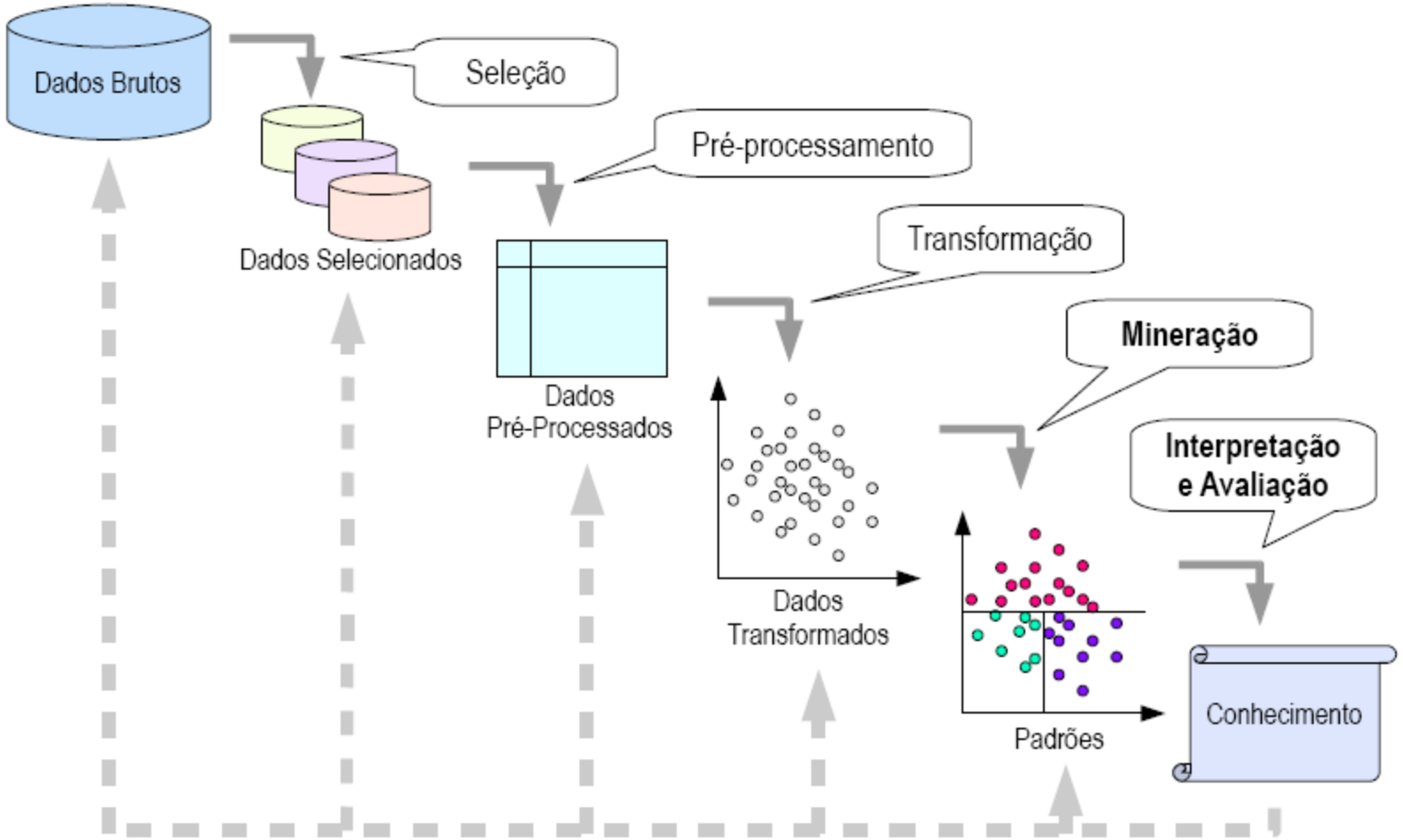
Introdução à Mineração de Dados

Leandro Guarino de Vasconcelos

KDD - *Knowledge Discovery in Databases*

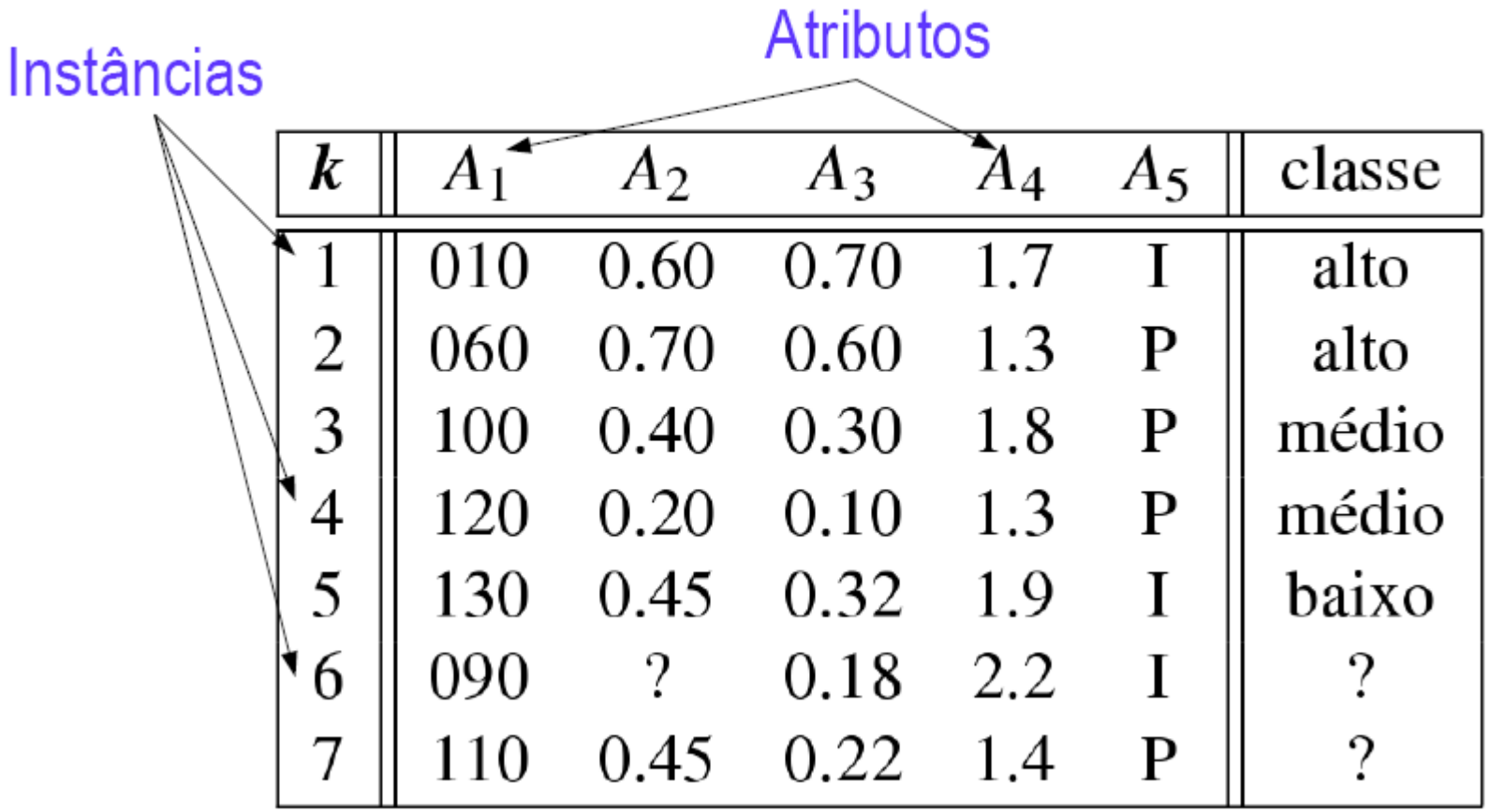
- **KDD: Processo geral de descoberta de conhecimentos úteis** previamente desconhecidos a partir de grandes bancos de dados (adaptado de Fayyad *et al).*
- O **principal foco** do KDD é automatizar o processamento de dados, permitindo que os usuários sejam mais eficientes na análise dos dados e encontrem fatos e relações entre os dados.

http://www.lac.inpe.br/~rafael.santos/



Exemplo

http://www.lac.inpe.br/~rafael.santos/



k	A_1	A_2	A_3	A_4	A_5	classe
1	010	0.60	0.70	1.7	I	alto
2	060	0.70	0.60	1.3	P	alto
3	100	0.40	0.30	1.8	P	médio
4	120	0.20	0.10	1.3	P	médio
5	130	0.45	0.32	1.9	I	baixo
6	090	?	0.18	2.2	I	?
7	110	0.45	0.22	1.4	P	?

Definições

- Dados em uma única tabela.
- Cada linha na tabela é uma **instância ou amostra (registros).**
- Cada coluna na tabela é um **atributo (campos).**
- Podemos ter vários tipos de atributos.
- Cada instância da base de dados tem os mesmos campos e que cada campo tem o mesmo tipo de valor.
- Eventualmente um atributo para uma instância pode ser desconhecido ou estar faltando.

k	A_1	A_2	A_3	A_4	A_5	classe
1	010	0.60	0.70	1.7	I	alto
2	060	0.70	0.60	1.3	P	alto
3	100	0.40	0.30	1.8	P	médio
4	120	0.20	0.10	1.3	P	médio
5	130	0.45	0.32	1.9	I	baixo
6	090	?	0.18	2.2	I	?
7	110	0.45	0.22	1.4	P	?

- Existe algum padrão? Existe algo fora de um padrão?
- Quais atributos influenciam nas classes?
- Podemos escolher a classe em função dos valores dos atributos?
- Podemos prever o valor de um atributo em função de outros?

Técnicas

- **Classificação: aprendizado de uma função que pode ser** usada para mapear dados em uma de várias classes discretas definidas previamente.
  - A classe é **alto se A1 < 70 e A2 > 0.5**
- **Regressão ou Predição: aprendizado de uma função que** pode ser usada para mapear os valores associados aos dados em um ou mais valores reais.
  - A3 pode ser calculado em função de A2?

Técnicas

- **Agrupamento (ou *clustering):* *identificação de grupos de*** dados onde os dados tem características semelhantes aos do mesmo grupo e onde os grupos tenham características diferentes entre si.
- **Sumarização: descrição do que caracteriza um conjunto de** dados (ex. conjunto de regras que descreve o comportamento e relação entre os valores dos dados).

Técnicas

- **Detecção de desvios ou *outliers*: identificação de dados** que deveriam seguir um padrão esperado mas não o fazem.
- **Regras de associação: identificação de grupos de** dados que apresentam co-ocorrência entre si (ex. cesta de compras).

Técnicas

- **Redes neurais artificiais:** são usadas principalmente em:
  - **Classificação:** treinando uma RNA, a saída pode ser usada como um vetor característico. **Exemplo:** classificação de risco de cliente em um banco. A partir de dados sobre os clientes, a classificação resultado em “risco” e “não-risco”.
  - **Predição:** dado um conjunto de treinamento, uma RNA pode ser treinada para representar a função aproximada que modela uma situação.

Técnicas

- **Árvore de decisão:** é uma estrutura de dados em que a busca dos dados classifica-os de acordo com um critério de seleção de cada ramo.
- **Sequential Pattern:** dado um conjunto de sequência em que cada sequência consiste de uma lista de elementos e cada elemento consiste de um conjunto de itens, a mineração de padrões sequenciais encontra subsequências frequentes.
- **Contiguous Sequential Pattern (CSP):** tipo específico de padrões sequenciais em que os itens que aparecem na sequência devem ser adjacentes (é definida uma ordem na sequência).

Web Mining

- Web mining não é uma tarefa trivial, considerando que a web é uma enorme coleção de dados heterogêneos, não-rotulados, distribuídos, variáveis no tempo, semi-estruturados e multidimensionais.
- **Classificação:**
  - **Web content mining:** o conteúdo da página web é usado como entrada para os algoritmos.
  - **Web structure mining:** usa a estrutura dos links como entrada para os algoritmos.
  - **Web usage mining:** usa client logs ou server logs para analisar o comportamento do usuário. Os dados podem ser complementados com histórico de compras, páginas que o usuário visitou, etc.

Aplicações de KDD na Web

- Próxima página a ser visitada
- Recomendação de páginas para visitar
- Recomendação de produtos em e-commerce
- Classificação de perfis de usuários
- Classificação de páginas Web
- Personalização Web (*Web Personalization*)

Web Personalization

- É qualquer ação que adapta a informação ou os serviços fornecidos pelo web site às necessidades de um usuário ou conjunto de usuários, tomando vantagem do conhecimento adquirido sobre o comportamento dos usuários nas interfaces e seus interesses individuais, combinando com o conteúdo e a estrutura do web site.
- Como menciona Mulvenna (2000), “O objetivo de um sistema de personalização Web é fornecer aos usuários a informação que eles querem ou precisam, sem esperar que eles perguntem explicitamente por ela”.

Recomendação de Conteúdo na Web

- Para a recomendação de conteúdo na Web, um método é a definição de **RULES**.

- Uma RULE é:
  - Se < condição > Então < recomendação >

- **Exemplos de RULES**
- R1: If VisitPage(p1) and SpentTime(t1) Then RecommendationPage(p10)
- R2: If CountPageVisit(pi)< D Then DeletePage(pi)

Referências

- Velásquez, J. D. & Palade, V. Jain, L. & Howlett, R. *(Eds.)* Adaptive Web Sites: A Knowledge Extraction from Web Data Approach
  IOS Press, 2008*.*
- Maria Paula Gonzalez, Jesus Lores, A. G. Enhancing usability testing through datamining techniques: A novel approach to detecting usability problem patternsfor a context of use, 2007.
- Chen, J. & Cook, T. Mining Contiguous Sequential Patterns from Web Logs. *Proceedings of the 16th International Conference on World Wide Web, ACM,* **2007**, 1177-1178.
- http://www.lac.inpe.br/~rafael.santos/
- Nagy, I. K. & Gaspar-Papanek, C. User Behaviour Analysis Based on Time Spent on Web Pages