Agrupamento II
(Algoritmos Sequenciais)

Tsang Ing Ren
George Darmiton da Cunha Cavalcanti
CIn/UFPE

Roteiro

- Introdução
- Categorias de Algoritmos de Agrupamento
- Algoritmo de Agrupamento Sequencial
  - Algoritmo de Agrupamento Sequencial Básico (BSAS)
  - MBSAS, uma modificação do BSAS
  - O Algoritmo maxmin
  - *"Two-threshold sequencial scheme (TTSAS)"*

Introdução

Algoritmos de agrupamento

– Número de possíveis agrupamentos

Seja $X=\{\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_N\}$.

Questão: De quantas maneiras N pontos podem ser organizados em m grupos?

Resposta:

$$S(N,m)=\frac{1}{m!}\sum_{i=0}^{m}(-1)^{m-1}\binom{m}{i}i^N$$

Exemplos

$$S(15,3)=2\ 375\ 101 \qquad S(20,4)=45\ 232\ 115\ 901$$

$$S(100,5)=10^{68}!!$$

Introdução

- Solução:
  - Considere apenas uma pequena fração de grupos de *X* e selecione uma quantidade "razoável" dentre eles.
    - Questão 1: Qual fração dos grupos devemos considerar?
    - Questão 2: O que "razoável" significa?
    - A resposta depende do algoritmo de agrupamento específico e do critério específico que é adotado.

Categorias Principais de Algoritmos de Agrupamento

- Sequencial
- Hierárquico
- Otimização da função de custo
- Outros Modelos

Sequencial

- Um único agrupamento é produzido
- Os dados são apresentados ao algoritmo uma ou poucas vezes
- O resultado final, geralmente, depende da ordem que os dados são apresentados ao algoritmo
- Este esquema tende a gerar agrupamentos compactos e em forma de elipse ou de circunferência, dependendo da distância usada.

Hierárquico

- Uma seqüência de agrupamentos aninhados (“*nested*”) é produzido.
  - Aglomerativo
  - Divisivo

Otimização da função de custo

- Para a maioria dos casos um *único* agrupamento é obtido.
  - Agrupamento Rígido (*Hard Clustering*)
  - Agrupamento Difuso (*Fuzzy Clustering*) – Cada ponto pertence a mais de um grupo simultaneamente

Outros Modelos

- Algoritmos baseado na teoria dos grafos
- Algoritmos de aprendizagem competitiva (modelos básicos de aprendizagem competitiva, mapas auto-organizadas Kohonen)
- Algoritmos de agrupamento por sub-espaço
- Algoritmos de agrupamento de morfologia binária

Algoritmo de Agrupamento Sequencial

- As características comuns compartilhadas por estes algoritmos são:
  - Necessidade de um ou poucos passos
  - O número de grupos não é conhecido *a priori*, exceto (possivelmente) um limite superior, *q*
  - Os grupos são definidos com ajuda de:
    - Uma distância apropriada $d(\underline{x}, C)$ entre um ponto e um agrupamento
    - Um limiar Θ associado à distância

Algoritmo de Agrupamento Sequencial

- *Basic Sequential Clustering Algorithm (BSAS)*

  - m=1 \{number of clusters}\
  - $C_m=\{\underline{x}_1\}$
  - For i=2 to N
    - Find C_k: $d(\underline{x}_i, C_k)=min_{1\leq j\leq m} d(\underline{x}_i, C_j)$
    - If $(d(\underline{x}_i, C_k)>\Theta)$ AND $(m<q)$ then
      - $m=m+1$
      - $C_m=\{\underline{x}_i\}$
    - Else
      - $C_k=C_k \cup \{\underline{x}_i\}$
      - Where necessary q, update representatives (*)
    - End {if}
  - End {for}

Θ limiar de dissimilaridade, m número grupos, q limiar para o número de grupos

Algoritmo de Agrupamento Sequencial

- Observações:
  - A ordem de apresentação dos dados no algoritmo é importante para o resultado do agrupamento. Diferente ordem de apresentação pode gerar um resultado totalmente diferente no agrupamento, em termos do número de grupos como também nos grupos em si
  - No BSAS, a decisão para cada vetor $\underline{x}$ é alcançada antes da formação final do grupo
  - BSAS realiza um único passo nos dados. A complexidade é *O(N)*
  - Se grupos são representados por pontos representativos, agrupamentos compactos são favorecidos.

Algoritmo de Agrupamento Sequencial

FIGURE 12.1: Three clusters are formed by the feature vectors. When q is constrained to a value less than 3, the BSAS algorithm will not be able to reveal them.

Algoritmo de Agrupamento Sequencial

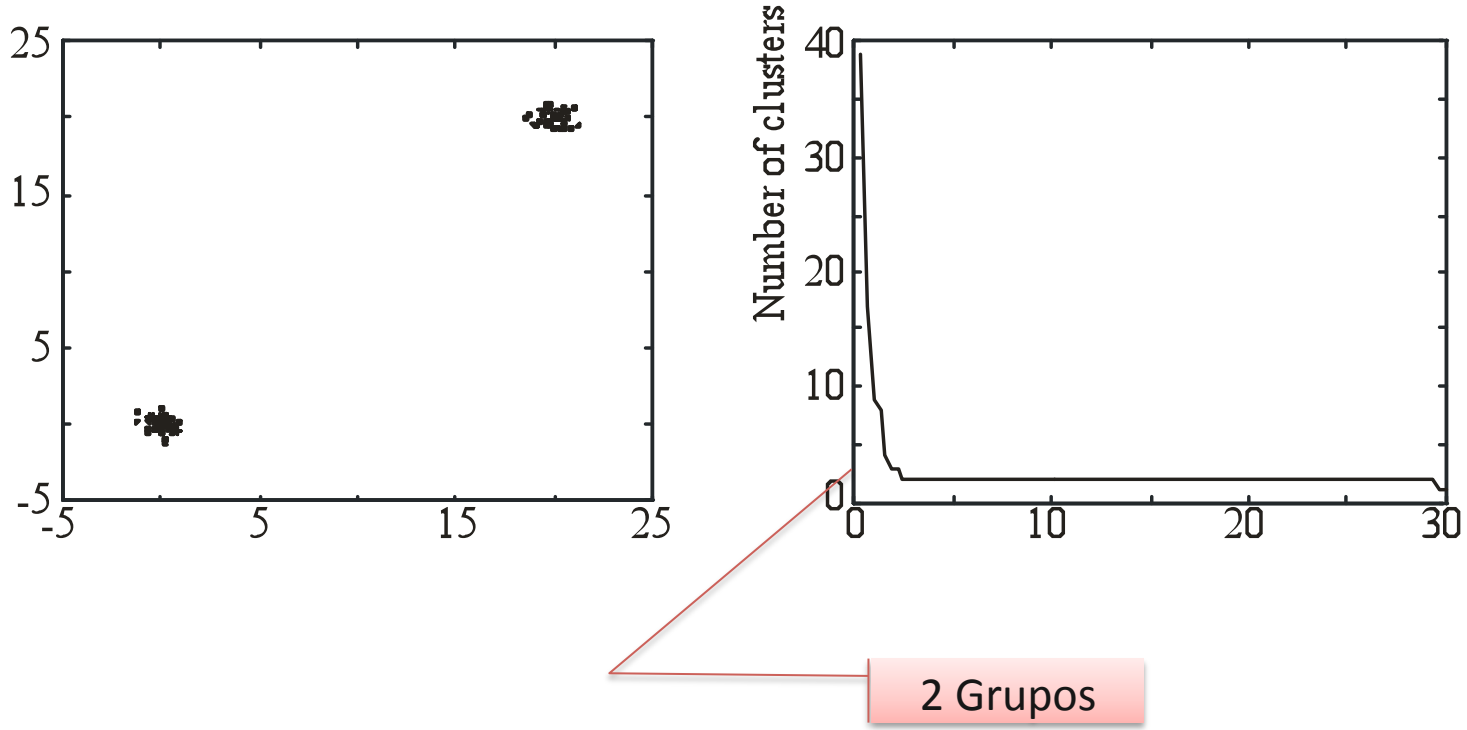
- Estimando o número de grupos para um conjunto de dados:

Seja $BSAS(\Theta)$, denota o algoritmo $BSAS$ quando o limiar de dissimilaridade é Θ.

- For $\Theta=a$ to b step c
  - Rode s vezes $BSAS(\Theta)$, cada vez apresentando os dados em uma ordem diferente.
  - Estime o número de grupos m_{Θ}, como o número mais freqüente resultante de rodar $BSAS(\Theta)$ s vezes.
- Next Θ

Algoritmo de Agrupamento Sequencial

Construir um gráfico m_Θ versus Θ e identificar o número de grupos m como um dos correspondentes à região larga mais plana no gráfico abaixo:



Algoritmo de Agrupamento Sequencial

- MBSAS, uma modificação do BSAS
  - No BSAS, a decisão para o vetor de dado x é obtida antes da formação final do grupo, o qual é determinado após a apresentação de todos os vetores ao algoritmo
  - MBSAS lida com este problema, a um custo de apresentar os dados duas vezes ao algoritmo
  - MBSAS consiste em:
    - Uma fase de determinação do grupo (primeira passada nos dados), que é o mesmo que BSAS exceto que nenhum vetor é designado para um grupo já formado
    - Uma fase de classificação de padrões (segunda passada nos dados), cada um dos vetores não designados é associado ao seu grupo mais próximo

Algoritmo de Agrupamento Sequencial

- *Modified Basic Sequential Clustering Algorithm (BSAS)*
- Determinação do Grupo

  - m=1 \{number of clusters}\
  - $C_m=\{\underline{x}_1\}$
  - For i=2 to N
    - Find C_k: $d(\underline{x}_i, C_k)=min_{1\le j\le m} d(\underline{x}_i, C_j)$
    - If $(d(\underline{x}_i, C_k)>\Theta)$ AND $(m<q)$ then
      - $m=m+1$
      - $C_m=\{\underline{x}_i\}$
    - End {if}
  - End {for}

Algoritmo de Agrupamento Sequencial

- *Modified Basic Sequential Clustering Algorithm (BSAS)*
- Classificação do Padrão

  - For i=1 to N
    - If $\underline{x}_i$ has not been assigned to a cluster, then
      - Find C_k: $d(\underline{x}_i, C_k)=min_{1\le j\le m} d(\underline{x}_i, C_j)$
      - $C_k=C_k\cup\{\underline{x}_i\}$
      - Where necessary q, update representatives (*)
    - End {if}
  - End {for}

Θ limiar de dissimilaridade , m número grupos, q limiar para o número de grupos

Algoritmo de Agrupamento Sequencial

- Observações:
  - No MBSAS, a decisão do vetor x durante a fase de classificação de padrão é obtida tomando em consideração todos os grupos
  - MBSAS é sensível a ordem de apresentação dos vetores
  - MBSAS requer dois passos nos dados. Sua complexidade é O(N)

Algoritmo de Agrupamento Sequencial

- O algoritmo *maxmin*

Seja W um conjunto de pontos escolhidos para formar grupos até a etapa atual da iteração. A formação dos grupos é realizado da seguinte forma:

- Para cada $\underline{x} \in X\text{-}W$ determine $d_x = min_{\underline{z} \in W}\, d(\underline{x},\underline{z})$
- Determine $\underline{y}$: $d_y = max_{\underline{x} \in X\text{-}W} d_x$
- If d_y é maior do que o limiar prefixado então
  - Este vetor forma um novo grupo
- Else
  - A fase de determinação do grupo do algoritmo termina.
- End {if}

Após a formação dos grupos, cada vetor não designado é associado ao seu grupo mais próximo.

Algoritmo de Agrupamento Sequencial

- Observações:
  - O algoritmo maxmin é computacionalmente mais custoso do que MBSAS
  - Entretanto, é de esperar que produza um melhor resultado no agrupamento

Algoritmo de Agrupamento Sequencial

- *Two-threshold sequencial scheme (TTSAS)*
  - A formação de agrupamentos, assim como a designação dos vetores para o grupo, é feito concorrentemente (como BSAS e diferentemente de MBSAS)
  - Dois limiares Θ_1 and Θ_2 ($\Theta_1<\Theta_2$) são empregados
  - A idéia geral é a seguinte:
    - Se a distância $d(\underline{x},C)$ de $\underline{x}$ ao seu grupo mais próximo, C, é maior que Θ_2 então:
      - Um novo grupo representado por $\underline{x}$ é formado.
    - Senão se $d(\underline{x},C)<\Theta_1$ então
      - $\underline{x}$ é designado para C.
    - Senão
      - A decisão é adiada para um estágio posterior.
    - Fim {Se}

Algoritmo de Agrupamento Sequencial

Os vetores não designados são apresentados interativamente para o algoritmo até que todos são classificados.

- Observações:
  - Na prática, alguns poucos passos (≥ 2) do conjunto de dados são requeridos
  - TTSAS é menos sensível a ordem de apresentação dos dados, em comparação a BSAS

Algoritmo de Agrupamento Sequencial

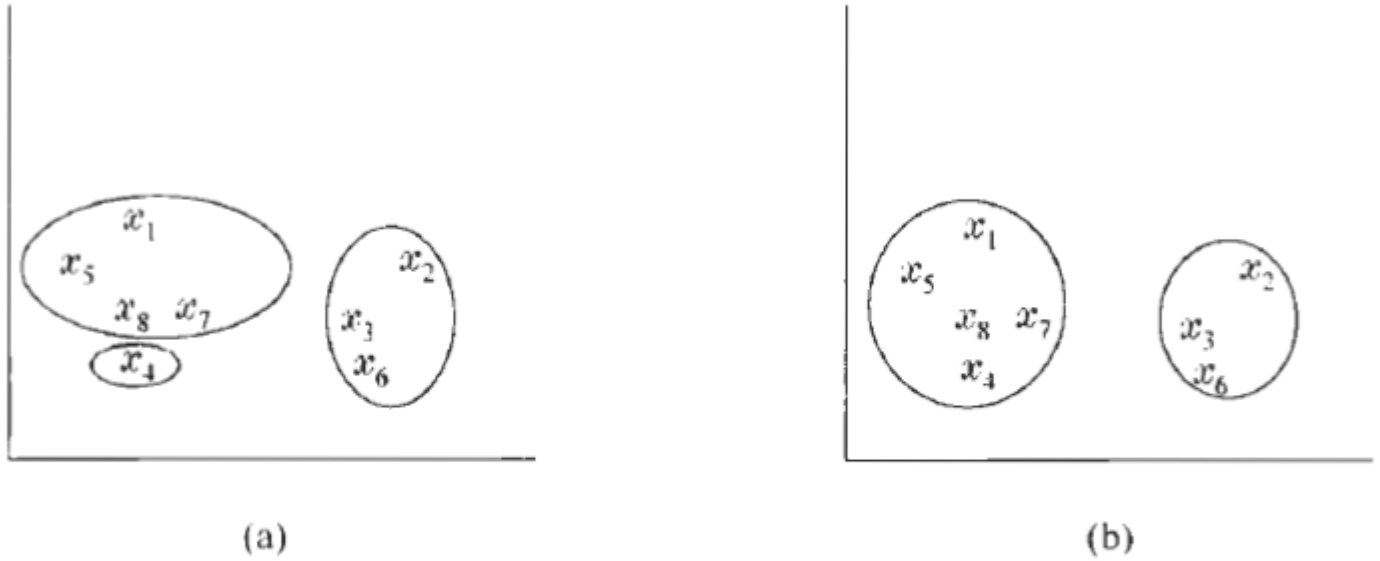


FIGURE 12.3: (a) The clustering produced by the MBSAS. (b) The clustering produced by the TTSAS.

Algoritmo de Agrupamento Sequencial

- Estágio de Refinamento:

O problema da proximidade dos grupos (*closeness of cluster*): *"Em todos os algoritmos anteriores pode acontecer que dois grupos se encontram muito perto um do outro".*

– Um procedimento simples de unir (*merging procedure*)

- (*A*) Find C_i, C_j ($i<j$) such that $d(C_i,C_j)=min_{k,r=1,\ldots,m,k\neq r}d(C_k,C_r)$
- If $d(C_i,C_j)\leq M_1$ then \{ M_1 is a user-defined threshold \}
  - Merge C_i, C_j to C_i and eliminate C_j .
  - If necessary, update the cluster representative of C_i .
  - Rename the clusters $C_{j+1},\ldots,C_m$ to $C_j,\ldots,C_{m-1}$, respectively.
  - $m=m-1$
  - Go to (*A*)
- Else
  - Stop
- End {if}

Algoritmo de Agrupamento Sequencial

- O problema da sensibilidade em relação à ordem de apresentação dos dados:

"Um vetor $\underline{x}$ pode ser designado para um grupo C_i no estágio atual mas outro grupo C_j pode ser formado no estágio mais adiante, e este está mais perto de $\underline{x}$"

- Um procedimento simples de re-designação

  - For $i=1$ to N
    - Find C_j such that $d(\underline{x}_i, C_j)=min_{k=1,\ldots,m} d(\underline{x}_i, C_k)$
    - Set $b(i)=j$ \{ $b(i)$ is the index of the cluster that lies closet to $\underline{x}_i$ \}
  - End {for}
  - For $j=1$ to m
    - Set $C_j=\{\underline{x}_i \in X: b(i)=j\}$
    - If necessary, update representatives
  - End {for}