Agrupamento
(Conceitos Básicos)

Tsang Ing Ren
George Darmiton da Cunha Cavalcanti
CIn/UFPE

Roteiro

- Conceitos Básicos
- Estágios do Processo de Agrupamento
- Áreas de Aplicação Básica para Agrupamento
- Tipos de Características
- Definições de Agrupamento
- Medidas de Proximidades
  - Entre vetores
  - Entre vetores e um conjunto
  - Entre conjuntos

Conceitos Básicos

- Em agrupamento ou aprendizagem não-supervisionada, uma base de treinamento com classes identificadas não está à disposição

- O objetivo se torna agrupar os dados em um número de grupos *razoável*. Com isto, pode-se descobrir similaridades e diferenças entre os dados disponíveis.

- Aplicações:
  - Engenharia
  - Bioinformática
  - Ciências Sociais
  - Medicina
  - Mineração de Dados e Web

Conceitos Básicos

- Para aplicar técnicas de agrupamento em um conjunto de dados, é necessário adotar um critério de agrupamento

- Critérios de agrupamentos diferentes, em geral, geram grupos diferentes

Conceitos Básicos

(um exemplo simples)

Cachorro

Gato

Ovelha

Peixe-dourado

Pardal

Lagarto

Víbora

Tubarão

Gaivota

Sapo

Tainha

Conceitos Básicos

(um exemplo simples)

peixe dourado,
tainha, tubarão

ovelha, cachorro,
pardal, gato, gaivota,
largato, víbora, sapo

1. Dois grupos
2. Critério de agrupamento:
   Existência de pulmão.

ovelha, cachorro,
gato, gaivota, víbora,
pardal, lagarto

Peixe-dourado,
tainha, tubarão

Sapo

1. Três grupos
2. Critério de agrupamento:
   O ambiente em que os
   animais vivem.

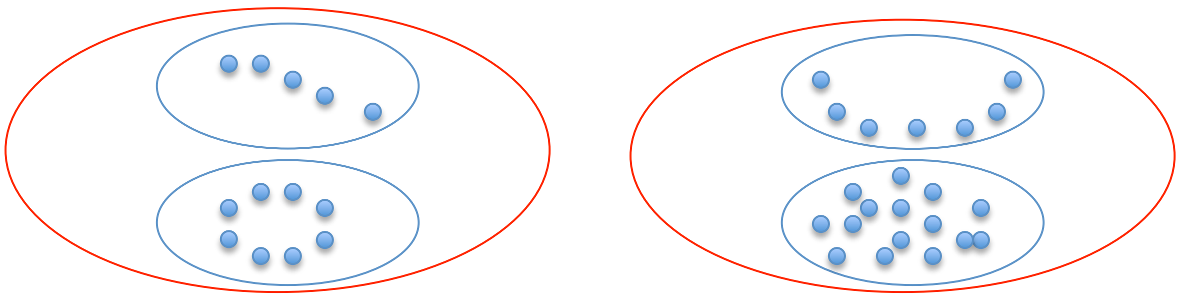
Estágios do Processo de Agrupamento

- **Seleção de Características**
  - características ricas em informação - Parcimônia (poupar)
- **Medidas de Proximidades**
  - quantifica o termo similar e dissimilar
- **Critério de Agrupamento**
  - consiste numa função de custo ou algum tipo de regra
- **Algoritmo de Agrupamento**
  - consiste num conjunto de passos seguidos que revela a estrutura, baseado na medida de similaridade e no critério adotado
- **Validação dos Resultados**
  - utilizar testes apropriados
- **Interpretação dos Resultados**
  - integração do resultado com evidências experimentais

Estágios do Processo de Agrupamento

- Dependendo da medida de similaridade, do critério e do algoritmo de agrupamento têm-se diferentes grupos como resultado.
- Subjetividade é intrínseco nestes problemas de agrupamento.

- Um exemplo simples: Quantos grupos?



2 ou 4?

Áreas de Aplicação Básica para Agrupamento

- Redução de Dados
  - Todos os vetores de dados dentro de um grupo são substituídos (representados) por representantes do grupo
- Geração de Hipóteses
  - Análise de grupo é utilizado no conjunto de dados para inferir alguma hipótese em relação à natureza dos dados
- Avaliar Hipóteses
  - Análise de grupo é utilizado para verificar a validade de uma hipótese específica. Ex: "uma grande companhia investe fora"
- Predição Baseado em Grupos
  - Análise de grupo é aplicada no conjunto de dados disponível, e os grupos resultantes são qualificados baseado nas características dos padrões que eles foram formados

Definições de Agrupamento

- Agrupamento Rígido (*Hard Clustering*)
  - Cada ponto pertence a um único grupo.
  - Seja $X = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$
  - Um m-agrupamento R de X é definido como uma partição de X em m conjuntos (grupos), $C_1, C_2, \ldots, C_m$ de forma que

$$C_i \neq \varnothing,\ i = 1,2,\ldots,m$$

$$C_i \cap C_j = \varnothing,\ \ i \neq j,\ \ i,j = 1,2,\ldots,m$$

$$\bigcup_{i=1}^{m} C_i = X$$

Além disso, dados em C_i, são mais similares entre si e menos similares do que o resto dos dados do grupo. Quantificar o termo similar-dissimilar depende dos tipos de grupos esperados que estão por trás das estruturas de X.

Definições de Agrupamento

- **Agrupamento Difuso (*Fuzzy Clustering*)**
  - Cada ponto pertence a todos os grupos e essa pertinência é definida por um certo grau.
  - Um agrupamento difuso de X em m grupos é caracterizado por m funções

$$u_j : \underline{x} \rightarrow [0,1], \;\; j = 1,2,\ldots,m$$

$$\sum_{j=1}^{m} u_j(\underline{x}_i) = 1, \;\; i = 1,2,\ldots,N$$

$$0 < \sum_{i=1}^{N} u_j(\underline{x}_i) < N, \;\; j = 1,2,\ldots,m$$

Definições de Agrupamento

São conhecidos como funcões de pertinências. Então, cada $\underline{x}_i$, pertence a qualquer grupo "até um certo grau", dependendo do valor de:

$$u_j(\underline{x}_i), \; j = 1,2,\ldots,m$$

$u_j(\underline{x}_i)$ perto de $1 \Rightarrow$ alto grau de pertinência de $\underline{x}_i$ para o grupo j.

$u_j(\underline{x}_i)$ perto de $0 \Rightarrow$ baixo grau de pertinência.

Definições de Agrupamento

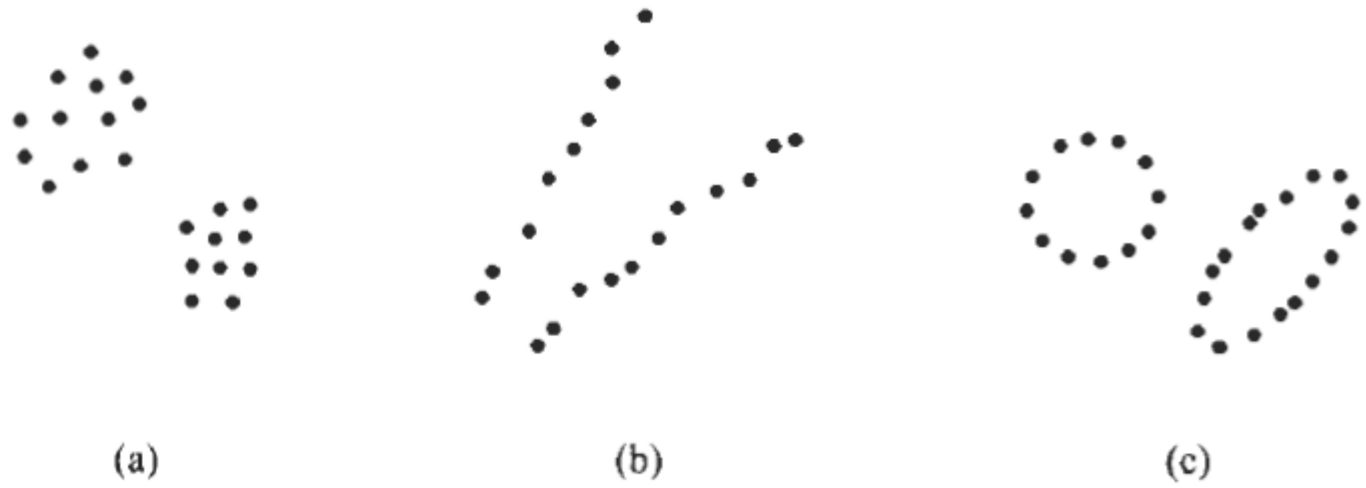


FIGURE 11.3: (a) Compact clusters. (b) Elongated clusters. (c) Spherical and ellipsoidal clusters.

Medidas de Proximidades

- Entre Vetores

  - Medida de dissimilaridade (entre vetores de X) é uma função

$$d: \; X \times X \longrightarrow \Re$$

com as seguintes propriedades

$$\exists d_0 \in \Re: \; -\infty < d_0 \leq d(\underline{x},\underline{y}) < +\infty, \; \forall \underline{x},\underline{y} \in X$$

$$d(\underline{x},\underline{x}) = d_0, \; \forall \underline{x} \in X$$

$$d(\underline{x},\underline{y}) = d(\underline{y},\underline{x}), \; \forall \underline{x},\underline{y} \in X$$

Medidas de Proximidades

Se além disso

- $d(\underline{x},\underline{y}) = d_0 \;\; if \;\; and \;\; only \;\; if \;\; \underline{x} = \underline{y}$
- $d(\underline{x},\underline{z}) \leq d(\underline{x},\underline{y}) + d(\underline{y},\underline{z}), \;\; \forall \underline{x},\underline{y},\underline{z} \in X$

(inequalidade triangular)

d é chamado de **métrica** de dissimilaridade

Medidas de Proximidades

Example 11.2. Let us consider the well-known Euclidean distance, d_2

$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{l}(x_i - y_i)^2}$$

where $\boldsymbol{x}, \boldsymbol{y} \in X$ and x_i, y_i are the ith coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. This is a dissimilarity measure on X, with $d_0 = 0$; that is, the minimum possible distance between two vectors of X is 0. Moreover, the distance of a vector from itself is equal to 0. Also, it is easy to observe that $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$.

The preceding arguments show that the *Euclidean distance is a dissimilarity measure*. In addition, the Euclidean distance between two vectors takes its minimum value $d_0 = 0$, when the vectors coincide. Finally, it is not difficult to show that the triangular inequality holds for the Euclidean distance (see Problem 11.2). Therefore, the Euclidean distance is a metric dissimilarity measure.

It is worth pointing out that for other measures the values d_0 (s_0) may be positive or negative.

Medidas de Proximidades

Medida de similaridade (entre vetores de X) é uma função

$$s: \; X \times X \longrightarrow \Re$$

com as seguintes propriedades

$$\exists s_0 \in \Re : \; -\infty < s(\underline{x},\underline{y}) \leq s_0 < +\infty, \; \forall \underline{x},\underline{y} \in X$$

$$s(\underline{x},\underline{x}) = s_0, \; \forall \underline{x} \in X$$

$$s(\underline{x},\underline{y}) = s(\underline{y},\underline{x}), \; \forall \underline{x},\underline{y} \in X$$

Medidas de Proximidades

Se além disso

$$s(\underline{x},\underline{y}) = s_0 \ \textit{if and only if} \ \underline{x} = \underline{y}$$

$$s(\underline{x},\underline{y})s(\underline{y},\underline{z}) \leq [s(\underline{x},\underline{y}) + s(\underline{y},\underline{z})]s(\underline{x},\underline{z}), \ \forall \underline{x},\underline{y},\underline{z} \in X$$

s é chamado de **métrica** de similaridade

Medidas de Proximidades

- Entre Conjuntos

Seja $D_i \subset X,\ i=1,\ldots,k$ and $U=\{D_1,\ldots,D_k\}$

A medida de proximidade $\wp$ em *U* é uma função

$$\wp : U \times U \longrightarrow \Re$$

Medidas de Proximidades

Example 11.3. Let $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ and $U = \{\{x_1, x_2\}, \{x_1, x_4\}, \{x_3, x_4, x_5\}, \{x_1, x_2, x_3, x_4, x_5\}\}$. Let us define the following dissimilarity function:

$$d_{min}^{ss}(D_i, D_j) = \min_{x \in D_i,\ y \in D_j} d_2(x, y)$$

where d_2 is the Euclidean distance between two vectors and $D_i, D_j \in U$.

The minimum possible value of d_{min}^{ss} is $d_{min,0}^{ss} = 0$. Also, $d_{min}^{ss}(D_i, D_i) = 0$, since the Euclidean distance between a vector in D_i and itself is 0. In addition, it is easy to see that the commutative property holds. Thus, this dissimilarity function is a measure. It is not difficult to see that d_{min}^{ss} is not a metric. Indeed, Eq. (11.7) for subsets of X does not hold in general, since the two sets D_i and D_j may have an element in common. Consider for example the two sets $\{x_1, x_2\}$ and $\{x_1, x_4\}$ of U. Although they are different their distance d_{min}^{ss} is 0, since they both contain x_1.

Medidas de Proximidades

- Entre Dois Pontos – Vetores de Valores Reais
  - Medida de dissimilaridade (DMs)
    - Métrica DMs *"Weighted"* (ponderada) l_p

$$d_p(\underline{x},\underline{y}) = \left(\sum_{i=1}^{l} w_i \,|\, x_i - y_i \,|^p \right)^{1/p}$$

- Casos interessantes são obtidos para:
  - p=1 (*weighted Manhattan* norm)
  - p=2 (*weighted Euclidean* norm)
  - p=∞ ($d_\infty(\underline{x},\underline{y})=\max_{1\leq i\leq l} w_i|x_i-y_i|$)

Medidas de Proximidades

Outras medidas

$$d_G(\underline{x},\underline{y}) = -\log_{10}\left(1 - \frac{1}{l}\sum_{j=1}^{l}\frac{|x_j - y_j|}{b_j - a_j}\right)$$

- b_j e a_j são os valores máximo e mínimo da j-ésima característica, entre os vetores de X (dependência do conjunto de dados atual)

$$d_Q(\underline{x},\underline{y}) = \sqrt{\frac{1}{l}\sum_{j=1}^{l}\left(\frac{x_j - y_j}{x_j + y_j}\right)^2}$$

Medidas de Proximidades

Example 11.4. Consider the three-dimensional vectors $\boldsymbol{x} = [0, 1, 2]^T$, $\boldsymbol{y} = [4, 3, 2]^T$. Then, assuming that all w_i's are equal to 1, $d_1(\boldsymbol{x}, \boldsymbol{y}) = 6$, $d_2(\boldsymbol{x}, \boldsymbol{y}) = 2\sqrt{5}$, and $d_\infty(\boldsymbol{x}, \boldsymbol{y}) = 4$. Notice that $d_\infty(\boldsymbol{x}, \boldsymbol{y}) < d_2(\boldsymbol{x}, \boldsymbol{y}) < d_1(\boldsymbol{x}, \boldsymbol{y})$.

Assume now that these vectors belong to a data set X that contains N vectors with maximum values per feature 10, 12, 13 and minimum values per feature 0, 0.5, 1, respectively. Then $d_G(\boldsymbol{x}, \boldsymbol{y}) = 0.0922$. If, on the other hand, $\boldsymbol{x}$ and $\boldsymbol{y}$ belong to an X' with the maximum (minimum) values per feature being 20, 22, 23 (-10, -9.5, -9), respectively, then $d_G(\boldsymbol{x}, \boldsymbol{y}) = 0.0295$.

Finally, $d_Q(\boldsymbol{x}, \boldsymbol{y}) = 0.6455$.

Medidas de Proximidades

Medida de similaridade (SMs)

- Produto Interno

$$s_{inner}(\underline{x},\underline{y}) = \underline{x}^T \underline{y} = \sum_{i=1}^{l} x_i y_i$$

- Medida de Tanimoto

$$s_T(\underline{x},\underline{y}) = \frac{\underline{x}^T \underline{y}}{\|\underline{x}\|^2 + \|\underline{y}\|^2 - \underline{x}^T \underline{y}}$$

- Outra medida

$$s_C(\underline{x},\underline{y}) = 1 - \frac{d_2(\underline{x},\underline{y})}{\|\underline{x}\| + \|\underline{y}\|}$$

Medidas de Proximidades

- Funções de Proximidades entre um Vetor e um Conjunto

  Seja $X=\{\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_N\}$ e $C \subset X$, $\underline{x} \in X$

  – ***Todos os pontos de C contribuem para a definição de*** $\wp(\underline{x}, C)$

    - Max função de proximidade

$$\wp^{ps}_{\max}(\underline{x},C) = \max_{\underline{y}\in C} \wp(\underline{x},\underline{y})$$

    - Min função de proximidade

$$\wp^{ps}_{\min}(\underline{x},C) = \min_{\underline{y}\in C} \wp(\underline{x},\underline{y})$$

    - Média função de proximidade

$$\wp^{ps}_{avg}(\underline{x},C) = \frac{1}{n_C}\sum_{\underline{y}\in C} \wp(\underline{x},\underline{y})$$

Medidas de Proximidades

- *Um representante(s) de C, r_c, contribui para a definição de $\wp(x, C)$*

Neste caso: $\wp(\underline{x}, C) = \wp(\underline{x}, \underline{r}_C)$

Representantes típicos são:

- O vetor médio:

$$\underline{m}_p = \left(\frac{1}{n_C}\right) \sum_{\underline{y} \in C} \underline{y}$$

- A média do centro:

$$\underline{m}_C \in C: \sum_{\underline{y} \in C} d(\underline{m}_C, \underline{y}) \leq \sum_{\underline{y} \in C} d(\underline{z}, \underline{y}), \ \forall \underline{z} \in C$$

- A mediana do centro:

$$\underline{m}_{med} \in C: \ med(d(\underline{m}_{med}, \underline{y}) \mid \underline{y} \in C) \leq med(d(\underline{z}, \underline{y}) \mid \underline{y} \in C), \ \forall \underline{z} \in C$$

d: a dissimilarity measure

Medidas de Proximidades

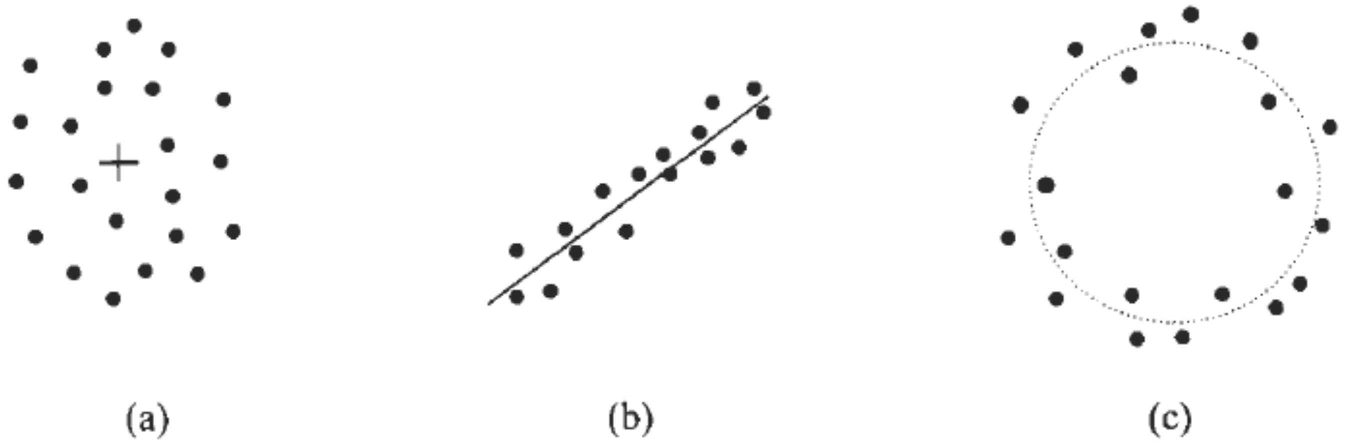


FIGURE 11.7: (a) Compact cluster. (b) Hyperplanar (linear) cluster. (c) Hyper-spherical cluster.

Medidas de Proximidades

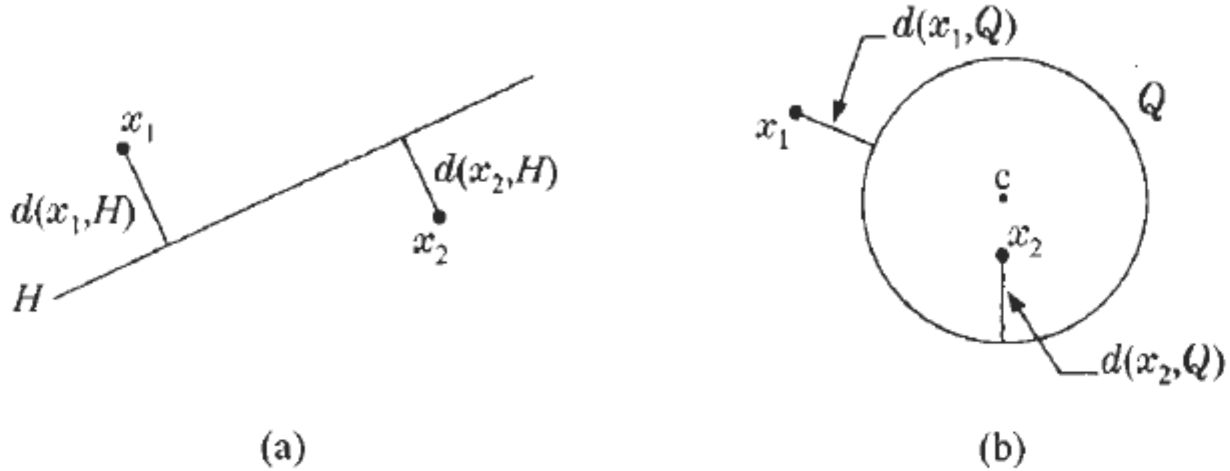


FIGURE 11.9: (a) Distance between a point and a hyperplane. (b) Distance between a point and hypersphere.

Medidas de Proximidades

- Funções de Proximidades entre Conjuntos
  - Seja $X=\{\underline{x}_1,\ldots,\underline{x}_N\}$, D_i, $D_j \subset X$ e $n_i=|D_i|$, $n_j=|D_j|$
  - Todos os pontos de cada conjunto contribuem para $\wp(D_i,D_j)$
    - Max função de proximidade (medida mas não métrica, se apenas $\wp$ é uma medida de similaridade)

$$\wp^{ss}_{\max}(D_i,D_j) = \max_{\underline{x}\in D_i,\underline{y}\in D_j} \wp(\underline{x},\underline{y})$$

    - Min função de proximidade (medida mas não métrica, se apenas $\wp$ é uma medida de dissimilaridade)

$$\wp^{ss}_{\min}(D_i,D_j) = \min_{\underline{x}\in D_i,\underline{y}\in D_j} \wp(\underline{x},\underline{y})$$

Medidas de Proximidades

- Média função de proximidade (não é uma medida, até se $\wp$ for uma medida)

$$\wp_{avg}^{ss}(D_i, D_j) = \left(1 / n_i n_j \right) \sum_{x \in D_i} \sum_{x \in D_j} \wp(\underline{x}, \underline{y}) \quad (n_i \text{ é a cardinalidade})$$

Medidas de Proximidades

- Observações

Escolhas de diferentes de função de proximidade entre conjuntos pode levar a agrupamentos totalmente diferentes.

Diferentes medidas de proximidades entre vetores na mesma função de proximidade entre conjuntos pode levar a agrupamentos totalmente diferente.

A única forma de alcançar um agrupamento correto é

- por tentativa e erro e,
- tomando em consideração a opinião de um especialista na área de aplicação.