UNIPAC - CENTRO UNIVERSITÁRIO PRESIDENTE ANTÔNIO CARLOS
CAMPUS BARBACENA

Bacharelado em Ciência da Computação

Mineração de dados

Material de Apoio

Parte II – Preparação dos Dados

Prof. Felipe Roncalli de Paula Carneiro
felipecarneiro@unipac.br

2º sem / 2023

Material cedido pela Profª Lívia e Profº Osvano

KDD

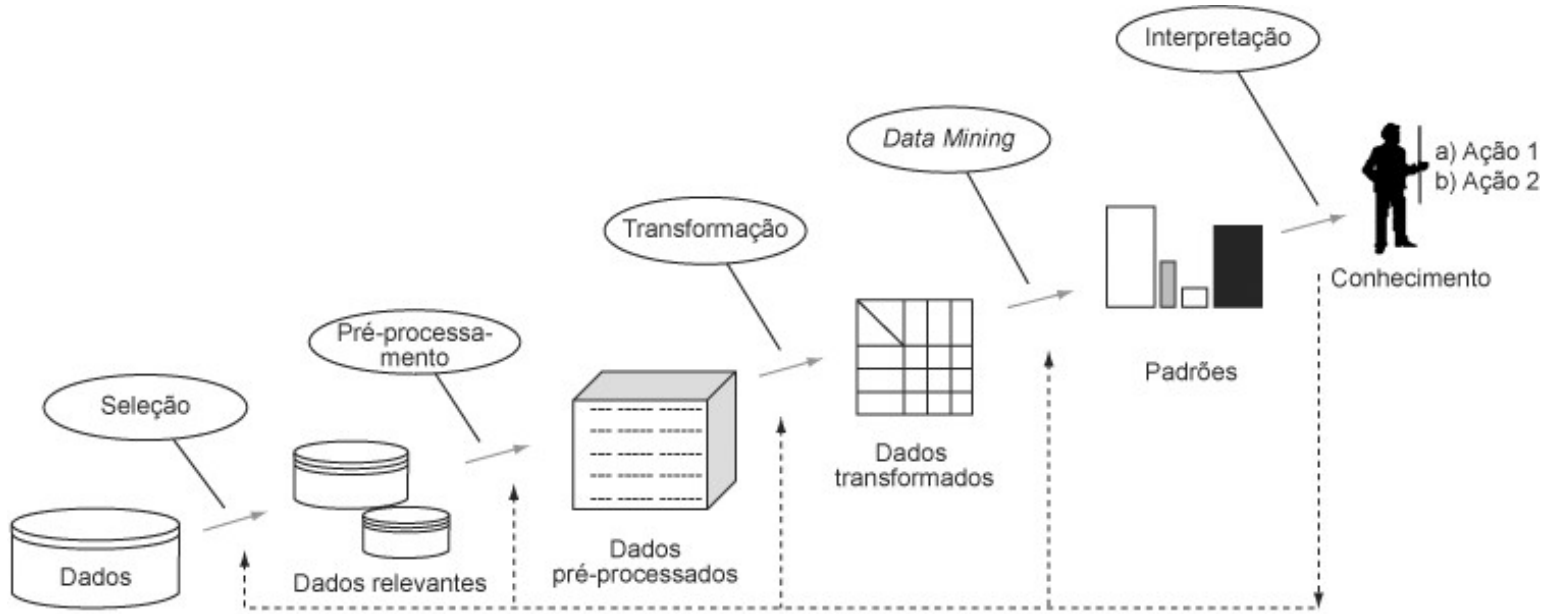


Figura 1. Etapas do processo *KDD* (Fayyad et al. (1996)).

KDD

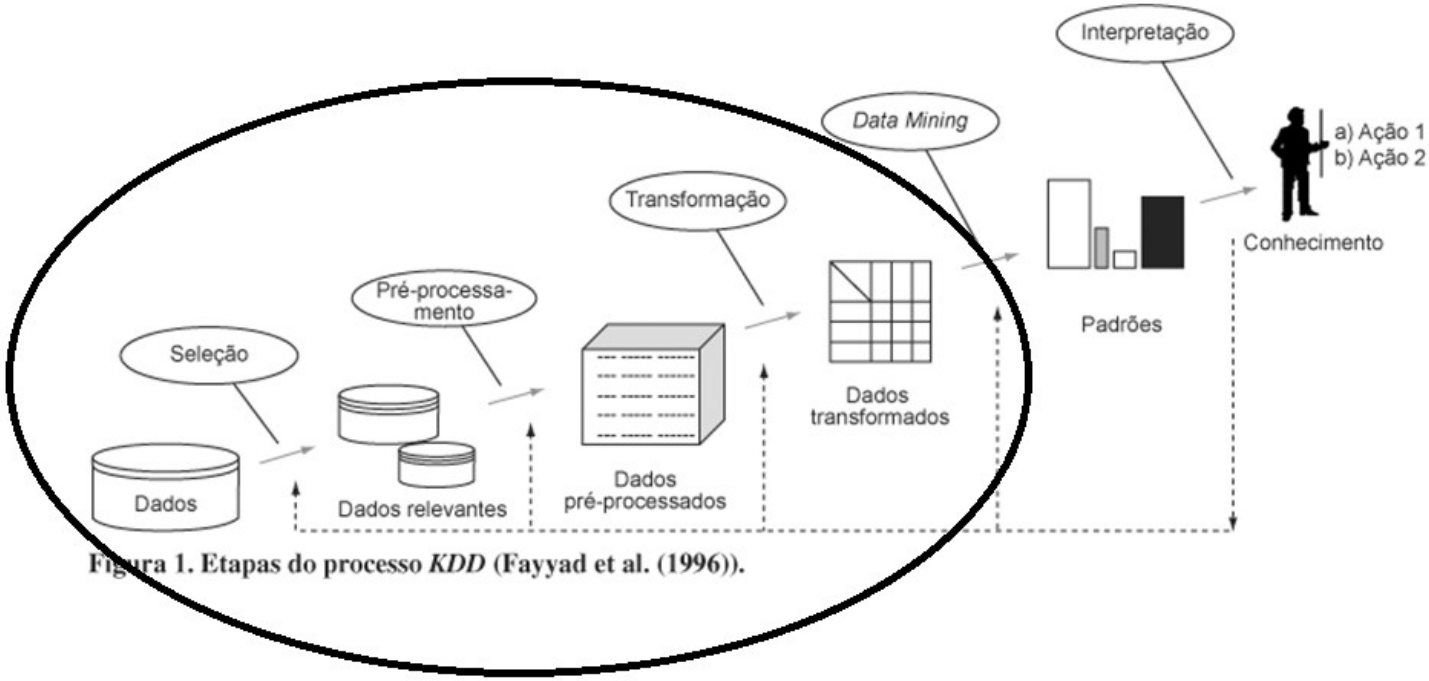


Figura 1. Etapas do processo *KDD* (Fayyad et al. (1996)).

Preparação dos dados

- No mundo real, dados coletados e organizados tendem a ser:
  - incompletos;
  - fora de padrões;
  - redundantes; e
  - inconsistentes.
- A fase de pré-processamento de dados inicia-se após a coleta e organização desses dados.
- Esta fase pode consumir até 60% do tempo disponível para exploração de dados.

Importância da preparação dos dados

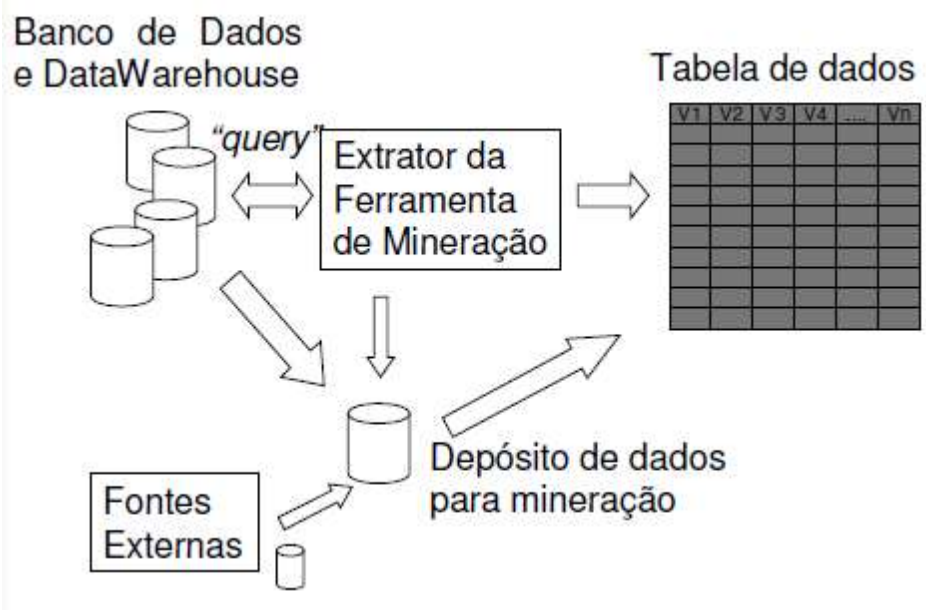
- Preparação de dados: a etapa que consome a maior parte de tempo no processo de KDD.
- O sucesso ou fracasso de um projeto de mineração de dados está relacionado à preparação de dados.
- A preparação de dados ajuda um analista:
  - Interpretar melhor os resultados;
  - Entender os limites nos dados.

Preparação dos dados

- Principais atividades:

1) Procura dos dados
2) Caracterização dos Dados
3) Montagem do Conjunto de Dados

Fonte e origem dos dados



Objetos e atributos

- Um *dataset* é uma coleção de objetos e seus atributos.
- Um atributo é uma propriedade ou característica de um objeto.
  - Exemplos: idade de uma pessoa, altura, etc.
  - Atributo é também conhecido como variável, campo, parâmetro ou “feature”.
- Uma coleção de atributos descrevem um objeto.
  - Objeto é também conhecido como registro, observação, ponto, entidade ou instância.

Objetos e atributos

- Exemplo:

Tid	Retorno	Estado Civil	Renda Anual	Mentiu
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

Melhoramento e enriquecimento dos dados

- Para uma satisfatória estruturação ou caracterização do problema é necessário introduzir nova informação que complemente à já existente.

- Dois processos podem ser utilizados:

1) Enriquecimento

2) Melhoramento

Melhoramento e enriquecimento dos dados

1) *Enriquecimento:* é o processo de inserir dados de fontes externas ao conjunto de dados.

- **Exemplo 1:** o perfil dos grupos pode não ser suficiente para decidir a liberação de crédito de uma pessoa. Pode ser necessário inserir o histórico de crédito e/ou de consumo.
- **Exemplo 2:** a valorização de um imóvel somente pelas suas características intrínsecas pode não ser suficiente para decidir seu valor. É necessário inserir informações sobre dados relativos ao lazer, índice de criminalidade, projetos de expansão futura, etc.

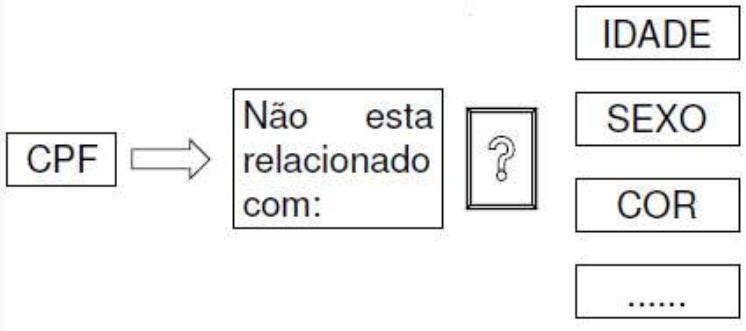
Melhoramento e enriquecimento dos dados

2) *Melhoramento:* é o processo de realçar características dos dados sem adição de fontes externas.

- **Exemplo 1:** do campo "observações clínicas" (campo textual) podem ser ***extraídas características adicionais*** para definir melhor o perfil de cada paciente.
- **Exemplo 2:** em processos físicos, quando a variabilidade de um parâmetro é grande (por exemplo, temperatura) pode ser necessário medir esse parâmetro com períodos de amostragem menor. Isso realça as características do parâmetro.

Seleção dos dados

- É necessário retirar dados irrelevantes que podem trazer conhecimento falso ou aumentar o tempo de processamento dos algoritmos de *Data Mining*.



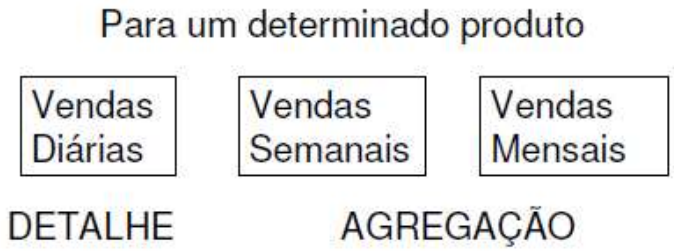
Problemas de acesso aos dados

- Fatores Legais
- Razões Políticas
- Formato dos Dados
- Conectividade
- Arquiteturas dos Banco de Dados

Caracterização dos dados

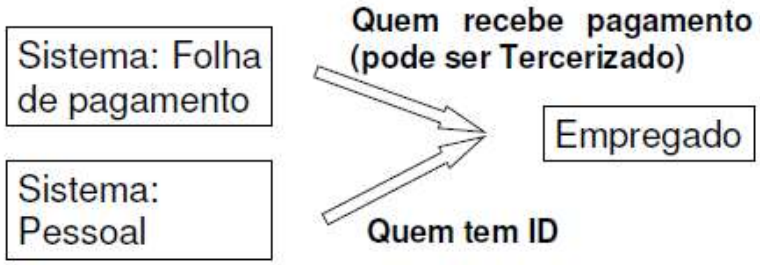
- GRANULARIDADE
- Granularidade = nível (detalhes/agregação)

Dados detalhados é preferível a dados agregados



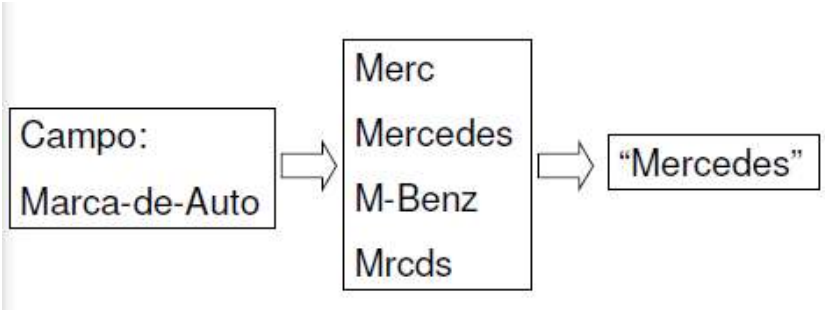
Caracterização dos dados

- CONSISTÊNCIA e INCONSISTÊNCIA
- Diferentes "*coisas*" representados pelo mesmo nome em diferentes sistemas



Caracterização dos dados

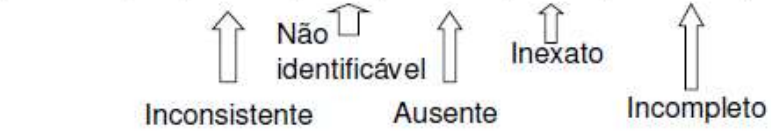
- CONSISTÊNCIA e INCONSISTÊNCIA
- A mesma *"coisa"* representada por diferentes nomes em diferentes sistemas.



Caracterização dos dados

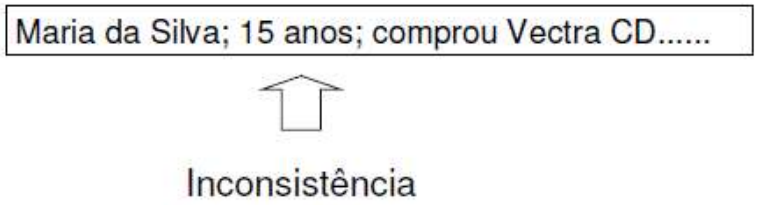
- POLUIÇÃO
- Os campos podem conter espaços em branco, estar incompletos, inexatos, inconsistentes ou não identificáveis.

Pessoa Física	EC	Data de Nasc.	Idade	Dependente	Escola/ salário	Telefone
Maria da Silva	C	28/02/03	15		30%	(xxx)45678 90



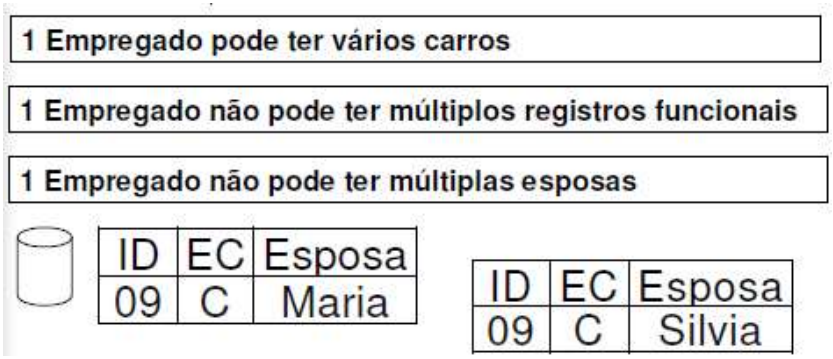
Caracterização dos dados

- RELAÇÕES
- É importante observar e analisar a consistência das instâncias dos objetos da estrutura problema.



Caracterização dos dados

- INTEGRIDADE
- Deve ser observada a integridade das relações avaliadas.



Caracterização dos dados

- DUPLICAÇÕES OU REDUNDÂNCIAS
- Ocorre principalmente quando as instâncias dependem de diferentes fluxos de dados
- As variáveis Duplicadas ou Redundantes exigem maior esforço computacional e dependendo do caso podem ser reduzidas.

Data de Nascimento => Idade Atual
Preço Unitário * Quantidade => Preço total

Principais tarefas no pré-processamento

- Limpeza dos dados
  - Preencher valores faltantes;
  - Reduzir ruídos nos dados;
  - Identificar e remover outliers;
  - Identificar e eliminar inconsistências;
  - Identificar e eliminar redundâncias.
- Integração de dados
  - Integração de múltiplos repositórios;
  - Integração de arquivos.

Principais tarefas no pré-processamento

- Transformação de dados
  - Normalização;
  - Agregação.
- Redução de dados
  - Obtenção da representação reduzida em volume, mas que produza resultados analíticos similares.
- Discretização de dados
  - Uma forma de redução de dados, mas com interesse particular, especialmente para dados numéricos.

Limpeza dos dados

- Preencher valores faltantes;
- Identificar outliers e remover ruídos nos dados;
- Corrigir e eliminar inconsistências.
- Remover redundâncias causadas pela integração de dados.

Valores faltantes

- Em muitos casos, dados podem ser incompletos.
- Valores faltantes ocorrem devido:
  - Problemas com equipamentos (perdas de dados);
  - Inconsistência com outros registros e portanto são deletados;
  - Dados não digitados por causa de má interpretação;
  - Alguns dados não são importantes no momento da entrada;
  - Falta de registros históricos ou mudança nos dados.
- Em muitos casos, valores faltantes podem ser inferidos.

Lidando com valores faltantes

- Método 1: Ignorar as observações (registros):
- A alternativa mais simples.
- Deve ser usado somente se a observação possui vários atributos com valores faltantes.
- É um método ineficiente:
  - Parte da informação é perdida;
  - É um método pobre quando a porcentagem de valores faltantes varia entre os atributos.

Lidando com valores faltantes

- Método 2: Preencher os valores manualmente.
- Essa alternativa só vale a pena se o dataset for muito pequeno.
- Ineficiência desse método:
  - Consome muito tempo;
  - Impraticável para grandes datasets.

Lidando com valores faltantes

- Método 3: Usar a média do atributo para preencher os valores faltantes.
  - Exemplo: se idade média de um grupo de pessoas é 45, esse valor deve ser usado para preencher os valores faltantes.
- Vantagem:
  - Procedimento simples de ser implementado.

Lidando com valores faltantes

- Método 4: Para atributo nominal, usar a moda para preencher os valores faltantes.
- A moda é o valor mais frequente em um conjunto de valores.
- Pode não ser uma boa alternativa quando o atributo considerado é o atributo classe.

Lidando com valores faltantes

- Método 5: Usar a média para observações pertencentes a uma mesma classe.
  - Nesse caso, o valor faltante não está no atributo classe.
- Exemplo: se um cliente não possui informação sobre o consumo mensal de cartão de crédito, substitua o valor faltante pela média de consumo de clientes na categoria (mesma classe).
- Em caso de atributo nominal (não-classe), use a moda do atributo considerando as observações que pertencem a mesma classe.

Lidando com valores faltantes

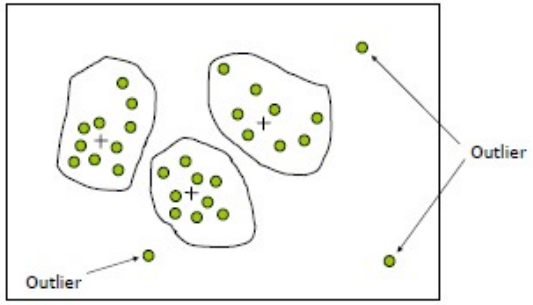
- Método 6: Usar o valor mais provável que é baseado em inferência.
- Exemplo: Determinar o valor faltante usando uma árvore de decisão, um modelo Bayesiano, etc.
- O método é muito eficiente, mas é também muito caro computacionalmente.

Ruído nos dados

- Ruído: erro aleatório ou variância nos valores de uma determinada variável.
- Valores incorretos de atributos podem ocorrer devido:
  - Falhas nos equipamentos de coleta de dados;
  - Problemas na entrada de dados;
  - Problemas na transmissão de dados;
  - Limitação tecnológica;
  - Inconsistência na convenção de nomes;
  - Transformações erradas aplicadas aos dados.

Ruído nos dados

- Outliers são objetos com características diferentes da maioria dos outros objetos em um conjunto de dados.
- Podem ser detectados por meios de agrupamentos (clusters). Intuitivamente, objetos que estão fora dos clusters são outliers.

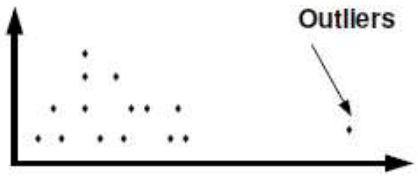


Análise de Outliers

- O que pode produzir outliers:
  - Erros de medição;
  - Valores default assumidos durante o preenchimento de uma base de dados (para o campo salário o default pode ser 0,00)
  - Podem corresponder a valores reais mas pertencentes a uma base de dados desbalanceada.

Análise de Outliers

- Outliers de uma Variável
  - É uma valor (ou valores) com ocorrência de baixa frequencia localizado longe das maiores concentrações dos valores da variável.
  - A grande questão é saber se os *"outliers" são um erro.*
- Pois estes, por exemplo, podem distorcer a resposta de uma rede neural.



Análise de Outliers

- A detecção de outliers não é um processo trivial.

- Método estatístico para remover outliers:

$$\text{Limiar} = \text{Média} \pm 2 * \text{Desvio padrão}$$

Base de dados desbalanceadas

- Uma bases de dados é considerada desbalanceada se existe nela objetos (instâncias) pertencentes a uma classe, em menor número em relação a outra.

- É considerado base desbalanceada se a relação do número de objetos das classes são da ordem de 1:100,1:1000 ou 1:10000.

Base de dados desbalanceadas

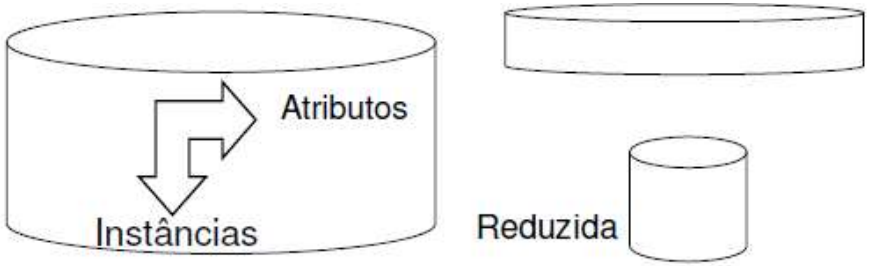
- Para solucionar o problema de classes desbalanceadas, pode-se utilizar algoritmos do tipo:
  - Undersampling: reduzem os objetos da classe majoritária em quantidade suficiente (comparada à quantidade da classe minoritária).
    - Ex: reduzir aleatoriamente (rand-under); dividir em clusters e escolher o centróide do cluster (M-clus).
  - Oversampling: aumentam os objetos da classe minoritária de forma a ser tornar uma base balanceada.
    - Ex: duplicar registros aleatórios (rand-over); criar registros a partir de exemplos vizinhos (smote).

Redução de dados

- Uma base de dados é considerada gigantesca se esta possui duas características: alta dimensionalidade e grande número de registros.
- Para gigantescas bases de dados, pode ser necessária uma etapa de redução de dados, antes de aplicar as técnicas de Data Mining.
- Enquanto grandes bases de dados tem potencial para melhorar os resultados da mineração, não existe garantia que estas levem para um melhor conhecimento extraído, que as bases com menos dados.

Redução de dados

- Um grande número de *atributos* (várias centenas) exige um grande número de ***instâncias***.
- Se existem poucas centenas de instâncias é necessária a redução da dimensionalidade para tornar possível a mineração e proporcionar algum uso prático ao conhecimento extraído.



Referências

- RUSSEL, S., NORVIG, P. *Inteligência Artificial*, Editora Campus, 2ª. edição.
- OLIVEIRA, S. R. M. *Introdução à mineração de dados*, Material para aulas, 2012.
- ZARATE, L.E. *Descoberta de Conhecimento em Banco de Dados e Data Mining*, Material para aulas, 2008.
- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, MIT, 1996.