CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Pré-processamento

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com

Por que fazer pré-processamento?

- Dados reais são problemáticos
  - **Incompletos**
    - Valores ausentes, atributos ausentes
  - **Ruídos**
    - Valores errados ou outliers
  - **Inconsistentes**
    - Contém discrepâncias em códigos ou nomes
  - **Sem qualidade** nos dados - não há resultados
    - Decisões de qualidade devem ser baseadas em dados de qualidade
    - Bases precisam de integração consistente de dados

Principais tarefas de pré-processamento

- **Limpeza de dados**
  - Valores ausentes, ruídos, *outliers*, inconsistências
- **Integração de dados**
- **Redução de dados**
  - Dimensionalidade e numerosidade
- **Transformação de dados**
  - Normalização e agregação
- **Discretização de dados**
  - Redução de dados com particular importância para dados numéricos

Limpeza de dados

- **Problemas em dados reais**
  - Incompletos: ausência de valores
    - Endereço: “ ”
- **Ruídos, erros e *outliers***
  - Salário = “-5000,00”
- **Inconsistentes: discrepâncias em códigos ou nomes**
  - Idade = “26” e Nascimento = “06/08/1990”

Dados incompletos – Valores Ausentes

- **Dados nem sempre estão disponíveis**
  - Várias instâncias (tuplas) com valores ausentes
- **Motivos mais comuns**
  - Mau funcionamento do equipamento
  - Erro na entrada de dados
  - Inconsistência com outros dados registrados e assim o dado torna-se ausente
  - Certos dados não são considerados importantes

Processamento de valores ausentes

- Ignorar a instância (tupla)
- Preencher os valores ausentes manualmente
- Preencher os valores automaticamente
  - Usar uma constate global para representar os valores ausentes
    - Ex: “Valor Desconhecido” - nova classe
  - Usar uma média
  - Usar uma média por classe
  - Usar o valor mais provável baseada por inferência (árvore de decisão)

Características de dados com ruídos

- **Erro aleatório**
  - Os dados estão incorretos
- **Motivos**
  - Problemas com instrumentos de entrada de dados
  - Problema na transmissão de dados
  - Limitação tecnológica
  - Inconsistência na padronização da nomenclatura

Processamento de dados com ruídos

- **Regressão**
  - Fazer um *fitting* dos dados usando uma função
- **Agrupamento**
  - Detectar e remover *outliers*
- **Combinar** inspeção automática com inspeção manual
  - Detectar valores estranhos e deixar que a verificação seja feita por humanos

Integração de dados

- Combinar diferentes fontes e formatos
- Resolver problemas de identificação e duplicação
- Identificar dados repetidos de fontes diferentes (sistema métrico ou escala)
- Identificar dados redundantes (correlação)

Redução de dados

- Obter uma representação reduzida dos dados
- Grandes bases exigem alto custo computacional
- Estratégias
  - Redução da dimensionalidade
  - Redução de dados (numerosidade)

Redução da dimensionalidade

- Ajuda a reduzir dados irrelevantes
- Reduz o tempo de processamento dos algoritmos de mineração
- Facilita a visualização dos dados
- Técnicas
  - PCA (*Principal Component Analysis*)
  - Seleção de Atributos
    - Alguns dados não precisam ser minerados
    - Remove atributos irrelevantes para o processo, como por exemplo os identificadores

Redução de Dados

- Por quê? Permite que os algoritmos de mineração de dados sejam mais eficientes
- Pode ser feita de pelo menos duas maneiras:
  - Escolhe formas alternativas de representar os dados
    - Agrupar os dados e escolher uma representação para cada grupo
  - Amostragem
    - Obter uma amostra m capaz de representar o conjunto completo de dados N
    - Como amostrar os dados???
      - Aleatoriamente (não funciona bem para dados com classes desbalanceadas)
      - Estratificada (mantém a distribuição das classes nos dados originais)

Transformação de Dados

- Encontrar uma função que mapeie todos os valores de um atributo para um novo conjunto de valores
- Técnicas
  - Construção de atributos
  - Normalização
  - Padronização
  - Discretização

Discretização

- A discretização divide o intervalo de um atributo numérico em intervalos
  - Os “nomes” de cada intervalo podem então substituir os valores numéricos
  - Pode levar em conta a classe dos exemplos ou não
- Formas
  - Supervisionada
  - Não supervisionada

Discretização Supervisionada

- Divisão de acordo com alguma heurística
  - MDL (*Minimum Description Length*)
- Partição em intervalos de valores que gera o maior ganho de informação em relação à classe
- Pode não gerar partições

Discretização Não Supervisionada

- Divisão em intervalos
- Partição em intervalos de mesmo tamanho
  - Divide os valores em ***n*** intervalos de mesmo tamanho
  - Se A é o menor e B o maior valor do atributo, o intervalo é representado por (A-B)/n
- Partição em intervalos de mesma frequência
  - Divide os valores em ***n*** intervalos com o mesmo número de amostras

CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Pré-processamento

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com