Universidade do Minho
Escola de Engenharia
Departamento de Informática

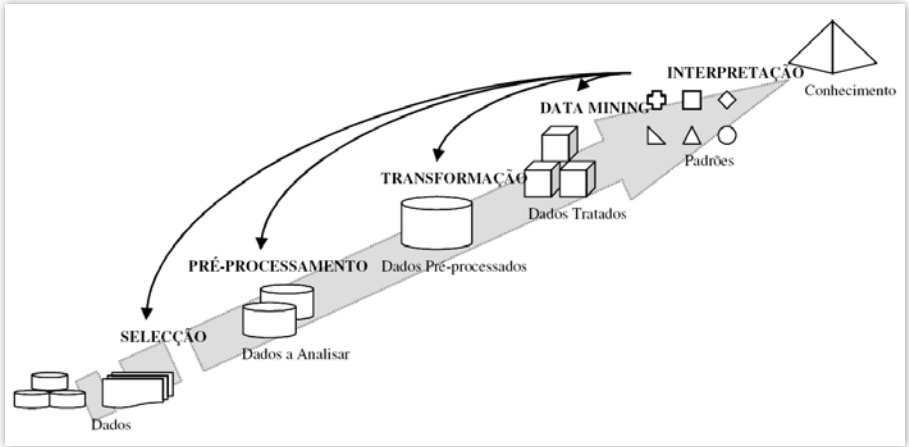
Preparação de Dados para Extração de Conhecimento

Cesar Analide, Paulo Novais, José Neves

Preparação de Dados para Extração de Conhecimento

Preparação de dados

- Preparação dos Dados (Pré-processamento), porquê?
  - Discretização;
    (classes etárias)
  - Limpeza;
    (nº BI)
  - Integração e Transformação;
    (fontes; diários/mensais)
  - Redução de dados.
    (moradas/regiões)



Data Mining – Descoberta de Conhecimento em Bases de Dados
Manuel Filipe Santos, Carla Azevedo

Preparação de Dados para Extração de Conhecimento

Porquê preparar os dados?

- **Porque SIM!**
- O principal objetivo da **preparação dos dados consiste em transformar** os *data sets* por forma a que a **informação** neles contida esteja **adequadamente exposta à ferramenta** de extração de conhecimento;
- A preparação dos dados também "prepara o preparador" por forma selecionar os modelos de EC mais adequados;
- Os dados têm de ser formatados para se adequarem a uma determinada ferramenta de EC;
- Os dados recolhidos do "mundo real":
  - são incompletos:
  - contêm lixo;
  - podem conter inconsistências.

Preparação de Dados para Extração de Conhecimento

Porquê preparar os dados?

- Os dados recolhidos do "mundo real":
  - são incompletos:
    - falta de valores em alguns atributos, falta de alguns atributos ou dados agregados ou generalizados;
    - Código postal: 4710-... Braga;
    - Nº de filhos: "";
  - contêm lixo;
  - podem conter inconsistências.

Preparação de Dados para Extração de Conhecimento

Porquê preparar os dados?

- Os dados recolhidos do "mundo real":
  - são incompletos;
  - contêm lixo:
    - identificam valores impossíveis;
    - Salário: -1.000EUR;
    - Idade: 321;
    - Data: 31/abril/2005;
    - País: Madeira;
  - podem conter inconsistências.

Preparação de Dados para Extração de Conhecimento

Porquê preparar os dados?

- Os dados recolhidos do "mundo real":
  - são incompletos;
  - contêm lixo;
  - podem conter inconsistências:
    - encontram-se discrepâncias entre valores ou nomes;
    - Idade = 35; Data de nascimento = 31/maio/1969;
    - Sexo: "M/F"; "0/1"; "Masculino/Feminino/Desconhecido";
    - diferenças entre valores de registos duplicados.

Preparação de Dados para Extração de Conhecimento

Tarefas na preparação de dados

- Discretização:
  - Redução de dados com importante aplicação a dados numéricos;
- Limpeza;
- Integração;
- Transformação;
- Redução.

Preparação de Dados para Extração de Conhecimento

Tarefas na preparação de dados

- Discretização;
- Limpeza:
  - Preenchimento de valores de atributos;
  - Remoção de lixo dos dados;
  - Remoção de valores impossíveis;
  - Resolução de inconsistências;
- Integração;
- Transformação;
- Redução.

Preparação de Dados para Extração de Conhecimento

Tarefas na preparação de dados

- Discretização;
- Limpeza;
- Integração:
  - Integração de dados provenientes de múltiplas fontes (BD's, ficheiros, papel, web, etc.);
- Transformação;
- Redução.

Preparação de Dados para Extração de Conhecimento

Tarefas na preparação de dados

- Discretização;
- Limpeza;
- Integração;
- Transformação:
  - Normalização e agregação de dados;
- Redução.

Preparação de Dados para Extração de Conhecimento

Tarefas na preparação de dados

- Discretização;
- Limpeza;
- Integração;
- Transformação;
- Redução:
  - obtenção de representações de dados menos volumosas, mas com capacidade para produzir idênticos resultados analíticos;
  - agregação, redução de dimensões e compressão de dados.

Preparação de Dados para Extração de Conhecimento

Tipos de dados

- Os tipos dos dados diferem na sua natureza e na quantidade de informação que proporcionam:
- **Qualitativos ou Quantitativos**.

Preparação de Dados para Extração de Conhecimento

Tipos de dados

- Nominais:
  - Atribui nome únicos a objetos:
    - Não existe outra informação que se possa deduzir;
    - Nomes de pessoas;
    - Códigos de identificação;
- Categorias;
- Ordinais;
- Intervalos;
- Rácios.

Preparação de Dados para Extração de Conhecimento

Tipos de dados

- Nominais;
- Categorias:
  - Atribui categorias a objetos:
    - Podem ser valores numéricos, mas são **não ordenados**;
    - Código postal;
    - Sexo;
    - Cor dos olhos;
- Ordinais;
- Intervalos;
- Rácios.

Preparação de Dados para Extração de Conhecimento

Tipos de dados

- Nominais;
- Categorias;
- Ordinais:
  - Os valores podem ser ordenados naturalmente;
    - Classificação: Excelente, Bom, Suficiente, etc.;
    - Temperatura: frio, morno, quente;
- Intervalos;
- Rácios.

Preparação de Dados para Extração de Conhecimento

Tipos de dados

- Nominais;
- Categorias;
- Ordinais;
- Intervalos:
  - É possível calcular a distância entre dois valores;
    - Temperatura;
- Rácios.

Preparação de Dados para Extração de Conhecimento

Tipos de dados

- Nominais;
- Categorias;
- Ordinais;
- Intervalos;
- Rácios:
  - Os valores podem ser utilizados para determinar um rácio significativo entre eles:
    - Salário;
    - Balanço bancário.

Preparação de Dados para Extração de Conhecimento

Tipos de dados

- Nominais;
- Categorias;
- Ordinais;

Qualitativos

- Intervalos;
- Rácios.

Quantitativos

Discretos

Contínuos

Maior conteúdo informativo

Preparação de Dados para Extração de Conhecimento

Tarefas na preparação de dados

- Discretização;
- Limpeza;
- Integração;
- Transformação;
- Redução.

Preparação de Dados para Extração de Conhecimento

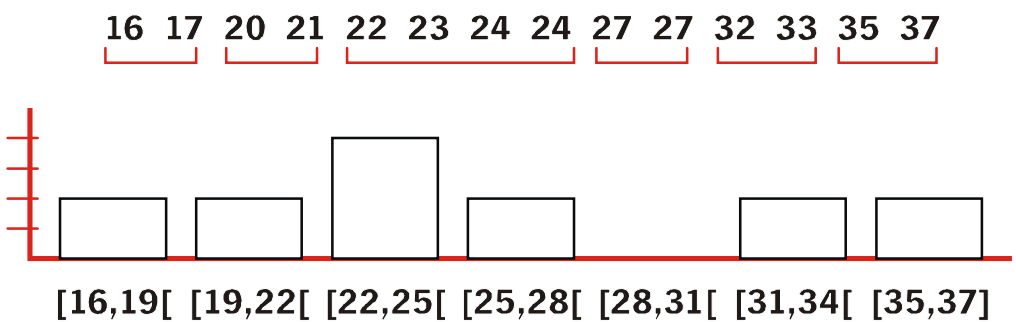
Discretização

- Utiliza-se a discretização (ou enumeração) para reduzir o número de valores de um atributo contínuo, dividindo-o em intervalos;
  - Os métodos mais utilizados (Naïve Bayes, CHAID, etc.), requerem valores discretos;
  - Redução do tamanho dos dados;
  - Método utilizado para produzir sumariação dos dados;
  - (Sinónimo de *binning*.)

Preparação de Dados para Extração de Conhecimento

Discretização de igual largura

- *Equal-width binning*:
- Divide a gama de valores em N intervalos de igual largura, resultando numa grelha uniforme;
- Sendo A e B os limites da gama de valores, a largura dos intervalos será L = (B - A) / N:



Preparação de Dados para Extração de Conhecimento

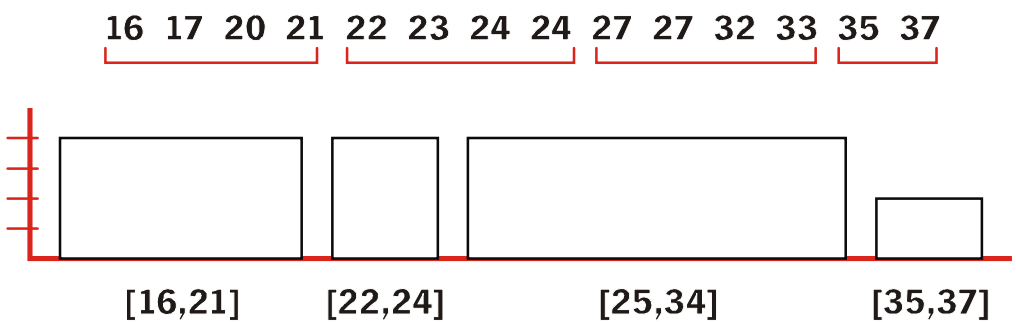
Discretização de igual largura

- Vantagens:
  - Simples e fácil de implementar;
  - Produz abstrações de dados razoáveis;
- Desvantagens:
  - Não supervisionado;
  - Quem determina N?;
  - Sensível a valores fronteira.

Preparação de Dados para Extração de Conhecimento

Discretização de igual altura

- *Equal-height binning*:
- Divide a gama de valores em N intervalos, contendo, cada um, **aproximadamente** a mesma quantidade de valores:



Preparação de Dados para Extração de Conhecimento

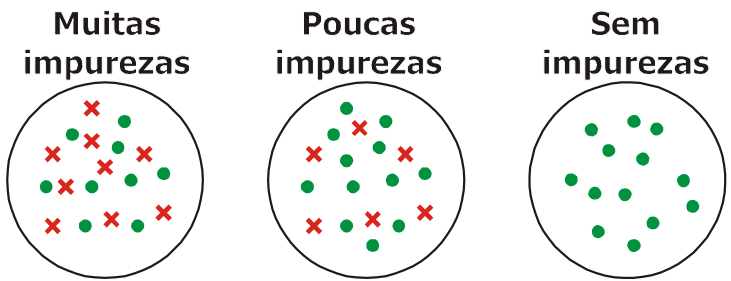
Discretização de igual altura

- Normalmente preferida à discretização por igual largura, uma vez que permite evitar o "amontoar" de valores;
- Na prática, utiliza-se uma discretização de "quase-igual" altura, garantindo intervalos mais intuitivos;
- Não deverá permitir a dispersão de valores frequentes por diferentes intervalos;
- Deverá criar intervalos separados para valores especiais ("0").

Preparação de Dados para Extração de Conhecimento

Discretização: outros métodos

- Método 1R:
  - Método supervisionado, baseado na divisão por *binning*;
- Discretização baseada em Entropia;
- Discretização baseada em Impurezas;



- Deteção de limites;
- etc.

Preparação de Dados para Extração de Conhecimento

Limpeza de dados

- Ausência de valores em determinados atributos devido a:
  - inconsistência;
  - dados não registados;
  - análise incorreta;
  - dados registados de forma errada;
  - etc.
- **A ausência de dados pode revelar algo sobre que campos não foram preenchidos!**

Preparação de Dados para Extração de Conhecimento

Limpeza de dados:

como tratar a ausência de dados

- Ignorar os registos onde faltam os dados e lidar, apenas com os dados conhecidos;
  - não aconselhável se a quantidade de dados em falta em cada atributo for elevada;
- Ignorar os atributos onde faltam os dados;
  - não aconselhável se os atributos onde acontece revelarem informação importante;
- Preencher (manualmente) os dados em falta:
  - é mais trabalhoso preencher ou é mais difícil adivinhar?
- Preencher os dados em falta com um mesmo valor ("talvez") pode criar novas classes;
- Preencher com o valor médio do atributo:
  - pouco impacto negativo, desde que o desvio padrão não seja grande;
- Preencher com o valor mais frequente do atributo;
- IMPORTANTE: evitar adicionar distorção aos dados.
- Quantos mais valores "inventados", maior o desvio dos dados que caracterizam o problema face à realidade que o problema ilustra!

Preparação de Dados para Extração de Conhecimento

Transformação de dados

- Alisamento (smoothing):
  - remover lixo/ruído dos dados (binning, regressão, clustering);
- Agregação;
- Generalização;
- Construção de Atributos;
- Normalização.

Preparação de Dados para Extração de Conhecimento

Transformação de dados

- Alisamento (smoothing);
- Agregação:
  - A agregação dos dados pressupõe que o resultado sumaria os dados iniciais (resumo de vendas trimestrais, durante 5 anos, em valores anuais);
- Generalização;
- Construção de Atributos;
- Normalização.

Preparação de Dados para Extração de Conhecimento

Transformação de dados

- Alisamento (smoothing);
- Agregação;
- Generalização:
  - Hierarquização de conceitos:
    - distrito → cidade → rua;
    - Valores diferentes: 18 → centenas → (largos) milhares
- Construção de Atributos;
- Normalização.

Preparação de Dados para Extração de Conhecimento

Transformação de dados

- Alisamento (smoothing);
- Agregação;
- Generalização;
- Construção de Atributos:
  - Construção de novos atributos a partir de outros (cálculo do preço líquido baseado no preço ilíquido e no IVA);
- Normalização.

Preparação de Dados para Extração de Conhecimento

Transformação de dados

- Alisamento (smoothing);
- Agregação;
- Generalização;
- Construção de Atributos;
- Normalização:
  - pretende evitar que atributos com uma gama alargada de valores sobressaiam em relação a outros atributos com menor quantidade de valores.

Preparação de Dados para Extração de Conhecimento

Integração de dados

- Os dados que caracterizam o problema podem ter proveniências diversas;
- O objetivo da integração é o de compor um conjunto de peças de informação numa coleção coerente e integrada de dados.
- Detetar e resolver conflitos entre os dados:
  - qual a fonte de dados mas fiável, quando os valores que transportam são inconsistentes?
- Integração exige "**conhecimento do negócio**".

Preparação de Dados para Extração de Conhecimento

Redução de dados

- Um *Data Warehouse* pode armazenar largos terabytes de dados;
- Realizar tarefas de EC em tais quantidades de dados pode tornar-se impraticável!
- A Redução de dados pretende obter uma representação reduzida do volume de dados, mas produzindo os mesmos (ou quase os mesmos) resultados analíticos.

Preparação de Dados para Extração de Conhecimento

Redução de dados: estratégias

- Construção de cubos de dados:
  - as operações de agregação são aplicadas de modo a construir cubos de dados;
- Redução de dimensões:
  - remoção de atributos que se mostrem irrelevantes, redundantes ou pouco interessantes para a análise;
- Compressão de dados:
  - aplicação de técnicas de compressão ou de transformação para comprimir a representação dos dados originais;
- Redução de quantidade:
  - redução do volume de dados (técnicas paramétricas ou não paramétricas);
- Discretização e generalização de conceitos:
  - redução da quantidade de valores por atributo.

Preparação de Dados para Extração de Conhecimento

Referências bibliográficas

- Data Preparation for Data Mining
  Dorian Pyle
- Data Mining: Concepts and Techniques
  Jiawei Han, Micheline Kamber
- Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations
  Ian Witten, Eibe Frank
- Data Mining: Descoberta de Conhecimento em BDs
  Manuel Filipe Santos, Carla Azevedo

Intelligent Systems Lab

Contactos

- Universidade do Minho
- Escola de Engenharia
- Departamento de Informática
- http://islab.di.uminho.pt
- DI-3.22