Mineração de Dados em Biologia Molecular

KDD

Docente: André C. P. L. F. de Carvalho
PAE: Victor Hugo Barella

Tópicos do Módulo

- Introdução
- Descoberta de Conhecimento em Bases de Dados
- Etapas de KDD
- Mineração de Dados
- Aplicações

Introdução

- Bases de Dados podem conter (esconder) dados preciosos
- Existe um interesse crescente em explorar esses dados armazenados
  - Descobrir conhecimento novo
  - Apoio à tomada de decisão

O que é um dado?

- Dado
  - Exemplo, objeto, registro
  - Estrutura fundamental sobre a qual um sistema de informação é construído
  - Não tem um significado associado a ele
    - Ex.: 345,43

Exemplo

muito alto,muito alto,2,2,pequeno,baixo,ruim
muito alto,alto,3,5 ou mais,grande,baixo,ruim
muito alto,baixo,3,4,grande,baixo,ruim
médio,baixo,4,2,pequeno,alto,ruim
médio,baixo,3,4,pequeno,médio,médio
alto,alto,2,4,grande,médio,médio
baixo,baixo,5 ou mais,4,pequeno,médio,médio
baixo,médio,4,4,pequeno,médio,médio
baixo,médio,4,4,grande,médio,bom
baixo,baixo,4,5 ou mais,grande,médio,bom
médio,baixo,2,4,pequeno,alto,bom
baixo,médio,4,4,grande,alto,muito bom
médio,médio,2,4,grande,alto,muito bom
baixo,baixo,5 ou mais,5 ou mais,grande,alto,muito bom

Informação

- Dados com um significado associado
  - Tornam os dados úteis em uma tomada de decisão
  - Transformação de dados em informação
    - Geralmente pela apresentação dos dados em uma forma compreensível para o usuário
  - Informação é criada quando é associado um significado a um conjunto de dados

Exemplo - Carros

- Preço
  - Compra: muito-alto, alto, médio, baixo
  - Manutenção: muito-alto, alto, médio, baixo
- Características técnicas
  - Conforto
    - # portas: 2, 3, 4, 5-5 ou mais
    - # pessoas: 2, 4, 5 ou mais
    - Espaço do porta malas: pequeno, médio, grande
  - Segurança: baixo, médio, alto
- Aval. do carro: ruim, médio, bom, muito bom

Exemplo - Carros

muito alto,muito alto,2,2,pequeno,baixo,ruim
muito alto,alto,3,5 ou mais,grande,baixo,ruim
muito alto,baixo,3,4,grande,baixo,ruim
médio,baixo,4,2,pequeno,alto,ruim
médio,baixo,3,4,pequeno,médio,médio
alto,alto,2,4,grande,médio,médio
baixo,baixo,5 ou mais,4,pequeno,médio,médio
baixo,médio,4,4,pequeno,médio,médio
baixo,médio,4,4,grande,médio,bom
baixo,baixo,4,5 ou mais,grande,médio,bom
médio,baixo,2,4,pequeno,alto,bom
baixo,médio,4,4,grande,alto,muito bom
médio,médio,2,4,grande,alto,muito bom
baixo,baixo,5 ou mais,5 ou mais,grande,alto,muito bom

- Preço
  - Compra: muito-alto, alto, médio, baixo
  - Manutenção: muito-alto, alto, médio, baixo
- Características técnicas
  - Conforto
    - # portas: 2, 3, 4, 5-5 ou mais
    - # pessoas: 2, 4, 5 ou mais
    - Espaço porta malas: pequeno, médio, grande
  - Segurança: baixo, médio, alto
- Aval. do carro: ruim, médio, bom, muito bom

Exemplo

tactagcaatacgcttgcgttcggtggttaagtatgtataatgcgcgggcttgtcgt
tgctatcctgacagttgtcacgctgattggtgtcgttacaatctaacgcatcgccaa
gtactagagaactagtgcattagcttatttttttgttatcatgctaggcggcg
aattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaatactaacaaactc
Tcgataattaactattgacgaaaagctgaaagactagaatgcgcctccgtggtag
aggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgtttaggt
aggcatgtaaacgtcttcgtagcgcatcagtgctttcttactgtgagtacgcgag
ccgagtaggcttagagagcatgtcagcctcgacaacttgcataaatgctttcttg
cgctaggactttcttgttgattttccatgcggtgttttgcgcaatgttaatcgcttt
tatgggaacgagtcaatcagggctttgactctggtattactgtgaacattatt
agagggtgtactccaagaagaggaagatgaggctagacgtctctgcatggagtatga
gagagcatgtcagcctcgacaacttgcataaatgctttcttgtagacgtgccctacg

Exemplo - Promotores

+,S10,tactagcaatacgcttgcgttcggtggttaagtatgtataatgcgcgggcttgtcgt
+,AMPC, tgctatcctgacagttgtcacgctgattggtgtcgttacaatctaacgcatcgccaa
+,AROH,gtactagagaactagtgcattagcttatttttttgttatcatgctaggcggcg
+,DEOP2,aattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaatactaacaaactc
+,LEU1_TRNA,tcgataattaactattgacgaaaagctgaaagactagaatgcgcctccgtggtag
+,MALEFG,aggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgtttaggt
-, 296,aggcatgtaaacgtcttcgtagcgcatcagtgctttcttactgtgagtacgcgag
-, 648,ccgagtaggcttagagagcatgtcagcctcgacaacttgcataaatgctttcttg
-, 230,cgctaggactttcttgttgattttccatgcggtgttttgcgcaatgttaatcgcttt
-,1163,tatgggaacgagtcaatcagggctttgactctggtattactgtgaacattatt
-,1321,agagggtgtactccaagaagaggaagatgaggctagacgtctctgcatggagtatga
-, 663,gagagcatgtcagcctcgacaacttgcataaatgctttcttgtagacgtgccctacg

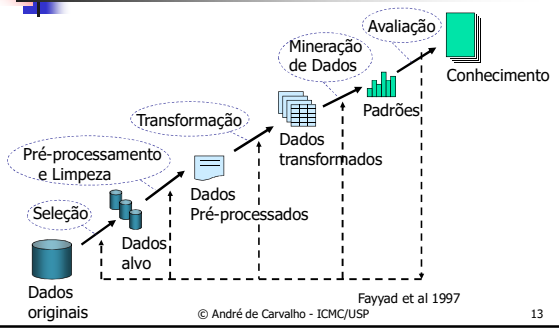
KDD

- Descoberta de conhecimento em Bancos de dados (BDs)
  - *Knowledge Discovery in Databases*
- Área de pesquisa em expansão
- Teorias e ferramentas computacionais para extrair informação útil de BDs
  - Informação útil ⇒ conhecimento

KDD

- Processo de encontrar em dados padrões que são
  - Úteis
  - Novos
  - Válidos
  - Potencialmente compreensíveis
- Processo interativo e iterativo
  - Várias etapas

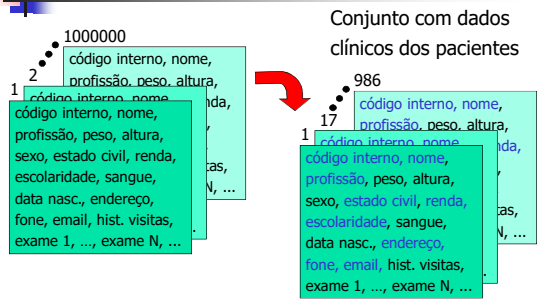
KDD



Seleção

- Cria um conjunto de dados para extração de conhecimento
  - Seleciona "manualmente" um subconjunto dos dados disponíveis
    - Subconjunto de registros (instâncias ou exemplos)
    - Subconjunto de atributos considerados relevantes para o problema
      - Elimina atributos que sejam claramente irrelevantes

Exemplo



Pré-processamento e Limpeza

- Melhora a qualidade dos dados e facilita sua posterior utilização
- Engloba várias operações
  - Amostragem "automática" de objetos
  - Seleção "automática" de atributos
  - Conversão de valores
  - Lidar com atributos ausentes
  - Eliminar dados duplicados
  - Detectar ruído

Transformação

- Inclui operações que modificam valores para um dado atributo
  - Cada operação deve ser aplicada a todos os valores do atributo
    - Em todos os objetos
  - Ex.: normalização, valor absoluto, ...

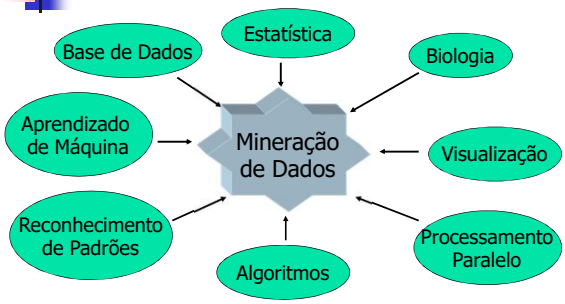
Mineração de Dados

- Principal passo no processo de KDD
  - Mineração de Dados (DM) e KDD são frequentemente utilizados como sinônimos
- Fronteiras da etapa de MD no processo de KDD são de difícil identificação
  - Pré-processamento e transformação de dados são geralmente vistos como uma parte da MD

Mineração de Dados

- MD X KDD
  - MD: ferramentas básicas utilizadas para extrair padrões de dados
  - KDD: processo que engloba o uso dessas ferramentas, além de:
    - Pré-processamento, seleção e transformação dos dados
    - Interpretação dos padrões
      - Geração de conhecimento
      - Suporte à tomada de decisão

Mineração de Dados



Mineração de Dados

- Outros termos utilizados para MD e KDD
  - Extração de conhecimento
  - Descoberta de informação
  - Extração de padrões
  - Análise exploratória de dados
  - Analítica (Data analytics ou analytics)

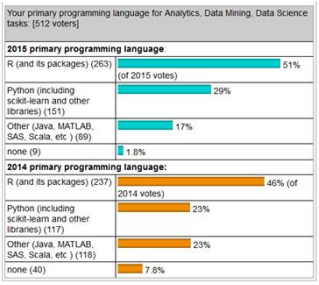
Analítica

- Ciência que analisa dados crus para extrair padrões desses dados
  - Pode englobar coleta e organização dos dados
- Analítica preditiva (predictive analytics)
  - Extrai modelos a partir de dados para fazer predições futuras
- Analítica descritiva (descriptive analytics)
  - Sumariza ou condensa dados para extrair informações

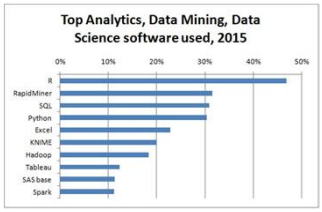
Interpretação / Avaliação

- Interpretação dos padrões encontrados na etapa de MD
  - Possível retorno a qualquer uma das etapas anteriores para iteração adicional
- Valida padrões encontrados
  - Importante consulta a um especialista
- Inclui análise estatística
- Ferramentas de visualização fornecem um suporte importante

Ferramentas



Ferramentas



Custos em MD Preditivo

- 15% - coleta de dados
- 60% - limpeza e pré-processamento de dados
- 15% - construção e análise de modelos
- 5% - aplicação
- 5% - melhorias contínuas

CRISP-DM

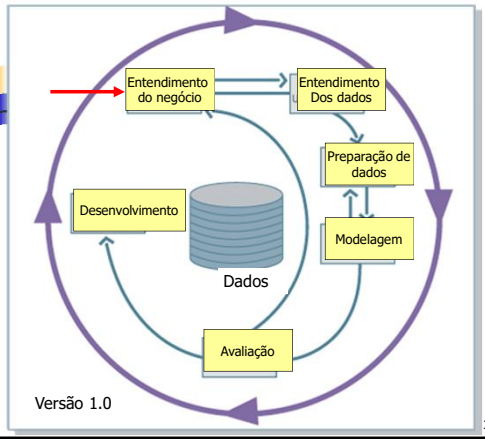
- Projeto CRISP-DM
  - *CRoss-Industry Standard Process for Data Mining*
  - Concebido em 1996 por:
    - Daimer-Chrysler
      - Aplicava MD em suas operações de negócios
    - SPSS
      - Prestava serviço de MD desde 1990
      - Desenvolveu primeira ferramenta comercial de MD (*Clemetine*)
    - NDR
      - Tinha o propósito de adicionar valor a sua enorme BD

CRISP-DM

- Projeto CRISP-DM
  - Desenvolveu um novo fluxo de processo para descoberta de conhecimento
    - A partir do processo anterior
      - Fayyad, Piatesky-Shapiro and Smyth
  - Em resposta a requisitos de usuários
    - Definiu e validou processo de MD utilizado em vários setores industriais

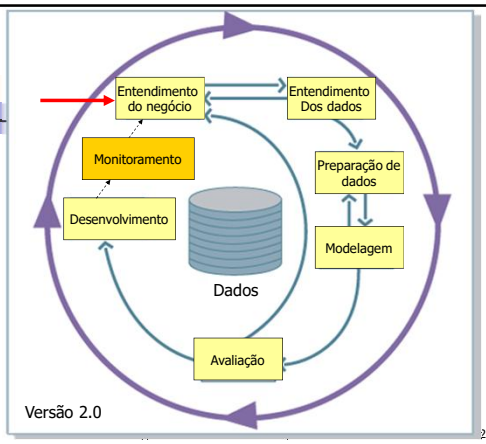
CRISP-DM

- Metodologia procura tornar os projetos
  - Mais rápidos
  - Mais baratos
  - Mais confiáveis
  - Mais facilmente gerenciáveis
- Pode ser aplicada a pequenos projetos
- Metodologia padrão da indústria



CRISP-DM 2.0

- SIG (*special interest group*) formado entre 2006 e 2008
- Mudanças estudadas
  - Divisão da fase de preparação de dados
  - Métodos de avaliação dentro da fase de modelagem
  - Fase de avaliação associada a avaliação na empresa
    - Centro de pesquisa, laboratório, hospital,...
  - Inclusão de fase de monitoramento



CRISP-DM 2.0

- Não foi lançado até 2005
- SIG foi desfeito
  - Website CRISP-DM.org não esta mais ativo
- IBM criou nova metodologia
  - Refina e estende CRISP-DM
  - Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM)

Produtos para MD

Mais Produtos

Considerações Finais

- Expansão do volume de dados armazenados
  - Necessidade de extrair conhecimento dos dados
- KDD é cada vez mais usado por órgãos privados e públicos
- Cuidado com promessas exageradas
- Leitura
  - Knowledge Discovery and Data Mining: Towards a Unifying Framework, U. Fayyad, P. Smyth, and G. Piatetsky-Shapiro, .2nd International Conference on Knowledge Discovery and Data Mining, 1996

Perguntas