DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

Claudio Napolis Costa [1], Jonatas Vieira Coutinho[2], Lúcia Helena de Magalhães[3], Márcio Aarestrup Arbex[4]

RESUMO

Vários métodos de aprendizado supervisionado ou não supervisionado têm como objetivo analisar uma base de dados em que se pode ou não haver uma classificação prévia que se torne objetivo da análise. Em muitos casos, não há um conhecimento das classes às quais os dados podem pertencer, e, nem mesmo, separá-los em classes resultará em informação importante. Existem análises que atendem a diversas demandas, por exemplo, aplicando métodos que fazem com que um conjunto de dados seja subdividido em grupos cujos elementos possuam alguma similaridade, portanto, neste tipo de análise, os próprios dados ditam as classes às quais pertencem, resultando em informação nova e potencialmente útil. O objetivo deste artigo é abordar os principais conceitos para descoberta de conhecimento, bem como os tipos de aprendizado de máquina.
PALAVRAS-CHAVE: Descoberta de Conhecimento. Base de Dados. Informação. Tomada de Decisão. Mineração de Dados.

ABSTRACT

Several methods of supervised and unsupervised learning are designed to analyze data set. The analysis specify a relationship between explanatory variables and dependent variable. But, in many cases, there is no knowledge of classes which data may belong. Then separate into classes will not result in important information. Tests meet various demands, for example, using data division method into groups whose elements have some similarity. Supervised learning requires that the target variable is well defined and that a sufficient number of its values are given. For unsupervised learning typically either the target variable is unknown or has only been recorded for too small a number of cases. This paper address the main concepts for Knowledge Discovery in Database and the types of machine learning.
KEY WORDS: Discovery of Knowledge. Database. Information. Decision Making. Data Mining

---

[1] Doutor em Animal Breeding pela Cornell University - Estados Unidos, mestre em Zootecnia pela Universidade Federal de Viçosa. cnc55@terra.com.br

[2] Graduado em Sistema de Informação pela Faculdade Doctum Cataguases. jonnyleg@gmail.com

[3] Pós Graduada Desenvolvimento de Aplicações para Web pelo Centro de Ensino Superior de Juiz de Fora, Pós Graduada em Matemática e Estatística pela Universidade Federal de Lavras, mestre em Sistemas Computacionais – Computação de Auto-Desempenho pela Universidade Federal do Rio de Janeiro. lhm@powermail.com.br.

[4] Pós Graduado em Ciência da Computação pela Universidade Federal de Viçosa, mestrando em Sistemas Computacionais – Computação de Auto-Desempenho pela Universidade Federal do Rio de Janeiro. marcio.arbex@gmail.com

INTRODUÇÃO

Com a globalização e a alta competitividade, a informação tornou-se um dos bens mais valiosos para uma organização. Ter acesso à informação precisa de maneira rápida e eficiente é um dos grandes diferenciais que podem levar ao sucesso. Neste aspecto, a Tecnologia da Informação tem evoluído, proporcionando aos tomadores de decisão uma infra-estrutura e ferramentas que aliadas às metodologias de coleta, organização, processamento e utilização da informação, se tornam um apoio vital para a sobrevivência das organizações.

Organizações são constituídas de vários processos e mudanças em seu ambiente em função das demandas do mercado. Ter o controle desses processos e mudanças é uma tarefa difícil, pela vasta coleção de dados e a grande necessidade de métodos eficazes para organizá-los visando uma tomada de decisão bem fundamentada e planejada. O uso da Tecnologia da Informação iniciou-se nos anos 60, e cada vez mais tem se tornado essencial para a sobrevivência de empresas no mercado atual, pois disponibiliza meios para coleta, armazenamento e tratamento dos dados vindos de todos os seus processos. Para dar o devido apoio à tomada de decisão, existem softwares e ferramentas que aliados a uma metodologia bem definida, dão todo o apoio ao processo decisório.

Para JESUS et. al (2004), a decisão estratégica deve ser um processo bem definido e contínuo na empresa, métodos objetivos e bem estruturados para a tomada de decisão devem fazer parte do dia a dia da empresa, que através das ferramentas de T.I. (Tecnologia da Informação), podem não só armazenar uma grande quantidade de dados, também filtrando, traduzindo e consolidando-os para obter, como produto final deste processo, a informação que é necessária para que a tomada de decisão deixe de ser subjetiva para tornar-se um processo sólido e fundamentado em informações precisas.

Dentre estas ferramentas e metodologias, destaca-se o *Data Mining*, ou Mineração de Dados, que é um processo, como o próprio nome sugere, dedicado a um trabalho de extração de dados mais elaborado e minucioso, buscando padrões não evidentes em uma simples pesquisa em uma base de dados. Com a enorme quantidade de dados que se tem hoje em dia através dos recursos computacionais, recuperá-los de maneiras convencionais, como por exemplo, através de consultas SQL (*Structured Query Language*) feitas diretamente nas bases de dados, pode não extrair todas as informações que essa grande massa de dados pode

proporcionar. "...apenas recuperar informação não propicia todas as vantagens possíveis. O processo de Data Mining permite que se investigue esses dados à procura de padrões que tenham *valor* para a empresa" NAVEGA, (2002: 1).

Para ELMASRI, NAVATHE (2005: 626)

> O conhecimento é classificado em indutivo e dedutivo. O **conhecimento dedutivo** deduz novas informações baseadas na aplicação de regras lógicas *predefinidas* de dedução sobre dados existentes. O Data Mining apóia o **conhecimento indutivo**, que descobre novas regras e padrões nos dados fornecidos.

AMO (2003: 1) afirma que

> Mineração de Dados é uma área de pesquisa multidisciplinar, incluindo tecnologia de bancos de dados, inteligência artificial, aprendizado de máquina, redes neurais, estatística, reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados.

"Data Mining é a concepção de modelos computacionais capazes de identificar e revelar padrões desconhecidos, mas existentes entre dados pertencentes a uma ou mais bases de dados distintas". THOMÉ (2002: 13). "Data Mining se refere à mineração ou a descoberta de novas informações em função de padrões ou regras em grandes quantidades de dados" ELMASRI, NAVATHE (2005: 620).

1 DATA WAREHOUSE

O *Data Mining* se aplica a grandes bases de dados, e é mais eficaz se aliada ao *Data Warehouse*, ou Armazém de dados, que consiste em consolidar e agregar os dados vindos, na maioria das vezes, dos bancos de dados transacionais. ELMASRI, NAVATHE (2005: 620) afirmam que

> A proposta de um Data Warehouse é sustentar a tomada de decisão com dados. O Data Mining pode ser usada em conjunto com o Data Warehouse para auxiliar certos tipos de decisão. Data Mining pode ser aplicada a bancos

de dados operacionais com transações individuais. Para fazer data Mining mais eficiente, o Data Warehouse deve ter uma coleção de dados agregados ou sumarizados.

Segundo INMON (1997), *Data Warehouse* é "uma coleção de dados integrados, orientados por assunto, variáveis com o tempo e não voláteis, usados para dar suporte ao processo gerencial de tomada de decisão".

O *Data Warehouse* trata-se de uma coleção de dados derivados de bancos de dados operacionais heterogêneos, com o objetivo de dar suporte à tomada de decisão, apresentando-os de forma analítica, detalhados ou resumidos. Não é uma base de dados transacional, os acessos são somente para carga, através de aplicativos extratores, no processo chamado de ETL (*Extration, Transformation and Load*), e consultas por parte da equipe que gerenciará a informação.

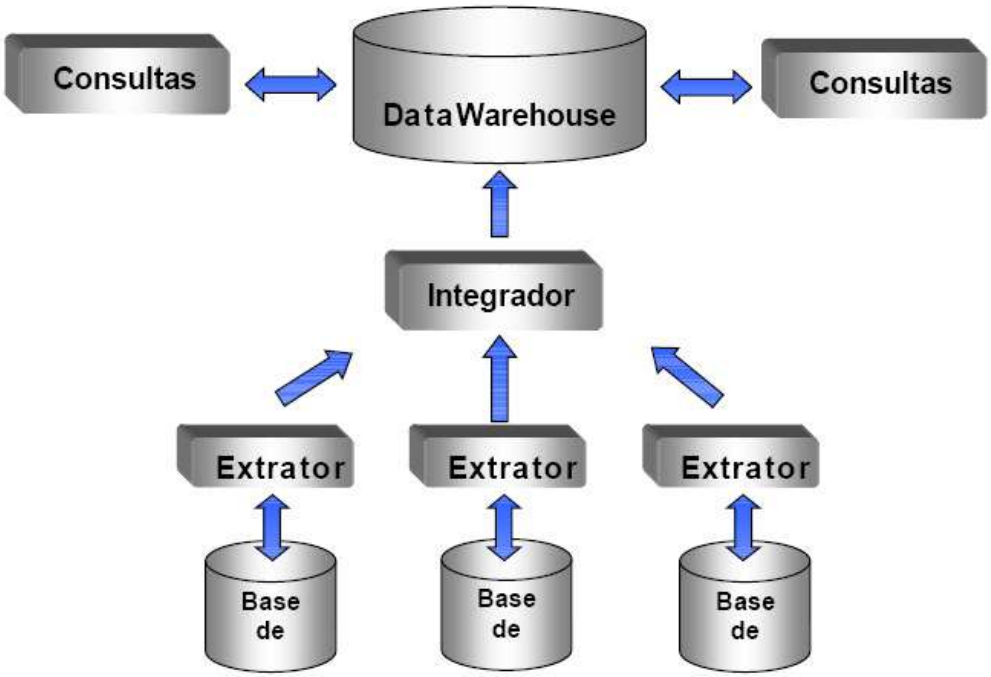

Figura 1: Esquema de emprego de um Data Warehouse. THOMÉ (2002, 8)

Os dados de *um Data Warehouse* proporcionam base para aplicações OLAP (*On Line Analytical Processing*), DSS (*Decision Support Systems*) e ferramentas de *Data Mining*. Por seu alto custo de implementação, algumas organizações não adotam o *Data Warehouse*, optando pelo uso do *Data Mart*, que é menos abrangente, armazenando dados apenas de assuntos distintos. INMON (1997) caracteriza os *Data Marts* como "subconjuntos de dados da empresa armazenados fisicamente em mais de um local, geralmente divididos por

departamento (*data marts* "departamentais")".

O *Data Mining* não necessita obrigatoriamente de um *Data Warehouse* ou um *Data Mart* para ser executada, porém seu uso é mais eficiente nestas bases, pois nelas os dados se encontram agregados ou sumarizados, facilitando e otimizando todo o processo de mineração.

2 O PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

Para um melhor entendimento do *Data Mining*, deve-se conhecer o contexto do qual se trata, que é o Processo de Descoberta de conhecimento em bases de dados, ou KDD (*Knowledge Discovery in Databases*), nome dado ao processo de busca e extração de conhecimento em bases de dados. Ainda não há um consenso sobre o assunto, e várias nomenclaturas podem ser encontradas para este processo, inclusive o próprio termo *Data Mining* é utilizado por alguns para descrever todo o processo.

> KDD, portanto, se caracteriza por ser um processo não trivial, que busca gerar conhecimento que seja novo e potencialmente útil para aumentar os ganhos, reduzir os custos ou melhorar o desempenho do negócio, através da procura e da identificação de padrões a partir de dados armazenados em bases muitas vezes dispersas e inexploradas. THOMÉ (2002: 11)

Segundo ELMASRI, NAVATHE (2005: 621), o KDD é composto de seis fases: seleção de dados, limpeza, enriquecimento, transformação ou codificação, *Data Mining* e construção dos relatórios de apresentação.

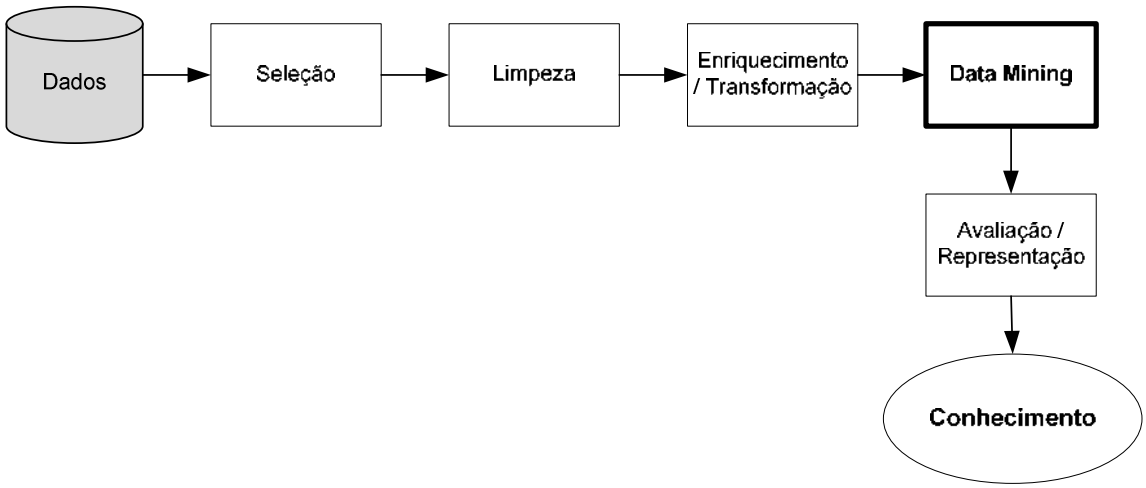

Figura 2: Etapas do Processo de Descoberta de Conhecimento em Bases de Dados

A fase de seleção de dados é onde itens específicos em um banco de dados são selecionados para o Processo de Descoberta do Conhecimento. Um exemplo é um banco de dados comercial onde diversos dados de clientes e transações podem ser recuperados, e dependendo do assunto a ser pesquisado, os dados a ele relevantes são selecionados. A fase de limpeza, também chamada de pré-processamento, corrige as inconsistências encontradas nos dados para garantir a confiabilidade nos dados que serão utilizados pela mineração. Segundo NAVEGA (2002: 2), as bases de dados são dinâmicas, incompletas, redundantes, ruidosas e esparsas, necessitando de um pré-processamento para "limpá-las". A fase de enriquecimento adiciona novos dados agregando-os aos existentes, como por exemplo, a informação da cidade e região vinda da análise dos prefixos dos telefones. A codificação ou transformação é o processo onde a quantidade de dados é reduzida, agrupando valores em outras categorias sumarizadas. Utilizando o exemplo anterior dos números de telefones, os mesmos poderiam ser substituídos pela cidade e região encontrada analisando seus prefixos. As fases de enriquecimento e transformação vistas anteriormente podem ser interpretadas como uma única fase. Na fase *Data Mining*, a garimpagem dos dados deve acontecer, após todo o pré-processamento das grandes quantidades de dados que tratadas e sumarizadas, estarão livres de ruídos e conterão somente os dados relevantes à pesquisa a ser feita.

> Para ser genérico, é necessário "perder" um pouco dos dados, para só conservar *a essência* da informação. O processo de Data Mining localiza padrões através da judiciosa aplicação de processos de generalização, algo que é conhecido como *indução.* NAVEGA (2002, 3).

Para ELMASRI, NAVATHE (2005: 625) os propósitos do Data Mining se enquadram, de forma geral nas seguintes classes:

- **Predição**: Projeções feitas para identificar o comportamento de certos atributos no futuro;
- **Identificação**: Padrões de dados que podem identificar a presença de um item, um evento ou uma atividade;
- **Classificação:** Particionamento dos dados, onde as classes ou categorias podem ser identificadas através de combinações de parâmetros;
- **Otimização:** Realiza tarefas semelhantes às das técnicas de pesquisa operacional para otimizar recursos limitados, maximizando variáveis de saída.

A mineração de dados será feita utilizando, dentre as técnicas disponíveis, a que melhor se aplica ao tipo de informação a ser encontrada. ELMASRI, NAVATHE (2005: 626) identificam as seguintes tarefas que descrevem o conhecimento descoberto durante o *Data Mining*: Regras de associação, hierarquias de classificação, padrões seqüenciais, padrões com séries temporais e *clustering* (agrupando). HARISSON (1998) apud DIAS (20--) afirma que

> Não há uma técnica que resolva todos os problemas de mineração de dados. Diferentes métodos servem para diferentes propósitos; cada método oferece suas vantagens e suas desvantagens. A familiaridade com as técnicas é necessária para facilitar a escolha de uma delas de acordo com os problemas apresentados.

3 APRENDIZADO DE MÁQUINA

O *Data Mining*, através de suas técnicas, realiza o que é chamado de Aprendizado de Máquina, descrito por MONARD, et. al. (2003: 1170) como

> Uma área de IA[5] cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática. Um sistema de aprendizado é um programa de computador que toma decisões baseado em experiências acumuladas através da solução bem sucedida de problemas anteriores. Os diversos sistemas de aprendizado de máquina possuem características particulares e comuns que possibilitam sua classificação quanto à linguagem de descrição, modo, paradigma e forma de aprendizado utilizado.

Através da análise de um conjunto de dados, pode-se projetar, classificar e agrupar dados de forma a garimpar novos conhecimentos, e o Aprendizado de máquina, contexto ao qual o *Data Mining* está inserida, realiza esta tarefa por indução. MONARD, et. al. (2003: 2) definem indução como

> A forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos. Ela é caracterizada como o raciocínio que se origina em um conceito específico e o generaliza, ou seja, da parte

[5] Inteligência Artificial

> para o todo. Na indução, um conceito é aprendido efetuando-se inferência indutiva sobre os exemplos apresentados. Portanto, as hipóteses geradas através da inferência indutiva podem ou não preservar a verdade. Mesmo assim, a inferência indutiva é um dos principais métodos utilizados para derivar conhecimento novo e predizer eventos futuros. Foi através da indução que Arquimedes descobriu a primeira lei da hidrostática e o princípio da alavanca, que Kepler descobriu as leis do movimento planetário, que Darwin descobriu as leis da seleção natural das espécies.

O aprendizado de máquina pode ser dividido em supervisionado e não supervisionado.

4 APRENDIZADO SUPERVISIONADO

É o aprendizado de máquina onde já são rotuladas as classes a serem verificadas É fornecido um banco de dados de amostra, onde os dados já estão classificados através de um atributo que representa a supervisão, e através da observação dos outros atributos destes dados, pode-se definir sua relevância para a classificação. Através disso, ao entrar um novo elemento, é possível classificá-lo de acordo com estes atributos. Um exemplo é a análise dos dados de clientes inadimplentes e não inadimplentes, caracterizados por um atributo que os identifica em uma das situações. Através de métodos de aprendizado supervisionado, pode-se chegar à conclusão que atributos como renda, faixa etária, região ou profissão estão ligados à situação de inadimplência de um cliente, portanto, podendo identificar entre os novos clientes, os inadimplentes em potencial. Algoritmos para classificação como indução de árvores decisórias e classificadores *bayeseanos* realizam aprendizado supervisionado.

5 APRENDIZADO NÃO SUPERVISIONADO

Diferentemente do aprendizado supervisionado, o aprendizado não supervisionado se aplica em situações onde não se conhecem as classes dos dados a serem analisados, tendo como objetivo agrupá-los com elementos que possuem alguma característica em comum, para então, dada a relevância desta característica, rotular os dados com as classes encontradas.

Aplicações para o aprendizado não supervisionado são várias, por exemplo, o

agrupamento de clientes de acordo com suas preferências de consumo ou pacientes de acordo com as reações a um medicamento. Analisando a similaridade entre determinados atributos, podem-se determinar grupos e a qual deles um indivíduo pertence.

CONCLUSÃO

As Tecnologias e os Sistemas de Informação são fundamentais na obtenção e extração de informações estratégicas de bases de dados que auxiliem na tomada de decisão, já que o papel que a informação desempenha na atualidade e dentro das organizações é de extrema importância e seus benefícios são evidentes. A técnica de Data Mining, especificamente para consulta de base de dados, tem mostrado eficaz no processo de extração de informações relevantes. Este artigo demonstrou o processo de descoberta de conhecimento em bases de dados e as principais técnicas utilizadas nesta metodologia.

REFERÊNCIAS

AMO, Sandra de. **Técnicas de Mineração de Dados**. Uberlândia, MG: 2003. Programa de Mestrado em Ciência da Computação, Universidade Federal de Uberlândia.

ELMASRI, Ramez.; NAVATHE, Shamkant**. Conceitos de Data Mining. In:** Sistemas de Banco de Dados. São Paulo: Pearson Addison Wesley, 2005, p. 624-645**.**

EVSUKOFF, Alexandre G.; ARBEX, Márcio A. **Inteligência Computacional**. Rio de Janeiro, RJ: 2008. Universidade Federal do Rio de Janeiro.

INMON, W. H.; HACKATHORN, R. D. **Como usar o Data Warehouse**. Rio de Janeiro: Infobook, 1997.

JESUS, Cláudia S. et al. **A Informação, o Processo Decisório e as Ferramentas para este Fim**. Salvador, BA: 2004. Curso de Administração, Faculdade Ruy Barbosa.

MONARD, Carolina M.; BARANAUSKAS, José A. **Conceitos sobre aprendizado de máquina**. São Paulo, SP: 2003. Universidade de São Paulo.

NAVEGA, Sérgio. **Princípios essenciais do Data Mining**. São Paulo, SP: 2002. Publicada nos Anais do Infoimagem.

THOMÉ, Antônio C. G. Data Warehouse, Data Mining. In: **Redes Neurais – Uma ferramenta para KDD e Data Mining**. [s.i.]: 20--.