2
Mineração de Dados

A mineração de dados, ou *data mining*, é uma das principais etapas do processo de busca de conhecimento. Este conceito é utilizado para identificar técnicas avançadas de análise dos dados, que podem ou não utilizar a inteligência artificial, refinando os dados em busca de padrões nos dados através de um modelo do mundo real. Essas técnicas são apenas uma parte do processo de busca de conhecimento dentro de um banco de dados, em que o objetivo maior é obter regras e padrões que se aplicam aos dados.

Normalmente, as técnicas de mineração de dados são aplicadas em dados armazenados em uma *data warehouse* (DW) ou em um *data mart*, mas também é possível aplicá-las em dados operacionais. Os resultados da mineração de dados podem ser usados em tomadas de decisão, em gerenciamento de informações, controle de processo, dentre outros. Sua aplicação pode ser em um processo de verificação onde o usuário tenta provar sua hipótese acerca da relação entre os dados ou como um processo de descoberta de padrões, fazendo uso de técnicas como redes neurais, algoritmos genéticos, regras de associação, árvores de decisão, regressão, entre outros.

2.1.
Processo de Busca de Conhecimento (KDD)

O avanço da tecnologia tanto no âmbito do hardware quanto do software permitiu que a capacidade de armazenamento e processamento de dados crescesse em velocidade muito grande. A análise manual ou semi-automática de grandes volumes de dados tornou-se impraticável, prejudicando a tomada de decisão. Nesse sentido, a aplicação de métodos que facilitem este processo se faz muito importante nos dias de hoje.

O processo de busca de conhecimento em banco de dados também é conhecido como KDD – Knowledge Discovery and Data Mining, cujo objetivo segundo Fayyad et. al [4] *"(...) é a extração de conhecimento de alto nível a partir de dados de baixo nível disponíveis em grandes bancos de dados (...) processo não trivial de identificação, a partir de dados, de padrões que sejam válidos, novos, potencialmente úteis e compreensíveis (...)".* Em outras palavras, o KDD é o processo de extração de conhecimento de grandes bases de dados, convencionais ou não.

Fayyad et. al [4] divide o processo de KDD em seis passos:

i. Preparação dos Dados: consiste em incluir o conhecimento relevante para a aplicação além de definir quais as metas que o processo precisa atingir.

ii. Limpeza dos Dados: consiste em retirar os dados que possam distorcer a análise. Assim, utiliza estratégias para remover ruídos, tratar atributos perdidos e até mesmo métodos de transformação para diminuir o número de variáveis envolvidas no processo, visando com isto melhorar o desempenho do algoritmo de análise.

iii. Seleção de Dados: consiste em escolher sobre qual conjunto ou subconjunto de dados em que o processo será aplicado.

iv. *Data Mining*: consiste em decidir qual tarefa de *data mining* será aplicada para atingir os objetivos do processo e qual a melhor técnica a ser utilizada (ver seção 3).

v. Incorporação do conhecimento anterior: consiste em interpretar o modelo descoberto a fim de verificar sua acuracidade em busca de melhorias, possibilitando o retorno para qualquer etapa anterior do processo, retirando padrões redundantes ou irrelevantes.

vi. Interpretação dos resultados: neste ponto o resultado obtido é incorporado ao sistema possibilitando a tomada de ações baseadas no conhecimento ou documentando-os e relatando-o às partes interessadas.

Em particular, o passo da mineração de dados utiliza técnicas de inteligência artificial que procuram relações de similaridade ou discordância entre dados com o objetivo de encontrar, automaticamente, padrões, anomalias e regras, focando em transformar dados, aparentemente ocultos, em informações úteis para a tomada de decisão ou avaliação de resultados.

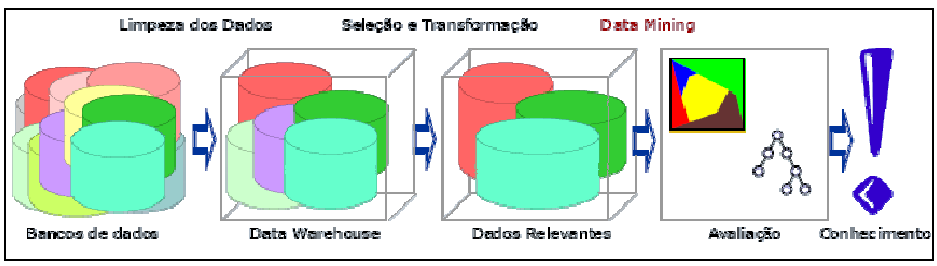

Figura 1 Etapas para busca de conhecimento

2.2. Tipos de Predição

Por tipos de predição entende-se os tipos de relacionamento que estabelecemos entre os dados para a obtenção de conhecimento. Sabendo a que resultado se deseja chegar, é fácil identificar a tarefa de mineração que mais auxiliará o processo de busca de uma solução para o problema. Para atender aos objetivos e gerar resultados esperados, há uma coleção de técnicas que podem ser utilizadas; cada técnica possui ainda uma gama de algoritmos que irão, efetivamente, manipular os dados. A Figura 2 a seguir, extraída de [20], ilustra a iteratividade entre funcionalidades, técnicas e algoritmos de mineração de dados.

O resto desta seção discute tipos de predição de dados.

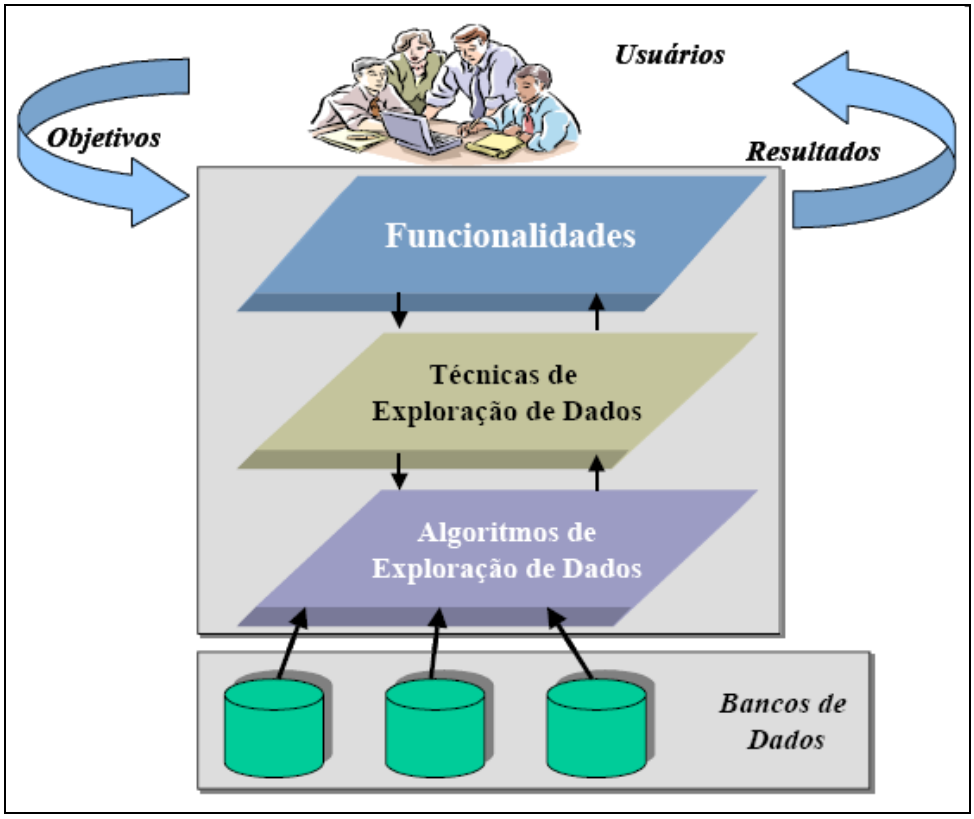

Figura 2 Iteratividade entre Funcionalidades, Técnicas e Algoritmos de Mineração de Dados

2.2.1. Classificação

Categorizar os clientes de acordo com seu perfil de compras é um exemplo de tarefa de classificação. Um modelo de classificação é criado e os atributos dos registros (no caso do exemplo acima, clientes) são analisados de acordo com as condições e

características das classes pré-determinadas pelo modelo. Caso esses atributos casem, o registro é então classificado na classe em que melhor se encaixa.

2.2.2. Segmentação

Essa tarefa é mais conhecida como *clustering* e se diferencia da classificação porque não pressupõe a existência de classes pré-definidas. Os registros são agrupados de acordo com a semelhança nos seus atributos, segmentando-os em *clusters* (subgrupos homogêneos) baseando-se no princípio de que os registros classificados em um grupo devem ser homogêneos entre si, e que os grupos devem ser heterogêneos entre si.

2.2.3. Regressão

Estimar um acontecimento ou um padrão não é uma tarefa simples. Esse é o objetivo da regressão: prever padrões para uma variável de valor contínuo, com base nos outros atributos disponíveis e nas outras ocorrências disponíveis para análise.

2.2.4. Associação

Através da associação é possível identificar transações que sempre ocorrem em conjunto. Na literatura também é conhecida como grupos de afinidade ou análise de cesta. O exemplo clássico da literatura é o MBA (*Market Basket Analysis*), que busca encontrar padrões nos produtos em um carrinho de compras através de regras de associação. Esse estudo auxilia as lojas a encontrar a melhor maneira de organizar seus produtos, de modo que a disposição deles nas prateleiras estimule as compras dos clientes.

2.3. Técnicas de Mineração de Dados

As técnicas de mineração de dados são os fundamentos computacionais que possibilitam a construção dos algoritmos que realizarão a busca por padrões nos dados. Diversas técnicas podem ser utilizadas para atender a uma tarefa de mineração de dados. Entretanto, cada técnica possui características específicas e é necessário ter o conhecimento do funcionamento e do objetivo das mesmas para interpretar os resultados obtidos.

No resto desta seção listamos algumas técnicas de mineração de dados.

2.3.1. Algoritmo Genético

Simulando o processo natural da evolução, os algoritmos genéticos (AG's) têm por objetivo realizar a busca e a otimização da descoberta de padrões. Diferentemente dos métodos convencionais de mesmo objetivo, os AG's trabalham simultaneamente em conjuntos de soluções diferentes, realizando pesquisas adaptativas nos dados [20], modelando uma solução para um problema específico em estruturas de dados que são semelhantes a um cromossomo. Operadores são aplicados recombinando essas estruturas, gerando assim novas combinações de regras de associação.

Essa técnica é utilizada na classificação e na segmentação de dados, formulando hipóteses sobre a dependência dos atributos dos dados; com operadores de mutação e cruzamento desenvolvem várias mutações para a solução do problema. Ao longo do tempo, o algoritmo tende a "aprender" e a se aperfeiçoar, de maneira que somente as soluções com maior poder de acerto na previsão são aceitas.

2.3.2. Redes Neurais

A técnica de Redes Neurais é bastante utilizada em tarefas de classificação, regressão e segmentação. Os dados são trabalhados com base no funcionamento do cérebro humano, aprendendo a tomar decisões baseadas nas experiências anteriores – nas instâncias anteriores dos dados. Os neurônios do cérebro são representados por nodos que estão conectados em outros nodos por sinapses, formando uma rede de processamento. Os valores das entradas são multiplicados nos neurônios pelos pesos de suas sinapses, conforme vão caminhando na rede. Ao final, temos a classificação ou a previsão da entrada.

2.3.3. Árvores de Decisão

As árvores de decisão têm como objetivo principal dividir as instâncias em classes. Cada nó da árvore testa o domínio de uma variável da entrada e o redireciona para o nó seguinte. Cada sub-árvore representa o resultado de um teste e a folha é a classificação que aquele registro recebeu. Ao final, cada nó terminal terá os registros da entrada que se adéquam às regras regidas por esse nó, representando assim, uma classe.

2.3.4. Regras de Associação

Basicamente, as regras de associação são definidas por uma correlação estatística entre alguns atributos da entrada com o objetivo de descobrir relações que

ocorrem em comum dentro de um conjunto de dados. Cada registro é visto como uma transação e cada variável como um item dessa transação, deixando subentendido que a presença de um item implica necessariamente na presença de outro na mesma transação. Esse conceito será mais detalhado na Seção 3.1

2.3.5. Análise de Vizinhança

Através de uma função definida para determinar a "distância" entre duas instâncias, ou seja, de uma função para identificar um conjunto de registros que estão próximos por determinada característica, essa técnica é empregada em análise de prognósticos e não para descoberta de conhecimento. Não é muito explorada na literatura.

2.4. Aplicações

Além da aplicação no campo dos profissionais de marketing, auxiliando na busca de padrões para melhorar o processo de recomendação e alavancar as vendas, as técnicas de *data mining* podem ser utilizadas em: (i) redes de telecomunicações, para evitar fraudes em ligações pré- ou pós-pagas, detecção de falhas, dimensionamento de sistemas; (ii) saneamento básico, detectando fraudes em ligações de água; (iii) monitoramento ambiental, para prevenção de desequilíbrios; (iv) na indústria, prevendo a demanda, planejando a produção; (v) na educação, auxiliando na identificação da evasão escolar; (vi) na medicina, atuando no diagnóstico e prevenção de doenças, fraudes de plano em planos de saúde; (vii) no comércio, definindo o perfil do consumidor, segmentando o mercado, sugestão de produtos, entre vários outros campos de atuação; (viii) no âmbito financeiro, ajudando no combate as fraudes de cartão de crédito, análise de investimentos e de crédito.