Mineração de Dados em Biologia Molecular

Técnicas de visualização

Docente: André C. P. L. F. de Carvalho
PAE: Victor Hugo Barella

Tópicos

- Técnicas de visualização
- Como podem ser classificadas
- Exemplos
  - Poucos atributos
  - Muitos atributos
  - Dados temporais

Introdução

- Visualização de dados e resultados são obtidas por técnicas de visualização
  - Frequentemente voltadas para o tipo de dado sendo analisado
- Novas técnicas e abordagens de visualização são continuamente criadas
  - Assim como variações especializadas das abordagens existentes
  - Em resposta a novos tipos de dados e tarefas

Técnicas de visualização

- Podem ser classificadas de várias formas
  - Baseadas no número de atributos utilizados
    - 1, 2, 3 ou vários
  - Se os dados têm alguma característica especial
    - Ex.: estrutura de grafo ou hierárquica
  - Pelo tipo dos atributos
  - De acordo com o tipo de aplicação
    - Ex.: científica, estatística ou visualização de informação

Técnicas de visualização

- Técnicas a serem analisadas no curso:
  - Visualização de um pequeno número de atributos
  - Visualização de dados com atributos espaciais ou temporais
  - Visualização de dados com muitos atributos

Poucos atributos

- Várias técnicas
  - Algumas ilustram a distribuição dos valores para um atributo
    - Ex.: histogramas, caule e folha
  - Outras mostram o relacionamento entre valores de dois atributos
    - Ex.: *Scatter plots*

Poucos atributos

- Ilustram a distribuição dos valores para um ou mais atributos
  - Caule e folha
  - Histograma
  - Histograma bi-dimensional
  - Box plot
  - Scatter plot
  - Contour plot

Caule e folha

- Resume valores de um atributo em duas partes
  - Caule
    - Cada caule contém grupo de valores que têm o mesmo primeiro dígito
    - Rotulado pelo primeiro dígito
    - Pode ter mais de um caule para o mesmo dígito
  - Folha
    - Exibe o segundo dígito de cada valor agrupado no caule
- É um tipo de histograma

Caule e folha

- Tamanho da pétala do conjunto iris

```
The decimal point is at the |

1 | 01223333333444444444444
1 | 55555555555555666666777799
2 |
2 |
3 | 033
3 | 55678999
4 | 00000111222233444 4
4 | 555555556667777788889999
5 | 00001111111223344
5 | 55566666677788899
6 | 0011134
6 | 6779
```

Para os valores

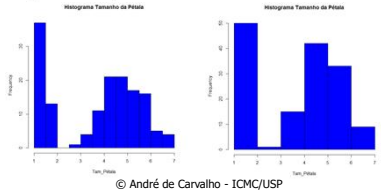
1,0 1,1 1,2 1,2 1,3 1,3
…
6,3 6,4 6,7 6,7 6,7 6,9

Histograma

- Ilustra por barras quantas vezes um valor ou conjunto de valores aparece em um atributo
  - Altura da barra proporcional ao número de vezes
  - Poucos valores (geralmente qualitativos)
    - Uma barra para cada valor
  - Muitos valores (geralmente quantitativos)
    - Divide os valores em cestas (intervalos) e associa uma barra para cada cesta
      - Barras têm larguras diferentes se intervalos das cestas são diferentes

Histograma

- Formato depende do número de cestas
  - Conjunto de dados iris
  - Largura das pétalas usando 12 e 6 cestas

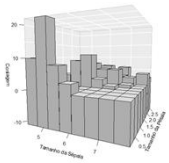


Variações de histogramas

- Histograma relativo ou de frequência
  - Usa frequência relativa ao invés de contagem
    - Muda apenas a escala, não o formato
- Histograma de Pareto
  - Usado principalmente para dados nominais
    - Barras têm tamanhos decrescentes
- Histograma bidimensional
  - Associa valores de dois atributos a cada cesta

Histogramas bidimensionais

- Conjunto de dados iris
  - Largura e tamanho das pétalas



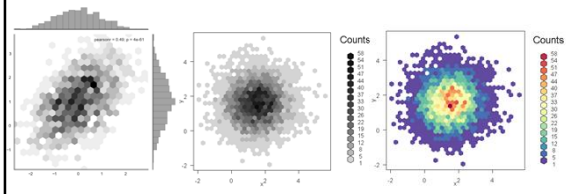
O que mostra?

Histogramas bidimensionais

- Pode ajudar a descobrir fatos interessantes
  - Atributos que estão relacionados
- Pode tornar a visualização mais difícil
  - Algumas das colunas pode ter seus valores escondidos por outras

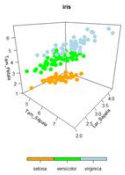
Alternativa bidimensional

- Hexbin
  - Usa tonalidade ao invés de altura



Alternativa tridimensional

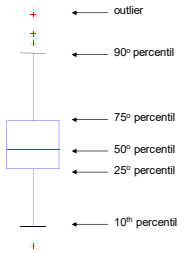
- Permite girar figuras
  - Visualização interativa



Boxplot

- Outra forma de mostrar a distribuição dos valores de um atributo
- Ilustra percentis e outliers

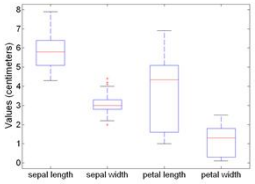
Boxplot



- O que indicam as linhas:
  - Inferiores e superiores da caixa: 25º e 75º percentis
  - Dentro da caixa: 50º percentil
  - Nos extremos:
    - 1,5 x quartis ou
    - 10º e 90º percentis
- Símbolo + indica outlier

Boxplot

- Pode ser usado para comparar atributos
  - Isso pode ser feito para cada classe



Scatter Plot

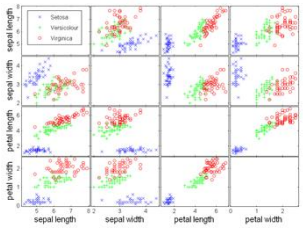
- Usado para ilustrar correlação linear entre variáveis (atributos)
- Cada objeto é associado a uma posição em um gráfico
  - Número de atributos define número de dimensões
    - Valores dos atributos definem sua posição
      - Os valores podem ser inteiros ou reais

Scatter Plot

- Número de atributos (dimensões)
  - Em geral, 2 dimensões
  - Mas também existem com 3 dimensões
- Atributos adicionais podem ser exibidos mesmo com 2 dimensões
  - Usando tamanho, formato e cor nos marcadores que representam os objetos
- Matriz de *scatter plot* resume o relacionamento entre vários pares de atributos

Scatter Plot

- Matriz de scatter plots (conjunto iris)



Diferentes classes são indicadas por diferentes

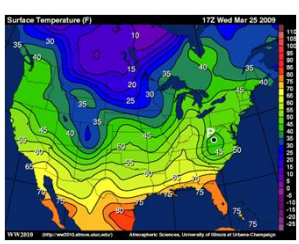
Scatter Plot

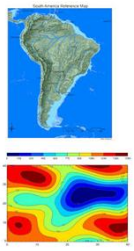
- Se as classes dos objetos são conhecidas
  - *Scatter plot* ilustra o grau com que 2 atributos separam as classes
    - Pode mostrar se é possível separar maioria dos objetos de uma das classes com uma reta

Contour Plots

- Gráficos de contorno
- Utilizados para dados com 3 atributos:
  - Dois atributos indicam uma posição no plano
  - Terceiro atributo assume valor contínuo
    - Dividem um plano em regiões com valor semelhante para o terceiro atributo
    - Podem representar características de regiões
      - Ex.: temperatura, índice pluviométrico, elevação ...

Contour Plots





Muitos Atributos

- Utilizadas para dados com mais que 3 atributos
  - Mostram apenas alguns aspectos dos dados
- Principais técnicas:
  - Gráficos de matrizes (*matrix plots*)
  - Coordenadas paralelas
  - Coordenadas estrela (*star plots*)
  - Heatmaps
  - Faces de *Chernoff*

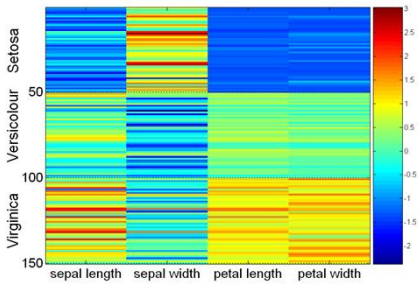
Matrix Plots

- Uma matriz de dados é uma matriz de valores
  - Cada elemento pode ser associado a uma região de uma imagem (pixel)
    - Imagem é representada por uma matriz de pixels
      - Pixel é a unidade básica de uma imagem
  - Brilho ou cor do *pixel* é representado pelo valor do elemento correspondente na matriz

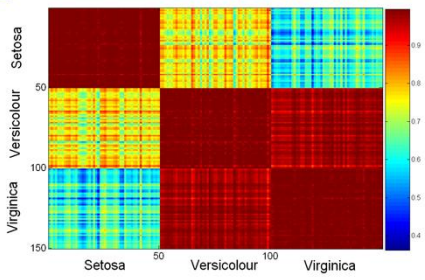
Matrix Plots

- Se as classes são conhecidas, matriz pode ser re-ordenada para deixar juntos objetos de uma mesma classe
  - Fica mais fácil ver que os objetos em uma classe têm valores similares para alguns atributos
  - Atributos com extremos diferentes são padronizados para média 0 e variância 1
    - Para que atributo com maior magnitude não sobressaia no gráfico

Matrix Plot - Iris

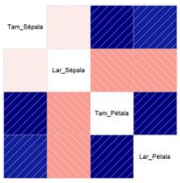


Matrix de correlação - Iris



Corgeam

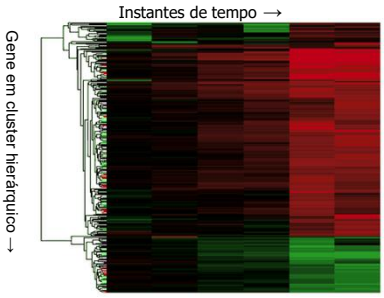
- Mostra de uma forma mais simples a matriz de correlação



Matrix Plots

- Quando as classes não são conhecidas
  - Várias técnicas re-ordenam as linhas e colunas da matriz de similaridade
    - De modo que grupos de objetos e atributos semelhantes fiquem juntos
  - É uma forma simples de agrupamento de dados

Heatmap



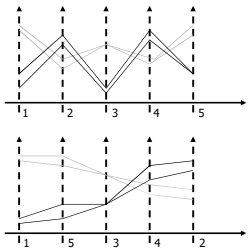
Coordenadas paralelas

- Associam um eixo de coordenada para cada atributo
  - Ao invés de perpendiculares, os eixos são paralelos
  - Cada objeto é representado por uma linha, ao invés de um ponto
  - O valor de cada atributo de um objeto é mapeado para um ponto do eixo associado àquele atributo
    - Os pontos são conectados para formar uma linha, que representa o objeto

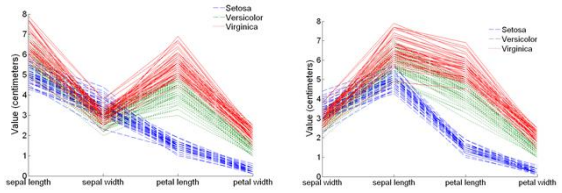
Coordenadas paralelas

- Geralmente, as linhas representando objetos de uma mesma classe estão agrupadas
  - Pelo menos para alguns atributos
- Detecção de padrões depende da ordem dos atributos no gráfico
  - Se as linhas se cruzam bastante, o gráfico pode ficar confuso
  - Pode ser desejável mudar a ordem dos eixos para obter menos cruzamentos

Coordenadas paralelas



Coordenadas paralelas



Ordem original dos atributos

Atributos re-ordenados

Exercício

- Representar visualmente o conjunto de dados ao lado utilizando coordenadas paralelas

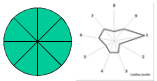
Peso	*Altura*	*Idade*	*Salário*	*Atividade*
70	*168*	*40*	*2000*	*Professor*
58	*185*	*32*	*3500*	*Jogador*
85	*190*	*25*	*3000*	*Jogador*
60	*170*	*34*	*1000*	*Professor*
80	*165*	*37*	*1000*	*Professor*
65	*170*	*26*	*4500*	*Jogador*
90	*190*	*22*	*6000*	*Jogador*
49	*74*	*44*	*1300*	*Professor*
68	*188*	*30*	*3200*	*Professor*
75	*192*	*24*	*4000*	*Jogador*

Star plots

- Gráficos de estrela
- Também permite exibir dados multidimensionais
  - Codificam objetos como ícones ou símbolos que transmitem informação
    - Cada atributo de um objeto é mapeado para uma característica específica do ícone
    - Valor do atributo determina a natureza exata da característica

Star plots

- Usam um eixo para cada atributo
  - Todos os eixos se originam em um ponto central e são igualmente espaçados
    - Como os raios de uma roda

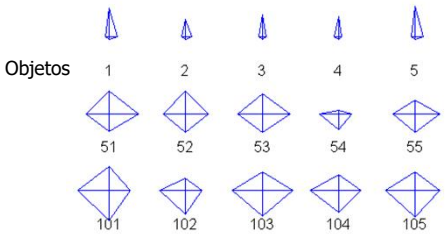


  - Geralmente os valores dos atributos são mapeados para o intervalo [0, 1]

Star plots

1 Cada valor de atributo do objeto é convertido para uma fração que representa sua posição entre os valores mínimos e máximos do atributo
2 Essa fração é mapeada para um ponto no eixo correspondente a esse atributo
3 Cada ponto é conectado com um segmento de linha ao ponto no eixo precedente ou seguinte, formando um polígono
4 Tamanho e formato desse polígono ilustra visualmente o valor dos atributos do objeto
5 Um conjunto separado de eixos é usado para cada objeto

Star plots - Iris



Exercício

- Representar visualmente o conjunto de dados ao lado usando *star plots*

Peso	*Altura*	*Idade*	*Salário*	*Atividade*
70	*168*	*40*	*2000*	*Professor*
58	*185*	*32*	*3500*	*Jogador*
85	*190*	*25*	*3000*	*Jogador*
60	*170*	*34*	*1000*	*Professor*
80	*165*	*37*	*1000*	*Professor*
65	*170*	*26*	*4500*	*Jogador*
90	*190*	*22*	*6000*	*Jogador*
49	*74*	*44*	*1300*	*Professor*
68	*188*	*30*	*3200*	*Professor*
75	*192*	*24*	*4000*	*Jogador*

Star plots

- O que acontece se o número de classes for muito grande?

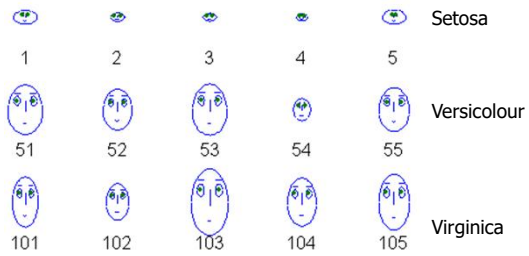
Faces de Chernoff

- Criado por Herman Chernoff
- Mapeia os valores dos atributos para imagens mais familiares: faces
- Baseia-se na habilidade humana de distinguir faces

Faces de Chernoff

- Cada objeto é representado por uma face
- Cada atributo é associado a uma característica específica de uma face
- Valor do atributo é usado para determinar a maneira como a característica é expressa
  - Aparência da característica facial correspondente
    - Ex.: Formato da face pode se tornar mais alongado quando o valor da característica correspondente aumenta

Faces de Chernoff



Faces de Chernoff

Atributo	Característica facial
Tamanho sépala	Tamanho da face
Largura sépala	Tamanho do arco relativo testa / boca
Tamanho pétala	Formato da testa
Largura pétala	Formato da boca

Outras características da face recebem valores *default*

Faces de Chernoff

- Representar visualmente o conjunto de dados ao lado usando faces de Chernoff

Peso	Altura	Idade	Salário	Atividade
70	168	40	2000	Professor
58	185	32	3500	Jogador
85	190	25	3000	Jogador
60	170	34	1000	Professor
80	165	37	1000	Professor
65	170	26	4500	Jogador
90	190	22	6000	Jogador
49	74	44	1300	Professor
68	188	30	3200	Professor
75	192	24	4000	Jogador

Observações

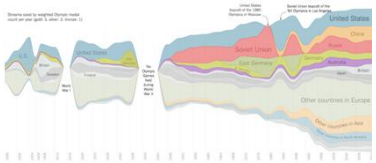
- Star plots e Chernoff faces não escalam bem
  - Pouco uso para vários problemas de mineração de dados
  - Mas podem ser úteis para rapidamente comparar pequenos conjuntos de objetos selecionados por outras técnicas

Word cloud

- Nuvem de palavras
  - Usada para mostrar termos mais frequentemente associados a um tema

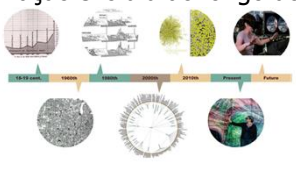
Gráficos de fluxos

- Mudança de valores ao longo do tempo
  - Streamographs
  - Medalhas ganhas em olimpíadas



Evolução

- Visualização evoluiu ao longo do tempo



E. Olshannikova, A. Ometov, Y. Koucheryavy and T. Olsson, Visualizing Big Data with augmented and virtual reality: challenges and research agenda Journal of Big Data2015 - http://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0031-2

Conclusão

- Técnicas de visualização
  - Poucos Atributos
  - Muitos atributos
- Exermplos
- Visualizar fluxos de dados
- Visualização para big data
- Visualização interativa

Perguntas