

Hierarchical Cluster Analysis (HCA): como funciona o algoritmo KNN?

¹Edilton de Souza Barcellos^{1*} (PQ), Márcia Miguel Castro Ferreira² (PQ).

¹Departamento de Química/UFV, ² Instituto de Química/Unicamp – barcello@ufv.br

Palavras Chave: HCA, KNN, Análise de Agrupamentos, single linkage, similaridade, medida de distâncias.

Introdução

HCA é uma técnica do subgrupo de Métodos Não Supervisionados de Reconhecimento de Padrão¹.

Dentre os métodos de conexão, destaca-se aqui o “single linkage” (ou KNN), bastante usado para agrupar objetos em função, tanto de similaridades quanto de distâncias¹, com aplicação em vários campos (psicologia, biologia, química, ciência de alimentos, medicina, ciência ambiental etc²).

Há casos nos quais entender o que está por trás do software leva a resultados melhores do que apenas “rodar o software”,

Nesse sentido, apresenta-se um passo a passo envolvendo cálculos de distâncias necessários à obtenção dos resultados na forma de um dendrograma.

Resultados e Discussão

A matriz X(7x2) abaixo, representa concentrações de 2 analitos (2 variáveis) nas 7 amostras coletadas em pontos diferentes de um dado local (lago, rio, solo, etc); em alimentos (seriam 7 marcas distintas de um produto); dentre outros exemplos.

$$X = \begin{bmatrix} 1 & 10; 13 & 12; 4 & 8; 3 & 6.5; 10 & 5; 10 & 3.5; 16 & 4.5 \end{bmatrix};$$

A Fig.1 representa um gráfico da variável 1 versus variável 2 da matriz de dados originais, destacando a junção dos pontos (5 e 6), como se vê abaixo.

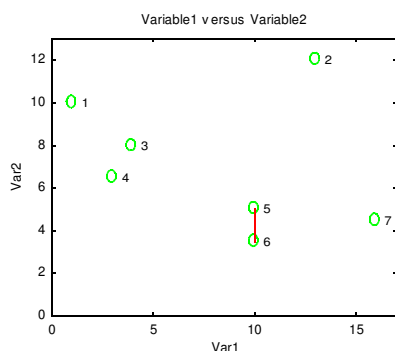


Figura 1. Junção dos pontos (5,6). Eles possuem menor distância entre si. Esses pontos são similares entre si. A seguir os mais próximos são os pontos (3,4), os quais são similares entre si, mas diferem de (5,6). A amostra mais dissimilar é a do ponto 2.

O resultado do 1º passo, gera a matriz de distâncias, D1, adiante.

D1 =

12.1655	3.6056	4.0311	10.2956	11.1018	15.9765
0	9.8489	11.4127	7.6158	9.0139	8.0777
0	0	1.8028	6.7082	7.5000	12.5000
0	0	0	7.1589	7.6158	13.1529
0	0	0	0	1.5000	6.0208
0	0	0	0	0	6.0828

Calculam-se distâncias para a construção das matrizes D2, D3, D4, e assim em diante até que todos os pontos estejam ligados.

Com base nas expressões matemáticas desses cálculos² (omitidas aqui) chega-se ao Dendrograma conforme se vê na Figura abaixo.

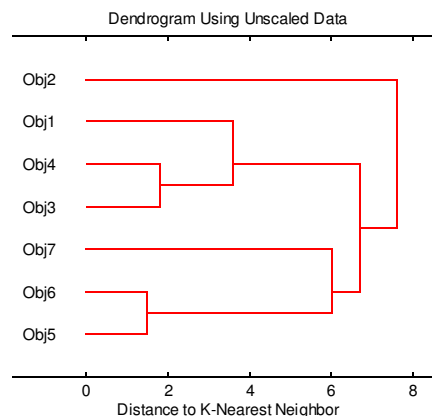


Figura 2. Dendrograma do agrupamento (“single linkage”) de dados alocados em uma matriz de 7 objetos e 2 variáveis.

Comparando-se os pontos (5,6) em cada figura vê-se claramente que eles são os mais próximos. Na Fig.2, vê-se na escala horizontal que esses dois pontos possuem o menor valor (~1,5). Depois são os pontos (3,4) cujo valor é (~2). Nota-se que (5, 6) e (3,4) pertencem a grupos diferentes.

Conclusões

Vê-se, nas Fig. 1 e 2 uma perfeita concordância para todos os pontos. Tanto as matrizes quanto os gráficos obtidos a partir delas são simples. A construção do dendrograma se dá de forma natural.

Agradecimentos

À FAPEMIG pelo financiamento das despesas de passagens e diárias.

¹Sharaf, M. A.; Illman, D. L. e Kowalski, B. R. John Wiley & Sons, 1986.

²Everitt, B. , Heinemann Educational Books Ltd, Social Science Research Council, London, 1974.