**ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN KHOA HỌC MÁY TÍNH**

**Bài tập thực hành:**

**Phân lớp dữ liệu**

**Thông tin nhóm:**

**Đàm Thiệu Quang 1241393**

**Nguyễn Thị Yến 1241444**

# Bài 1: Trình bày và cho ví dụ các khái niệm Single Link, Compelete Link, Mean Link, Average Link, Centroid Link, Ward Link trong gom nhóm phân cấp

1. Single Link

* Với Single Link, khoảng cách giữa 2 cụm được tính bằng khoảng cách ngắn nhất giữa những đối tượng gần nhau nhất trong các cụm
* Ví dụ:

Cụm 1 gồm A, B

Cụm 2 gồm C,D,E

Các khoảng cách giữa các đối tượng trong 2 cụm như sau:

KC(A,C) =1

KC(A,D)=2

KC(A,E) =3

KC(B,C) =4

KC(B,D)=5

KC(B,E) =6

Khoảng cách giữa cụm 1 và 2 được tính như sau:

KhoangCach(C1,C2)= min(KC(A,C), KC(A,D), KC(A,E), KC(B,C)

, KC(B,D), KC(B,E))=KC(A,C)=1

1. Complete Link

* Với Complete Link, khoảng cách giữa 2 cụm được tính bằng khoảng cách lớn nhất giữa các đối tượng trong các cụm
* Ví dụ:

Cụm 1 gồm A, B

Cụm 2 gồm C,D,E

Các khoảng cách giữa các đối tượng trong 2 cụm như sau:

KC(A,C) =1

KC(A,D)=2

KC(A,E) =3

KC(B,C) =4

KC(B,D)=5

KC(B,E) =6

Khoảng cách giữa cụm 1 và 2 được tính như sau:

KhoangCach(C1,C2)= max(KC(A,C), KC(A,D), KC(A,E), KC(B,C)

, KC(B,D), KC(B,E))=KC(B,E)=6

1. Mean Link
2. Average Link

* Với Average Link, khoảng cách giữa 2 cụm được tính bằng trung bình cộng khoảng cách của tất cả đối tượng của cụm này với tất cả đối tượng của cụm
* Ví dụ:

Cụm 1 gồm A, B

Cụm 2 gồm C,D,E

Khoảng cách giữa cụm 1 và 2 được tính như sau:

KhoangCach(C1,C2)=(KC(A,C) + KC(A,D) + KC(A,E)

+ KC(B,C) + KC(B,D) + KC(B,E))/6

1. Centroid Link

* Với Centroid Link, khoảng cách giữa 2 cụm được tính bằng bình phương khoảng cách Euclide của hai tâm của hai cụm này
* Ví dụ:

Cụm 1 gồm A, B,T có tâm cụm là T

Cụm 2 gồm C,D,E có tâm cụm là E

Khoảng cách giữa cụm 1 và 2 được tính như sau:

KhoangCach(C1,C2)=(KC(T,E))2

1. Ward Link

* Ward Link sử dụng hàm tổng bình phương lỗi để xác định khoảng cách giữa 2 cụm
* Ví dụ:

# Bài 2: Trình bày cải tiến của K-mean++:

K-Means++ là một cải tiến của K-means trong đó có sự cải tiến việc khởi tạo tâm ban đầu theo tiêu chí chọn các đặc trưng cách xa nhau và tránh được tình trạng bị ảnh hưởng bởi outlier.

Cải tiến như sau:

Xét tập đặc trưng T: (x1,x2,x3,..,xn)

* Gọi Dist(x) là khoảng cách từ đặc trưng x (thuộc tập T) đến tâm gần nhất đã được xác định trước đó.
* Đầu tiên, chọn ngẫu nhiên tâm t1 trong tập T
* Ta chọn tâm tiếp theo ti = x’∈ T với xác suất chọn là:
* Chọn tiếp các tâm tiếp theo cho đến khi đủ k tâm