Gom nhóm (Clustering)

Mục tiêu: Chúng ta có 1 tập dữ liệu không biết tính chất➔ gom nhóm dữ liệu so cho những đối tượng giống nhau nhất sẽ thành 1 nhóm.

Thuật toán gom nhóm k-means:

	F1	F2
O1	1	2
O2	3	2
O3	1	1
O4	2	3
O5	4	1

-Thực hiện min-max normalization [0,1]

	F1	F2
O1	0	0.5
O2	0.67	0.5
O3	0	0
O4	0.33	1
O5	1	0

Giả sử số nhóm cần gom nhóm là k=2

B0. Chọn ngẫu nhiên 2 trọng tâm nhóm trong bộ dữ liệu:

Trọng tâm nhóm 1: O2; Trọng tâm nhóm 2: O5.

B1: Xác định nhóm của các đối tượng trong bộ dữ liệu

-O1,O2,O3,O4,O5 thuộc nhóm nào 1 hay 2??

-Xác định nhóm cho O1: so sánh khoảng cách từ O1 tới trọng tâm nhóm 1 và khoảng cách từ O1 tới trọng tâm nhóm 2→ quyết định O1 thuộc nhóm 1 hay 2.

-Xác định nhóm cho O2

-Xác định nhóm cho O3

-Xác định nhóm cho O4

-Xác định nhóm cho O5

Ví dụ: Nhom1 ={O2,O3,O4} Nhom2={O1,O5}

-Tính lại trọng tâm nhóm: Trọng tâm nhóm 1=(O2+O3+ O4)/3; Trọng tâm nhóm 1=(O1+O5)/2.

-So sánh trọng tâm nhóm trước B1 và sau B1 có thay đổi hay không? Nếu không thì dừng thuật toán, nếu có lặp lại bước 1.

Thuật toán tính khoảng cách giữa 2 điểm:

A(1,2,3) B(2,3,1)

Euclid= sqrt ((1-2)^2+(2-3)^2 + (3-1)^2)

→Càng nhỏ thì càng gần nhau

Cos = (1*2+2*3+3*1): [sqrt(1^2+2^2+3^2)*sqrt(2^2+3^2+1^2)]

→Càng lớn thì càng gần.