Agrupamento de Dados

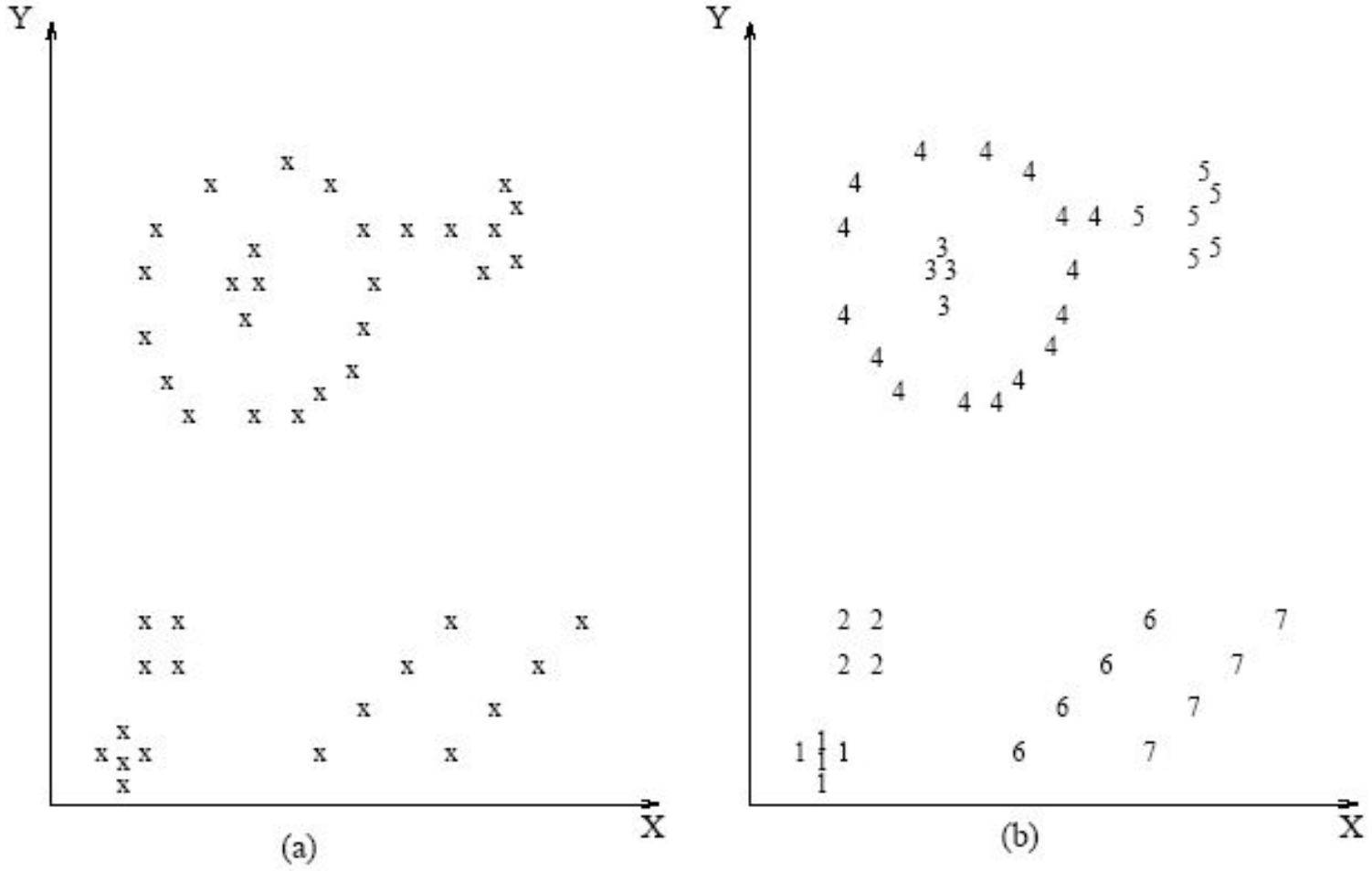
Ricardo Prudêncio

Clustering (Agrupamento)

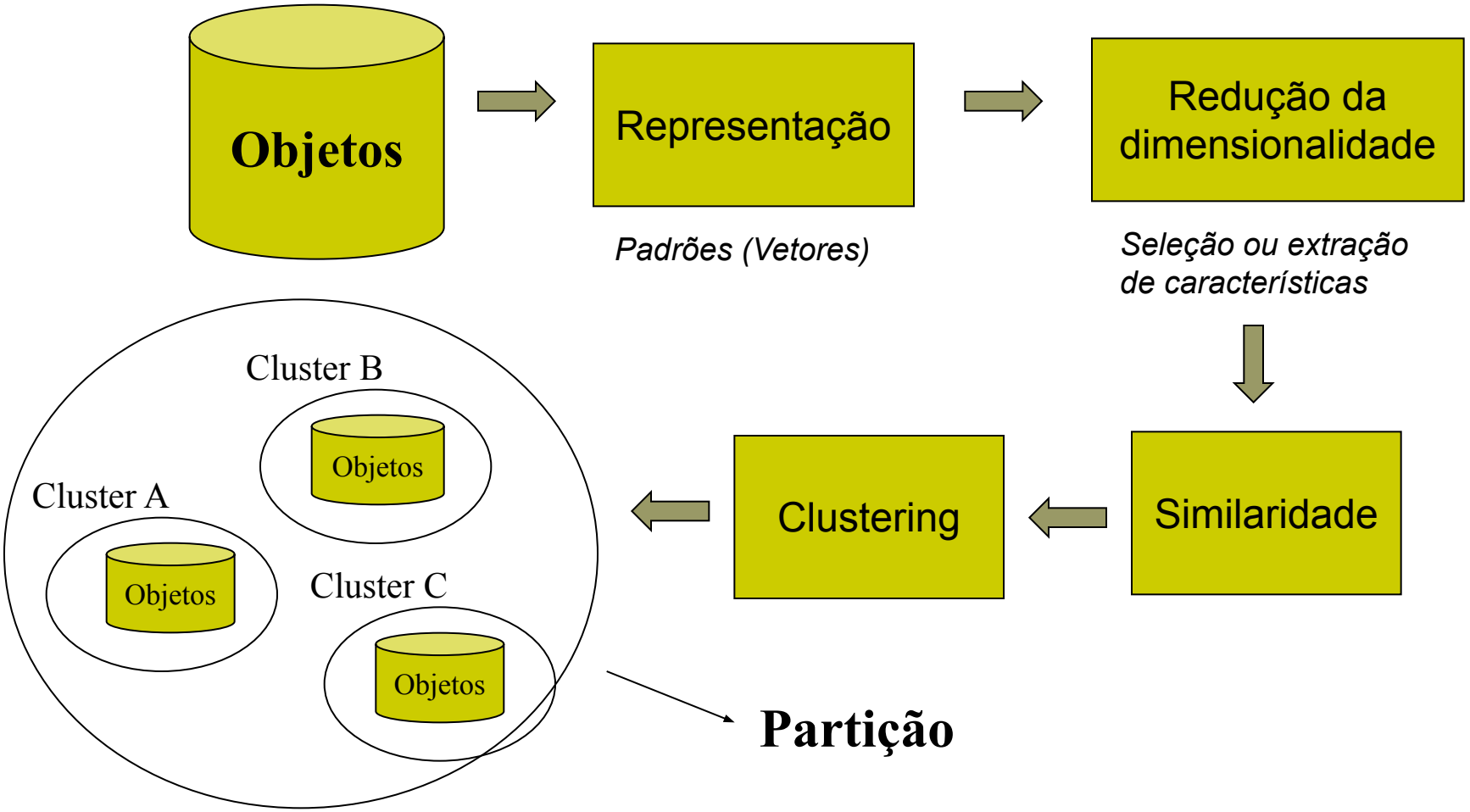
- Particionar objetos em *clusters* de forma que:
  - Objetos dentro de um cluster são similares
  - Objetos de clusters diferentes são diferentes

- Descobrir novas categorias de objetos de uma maneira *não-supervisionada*
  - Rótulos de classes não são fornecidos a priori

A. Jain et al.



Clustering - Etapas



Tipos de Clustering

- Algoritmos Flat (ou Particional)
  - Geram partição “plana”, i.e. não existe relação hierárquica entre os clusters

- Algoritmos Hierárquicos
  - Geram uma hierarquia de clusters, i.e. cada cluster é associado a um cluster-pai mais genérico
    - Vantagem: diferentes visões dos dados

Tipos de Clustering

- Hard
  - Cada objeto pertence exclusivamente a um único grupo na partição
- Fuzzy
  - Cada objeto está associado a um cluster com certo grau de pertinência
    - Partição Fuzzy pode ser convertida facilmente para uma partição hard

Tipos de Clustering

- Incremental
  - Partição é atualizada a cada novo objeto observado
    - Em geral, apenas um número pequeno de clusters é modificado
- Não-incremental
  - Partição é gerada de uma única vez usando todos os objetos disponíveis

Tipos de Clustering

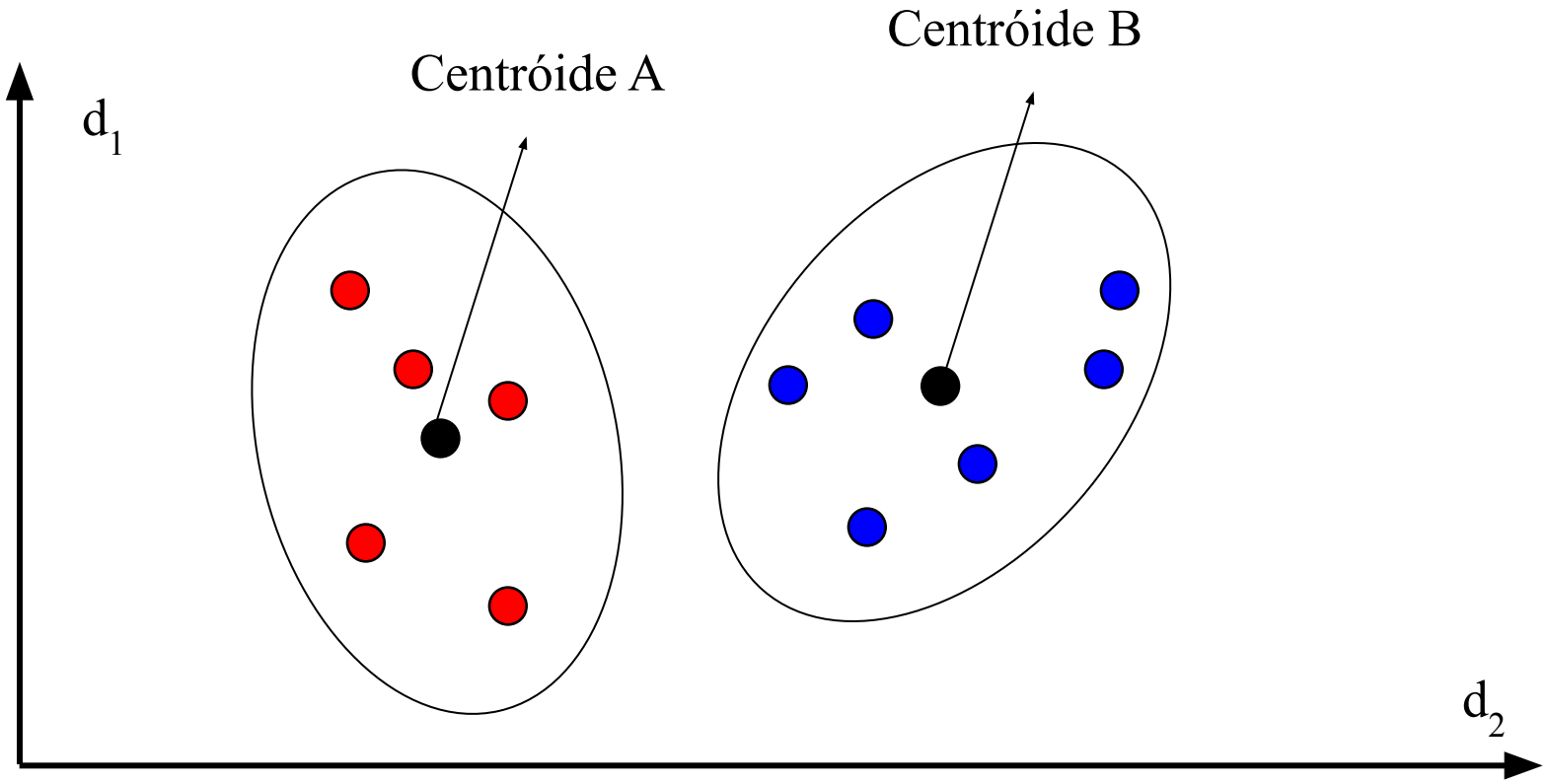
- Determinísticos
  - Algoritmo gera um única partição, independente se é executado várias vezes
- Estocásticos
  - Diferentes execuções do algoritmo podem gerar diferentes partições

Tipos de Clustering

- Baseado em Erro Quadrado
- Baseado em Grafos
- Mistura de Modelos
- Baseado em Redes Neurais
- Técnicas de Busca Combinatória
- Etc, etc,…

Algoritmo k-Means

- Encontra de forma interativa os *centróides* dos clusters



Algoritmo k-Means

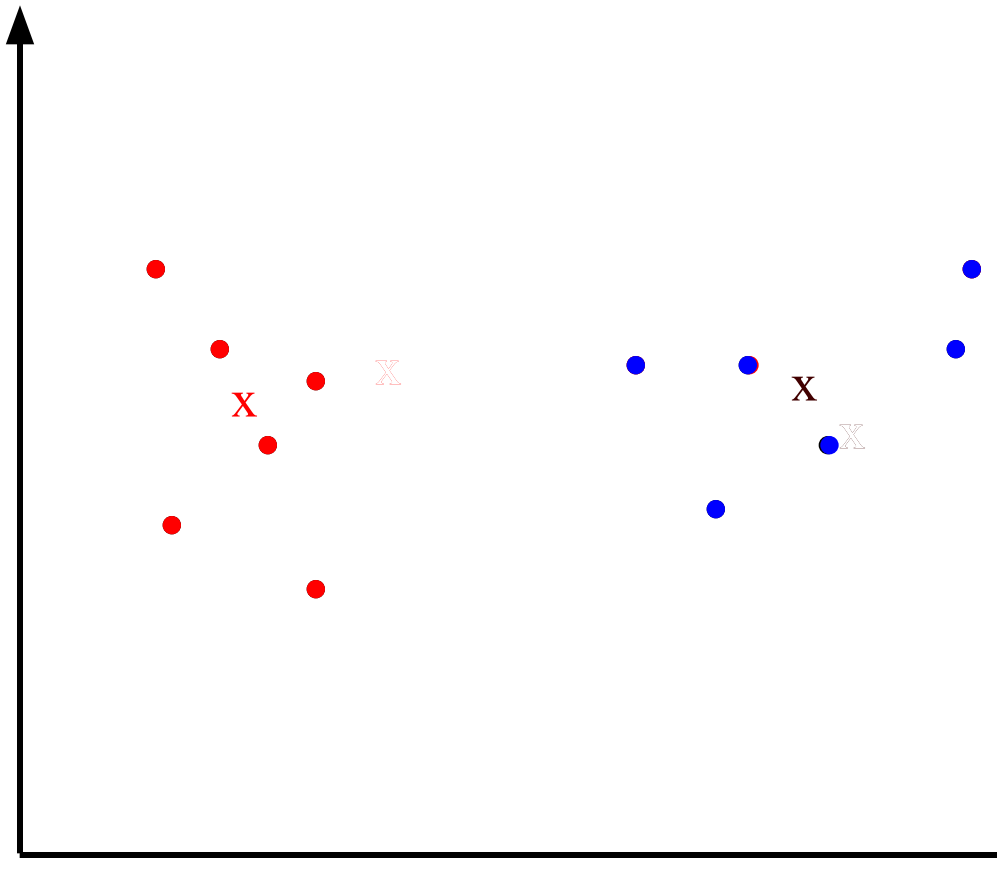
- Clusters definidos com base nos *centróides* (*centro de gravidade*, ou o ponto médio dos cluster:

- Alocação dos objetos nos clusters feita com base na similaridade com o centróide até critério de parada

Algoritmo k-Means

- **Passo 1:** Defina k centróides iniciais, escolhendo k objetos aleatórios;
- **Passo 2:** Aloque cada objeto para o cluster correspondente ao centróide mais similar;
- **Passo 3:** Recalcule os centróides dos clusters.
- **Passo 4:** Repita passo 2 e 3 até atingir um critério de parada
  - e.g. até um número máximo de iterações ou;
  - até não ocorrer alterações nos centróides (i.e. convergência para um mínimo local da função de erro quadrado)

k-Means (Exemplo com K=2)



Inicializar centróides

Alocar objetos

Computar centróides

Realocar objetos

Computar centróides

Realocar objetos

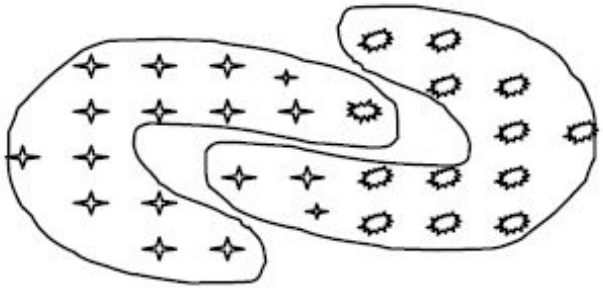
Convergiu!

Algoritmo k-Means

- O *k*-Means tende a gerar clusters esféricos

- Assim pode falhar para clusters naturais com formas mais complexas

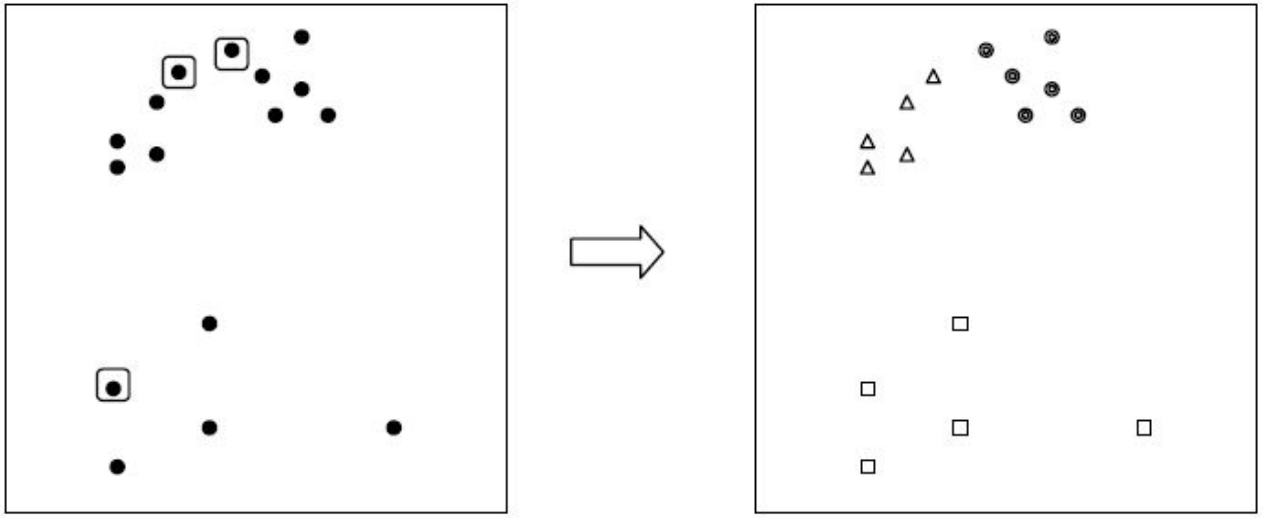
  - Exemplo -->

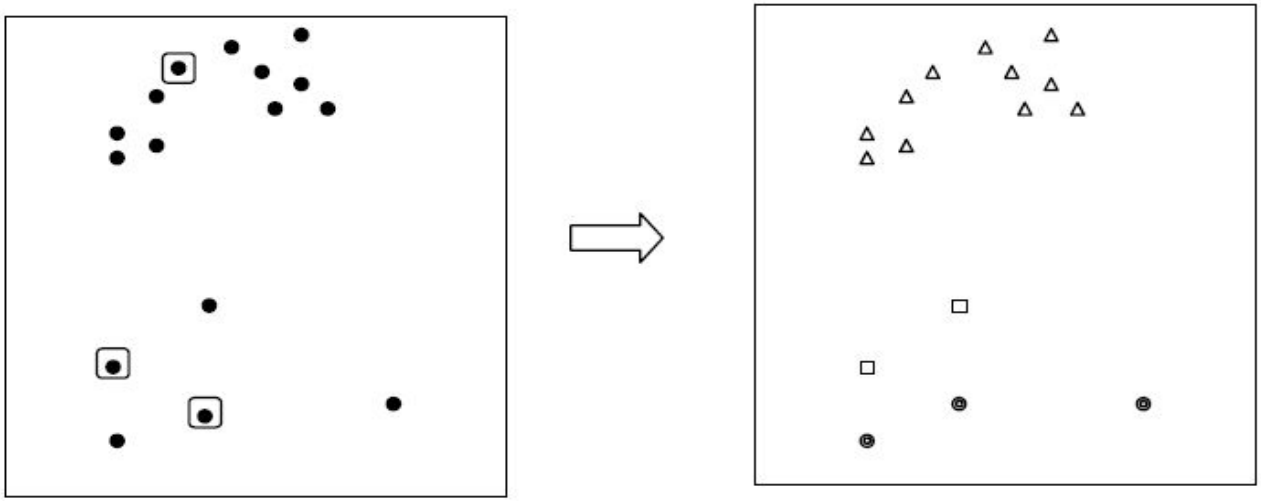


Algoritmo k-Means

- O *k*-Means é popular pela facilidade de implementação, e eficiência no tempo
  - *O(nK)*, onde *n* é o número de objetos e K é o número de clusters

- Comentários:
  - Não adequado para atributos categóricos
  - Sensível a outliers e ruído
  - Converge para mínimos locais
  - Desempenho do algoritmo é dependente da escolha dos centróides iniciais





Algoritmo k-Means – Escolha dos Centróides Iniciais

- Realizar várias execuções com inicializações diferentes
  - Escolher a melhor partição dentre as obtidas
- Gerar uma partição aleatória e usar centróides da partição como pontos iniciais do k-Means
- Usar algoritmos de clustering mais leves para gerar os centróides iniciais

Algoritmo ISODATA

- Similar ao k-Means porém o número de clusters é ajustado dinamicamente
- Procedimentos:
  - Juntar clusters com centróides similares
  - Dividir cluster com muita variação entre objetos
- Crítica:
  - Menos sensível a ruídos e a outliers
  - Porém limiares para junção e divisão de clusters devem ser definidos

Algoritmo k-Medoid

- Similar ao k-Means mas cada cluster é representado por um objeto que realmente existe (medoid)
- Medoid é o objeto do grupo cuja similaridade média com os outros objetos possui o valor máximo
- Comentários:
  - Tolerante a outliers e adequado para atributos categóricos
  - Porém, custo mais alto

Referências

- Jain, A. K., Murty, M. N., and Flynn, P. (1999). Data clustering: a review. ACM Computing Surveys, 3(31):264–323.
- Xu, R. and Wunsch II, D. (2005). Survey of Clustering Algorithms, IEEE Trans. on Neural Networks, 16(3):645-677.
- Jiang, D., T., Tang, and Zhang, A. (2004). Cluster Analysis for Gene Expression Data: A Survey, IEEE Trans. on Knowledge and Data Engineering, 16(11).