**Seminário de Agrupamento: Roteiro**

* **Conceitos iniciais**
  + **Análise de clusters:** A análise de clusters consiste em uma técnica de aprendizado não supervisionado com objetivo de agrupar um conjunto de objetos de acordo com a similaridade entre si
  + **Algoritmos Particionais:** Os algoritmos particionais dividem o conjunto de dados em um número predefinido de clusters. Eles são baseados na ideia de otimizar um critério de particionamento, como minimizar a soma das distâncias quadradas dentro dos clusters.
  + **Algoritmos Hieráquicos Divisivos:** Os algoritmos hierárquicos divisivos começam com todos os dados em um único cluster e, iterativamente, dividem os clusters em subclusters menores.
  + **Algoritmos Hierárquicos Aglomerativos (AHC):** Os algoritmos hierárquicos aglomerativos começam com cada ponto de dados como um cluster separado e, iterativamente, mesclam os clusters mais semelhantes até formar um único cluster.
* **Técnicas de agrupamento aglomerativo**

No contexto das técnicas de agrupamento hierárquico aglomerativo, a forma como a distância entre clusters é medida pode variar. Três métodos comuns para definir essa distância são single linkage, complete linkage e average linkage.

* + **Single linkage:** No método de single linkage, a distância entre dois clusters é definida como a menor distância entre qualquer par de pontos, onde cada ponto é de um cluster diferente
  + **Complete linkage:** No método de complete linkage, a distância entre dois clusters é definida como a maior distância entre qualquer par de pontos, onde cada ponto é de um cluster diferente.
  + **Average linkage:** No método de average linkage, a distância entre dois clusters é definida como a média das distâncias entre todos os pares de pontos, onde cada par consiste em um ponto de cada cluster.
* **Hierarquia aglomerativa**
  + **Aplicação para o tema do artigo:**
  + **Estrutura hierárquica:** Dentre os algoritmos hierárquicos, o agrupamento hierárquico aglomerativo (AHC) tornou-se um dos principais algoritmos de agrupamento devido à menor complexidade de tempo e melhor estabilidade computacional. Os algoritmos hierárquicos podem oferecer mais resultados de agrupamento mais satisfatórios que algoritmos particionais.
  + **Relação entre o nível e a quantidade dos clusters:** Como cada camada de uma hierarquia de cluster é determinada por um valor limite, que corresponde ao número de clusters, o problema de particionamento ótimo pode ser caracterizado como uma questão de determinação do número de cluster ideal.
  + **Motivação:** Para este artigo, com base no algoritmo AHC, a motivação consiste na integração e otimização das propostas apresentadas por trabalhos relacionados, gerando de um novo índice de validade de agrupamento, o qual pode ser aplicado em um algoritmo para determinar o número ideal de clusters.
* **Estudos prévios**
  + **Conjunto de partições estendidas** – propuseram a busca pelo método do conjunto de partições estendidas e um novo índice de validade de cluster para encontrar a melhor partição no cluster hierárquico, integrando o operador de média ponderada ordenada (OWA) com agrupamento hierárquico para determinar a distância entre clusters. Contudo, não foram feitas pesquisas suficientes sobre o ajuste dos pesos ideais do operador que determina a distância entre clusters para obter os melhores resultados de agrupamento
  + **Proposta de novo agrupamento hierárquico** – propuseram um método hierárquico, que primeiro obteve o recurso de agrupamento por meio da varredura do conjunto de dados e gerou aglomerativamente partições hierárquicas do conjunto de dados, depois construiu incrementalmente uma curva de qualidade de agrupamento para diversas partições e, finalmente, usou as partições correspondentes ao extremo da curva para estimar o número ideal de clusters. Eles afirmaram que o método proposto poderia ser usado para medir dados de estruturas não convexas. Contudo não descreveram a distribuição detalhada dos conjuntos de dados experimentais
  + **Proposta de um novo índice de validade de cluster** – propuseram um novo índice de validade de cluster para determinar o número ideal de clusters em cluster hierárquico, definindo a compactação e separabilidade dos resultados do cluster. Contudo o artigo de estudo não especificou o método de medida de similaridade para construir uma matriz de similaridade
* **Objetivo:** Descrito no próprio slide
* **Índice de proporção de compacidade separada (CSP)**
  + **Análise de validade do agrupamento:** consiste na avaliação dos resultados de um dado agrupamento, determinando se os clusters formados são significativos, coesos internamente e bem separados uns dos outros, por meio da análise da estrutura interna de um conjunto dos dados em um cluster
  + **Conceito do índice CSP:** Buscaavaliar efetivamente os resultados de agrupamento de múltiplos tipos de conjuntos de dados, como estruturas lineares, múltiplas, anulares e convexas.
  + **Índices de validade existentes:** Descrito no próprio slide
* **Definições do índice CSP**
  + **Definição 1 –** Descrito no próprio slide
  + **Definição 2 –** Descrito no próprio slide
  + **Definição 3 –** Descrito no próprio slide
  + **Definição 4 –** Descrito no próprio slide
  + **Definição 5 –** Descrito no próprio slide
* **Análise do agrupamento por meio do índice CSP dos demais clusters no conjunto de dados**

Define o índice CSP geral para todos os demais agrupamentos com base na média aritmética dos índices individuais para cada cluster

* **Utilização do algoritmo *optimal number of cluster determination* (ONCD):** Explicar o algoritmo com base no próximo slide, por meio do **Exemplo do *step 3***
* **Resultados: Verificação do desempenho do índice CSP e ONCD –** Descrito no próprio slide