Aula 3 – Tópicos Especiais em Computação: Agrupamento de Dados Representação dos Dados

Profa. Elaine Faria

UFU - 2020

Agradecimentos

Este material é baseado

- No livro Tan et al, 2006
- Nos slides do prof. Andre C. P. L. F. Carvalho

- Agradecimentos
  - Ao professor André C. P. L. F. Carvalho que gentilmente cedeu seus slides

O que é um conjunto de dados?

- Atributos
  - Descreve os objetos
  - Variável, característica, campo ou dimensão
- Objetos
  - Descritos por seus atributos
  - Registros, pontos, vetor, padrões, eventos, exemplos, observações, instâncias ou entidades
- Conjunto de dados (*data set* ou base de dados)
  - Coleção de objetos de dados

O que é um conjunto de dados?

- Conjunto de Dados
  - É um arquivo no qual os objetos são os registros (linhas) do arquivo
  - Cada campo (coluna) corresponde a um atributo

Matricula	Período	Nota
1010	1	95
1011	1	66
2010	2	55
2012	2	87
...	...	...
7010	7	72

Conjunto de dados sobre estudantes

Atributos e Medidas

- Atributo
  - É uma propriedade ou característica de um objeto
  - Pode variar de um objeto para outro ou de tempos em tempos
  - Ex: cor dos olhos → símbolo (azul, preto,...)

    temperatura → número (35, 30,...)

Atributos e Medidas

Atributos de entrada

	Nome	Temp.	Idade	Peso	Altura
Exemplos (objetos, padrões)	João	37	70	94	190
	Maria	38	65	60	172
	José	39	19	70	185
	Sílvia	38	25	65	160
	Pedro	37	70	90	168

Figura adaptada dos slides do prof. André C. P. L. F. Carvalho – disciplina Mineração de Dados em Biologia Molecular – ICMC-USP

Atributos e Medidas

Na tabela de informações sobre alunos podemos aplicar a função média nos atributos matrícula e nota?

Tipos de Atributos

- Categóricos (qualitativo)
  - Nominal
    - Valores são apenas nomes diferentes
    - Ex: sexo ({M,F} = {0,1}),
    - Ex: cor do olho ({preto,azul, verde} = {0,1,2})
  - Ordinal
    - Fornecem informações para ordenar os objetos
    - Ex: tamanho: ({pequeno, médio e grande}={1,2,3})

Tipos de Atributos

- Numérico (quantitativo)
  - Intervalar
    - As diferenças entre valores fazem sentido
    - Interpretação depende de uma unidade de medida, cujo zero é arbitrário
    - Ex: temperatura em graus Celsius, datas
  - Racional
    - Razão e diferença entre valores fazem sentido
    - Ex: temperatura em Kelvin, quantidades monetárias, idade e altura

Tipos de Atributos

- Propriedades:
  - Distinção: = e ≠
  - Ordem: > e <
  - Adição: + e –
  - Multiplicação: * e /

  - Atributo Nominal: distinção
  - Atributo Ordinal: distinção e ordem
  - Atributo Intervalar: distinção, ordem e adição
  - Atributo Racional: as 4 propriedades

Exemplos

Nome	Temp	Enjôo	Mancha	Dor	Salário	Diagnóstico
João	37.7	sim	pequena	sim	1000	doente
Pedro	37.0	não	pequena	não	1100	saudável
Maria	38.2	sim	grande	não	600	saudável
José	39.0	não	pequena	sim	2000	doente
Ana	37.3	não	grande	sim	1800	saudável
Leila	37.7	não	grande	sim	900	doente

Nominal Intervalar Ordinal Racional

Figura retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Mineração de Dados em Biologia Molecular – ICMC-USP

Exercício

- Defina o tipo dos seguintes atributos
  - Renda mensal
  - Número de palavras de um texto
  - Número de RG
  - Data de nascimento
  - Código de disciplina
  - Posição em uma corrida

Quantidade de valores de um atributo

- Discreto
  - Tem um número finito de valores ou um conjunto infinito enumerável de valores
  - Ex: CEP e ID
  - **Binário**
    - Tipo especial de atributo discreto
    - Assume somente dois valores
    - Ex: true/false, sim/não ou 0/1
- Contínuo
  - Possui valores numéricos reais
  - Ex: temperatura, peso, altura

Atributos Assimétricos

- Caso especial de atributo discreto
- Somente a presença de um dos valores é considerado importante
- Ex: base de dados de estudantes, vetor de atributos indica se o aluno cursa ou não uma disciplina
- Atributo binário assimétrico → atributos binários nos quais somente valores não-zero são importantes

Características de uma base de dados

- Dimensionalidade
  - Número de atributos
  - Bases de dados com alta dimensionalidade → dificuldades
  - Pré-processamento: redução da dimensionalidade
- Esparsidade
  - Conjuntos de dados no qual os objetos têm característica assimétrica
  - A maioria dos atributos do objeto tem valores 0, mas o que interessa são apenas os valore não-zeros

Características de uma base de dados

- Resolução
  - Diferentes níveis → diferentes resoluções e diferentes propriedades de dados
  - Ex: calcular a variação atmosférica na escala de horas ou meses, o primeiro consegue-se ver o movimento de tempestades, já o ultimo o fenômeno não é detectável

Qualidade dos Dados

- Dados usados na mineração de dados foram coletados para outros propósitos
- Problemas nos dados precisam ser corrigidos
- Algoritmos podem tratar problemas nos dados
  - Ex: tratamento de ruídos e *outliers*

Qualidade dos dados

- Problemas nos dados são devidos a
  - Erros humanos
  - Limitações dos dispositivos de medição
  - Falhas na coleta dos dados
- Exemplos de problemas nos dados
  - Atributo com valores ausente
  - Objetos espúrios ou duplicados
  - Valores incorretos

Qualidade dos Dados

- Causa dos erros nos dados
  - Erro de medição
    - O valor registrado difere do valor real
  - Erro de coleta
    - Omissão de objetos de dados ou valores dos atributos ou inclusão de objetos inapropriadamente
    - Erros de digitação

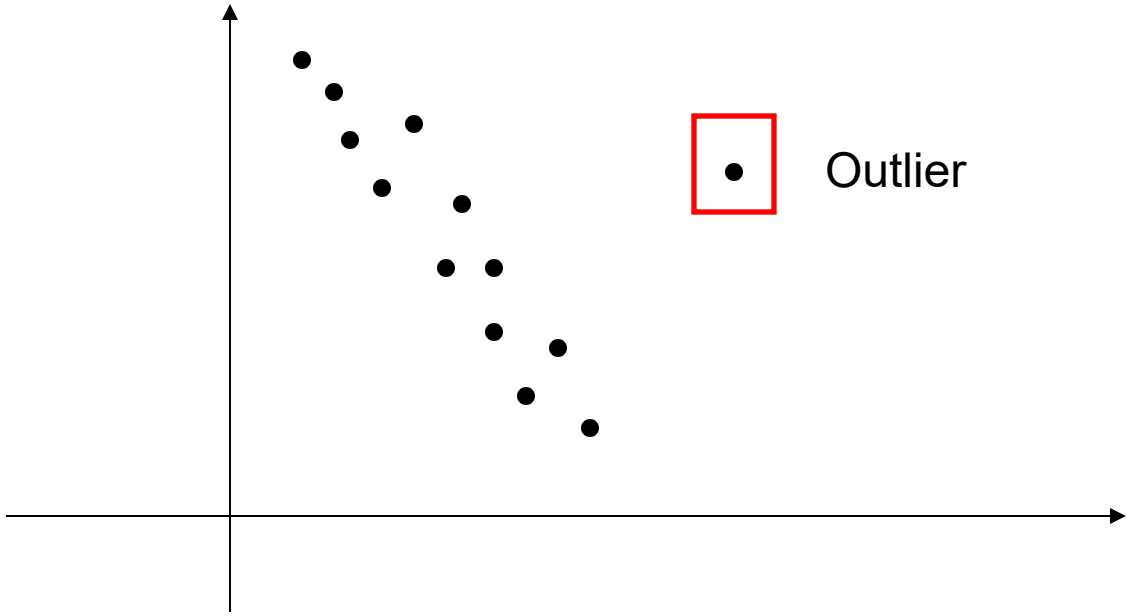
Qualidade dos Dados

- Ruídos
  - Componente aleatória de uma medida de erro
  - Envolve a distorção de um valor ou a adição de objetos espúrios
- Artefatos
  - Distorção determinística nos dados

Qualidade dos Dados

- *Outliers*
  - Objetos de dados que tem características que são diferentes da maioria dos outros objetos no conjunto de dados
  - Valores de um atributo que são não usuais em relação aos valores típicos para aquele atributos
  - Em alguns casos, os *outliers* podem ser de interesse
    - Ex: detecção de fraudes

Qualidade dos Dados



Qualidade dos Dados

- Valores Ausentes
  - Não é incomum objetos terem valores ausentes para um ou mais atributos
  - Causas
    - Informação não coletada
    - Atributo não é aplicável a todos os objetos
    - Problema na coleta

Qualidade dos Dados

- Estratégias para lidar com valores ausentes
  - Eliminar os objetos ou atributos com valores ausentes
    - Vantagens? Desvantagens?
  - Estimar os valores ausentes
    - Vantagens? Desvantagens?
  - Ignorar os valores ausentes
    - Vantagens? Desvantagens?
  - Modificar o algoritmo para lidar com valores ausentes
    - Vantagens? Desvantagens?

Qualidade dos Dados

- Ex: modificando o algoritmo de AM para ignorar valores ausentes
  - Agrupamento: similaridade entre pares de objetos precisam ser calculadas. Se um ou mais objetos tem valores ausentes para alguns atributos, então a similaridade pode ser calculada usando os outros atributos

Qualidade dos dados

- Instâncias duplicadas
  - Instâncias idênticas ou que diferem de forma não significativa
  - Ex: uma pessoa aparece duas vezes na base de dados com o campo nome com pequenas diferenças
  - Ex: duas instâncias com os mesmos valores

- Deduplicação
  - Processo de lidar com instâncias duplicadas
    - Detectar e tratar o problema

Qualidade dos dados

Nome	Idade	Altura	Peso
João	15	1,72	80
Maria	20	1,54	50
Pedro	15	1,72	80
Ana	32	1,60	58

Conhecer os dados é muito importante!!!!

Quais são os atributos?

Há dados duplicados?

Há valores ausentes?

Há ruídos?

Os *outliers* são de interesse para minha aplicação?

Pré-processamento de dados

- Pré-processamento
  - Tornar os dados mais adequados para a tarefa de mineração
  - Objetivo: melhorar a mineração com relação a tempo, custo e qualidade
  - Diferentes técnicas podem ser usadas
    - Agregação
    - Amostragem
    - Redução de dimensionalidade
    - Seleção de atributos
    - Discretização ou binarização
    - Transformação de variáveis

Agregação

- Combinar duas ou mais instâncias em um única instância
- Objetivo
  - Redução de dados
    - Menos memória e tempo de processamento
    - Uso de algoritmos mais “caros”
  - Comportamento mais estável
    - Quantidades agregadas tem menor variabilidade que objetos individuais
  - Desvantagem: perda de detalhes
- Ex: Agregar a soma de produtos vendidos nas cidades de uma região

Amostragem

- Selecionar um subconjunto das instâncias
- Técnica muito útil em MD

- Visão da Estatística
  - É caro e exige muito tempo obter todos os dados
- Visão da MD
  - É caro e consome muito tempo processar todos os dados

Amostragem

- Princípio básico
  - O uso da amostragem irá produzir resultados tão bons quanto usar o conjunto de dados inteiro, se a amostragem for **representativa**
  - Menor esforço computacional para processar os dados

- Amostragem representativa
  - Possui as mesmas propriedades (de interesse) da base de dados original
  - Ex: mesma média na base original e na amostra

Técnicas de Amostragem

- Principais técnicas
  - Aleatória
  - Estratificada
  - Progressiva

Técnicas de Amostragem

- Aleatória
  - Sem reposição
    - Cada instância selecionada é removida do conjunto de dados que constituem a população
  - Com reposição
    - Instâncias não são removidas da população quando ela são selecionadas
    - A mesma instância pode ser selecionada mais de uma vez
    - A probabilidade de selecionar um objeto se mantêm constante

Técnicas de Amostragem

- Estratificada
  - Usada para garantir que todas as classes do problema serão representadas
  - Variações
    - O mesmo número de instâncias de cada classe são selecionadas
    - O número de instâncias selecionadas de cada classe é proporcional ao número de instâncias da classe

Técnicas de Amostragem

- Progressiva
  - O tamanho da amostra é difícil de ser determinado
  - Começar com uma pequena amostra e aumentar o tamanho da amostra até que um tamanho suficiente seja encontrado

Como avaliar se o tamanho é suficiente?

Amostragem e Perda de Informação

- Como escolher o tamanho da amostra?
  - Amostras grandes
    - Aumentam a probabilidade de que a amostra será representativa
    - Eliminam muitas das vantagens da amostragem
  - Amostras pequenas
    - Aumenta a chance de perda de informação

Técnicas de Amostragem

8000 points

2000 points

500 points

Imagem reitrada de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Redução de Dimensionalidade

- Conjuntos de dados podem ter um grande nro de atributos
  - Ex: mineração de texto

  Reduzir a dimensionalidade pode ser a solução para trabalhar com conjuntos de dados contendo muitos atributos

Redução de Dimensionalidade

- Benefícios
  - Algoritmos trabalham melhor com poucos atributos
  - Redução de atributos irrelevantes e ruídos
  - Modelo mais compreensível
  - Dados são mais facilmente visualizados
  - Quantidade de tempo e memória usados pelo algoritmo de MD é reduzida

Redução de Dimensionalidade

- Extração de Atributos
  - Reduzir a dimensão criando novos atributos que são uma combinação dos atributos antigos
- Seleção de atributos
  - Reduzir a dimensão selecionando novos atributos que são um subconjunto dos atributos antigos

Maldição da dimensionalidade

- Fenômeno no qual a análise dos dados torna-se mais complicada com o aumento da dimensionalidade

- Dimensão aumenta → dados mais esparsos
  - Densidade e distância entre pontos tornam-se menos significativas
  - Qualidade dos clusters pode ser pobre

- Número de instâncias para manter o desempenho cresce com o nro de atributos

Seleção de Atributos

- Usar somente um subconjunto de atributos → reduz a dimensionalidade

- Atributos redundantes
  - Duplicar a informação contida em um ou mais atributos
- Atributo irrelevante
  - Não contém informação útil para a tarefa preditiva

Como selecionar um conjunto de atributos?

Seleção de Atributos

- Propostas *Embedded*
  - Seleção de atributos ocorre naturalmente como parte do algoritmo de AM
- Filtros
  - Atributos são selecionados antes da execução do algoritmo de AM
- *Wrappers*
  - Seleção de atributos usa o algoritmo de AM para encontrar o melhor subconjunto de atributos

Discretização e Transformação de Variáveis → próxima aula

Tarefa

- Leitura do Capítulo (seções 2.1 a 2.3) do livro Tan et al, 2006

Referências

- Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006.