#### Ghép nhóm(clustering)

 Ghép nhóm là phương pháp phân chia các điểm trong một đám mây điểm thành các nhóm nhỏ, qua đó giảm đáng kể thời gian để xử lý toàn bộ lượng dữ liệu ban đầu. Các phương pháp ghép nhóm đơn giản sử dụng cách tìm các đường biên hay so sánh về khoảng cách đến các điểm lân cận để nhóm các điểm gần nhau lại với nhau.

Giả sử trong tập dữ liệu $P=\left\{p\_{1,}p\_{2,},,,,,,p\_{n,}\right\}$, hai nhóm $O\_{i}=\left\{p\_{i}\in P\right\}$ và $O\_{j}=\left\{p\_{j}\in P\right\}$ là hai nhóm riêng biệt nếu :

$$min=\left‖p\_{i}-p\_{j}\right‖<Є$$

Với ngưỡng $Є $là khoảng cách giới hạn.Nói cách khác, nếu khoảng cách nhỏ nhất giữa hai tập các điểm $O\_{i}=\left\{p\_{i}\in P\right\}$ và $O\_{j}=\left\{p\_{j}\in P\right\}$ lớn hơn một ngưỡng giới hạn cho trước thì khi đó $O\_{i}$và$ O\_{j}$là hai nhóm khác nhau.

 

*Hình 1. Các cụm điểm thành nhóm riêng biệt*

Ví dụ về việc ghép nhóm trong một đám mây điểm là trong bài toán xác định các vật thể đặt trên bàn. Khi đó đám mây điểm đầu vào sẽ bao gồm một mặt phẳng lớn – là mặt bàn – và các nhóm điểm biểu diễn cho các vật nằm trên mặt bàn đó. Sau khi tách được mặt bàn thì đám mây điểm chỉ còn bao gồm các nhóm điểm riêng biệt rõ ràng, mỗi nhóm là dữ liệu quét của một vật thể.

Trong bài toán như trên, phương pháp ghép nhóm đơn giản nhất là phương pháp sử dụng khoảng cách vật lý và giải thuật tìm kiếm điểm lân cận. Các bước để thực hiện giải thuật này như sau:

* Với dữ liệu đầu vào là đám mây điểm ***P***, tạo cây kd-tree biểu diễn dữ liệu để thuận lợi cho việc tìm kiếm lân cận.
* Tạo ra một danh sách nhóm ***C***, và các điểm cần được khảo sát ***Q***. Ban đầu, các điểm cần được khảo sát là toàn bộ dữ liệu đầu vào.
* Với mỗi điểm **PiЄP** , thực hiện các bước sau:

+ Thêm **Pi** vào danh sách các điểm cần khảo sát ***Q***.

+ Với mỗi **Pi Є Q** , tìm kiếm các điểm lân cận của Pi trong bán kính. Sau đó kiểm tra các điểm lân cận này đã được xử lý hay chưa, nếu chưa được xử lý thì thêm điểm đó vào ***Q***.

+ Khi tất cả các điểm trong ***Q*** đã được xử lý, điều đó có nghĩa là không còn điểm nào trong ***P*** có khoảng cách đến ***Q*** nhỏ hơn *ε*. ***Q*** được coi là một nhóm.

* Quá trình trên được lặp lại cho đến khi tất cả các điểm đều thuộc một nhóm nào đó.