



ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

CHƯƠNG 8

Hệ thống phân cụm cho doanh nghiệp

Biên soạn: ThS. Nguyễn Thị Anh Thư



Nội dung

1. Giới thiệu
2. Quy trình triển khai
3. Lựa chọn thuật toán
4. Đánh giá hiệu quả
5. Triển khai hệ thống
6. Giám sát và bảo trì
7. Tổng kết



1. Giới thiệu



Phân tích thị trường

Phát hiện gian lận

Khai phá kiến thức

...

Hệ thống phân cụm được sử dụng để phân chia một tập dữ liệu thành các nhóm riêng biệt (gọi là cụm) mà không cần có nhãn cho từng điểm dữ liệu.

Được sử dụng trong nhiều lĩnh vực khác nhau.



Phân cụm (Clustering)

Học không giám sát (Unsupervised Learning)

Phân cụm (Clustering)

Khám phá cấu trúc ẩn trong dữ liệu bằng cách **nhóm các dữ liệu tương đồng** vào cùng một cụm.

Mục đích:

- Tìm kiếm các mẫu (pattern) trong dữ liệu.
- Hiểu rõ hơn về cấu trúc dữ liệu.
- Phân chia dữ liệu thành các nhóm có ý nghĩa.
- Tăng hiệu quả của các thuật toán khác như phân loại, dự đoán.



1. Giới thiệu

Hệ thống phân cụm dựa trên học không giám sát sử dụng các thuật toán để tự động khám phá cấu trúc ẩn trong dữ liệu. Các thuật toán này thường dựa trên các tiêu chí như:

- **Giống nhau:** Các điểm dữ liệu trong cùng một cụm nên có nhiều điểm chung với nhau, ví dụ như vị trí gần nhau trong không gian đa chiều hoặc có các thuộc tính tương tự nhau.
- **Khác biệt:** Các điểm dữ liệu thuộc các cụm khác nhau nên có sự khác biệt đáng kể so với nhau.



1. Giới thiệu

Phân tích thị trường:

- **Phân khúc khách hàng:** Phân chia khách hàng thành các nhóm có đặc điểm, sở thích và hành vi mua sắm tương tự nhau. Điều này giúp doanh nghiệp có thể nhắm mục tiêu quảng cáo và chiến lược tiếp thị hiệu quả hơn, đồng thời phát triển sản phẩm và dịch vụ phù hợp với nhu cầu của từng nhóm khách hàng.
- **Phân tích xu hướng thị trường:** Xác định các xu hướng mới trong thị trường bằng cách phân tích hành vi mua sắm của khách hàng. Ví dụ, một hệ thống phân cụm có thể phát hiện ra rằng một nhóm khách hàng nhất định đang mua nhiều sản phẩm mới ra mắt gần đây. Doanh nghiệp có thể sử dụng thông tin này để điều chỉnh chiến lược kinh doanh của mình cho phù hợp.



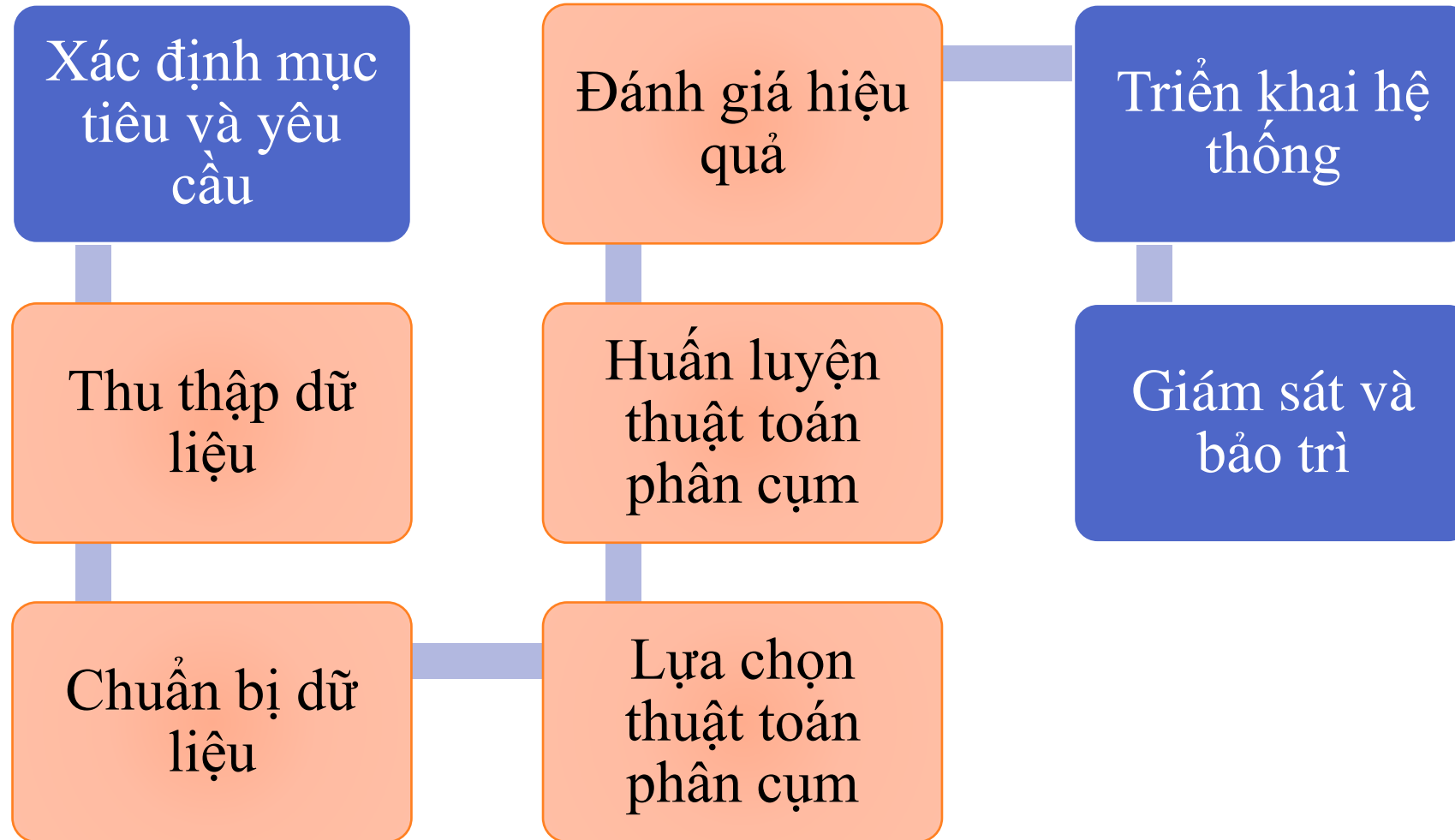
1. Giới thiệu

Phát hiện gian lận:

- **Phát hiện gian lận thẻ tín dụng:** Xác định các giao dịch thẻ tín dụng có khả năng gian lận bằng cách phân tích các mẫu chỉ tiêu. Ví dụ, một hệ thống phân cụm có thể phát hiện ra rằng một số giao dịch thẻ tín dụng được thực hiện từ các địa điểm khác nhau trong một khoảng thời gian ngắn. Điều này có thể là dấu hiệu của hoạt động gian lận.
- **Phát hiện gian lận bảo hiểm:** Phát hiện các yêu cầu bảo hiểm gian lận bằng cách phân tích lịch sử yêu cầu bảo hiểm của khách hàng. Ví dụ, một hệ thống phân cụm có thể phát hiện ra rằng một số khách hàng đã nộp nhiều yêu cầu bảo hiểm cho cùng một loại thiệt hại trong một khoảng thời gian ngắn. Điều này có thể là dấu hiệu của hoạt động gian lận.



2. Quy trình triển khai





2. Quy trình triển khai

1. Xác định mục tiêu và yêu cầu:

- Xác định rõ mục tiêu sử dụng hệ thống phân cụm. Mục tiêu này sẽ giúp lựa chọn thuật toán phân cụm phù hợp và đánh giá hiệu quả của hệ thống.
- Xác định các yêu cầu về hiệu suất, độ chính xác, khả năng mở rộng, v.v. của hệ thống.

2. Thu thập dữ liệu:

- Thu thập dữ liệu cần thiết cho việc phân cụm. Dữ liệu nên được thu thập từ nguồn đáng tin cậy và có chất lượng cao.
- Đảm bảo rằng dữ liệu có đủ số lượng và độ đa dạng để có thể huấn luyện mô hình phân cụm hiệu quả.



2. Quy trình triển khai

3. Chuẩn bị dữ liệu:

- Tiền xử lý dữ liệu để loại bỏ nhiễu, thiếu sót và chuẩn hóa dữ liệu nếu cần thiết.
- Biến đổi dữ liệu nếu cần thiết để phù hợp với thuật toán phân cụm được lựa chọn.

4. Lựa chọn thuật toán phân cụm:

- Có nhiều thuật toán phân cụm khác nhau, mỗi thuật toán có ưu và nhược điểm riêng.
- Lựa chọn thuật toán phù hợp dựa trên mục tiêu sử dụng, đặc điểm của dữ liệu và tài nguyên tính toán.



2. Quy trình triển khai

5. Huấn luyện thuật toán phân cụm:

- Chia dữ liệu thành tập huấn luyện và tập đánh giá.
- Huấn luyện mô hình phân cụm trên tập huấn luyện.
- Điều chỉnh các tham số của thuật toán nếu cần thiết để đạt được kết quả tốt nhất.

6. Đánh giá hiệu quả:

- Đánh giá hiệu quả của mô hình phân cụm trên tập đánh giá.
- Sử dụng các chỉ số đánh giá như Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index, v.v. để đánh giá chất lượng của các cụm.



2. Quy trình triển khai

7. Triển khai hệ thống:

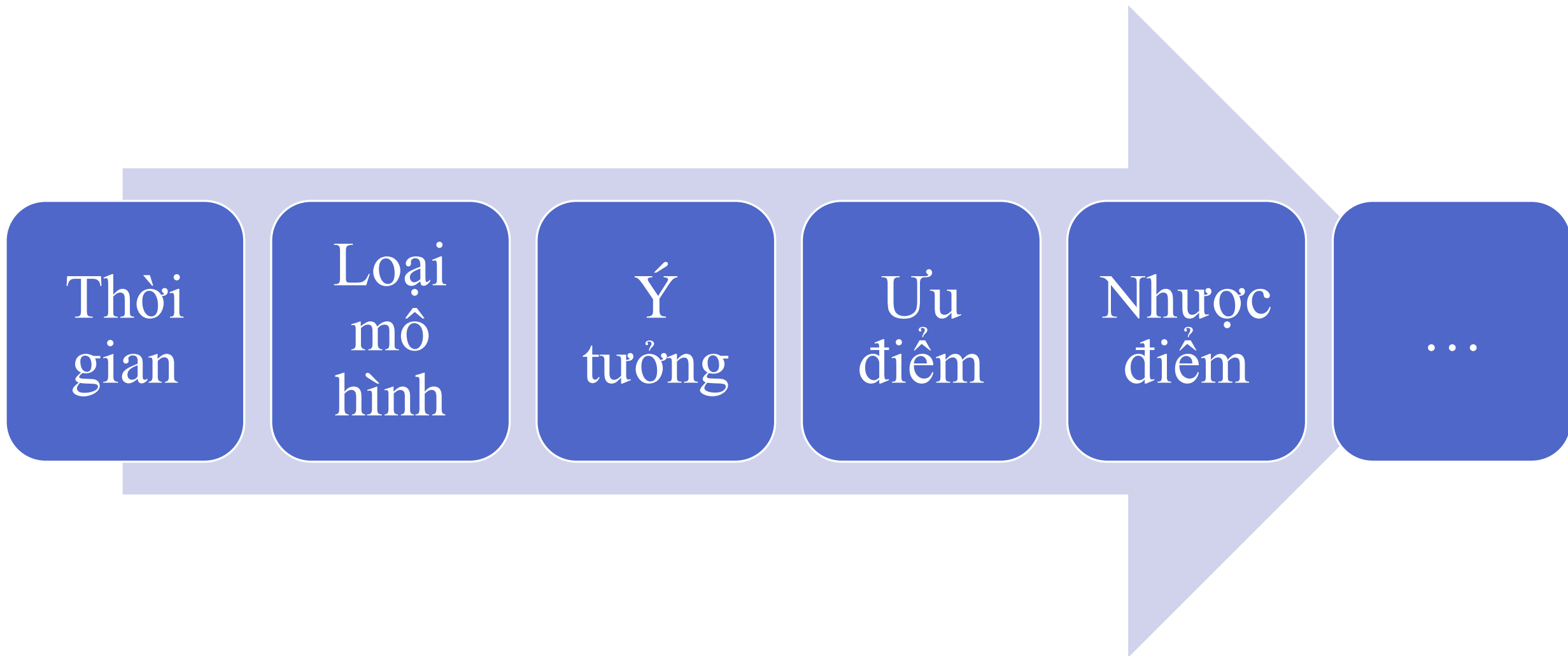
- Tích hợp mô hình phân cụm vào hệ thống ứng dụng.
- Phát triển giao diện người dùng để người dùng có thể tương tác với hệ thống.
- Triển khai hệ thống ứng dụng ra môi trường thực tế.

8. Giám sát và bảo trì:

- Theo dõi hiệu quả hoạt động của hệ thống.
- Giám sát chất lượng dữ liệu đầu vào.
- Cập nhật và cải tiến mô hình phân cụm khi cần thiết.



3. Lựa chọn thuật toán





3. Lựa chọn thuật toán

- **K-means:** K-means++, K-means mềm, K-medoids.
- **Phân cụm phân cấp:** Phân cụm phân cấp kết hợp (Agglomerative Clustering), Phân cụm phân cấp phân chia (Divisive Clustering), BIRCH, ...
- **Phân cụm dựa trên mật độ:** DBSCAN, OPTICS, DENCLUE, ...
- **Phân cụm dựa trên mô hình:** Gaussian Mixture Model (GMM), Finite Mixture Model (FMM), Hidden Markov Model (HMM), ...
- **Phân cụm quang phổ:** Spectral Clustering, Normalized Cut, Cheeger Cut, ...
- **Phân cụm dựa trên logic mờ:** Fuzzy c-means, Possibilistic c-means, Rough c-means, ...
- **Phân cụm dựa trên mạng nơ-ron:** Self-Organizing Maps (SOMs), Kohonen's Growing Neural Networks (KGNNs), Competitive Neural Networks (CNNs), ...
- **Phân cụm dựa trên gen:** Phân cụm dựa trên thuật toán di truyền (Genetic Algorithm-Based Clustering - GABC), Phân cụm dựa trên lập trình tiến hóa (Evolutionary Programming-Based Clustering - EPBC), Phân cụm dựa trên lập trình đàn kiến (Ant Colony Optimization-Based Clustering - ACOBC), ...



3. Lựa chọn thuật toán

Thuật toán	Ưu điểm	Nhược điểm	Thích hợp cho
K-means	Đơn giản, hiệu quả, dễ hiểu, dễ triển khai	Yêu cầu số lượng cụm được xác định trước, nhạy cảm với giá trị khởi tạo tâm cụm, khó xử lý dữ liệu nhiễu	Dữ liệu có cấu trúc đơn giản, số lượng cụm rõ ràng
Phân cụm phân cấp	Có thể tự động xác định số lượng cụm, có thể xử lý dữ liệu nhiễu	Phân cụm có thể không cân bằng, khó trực quan hóa kết quả	Dữ liệu có cấu trúc phân cấp, số lượng cụm không xác định trước
Phân cụm dựa trên mật độ	Có thể tự động xác định số lượng cụm, có thể xử lý dữ liệu nhiễu, hiệu quả với dữ liệu có mật độ cao	Phân cụm có thể không cân bằng, khó điều chỉnh tham số mật độ	Dữ liệu có điểm dữ liệu tập trung thành cụm rõ ràng
Phân cụm dựa trên mô hình	Có thể mô hình hóa mối quan hệ phi tuyến tính giữa các điểm dữ liệu, linh hoạt	Phức tạp hơn so với các thuật toán khác, đòi hỏi dữ liệu có chất lượng cao	Dữ liệu có mối quan hệ phi tuyến tính phức tạp

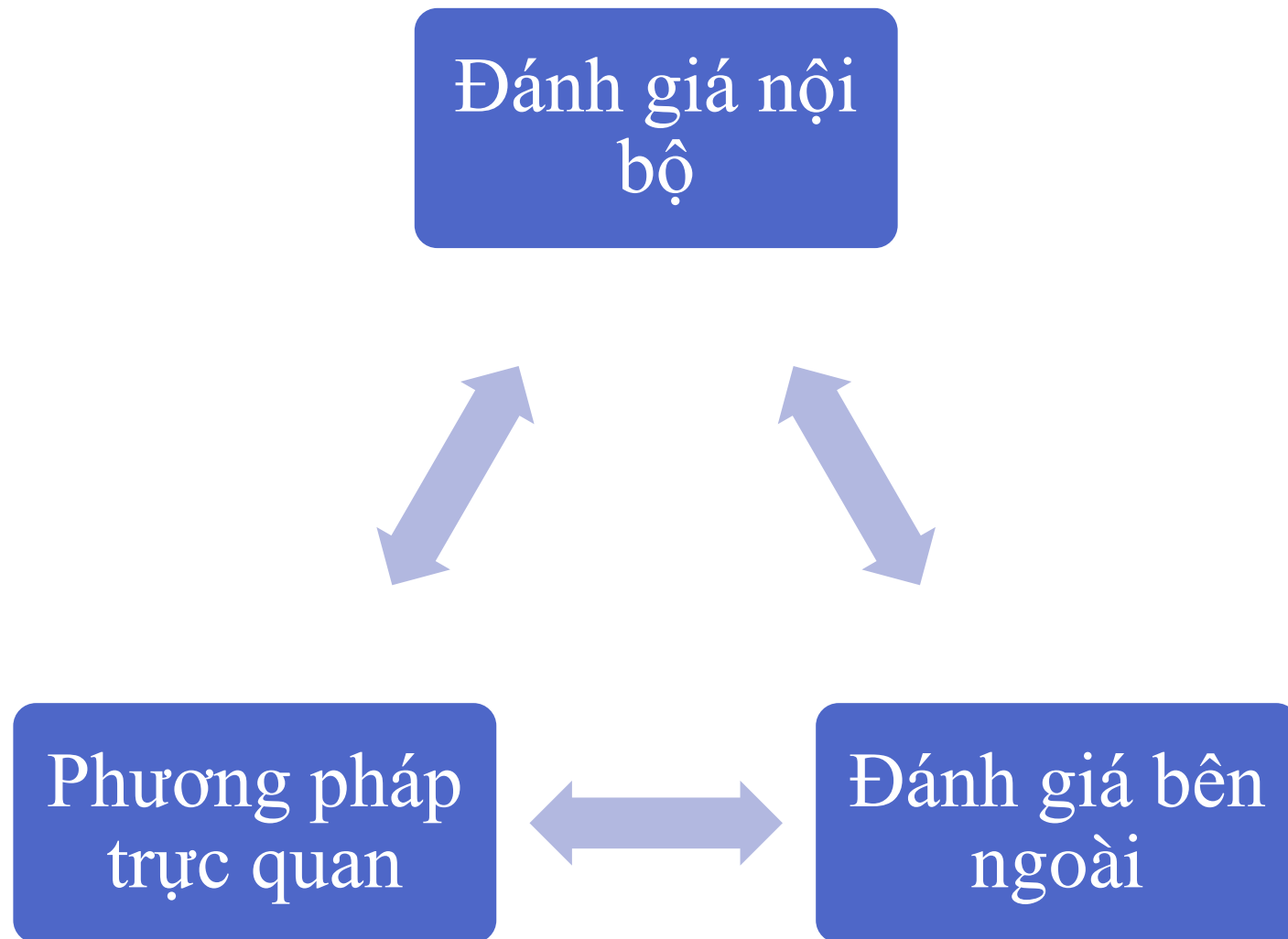


3. Lựa chọn thuật toán

Thuật toán	Ưu điểm	Nhược điểm	Thích hợp cho
Phân cụm quang phổ	Có thể xử lý dữ liệu có cấu trúc phức tạp, hiệu quả với dữ liệu đồ thị	Phức tạp hơn so với các thuật toán khác, đòi hỏi dữ liệu có chất lượng cao	Dữ liệu có cấu trúc phức tạp, dữ liệu đồ thị
Phân cụm dựa trên logic mờ	Có thể xử lý dữ liệu không chắc chắn, linh hoạt	Phức tạp hơn so với các thuật toán khác, đòi hỏi kiến thức về logic mờ	Dữ liệu không chắc chắn, dữ liệu có thuộc tính không rõ ràng
Phân cụm dựa trên mạng nơ-ron	Có thể học các mô hình phân cụm phức tạp từ dữ liệu, hiệu quả với dữ liệu có chiều cao cao	Phức tạp hơn so với các thuật toán khác, đòi hỏi dữ liệu có chất lượng cao, có thể bị mắc kẹt trong cực tiểu cục bộ	Dữ liệu có chiều cao cao, dữ liệu có mô hình phân cụm phức tạp
Phân cụm dựa trên gen	Có thể tìm kiếm các giải pháp tối ưu cho bài toán phân cụm phức tạp	Phức tạp hơn so với các thuật toán khác, đòi hỏi kiến thức về thuật toán di truyền	Dữ liệu có cấu trúc phức tạp, bài toán phân cụm phức tạp



4. Đánh giá hiệu quả





Khoảng cách giữa các điểm

1. Khoảng cách Euclid:

Đây là phương pháp đơn giản và phổ biến nhất, được sử dụng để đo khoảng cách giữa hai điểm trong không gian Euclidean. Công thức tính toán khoảng cách Euclid giữa *hai điểm* x và y với n chiều là:

$$d(x, y) = \sqrt{\sum_{i=1, 2, \dots, n} (x_i - y_i)^2}$$

2. Khoảng cách Manhattan:

Khoảng cách Manhattan (còn gọi là khoảng cách thành phố) đo khoảng cách giữa hai điểm bằng tổng giá trị tuyệt đối của các chênh lệch tọa độ tương ứng. Công thức tính toán khoảng cách Manhattan giữa *hai điểm* x và y với n chiều là:

$$d(x, y) = \sum_{i=1, 2, \dots, n} |x_i - y_i|$$



Khoảng cách giữa các điểm

3. Khoảng cách Hamming:

Khoảng cách Hamming đo số lượng bit khác nhau giữa hai chuỗi có cùng độ dài. Nó thường được sử dụng để so sánh các chuỗi văn bản hoặc mã nhị phân. Công thức tính toán khoảng cách Hamming giữa hai chuỗi x và y có độ dài n là:

$$d(x, y) = \sum (x_i \neq y_i, i = 1, 2, \dots, n)$$

4. Khoảng cách Jaccard:

Khoảng cách Jaccard đo tỷ lệ phần trăm các phần tử không chung nhau giữa hai tập hợp. Nó thường được sử dụng để so sánh các tập dữ liệu dạng tập hợp. Công thức tính toán khoảng cách Jaccard giữa hai tập hợp X và Y là:

$$d(X, Y) = 1 - |X \cap Y| / |X \cup Y|$$




Khoảng cách giữa các điểm

Phụ thuộc vào nhiều yếu tố: loại dữ liệu, mục tiêu phân nhóm và thông tin có sẵn.



Phụ thuộc thuật toán gom cụm.



Dữ liệu là các điểm trong không gian Euclidean → khoảng cách Euclid; các chuỗi văn bản → khoảng cách Hamming hoặc Jaccard.



Nhãn cụm (cluster labels)

Nhãn cụm này đại diện cho việc điểm dữ liệu đó thuộc về cụm nào. Có hai cách chính để lấy nhãn cụm:

1. Sử dụng thuật toán gom nhóm:

- Có thể sử dụng một thuật toán gom nhóm để phân chia tập dữ liệu thành các cụm và gán nhãn cụm cho từng điểm dữ liệu.
- Phân cụm khách quan về cấu trúc vốn có của dữ liệu, tự động hóa.

2. Sử dụng dữ liệu được gán nhãn sẵn:

- Nếu có tập dữ liệu được gán nhãn sẵn, có thể trực tiếp sử dụng những nhãn này làm nhãn cụm.
- Tạo trước nhãn cụm, mang tính chủ quan.



Đánh giá nội bộ

Silhouette Score: Đo lường mức độ trung bình của mỗi điểm dữ liệu nằm trong cụm của nó so với các cụm lân cận. *Giá trị cao hơn cho thấy sự phân nhóm tốt hơn.*

Với mỗi điểm dữ liệu i : $s(i) = (b(i) - a(i)) / \max(a(i), b(i))$

- **$a(i)$** : Đây là khoảng cách trung bình giữa điểm i và các điểm khác trong cùng cụm với nó.
- **$b(i)$** : Đây là khoảng cách trung bình nhỏ nhất giữa điểm i và các điểm trong một cụm khác.



Đánh giá nội bộ

Silhouette Score: Đo lường mức độ trung bình của mỗi điểm dữ liệu nằm trong cụm của nó so với các cụm lân cận. *Giá trị cao hơn cho thấy sự phân nhóm tốt hơn.*

Giá trị $s(i)$ có thể nằm trong khoảng $[-1, 1]$:

- $s(i) > 0$: Điểm i nằm tốt trong cụm của nó.
- $s(i) = 0$: Điểm i nằm trên ranh giới giữa hai cụm.
- $s(i) < 0$: Điểm i có thể được phân nhóm tốt hơn vào một cụm khác.

Tính toán Silhouette Score cho tất cả các điểm → Tính toán Silhouette Score trung bình cho toàn bộ tập dữ liệu → Giá trị Silhouette Score trung bình cao hơn cho thấy sự phân nhóm tốt hơn.



Đánh giá nội bộ

Calinski-Harabasz Index: Đo lường sự phân chia giữa các cụm và độ nén của các điểm dữ liệu trong mỗi cụm. *Giá trị cao hơn cho thấy sự phân nhóm tốt hơn.*

Để tính toán CHI cho một tập dữ liệu được phân nhóm thành k cụm, ta thực hiện các bước sau:

- **B:** Tổng bình phương khoảng cách giữa mỗi điểm dữ liệu và tâm của cụm mà nó thuộc về.
- **W:** Tổng bình phương khoảng cách giữa các tâm của các cụm.
- Tính toán CHI: $\text{CHI} = (\mathbf{B} / (\mathbf{k} - 1)) / (\mathbf{W} / \mathbf{k})$



Đánh giá nội bộ

Calinski-Harabasz Index: Đo lường sự phân chia giữa các cụm và độ nén của các điểm dữ liệu trong mỗi cụm. *Giá trị cao hơn cho thấy sự phân nhóm tốt hơn.*

Giá trị CHI có thể nằm trong khoảng $[0, \infty]$:

- **CHI > 1:** Sự phân chia giữa các cụm tốt hơn độ nén của các điểm dữ liệu trong mỗi cụm.
- **CHI = 1:** Sự phân chia giữa các cụm và độ nén của các điểm dữ liệu trong mỗi cụm tương đương nhau.
- **CHI < 1:** Độ nén của các điểm dữ liệu trong mỗi cụm tốt hơn sự phân chia giữa các cụm.



Đánh giá nội bộ

Davies-Bouldin Index: Đo lường độ tách biệt giữa các cụm và độ nén của các điểm dữ liệu trong mỗi cụm. *Giá trị thấp hơn cho thấy sự phân nhóm tốt hơn.*

Để tính toán DBI cho một tập dữ liệu được phân nhóm thành k cụm, ta thực hiện các bước sau:

- **S_i :** Đây là tổng bình phương khoảng cách giữa mỗi điểm dữ liệu trong cụm i và tâm của cụm i .
- **$R_{\{ij\}}$:** Đây là giá trị trung bình của S_i và S_j , nơi i và j là hai cụm khác nhau.
- **D_i :** Đây là khoảng cách trung bình giữa tâm của cụm i và tâm của các cụm khác.
- Tính toán DBI: $DBI = (1 / k) * \sum(D_i / \max(R_{\{ij\}}))$



Đánh giá nội bộ

Davies-Bouldin Index: Đo lường độ tách biệt giữa các cụm và độ nén của các điểm dữ liệu trong mỗi cụm. *Giá trị thấp hơn cho thấy sự phân nhóm tốt hơn.*

Giá trị DBI có thể nằm trong khoảng $[0, \infty]$:

- **DBI < 1:** Sự phân chia giữa các cụm tốt và độ nén của các điểm dữ liệu trong mỗi cụm tốt.
- **DBI = 1:** Sự phân chia giữa các cụm và độ nén của các điểm dữ liệu trong mỗi cụm tương đương nhau.
- **DBI > 1:** Độ nén của các điểm dữ liệu trong mỗi cụm tốt hơn sự phân chia giữa các cụm.



Đánh giá bên ngoài

Purity: Đo tỷ lệ phần trăm các điểm dữ liệu trong mỗi cụm có cùng nhãn. *Giá trị cao hơn cho thấy sự phân nhóm tốt hơn.*

Cách tính Purity:

- 1. Tính số lượng điểm dữ liệu trong mỗi cụm:** Đối với mỗi cụm C_i , đếm số lượng điểm dữ liệu n_i thuộc về cụm đó.
- 2. Tính số lượng điểm dữ liệu có cùng nhãn trong mỗi cụm:** Đối với mỗi nhãn l , đếm số lượng điểm dữ liệu n_{il} trong cụm C_i có nhãn l .
- 3. Tính Purity cho mỗi cụm:** $\text{Purity}(C_i) = \max(n_{il}) / n_i$
- 4. Tính Purity tổng thể:** $\text{Purity}(D) = \text{mean}(\text{Purity}(C_i))$



Đánh giá bên ngoài

Purity: Đo tỷ lệ phần trăm các điểm dữ liệu trong mỗi cụm có cùng nhãn. *Giá trị cao hơn cho thấy sự phân nhóm tốt hơn.*

➤ **Purity cao:** Khi Purity cao, nghĩa là có nhiều cụm chứa phần lớn các điểm dữ liệu có cùng nhãn. Điều này cho thấy thuật toán phân cụm đã phân chia dữ liệu hiệu quả, tách biệt các lớp khác nhau.

➤ **Purity thấp:** Khi Purity thấp, nghĩa là có nhiều cụm chứa nhiều điểm dữ liệu thuộc về nhiều nhãn khác nhau. Điều này cho thấy thuật toán phân cụm không tách biệt các lớp hiệu quả, dẫn đến các cụm không đồng nhất.

➤ *Purity có thể bị ảnh hưởng bởi tỷ lệ phần trăm các lớp trong tập dữ liệu.* Ví dụ, nếu một lớp chiếm phần lớn dữ liệu, Purity có thể cao mặc dù thuật toán phân cụm không hiệu quả.



Đánh giá bên ngoài

Entropy: Đo mức độ hỗn loạn của các nhãn trong mỗi cụm. *Giá trị thấp hơn cho thấy sự phân nhóm tốt hơn.*

Cách tính Entropy:

- 1. Tính tỷ lệ phần trăm của mỗi nhãn trong mỗi cụm:** Đối với mỗi nhãn l trong cụm C_i , tính tỷ lệ phần trăm p_{il} của điểm dữ liệu có nhãn l trong cụm C_i .
- 2. Tính Entropy cho mỗi cụm:** $\text{Entropy}(C_i) = - \sum p_{il} * \log_2(p_{il})$
- 3. Tính Entropy tổng thể:** $\text{Entropy}(D) = \text{mean}(\text{Entropy}(C_i))$



Đánh giá bên ngoài

Entropy: Đo mức độ hỗn loạn của các nhãn trong mỗi cụm. *Giá trị thấp hơn cho thấy sự phân nhóm tốt hơn.*

- **Entropy cao:** Khi Entropy cao, nghĩa là có nhiều nhãn phân bố không đồng đều trong các cụm. Điều này cho thấy thuật toán phân cụm không tách biệt các lớp hiệu quả, dẫn đến các cụm hỗn loạn.
- **Entropy thấp:** Khi Entropy thấp, nghĩa là các nhãn phân bố đều hơn trong các cụm. Điều này cho thấy thuật toán phân cụm đã phân chia dữ liệu hiệu quả, tách biệt các lớp khác nhau.
- *Entropy có thể phát hiện các trường hợp phân nhóm không hiệu quả, ngay cả khi Purity cao.*



Đánh giá bên ngoài

F1 Score: Đo sự cân bằng giữa độ chính xác và độ thu hồi của thuật toán gom nhóm trong việc phân loại các điểm dữ liệu. *Giá trị cao hơn cho thấy sự phân nhóm tốt hơn.*

- **Độ chính xác (Precision):** Tỷ lệ các điểm dữ liệu được phân loại chính xác vào một cụm cụ thể.
- **Độ thu hồi (Recall):** Tỷ lệ các điểm dữ liệu thực sự thuộc về một cụm cụ thể được phân loại chính xác vào cụm đó.
- **Cách tính F1 Score:** $F1 = 2 * (Precision * Recall) / (Precision + Recall)$



Đánh giá bên ngoài

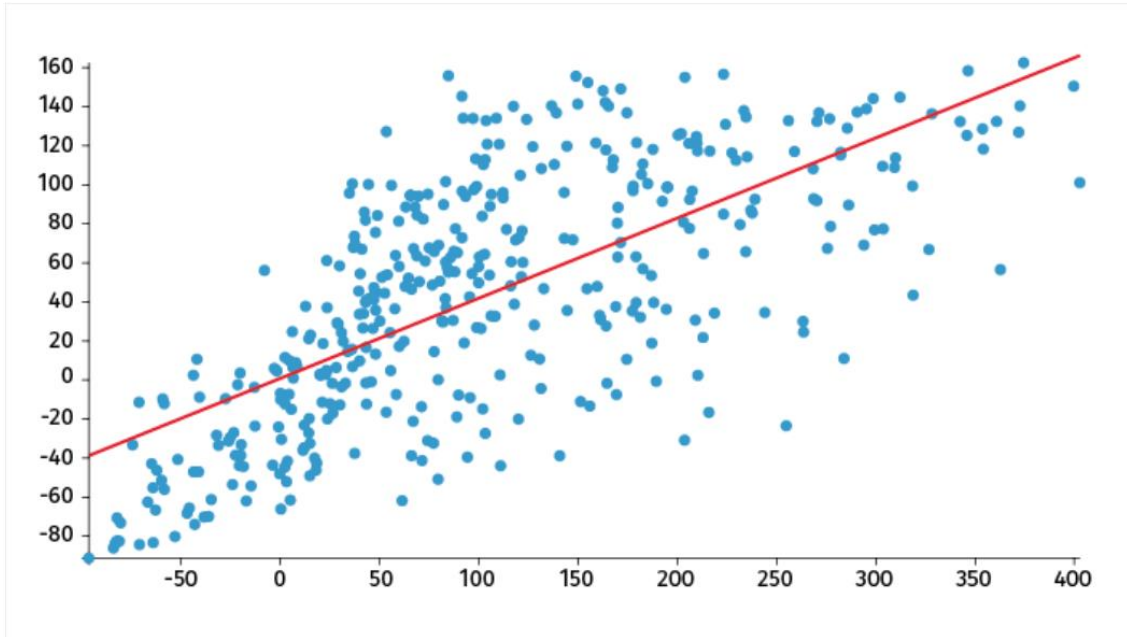
F1 Score: Đo sự cân bằng giữa độ chính xác và độ thu hồi của thuật toán gom nhóm trong việc phân loại các điểm dữ liệu. *Giá trị cao hơn cho thấy sự phân nhóm tốt hơn.*

- **F1 cao:** Chỉ ra rằng thuật toán phân cụm có khả năng phân loại chính xác các điểm dữ liệu, cân bằng tốt giữa việc giảm thiểu lỗi phân loại sai (False Negative) và lỗi phân loại dương tính sai (False Positive).
- **F1 thấp:** Chỉ ra rằng thuật toán phân cụm không hiệu quả trong việc phân loại các điểm dữ liệu, có thể thiên về một hoặc cả hai yếu tố Độ chính xác và Độ thu hồi.
- *F1 Score, Purity và Entropy là ba chỉ số phổ biến để đánh giá chất lượng phân cụm. Việc sử dụng kết hợp cả ba chỉ số này có thể cung cấp cái nhìn toàn diện hơn về hiệu quả của thuật toán phân cụm.*

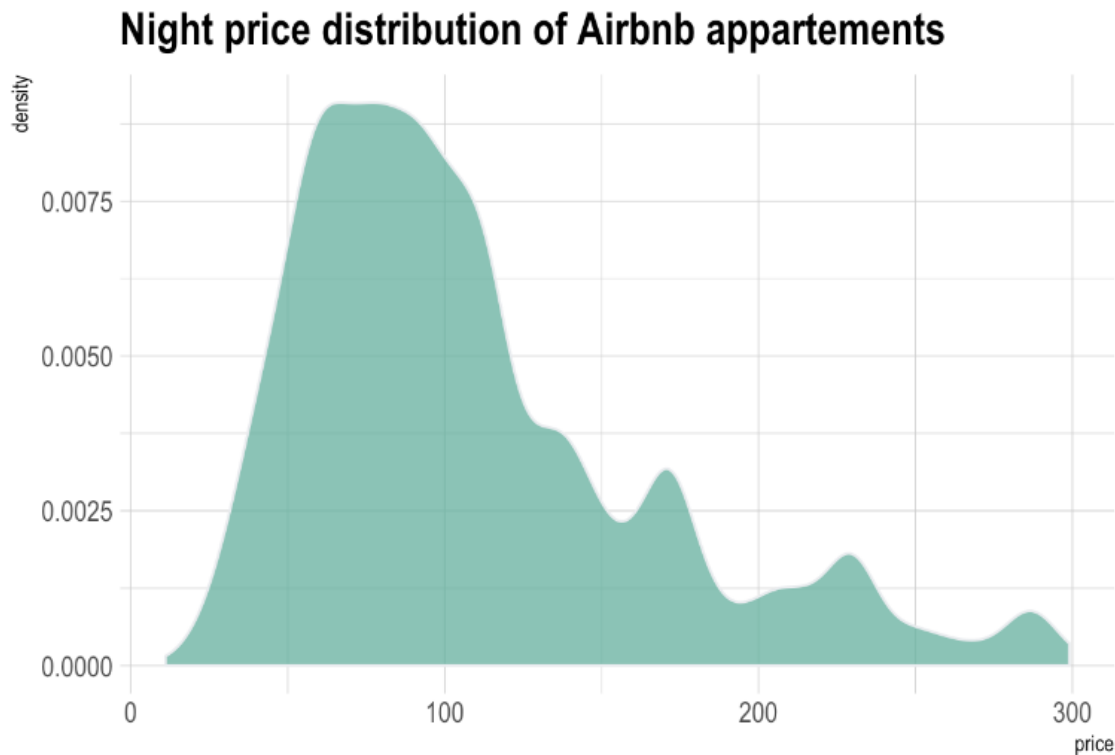
Phương pháp trực quan

Vizuualize: Biểu đồ phân bố các điểm dữ liệu trong không gian hai chiều hoặc ba chiều có thể giúp đánh giá trực quan chất lượng của các cụm.

➤ **Biểu đồ tán xạ (Scatter Plot):** Hiện thị mỗi điểm dữ liệu dưới dạng một điểm trong không gian hai chiều, với các điểm có cùng nhãn được tô màu hoặc biểu tượng hóa giống nhau.

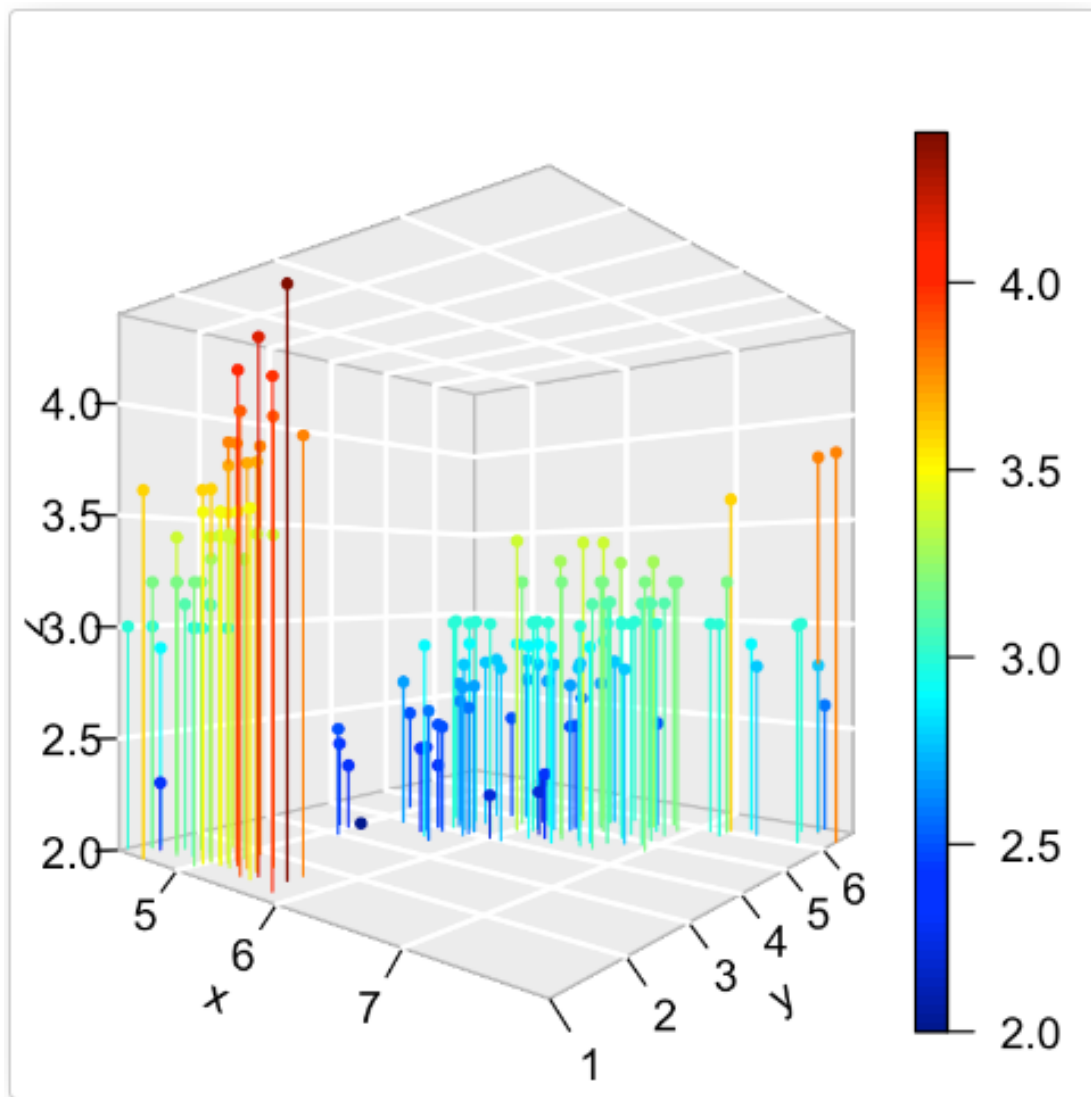


Phương pháp trực quan



Vizuualize: Biểu đồ phân bố các điểm dữ liệu trong không gian hai chiều hoặc ba chiều có thể giúp đánh giá trực quan chất lượng của các cụm.

➤ **Biểu đồ mật độ (Density Plot):** Ước tính mật độ phân bố của các điểm dữ liệu trong không gian hai chiều, giúp hiển thị rõ ràng các cụm tập trung.

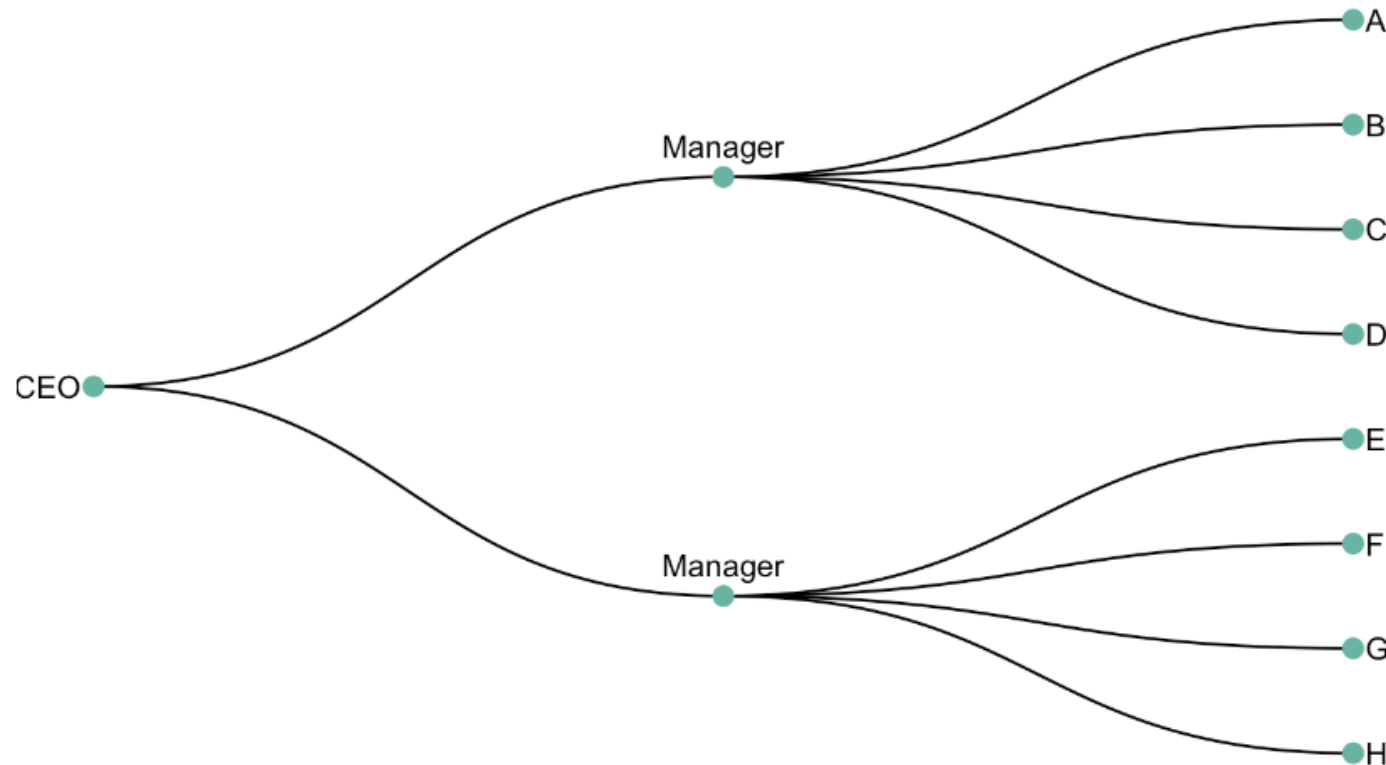


Phương pháp trực quan

Vizuualize: Biểu đồ phân bố các điểm dữ liệu trong không gian hai chiều hoặc ba chiều có thể giúp đánh giá trực quan chất lượng của các cụm.

➤ **Biểu đồ 3D:** Hiển thị các điểm dữ liệu trong không gian ba chiều, cho phép quan sát các cụm từ nhiều góc độ khác nhau.

Phương pháp trực quan



Dendrogram: Biểu đồ dạng cây biểu thị mối quan hệ phân cấp giữa các cụm.

➤ Hiện thị thứ tự phân cấp của các cụm, giúp hiểu rõ cách các cụm được liên kết với nhau.



Phương pháp trực quan

Cách sử dụng biểu đồ để đánh giá chất lượng phân cụm:

- **Quan sát sự phân bố của các điểm dữ liệu:** Các điểm dữ liệu trong cùng một cụm nên được nhóm lại gần nhau và tách biệt với các điểm dữ liệu của các cụm khác.
- **Kiểm tra hình dạng của các cụm:** Các cụm có hình dạng tốt (ví dụ: hình cầu, hình elip) thường cho thấy sự phân nhóm hiệu quả hơn so với các cụm có hình dạng không đều.
- **Đánh giá sự chồng chéo giữa các cụm:** Nếu có nhiều điểm dữ liệu nằm giữa các cụm hoặc thuộc về nhiều cụm, điều đó có thể cho thấy thuật toán phân cụm không hiệu quả.
- **Sử dụng các chỉ số khác:** Kết hợp trực quan hóa dữ liệu với các chỉ số chất lượng phân cụm như Purity, Entropy và F1 Score để có đánh giá toàn diện hơn.



5. Triển khai hệ thống

1. Chuẩn bị:

- **Mô hình:** Mô hình đã được huấn luyện và đánh giá hiệu quả.
- **Ứng dụng web / Ứng dụng di động:** Đã được phát triển và sẵn sàng để tích hợp mô hình.

2. Triển khai mô hình trên server:

- **Môi trường:** Cloud Platforms / On-Premise Servers
- **Mô hình được triển khai trên server và ứng dụng sẽ gửi yêu cầu đến server để thực hiện tác vụ.**
 - **Lưu trữ mô hình:** Mô hình được lưu trữ trên server.
 - File format: .h5 (HDF5), .pb (Protocol Buffer), .pkl (Pickle), v.v.
 - **Tạo API:** API được tạo ra để nhận yêu cầu từ ứng dụng và gửi kết quả.
 - REST API: Tạo RESTful API bằng các frameworks như Flask, FastAPI, v.v.
 - **Gọi API:** Ứng dụng gửi yêu cầu đến API và nhận kết quả.



6. Giám sát và bảo trì

Giám sát:

- **Theo dõi các chỉ số chính:** Theo dõi thường xuyên các chỉ số phản ánh hiệu suất và tình trạng của mô hình.
- **Phát hiện dữ liệu bất thường:** Phân phối dữ liệu thực tế được cung cấp cho mô hình bắt đầu khác với dữ liệu mà nó được đào tạo.
- **Cảnh báo và ghi nhật ký:** Theo dõi hành vi của mô hình và chẩn đoán các vấn đề dễ dàng hơn.

Bảo trì:

- **Đào tạo lại:** Thường xuyên đào tạo lại mô hình với dữ liệu mới, đặc biệt nếu phát hiện dữ liệu bất thường. Điều này giúp mô hình thích ứng với những thay đổi trong thế giới thực và duy trì hiệu suất tối ưu.

Công cụ và tài nguyên: TensorBoard, Mlflow, Cloud Monitoring Services.



7. Tổng kết

Khoảng cách
giữa các điểm

Nhãn cụm
(cluster labels)

Question & Answer
