Fundamentos dos Dados

SCC5836 – Visualização Computacional

Prof. Fernando V. Paulovich
http://www.icmc.usp.br/~paulovic
paulovic@icmc.usp.br

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

VICG Grupo de Visualização, Imagens e Computação Gráfica

Sumário

Introdução

- Dados podem vir de **sensores** ou serem **coletados**, também podem ser gerados por **simulações ou computações**

Introdução

- Dados podem vir de **sensores** ou serem **coletados**, também podem ser gerados por **simulações ou computações**

- Dados podem ser **crus** (não tratados) ou podem ser **derivados** por meio de algum processo, como suavização, remoção de ruído, escala ou interpolação

Introdução

- Dados consistem de n instâncias

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

- Cada uma consistindo de m (uma ou mais) observações ou variáveis

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

- Cada observação pode ser um 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, o tempo em uma série temporal
  - **Dependente** (dv_j): é afetada pela variação de uma ou mais variáveis independentes

- Assim uma instância pode ser definida como

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

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

Sumário

Tipos de Dados

- Cada variável representa uma parcela de informação e pode ser classificada como **ordinal** (númerica) e **nominal** (não-numérica)

Tipos de Dados

- Cada variável representa uma parcela de informação e pode ser classificada como **ordinal** (númerica) e **nominal** (não-numérica)

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

Tipos de Dados

- Cada variável representa uma parcela de informação e pode ser classificada como **ordinal** (númerica) e **nominal** (não-numérica)

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

- Dados nominais
  - **categóricos**: valor em 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 uma 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**: os dados podem ser ordenados – variáveis ranqueadas nominais e todas ordinais

Tipos de Dados

- Os dados podem também ser classificados considerando

- **Relação de ordem**: os dados podem ser ordenados – variáveis ranqueadas nominais e todas ordinais

- **Métrica de distância**: é possível calcular distâncias entre instâncias – variáveis ordinais, mas não é comum em variáveis nominais

Tipos de Dados

- Os dados podem também ser classificados considerando

- **Relação de ordem**: os dados podem ser ordenados – variáveis ranqueadas nominais e todas ordinais

- **Métrica de distância**: é possível calcular distâncias entre instâncias – variáveis ordinais, mas não é comum em variáveis nominais

- **Existências 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 $\div$)

Tipos de Dados

Importância

- Os elementos gráficos tem uma escala associada, a qual deve (preferencialmente) ser compatível com a escala das variáveis que representam

Sumário

Estrutura Dentro e Entre Instâncias

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

Escalares, Vetores e Tensores

Escalar

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

Escalares, Vetores e Tensores

Escalar

- Um número 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**, por exemplo, velocidade, aceleração e força

Escalares, Vetores e Tensores

Tensores

- Escalares e vetores são tipos simples de **tensores**
  - Um tensor é definido por um *rank* e pelaa dimensão do espaço em está definido
  - Representado como uma matriz ou vetor

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

Geometria e Grades

- 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** é outra maneira de **estruturar dados** (*timestamp*), por meio de medidas tomadas a intervalos uniformes ou não-uniformes

Outras Formas de Estruturas

- **Tempo** é outra maneira de **estruturar dados** (*timestamp*), por meio de medidas tomadas a intervalos uniformes ou não-uniformes

- **Topologia** é ainda outra maneira de estruturar dados, definindo como as instâncias são **conectadas**
  - Pode ser definida como uma **vizinhança em uma grade**, ou por uma **hierarquia** ou **grafo**
  - Importante para **reamostragem** 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

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 uma 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 uma 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 uma 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 uma 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 uma 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 tipos, com vários atributos de conectividade que podem ser espacial, temporal ou dependente de outros atributos

Sumário

Processamento dos Dados

- Em alguns casos, pode ser preferível visualizar os dados “crus”, por exemplo, em aplicações médicas, mas em outros pode ser **necessário** aplicar algum **pré-processamento**

Sumário

Metadados e Estatística

- **Metadados** podem ajudar na interpretação dos dados, **provendo informação** como
  - ponto de referência para as medidas,
  - unidades usadas nas medidas
  - símbolo para indicar valores ausentes
  - resolução das medidas
  - etc.

Metadados e Estatística

- **Análise estatística** pode prover informação útil como
  - detecção de valores **espúrios** (dados medidos errados)
  - identificação de **agrupamentos** similares
  - identificação de campos **redundantes** por meio de **correlação**

Metadados e Estatística

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

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

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

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

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

com $cov(x_i, x_j)$ a covariância entre x_i e x_j

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

e $var(x_i)$ a variância de x_i

$$var(x_i) = \sigma_i^2$$

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

Sumário

Valores Faltando e Limpeza dos Dados

- Em dados do mundo “real” é normal que **dados** estejam **faltando** ou que estejam **errados**

Valores Faltando e Limpeza dos Dados

- Em dados do mundo “real” é normal que **dados** estejam **faltando** ou que estejam **errados**

- Estratégias comuns para lidar com esses problemas
  - **Descartar** as instâncias com **defeito**: pode representar grande perda de informação

Valores Faltando e Limpeza dos Dados

- Em dados do mundo “real” é normal que **dados** estejam **faltando** ou que estejam **errados**

- Estratégias comuns para lidar com esses problemas
  - **Descartar** as instâncias com **defeito**: pode representar grande perda de informação
  - Assinalar um valor **sentinela**: cuidado para não usar a sentinela nos cálculos

Valores Faltando e Limpeza dos Dados

- Em dados do mundo “real” é normal que **dados** estejam **faltando** ou que estejam **errados**

- Estratégias comuns para lidar com esses problemas
  - **Descartar** as instâncias com **defeito**: pode representar 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
  - Assinalar o valor **médio**: pode obscurecer valores espúrios

Data Imputation

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

Sumário

Normalização

Problema

- Em aplicações que envovlam a comparação entre instâncias de dados dois diferentes cenários podem distorcer os resultados ou torná-los tendenciosos
  - Quando as normas Euclideanas dos vetores que representam as instâncias são muito diferentes
  - Quando uma (ou mais) coordenadas desses vetores está em uma escala diferente das outras coordenadas

Normalização

Problema

- Em aplicações que envovlam a comparação entre instâncias de dados dois diferentes cenários podem distorcer os resultados ou torná-los tendenciosos
  - Quando as normas Euclideanas dos vetores que representam as instâncias são muito diferentes
  - Quando uma (ou mais) coordenadas desses vetores está em uma escala diferente das outras coordenadas

- Uma possível solução nesse cenário é aplicar um processo conhecido como **Normalização**
  - Transformar os dados de forma que os mesmos obedeçam alguma **propriedade estatística**

Normalização

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

Normalização

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

- Os vetores podem ser a linhas ou colunas da matriz que representa os dados

Normalização

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

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

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

Normalização

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

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

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

- Em casos específicos é interessante **usar** não os valores máximos e mínimos contidos nos dados, mas **valores conhecidos**, como nas porcentagens

Normalização

- **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”** faz com que a média dos valores 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"** faz com que a média dos valores seja 0 e o desvio padrão 1

- Considerando a média μ_j e o desvio padrão σ_j do atributo j, calculamos

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

Sumário

Re-Amostragem e Extração de Sub-conjuntos

- Em muitos casos pode ser interessante preencher “espaços” dentre duas amostras; para isso utiliza-se **interpolação** para reamostragem

Re-Amostragem e Extração de Sub-conjuntos

- Em muitos casos pode ser interessante preencher “espaços” dentre duas amostras; para isso utiliza-se **interpolação** para reamostragem

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

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

Re-Amostragem e Extração de Sub-conjuntos

- Alternativamente, pode-se usar a **equação paramétrica da reta** (independente de dimensão)

Re-Amostragem e Extração de Sub-conjuntos

- Alternativamente, pode-se 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 é o vetor de P_B até P_A, $V = P_B - P_A$

Re-Amostragem e Extração de Sub-conjuntos

- Alternativamente, pode-se 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 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 $U = d_b - d_a$

Re-Amostragem e Extração de Sub-conjuntos

Interpolação Bilinear

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

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 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 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


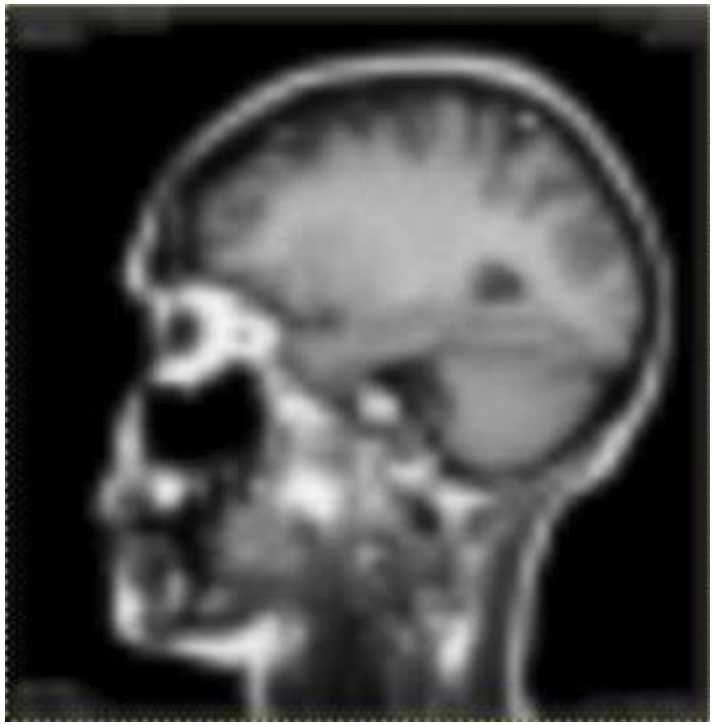

Figura: Imagens em menor resolução da caveira apresentada anteriormente usando replicação de pixels e cálculo de média.

Sumário

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

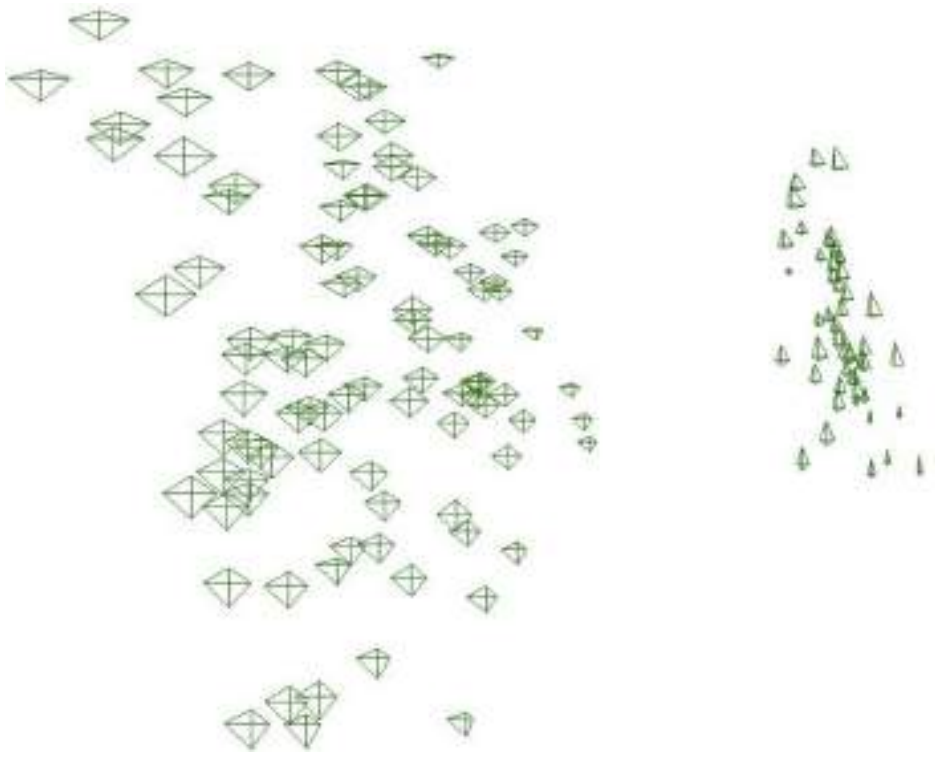

Figura: Projeção usando PCA do conjunto de dados Iris. Os elementos gráficos (*glifos*) representam as 4 variáveis originais – cada linha emanando do centro é proporcional ao valor do atributo.

Sumário

Mapeando Dimensões Nominais para Números

- No caso de valores **nominais ranqueados** (qualidade do ar), o mapeamento é **direto**, mas para valores **não ranqueados** é um problema 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** é um problema 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** é um problema 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, pode-se mostrar mais densamente os rótulos perto do cursor
  - Pode-se **assinalar um valor numérico** à instância, verificando as outras variáveis (distâncias) e então usar MDS para definir as coordenadas unidimensionais (*correspondence analysis*)

Sumário

Agregação e Sumarização

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 e Sumarização

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 um usuário a inspecionar **analisar mais a fundo** os dados de um grupo
  - Suporte à detecção da variabilidade, presença de valores espúrios, etc. é essencial

Agregação e Sumarização

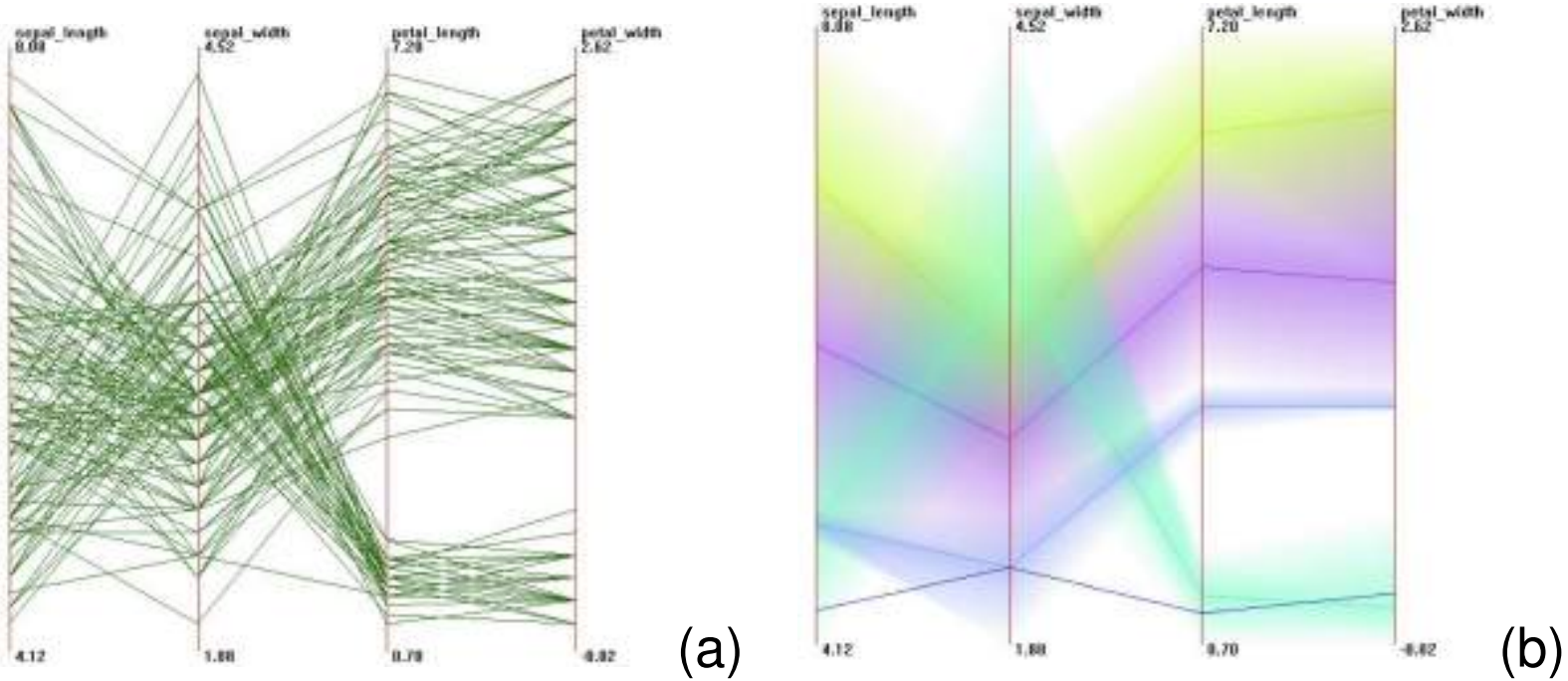

Figura: Conjunto Iris apresentado por Coordenadas Paralelas. (a) conjunto de dados original. (b) os centros e as extensões dos grupos após agregação.

Observação Final

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

Sumário

Distâncias

- A forma como a distância ($\delta(\mathbf{x}_i, \mathbf{x}_j)$) entre os objetos multidimensionais $\mathbf{X}$ é calculada desempenha papel central

- Distância de *Minkowski* – família de métricas de distância denominadas normas L_p

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

- 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 ($L_\infty(\mathbf{x}_i, \mathbf{x}_j) = \max_{k=1}^{m} |x_{ik} - x_{jk}|$)

Distâncias

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 existem certos tipos de métricas específicas que devem ser usadas

- `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 variáreis 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 corfficient)
- $d_{ij} = 2p/(2p + q + r)$
- $d_{ij} = 2(p + s)/(2(p + s) + q + r)$
- $d_{ij} = p/(q + r)$
- $d_{ij} = (p + s)/(q + r)$

Sumário

Referências

- Ward, M., Grinstein, G. G., Keim, D. **Interactive data visualization foundations, techniques, and applications**. Natick, Mass., A K Peters, 2010.