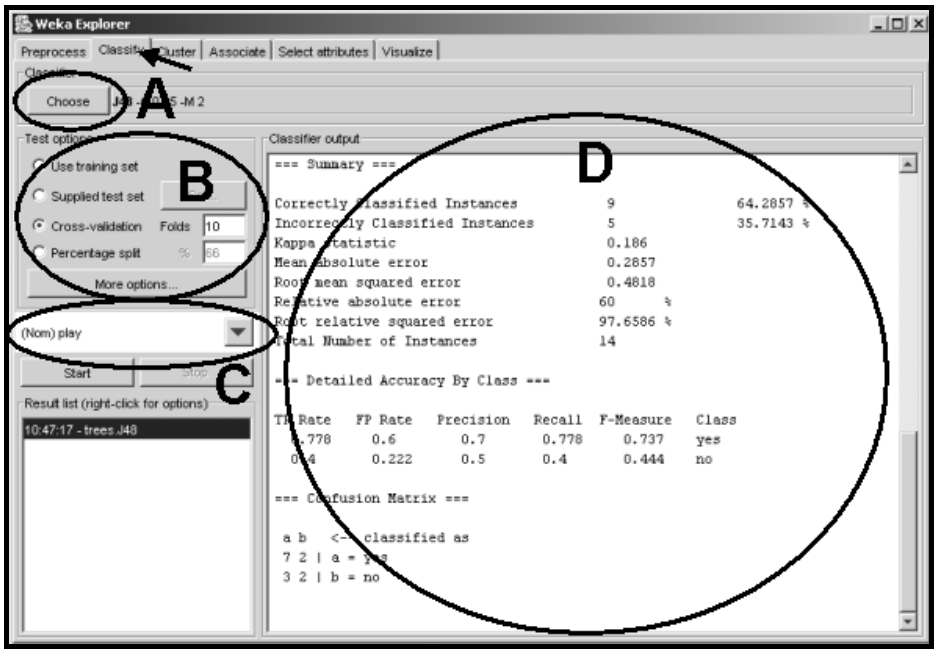

Figura 9. Classificação no Weka

- (A) Seleção e parametrização do algoritmo a ser utilizado (Id3, C45, J48, BayesNet, Prism, Part etc);
- (B) Permite selecionar a opção de teste e validação do modelo gerado (o próprio conjunto de dados do treinamento, um outro conjunto só para testes, cross-validation, separar parte do conjunto de treinamento para teste);
- (C) Seleção do atributo classe para a tarefa de classificação;
- (D) Resumo da tarefa efetuada, com dados estatísticos, modelo, matriz de confusão etc.

As opções “Cluster”, “Associate” e “Select attributes” possuem interfaces semelhantes, fornecendo algumas opções peculiares a estas tarefas. No caso de tarefas de agrupamento (“Cluster”) a interface fornece a opção de ignorar atributos, pois é muito comum que neste tipo de tarefa um ou mais atributos gerem viés ou ruídos no processo de agrupamento. Já na seleção de atributos (“Select attributes”), é possível escolher o

algoritmo avaliador de atributos e o método de busca para a tarefa (Figura 10). Faz-se necessário salientar que alguns avaliadores demandam métodos de busca específicos.

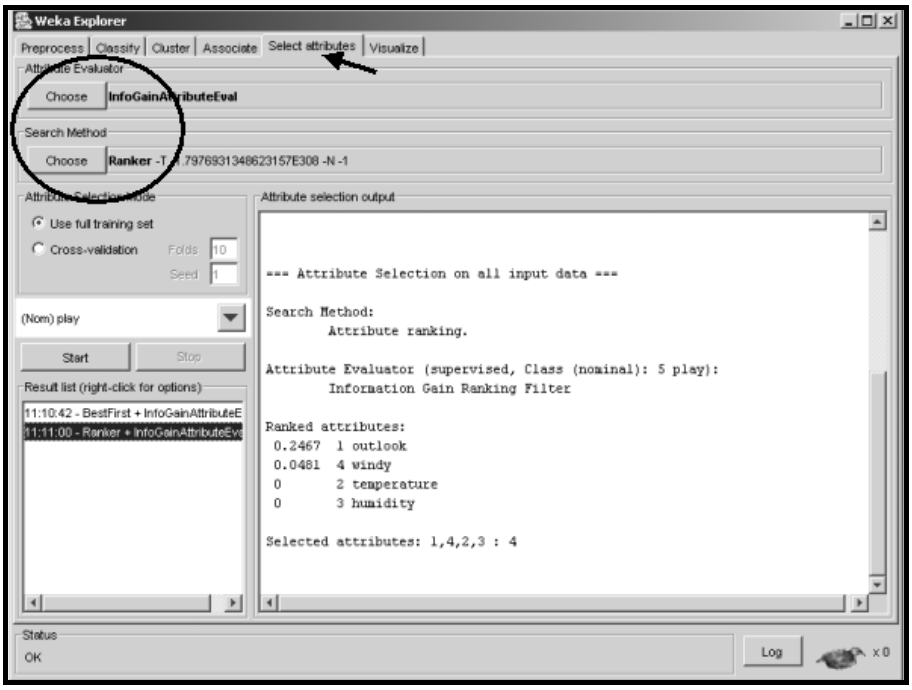

Figura 10. Seleção de atributos

4.2 Instalação, Configuração e Documentação do Weka

A instalação do software é simples. Basta baixar o pacote de [Waikato 2004] e executar o instalador. Atividades de configuração podem ser encaradas como a própria parametrização dos algoritmos utilizados. O processo de escolha de algoritmos e a respectiva parametrização destes constituem um dos desafios na mineração de dados, pois dependem muito do conhecimento de cada algoritmo, da experiência do minerador e do domínio do especialista da área minerada (dados comerciais, científicos etc.). Na documentação abordada a seguir é possível encontrar informações que muito auxiliarão nesta tarefa.

Diferentes recursos de documentação podem ser encontrados no software e no site do projeto. Na instalação do Weka um pacote de documentação é disponibilizado, o qual contem informações da API (Figura 11). No pacote ainda está incluso um tutorial, que na realidade é o oitavo capítulo do livro escrito pelos líderes do projeto [Witten & Frank 2000].

No site do Weka diferentes recursos agregam informações e ajuda ao usuário (Figura 12):

- Página de trouble-shooting;
- Fórum de discussões (com arquivo das mensagens);
- Guia explicativo do formato ARFF adotado pelo Weka (e outros softwares);
- Introdução ao uso do Weka a partir da linha de comando (chamando diretamente componentes Java);

- Guia do usuário para a interface do Explorer;
- Descrição do pacote Bayes Net;
- Tutorial do Experimenter;
- Descrição de como carregar bases de dados do MS Access para o Weka.

São disponibilizadas ainda bases de dados para testes e aprendizagem, além de uma lista de projetos relevantes relacionados ao Weka.

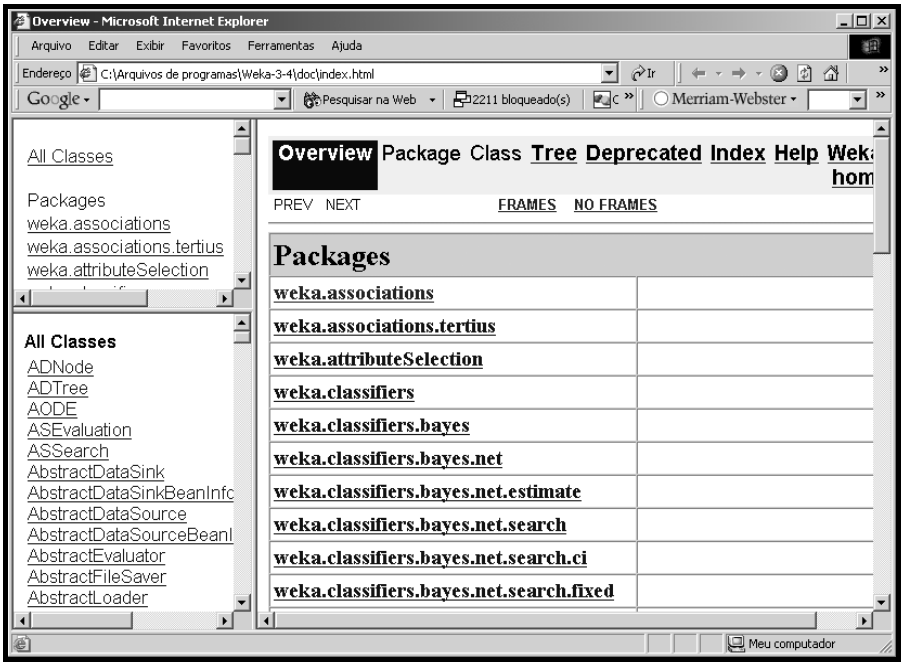

Figura 11. Documentação da API

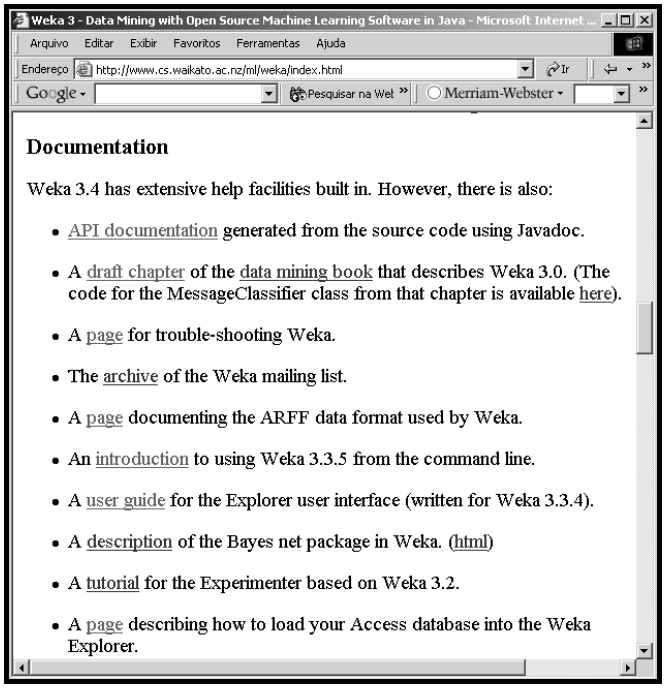

Figura 12. Recursos e ajuda no site do Weka

4.3 Explorando Potencialidades

Por mais aplicada que seja a análise documental do Weka, torna-se imprescindível a plena utilização do software. Neste minicurso, com a finalidade de fixar os conceitos apresentados e propiciar uma melhor percepção do potencial desta recente e desafiadora área de pesquisa, atividades de mineração de dados serão realizadas diretamente no Weka visando um enfoque prático e aplicado.

5. Kdnuggets: Um Portal de Recursos para KDD

O Kdnuggets [Kdnuggets 2004] é um portal com informações e recursos pra vários tópicos ligados a KDD e suporte à decisão, com conteúdos e/ou links para artigos, softwares, cursos, publicações, empresas do ramo, notícias, encontros e congressos, oportunidades de trabalho, dentre outros (Figura 13).

Figura 13. Portal Kdnuggets

6. Considerações Finais

Mineração de dados, e conseqüentemente KDD, possuem uma vasta aplicação nos mais diferentes segmentos, tanto acadêmicos como corporativos, além de uma série de desafios relevantes que podem motivar excelentes trabalhos científicos.

Este minicurso não esgota em momento algum os diferentes tópicos da mineração de dados, mas antes procura fornecer uma visão geral do assunto bem como seus fundamentos, apresentando ainda diferentes áreas de pesquisa e aplicação desta tecnologia.

O avanço tecnológico e a oferta de ferramentas não dispensam de forma alguma o especialista do domínio minerado. A experiência profissional, a convivência com os processos e a leitura dos padrões descobertos são atributos que propiciam ao(s) minerador(es) amplas chances de sucesso nos processos de KDD.

O breve contato com o software Weka permite que algumas tarefas de mineração (e KDD) sejam de fato desenvolvidas, fixando conceitos e apresentando uma ferramenta de qualidade e de código aberto, possibilitando ainda a quebra de paradigmas em relação à mineração de dados.

Aqueles que de fato se identificarem com esta área de pesquisa devem continuar a exploração do Weka e de outras ferramentas, buscando nas referências conteúdos e subsídios para ampliar o conhecimento e a visão crítica deste promissor segmento da computação. A partir deste ponto, o desenvolvimento de excelentes projetos, dissertações, teses e aplicações será uma conseqüência natural do envolvimento acadêmico e da dedicação pessoal.

Referências

Agrawal, R.; Srikant, R. "Fast algorithms for mining association rules in large databases". Proceedings of the International Conference on Very Large Databases, Santiago, Chile, 1994

Dbminer Technology Inc. DBMiner Interprise 2.0 (2000). Disponível no site da DBMiner Technology. URL: http://www.dbminer.com/

Fayyad, U. M.; Piatesky-Shapiro, G.; Smyth, P. "From Data Mining to Knowledge Discovery: An Overview". In: Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

Han, J.; Koperski, K.; Stefanovic, N. GeoMiner: "A System Prototype for Spatial Data Mining", ACM SIGMOD International Conference on Management of Data, Arizona, 1997.

Han, J.; Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.

Holder, L.; Cook, D.; Gonzalez, J.; Jonyer, I. "Structural Pattern Recognition in Graphs, in Pattern Recognition and String Matching", Kluwer Academic Publishers, 2002.

Kdnuggets. Data Mining and Knowledge Discovery. Disponível no site da Kdnuggets (2004). URL: http://www.kdnuggets.com

Sarawagi, S.; Agrawal, R.; Megiddo, N. "Discovery-Driven Exploration of OLAP Data Cubes". IBM Almaden Research Center, 1998.

Silva, M. P. S.; Robin, J. R. "SKDQL – Uma Linguagem Declarativa de Especificação de Consultas e Processos para Descoberta de Conhecimento em Bancos de Dados e sua Implementação" (2002). Dissertação de Mestrado. UFPE, 2002.

Silva, M. P. S.; Robin, J. R. SKDQL: "A Structured Language to Specify Knowledge Discovery Processes and Queries" (2004). XVII Brazilian Symposium on Artificial Intelligence - SBIA'04.

Simoff, S.; Djeraba, C.; Zaiane, O. "Multimedia Data Mining between Promisses and Problems" (2002). SIGKDD Explorations.

University of Waikato. Weka 3 – Machine Learning Software in Java. Disponível no site da University of Waikato (2004). URL: http://www.cs.waikato.ac.nz/ml/weka

Witten, I.; Frank, E. Data Mining – Practical Machine Learning Tools. Morgan Kaufmann, 2000.