

Fundamentos dos Dados

SCC5836/SCC0252 – Visualização Computacional

Profa. Maria Cristina

sala4-205

cristina@icmc.usp.br

Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)

VICG Grupo de Visualização,
ICMC Imagens e Computação Gráfica
USP

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias
- 4 Processamento dos Dados
 - Metadados e Estatística
 - Valores Ausentes
 - Normalização
 - Re-Amostragem
 - Redução de Dimensionalidade
 - Mapeando Dimensões Nominais para Números
 - Agregação e Sumarização
- 5 Distâncias e Similaridades
- 6 Referências

- Fontes de dados: **sensores, medições ou coletas, simulações ou computações**

Introdução

- Fontes de dados: **sensores, medições ou coletas, simulações ou computações**
- Dados podem ser **crus** (não tratados) ou podem ter sido processados de alguma maneira (suavização, normalização, remoção de ruído, interpolação, escala...)

Introdução

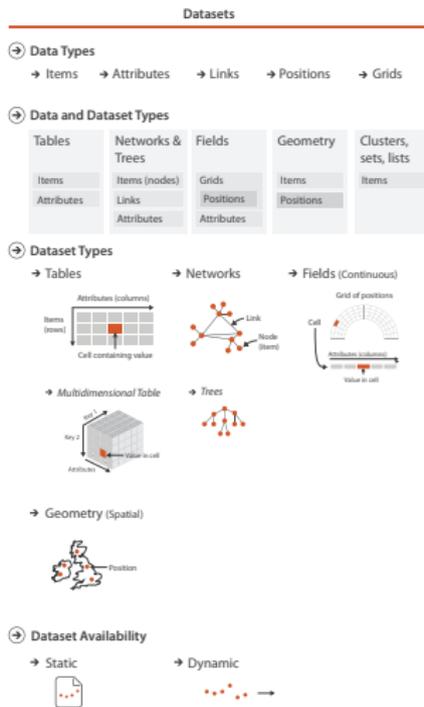


Figura: O que pode ser visualizado: dados, conjuntos de dados, e atributos.

Fonte: T. Munzner Visualization Analysis & Design (Fig. 2.2)

Introdução

- Um conjunto de dados consiste de n instâncias (registros, observações, elementos)

$$(r_1, r_2, r_3, \dots, r_n)$$

- Cada qual descrita por m (uma ou mais) atributos, ou variáveis

$$(v_1, v_2, v_3, \dots, v_m)$$

Introdução

- Um conjunto de dados consiste de n instâncias (registros, observações, elementos)

$$(r_1, r_2, r_3, \dots, r_n)$$

- Cada qual descrita por m (uma ou mais) atributos, ou variáveis

$$(v_1, v_2, v_3, \dots, v_m)$$

- Matriz de dados $n \times m$: cada linha descreve uma instância, cada coluna descreve um atributo

Introdução

- Um conjunto de dados consiste de n instâncias (registros, observações, elementos)

$$(r_1, r_2, r_3, \dots, r_n)$$

- Cada qual descrita por m (uma ou mais) atributos, ou variáveis

$$(v_1, v_2, v_3, \dots, v_m)$$

- Matriz de dados $n \times m$: cada linha descreve uma instância, cada coluna descreve um atributo

- Cada atributo (variável) pode ser um valor único (**número**, **símbolo** ou **cadeia de caracteres**), ou uma estrutura mais **complexa**

Introdução

- Uma variável pode ser
 - **Independente** (iv_i): seu valor não é afetado ou controlado por outra variável, por exemplo, tempo em uma série temporal
 - **Dependente** (dv_j): seu valor é afetado/determinado por uma ou mais variáveis independentes

Introdução

- Uma variável pode ser
 - **Independente** (iv_i): seu valor não é afetado ou controlado por outra variável, por exemplo, tempo em uma série temporal
 - **Dependente** (dv_j): seu valor é afetado/determinado por uma ou mais variáveis independentes

- Assim, uma instância pode ser definida como

$$r_i = (iv_1, iv_2, \dots, iv_{m_i}, dv_1, dv_2, \dots, dv_{m_d})$$

- Com m_i o número de variáveis independentes e m_d o número de variáveis dependentes e $m = m_i + m_d$

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias
- 4 Processamento dos Dados
 - Metadados e Estatística
 - Valores Ausentes
 - Normalização
 - Re-Amostragem
 - Redução de Dimensionalidade
 - Mapeando Dimensões Nominiais para Números
 - Agregação e Sumarização
- 5 Distâncias e Similaridades
- 6 Referências

Tipos de Dados

- Cada atributo/variável representa um componente de informação, e pode ser categorizado como **ordenável** ou **categorício**

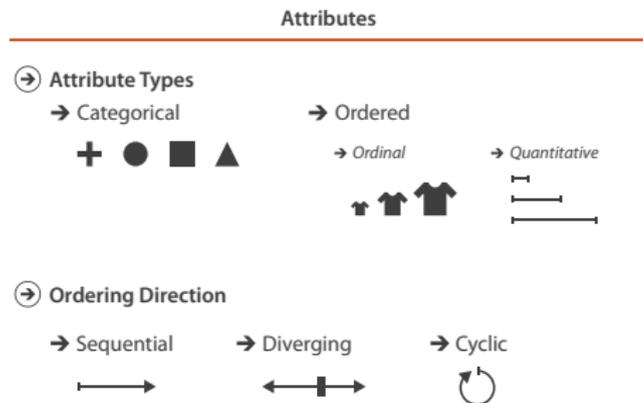


Figura: Tipos de atributos. Fonte: T. Munzner, Visualization Analysis & Design (Fig. 2.7)

Tipos de Dados

- Cada atributo/variável representa um componente de informação, e pode ser classificada como **ordinal** (numérica) ou **nominal** (categórica, ou não-numérica)

Tipos de Dados

- Cada atributo/variável representa um componente de informação, e pode ser classificada como **ordinal** (numérica) ou **nominal** (categórica, ou não-numérica)
- Atributos ordinais
 - **binários**: valores 0 e 1
 - **discretos**: valores inteiros
 - **contínuos**: valores reais

Tipos de Dados

- Cada atributo/variável representa um componente de informação, e pode ser classificada como **ordinal** (numérica) ou **nominal** (categórica, ou não-numérica)

- Atributos ordinais

- **binários**: valores 0 e 1
- **discretos**: valores inteiros
- **contínuos**: valores reais

- Atributos nominais

- **categóricos**: valor de uma lista finita de possibilidades (ex. vermelho, azul e verde)
- **ranqueados**: valor categórico com uma ordem (ex. pequeno, médio e grande)
- **arbitrários**: faixa infinita de valores sem ordem (ex. endereço)

Tipos de Dados

- Os dados podem também ser classificados considerando

Tipos de Dados

- Os dados podem também ser classificados considerando
- **Relação de ordem:** se podem ser ordenados – variáveis ranqueadas nominais e todas as ordinais

Tipos de Dados

- Os dados podem também ser classificados considerando
- **Relação de ordem:** se podem ser ordenados – variáveis ranqueadas nominais e todas as ordinais
- **Métrica de distância:** se é possível calcular distâncias entre instâncias – não é comum em variáveis nominais

Tipos de Dados

- Os dados podem também ser classificados considerando
- **Relação de ordem:** se podem ser ordenados – variáveis ranqueadas nominais e todas as ordinais
- **Métrica de distância:** se é possível calcular distâncias entre instâncias – não é comum em variáveis nominais
- **Existência de zero absoluto:** menor valor fixo – diferenciar variáveis ordinais (ex. peso e saldo bancário) – somente possui zero absoluto se as 4 operações matemáticas puderem ser aplicadas (+, -, \times e \div)

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias**
- 4 Processamento dos Dados
 - Metadados e Estatística
 - Valores Ausentes
 - Normalização
 - Re-Amostragem
 - Redução de Dimensionalidade
 - Mapeando Dimensões Nominiais para Números
 - Agregação e Sumarização
- 5 Distâncias e Similaridades
- 6 Referências

Estrutura Dentro e Entre Instâncias

- Os dados são estruturados em termos de **representação** (sintaxe) e tipos de **relacionamentos** dentro e entre instâncias (semântica)

Escalar

- Um valor numérico individual em uma instância é normalmente chamado de **escalar**
 - Múltiplos escalares podem compor um item de dados, referido como multidimensional, multivariado ou multicampo

Escalar

- Um valor numérico individual em uma instância é normalmente chamado de **escalar**
 - Múltiplos escalares podem compor um item de dados, referido como multidimensional, multivariado ou multicampo

Vetor

- Múltiplos escalares relacionados podem compor um **vetor**
 - Vetores normalmente codificam **direção e magnitude**, como velocidade, aceleração e força

Tensores

- Escalares e vetores são tipos simples de **tensores**
 - Um tensor é definido por um *rank* e a dimensão do espaço no qual está definido
 - Representado como uma matriz ou vetor

- Uma matriz 3×3 pode representar um tensor de *rank* 2 no espaço 3D

- Geometria pode ser definida, por exemplo, **adicionando as coordenadas** de uma instância

Geometria e Grades

- Geometria pode ser definida, por exemplo, **adicionando as coordenadas** de uma instância
- Geometria também pode ser definida por meio de uma **grade** (*grid*)
 - O conjunto de dados é **estruturado** de forma que instâncias de dados sucessivas estão **localizadas em posições sucessivas** na grade

Geometria e Grades

- Geometria pode ser definida, por exemplo, **adicionando as coordenadas** de uma instância
- Geometria também pode ser definida por meio de uma **grade** (*grid*)
 - O conjunto de dados é **estruturado** de forma que instâncias de dados sucessivas estão **localizadas em posições sucessivas** na grade
- Diferentes **sistemas de coordenadas** são usados em conjuntos de dados estruturados como grades
 - Cartesiano, esférico e hiperbólico

Geometria e Grades

- Geometria pode ser definida, por exemplo, **adicionando as coordenadas** de uma instância
- Geometria também pode ser definida por meio de uma **grade** (*grid*)
 - O conjunto de dados é **estruturado** de forma que instâncias de dados sucessivas estão **localizadas em posições sucessivas** na grade
- Diferentes **sistemas de coordenadas** são usados em conjuntos de dados estruturados como grades
 - Cartesiano, esférico e hiperbólico
- Geometria **não-uniforme** (ou irregular) também é comum
 - A densidade de cálculos, ou medidas, varia sobre o domínio
 - As coordenadas geométricas devem fazer parte dos dados

Outras Formas de Estruturas

- **Tempo:** *timestamp*, medidas podem ser tomadas de maneira uniforme ou não-uniforme

Outras Formas de Estruturas

- **Tempo:** *timestamp*, medidas podem ser tomadas de maneira uniforme ou não-uniforme
- **Topologia:** define como as instâncias de dados são **conectadas**
 - Pode ser definida como uma **vizinhança em uma grade**, ou uma **hierarquia** ou **grafo**
 - Importante para **re-amostragem** e **interpolação**
 - Pode ser definida **explicitamente** ou por meio da **estrutura dos dados**

Outras Formas de Estruturas

- Exemplos de dados estruturados
 - **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em uma grade 3D

- Exemplos de dados estruturados
 - **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em uma grade 3D
 - **CFD (*computational fluid dynamics*)**: três dimensões para deslocamento, com um atributo temporal e três espaciais, conectividade em uma grade 3D (uniforme e não-uniforme)

Outras Formas de Estruturas

- Exemplos de dados estruturados
 - **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em uma grade 3D
 - **CFD (*computational fluid dynamics*)**: três dimensões para deslocamento, com um atributo temporal e três espaciais, conectividade em uma grade 3D (uniforme e não-uniforme)
 - **Financeiros**: sem estrutura geométrica, n componentes possivelmente independentes, nominal e ordinal, com atributo temporal

Outras Formas de Estruturas

- Exemplos de dados estruturados
 - **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em uma grade 3D
 - **CFD (*computational fluid dynamics*)**: três dimensões para deslocamento, com um atributo temporal e três espaciais, conectividade em uma grade 3D (uniforme e não-uniforme)
 - **Financeiros**: sem estrutura geométrica, n componentes possivelmente independentes, nominal e ordinal, com atributo temporal
 - **Sensoriamento remoto**: múltiplos canais, com dois ou três atributos espaciais, um temporal e conectividade por uma grade

Outras Formas de Estruturas

- Exemplos de dados estruturados
 - **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em uma grade 3D
 - **CFD (*computational fluid dynamics*)**: três dimensões para deslocamento, com um atributo temporal e três espaciais, conectividade em uma grade 3D (uniforme e não-uniforme)
 - **Financeiros**: sem estrutura geométrica, n componentes possivelmente independentes, nominal e ordinal, com atributo temporal
 - **Sensoriamento remoto**: múltiplos canais, com dois ou três atributos espaciais, um temporal e conectividade por uma grade
 - **Censo**: múltiplos campos de todos os tipos, atributos espacial e temporal, e conectividade dada pela similaridade dos campos

Outras Formas de Estruturas

- Exemplos de dados estruturados
 - **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em uma grade 3D
 - **CFD (*computational fluid dynamics*)**: três dimensões para deslocamento, com um atributo temporal e três espaciais, conectividade em uma grade 3D (uniforme e não-uniforme)
 - **Financeiros**: sem estrutura geométrica, n componentes possivelmente independentes, nominal e ordinal, com atributo temporal
 - **Sensoriamento remoto**: múltiplos canais, com dois ou três atributos espaciais, um temporal e conectividade por uma grade
 - **Censo**: múltiplos campos de todos os tipos, atributos espacial e temporal, e conectividade dada pela similaridade dos campos
 - **Redes sociais**: campos de todos os tipos, com vários atributos de conectividade que podem ser espacial, temporal ou dependente de outros atributos

Processamento dos Dados

- Em algumas situações, por exemplo, aplicações médicas, pode ser preferível visualizar dados “crus”
- Mas, em geral é **necessário** algum tipo de **pré-processamento**

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias
- 4 Processamento dos Dados**
 - **Metadados e Estatística**
 - Valores Ausentes
 - Normalização
 - Re-Amostragem
 - Redução de Dimensionalidade
 - Mapeando Dimensões Nominais para Números
 - Agregação e Sumarização
- 5 Distâncias e Similaridades
- 6 Referências

- **Metadados** podem apoiar a interpretação dos dados, **provendo informação** sobre sua semântica
 - ponto de referência para as medidas,
 - unidades de medidas
 - símbolo usado para indicar valores faltantes
 - resolução das medidas
 - etc.

- **Análise estatística** pode prover informação útil, como
 - detecção de valores **espúrios** (erros de medida/coleta)
 - identificação de **agrupamentos** (medidas similares)
 - identificação de atributos **redundantes** por meio de **análise de correlação**

- Cálculos estatísticos mais comuns incluem a média (considerando o atributo j)

$$\mu_j = \frac{1}{n} \sum_{i=1}^n (x_{ij})$$

- e o desvio padrão

$$\sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \mu_j)^2}$$

- Histogramas permitem analisar a distribuição dos dados

Método Simples de Detecção de Anomalia

- Supondo que o atributo j tenha uma distribuição Gaussiana $N(\mu_j, \sigma_j)$, primeiro transforme os dados para que $\mu_j = 0$ e $\sigma_j = 1$
- Se x_{ij} for o valor do atributo j da instância i e c for uma constante, a probabilidade de que $|x_{ij}| \geq c$ cai rapidamente conforme c aumenta

- Se $\alpha = \text{prob}(|x_{ij}| \geq c)$, uma instância de dados é um *outlier* se

$$|x_{ij}| \geq c$$

com c uma constante escolhida de forma que $\text{prob}(|x_{ij}| \geq c) \geq \alpha$

Método Simples de Identificação de Atributos Redundantes

- A correlação indica o quanto dois atributos são (linearmente) relacionados

- Sejam x_i e x_j dois atributos, a correlação entre eles é calculada como

$$\text{cor}(x_i, x_j) = \frac{\text{cov}(x_i, x_j)}{\text{var}(x_i)\text{var}(x_j)}$$

com $\text{cov}(x_i, x_j)$ a covariância entre x_i e x_j

$$\text{cov}(x_i, x_j) = \frac{1}{m-1} \sum_{k=1}^m (x_{ki} - \mu_i)(x_{kj} - \mu_j)$$

e $\text{var}(x_i)$ a variância de x_i

$$\text{var}(x_i) = \sigma_i^2$$

- Valores de $|\text{cor}(x_i, x_j)|$ próximos de 1 indicam alta correlação entre atributos, de forma que podemos descartar x_i ou x_j

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias
- 4 Processamento dos Dados**
 - Metadados e Estatística
 - Valores Ausentes**
 - Normalização
 - Re-Amostragem
 - Redução de Dimensionalidade
 - Mapeando Dimensões Nominais para Números
 - Agregação e Sumarização
- 5 Distâncias e Similaridades
- 6 Referências

Valores Ausentes e Limpeza dos Dados

- Em dados “reais” é normal que **dados** estejam **faltando** ou que estejam **errados**

Valores Ausentes e Limpeza dos Dados

- Em dados “reais” é normal que **dados** estejam **faltando** ou que estejam **errados**
- Estratégias comuns para lidar com esses problemas
 - **Descartar** as instâncias **incompletas**: pode implicar em grande perda de informação

Valores Ausentes e Limpeza dos Dados

- Em dados “reais” é normal que **dados** estejam **faltando** ou que estejam **errados**
- Estratégias comuns para lidar com esses problemas
 - **Descartar** as instâncias **incompletas**: pode implicar em grande perda de informação
 - Assinalar um valor **sentinela**: cuidado para não usar a sentinela nos cálculos

Valores Ausentes e Limpeza dos Dados

- Em dados “reais” é normal que **dados** estejam **faltando** ou que estejam **errados**
- Estratégias comuns para lidar com esses problemas
 - **Descartar** as instâncias **incompletas**: pode implicar em grande perda de informação
 - Assinalar um valor **sentinela**: cuidado para não usar a sentinela nos cálculos
 - Calcular um valor **substituto**: “*data imputation*”

Data Imputation

- Dois métodos simples de “*imputation*” são
 - Atribuir o valor **médio**: pode obscurecer valores espúrios

- Dois métodos simples de “*imputation*” são
 - Atribuir o valor **médio**: pode obscurecer valores espúrios
 - Atribuir um valor baseado no **vizinho** mais próximo: no cálculo da vizinhança é difícil saber se existem atributos mais importantes

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias
- 4 Processamento dos Dados**
 - Metadados e Estatística
 - Valores Ausentes
 - Normalização**
 - Re-Amostragem
 - Redução de Dimensionalidade
 - Mapeando Dimensões Nominais para Números
 - Agregação e Sumarização
- 5 Distâncias e Similaridades
- 6 Referências

Problema

- Em aplicações que envolvem a comparação entre instâncias de dados, duas situações podem distorcer os resultados, ou torná-los tendenciosos
 - Quando as normas Euclidianas dos vetores que representam as instâncias são muito diferentes
 - Quando um (ou mais) atributos desses vetores assume valores em uma escala muito diferente dos demais

Problema

- Em aplicações que envolvem a comparação entre instâncias de dados, duas situações podem distorcer os resultados, ou torná-los tendenciosos
 - Quando as normas Euclidianas dos vetores que representam as instâncias são muito diferentes
 - Quando um (ou mais) atributos desses vetores assume valores em uma escala muito diferente dos demais
- Uma possível solução é aplicar um processo conhecido como **Normalização**
 - Transformar os valores de forma que satisfaçam alguma **propriedade estatística**

Normalização

- Para evitar o primeiro cenário, pode-se transformar os dados de modo que os vetores que representam as instâncias tenham norma unitária
 - $x'_{ij} = x_{ij}/\|\mathbf{x}_i\|$ para $1 \leq j \leq m$

Normalização

- Para evitar o primeiro cenário, pode-se transformar os dados de modo que os vetores que representam as instâncias tenham norma unitária
 - $x'_{ij} = x_{ij}/\|\mathbf{x}_i\|$ para $1 \leq j \leq m$
- Os vetores podem ser as linhas ou as colunas da matriz (tabela) de dados

Normalização

- Outro processo de normalização consiste em transformar os dados para que os valores fiquem no intervalo $[0, 1]$
- Nesse, se os valores máximo d_{max} e mínimo d_{min} forem conhecidos, fazemos

$$d_{normalizado} = (d_{original} - d_{min}) / (d_{max} - d_{min})$$

Normalização

- Outro processo de normalização consiste em transformar os dados para que os valores fiquem no intervalo $[0, 1]$

- Nesse, se os valores máximo d_{max} e mínimo d_{min} forem conhecidos, fazemos

$$d_{normalizado} = (d_{original} - d_{min}) / (d_{max} - d_{min})$$

- Em casos específicos é interessante **usar** não os valores máximos e mínimos nos dados (intervalo de variação), mas **valores conhecidos**, como nas porcentagens

- **Normalização** pode **distorcer** os dados, por exemplo, na presença de valores espúrios, os dados podem ser “achatados”

Normalização

- **Normalização** pode **distorcer** os dados, por exemplo, na presença de valores espúrios, os dados podem ser “achatados”
- Outra normalização conhecida, a “**standardization**” transforma os valores de um atributo de modo que a média seja 0 e o desvio padrão 1

Normalização

- **Normalização** pode **distorcer** os dados, por exemplo, na presença de valores espúrios, os dados podem ser “achatados”
- Outra normalização conhecida, a “**standardization**” transforma os valores de um atributo de modo que a média seja 0 e o desvio padrão 1
- Considerando a média μ_j e o desvio padrão σ_j do atributo j , calculamos

$$x'_{ij} = (x_{ij} - \mu_j) / \sigma_j$$

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias
- 4 Processamento dos Dados**
 - Metadados e Estatística
 - Valores Ausentes
 - Normalização
 - Re-Amostragem**
 - Redução de Dimensionalidade
 - Mapeando Dimensões Nominiais para Números
 - Agregação e Sumarização
- 5 Distâncias e Similaridades
- 6 Referências

Re-Amostragem e Extração de Sub-conjuntos

- Pode ser interessante, em muitos casos preencher “espaços” dentre duas amostras: re-amostragem por **interpolação**

Re-Amostragem e Extração de Sub-conjuntos

- Pode ser interessante, em muitos casos preencher “espaços” dentre duas amostras: re-amostragem por **interpolação**

Interpolação Linear

- Dados os valores de uma variável d em duas posições A e B , é possível estimar seu valor na posição C fazendo (porcentagem da distância entre A e B)

$$(x_C - x_A)/(x_B - x_A) = (d_C - d_A)/(d_B - d_A)$$

- isto é

$$d_C = d_A + (d_B - d_A) * (x_C - x_A)/(x_B - x_A)$$

- Outra maneira é usar a **equação paramétrica da reta** (independente de dimensão)

Re-Amostragem e Extração de Sub-conjuntos

- Outra maneira é usar a **equação paramétrica da reta** (independente de dimensão)

- Um ponto $P(t)$ pode ser descrito em uma reta começando em P_A e indo até P_B como

$$P(t) = P_A + Vt$$

- Onde V é o vetor de P_B até P_A , $V = P_B - P_A$

Re-Amostragem e Extração de Sub-conjuntos

- Outra maneira é usar a **equação paramétrica da reta** (independente de dimensão)

- Um ponto $P(t)$ pode ser descrito em uma reta começando em P_A e indo até P_B como

$$P(t) = P_A + Vt$$

- Onde V é o vetor de P_B até P_A , $V = P_B - P_A$

- Sabendo a localização P_C , calculamos t e usamos o mesmo tipo de equação para calcular o valor do atributo

$$d(t) = d_A + Ut$$

- Onde $U = d_b - d_a$

Interpolação Bilinear

- É possível estender esse conceito para **duas dimensões**, repetindo esse procedimento para cada dimensão
 - Interpola inicialmente na horizontal (vertical) e então na vertical (horizontal)

- Além de expandir os dados também é possível **reduzi-los por meio de amostragem**
 - Isso pode ser feito, por exemplo, selecionando **dados regularmente espaçados**, mas pode acarretar em perda de informação (mapas)

Re-Amostragem e Extração de Sub-conjuntos

- Além de expandir os dados também é possível **reduzi-los por meio de amostragem**
 - Isso pode ser feito, por exemplo, selecionando **dados regularmente espaçados**, mas pode acarretar em perda de informação (mapas)
- Outras abordagens envolvem o cálculo das **médias** em vizinhanças ou a **seleção aleatória** de um pixel em uma determinada região

Re-Amostragem e Extração de Sub-conjuntos

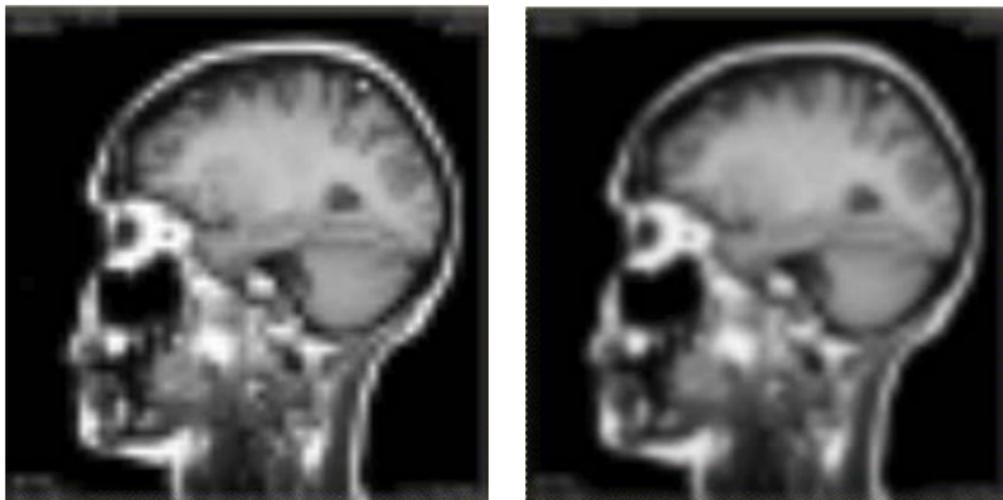


Figura: Imagens em menor resolução da caveira apresentada anteriormente usando replicação de pixels e cálculo de média.

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias
- 4 Processamento dos Dados**
 - Metadados e Estatística
 - Valores Ausentes
 - Normalização
 - Re-Amostragem
 - Redução de Dimensionalidade**
 - Mapeando Dimensões Nominiais para Números
 - Agregação e Sumarização
- 5 Distâncias e Similaridades
- 6 Referências

Redução de Dimensionalidade

- Nas situações em que a **dimensionalidade excede a capacidade** da técnica de visualização é necessário uma redução **preservando**, o máximo possível, **informação** contida nos dados

Redução de Dimensionalidade

- Nas situações em que a **dimensionalidade excede a capacidade** da técnica de visualização é necessário uma redução **preservando**, o máximo possível, **informação** contida nos dados
- Essa redução pode ser feita **manualmente**, selecionando os atributos, ou usando alguma **técnica**, como
 - Principal Component Analysis (PCA)
 - Multidimensional Scaling (MDS)
 - Self-organizing Maps (SOM)

Redução de Dimensionalidade

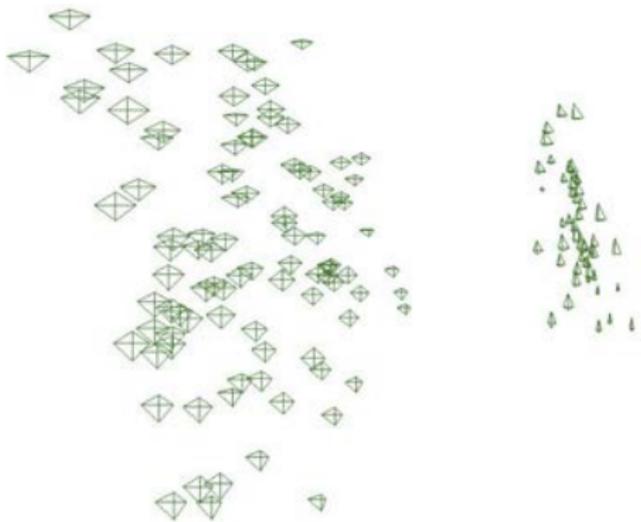


Figura: Projeção usando PCA do conjunto de dados Iris. Os elementos gráficos (*glyphs*) representam as 4 variáveis originais – cada linha emanando do centro é proporcional ao valor do atributo.

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias
- 4 Processamento dos Dados**
 - Metadados e Estatística
 - Valores Ausentes
 - Normalização
 - Re-Amostragem
 - Redução de Dimensionalidade
 - Mapeando Dimensões Nominais para Números**
 - Agregação e Sumarização
- 5 Distâncias e Similaridades
- 6 Referências

Mapeando Dimensões Nominais para Números

- No caso de valores **nominais ranqueados** (qualidade do ar), o mapeamento é **direto**, mas para valores **não ranqueados** o problema é mais **complexo** (montadoras de carros)

Mapeando Dimensões Nominais para Números

- No caso de valores **nominais ranqueados** (qualidade do ar), o mapeamento é **direto**, mas para valores **não ranqueados** o problema é mais **complexo** (montadoras de carros)
- Se as variáveis apresentarem **poucos valores** diferentes pode-se usar **cor** ou **forma** (não implica em uma relação de ordem)

Mapeando Dimensões Nominais para Números

- No caso de valores **nominais ranqueados** (qualidade do ar), o mapeamento é **direto**, mas para valores **não ranqueados** o problema é mais **complexo** (montadoras de carros)
- Se as variáveis apresentarem **poucos valores** diferentes pode-se usar **cor** ou **forma** (não implica em uma relação de ordem)
- Se existir **apenas uma variável nominal** podemos
 - Usar seu valor como um **rótulo para a instância** – para evitar problemas de oclusão podemos mostrar mais densamente os rótulos perto do cursor
 - Podemos **atribuir um valor numérico** à variável verificando as outras variáveis (distâncias) e então usar MDS para definir as coordenadas unidimensionais (*correspondence analysis*)

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias
- 4 Processamento dos Dados**
 - Metadados e Estatística
 - Valores Ausentes
 - Normalização
 - Re-Amostragem
 - Redução de Dimensionalidade
 - Mapeando Dimensões Nominais para Números
 - **Agregação e Sumarização**
- 5 Distâncias e Similaridades
- 6 Referências

Agregação

- Pode ser útil **agrupar** dados e mostrar um **representativo** dos mesmos
 - Pode-se mostrar uma **média** ou algum tipo de **informação extra**, como o número de elementos do grupo
- A agregação tem dois elementos
 - O método para **definir os grupos**
 - A abordagem para **apresentar os grupos** resultantes

Agregação

- Pode ser útil **agrupar** dados e mostrar um **representativo** dos mesmos
 - Pode-se mostrar uma **média** ou algum tipo de **informação extra**, como o número de elementos do grupo
- A agregação tem dois elementos
 - O método para **definir os grupos**
 - A abordagem para **apresentar os grupos** resultantes
- Ponto central é **prover informação** para ajudar o usuário a **analisar mais a fundo** os dados de um grupo
 - Suporte na detecção da variabilidade, presença de valores espúrios, etc. é essencial

Agregação e Sumarização

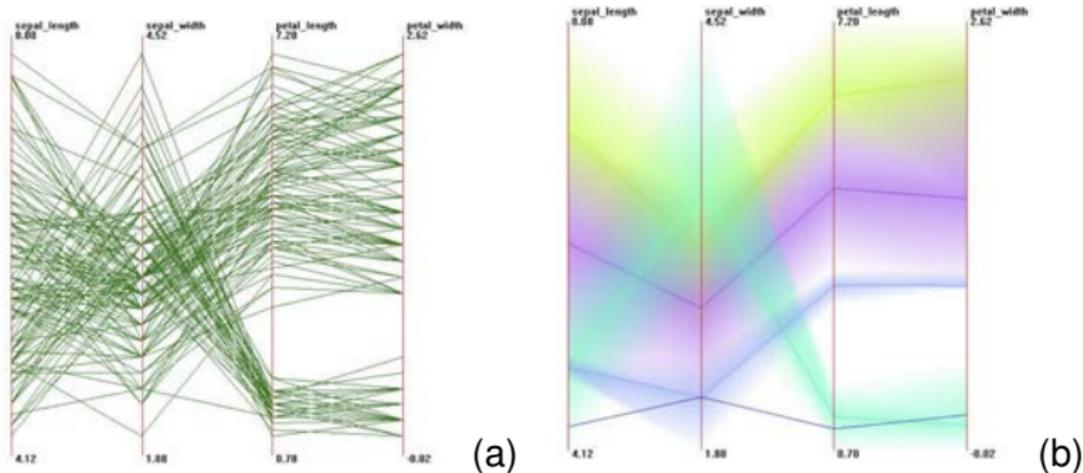


Figura: Conjunto Iris apresentado usando coordenadas paralelas. (a) conjunto original. (b) os centros e as extensões dos grupos após a agregação.

Observação Final

- Se os dados forem processados e sofrerem alguma mudança, isso deve ser comunicado ao usuário analista

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias
- 4 Processamento dos Dados
 - Metadados e Estatística
 - Valores Ausentes
 - Normalização
 - Re-Amostragem
 - Redução de Dimensionalidade
 - Mapeando Dimensões Nominais para Números
 - Agregação e Sumarização
- 5 Distâncias e Similaridades
- 6 Referências

Distâncias

- A maneira de calcular a distância ($\delta(\mathbf{x}_i, \mathbf{x}_j)$) entre os objetos multidimensionais \mathbf{X} desempenha papel central
- Distância de *Minkowski* – família de métricas de distância denominadas normas L_p

$$L_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^m |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}} \quad (1)$$

- Com $p = 1$ obtém-se a distância *Manhattan* (*City Block*)
- Com $p = 2$ tem-se a distância Euclideana
- Com $p = \infty$ obtém-se a distância do infinito ($L_\infty(\mathbf{x}_i, \mathbf{x}_j) = \max_{k=1}^m |x_{ik} - x_{jk}|$)

Propriedades de uma Métrica (Distância)

- 1 **Não-Negatividade:** $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}, \delta(\mathbf{x}_i, \mathbf{x}_j) \geq 0$
- 2 **Identidade:** $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}, \mathbf{x}_i = \mathbf{x}_j \Leftrightarrow \delta(\mathbf{x}_i, \mathbf{x}_j) = 0$
- 3 **Simetria:** $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}, \delta(\mathbf{x}_i, \mathbf{x}_j) = \delta(\mathbf{x}_j, \mathbf{x}_i)$
- 4 **Desigualdade Triangular:** $\forall \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \in \mathbf{X},$
 $\delta(\mathbf{x}_i, \mathbf{x}_k) \leq \delta(\mathbf{x}_i, \mathbf{x}_j) + \delta(\mathbf{x}_j, \mathbf{x}_k)$

Distâncias

- Nem toda dissimilaridade é uma distância (métrica) – não precisa obedecer as propriedades métricas

- Uma dissimilaridade pode ser o inverso de uma similaridade $s(\mathbf{x}_i, \mathbf{x}_j)$

- $\delta(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{s(\mathbf{x}_i, \mathbf{x}_j) + 1}$

- $\delta(\mathbf{x}_i, \mathbf{x}_j) = e^{-s(\mathbf{x}_i, \mathbf{x}_j)}$

- $\delta(\mathbf{x}_i, \mathbf{x}_j) = 1 - s'(\mathbf{x}_i, \mathbf{x}_j)$

- com $s'(\mathbf{x}_i, \mathbf{x}_j) = \frac{s(\mathbf{x}_i, \mathbf{x}_j) - s_{min}}{s_{max} - s_{min}}$

- Exemplo conhecido: dissimilaridade do cosseno
 - $1 - \cos(\mathbf{x}_i, \mathbf{x}_j)$

Distâncias Binárias

- Para dados binários, devem ser usadas certos tipos específicos de métricas
- <http://people.revoledu.com/kardi/tutorial/Similarity/BinaryVariables.html>

Distâncias Binárias

- p : número de variáveis 1 para ambas instâncias
- q : número de variáveis 1 para uma instância e 0 para a outra
- r : número de variáveis 0 para uma instância e 1 para a outra
- s : número de variáveis 0 para ambas instâncias
- $t = p + q + r + s$: número total de instâncias

- $d_{ij} = (q + r)/t$ (simple matching)
- $d_{ij} = (q + r)/(p + q + r)$ (Jaccard's distance)
- $d_{ij} = (q + r)$ (Hamming distance)
- $d_{ij} = (p + s)/t$ (simple matching coefficient)
- $d_{ij} = p/t$
- $d_{ij} = p/(p + q + r)$ (Jaccard's coefficient)
- $d_{ij} = 2p/(2p + q + r)$
- $d_{ij} = 2(p + s)/(2(p + s) + q + r)$
- $d_{ij} = p/(q + r)$
- $d_{ij} = (p + s)/(q + r)$

- 1 Introdução
- 2 Tipos de Dados
- 3 Estrutura Dentro e Entre Instâncias
- 4 Processamento dos Dados
 - Metadados e Estatística
 - Valores Ausentes
 - Normalização
 - Re-Amostragem
 - Redução de Dimensionalidade
 - Mapeando Dimensões Nominais para Números
 - Agregação e Sumarização
- 5 Distâncias e Similaridades
- 6 Referências

Referências

- Ward, M., Grinstein, G. G., Keim, D. **Interactive data visualization foundations, techniques, and applications.** Natick, Mass., A K Peters, 2010, Cap. 2.
- **[Munzner, 2015]** Tamara Munzner, Visualization Analysis & Design, CRC Press, Cap. 2.