DEPARTAMENTO DE COMPUTAÇÃO
E TECNOLOGIA

Sistemas de Informação
UFRN

Preparação dos Dados

Karliane Vale (karliane.vale@ufrn.br)
Huliane Medeiros (hulianeufrn@gmail.com)

Sumário

- Conjunto de Dados
- Análise de Dados

Conjunto de Dados

- A cada dia uma enorme quantidade de dados é gerada.
- Oriundos de transações financeiras, monitoramento ambiental, obtenção de dados clínicos, captura de imagens, navegação na internet, etc.
- Dados podem estar em formatos diferentes: séries temporais, páginas na web, imagens, vídeos, áudios, textos, grafos, redes sociais, entre outros.
- Com o crescente aumento da quantidade de dados gerada, tem aumentado muito a distâncias entre a quantidade de dados existentes e a porção desse dados que é analisada e compreendida.

Conjunto de Dados

- Conjuntos de dados são formados por objetos que podem representar um objeto físico ou abstrato. Por exemplo?
- Cada objeto é descrito por uma conjunto de atributos (características)
- Cada objeto corresponde a uma ocorrência dos dados
- Cada atributo está associado a uma propriedade do objeto

Conjunto de Dados

- Apesar do grande número de bases de dados, na maioria das vezes não é possível aplicar os algoritmos de AM diretamente sobre elas.
- Técnicas de pré-processamento de dados são usadas para tornar os dados mais adequados para o uso de algoritmos de AM.
- Essas técnicas podem ser agrupadas em:
  - Eliminação manual de atributos
  - Integração de dados
  - Amostragem de dados
  - Balanceamento de dados
  - Limpeza de dados
  - Redução de dimensionalidade
  - Transformação de dados

Exemplos de situações

- Empresas do governo ou outras instituições têm seus dados armazenados em mais de uma base de dados ou conjunto de dados.
- Dados podem vir de mais de uma fonte.
- Precisam ser utilizados por algoritmos de AM, os conjunto de dados precisam ser integralizados de forma a construir um único conjunto ou tabela.
- Pode levar a inconsistências e redundâncias.

Exemplos de situações

- Os algoritmos podem ter dificuldade em lidar com conjuntos de dados grandes.
  - Quantidade pode estar relacionada a número de objetos, atributos ou ambos
- Problemas como redundância e inconsistência muitas vezes estão relacionados a grande quantidade de dados

- Técnicas de amostragem e seleção de atributos podem ser usadas

Exemplos de situações

- Os algoritmos podem ter dificuldade em lidar com conjuntos de dados grandes.
  - Quantidade pode estar relacionada a número de objetos, atributos ou ambos
- Problemas como redundância e inconsistência muitas vezes estão relacionados a grande quantidade de dados

- Técnicas de amostragem e seleção de atributos podem ser usadas

Exemplos de situações

- A distribuição de objetos entre as classe pode não ser uniforme

- Alguns algoritmos têm dificuldades de induzir um bom modelo a partir de um conjunto de dados desbalanceados

- Boa parte dos conjuntos de dados apresentam problemas como a presença de ruídos e dados incompletos e/ou inconsistentes.
  - Dados incompletos por causa da ausência de valores
  - Inconsistentes por causa de erros na geração, captação ou entrada

- Isso afeta o desempenho de grande parte dos algoritmos de AM. Diversas técnicas de limpeza de dados têm sido propostas e investigadas.

Exemplos de situações

- Após a eliminação de atributos por especialistas, os atributos restantes pode dificultar a tarefas dos algoritmos de AM, devido a diferentes motivos:
  - Presença de um número grande de atributos
  - Atributos redundantes, irrelevantes e/ou inconsistentes

- Vários algoritmos têm dificuldades em trabalhar com os dados originais
  - Transformar os dados originais antes que sejam utilizados pelo algoritmos
  - Exemplo: transforma dados de valores simbólicos em dados numéricos.

Análise de Dados

- A análise das características presentes em um conjunto de dados permite a descoberta de padrões e tendências que podem fornecer informações valiosas que ajudem a compreender o processo que gerou os dados.

- Muitas dessas características podem ser obtidas por meio de aplicação de técnicas estatísticas simples

- Podem ser observadas por técnicas de visualização de dados

Caracterização dos Dados

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Caracterização dos Dados

- Os valores que um atributos pode assumir podem ser definidos de diferentes formas. Por exemplo: tipo e escala.

- Tipo diz respeito ao grau de quantização nos dados

- Escala indica a significância relativa dos valores

- Conhecer o tipo/escala dos atributos auxilia a identificar a forma adequada de preparar os dados e posteriormente de modelá-los

Tipo

- Define se o atributo representa quantidade (quantitativo ou numérico) ou qualidade (qualitativo, simbólico ou categórico)
- Exemplos de conjuntos de valores qualitativos:
  - {pequeno, médio, grande}
  - {física, matemática, química}
- Exemplo de conjunto de valores quantitativo:
  - {123, 45, 12}
  - Valores quantitativos pode ser contínuos ou discretos:
    - Contínuos: atributos que representam o peso, tamanho, distância
    - Discretos: atributos binários (0/1, sim/não, verdadeiros/falso, ausência/presença)

Tipo

Atributo	Classificação
Id.	Qualitativo
Nome	Qualitativo
Idade	Quantitativo discreto
Sexo	Qualitativo
Peso	Quantitativo contínuo
Manchas	Qualitativo
Temp.	Quantitativo contínuo
#Int.	Quantitativo discreto
Est.	Qualitativo
Diagnóstico	Qualitativo

Tipo

- Atributos quantitativo assumem valores binários, inteiros ou reais
- Atributos qualitativo são representados por número finito de símbolos ou nomes

Escala

- Define as operações que podem ser realizadas sobre os valores dos atributos

- Em relação a escala, os atributos podem ser: nominais, ordinais, intervalares e racionais.

- Os dois primeiros são qualitativo, os dois últimos são quantitativo

Escala nominal

- Os valores são apenas nomes diferentes, carregando a menos quantidade de informação possível

- Não existe relação de ordem entre seus valores

- Operações mais utilizadas: igualdade e desigualdade
  - Exemplo: atributos que representa continente do planeta, é possível apenas ver se dois valores são iguais ou diferentes. Caso queira ordenar, nesse caso seria do tipo ordinal.

Escala nominal

- Exemplos:
  - Nome de paciente
  - RG
  - CPF
  - Número da conta
  - CEP
  - Cores
  - Sexo

Escala ordinal

- Refletem uma ordem das categorias representadas
- Pode-se usar operadores relacionais
- Exemplo: atributo possui como valores pequeno, médio e grande, além dos valores serem categóricos, é possível definir se é igual, maior ou menor que outro.
- Atributos com escala ordinal: hierarquia militar e avaliações qualitativas de temperatura, como frio, quente e morno.

Escala intervalar

- Os atributos são representados por números que variam dentro de um intervalo
- Possível definir a ordem e a diferença em magnitude entre dois valores
- A diferença em magnitude indica a distância que separa dois valores no intervalos de possíveis valores

Escala racional

- São os que carregam mais informações. Os números têm um significado absoluto, ou seja, existe o zero absoluto junto com uma unidade de medida. A razão tem significado

- Por exemplo: número de vezes que uma pessoa vai ao hospital
  - No ponto zero = não foi nenhum vez
  - Um paciente teve duas internações, outro teve oito, corretos afirmar que o segundo teve internado 4 vezes mais que o primeiro

- Para esses atributos, faz sentido usar a razão entre dois valores

Escala racional

- Outros exemplos de atributos com escala de razão são tamanho, distância e valores financeiros, como salário e saldo em conta corrente.

Atributo	Classificação
Id.	Nominal
Nome	Nominal
Idade	Racional
Sexo	Nominal
Peso	Racional
Manchas	Nominal
Temp.	Intervalar
#Int.	Racional
Est.	Nominal
Diagnóstico	Nominal

Exploração de Dados

- Uma grande quantidade de informações úteis pode ser extraída de um conjunto de dados por meio de sua análise e exploração

- Informações obtidas na exploração podem ajudar na seleção da técnica mais apropriada para o pré-processamento dos dados e para o aprendizado

- Uma forma simples é a extração de medidas de uma área da estatística chamada estatística descritiva

Exploração de Dados

- A estatística descritiva resume de forma quantitativa as principais características de um conjunto de dados

- Muitas medidas são calculadas rapidamente

- Por exemplo: no conjunto de dados pacientes, duas medidas estatísticas podem ser facilmente calculadas: a idade média dos pacientes e a porcentagem de pacientes do sexo masculino.

Considerações Finais

- Antes de aplicar algoritmos de AM é importante que os dados sejam analisados
  - Técnicas estatísticas
  - Técnicas de visualização

- Permite melhor compreensão da distribuição dos dados e pode dá suporte a formas de abordar o problema.