

Chapter Clustering

K-means based method



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

Tổng kết

Nguyễn Giáp Nguyên Sinh
Nguyễn Quốc Long
Khoa Khoa học và Kỹ thuật Máy tính
Đại học Bách Khoa TP.HCM

Nội dung

① Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

② Vấn đề & K-means

K-means

Mô phỏng K-means

③ Ứng dụng

④ Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

⑤ Tổng kết



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

Điểm mạnh

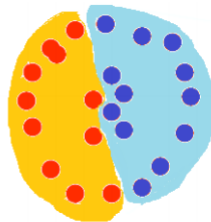
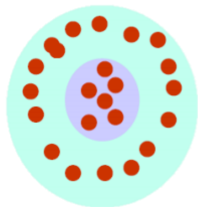
Điểm yếu

Tổng kết



Phân cụm là gì ?

- Việc tổ chức dữ liệu chưa được gán nhãn (label) vào các nhóm tương tự nhau được gọi là phân cụm
- Một cụm là một tập hợp các phần tử dữ liệu có sự giống nhau về mặt dữ liệu và sẽ khác với các phần tử dữ liệu ở các cụm khác.



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

Tổng kết



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

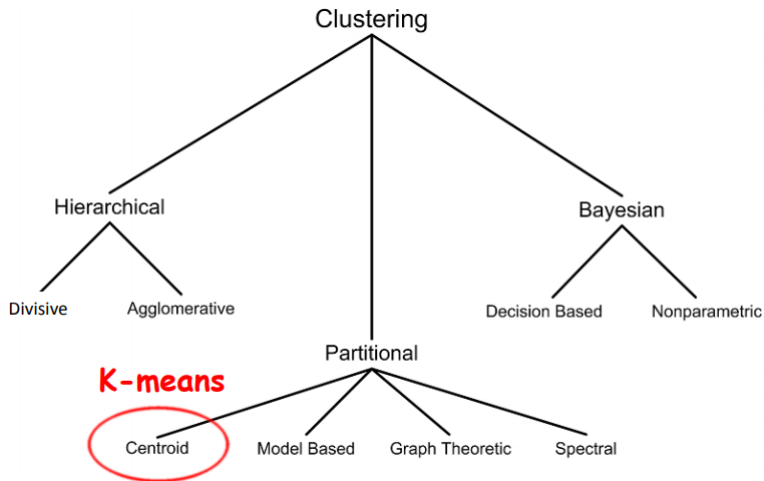
Ứng dụng

Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

Tổng kết





Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

Tổng kết

Khi nào nghĩ đến K-means

- Không biết nhãn(label) của điểm dữ liệu
- Mục đích : Phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong một cụm có tính chất giống nhau.

K-means cluster

- K-means được đề xuất bởi MacQueen năm 1967
- Giải thuật k-means chia tập dữ liệu thành k cụm (cluster)
 - Mỗi cluster có một điểm trung tâm , gọi là *centroid*
 - K được chỉ định bởi nhân viên phân tích dữ liệu (Data analytics)

Giải thật k-means

Nhập giá trị k , giải thuật k-means sẽ thực thi các bước như sau:

- 1 Chọn ngẫu nhiên k điểm dữ liệu làm *centroids*, điểm trung tâm của cụm dữ liệu.
- 2 Phân mỗi điểm dữ liệu vào cluster có điểm trung tâm (center) gần nó nhất.
- 3 Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
- 4 Cập nhật center cho từng cluster bằng cách lấy lấy trung bình cộng của tất cả các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
- 5 Quay lại bước 2.



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

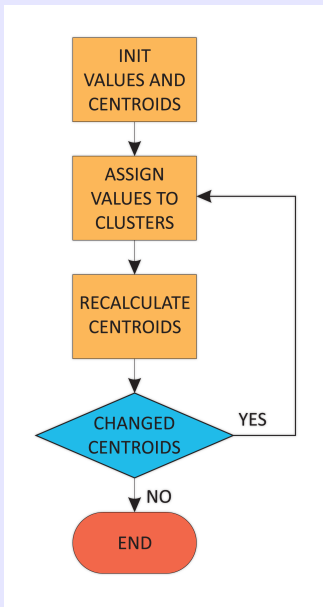
Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

Tổng kết

Cơ chế của giải thuật K-means có thể tổng quát bằng sơ đồ dưới đây:



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

Tổng kết

Cách tính khoảng cách và điểm trung tâm

Cách tính khoảng cách trong giải thuật K-means

Để tìm được điểm dữ liệu gần với điểm trung tâm nào nhất thì ta dựa vào giá trị nhỏ nhất của hàm tính khoảng cách Euclidean.

- Với dữ liệu 1 chiều, với 2 điểm dữ liệu p và q
$$\sqrt{p - q} = |p - q|$$
- Với dữ liệu 2 chiều, với 2 điểm dữ liệu $p(p_1, p_2)$ và $q(q_1, q_2)$
$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$
- Với dữ liệu 3 chiều, với 2 dữ liệu $p(p_1, p_2, p_3)$ và $q(q_1, q_2, q_3)$
$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$$
- Với dữ liệu n chiều, với 2 dữ liệu $p(p_1, p_2, \dots, p_n)$ và $q(q_1, q_2, \dots, q_n)$
$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}$$

Cách tính điểm trung tâm

Điểm trung tâm *centroid* của cluster là trung bình cộng của các điểm trong cluster đó.



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

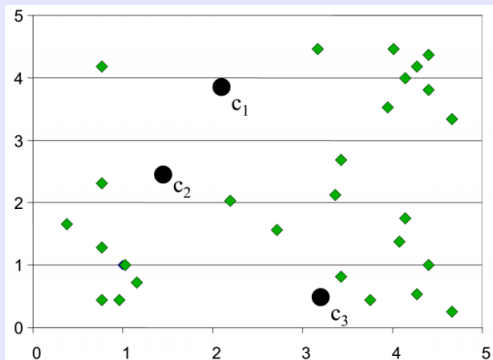
Điểm mạnh

Điểm yếu

Tổng kết

Mô phỏng K-means

Chọn ngẫu nhiên các centroid



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

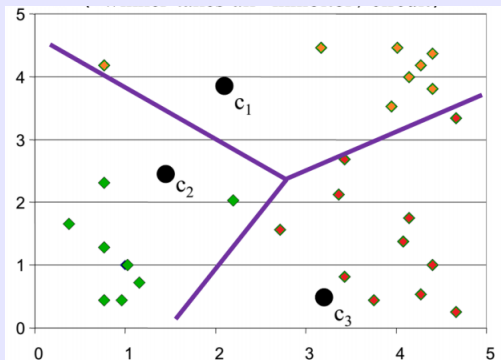
Điểm mạnh

Điểm yếu

Tổng kết

Mô phỏng K-means

Xác định cluster cho từng điểm dữ liệu



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

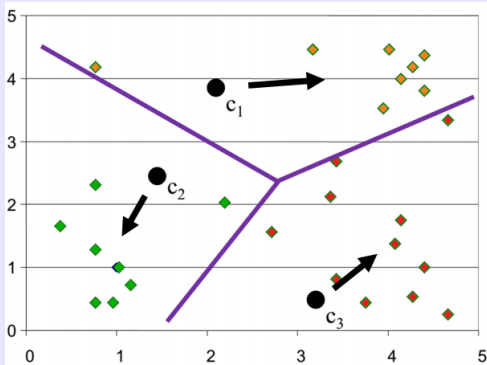
Điểm mạnh

Điểm yếu

Tổng kết

Mô phỏng K-means

Xác định lại centroid cho các cluster



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

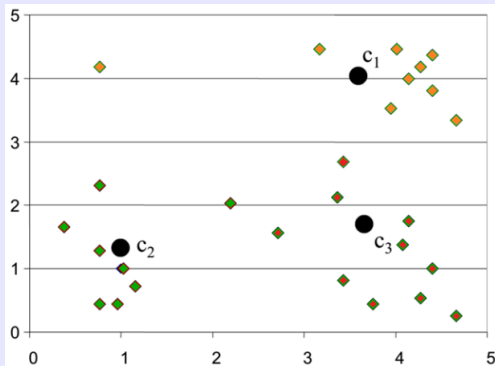
Điểm mạnh

Điểm yếu

Tổng kết

Mô phỏng K-means

Kết quả của vòng lặp thứ nhất



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

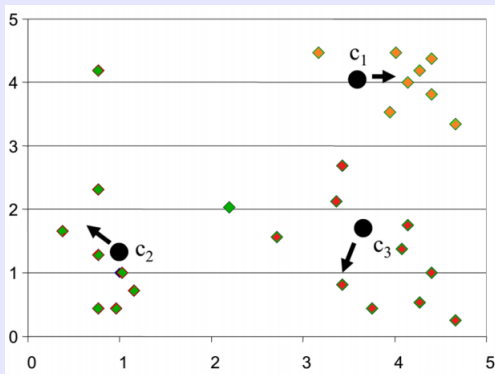
Điểm mạnh

Điểm yếu

Tổng kết

Mô phỏng K-means

Vòng lặp thứ 2, xác định lại các điểm centroid



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

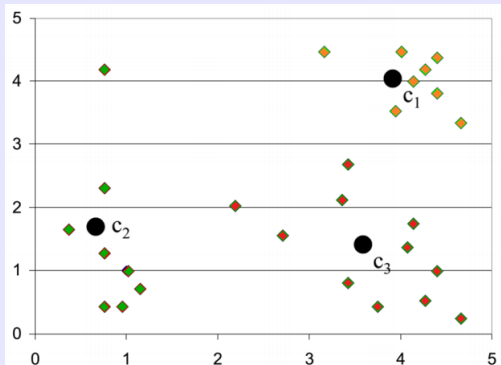
Điểm mạnh

Điểm yếu

Tổng kết

Mô phỏng K-means

Kết quả của vòng lặp thứ hai



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

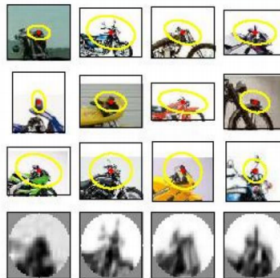
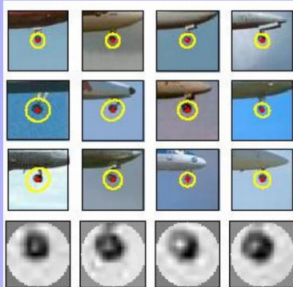
Điểm mạnh

Điểm yếu

Tổng kết

Ứng dụng phân loại hình ảnh

Examples of visual words



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

Tổng kết



Nội dung

[Giới thiệu về phân cụm](#)

[Phân cụm](#)

[Kỹ thuật phân cụm](#)

[Vấn đề & K-means](#)

[K-means](#)

[Mô phỏng K-means](#)

[Ứng dụng](#)

[Điểm mạnh và điểm yếu](#)

[Điểm mạnh](#)

[Điểm yếu](#)

[Tổng kết](#)

Điểm mạnh

- Đơn giản: Dễ dàng hiểu và thực thi
- Rất hiệu quả: Độ phức tạp chỉ là $O(tkn)$, trong đó:
 - k là số cụm
 - t là số lần lặp

Vì k và t đều nhỏ nên giải thuật K-mean thuộc giải thuật tuyến tính

K-means là một giải thuật phân cụm phổ biến

Điểm yếu

- Người phân tích dữ liệu phải chỉ định số cluster K
- Rất hiệu quả: Độ phức tạp chỉ là $O(tkn)$, trong đó:
 - k là số cụm
 - t là số lần lặp

Vì k và t đều nhỏ nên giải thuật K-mean thuộc giải thuật tuyến tính

- Giải thuật rất dễ sai với các điểm nhiễu *outliers*
 - Outlier là những điểm dữ liệu quá xa so với các điểm dữ liệu còn lại.
 - Outlier có thể xảy ra khi xảy ra lỗi khi ghi dữ liệu hoặc một số điểm dữ liệu đặc biệt với các giá trị rất khác nhau.



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

Tổng kết

Outliers



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

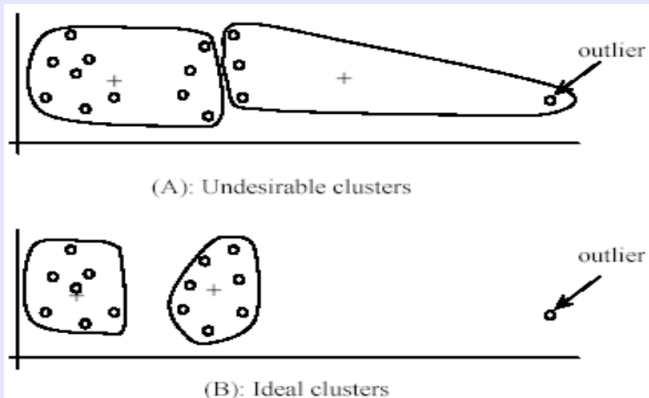
Ứng dụng

Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

Tổng kết



Giải quyết Outlier

- Xóa một số điểm dữ liệu ở rất xa centroid so với đa số các điểm dữ liệu khác
 - Để an toàn, chúng ta có thể theo dõi những điểm outlier này một vài lần lặp rồi mới quyết định loại bỏ chúng hay không.
- Thực hiện lấy mẫu ngẫu nhiên: bằng cách chọn một tập hợp con của các điểm dữ liệu, cơ hội lựa chọn một outlier là nhỏ hơn nhiều.
 - Gán những phần còn lại của các điểm dữ liệu cho các cluster bằng khoảng cách



Nội dung

[Giới thiệu về phân cụm](#)

Phân cụm

Kỹ thuật phân cụm

[Vấn đề & K-means](#)

K-means

Mô phỏng K-means

[Ứng dụng](#)

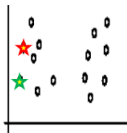
[Điểm mạnh và điểm yếu](#)

Điểm mạnh

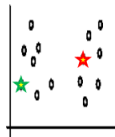
Điểm yếu

[Tổng kết](#)

Sensitivity to initial seeds



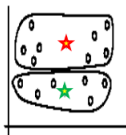
Random selection of seeds (centroids)



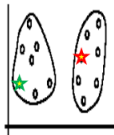
Random selection of seeds (centroids)



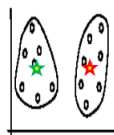
Iteration 1



Iteration 2



Iteration 1



Iteration 2



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

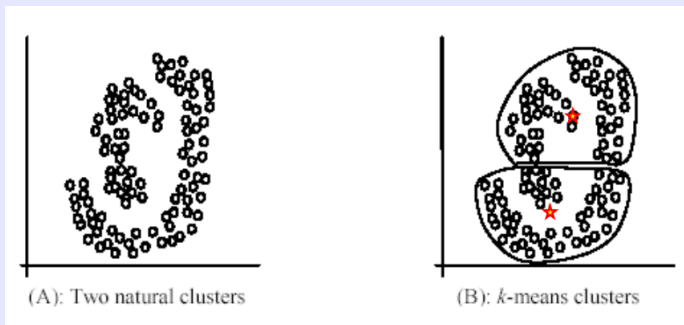
Điểm mạnh

Điểm yếu

Tổng kết

Cấu trúc dữ liệu đặc biệt

Giải thuật K-means không phù hợp cho xác định các cụm mà hình không phải là hình elip hay hình cầu



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

Tổng kết



Nội dung

Giới thiệu về phân cụm

Phân cụm

Kỹ thuật phân cụm

Vấn đề & K-means

K-means

Mô phỏng K-means

Ứng dụng

Điểm mạnh và điểm yếu

Điểm mạnh

Điểm yếu

Tổng kết

Tổng kết giải thuật K-means

- Mặc dù có điểm yếu nhưng k-means vẫn là giải thuật phổ biến do tính đơn giản và tính hiệu quả của nó.
- Chưa có bằng chứng nào chứng tỏ giải thuật phân cụm nào là tốt nhất.