

USP

LABIO

Data Mining (Etapas)

Solange Oliveira Rezende

USP São Carlos / ICMC
Departamento de Ciências de Computação e Estatística
Laboratório de Inteligência Computacional
<http://www.icmc.usp.br>

© Solange Oliveira Rezende

USP

LABIO

Extração de Conhecimento de Bases de Dados

- Extração de Conhecimento de Bases de Dados é um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados. [Fayyad-96]
- Área multidisciplinar
- Processo iterativo e iterativo

SOF

USP

LABIO

Data Mining

Processo de identificação de padrões válidos, inovadores, potencialmente úteis e, principalmente, compreensíveis em conjuntos de dados.
(Fayyad et al., 1996)

SOF

USP

LABIO

Processo Data Mining

```

graph TD
    A[IDENTIFICAÇÃO DO PROBLEMA] --> B[PRÉ-PROCESSAMENTO]
    B --> C[EXTRAÇÃO DE PADRÕES]
    C --> D[PÓS-PROCESSAMENTO]
    D --> E[UTILIZAÇÃO DO CONHECIMENTO]
    E --> A
  
```

SOF

USP

LABIO

Processo de Data Mining

IDENTIFICAÇÃO DO PROBLEMA

- Estudo do domínio da aplicação;
- Definição dos objetivos e metas a serem alcançados;
- Identificação e seleção dos conjuntos de dados.

SOF

USP

LABIO

Identificação do Problema

- Antes de iniciar o processo é imprescindível a obtenção de algum conhecimento inicial do domínio.
- Questões importantes devem ser respondidas:
 - Quais são as principais metas do processo?
 - Quais fontes de dados serão utilizadas?
 - Quais critérios de performance são importantes?
 - Qual deve ser a relação entre simplicidade e precisão do conhecimento extraído?
- As informações obtidas nessa etapa fornecem subsídios para todas as etapas seguintes do processo.

SOF



USP **LABIO**

Pré-Processamento

- Geralmente, os dados utilizados no processo não estão adequados para serem utilizados na etapa de Extração de Padrões.
- Os dados podem apresentar diversos problemas:
 - Ruído
 - Dados incompletos
 - Formato inadequado
 - Grande volume
- O Pré-Processamento consiste na aplicação de técnicas e métodos com o objetivo de adequar os dados para serem utilizados na etapa de Extração de Padrões.

SOF

USP **LABIO**

Seleção

- Escolhe ou segmenta os dados de acordo com alguns critérios.
 - Exemplo: todas as pessoas que possuem carro zero.
- Subconjunto de dados são determinados nesta etapa.

SOF

USP **LABIO**

Pré-processamento

- Nesta etapa é realizada a limpeza dos dados.
 - Informações podem ser removidas quando julgadas desnecessárias;
 - Os dados podem ser reconfigurados para assegurar a consistência de formatos uma vez que ele podem ser oriundos de várias fontes.
 - Exemplo: o sexo pode ser representado por f ou m , bem como por 0 ou 1.

SOF

USP **LABIO**

Transformação

- Nesta etapa os processos não são apenas transferência de dados.
- Em algumas ocasiões torna-se necessário acrescentar outras informações pertinente ao domínio da aplicação.
 - Exemplo: CEP, comumente usado nas pesquisas de mercado.

SOF

USP **LABIO**

Pré-Processamento

- Obtenção e Integração**
 - Os dados podem estar em diferentes formatos, como arquivos texto, arquivos no formato MS EXCEL, banco de dados relacionais, *Datawarehouse*.
 - Transformação para o formato atributo-valor.

	X_1	X_2	...	X_m	Y
T_1	x_{11}	x_{12}	...	x_{1m}	y_1
T_2	x_{21}	x_{22}	...	x_{2m}	y_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
T_l	x_{l1}	x_{l2}	...	x_{lm}	y_l

SOF

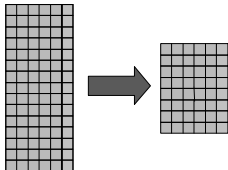
USP

LABIO

Pré-Processamento

■ Redução do volume de dados

- Limitações de espaço em memória, tempo de processamento, entre outras, podem inviabilizar o uso de algoritmos de extração de padrões.
- A redução pode ser realizada de três formas:
 - Número de exemplos



SOF

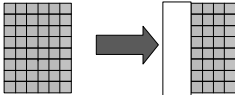
USP

LABIO

Pré-Processamento

■ Redução do volume de dados

- Limitações de espaço em memória, tempo de processamento, entre outras, podem inviabilizar o uso de algoritmos de extração de padrões.
- A redução pode ser realizada de três formas:
 - Número de exemplos
 - Número de atributos



SOF

USP

LABIO

Pré-Processamento

■ Redução do volume de dados

- Limitações de espaço em memória, tempo de processamento, entre outras, podem inviabilizar o uso de algoritmos de extração de padrões.
- A redução pode ser realizada de três formas:
 - Número de exemplos
 - Número de atributos
 - Número de valores de um atributo
 - Discretização

$$\begin{array}{l}
 \text{atr} \\
 1 \\
 1 \\
 2 \\
 3 \\
 3 \\
 3 \\
 4 \\
 5 \\
 5 \\
 7
 \end{array}
 \begin{array}{l}
 \\
 A \\
 \\
 B \\
 \\
 C
 \end{array}$$

A se $\text{atr} < 2,5$
 B se $2,5 \leq \text{atr} < 3,5$
 C se $3,5 \leq \text{atr}$

SOF

USP

LABIO

Pré-Processamento

■ Redução do volume de dados

- Limitações de espaço em memória, tempo de processamento, entre outras, podem inviabilizar o uso de algoritmos de extração de padrões.
- A redução pode ser realizada de três formas:
 - Número de exemplos
 - Número de atributos
 - Número de valores de um atributo
 - Discretização
 - Suavização

atr	Valor	mediano
1	1	1
1	1	1
2	1	1
3	3	3
3	3	3
3	3	3
4	5	5
5	5	5
5	5	5
7	5	5

SOF

USP

LABIO

Processo de *Data Mining*



IDENTIFICAÇÃO DO PROBLEMA

- Escolha da função:
 - descritiva ou preditiva
- Escolha do algoritmo:
 - algoritmo e parâmetros
- Transformação dos dados
- Obtenção de padrões:
 - aplicação do algoritmo aos dados

SOF

USP

LABIO

Data Mining

- DM, então, é responsável por encontrar padrões, modelos ou classificações dentro do conjunto de dados.
- Os modelos gerados por DM seguem (geralmente) os padrões estatísticos, neurais ou simbólicos.

SOF

USP

LABIO

Extração de Padrões

- A etapa de Extração de Padrões visa escolher, configurar e executar um ou mais algoritmos de extração de padrões, a fim de cumprir os objetivos do processo
- Essa etapa pode ser executada diversas vezes para ajustar os parâmetros dos algoritmos e dessa forma obter um resultado mais adequado
- Sub-etapas:
 - Escolha da Função
 - Escolha do Algoritmo
 - Transformação dos Dados
 - Extração dos Padrões

SOF

USP

LABIO

Técnicas usadas em Data Mining

- Classificação
- Regras de Associação
- Caracterização
- Predição
- Clustering (Categorização)
- Discriminação
- Evolução
- Desvio

SOF

USP

LABIO

Escolha das Atividades de Data Mining

- Os algoritmos para Extração de conhecimento de dados podem ser divididos em dois grandes grupos referentes as:
 - Atividades Preditivas (Aprendizado Supervisionado)
 - Atividades Descritivas (Aprendizado Não-Supervisionado)

SOF

USP

LABIO

Extração de Padrões

Escolha da Função

```

graph TD
    DM([Data Mining]) --> AP([Atividade Preditiva])
    DM --> AD([Atividade Descritiva])
    AP --> C([Classificação])
    AP --> R([Regressão])
    AD --> RA([Regras de Associação])
    AD --> CL([Clustering])
    AD --> S([Sumarização])
  
```

SOF

USP

LABIO

Extração de Padrões

Escolha da Função

Atividade Preditiva

Classificação

Reg

Definição: Consiste na generalização de exemplos com suas respectivas classes conhecidas em uma linguagem capaz de reconhecer a classe de um novo exemplo.

Exemplo:

X	X	C
1	3	a
3	5	a
7	6	b
8	6	b

Regras:

X2=5 -> C=a

X1=6 -> C=b

Classificação de um novo exemplo

X1=4, X2=5 e C=?

C=a

Aprendizado Supervisionado

SOF

USP

LABIO

Extração de Padrões

Escolha da Função

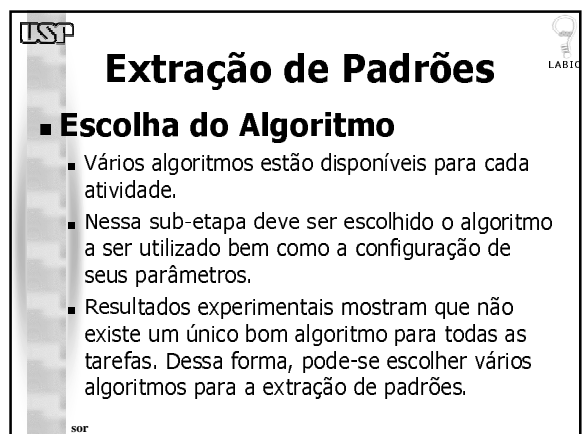
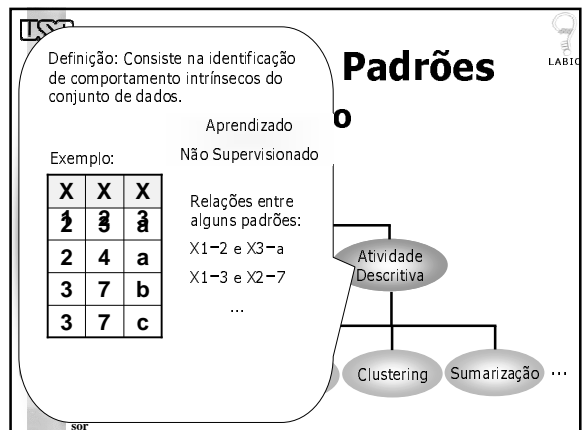
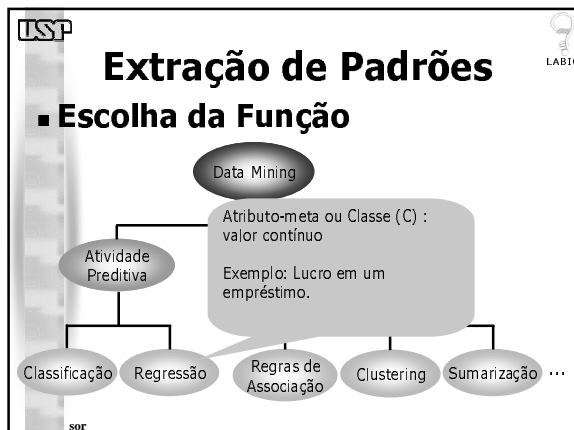
```

graph TD
    DM([Data Mining]) --> AP([Atividade Preditiva])
    DM --> AD([Atividade Descritiva])
    AP --> C([Classificação])
    AP --> R([Regressão])
    AD --> RA([Regras de Associação])
    AD --> CL([Clustering])
    AD --> S([Sumarização])
  
```

Atributo-meta ou Classe (C) : valor categórico

Exemplo: Se um cliente é bom ou ruim pagador.

SOF



USP **LABIO**

Escolha de Algoritmos

Quando se trata de algoritmos de AM, algumas características podem ser analisadas para a seleção de algoritmo.

Tipo de Aprendizado	Paradigmas de Aprendizado	Linguagens de Descrição	Integração de Novos exemplos
- Supervisionado	- Simbólico	- Exemplos ou Instâncias	- Incremental
- Não-supervisionado	- Estatístico	- Hipóteses ou Conceitos	- Não Incremental
	- Instance-Based	- Teoria de Domínio ou Conhecimento de Fundo	
	- Conexionista		
	- Genético		

SOF

USP **LABIO**

Extração de Padrões

■ Escolha do Algoritmo

Algumas características podem ser analisadas quando os algoritmos que serão selecionado forem algoritmos de AM.

Tipo de Aprendizado	Paradigmas de Aprendizado	Linguagens de Descrição	Integração de Novos Exemplo
•Supervisionado	•Simbólico	•Exemplos ou Instâncias	•Incremental
•Não-supervisionado	•Estatístico	•Hipóteses ou Conceitos	•Não Incremental
	•Instance-Based	•Teoria de Domínio ou Conhecimento de Fundo	
	•Case-Based		
	•Conexionista		
	•Evolutivo		

SOF

USP

LABIO

Pós-Processamento

- Algumas necessidades após a extração de padrões:
 - Filtrar o conhecimento extraído : Poda ou Truncagem.
 - Documentar ou modificar o conhecimento para torna-lo compreensível, possibilitando o seu uso em Sistemas Inteligentes ou como apoio em processos de tomada de decisão.
 - Se utilizado mais de um algoritmo, o conhecimento extraído pode ser combinado visando obter uma maior precisão e um melhor desempenho do sistema em sua totalidade.

SOF

USP

LABIO

Pós-Processamento

- Existem diversas medidas para avaliar o conhecimento extraído

```

graph TD
    A([Medidas de Avaliação]) --> B([Qualidade])
    A --> C([Desempenho])
    B --> D([Compreensibilidade])
    B --> E([Interessabilidade])
    E --> F([Objetivas])
    E --> G([Subjetivas])
          
```

SOF

USP

LABIO

Processo de Data Mining

```

graph TD
    A[IDENTIFICAÇÃO DO PROBLEMA] --> B[PRÉ-PROCESSAMENTO]
    B --> C[EXTRAÇÃO DE PADRÕES]
    C --> D[PÓS-PROCESSAMENTO]
    D --> A
    D --> E[Utilização do Conhecimento]
    E --> A
          
```

Processo Iterativo e Iterativo

SOF

USP

LABIO

Processo de Data Mining

```

graph TD
    A[IDENTIFICAÇÃO DO PROBLEMA] --> B[PRÉ-PROCESSAMENTO]
    B --> C[EXTRAÇÃO DE PADRÕES]
    C --> D[PÓS-PROCESSAMENTO]
    D --> A
    D --> E[Utilização do Conhecimento]
    E --> A
          
```

SOF

USP

LABIO

Processo Data Mining

```

graph TD
    A[IDENTIFICAÇÃO DO PROBLEMA] --> B[PRÉ-PROCESSAMENTO]
    B --> C[EXTRAÇÃO DE PADRÕES]
    C --> D[PÓS-PROCESSAMENTO]
    D --> A
    D --> E[Utilização do Conhecimento]
    E --> A
          
```

SOF

USP

LABIO

Processo KDD

```

graph TD
    DADOS[DADOS] --> SELEÇÃO[SELEÇÃO]
    SELEÇÃO --> DADO_ANALISADO[DADO ANALISADO]
    DADO_ANALISADO --> DADO_PROCESSADO[DADO PROCESSADO]
    DADO_PROCESSADO --> TRANSFORMAÇÃO[TRANSFORMAÇÃO]
    TRANSFORMAÇÃO --> DADO_TRANSFORMADO[DADO TRANSFORMADO]
    DADO_TRANSFORMADO --> DATA_MINING[DATA MINING]
    DATA_MINING --> PADRÕES[PADRÕES]
    PADRÕES --> INTERPRETACAO_AVALIACAO[INTERPRETAÇÃO/ AVALIAÇÃO]
    INTERPRETACAO_AVALIACAO --> CONHECIMENTO[CONHECIMENTO]
          
```

FAYYAD 1996

SOF

