Aplicações das Técnicas de Mineração de Dados

Jeferson José de Lima jefersonjl82@gmail.com
Profª.Drª Simone Nasser de Matos snasser@utfpr.edu.br
Prof.Dr. João Luiz Kovaleski kovaleski@utfpr.edu.br

Resumo:

A busca pelo conhecimento faz parte da rotina corporativa, porém devido à quantidade e complexibilidade das informações, torna-se cada vez mais inviável para capacidade humana a análise e interpretação destes dados. Uma alternativa são as ferramentas de mineração de dados. Este artigo faz uma pequena abordagem sobre alguns estudos em aplicações da extração de conhecimento em banco de dados nas mais variadas áreas de atuação.

Palavras chave: Mineração de Dados, Descoberta de Conhecimento em Banco de Dados, Banco de dados.

Applications of Data Mining Techniques

Abstract
The search for knowledge is a routine part enterprise, but because of the amount and complexity of information, it becomes increasingly impossible for the human capacity analysis and interpretation of these data. An alternative are the tools of data mining. This article is an approach to some small studies on applications of knowledge mining in databases in various fields.

Key-words: Data Mining, Knowledge Discovery in Databases, Database.

1 Introdução

Nos últimos anos, pode se observar um aumento expressivo na quantidade de informações armazenadas em bancos de dados, motivada em grande parte pelos avanços tecnológicos nos processos de armazenamento digital de informações.

Dados científicos em projetos de pesquisa, tais como missões espaciais da NASA e o Projeto do Genoma Humano, têm alcançado proporções gigantescas. Empresas como FedEx, Wal-Mart, UPS, Banco do Brasil, Caixa Econômica Federal e Sendas possuem base de dados da ordem de centenas de *terabytes* de informação (GOLDSCHMIDT & PASSOS, 2005).

Isso leva as instituições a uma visão cada vez mais preocupada com o valor da informação no processo de tomada de decisão, direcionando a área de Descoberta de Conhecimento em Banco de Dados (KDD - *Knowledge Discovery in Databases* em inglês) a um novo contexto, para proporcionar um melhor entendimento dos problemas.

Considerando a ampla aplicação das técnicas de KDD em processos subjetivos de manipulação de dados este presente artigo tem o objetivo mostrar as diversas aplicações onde se conseguiu transcrever a subjetividade para uma manipulação objetiva dos dados.

2 Referencial Teórico

2.1 Descoberta do Conhecimento em Base de Dados

A evolução da tecnologia da informação tornou possível o armazenamento de uma grande e variável base de dados, porém a funcionalidade destas informações está a quem do que as empresas têm utilizado. Nos métodos convencionais a análise de grande quantidade de informações torna-se inviável pelo homem sem o auxílio de ferramentas computacionais aprimoradas.

Este desafio começa a ser desvendado a partir dos anos de 1980 onde grandes descobertas foram feitas para tornar possível migrar estes dados para o campo do conhecimento, por meio de sistemas de informação, com a finalidade de analisar e organizar a informação para melhorar a decisão em processo críticos.

Para REZENDE (2005), o processo de gerar conhecimento resulta de um processo no qual a informação é comparada a outra e combinada em muitas ligações (hiperconexões) úteis e com significado. Isso implica que o conhecimento é dependente de nossos valores e nossa experiência e sujeito às leis universalmente aceitas.

A partir desta necessidade surge uma nova área denominada Descoberta do Conhecimento em Base de Dados, que tem o objetivo de encontrar conhecimento a partir de um conjunto de dados para serem utilizados num processo de tomada de decisão. O processo de KDD é composto por várias etapas operacionais. A figura 1 apresenta como estes processos são interligados, sendo a etapa de pré-processamento responsável às funções de captação, organização e tratamento dos dados. Posteriormente a mineração de dados realiza a busca efetiva por conhecimentos úteis no contexto de aplicação de KDD. Por fim, o pos-processamento abrange o tratamento do conhecimento obtido na mineração dos dados.

Figura 1 - Processo de KDD
Fonte: GOLDSCHMIDT & PASSOS, 2005

2.1.1 Pre-Processamento

Normalmente, os dados disponíveis para análise não estão em um formato adequado para a Extração do Conhecimento. Além disso, em razão de limitações de memória ou tempo de processamento, muitas vezes não é possível a aplicação direta dos algoritmos de extração de padrões dos dados, desta maneira, torna-se necessária a aplicação de métodos para tratamento, limpeza e redução do volume de dados. É importante salientar que a execução de transformação deve ser guiada pelos objetivos do processo de extração a fim de que o conjunto de dados gerado apresente as características necessárias para que o s objetos sejam cumpridos (REZENDE, 2005).

As etapas executadas no pré-processamento são:

- Extração e Integração. A utilização de apenas uma base de dados nem sempre é uma realidade, para isso é necessário a unificação formando uma única fonte de dados;
- Transformação. Após a junção dos dados, faz-se essencial a adequação a um padrão de representação homogêneo aos dados de mesma categoria para que possam ser interpretação pelos algoritmos;
- Limpeza. Devido a falhas de digitação, caracteres inválidos entre outros erros tornam a limpeza destes dados algo imprescindível;
- Seleção e Redução de Dados. Está etapa é feita de três maneiras, redução de número de exemplos, redução do número de atributos e redução do número de valores de atributos.

2.1.2 Mineração de dados

A execução da etapa de Mineração de Dados compreende a aplicação de algoritmos sobre dados procurando abstrair conhecimento. Estes algoritmos são fundamentados em técnicas que procuram, segundo determinados paradigmas, explorar os dados de forma a produzir modelos de conhecimento. Para o autor, todo conhecimento abstraído ao longo do processo de KDD será interpretado e referenciado pela expressão modelo de conhecimento. A forma de representação do conhecimento depende diretamente do algoritmo de Mineração de Dados utilizado (GOLDSCHMIDT & PASSOS, 2005).

A etapa de extração de padrões é direcionada ao cumprimento dos objetivos definidos quando feita a identificação do problema. A Extração de padrões compreende a escolha da tarefa de

mineração de dados a ser utilizada, que é feita de acordo com os objetivos desejáveis para a solução, conforme figura 2.

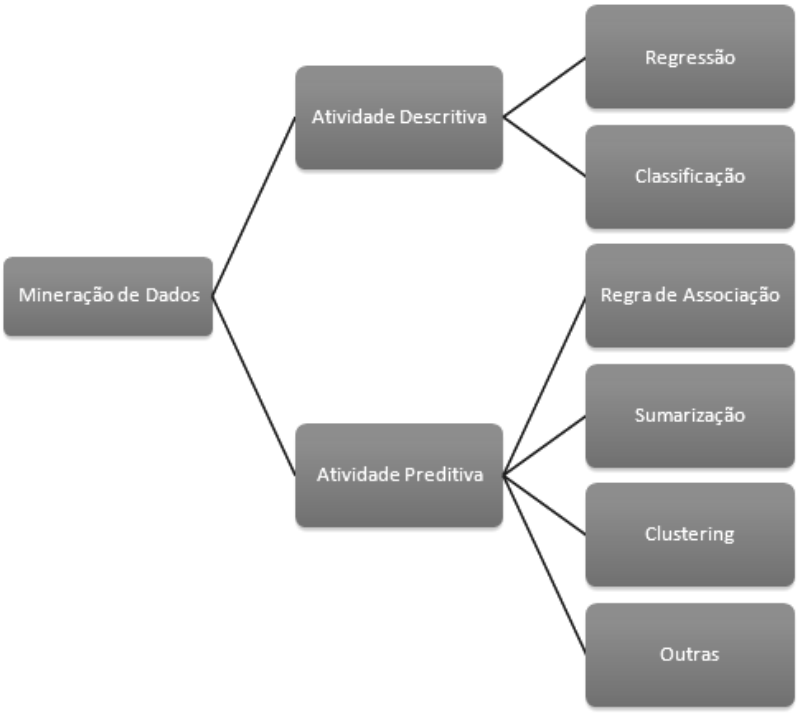

Figura 2 - Tarefas de Mineração de Dados
Fonte: GOLDSCHMIDT & PASSOS, 2005

Na atividade de predição acontece a generalização de exemplos ou experiências passadas com resposta conhecida. A tarefa de classificação consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de rótulos categóricos predefinidos, denominados classes. Já na regressão o atributo a ser predito consiste em um valor contínuo. Esta tarefa é similar à tarefa de classificação, sendo restrita apenas a atributos numéricos.

Atividade descritiva consiste na identificação de comportamentos intrínsecos do conjunto de dados, sendo que estes dados não possuem uma classe especificada, como no caso da regra de associação onde a busca acontece por itens que frequentemente ocorram de forma simultânea em transações do banco de dados. A tarefa de sumarização é muito comum em KDD, pois consiste na procura e identificação de características comuns entre conjuntos de dados. Já a tarefa de *clustering* é utilizada para separar os registros de uma base de dados em subconjuntos ou clusters, de tal forma que os elementos de um cluster compartilhem de propriedades comuns que os distingam de elementos em outro cluster.

Para cada tarefa a ser empregada, existe uma variedade de algoritmos para executá-la. A escolha do algoritmo é feita de forma subordinada a linguagem de representação dos padrões a serem encontrados. Entre os tipos mais frequentes de representar padrões está a árvore de decisão, regra de produção, modelos lineares, modelos não-lineares (Redes Neurais Artificiais), modelo baseado em exemplos e modelos de dependência probabilística.

2.1.3 Pos-Processamento

A resposta obtida pelo processo de mineração de dados precisa passar por alguns critérios de avaliação do especialista em KDD, caso o resultado não seja o desejado, o processo pode ser

inicializado inúmeras vezes até que se tenha o objetivo alcançado. Um dos fatos que exige a necessidade de novos processamentos é a geração de muitos padrões onde somente o necessário deve estar associado. Diversas são as medidas para avaliar a assertividade do modelo como:

- Medida de Desempenho. É utilizada para avaliar a informação gerada como precisão, erro, confiança negativa, sensitividade, especificidade, cobertura, suporte, satisfação, velocidade e tempo de aprendizado;
- Medida de Qualidade. São medidas para avaliar o fator de compreensibilidade e interessabilidade;
- Monitoração Direta. Monitora os resultados dos modelos diretamente;
- Monitoração Indireta. Mede as ações de negócio executadas com base nos resultados dos modelos.

3 Aplicações de Mineração de dados

Segue os exemplos de aplicação do processo de mineração de dados:

3.1 Franquia de *Fast-Food*

Para estabelecer uma associação entre produtos que possam ser vendidos combinados, muitas franquias utilizam as informações sobre o número de transações de venda de itens realizadas durante um determinado período observado em dias normais de venda. Através da regra de associação pode se definir quais produtos podem ser vendidos em forma de "pacote".

3.2 Ação Social

O projeto PRODERJ (Órgão de Tecnologia da Informação do Estado do Rio de Janeiro) promoveu um estudo sobre o processo de reintegração de pessoas de rua ao estado através de um banco de dados das informações de cada indivíduo. Após o processamento destas informações foi possível associar cada perfil a um programa de reintegração.

3.3 Educação

Em 2001 todas as escolas do Rio de Janeiro preencheram um questionário com mais de 600 perguntas referentes à sua gestão. Com este banco de dados foi possível descobrir varias questões como por que determinada escola tem uma maior procura que outra, os motivos do alto índice de evasão entre outras para que o governo estabelecer medidas efetivas sobre alguns dos problemas detectados.

3.4 Área Médica

A partir do SIM (Sistema de Informações de Mortalidade) implantada no Brasil em 1975, o SINASC (Sistema de Informações sobre Nascidos Vivos), criado em 1990, e o SIMI (Sistema de Investigação da Mortalidade Infantil) de 2000, VIANNA *et alii* (2010) agrupa toda esta base de dados no sistema de mineração de dados para descobrir diversas regras sobre a mortalidade infantil, como exemplo a regra da mãe adolescente (menor de 16 anos) mesmo a criança apresentando bom peso ao nascer (2.500 a 3.500kg), a mãe surge como alto critério de risco para óbito perinatal.

3.5 Manutenção

TRONCHONI *et alii* (2010) traz um estudo sobre descoberta do conhecimento em eventos de desligamento em empresas de distribuição. No trabalho são coletados 570.000 eventos de

intervenções da manutenção em redes de distribuição buscando a correta identificação das causas de desligamento forçado. Por fim, é apresentada uma tabela com os variados eventos e suas prováveis causas como apoio aos operadores na tomada de decisão para alocação de equipes e recursos na área de concessão da empresa.

4 Conclusão

Através deste artigo pretendeu-se verificar, numa breve abordagem sobre casos, que o processo de mineração de dados torna possível quantificar afirmações subjetivas de conhecimento humano ou até mesmo a descoberta de novas regras e associações que por meio da análise de um indivíduo torna-se quase impossível.

Mesmo que ainda as ferramentas comerciais de mineração de dados façam parte de um processo caro, por outro lado diversos aplicativos gratuitos tornam-se viável para pequenas e médias empresas.

Referências

GOLDSCHIMIDT, R e PASSOS, E. ***Data mining*: Um guia prático**. Rio de Janeiro: Campus, 2005.

PIMENTA, A., VALENTIM, P., SANTOS, D., NETO, M., "**WEKA-G: mineração de dados paralela em grades computacionais**", Revista de Sistemas de Informação da FSMA n. 4 (2009) pp. 2-9.

REZENDE, Solange Oliveira. **Sistemas Inteligentes: Fundamentos e Aplicações**. São Paulo: Barueri, 2003.

TRONCHONI, Alex B.; PRETTO, Carlos O.; ROSA, Mauro A. da e LEMOS, Flávio A. Becon. **Descoberta de conhecimento em base de dados de eventos de desligamentos de empresas de distribuição**. Sba Controle & Automação [online]. 2010, vol.21, n.2 ISSN 0103-1759

VIANNA, Rossana Cristina Xavier Ferreira; MORO, Claudia Maria Cabral de Barra; MOYSÉS, Samuel Jorge; CARVALHO, Deborah; NIEVOLA, Julio Cesar. **Mineração de dados e características da mortalidade infantil**. Cadernos de Saúde Pública (ENSP. Impresso), v. 26, p. 535-542, 2010.