
Cap. 3: Pré-processamento de Dados

Alunos: Gabriel Luciano, Geovane Fonseca,
Luigi Domenico e Paulo Junio

Eliminação Manual de Atributos

Motivo: Nem todos os atributos são necessários para resolver o problema

Solução:

- Definição feita por especialistas no domínio dos dados
- Eliminar atributos irrelevantes
- Eliminar atributos invariáveis

Integração de Dados

Motivo: Dados a serem utilizados estão distribuídos em diferentes estruturas de dados

Solução:

- Identificar os objetos presentes nos diferentes conjuntos
- Utilizar metadados
- Criar um depósito ou repositório de dados

Amostragem de Dados

Motivo: Dificuldade em lidar com número grande de objetos

Solução:

- Amostragem Aleatória Simples
- Amostragem Estratificada
- Amostragem progressiva

Dados Desbalanceados

Motivo: Possível favorecimento de classificação de novos dados na classe majoritária

Solução:

- Redefinir o tamanho do conjunto de dados
- Utilizar diferentes custos de classificação
- Induzir modelo para uma classe

Limpeza de Dados

Motivo: Dificuldades relacionadas à qualidade dos dados (e.g. dados ruidosos, inconsistentes, redundantes ou incompletos)

Solução:

- Dados Incompletos:
 - Eliminar os objetos com valores ausentes
 - Preencher manualmente os valores dos atributos
 - Heurística para definir os valores

Limpeza de Dados

- Dados Inconsistentes
 - Algoritmo simples
 - Remoção manual (caso tenha poucos dados)
- Dados Redundantes
 - Remoção de objetos semelhantes
- Dados com ruídos
 - Técnicas baseadas em agrupamento
 - Técnicas baseadas em distância

Transformação de Dados

Motivo: Algumas técnicas de AM estão limitadas a valores de determinados tipos

Solução:

- Conversão Simbólico-Numérico
- Conversão Numérico-Simbólico
- Transformação de atributos numéricos

Redução de Dimensionalidade

Motivo: Número elevado de atributos para serem tratados pelo AM

Solução:

- Agregação
- Seleção de Atributos
- Técnicas de Ordenação
- Técnicas de Seleção de Subconjunto