Mineração de Dados – Funcionalidades, Técnicas e Abordagens

Sérgio da Costa Côrtes[1]
scortes@inf.puc-rio.br

Rosa Maria Porcaro[2]
rporcaro@ibge.gov.br

Sérgio Lifschitz[3]
sergio@inf.puc-rio.br

PUC-RioInf.MCC10/02 Maio, 2002

Abstract

The subject of the study is Data Mining, with emphasis on its functionalities (results), techniques and application strategies. We underline the importance of data mining as part of a larger research process called Knowledge Discovery in Database (KDD), for which is presented the methodology for preparation and exploration of data, interpretation of results and assimilation of mined knowledge. Data mining is presented in the context of business intelligence; its forms of presentation and the difficulties in implementation in corporations and some applications suitable for the use of data mining as well as their fields of research are discussed.

Keywords: Data Mining, business intelligence,. KDD, knowledge discovery

Resumo

Apresentamos um estudo sobre Mineração de Dados (*data mining*), destacando suas funcionalidades (resultados), técnicas e abordagens de aplicação. Destacamos mineração de dados como parte de um processo maior de pesquisa denominado Busca de Conhecimento em Banco de Dados (*Knowledge Discovery in Database - KDD*), para o qual apresentamos sua metodologia para preparação e exploração dos dados, interpretação de seus resultados e assimilação dos conhecimentos minerados. Apresentamos a mineração de dados no contexto da inteligência de negócios, como se apresenta e quais suas dificuldades de implantação nas empresas e discutimos algumas aplicações candidatas à utilização da mineração de dados e suas áreas de pesquisas.

Palavras-chave: Mineração de dados, data mining, inteligência de negócios, KDD, busca de conhecimento

---

[1] *Doutorando, parcialmente apoiado pela Fundação IBGE e pela PUC-Rio*
[2] *Doutora em Ciência da Informação, pesquisadora DPE/DEMET – Fundação IBGE*
[3] *Parcialmente apoiado por bolsa de pesquisa do CNPq 300048/94-7*

1 Introdução

Atualmente as organizações têm se mostrado extremamente eficientes em capturar, organizar e armazenar grandes quantidades de dados, obtidos de suas operações diárias ou pesquisas científicas, porém, a maioria ainda não usa adequadamente essa gigantesca quantidade de dados para transformá-la em conhecimentos que possam ser utilizados em suas próprias atividades, sejam elas comerciais ou científicas.

O conceito de Mineração de Dados (*Data Mining*) está se tornando cada vez mais popular como uma ferramenta de descoberta de informações, que podem revelar estruturas de conhecimento, que possam guiar decisões em condições de certeza limitada. Recentemente, tem havido um interesse crescente em desenvolver novas técnicas de análise de dados, especialmente projetadas para tratar questões relativas a mineração de dados. No entanto, a mineração de dados ainda está baseada em princípios conceituais de Análise de Dados Exploratórios (*Exploratory Data Analysis - EDA*) e de modelagem.

Diversas definições de Mineração de Dados podem ser encontradas na literatura. Entre as diversas definições destacamos as seguintes:

- Mineração de dados é a busca de informações valiosas em grandes bancos de dados. É um esforço de cooperação entre homens e computadores. Os homens projetam bancos de dados, descrevem problemas e definem seus objetivos. Os computadores verificam dados e procuram padrões que casem com as metas estabelecidas pelos homens [WI99].

- Mineração de dados é a exploração e análise de dados, por meios automáticos ou semi-automáticos, em grandes quantidades de dados, com o objetivo de descobrir regras ou padrões interessantes [BL97].

- Mineração de dados, em poucas palavras, é a análise de dados indutiva [Men99].

- Mineração de dados é o processo de proposição de várias consultas e extração de informações úteis, padrões e tendências, freqüentemente desconhecidos, a partir de grande quantidade de dados armazenada em bancos de dados [BT99].

- Mineração de dados, de forma simples, é o processo de extração ou mineração de conhecimento em grandes quantidades de dados [HK01].

Um conceito muito difundido e *errado* sobre mineração de dados é o que define os sistemas de mineração de dados como sistemas que podem *automaticamente* minerar todos os conceitos valiosos que estão escondidos em um grande banco de dados sem intervenção ou direcionamento humano [HK01].

Para nós, *mineração de dados é um processo altamente cooperativo entre homens e máquinas, que visa a exploração de grandes bancos de dados, com o objetivo de extrair conhecimentos através do reconhecimento de padrões e relacionamento entre variáveis,*

conhecimentos esses que possam ser obtidos por técnicas comprovadamente confiáveis e validados pela sua expressividade estatística.

Freqüentemente, mineração de dados tem sido considerada e classificada como uma mistura de pesquisas em estatística, inteligência artificial e bancos de dados. Até recentemente, não era reconhecido como um campo de interesse para os estatísticos, sendo mesmo considerado, nesta área, como uma área de pesquisa 'pouco relevante'. Devido à sua importância prática, entretanto, o campo tem emergido como uma área de crescimento acentuado e de elevada importância, destacando-se pelo surgimento de diversos congressos científicos e produtos comerciais.

Mineração de Dados é parte de um processo maior de pesquisa denominado Busca de Conhecimento em Banco de Dados (*Knowledge Discovery in Database - KDD*), o qual possui uma metodologia própria para preparação e exploração dos dados, interpretação de seus resultados e assimilação dos conhecimentos minerados. No entanto, se tornou mais conhecida do que o próprio processo de KDD em função de ser a etapa onde são aplicadas as técnicas de busca de conhecimentos.

O objetivo deste trabalho é apresentar os principais conceitos da tecnologia de KDD, enfatizando o processo de Mineração de Dados com suas funcionalidades e principais técnicas utilizadas para obtenção de conhecimento, bem como apresentar uma metodologia para preparação dos dados para mineração, detalhando os processos de limpeza, integração, seleção e transformação de dados, etapas fundamentais para o sucesso da mineração. Além disso, inserir a Mineração de Dados no contexto da Inteligência de Negócios (*Business Intelligence – BI*) como uma ferramenta de apoio a tomada de decisão de nível mais elevado, sendo utilizada principalmente no planejamento estratégico das empresas.

Este trabalho está organizado da seguinte forma. Na seção 2, apresentam-se as funcionalidades (resultados), técnicas e abordagens da mineração de dados. A seção 3 trata da Busca de Conhecimento em Banco de Dados (*Knowledge Discovery in Database - KDD*), onde um estudo detalhado das etapas de limpeza, integração, seleção e transformação de dados é apresentado. Mineração de dados no contento da inteligência de negócios é apresentada na seção 4, enquanto que, na seção 5, são apresentadas as aplicações em potencial para mineração de dados. Finalmente, na seção 6, são apresentadas as considerações finais.

2 Funcionalidades da mineração de dados, suas técnicas e abordagens

O processo de aplicação da mineração de dados envolve vários estágios, conforme veremos neste trabalho, mas o principal estágio antes de se iniciar a busca do conhecimento oriundo dessa aplicação é definir, claramente, a que resultados deseja-se chegar. Uma vez definidos os resultados (seções 2.1, 2.2 e 2.3), é preciso definir que técnicas (seção 2.4) utilizar e como aplicar essas técnicas para obtenção dos conhecimentos desejados.

2.1 Funcionalidades em mineração de dados

Diversos tipos de armazenamentos de dados e de bancos de dados podem ser utilizados no processo de mineração. Em função do tipo de dados armazenado e disponível se pode

definir que tipo de padrões ou relacionamento queremos minerar. A funcionalidade da mineração de dados irá especificar que tipo de padrões ou relacionamentos entre os registros e suas variáveis podem ser utilizados na mineração. Essa funcionalidade é tratada, por alguns autores, como resultados (*outcomes*) ou tarefas (tasks).

A literatura, em muitos casos, não deixa claro as diferenças entre funcionalidades e técnicas. Por exemplo, uma coleção de técnicas podem ser utilizadas na *análise de cestas de produtos*, entre estas regras de associação. Essas técnicas são conhecidas como técnicas de análise de cestas de produtos, muito utilizadas em marketing. Entretanto, análise de cestas de produtos é também uma aplicação, que busca determinar que itens são vendidos juntos em supermercados. A figura 1 a seguir mostra, em camadas, as interações entre funcionalidades, técnicas e algoritmos, visando esclarecer a interatividade do objetivo da mineração de dados com as técnicas a serem empregadas.

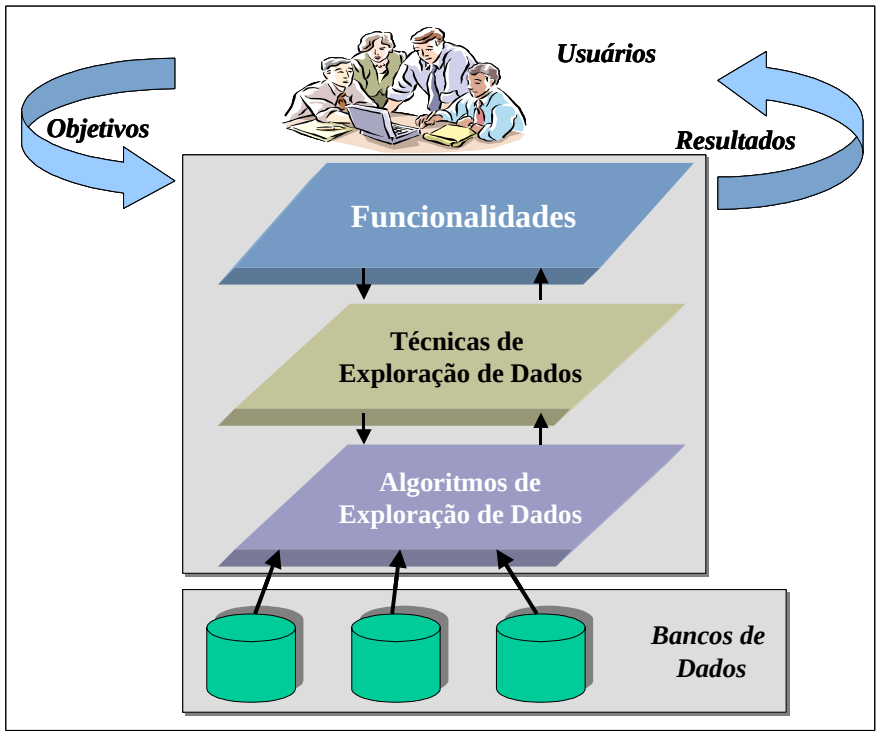

Figura 1: Interatividade entre as funcionalidades e técnicas da mineração de dados

Diversos autores tratam as funcionalidades da mineração de dados de forma diferente. A seguir algumas formas de tratamento por esses autores:

- [AZ96] - Descoberta de conhecimento e Predição
- [BL97] - Classificação, Estimação, Predição, Afinidade em grupos, Agrupamentos (*clustering*) e Descrição
- [BT99] - Classificação, Detecção de seqüência, Análise de dependência de dados e Análise de desvio
- [EN99] - Previsão, Identificação, Classificação e Otimização
- [HK01] - Descrição e Predição

- [Men99] - Predição, Classificação, Agrupamento (*clustering*), Segmentação, Associação, Visualização e Otimização

- [WB98] - Classificação, Estimação, Segmentação e Descrição

- [WI99] - Predição, Detecção de desvio, Segmentação, Agrupamento (*clustering*), Análise de ligações e Regras de associação, Sumarização e Visualização e Garimpagem em textos

Como podemos observar, a funcionalidade em mineração de dados não é um consenso e é tratada pelos autores muito mais pela sua área de atuação do que pelo formalismo necessário. No entanto, consideramos que definir bem os conceitos da funcionalidade da mineração de dados, a que resultados queremos chegar é fundamental para o processo como um todo. Uma vez bem definida, se pode melhor escolher as técnicas a serem aplicadas para se obter os resultados esperados. Assim, classificamos a funcionalidade em mineração de dados como *Análise Descritiva* e *Análise de Prognóstico*. A figura 2 ilustra essa forma de abordagem da funcionalidade na mineração de dados.

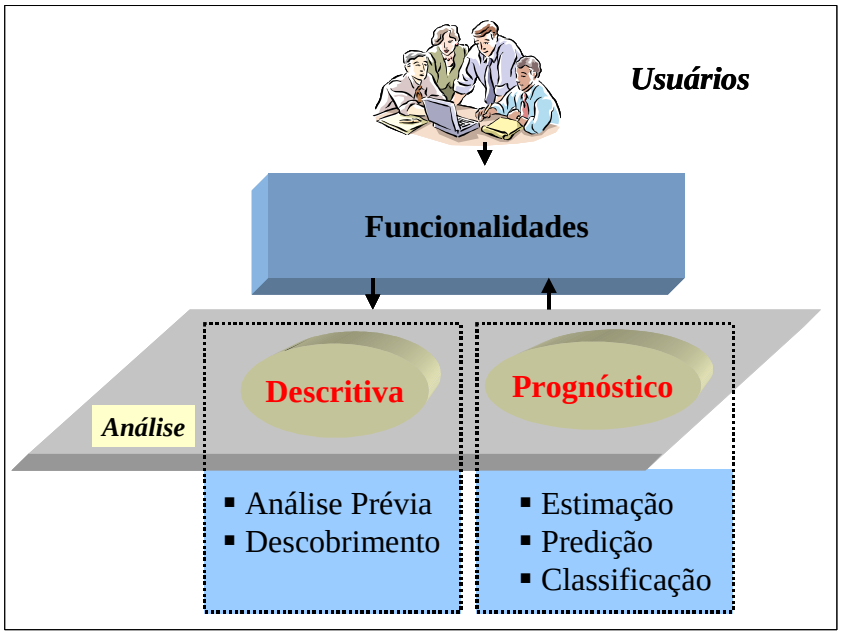

Figura 2: Funcionalidades em mineração de dados

Uma das vantagens de se focar as funcionalidades da mineração de dados dessa forma, diz respeito as facilidade que podem ser obtidas quando surge uma nova necessidade de análise de dados. Neste caso, basta identificar a que resultado se deseja chegar e imediatamente partir para identificação de que técnica aplicar. A seguir descreveremos os detalhes dessa abordagem.

2.2 Análise Descritiva

A *Análise Descritiva* representa a área de investigação nos dados que busca tanto descrever fatos relevantes, não-triviais e desconhecidos dos usuários, como analisar a base de dados, principalmente pelo seu aspecto de qualidade, para validar todo o processo da mineração e

seus resultados, ou seja, o conhecimento encontrado. Podemos subdividi-la em *Análise* e *Prévia Descobrimento*.

- *Análise Prévia* – é o processo de analisar uma base de dados com o objetivo de identificar anomalias ou resultados raros que possam influenciar os resultados da mineração de dados.

- *Descobrimento* – é o processo de examinar uma base de dados com o objetivo de encontrar padrões escondidos, sem que necessariamente exista uma idéia ou hipótese clara previamente estabelecida.

Para facilitar a aplicabilidade dos processos de mineração de dados, podemos especializar tanto a *análise prévia* quanto o *descobrimento* em outras *sub-funcionalidades* conforme a figura 3 a seguir.

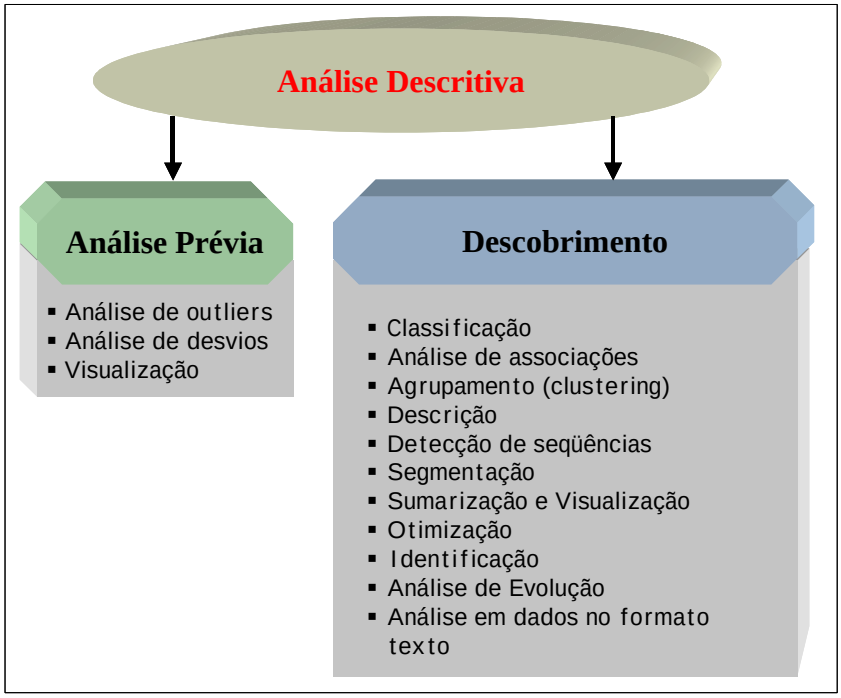

Figura 3: *Sub-funcionalidades* da análise prévia e do descobrimento.

À seguir especificaremos as sub-funcionalidades de cada funcionalidade descrita anteriormente.

2.2.1 Análise Prévia

A funcionalidade *análise prévia* pode ser aplicada usando-se uma das seguintes sub-funcionalidades:

2.2.1.1 Análise de Outliers ou detecção de desvios

Esta funcionalidade objetiva encontrar conjuntos de dados que não obedecem ao comportamento ou modelo dos dados. Uma vez encontrados podem ser tratados ou descartados para utilização no processo de mineração de dados. Trata-se de uma

importante avaliação nos dados no sentido de descobrir probabilidades crescentes de desvios ou riscos associados aos vários objetivos traçados inicialmente na mineração dos dados. Detectar esses desvios é muito análogo às técnicas utilizadas em análises estatística, onde são aplicados testes de significância que assumem uma distribuição, utilizando medidas estatísticas do tipo média aritmética e desvio padrão para aferir essas diferenças [HK01, WI99].

Como exemplo, podemos avaliar as vendas de uma determinada empresa para verificar o comportamento de suas vendas como um todo, bem como podemos avaliar suas vendas por produtos, regiões e estados, podendo encontrar outro tipo de comportamento. A figura 4 a seguir, extraída de [HK01], identifica visualmente a presença de *outliers*, onde os pontos externos aos polígonos são valores fora dos padrões da população (*vendas*) observada.

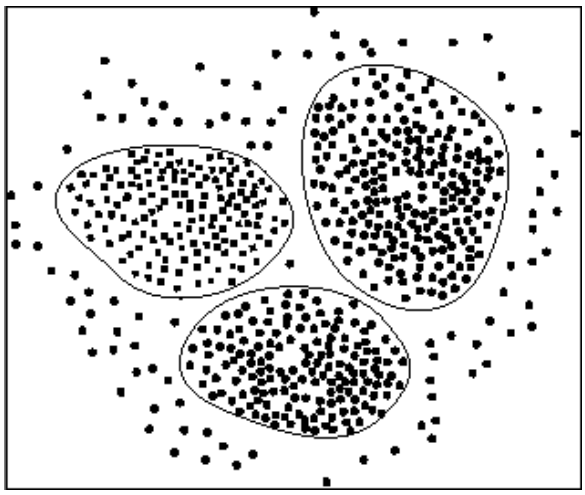

Figura 4: Detecção de *outliers* utilizando uma abordagem visual

2.2.1.2 Análise de desvios

Esta funcionalidade tem por objetivo detectar mudanças de comportamentos, comparando as ações com os padrões para detectar mudanças de comportamento [BT99]. Por exemplo, após avaliar o comportamento de clientes em um plano de saúde, qualquer alteração de comportamento pode ser imediatamente analisado e verificado. Essa mesma análise pode ser feita para fraudes em cartões de crédito, conhecendo-se o comportamento de compra dos clientes, entre outras diversas aplicações.

2.2.1.3 Visualização

Esta funcionalidade é utilizada, principalmente, quando não se tem nenhuma idéia da distribuição dos dados e se deseja encontrar algum tipo de disparidades nos dados. Por exemplo, construir histogramas por tempo de duração de chamadas telefônicas, no sentido de identificar os bairros de uma cidade onde o tempo de duração é maior ou menor do que nos outros bairro. Após essa análise, podemos identificar melhor como segmentar os dados ou selecionar atributos (*variáveis*) para formação de agrupamento (*clustering*). A figura 5 a seguir exemplifica a visualização empregada numa análise prévia deste tipo.

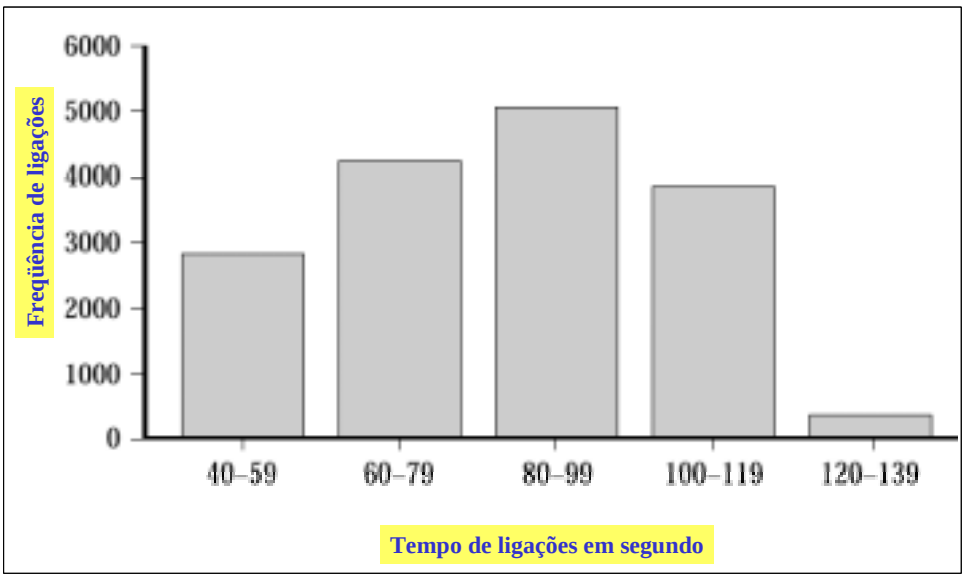

Figura 5: Detecção de *outliers* utilizando uma abordagem visual

2.2.2 Descobrimento

A funcionalidade *descobrimento* pode ser aplicada usando-se uma das seguintes sub-funcionalidades:

2.2.2.1 Classificação - descrição por classes e conceitos

A *classificação* consiste em examinar uma certa característica nos dados e atribuir uma classe previamente definida. Dados podem ser associados a classes ou a conceitos através de um processo de *discriminação* ou de *caracterização*. *Discriminação* se caracteriza por ter seu resultado obtido através da atribuição de um valor a um atributo no registro, em função de um ou mais atributos do mesmo. Por exemplo, em um supermercado podemos classificar os produtos por tipo como alimentício, vestuário, higiene e limpeza etc. Já *caracterização* é a sumarização de um atributo de estudo por uma característica de um ou mais atributos ([BL97], [HK01]. Por exemplo, podemos caracterizar um empregado pelo seu salário anual, identificando faixas da agregação mensal de seus salários em baixa, média e alta.

2.2.2.2 Análise de associações

Também conhecida na área de marketing como *grupos de afinidade* ou *análise de cestas de venda* esta funcionalidade objetiva determinar que “coisas” estão relacionadas, estão juntas, ou seja, descobrir as *regras de associação* condicionadas a valores de atributos que ocorrem juntos em um conjunto de dados. Se aplica nos casos em que deseja-se estudar preferências, afinidades, visando principalmente criar oportunidades para formação de “pacotes” para consumidores ([BL97], [HK01]). Por exemplo, uma vez observado que dois itens são freqüentemente adquiridos juntos num supermercado, pode-se preparar e oferecer estes produtos juntos, numa “mesma cesta”, pois existe grande afinidade na preferência de seus compradores.

2.2.2.3 Agrupamento (clustering)

Esta funcionalidade visa segmentar um conjunto de dados num número de subgrupos homogêneos ou clustering. Seu objetivo é formar grupos baseados no princípio de que esses grupos devem ser o mais homogêneos em si e mais heterogêneos entre si. A diferença fundamental entre a formação de agrupamento e a classificação é que no agrupamento não existem classes predefinidas para classificar os registros em estudo. Os registros são agrupados em função de suas similaridades básicas, ou seja, quando se deseja formar agrupamentos, seleciona-se um conjunto de atributos (*variáveis*) e em função da similaridade desses atributos são formados os grupos ([BL97], [HK01], [WI99]). Como exemplo, podemos utilizar dados de um recenseamento nacional para formar grupos de domicílios, utilizando os atributos escolaridade, profissão, faixa etária, sexo, número de filhos. Observa-se que não existem classes pré definidas e poderemos ter num mesmo grupo domicílios de estados geograficamente opostos, porém, semelhantes nestes atributos (*variáveis*). A figura 6 a seguir exemplifica três possíveis agrupamentos (*clustering*) formados à partir de um conjunto de dados.

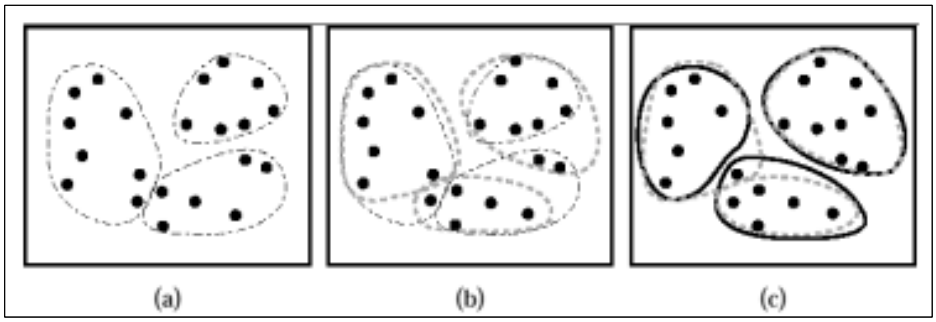

Figura 6: Três critérios diferentes de formação de agrupamentos (*clusters*)

2.2.2.4 Descrição

Esta funcionalidade pode ser empregada numa análise para tornar mais clara alguma idéia que está sendo utilizada, hipóteses ou fatos observados em um banco de dados. Pode ser executada em uma *análise de classificação* quando as classes não estiverem bem definidas ([BL97], [WB98]). Como exemplos, podemos estudar a afirmação de que "*mulheres tendem a votar em candidatas femininas em maior número do que os homens*" ou que as características de uma pessoa que fralda cartões de crédito é do tipo "*sexo masculino, idade entre 25 e 40 anos e possui nível superior*".

2.2.2.5 Detecção de seqüências

Esta funcionalidade tem por objetivo utilizar algum tipo de padrão nos dados para determinar que tipos de seqüências podem ser determinadas [BT99]. Por exemplo, clientes que compram determinado produto, três meses após retornarão para comprar algum outro produto associado ao primeiro (comprar um aparelho celular e três meses após um carregador de baterias para utilização em carros).

2.2.2.6 Segmentação

Nesta funcionalidade o conjunto de dados é subdividido em conjuntos menores, com comportamento similares nos atributos de segmentação. Com esses subconjuntos, pode-se determinar novos agrupamentos (*clustering*) ou mesmo fazer algum tipo de prognóstico. Difere da análise de agrupamento (*clustering)* pois a formação dos grupos é conduzida pelo usuário e não determinada pelo sistema, conforme descrito anteriormente ([Men99], [WB98], [WI99]). Por exemplo, podemos segmentar os registros dos bilhetes aéreos de uma companhia de aviação por cidade de origem, sexo e classe do assento. Após a segmentação, podemos formar agrupamentos (*clustering)* por profissão, faixa etária, estado de moradia, freqüência de viagens e faixa salarial para traçar um perfil de seus passageiros.

2.2.2.7 Sumarização e Visualização

Um dos principais objetivos da tecnologia de mineração de dados é oferecer seus resultados numa forma fácil de ser interpretado pelos usuários finais. Utilizar a sumarização de dados para facilitar o entendimento dos dados é uma estratégia muito usual que facilita e identifica inúmeras características nos dados em estudo. Uma das principais abordagens para descrição de informações é a visualização, principalmente quando o conjunto de dados a ser explorado não está organizado em uma forma padrão. Os resultados da sumarização e da visualização são normalmente utilizados em conjunto com outras funcionalidades [WI99]. Por exemplo, podemos imaginar um gráfico de colunas impresso num mapa do Brasil, indicando em cada estado o número de chamadas telefônicas realizadas no ano de 2000. Facilmente, podemos comparar esses resultados entre os estados. Se colocarmos os dados de dois anos, nossa análise será ainda mais rica. A figura 7 a seguir é um exemplo de mineração de dados fornecendo seus resultados com técnicas de sumarização e visualização.

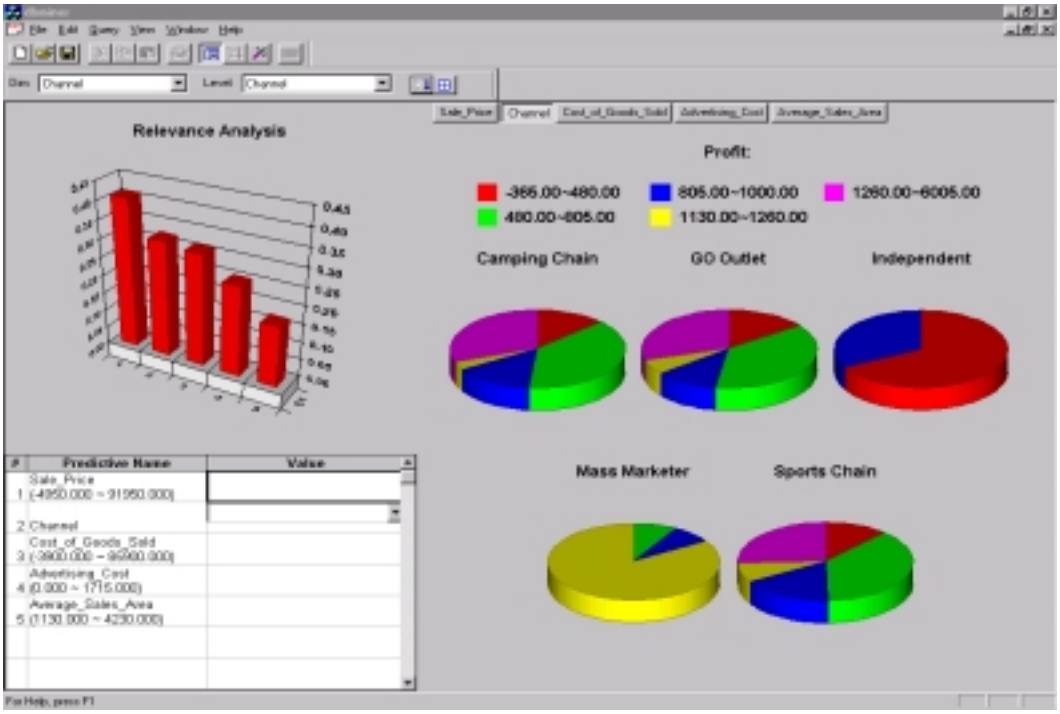

Figura 7: Mineração de dados com resultados da sumarização e visualização

2.2.2.8 Otimização

Esta funcionalidade visa otimizar recursos limitados como tempo, espaço, dinheiro, matéria-prima etc, buscando maximizar variáveis de resultado como vendas, lucros, distribuição, economia de espaço etc. Esta funcionalidade se aproxima dos estudos da área de pesquisa operacional, a qual trata de problemas de otimização, sempre sujeito a restrições ([EN99], [Men99]). Como exemplo, podemos estudar as vendas de um supermercado, no sentido de otimizar a distribuição de seus produtos em suas gôndolas, visando otimizar a exposição de um número cada vez maior de produtos.

2.2.2.9 Identificação

Esta funcionalidade tem por objetivo utilizar os padrões de dados para identificar a existência de um item, um evento ou uma atividade. Por exemplo, intrusos que tentam romper um sistema podem ser identificados através dos programas executados, dos arquivos acessados e do tempo de CPU por sessão. Em aplicações biológicas, a existência de um gene pode ser identificada através de certas seqüências de símbolos nucleotídeos (*nucleotides*) na seqüência de DNA. A área conhecida como autenticação é uma forma de identificação. Ela verifica se um usuário é de fato um usuário específico ou se pertence a uma classe autorizada; envolve uma comparação de parâmetros ou imagens ou sinais em relação ao banco de dados [EN99].

2.2.2.10 Análise de Evolução

Esta funcionalidade descreve e estuda a regularidade de modelos ou tendências para objetos cujo comportamento muda ao longo do tempo [HK01]. Como exemplo, podemos citar a evolução de estoques em que as empresas que necessitam constantemente decidir sobre a sua ampliação ou melhor distribuição de seus produtos, em função da regularidade das vendas da empresa.

2.2.2.11 Análise em dados no formato texto

Esta funcionalidade visa trabalhar os diversos dados armazenados no formato texto, tais como narrativas, processos judiciais etc, visando transformar esses textos em uma forma de uso e extrair seus resultados baseados em técnicas de tratamento e exploração de textos [WI99]. Como exemplo, podemos citar a exploração de dados no formato texto em processos criminais, no sentido de utilizá-los no reconhecimento de padrões e verificação de similaridade entre crimes.

Outras funcionalidades de descobrimento podem ser utilizadas tais como a comparação de imagens de satélites, as seqüências genéticas, a mineração em bancos de dados espaciais, a mineração em bancos de dados multimídias, a mineração dos dados da web etc., para citamos somente as mais utilizadas em negócios e já implementadas em ferramentas comerciais.

2.3 Análise de Prognóstico

A *Análise de Prognóstico* representa a área de investigação nos dados que busca inferir resultados a partir dos padrões encontrados na análise descritiva, ou seja prognosticar o comportamento de um novo conjunto de dados. Podemos subdividi-la em *Classificação*, *Estimação* e *Predição* [BL97, BT99 e HK01]. Embora muitos autores tratem a análise de prognóstico como simplesmente predição ou estimação, nos pareceu conveniente a separação para um melhor entendimento.

- *Estimação* – é o processo de predizer algum valor, baseado num padrão já conhecido. Por exemplo, conhecendo-se o padrão de despesas e a idade de uma pessoa, estimar seu salário e seu número de filhos.

- *Predição* – é o processo de predizer um comportamento futuro, baseado em vários valores. Por exemplo, baseado na formação escolar, no trabalho atual e no ramo de atividade profissional de uma pessoa, predizer que seu salário será de um certo montante até um determinado ano.

- *Classificação* – é o processo para predizer algum valor para uma variável categórica. Por exemplo, podemos num banco financeiro, determinar que conjunto de clientes oferecem risco ou não para contrair um empréstimo pessoal.

Não dividimos a análise de prognóstico em sub-funcionalidades conforme a análise descritiva, pois ao nível de funcionalidade as definições acima esgotam com bastante clareza a sua aplicação. Maiores detalhes sobre modelagem de prognósticos podem ser encontrados em ([BL97], [BT99] e [HK01]) e, principalmente, em [WI99].

2.4 Técnicas para obtenção das funcionalidades

Uma vez definidas as funcionalidades (*resultados*) a que se deseja chegar com o processo de mineração de dados, cabe agora escolher que *técnicas* devemos utilizar, que sejam mais aderentes para a obtenção dos resultados, com uma melhor precisão. Por exemplo, a funcionalidade de *estimação* pode ser feita utilizando-se a técnica de regressão linear ou regressão múltipla. Entretanto, sabemos que para uma estimativa de curtíssimo prazo e com poucas variáveis a regressão linear é mais fácil de ser utilizado e proporciona bons resultados, entretanto, para estimativas de longo prazo não é a mais indicada. A tabela 1 a seguir mostra um conjunto parcial de técnicas que podem ser utilizadas em cada funcionalidade:

Funcionalidade	Sub-funcionalidade	Técnica
Análise Prévia		
	Análise de outliers	Ferramentas de consulta e técnicas de estatística
		Indução por árvores de decisão
	Análise de desvios	Ferramentas de consulta e técnicas de estatística
		Indução por árvores de decisão
	Visualização	Agregações e gráficos diversos
Descobrimento		
	Classificação	Indução por árvores de decisão
	Análise de associações	Mineração de Regras de associação (Análise da cesta de venda - *Market basket analysis)*
		Minerando regras de associação booleanas unidimensionais a partir de bancos de dados transacionais
		Minerando regras de associação em múltiplos níveis a partir de bancos de dados transacionais
		Minerando regras de associação multidimensionais a partir de bancos de dados transacionais e data warehouse
		Da mineração de associação à análise de correlação
		Mineração de associação baseada em restrição

	Agrupamento (clustering)	Métodos de particionamento
		Métodos hierárquicos
		Métodos baseados em densidade
		Métodos baseados em grid
		Métodos de clustering baseados em modelos – abordagem estatística e redes neurais
		Análise de outliers
	Descrição do Conceito - (caracterização e comparação)	Sumarização e Generalização dos dados baseados em caracterização
		Caracterização analítica – análise da relevância do atributo
	Segmentação	Indução por árvores de decisão
	Sumarização e Visualização	Agregações e gráficos diversos
	Análise em dados no formato texto	Análise de dados textual e recuperação de informações
		Mineração de textos – classificação de documentos e associação por palavras chaves
Estimação/Predição	*Estimação/Predição*	Regressão Linear
		Regressão Múltipla
		Regressão não linear
		Regressão Logística
		Regressão de Poisson
		Outros modelos de regressão
Classificação	*Classificação*	Indução por árvores de decisão
		Classificação bayseana
		Classificação por backpropagation – *Redes Neurais Artificiais*
		Classificação baseada em conceitos da mineração de regras de associação
		Classificação por Backpropagation –

		Redes Neurais
		Análise de vizinhança (k-Nearest Neighbor)
		Casos baseados em Raciocínio
		Algoritmos genéticos
		Abordagem por conjuntos fuzzy

Tabela 1: Funcionalidades e suas técnicas

A seguir descreveremos algumas técnicas que são utilizadas no processo de mineração de dados.

2.4.1 Ferramentas de consulta e técnicas de estatística

O primeiro passo em um projeto de mineração de dados pode ser uma análise simples, preliminar, "grosseira" do conjunto de dados que será minerado, utilizando-se de ferramentas de consultas. Aplicando-se as funções *built-in* da linguagem SQL de um banco de dados relacional, podemos obter informações bastante ricas sobre a distribuição dos dados. Antes de aplicarmos algoritmos avançados de reconhecimento de padrões, precisamos conhecer os aspectos e estruturas do conjunto de dados que iremos minerar. Estatísticas como média aritmética, desvio padrão, valores máximos e mínimos e distribuição percentual de todo o conjunto de dados ou por grupos (*utilizando-se a cláusula group by*) representam os passos iniciais num processo de mineração de dados. Além dessas consultas e estatísticas, vários gráficos podem ser preparados utilizando-se os dados e estatísticas gerados para facilitar as análises iniciais [AZ96].

2.4.2 Visualização

A técnica de visualização de dados é extremamente útil como técnica de descobrimento de padrões em conjunto de dados e pode ser largamente utilizada no início do processo de mineração de dados. Embora possa parecer uma técnica não muito sofisticada, permite que se tenha uma medida inicial da qualidade dos dados e de onde os padrões possam ser encontrados.

Quando utilizada nos processos mais avançados da mineração de dados, possibilita a utilização de gráficos tri-dimensional de forma interativa, gráficos hierárquicos para segmentação da base de dados em formato de árvores, entre outras formas de visualização. A figura 8 a seguir apresenta um resultado da mineração de dados utilizando a técnica de visualização na disposição de produtos em depósitos de uma empresa.

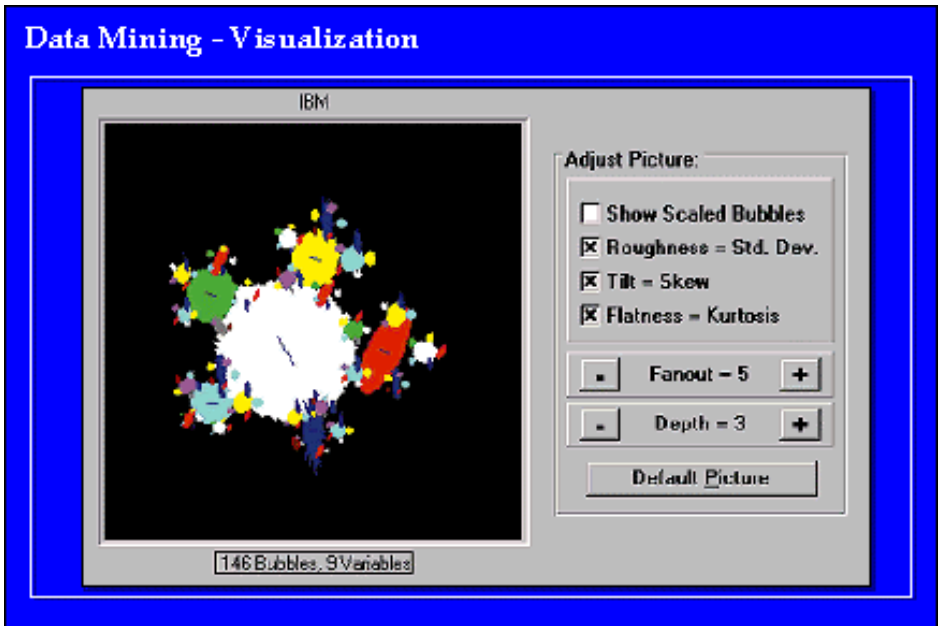

Figura 8: Um exemplo de visualização de dados

2.4.3 Análise de vizinhança (*K-nearest neighbor*)

Quando se interpreta dados como um ponto no espaço, nós precisamos definir o conceito de vizinhança, o qual significa identificar o conjunto de registros que estão próximos, que são "*fechados*" por alguma característica nos dados. Está técnica é tipicamente uma técnica de pesquisa e não de conhecimento que é empregada principalmente na análise de prognósticos. Por exemplo, podemos estimar a renda de um indivíduo de uma população, pesquisando *k=20* vizinhos mais próximos do mesmo pelos valores dos atributos bairro de moradia, profissão, escolaridade e idade. Um dos problemas da aplicação dessa técnica é a necessidade de existir nos registros um número de atributos suficientes para determinação da vizinhança.

[EN99],

2.4.4 Árvores de decisão

Uma árvore de decisão é um fluxograma (*flow-chat*) semelhante a uma estrutura de árvore, onde cada nó interno denota um teste em um atributo, cada ramo (sub-árvore) representa o resultado do teste e cada folha representa a distribuição dos registros. Quando utilizada na análise de prognóstico, em classificação, sua aplicação é chamada por alguns autores de *indução por árvore de decisão.* Sua utilização recomenda o treinamento do método, utilizando-se várias amostras nos dados, até que se conheça as melhores regras para segmentação do conjunto de dados. Um outro problema que deve ser estudado é a poda da árvore, ou seja, determinar quantas sub-árvores, particionamentos, será necessário gerar. A figura 9 a seguir apresenta uma classificação utilizando um algoritmo de árvore de decisão, para prognosticar o grupo de cliente mais propício a comprar um determinado produto. Examinado a figura 9, observa-se que 90% dos homens com salário superior a R$ 4.000,00 são candidatos a comprarem o produto, enquanto que apenas 5% das pessoas que ganham menos de R$ 4.000,00 e não possuem casa própria devem comprar o produto.

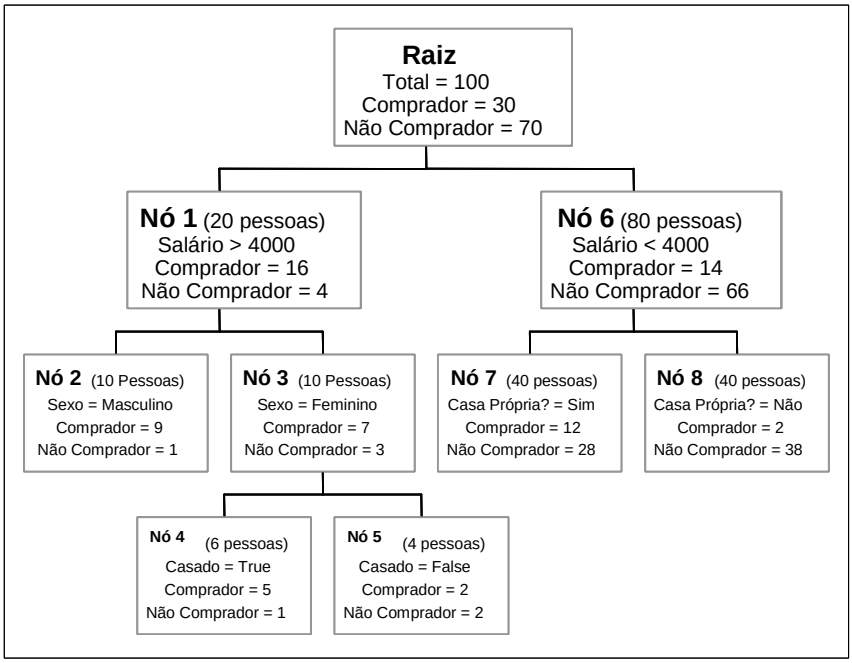

Figura 9: Um exemplo de visualização de uma árvore de decisão

2.4.5 Regras de associação

Análise de associação é o processo de interconexão de objetos na tentativa de expor características e tendências. Gera redes de interações e conexões presentes nos conjuntos de dados usando as associações item a item. Entende-se que a presença de um item implica necessariamente na presença do outro na mesma transação. O banco de dados é visto como uma coleção de transações, cada uma envolvendo um conjunto de itens. Essas regras

correlacionam a presença de um conjunto de itens com um outro intervalo de valores para um outro conjunto de variáveis. Um exemplo comum é aquele referente à cesta do supermercado. Neste caso, a cesta do supermercado corresponde àquilo que o consumidor compra em um supermercado durante uma visita [EN99, DN00]. Na área de marketing é conhecido como análises de transações de compras (*market basket analysis*). A figura 10 a seguir apresenta uma representa gráfica de um estudo de transações de compras para ser resolvido com regras de associação.

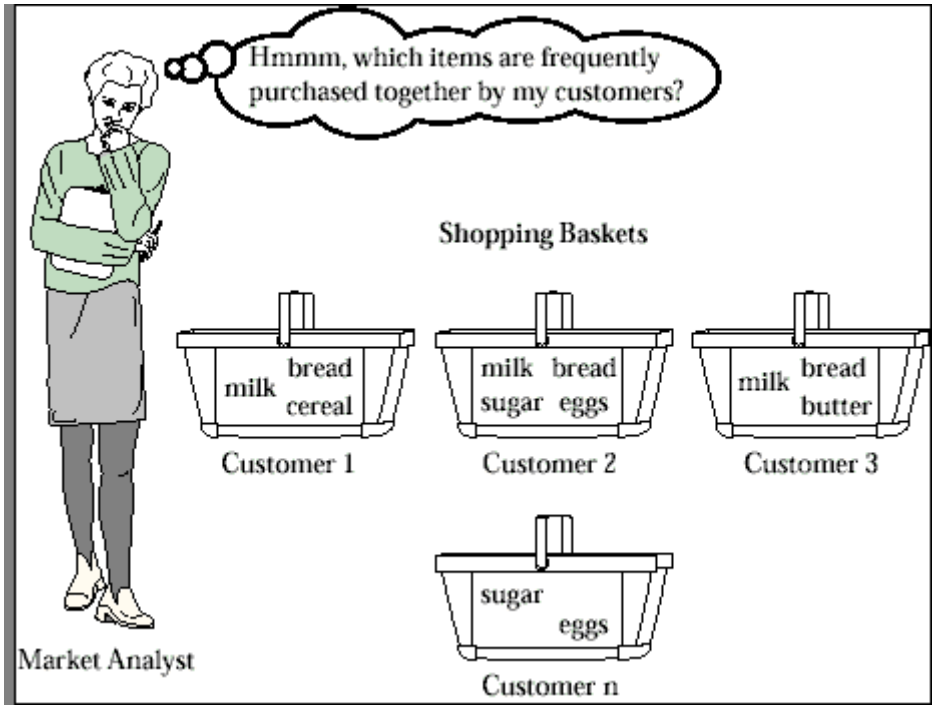

Figura 10: Um exemplo de análises de transações de compras extraído de [HK01]

Um estudo completo sobre regras de associação pode ser encontrado em [MPO01].

2.4.6 Redes neurais artificiais

As redes neurais compreendem procedimentos computacionais que envolvem o desenvolvimento de estruturas matemáticas com habilidade de aprendizado. Representam o esforço de investigações acadêmicas para implementar computacionalmente, a maneira pela qual o cérebro humano funciona. São programas que implementam detecções sofisticadas de padrões e algoritmos de aprendizado de máquina, para construir modelos, principalmente, de prognóstico de grandes bancos de dados históricos. Está baseada nos conceitos de como um cérebro humano está organizado e como ele aprende. Existem duas estruturas principais: (1) O nó, que corresponde ao neurônio; (2) O link, que corresponde as conexões entre neurônios. Segundo [Hay99] redes neurais pode ser definidas como:

"*Uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para uso. Ela se assemelha ao cérebro em dois aspectos":*

1. *O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem;*

2. *Forças de conexão entre neurônios, conhecidos como pesos sinápticos, são utilizados para armazenar o conhecimento adquirido.*"

A figura 11 a seguir, exemplifica as várias camadas que podem ser geradas num processamento de uma rede neural. Todas as camadas intermediárias representam os diferentes níveis de conhecimento que são adquiridos no seu processamento, numa tentativa de emitar o cérebro humano.

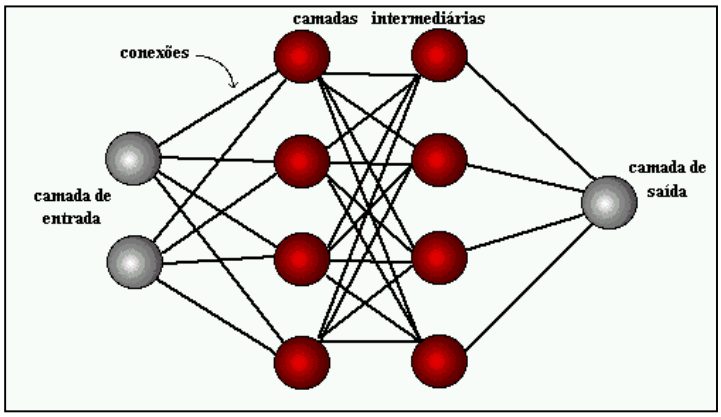

Figura 11: Uma representação de um processamento de uma rede neural

Um estudo completo sobre regras redes neurais pode ser encontrado em [Hay99].

2.4.7 Algoritmos Genéticos

Algoritmos Genéricos – AGs são algoritmos de otimização e busca baseados nos mecanismos de seleção natural e genética. Enquanto os métodos de otimização e busca convencionais trabalham geralmente de forma seqüencial, avaliando a cada instante uma possível solução, os AGs trabalham com um conjunto de possíveis soluções simultaneamente [BLC00]. Segundo [EN99], Algoritmos Genéticos (AGs) são uma classe de procedimentos de pesquisa aleatórios capazes de realizar pesquisas adaptativas e robustas sobre uma ampla gama de topologias de espaço de pesquisa. Modelados após o surgimento adaptativo de espécies biológicas a partir de mecanismos evolutivos e introduzidos por Holland,[4] AGs vêm sendo aplicados com sucesso em campos diversificados como análise de imagens, escalonamentos e projetos de engenharia.

As soluções produzidas por algoritmos genéticos (AGs) são diferenciadas da maioria das outras técnicas de pesquisa através das seguintes características:

- Uma pesquisa de AG utiliza um conjunto de soluções durante cada geração ao invés de uma única solução.

- A pesquisa no espaço de strings representa uma pesquisa paralela maior no espaço de soluções codificadas.

[4] O trabalho de Holland entitulado "Adaptation in Natural and Artificial Systems ("Adaptação em Sistemas Naturais e Artificiais"), apresentado em um seminário, introduziu a idéia de algoritmos genéticos.

- A memória da pesquisa realizada é representada unicamente através do conjunto de soluções disponíveis

- Um algoritmo genético é um algoritmo aleatório, uma vez que mecanismos de pesquisa utilizam operadores de probabilidade.

- Ao prosseguir de uma geração para a seguinte, um AG encontra o equilíbrio próximo ao ótimo entre aquisição e exploração de conhecimento, manipulando soluções codificadas.

Algoritmos genéticos são utilizados para resolver problemas e para agrupar problemas. Sua capacidade de resolver problemas em paralelo fornece uma ferramenta poderosa para Mineração de dados. As deficiências de AGs incluem a grande superprodução de soluções individuais, o caráter aleatório do processo de pesquisa e a elevada demanda no processamento computacional. Em geral, uma substancial demanda computacional é exigida para alcançar qualquer coisa significativa com algoritmos genéticos [EN99].

2.4.8 Técnicas de análise de agrupamento (*clustering*)

Os métodos de particionamento, hierárquicos, baseados em densidade, em grid, em modelos (abordagem estatística e redes neurais) são técnicas da análise de agrupamento (*clustering*) que visam detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso da existência, determinar estes grupos. São técnicas de mineração de dados que estão direcionadas aos objetivos de *identificação* e *classificação*. O *clustering* tenta identificar um conjunto finito de categorias ou *clusters* para os quais cada registro (elemento da população) possa ser mapeado. As categorias podem ser disjuntas (separadas) ou sobrepostas (não-disjuntas) e podem algumas vezes ser organizadas em árvores. A população forma um agrupamento que pode ser dividido em dois ou mais grupos, que podem ser novamente divididos em dois ou mais grupos, e assim por diante, incluindo a partição onde cada elemento é um único elemento do grupo.

Um estudo completo sobre análise de cluster pode ser encontrado em [BL97] e [HK01].

2.4.9 Conclusão sobre as técnicas de mineração de dados

Várias outras técnicas de mineração de dados estão em uso nos dias de hoje, conforme visto na tabela 1. Elas incluem lógica fuzzy, redução de dados, classificação baysiana entre outras. A figura 12 a seguir, extraída de [BT99], sintetiza um ciclo de vida para operação das técnicas de mineração de dados.

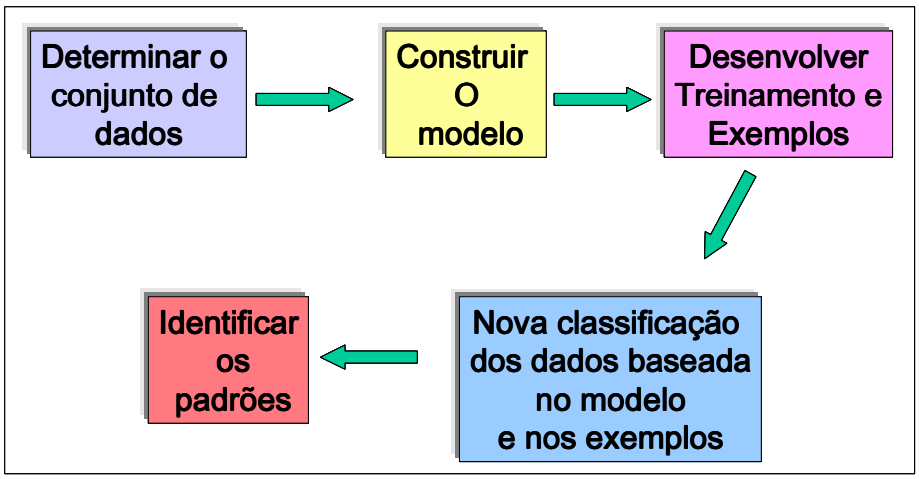

Figura 12: Ciclo de vida de operação das técnicas de mineração de dados

Um amplo estudos sobre as técnicas de mineração de dados pode ser encontrado em [BL97] e [HK01].

2.5 Abordagens da mineração de dados

As abordagens da mineração de dados ou metodologias de aplicação descrevem como o usuário irá conduzir o processo da mineração na obtenção de suas funcionalidades. Essencialmente existem as abordagens top-down e botton-up, e uma terceira que pode ser a combinação dessas abordagens chamada de híbrida. Na abordagem top-down, também chamada de *teste de hipótese*, o usuário parte do princípio que existe uma hipótese, uma idéia pré-concebida e que mesmo deseja confirmá-la ou refutá-la. Á na abordagem botton-up, também chamada de *busca de conhecimento*, o usuário inicia o processo de exploração dos dados na tentativa de descobrir alguma coisa que ainda não é de conhecimento [BT99, BL97].

Na aplicação de uma dessas abordagens o usuário decidirá se usará a abordagem para busca de conhecimento na forma *direta* ou *indireta*. A seguir é descrito essas duas formas de aplicação.

2.5.1 Busca de conhecimento direta

Na busca de conhecimento *direta* ou *supervisionada* sua meta é orientada. Existe um valor para ser prognosticado, uma classe a ser atribuída aos registros ou um determinado relacionamento para ser explorado. Existe apenas uma vaga idéia do que se estar procurando. Os passos para aplicação da busca de conhecimento direta são:

- Identificar as fontes dos dados selecionados para mineração;
- Prepara os dados para análise;
- Construir e trinar o modelo computacional;

- Avaliar o modelo computacional.

Maiores detalhes sobre esses passos podem ser encontrados em [BL97] e [BT99].

2.5.2 Busca de conhecimento indireta

Na busca de conhecimento *indireta* ou *não-supervisionada* não existe uma meta bem definida. As ferramentas são mais livres na sua aplicação sobre os dados e esperá-se que será descoberto alguma estrutura significante nos dados. Os passos para aplicação da busca de conhecimento direta são:

- Identificar as fontes dos dados;
- Prepara os dados para análise;
- Construir e trinar o modelo computacional;
- Avaliar o modelo computacional;
- Aplicar o modelo computacional no novo conjunto de dados;
- Identificar potenciais objetivos para busca de conhecimento direta;
- Gerar novas hipóteses para teste.

Maiores detalhes sobre esses passos podem ser encontrados em [BL97] e [BT99].

A figura 13 abaixo, extraída de [BT99], resume a forma de aplicação do processo de mineração de dados.

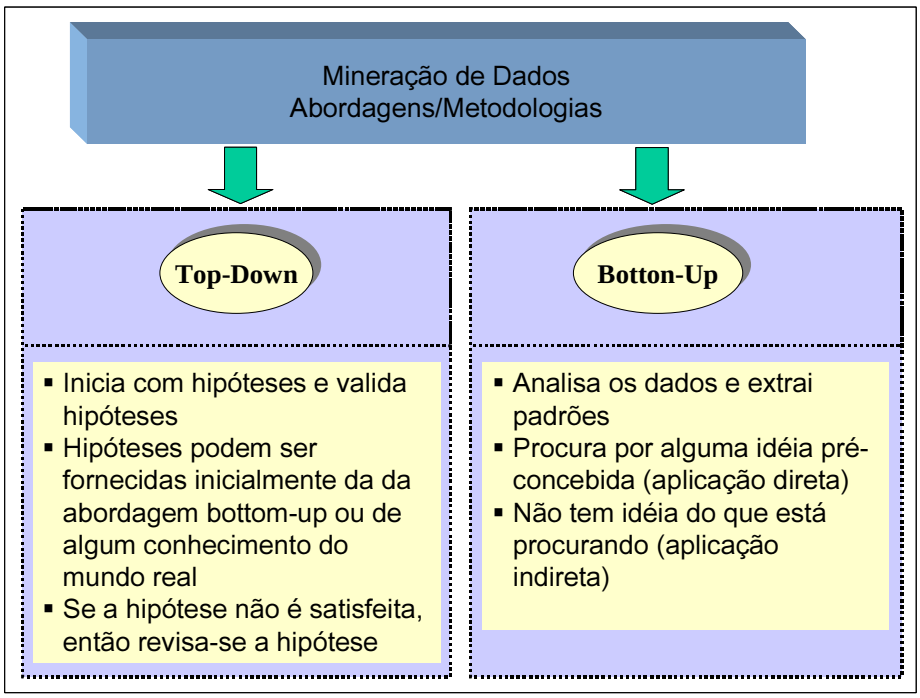

Figura 13: Abordagens para aplicação da mineração de dados

2.6 Conclusão

Conforme descrito nesta seção para utilização de um processo de mineração de dados, deve-se ter bem claro qual a funcionalidade ou os resultados a que se deseja chegar. A escolha da funcionalidade, na maioria dos casos, exige a participação de pessoas que efetivamente entendam do negócio em estudo, mesmo que não sejam especialistas na utilização e manuseio computacional dos dados. Uma vez definidas as funcionalidades parte-se para identificar a melhor técnica, a mais aderente para obtenção dos resultados. Diversas técnicas podem ser utilizadas para se chegar aos resultados pretendidos, entretanto, cada técnica possui suas características, suas peculiaridades e precisa de pessoas que saibam interpretar seus resultados. Uma vez identificadas as funcionalidades e as melhores técnicas a serem aplicadas, deve-se escolher uma abordagem, uma metodologia de aplicação para condução de todos os processos.

A próxima seção tratará dos processos de *Busca de Conhecimento em Banco de Dados*, também conhecido como *Knowledge Discovery in Database - KDD*, com destaque para as etapas de preparação dos dados para mineração propriamente dita.

3 O processo Busca de Conhecimento em Banco de Dados

Para alguns, mineração de dados representa o passo essencial, principal, no processo de Busca de Conhecimento em Banco de Dados (*Knowledge Discovery in Database - KDD*). Segundo [HK01], o processo de KDD consiste de uma seqüência iterativa dos seguintes passos:

1. **Limpeza de dados** - remove dados inconsistentes e fora dos padrões (*noise data*);

2. **Integração de dados** - possibilita a integração de várias fontes de dados, mantendo a consistência e coerência dos dados integrados;

3. **Seleção dos dados** - seleciona os dados relevantes para aplicação das técnicas de mineração de dados;

4. **Transformação de dados** – possibilita a transformação ou consolidação dos dados no formato apropriado para o processo de mineração (*mining*), através de operações do tipo sumarização ou agregação, entre outras técnicas;

5. **Mineração dos dados** – processo essencial, onde técnicas são aplicadas para análise e extração de padrões dos dados;

6. **Avaliação dos Padrões** – identifica os padrões verdadeiramente interessantes entre os diversos apresentados pelo processo de mineração de dados, baseados em algumas medidas de interesse;

7. **Apresentação e assimilação do conhecimento** – utiliza técnicas de visualização e representação do conhecimento para apresentar o conhecimento adquirido aos usuários, bem como introduzi-los no âmbito estudado.

A figura 14 à seguir, uma adaptação de [HK01] e [AZ96], apresenta a interação entre todas as etapas do processo de KDD. As etapas de limpeza e integração só são necessárias quando os dados que serão utilizados na mineração estão armazenados em vários bancos de dados ou arquivos do legado. Quando a seleção dos dados for em um Data Warehouse[5] podemos iniciar o processo selecionando e transformando os dados. Uma vez preparados, os dados são submetidos uma técnica de mineração de dados, conforme especificado na seção 2. Os resultados extraídos da aplicação dessas técnicas são avaliados e interpretados, e podem ser reconhecidos como padrões. Os padrões interessantes são apresentados aos usuários e são armazenados como uma nova base de conhecimentos Uma vez apresentados e assimilados transformam-se numa base de conhecimento e geram ações em seu negócio específico.

[5] Um Data Warehouse é um banco de dados que possui seus dados resultante da integração de dados de várias fontes, já possuindo um alto padrão de qualidade.

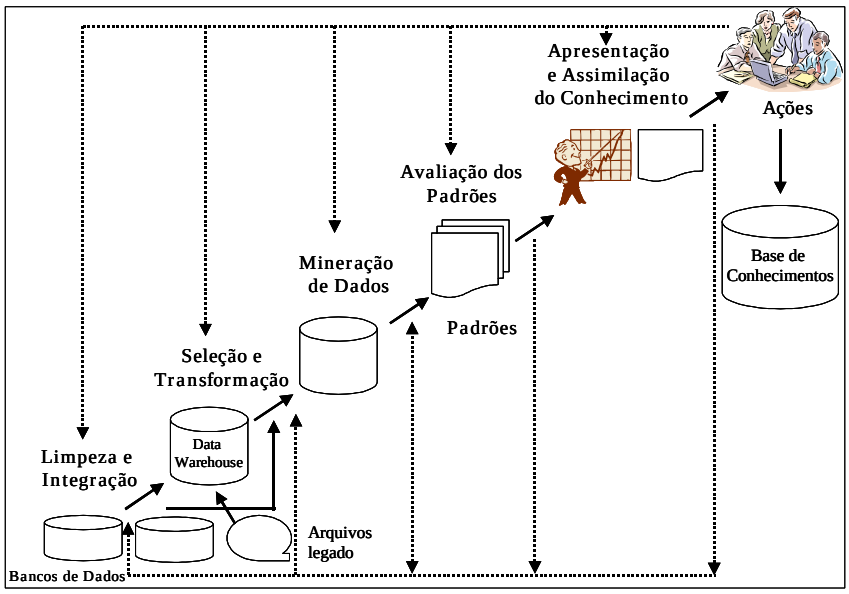

Figura 14: Mineração de dados como uma etapa no processo de KDD

Cada etapa do processo de KDD pode retornar a um processo anterior, conforme sua necessidade [AZ96]. Esta necessidade pode se dá em função de uma reavaliação nos dados, uma nova hipótese a ser testada etc. Note que segundo essa visão, mineração de dados é apenas uma etapa no processo de KDD, essencial para descobrir padrões para avaliação, padrões estes até então escondidos nas bases de dados.

Os bancos de dados do mundo real são altamente suscetíveis a armazenarem dados incoerentes, inconsistentes, grande quantidade de valores ausentes e geralmente armazenam uma quantidade de dados em torno de muitos gigabytes e terabytes. Preparar os dados para o processo de mineração de dados, significa melhorar a qualidade dos dados a serem processados e conseqüentemente a qualidade dos resultados obtidos. Preparar os dados para o processo de mineração de dados, envolve preparar e executar as fases de limpeza, integração, seleção e transformação de dados. Essas fases do processo de KDD estão detalhadamente examinadas a seguir.

3.1 Limpeza de dados (data cleaning)

Os dados no mundo real tendem a ser incompletos, fora de padrões e inconsistentes. As rotinas de limpeza de dados empreendem esforços no sentido de preencher os valores ausentes (*missing values*), aplainar dados (padronizar - *noise data*) enquanto identificam valores fora de padrões (*outliers*) e corrigem inconsistências nos dados [HK01]. Existem vários tipos de processos de limpeza que podem ser aplicados inicialmente, outros, no entanto, podem ser aplicados somente após a detecção de algum tipo de problema nas etapas subseqüentes do KDD, mineração de dado ou avaliação de padrões [AZ96].

3.1.1 Valores ausentes *(missing values)*

Valores ausentes se caracterizam por existirem em diversas tuplas (ou registros) atributos (campos) que não possuem valores armazenados, os quais podem ser importantes para o processo de mineração de dados. Como exemplo, o atributo rendimento em uma instância

de uma entidade cliente. A seguir apresentamos alguns métodos que podem ser utilizados para preencher (atribuir) valores a esses atributos [HK01].

1. *Ignorar o registro (tupla)* – usualmente utilizada quando o conteúdo da variável está ausente - null (assumindo que o processo de mining envolverá classificação ou descrição). Não é muito efetivo, a menos que a tupla possua muitos atributos com valores ausentes. É especialmente pobre quando o percentual de valores ausentes varia consideravelmente.

2. *Preencher (imputar) o valor manualmente* – em geral essa abordagem consome muito tempo e pode não ser possível em grandes bases de dados com muitos valores ausentes.

3. *Usar uma constante global para preencher os valores ausentes* – atualizar todos os valores ausentes com um único valor constante, tal qual "*desconhecido*" ou – *high values*. Embora simples, não é muito recomendado.

4. *Utilizar um atributo médio para preencher os valores ausentes* – utilizado quando o atributo é do tipo numérico e seu significado é passível de utilização de um valor médio. Calcular o valor médio do atributo em estudo (por exemplo, rendimento do cliente) e atribuir esse valor a todos os valores ausentes do atributo.

5. *Utilizar um atributo médio pertencente a mesma classe a qual a tupla pertença* – utilizado quando o atributo é do tipo numérico e seu significado é passível de utilização de um valor médio. Calcular valores médios do atributo em estudo segundo os valores (*classificação*) de um outro atributo (por exemplo profissão do cliente) e atribuir esse valor a todos os valores ausentes do atributo (por exemplo rendimento do cliente), segundo a classificação do atributo.

6. *Utilizar o valor de maior probabilidade para preencher os valores ausentes* – este valor pode ser determinado através da aplicação de uma técnica de regressão, ferramentas de inferência básica, utilizando um formalismo bayseano ou indução por árvores de decisão. Por exemplo, utilizar outros atributos do conjunto de dados de clientes para construir uma árvore de decisão para predizer (estimar) o rendimento dos mesmos.

Os métodos de 3 a 6 inferem sobre o dado. O valor imputado pode não ser correto. O método 6, entretanto, é o mais popular. Comparando com outros métodos, ele considera mais informações sobre os dados para predizer valores ausentes. Assim, a utilização de outros atributos na estimação dos valores ausentes, nos dá uma grande chance de preservar o relacionamento entre o atributo estimado e os demais atributos utilizados no processo de estimação.

3.1.2 Valores fora de padrão (noisy data)

Noisy data (valores extremos) é um erro aleatório ou uma variação acentuada na medição de uma variável. Ocorre em variáveis numéricas do tipo rendimento, faturamento etc e que precisam ser aplainadas (*smooth*), retirando-se esse erro de medição. A seguir apresentamos alguns métodos que podem ser utilizados para corrigir esses valores [HK01].

1. *Binning* – esse método ordena os valores do atributo para utilizar o conceito de vizinhança entre os dados. Após a ordenação os valores são distribuídos por grupos (bins ou buckets), onde cada grupo deverá ter o mesmo número de elementos (valores). Em cada grupo aplica-se um critério na escolha de uma medida para ajustar os valores dos grupos, tais como a média aritmética, a mediana ou um valor de limite. Assim, substituí-se os valores pelas medidas calculadas em cada grupo, ajustando assim os valores da série. Diversos métodos podem ser utilizados para ajustar os valores dos grupos.

2. *Agrupamento (Clustering) – Outliers* podem ser detectados quando valores similares são organizados em grupos ou clusters. Intuitivamente, valores que estão fora dos clusters podem ser considerados como outliers. A figura 15 à seguir, extraída de [HK01], mostra aplicação da técnica de agrupamento para detecção de outliers.

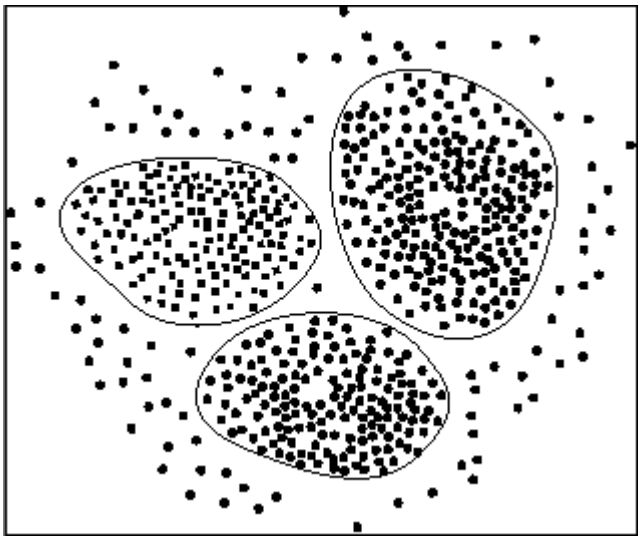

Figura 15: Detectando outliers com a técnica de análise de agrupamento (*clustering*)

3. *Combinação de inspeção humana e computador – outliers* podem ser identificados através da combinação de inspeção humana e do uso do computador. Em uma aplicação, uma medida teórica pode ser utilizada para ajudar a identificar padrões de *outliers*. Uma vez identificados os padrões de *outliers* e sendo estes considerados lixo, podem ser excluídos da utilização na etapa de mineração de dados.

4. *Regressão* – dados podem ser ajustados (*smoothed*) por funções de ajustamentos de dados, tais como funções de regressão. Regressão linear busca encontrar a melhor linha de ajustamento para duas variáveis, desde que uma possa ser estimada (predita) pela outra. Regressão linear múltipla é uma extensão da regressão linear, onde duas ou mais variáveis são envolvidas e os dados são combinados numa superfície (plano) multidimensional.

3.1.3 Dados inconsistentes

Podem haver inúmeras inconsistências nos dados armazenados. Alguns dos dados inconsistentes podem ser corrigidos manualmente utilizando referências externas, como

erros causados por entrada de dados manual. Ferramentas de software de engenharia de conhecimento podem também ser utilizadas para detectar violações nas restrições de integridades dos dados, tais como funções de dependência entre atributos. Podem também existir inconsistências causadas por integrações de dados, onde um atributo pode possuir nomes diferentes em seus bancos de dados. Podemos ter redundância de dados [HK01].

3.2 Integração de Dados

O processo de mineração de dados freqüentemente requer a integração (união, consolidação, fusão) de várias bases de dados. Neste processo, provavelmente, também existirá a necessidade de transformar os dados integrados em um formato apropriado para o processo de mineração de dados [HK01,WB98].

3.2.1 Integração de diversas fontes de dados

A integração de dados geralmente envolve a combinação de várias fontes de dados em um único armazenamento coerente, semelhante as bases de dados geradas no processo de construção de um data warehouse. Essas fontes de dados podem incluir vários bancos de dados, arquivos textos, flat files entre outros tipos de armazenamento. Existem três pontos importantes na integração de esquemas:

- *Integração de esquemas internos* – diversas entidades do mundo real podem ser semelhantes e estarem definidas em diversos esquemas com nomes e atributos diferentes, recaindo num problema típico de identificação de entidades. Quando se tem apenas bancos de dados operacionais e/ou data warehouses para integração, esta atividade se torna mais fácil, uma vez que esses armazenamentos devem possuir metadados, os quais ajudam a evitar esse tipo de problema.

- *Redundância de dados* – um atributo pode ser redundante se ele puder ser derivado de outro armazenamento (*tabela*), tal como o atributo agregado *salário anual*. Inconsistências em atributos ou nome de dimensões (salário, salário anual etc.) podem ser a causa de redundância em conjunto de dados. Uma técnica muito interessante para verificar redundância em conjunto de dados é a utilização da análise de correlação, a qual medirá o quanto dois atributos são correlatos. A redundância a nível de atributo também pode ser identificada através da geração de registros (tuplas) idênticas geradas numa mesma entrada de dados.

- *Detecção e resolução de valores conflitantes* – para as mesmas entidades do mundo real, os valores dos atributos podem diferir em diversas fontes de dados. Preços de produtos, diárias de hotéis, salários de empregados etc, numa mesma empresa podem ser registrados em unidades e moedas diferentes, incluindo ou não parcelas de valores tais quais impostos ou taxas. A heterogeneidade semântica dos dados pode causar grandes desafios na integração dos dados.

Outros fatores como os vários formatos de armazenamento dos dados, tais como armazenamentos em bancos de dados relacionais, de rede e hierárquico, arquivos textos, campos fixos e variáveis, entre outros formatos irão afetar a forma como se recupera e integra os dados. A variedade dos sistemas operacionais e plataformas de hardware

também são fatores que dificultam o acesso aos inúmeros protocolos para recuperação e integração dos dados.

Cuidados na integração dos dados oriundos de várias fontes podem ajudar a reduzir e evitar redundâncias e inconsistências no resultado do conjunto de dados gerado na integração. Certamente irá melhorar a precisão dos resultados e a velocidade das fases subseqüentes dos processos de garimpagem.

3.3 Seleção de dados

Nesta etapa será identificado e selecionado todos os dados que são necessários para o processo de mineração de dados. Vale a pena ressaltar que esta etapa ocorrendo após a etapa de integração, possibilita a seleção somente do conjunto de dados que possa ser efetivamente utilizado e que sua integração já garantiu a coerência entre as diversas fontes de dados utilizadas.

3.4 Transformação de Dados

Nesta etapa os dados são transformados e consolidados em formatos apropriados para a atividade de garimpagem (mining). A transformação de dados envolve:

- *Smooting (aplainamento)* – este trabalho remove os noisy data. Utiliza técnicas de binning, agrupamento e regressão.

- *Agregação* – aplica operações de sumarização e agregação nos dados. Por exemplo, vendas diárias são agregadas em vendas semanais, quinzenais e mensais. Tipicamente usada para geração de dados no formato multidimensional em dados com alta granularidade[6] (muitos detalhes).

- *Generalização* – generalização dos dados é a etapa que permite transformar os dados primitivos, como linhas de tabelas, em hierarquias de mais alto nível, como por exemplo, criar novas categorias de bairro, cidade e estado a partir do atributo logradouro, ou criança, adolescente, adulto e idade a partir do atributo idade.

- *Normalização* – normalização dos dados permite atribuir uma nova escala a um atributo de forma que os valores desse atributo possam cair na nova escala em um intervalo especificado, tal como entre –1.0 a 1.0 ou de 0.0 a 1.0 etc.

- *Construção de atributos* – nesta etapa novos atributos podem ser construídos a partir dos atributos existentes, no sentido de ajudar o processo de análise. Por exemplo, pode-se gerar um novo atributo levando-se em consideração os atributos idade, peso e altura de uma pessoa ou a aplicação de uma fórmula específica.

[6] A granularidade se refere ao nível de agregação dos dados. Quando se trabalha com os fatos observados (registros ou tuplas) estamos com uma alta granularidade. Quando agregamos esses fatos, diminuí-se a granularidade.

3.4.1 Redução de Dados

Redução de dados é uma técnica que pode ser aplicada para obtenção de uma representação reduzida (*compactada*) de um conjunto de dados, muito menor em volume, mantendo a integridade do conjunto de dados original. Isto é, garimpar nesse conjunto de dados reduzido pode produzir resultados mais eficiente do que no conjunto de dados originais. Podemos aplicar as seguintes técnicas para redução de dados:

1. *Agregação de dados em cubo* – operações de agregação de dados são aplicadas para construção de cubos de dados (estrutura multidimensional para análise de dados). A figura 16 á seguir, extraída de [HK01], mostra a transformação de dados relacionais em multidimensionais. Já a figura 17, mostra uma forma de visualização e interpretação dos dados no modelo multidimensional.

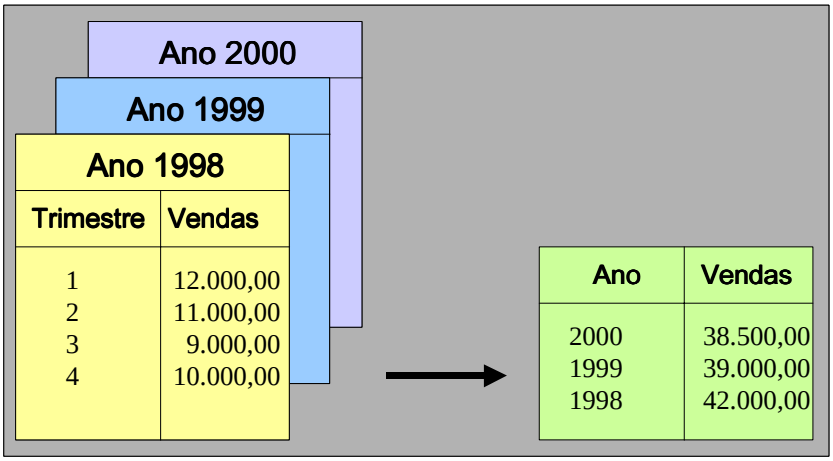

Figura 16: Agregação de dados em forma multidimensional

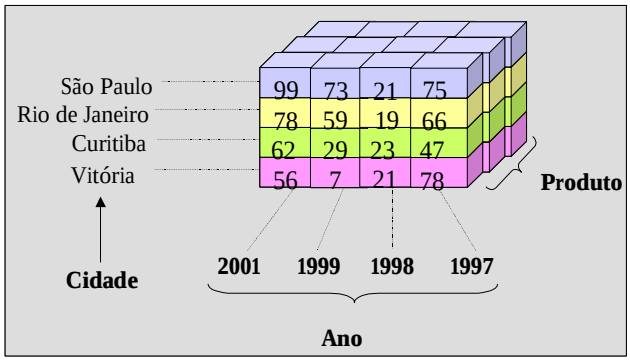

Figura 17: Representação de dados no modelo multidimensional

2. *Redução de dimensão* – atributos ou dimensões irrelevantes, fracas ou redundantes são identificadas e retiradas.

3. *Compressão de dados* – mecanismos de codificação são aplicados para reduzir o tamanho do conjunto de dados.

4. *Redução de numerosidade (numerosity reduction)* – dados são alterados ou estimados por valores alternativos, menores representações de dados tal como modelos paramétricos ou métodos não paramétricos tais como clustering, amostras e usos de histogramas.

5. *Discretização e geração de conceitos hierárquicos* – valores de atributos em linhas ou registros são alterados por intervalos ou níveis de conceitos mais elevados. Conceitos hierárquicos permitem a garimpagem de dados em vários níveis de abstração e são fortemente utilizados em ferramentas de Mineração de dados.

3.5 Mineração dos dados

A etapa de mineração de dados, conforme descrita na seção 2, envolve as etapas de identificação dos objetivos da mineração (sua funcionalidade), a identificação da melhor técnica a ser aplicada e da abordagem da aplicação de seus processos.

3.6 Avaliação dos Padrões

Nem todos os padrões obtidos no processo de mineração de dados podem ser considerados para o negócio em estudo. Nesta etapa, deverá ser feito um estudo e avaliação dos resultados, identificando claramente, quais padrões ou prognósticos podem ser utilizados, sempre baseados em sua expressividade estatística.

3.7 Apresentação e assimilação do conhecimento

Nesta etapa os resultados de todo o processo de mineração de dados deverá retornar em ações baseadas no conjunto de conhecimentos adquiridos em todo o processo. Consiste basicamente das seguintes etapas:

- Apresentar as descobertas obtidas
- Determinar a melhor forma de utilizar tais informações na tomada de decisão
- Definir as vantagens e desvantagens do projeto
- Reavaliar o projeto
- Criar novos projetos

3.8 Conclusão

Conforme descrito nesta seção o processo de *Busca de Conhecimento em Banco de Dados* (*Knowledge Discovery in Database – KDD)*, consiste de uma seqüência iterativa das etapas de limpeza, integração, seleção e transformação de dados, além da mineração dos dados, avaliação dos padrões e, apresentação e assimilação do conhecimento. Várias técnicas foram detalhadas principalmente para os quatros etapas iniciais. A qualidade dos dados que serão selecionados para mineração é de fundamental importância para o a qualidade do resultado final. Assim, quando se aplica o processo de mineração de dados a partir de um data warehouse, espera-se que a qualidade dos dados não seja mais um

problema e as etapas de limpeza, integração, seleção e transformação de dados precisam apenas de pequenos ajustes para ir de encontro a funcionalidade desejada.

A seção a seguir a apresenta a tecnologia de *mineração de dados* no contexto da inteligência de negócios (business intelligence) , no segmento de sistemas de apoio a decisão.

4 Mineração de dados no contexto da inteligência de negócios

Inteligência de Negócios ou Business Intelligence (BI) é um conjunto de conceitos e metodologias que, fazendo uso de acontecimentos (fatos) e sistemas baseados nos mesmos, apóia a tomada de decisões em negócios. Diversas tecnologias tem sido usadas conjuntamente em Inteligência de Negócios, entre elas se destacam as tecnologia de Data Warehousing (DW), de On-Line Analitical Processing (OLAP), de Análise e Exploração de Dados (AED) e de Mineração de Dados. Mineração de dados foi aclamada como uma das principais tecnologias para o futuro próximo e é considerada, atualmente, o ponto mais alto na *busca de conhecimentos* para tomada de decisões.

A figura 18 à seguir exibe as tecnologias que são utilizadas no contexto da inteligência de negócios. A mineração de dados, atualmente, representa, ao nível da informação, a principal tecnologia para tomada de decisão. Nesta área, não basta apenas trabalhar com informação, é cada vez mais importante ter conhecimento de seus negócios, conhecimentos esses que a mineração de dados proporciona.

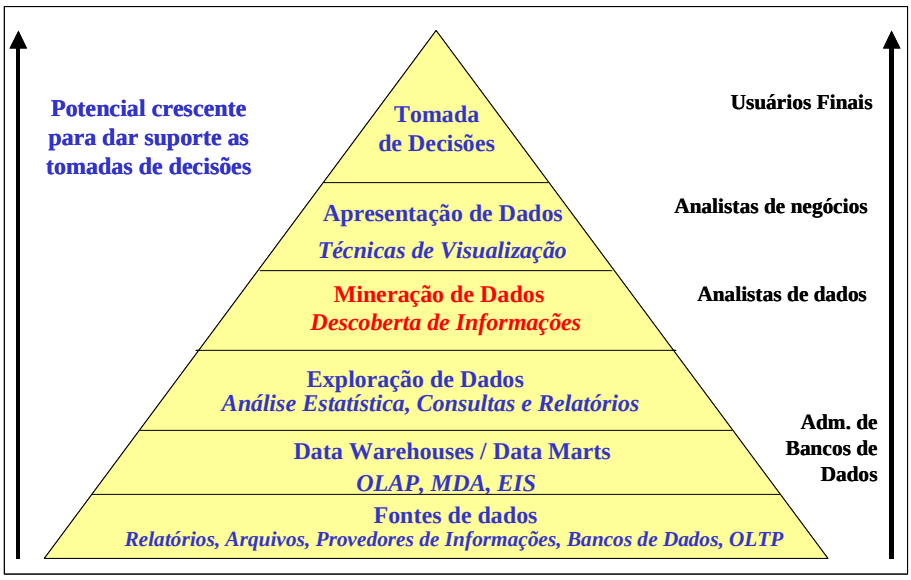

Figura 18: Mineração de dados no contexto da inteligência de negócios

As áreas de negócios das empresas, principalmente das grandes empresas, já estão iniciando a utilização de mineração de dados como busca de conhecimento. Estas soluções se apresentam, basicamente, de quatro formas:

- *Solução Direcionadas* – apresentam o poder de mineração de dados, mas são aplicadas a um problema ou indústria em particular, como por exemplo o HNC Falcon System, que é uma solução baseada em redes neurais, e direcionada especificamente a fraudes de cartões de crédito e risco de perdas e o sistema Churn Prophet, que foi criado especificamente para detecção de "agitação" de clientes (como no caso da telefonia celular).

- *Solução de negócio* - direcionadas a usuários finais de negócios com a intenção de apresentar o poder da mineração de dados de modo fácil o bastante para usar e compreensível o bastante para que os executivos consigam extrair algum valor da ferramenta, sem risco de erros causados pelo mal uso do produto.

- *Solução do analista de negócios* - direcionadas para usuários de aplicações de negócios com algum conhecimento de como a mineração de dados funciona e algumas variações diferentes. Geralmente apresentam os resultados de forma mais parecida com algoritmos de mineração de dados do que para uso final.

- *Solução de analistas de pesquisas* - Direcionadas para analistas de pesquisa ou estatísticos que desejam obter o máximo controle, bem como escolher o algoritmo. Geralmente oferecem bibliotecas de software estatístico, gráficos e visualização. São as primeiras a incluir técnicas mais modernas, recém-descobertas.

Nem sempre a mineração de dados agrega valor aos Sistemas de Apoio a Decisão - SAD. De fato, houve no passado (e ainda há, de certa forma) muitas barreiras para a mineração de dados se tornar uma função essencial dos SAD. As mais importantes têm sido superadas, mas outras ainda se mantêm. Fundamentalmente, as mais importantes foram: alto custo das soluções, a necessidade de grandes volumes de dados armazenados em poderosos servidores e a pouca amigabilidade das ferramentas de mineração de dados para pessoas que não fossem altamente especializadas. Outras que podem ser citadas são o desafio de preparar os dados para mineração, as dificuldades em se obter uma análise custo/benefício bem fundamentada antes do início do projeto e a preocupação quanto à viabilidade de fornecedores dessas ferramentas.

A seção a seguir apresentará algumas áreas de negócios e pesquisas que possuem um grande potencial para mineração de dados.

5 Aplicações potenciais em mineração de dados

As tecnologias de mineração de dados podem ser aplicadas a uma grande variedade de contextos de tomada de decisão no ramo dos negócios. Em particular, áreas que envolvem contrapartidas (retornos) significativas supostamente incluem o seguinte:

- Marketing – As aplicações incluem a análise do comportamento do consumidor com base em padrões de compra; a determinação de estratégias de marketing incluindo propaganda, localização de lojas e mala direta; a segmentação de clientes, lojas ou produtos; bem como o projeto de catálogos, o layout de lojas e campanhas publicitárias.

- Finanças – As aplicações incluem a análise da avaliação para concessão de crédito a clientes, segmentação de contas a receber, análise de desempenho de investimentos financeiros como ações, bônus e fundos mútuos; avaliação de opções financeiras e detecção de fraudes.

- Manufatura (Indústria) – As aplicações envolvem a otimização de recursos como equipamentos, força de trabalho e matéria-prima; o projeto ótimo de processos de produção, layouts de lojas e projetos de produtos, como por exemplo para automóveis, com base em exigências dos clientes.

- Saúde – As aplicações incluem a análise da eficácia de certos tratamentos; a otimização de processos dentro de um hospital, o relacionamento de dados sobre o estado de saúde do paciente com a qualificação médica; e a análise de efeitos colaterais de drogas.

- Área biomédica – Diversos aparelhos de diagnósticos estão sendo desenvolvidos segundo os padrões encontrados em populações observadas ao longo de vários anos. Seu objetivo principal é detectar e identificar, principalmente, grupos de riscos para os pacientes e trabalhar na prevenção de possíveis doenças.

- Outras aplicações– Diversas áreas do conhecimento começam a utilizar as técnicas de mineração de dados, visando conhecer e identificar padrões até então desconhecidos. Entre essas áreas destacam-se as áreas de seguros, bancos, comunicações, exploração de petróleo etc.

6 Comentários Finais

Neste trabalho apresentamos uma visão geral das funcionalidades e técnicas de mineração de dados. Neste contexto são apresentados os principais conceitos, uma forma bastante prática de se identificar às funcionalidades, ou seja, os resultados que se deseja obter com a mineração de dados, bem como uma identificação das técnicas que podem ser utilizadas para cada funcionalidade. Mostramos também, que uma mesma técnica pode ser empregada para se obter resultados em diferentes funcionalidades. Descrevemos duas abordagens de como um processo de mineração de dados pode ser conduzido, no sentido de obter melhores resultados.

Apresentamos o processo de *Busca de Conhecimento em Banco de Dados* (*Knowledge Discovery in Database – KDD)*, suas etapas e descrevemos detalhadamente cada etapa do processo, desde a preparação dos dados até a aplicação dos processos da mineração de dados e como apresentar e internalizar os resultados obtidos.

Finalmente apresentamos a mineração de dados no contexto da Inteligência de Negócios (Business Intelligence – BI), suas principais soluções e aplicações potenciais.

Deixamos claro que este trabalho não esgota o assunto, sendo apenas uma introdução a mineração de dados.

7 Referências Bibliográficas

[AZ96] Pieter Adriaans, Dolf Zantinge; “Data Mining”; Addison-Wesley, 1996

[BL97] Michael J. A. Berry; Gordon Linoff, “Data Mining Techiques for Marketing, Sales, and Customer Support”; John Wiley & Sons, Inc., 1997.

[BLC00] Antônio de Pádua Braga, Teresa Bernarda Ludermir, André Carlos P. de L. F. Carvalho; “Redes Neurais Arificiais – Teoria e Aplicações”, Editora LTC, 2000

[BT99] Bhavani Thuraisingham; “Data Mining”; CRC Press, 1999

[DN00] Carlos Alberto R. Diniz, Francisco Louzada Neto; “Data Mining: Uma Introdução”; 14ª Sinape – Caxambu – ABE – Associação Brasileira de Estatística, 2000

[EN99] Ramez Elmasri e Shamkant Navathe; “Fundamentals of Database Systems”; Addison-Wesley, 1999, 3rd Edition.

[HK01] Jiawei Han, Micheline Kamber; “Data Mining – Concepts and Techniques”; Morgan Kaufmann Publishers, Inc, 2001

[Men99] Jesus Mena; “Data Mining Your Website”; Digital Press, 1999

[WB98] Christopher Westphal, Teresa Blaxton, “Data Mining Solutions – Methodos and Tools for Solving real-Word Problems”; John Wiley & Sons, Inc., 1998.

[WI99] Sholom M. Weis, Nitim Indurkhya; “Predict Data Mining”; Morgan Kaufmann Publishers, Inc, 1999

[MPO01] Ilza Maria B. Mendes, Alexandre Plastino e Luiz Satoru Ochi, “Regras de Associação: suas Diferentes Formas e seus Algoritmos de Mineração”, Mini-curso apresentado no SBBD 2001.

[Hay99] Simon Haykin, “Redes Neurais - Princípios e Prática”, tradução da segunda edição, Editora Bookman, 1999.