**Seminário de Agrupamento: Roteiro**

* **Metadados**
  + Informações extras sobre o artigo a ser apresentado:
  + Artigo publicado na IEEE Transactions on Neural Networks and Learning Systems
  + Artigo citado em 132 artigos diferentes
  + Ano de publicação: 2016
* **Sumário –** Descritos no próprio slide
* **Conceitos iniciais**
  + **Análise de clusters:** O artigo utiliza extensivamente o conceito da análise de clusters para o desenvolvimento de sua proposta. Trata-se de uma técnica de mineração de dados e aprendizado de máquina utilizada para agrupar um conjunto de objetos de forma que os objetos dentro de um grupo (ou cluster) sejam mais semelhantes entre si do que aos objetos em outros grupos.

Existem diferentes tipos de algoritmos de agrupamento, cada um com suas próprias características e métodos de funcionamento, os quais podem ser caracterizados como:

* + **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 (AHD):** Cria uma hierarquia de clusters que pode ser representada como uma árvore (dendrograma), aqui todos os dados são iniciados como um único cluster e, iterativamente, dividem os clusters em subclusters menores.
  + **Algoritmos Hierárquicos Aglomerativos (AHC):** possuem o mesmo conceito de hierarquia proveniente do AHD, contudo, iniciam cada ponto de dados como um cluster separado e, iterativamente, mesclam os clusters mais semelhantes até formar um único cluster. Esta é a técnica implementada na metodologia do artigo
* **Técnicas de agrupamento aglomerativo**

No contexto das técnicas de agrupamento hierárquico aglomerativo, é necessário realizar a medição das distâncias entre os clusters a fim de definir quais serão aglomerados, contudo, a forma como tal distância é medida pode variar. Três métodos comuns para definir essa distância são:

* + **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. É eficiente para detectar formas não globulares de clusters e pode lidar bem com ruído e outliers, contudo, pode levar ao efeito de "cadeia", onde clusters longos e esparsos são formados
  + **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. Forma clusters compactos e globulares evitando o efeito de cadeia, contudo, pode ser influenciado por outliers
  + **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. Tende a encontrar um bom equilíbrio entre clusters compactos e bem separados, contudo, costuma ser computacionalmente mais intenso
* **Hierarquia aglomerativa**
  + **Estrutura hierárquica:** Foi a abordagem de agrupamento selecionada para desenvolver a metodologia proposta neste artigo. Sua utilização se mostra bastante vantajosa devido a menor complexidade de tempo e melhor estabilidade computacional, além disso os algoritmos hierárquicos podem oferecer 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.
* **Estudos prévios**

Alguns estudos foram citados como base para a produção da metodologia proposta neste artigo

* + **Conjunto de partições estendidas** – Propuseram um método que integra o operador de média ponderada ordenada (OWA) com agrupamento hierárquico para determinar a distância entre clusters, mas não investigaram suficientemente o ajuste dos pesos ideais do operador.
  + **Proposta de novo agrupamento hierárquico** – Desenvolveram um método hierárquico para a geração de partições, construindo uma curva de qualidade de agrupamento e usa o extremo da curva para estimar o número ideal de clusters, aplicável a dados não convexos. No entanto, não detalharam a distribuição dos conjuntos de dados experimentais.
  + **Proposta de um novo índice de validade de cluster** – Introduziram um novo índice para determinar o número ideal de clusters, definindo a compactação e separabilidade dos resultados. Contudo, não especificaram o método de medida de similaridade para construir a matriz de similaridade.
* **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.
* **Objetivo:** Descrito no próprio slide
* **Índice de proporção de compacidade separada (CSP)**
  + **Análise de validade do agrupamento:** consiste no processo de avaliar a qualidade e a eficácia de um agrupamento (cluster) resultante de um algoritmo de clustering, tendo como objetivo principal, determinar o quão bem os clusters encontrados representam a estrutura subjacente dos dados.
  + **Conceito do índice CSP:** Buscaavaliar efetivamente os resultados de agrupamento, tendo como diferencial, a capacidade de avaliar múltiplos tipos de conjuntos de dados, como os dados de estruturas lineares, múltiplas, anulares e convexas.
  + **Índices de validade existentes:** Descrito no próprio slide