TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC

NHẬP MÔN HỌC MÁY

BÀI KIỂM TRA GIỮA KỲ

Người hướng dẫn: **GV. LÊ ANH CƯỜNG**

Người thực hiện: **NGUYỄN ĐỨC TÍN – 51800248**

NGUYỄN VIỆT TÂN – 51800621

Lớp **: 18050303**

Khoá **: 22**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC

NHẬP MÔN HỌC MÁY

BÀI KIỂM TRA GIỮA KỲ

Người hướng dẫn: **GV. LÊ ANH CƯỜNG**

Người thực hiện: **NGUYỄN ĐỨC TÍN – 51800248**

NGUYỄN VIỆT TÂN – 51800621

Lớp : **18050303**

Khoá : **22**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022

LỜI CẢM ƠN

Em xin chân thành cảm ơn thầy Lê Anh Cường đã giảng dạy tận tình trong suốt quá trình học tập, để chúng em có đủ kiến thức về lý thuyết. Em xin chúc cho thầy thật nhiều sức khỏe và gặt hái được những thành công trong cuộc sống.

Với điều kiện thời gian cũng như kinh nghiệm còn hạn chế, bài giữa kỳ này không thể tránh được những thiếu sót. Chúng em mong sẽ nhận được sự đóng góp ý kiến từ thầy để có thể tìm ra những vấn đề còn hạn chế và bổ sung, nâng cao kiến thức của mình, phục vụ cho công tác thực tế sau này.

Chúng em xin chân thành cảm ơn.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi / chúng tôi và được sự hướng dẫn của Thầy Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm

Tác giả

(ký tên và ghi rõ họ tên)

Nguyễn Việt Tân

Nguyễn Đức Tín

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

LỜI MỞ ĐẦU

Kể từ khi ra đời, khai thác dữ liệu (Data mining) đã trở thành một trong những xu hướng nghiên cứu phổ biến nhất trong học tập máy tính và công nghệ tri thức. Nhiều kết quả nghiên cứu của khai phá dữ liệu được ứng dụng thực tế. Có nhiều hướng quan trọng để khai thác dữ liệu, một trong số đó là phân cụm dữ liệu. Phân cụm là một quá trình tìm kiếm để phân biệt các cụm dữ liệu, các mẫu dữ liệu từ nhiều cơ sở dữ liệu lớn.

MỤC LỤC

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC KÝ HIỆU

CÁC CHỮ VIẾT TẮT

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

DANH MỤC HÌNH

DANH MỤC BẢNG

CHƯƠNG 1 –PHÂN CỤM DỮ LIỆU

1.1 Khái niệm

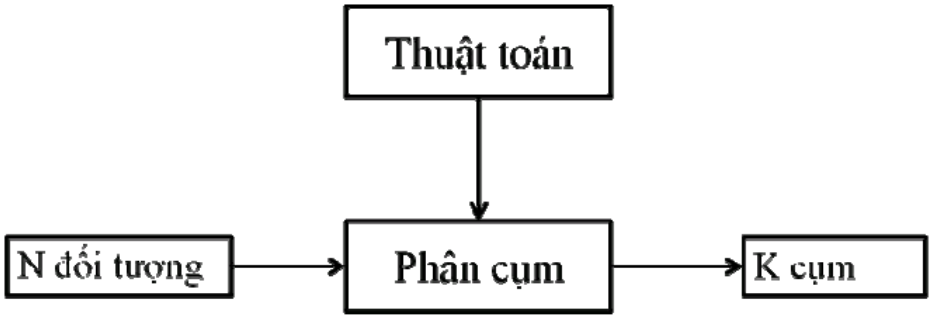
Phân cụm là một kỹ thuật rất quan trọng trong khai thác dữ liệu và thuộc về lớp phương pháp học không giám sát của học máy. Phân cụm là quá trình nhóm các đối tượng cụ thể thành một cụm và tìm cách làm cho các đối tượng trong cùng một cụm giống nhau và các đối tượng khác cụm thì trở nên khác biệt.

Mục đích của phân cụm là để tìm ra bản chất bên trong các nhóm của dữ liệu. Thuật toán phân cụm (Clustering Algorithms) đều sinh ra các cụm (Clusters).

Kỹ thuật phân cụm có thể áp dụng cho rất nhiều lĩnh vực như: Marketing, Libraries, Biology, Insurance, Finance, WWW, …

Các kỹ thuật phân cụm được phân loại như sau:

- Phương pháp phân hoạch
- Phương pháp phân cấp
- Phương pháp dựa trên mật độ
- Phương pháp dựa trên mô hình
- Phương pháp dựa trên lưới



Hình 1. 1 Quy trình phân cụm

1.2 Các kỹ thuật tiếp cận phân cụm dữ liệu

1.2.1 Phương pháp phân cụm phân hoạch

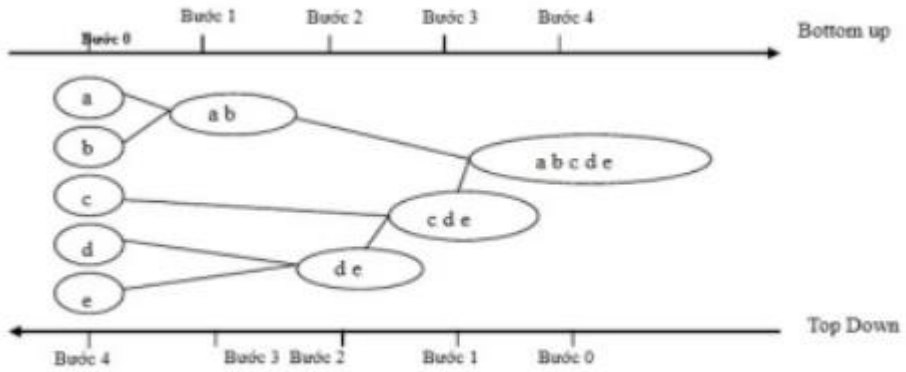
Kỹ thuật này chia một tập dữ liệu có n phần tử thành k nhóm cho đến khi số lượng cụm được xác định. Số lượng cụm được thiết lập là các đặc trưng được lựa chọn trước, phương pháp này thích hợp để tìm các cụm hình cầu trong không gian Euclide.

Các thuật toán phân hoạch dữ liệu có độ rất phức tạp trong việc xác định nghiệm giải pháp tối ưu toàn cục cho vấn đề phân cụm dữ liệu. Ý tưởng chính của thuật toán phân cụm phân hoạch tối ưu cục bộ là sử dụng chiến lược ăn tham (Greedy) để tìm kiếm nghiệm.

1.2.2 Phương pháp phân cụm phân cấp

Phương pháp này xây dựng một hệ thống phân cấp dựa trên các đối tượng dữ liệu được xem xét. Có nghĩa là, cây phân cấp này được xây dựng theo cách đệ quy để sắp xếp một tập dữ liệu cụ thể trong một cấu trúc giống như cây. Có hai cách tiếp cận chung cho cách tiếp cận này:

- Hòa nhập nhóm, thường được gọi là tiếp cận Bottom-Up
- Phân chia nhóm, thường được gọi là tiếp cận Top-Down



Hình 1. 2 Chiến lược phân cụm phân cấp

1.2.3 Phương pháp phân cụm dựa trên mật độ

Kỹ thuật này nhóm các đối tượng dữ liệu dựa trên một hàm mật độ cụ thể. Mật độ là số lượng vùng lân cận trong đối tượng dữ liệu. Kỹ thuật phân cụm có thể xác định các cụm dữ liệu dựa trên mật độ của các đối tượng và phát hiện các cụm dữ liệu có hình dạng bất kỳ. Mặc dù kỹ thuật này có thể khắc phục rất tốt các giá trị ngoại lai và nhiễu, nhưng rất khó xác định các tham số mật độ của thuật toán, nhưng các tham số này có tác động đáng kể đến kết quả phân cụm.

1.2.4 Phương pháp phân cụm dựa trên lưới

Kỹ thuật phân cụm dựa trên lưới phù hợp với dữ liệu đa chiều dựa trên cấu trúc dữ liệu lưới của phân cụm, và phương pháp này chủ yếu được áp dụng cho các lớp dữ liệu không gian. Mục đích của phương pháp này là định lượng dữ liệu trong các ô tạo thành cấu trúc dữ liệu lưới. Trong trường hợp đó, hoạt động phân cụm chỉ hoạt động trên các đối tượng trong mỗi ô của lưới, không phải các đối tượng dữ liệu.

Ưu điểm của phương pháp phân cụm dựa trên lưới là thời gian xử lí nhanh và độc lập với số đối tượng dữ liệu trong tập dữ liệu ban đầu, thay vào đó là chúng phụ thuộc vào số ô trong mỗi chiều của không gian lưới

1.2.5 Phương pháp phân cụm dựa trên mô hình

Các kỹ thuật phân cụm dựa trên mô hình cố gắng điều chỉnh dữ liệu phù hợp với một mô hình toán học dựa trên giả định rằng dữ liệu được tạo ra bằng cách sử dụng hỗn hợp các phân phối xác suất cơ bản. Có hai cách tiếp cận chính đối với các thuật toán phân cụm dựa trên mô hình: mô hình thống kê và mạng nơ-ron.

Phương pháp này tương tự như phân cụm dựa trên mật độ ở chỗ nó phát triển các cụm riêng lẻ để mở rộng mô hình đã xác định trước đó, nhưng thay vì bắt đầu với một số cụm cố định và không sử dụng cùng một khái niệm mật độ cho các cụm.

CHƯƠNG 2 – CÁC THUẬT TOÁN PHÂN CỤM DỮ LIỆU

2.1 Thuật toán phân cụm dữ liệu dựa vào phân cụm phân cấp

2.1.1 Thuật toán BIRCH

BIRCH: **B**alanced **I**terative **R**educing **C**lustering Using **H**ierarchies.

Được đề xuất năm 1996 bởi Tian Zhang, Amakrishnan và Livny.

BIRCH là thuật toán phân cụm phân cấp sử dụng chiến lược TOP DOWN

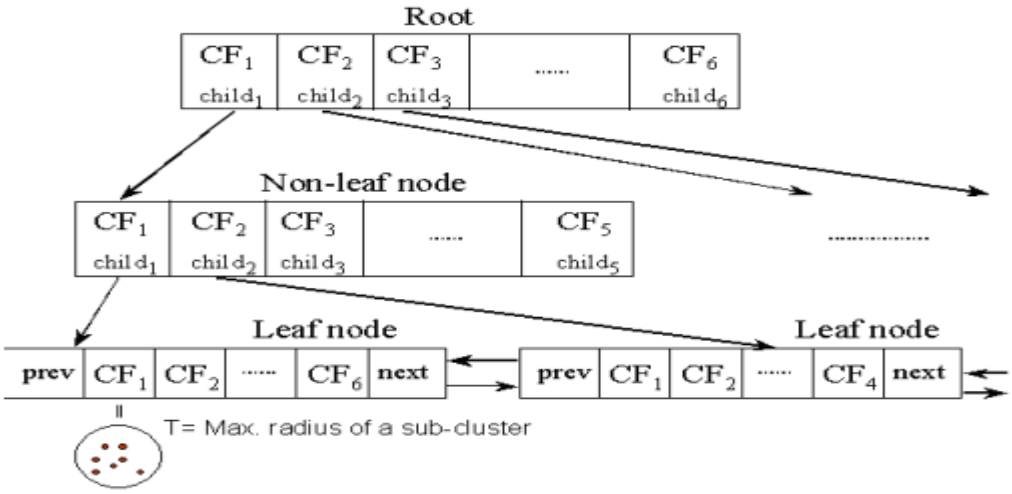
Phương pháp Top Down: Bắt đầu với trạng thái là tất cả các đối tượng được xếp trong cùng một cụm.

Mỗi vòng lặp thành công, một cụm được tách thành các cụm nhỏ hơn theo giá trị của một phép đo độ tương tự nào đó cho đến khi mỗi đối tượng là một cụm hoặc cho đến khi điều kiện dừng thỏa mãn.

Cách tiếp cận này sử dụng chiến lược chia để trị trong quá trình phân cụm.

Ý tưởng của thuật toán là không cần lưu toàn bộ các đối tượng dữ liệu của các cụm trong bộ nhớ mà chỉ cần lưu các đại lượng thống kê.

Đối với mỗi cụm dữ liệu, BIRCH chỉ lưu một bộ ba (n, LS, SS), với n là số các điểm trong phân hoạch cụm con, LS là tổng số các giá trị thuộc tích và SS là tổng bình phương của các điểm đó. Các bộ ba này được gọi là các đặc trưng của cụm CF = (n, LS, SS) (Cluster Features CF) và được lưu giữ trong một cây được gọi là cây CF.



Hình 2. 1 Ví dụ về cây CF

Hình phía trên là biểu thị một ví dụ về cây CF. Chúng ta thấy rằng tất cả các nút trong lưu tổng các đặc trưng cụm CF của nút con, trong khi đó các nút lá lưu trữ các đặc trưng của cụm dữ liệu.

Cây CF là cây cân bằng, nhằm để lưu trữ các đặc trưng của cụm (CF). Cây CF chứa các nút trong và nút lá, nút trong là nút chứa các nút con và nút là thì không có con. Nút trong lưu trữ tổng các đặc trưng cụm (CF) của các nút con của nó.

Một cây (CF) được đặc trưng bởi hai tham số :

- Yếu tố nhánh (Braching Factor – B) : Nhằm xác định tối đa các nútcon của một nút lá trong của cây
- Ngưỡng (Threshold – T) : khoảng cách tối đa giữa bất kỳ một cặp đốitượng trong nút lá của cây, khoảng cách này còn gọi là đường kính của cáccụm con được lưu tại các nút lá.

Hai tham số này có ảnh hưởng đến kích thước của cây CF..

Các bước cơ bản để thực hiện thuật toán BIRCH:

Bước 1: Các đối tượng dữ liệu lần lượt được chèn vào cây C, chèn tất cả các đối tượng và sau đó nhận được cây CF đã khởi tạo. Đối tượng được chèn trong nút gần nhất với sự hình thành của phân lớp. Nếu đường kính của phân lớp này lớn hơn T, nút lá sẽ bị tách ra. Khi đối tượng thích hợp được chèn vào nút lá, tất cả các nút trỏ đến gốc của cây đều được cập nhật thông tin cần thiết.

Bước 2: Nếu cây CF hiện tại hết bộ nhớ khi tạo cây CF nhỏ: Kích thước của cây CF được điều khiển bởi tham số F, vì vậy việc chọn giá trị lớn sẽ hợp nhất nhiều phân nhóm thành một. Việc này làm cho các cụm và cây CF nhỏ hơn. Quy trình này không yêu cầu bạn đọc dữ liệu từ đầu, nhưng nó đảm bảo rằng cây dữ liệu nhỏ sẽ được sửa đổi.

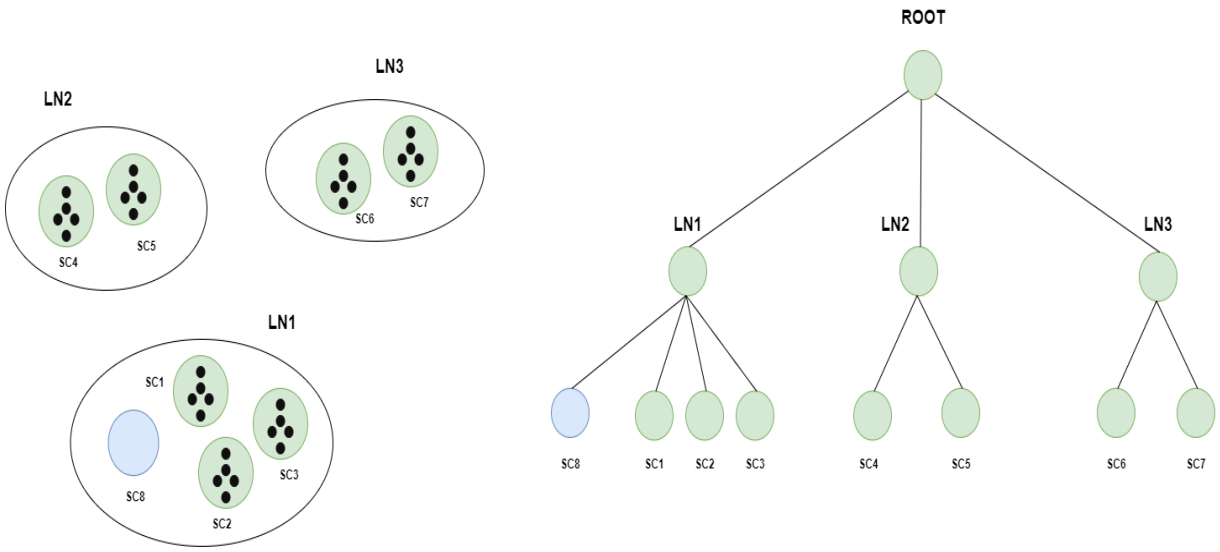
Bước 3: Thực hiện phân cụm: Các nút lá cây CF lưu trữ các đại lượng thống kê của các cụm con. BIRCH sử dụng các đại lượng thống kê này để áp dụng một số kỹ thuật phân cụm, ví dụ K-means và tạo ra một khởi tạo cho phân cụm.

Bước 4: Phân phối lại đối tượng dữ liệu bằng cách sử dụng đối tượng tiêu điểm của cụm được xác định ở bước 3. Đây là bước tùy chọn để xem lại tập dữ liệu và ánh xạ

lại một đối tượng dữ liệu thành một đối tượng dữ liệu khác. Ở bước này, ta đánh dấu dữ liệu khởi tạo và loại bỏ các đối tượng ngoại lai.

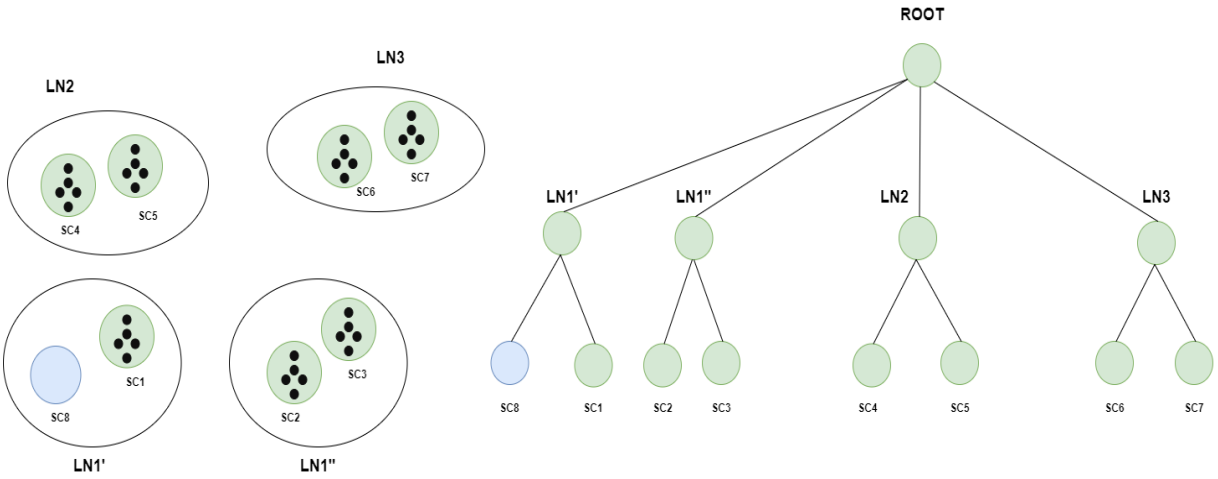
Khi hòa nhập 2 cụm ta có : CF=CF1+CF2=(n1+n2; LS1+LS2; SS1+SS2)

Khoảng cách giữa các cụm có thể đo bằng khoảng cách Euclid, Manhatta,…



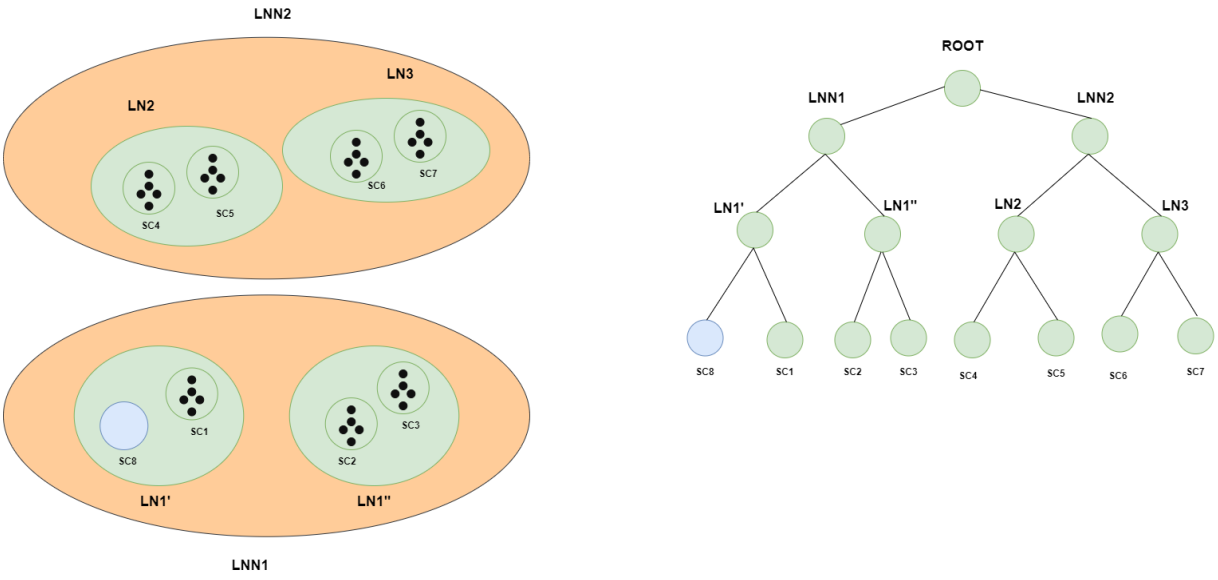
Hình 2. 2 Ví dụ hoạt động chèn nút lá

Vì các yếu tố phân nhánh của một nút lá không thể vượt quá ba nên sau đó LN1 được phân chia (Hình 2. 3).



Hình 2. 3 LN1 được phân chia

Các yếu tố phân nhánh của một nút không thể vượt quá 3 nên sau đó gốc được phân chia và chiều cao của cây CF tăng (Hình 2. 4).



Hình 2. 4 Gốc được phân chia và chiều cao của cây CF tăng

Ưu điểm của BIRCH:

- Sử dụng cấu trúc cây CF làm cho thuật toán có tốc độ thực hiện phia cụm dữ liệu nhanh và có thể áp dụng đối với tập dữ liệu lớn, đặc biệt áp dụng với dữ liệu tăng trưởng theo thời gian.

- BIRCH chỉ duyệt toàn bộ dữ liệu một lần quét thêm tùy chọn, nghĩa độ phức tạp của nó là O(n) với n là số đối tượng dữ liệu.

Hạn chế của BIRCH:

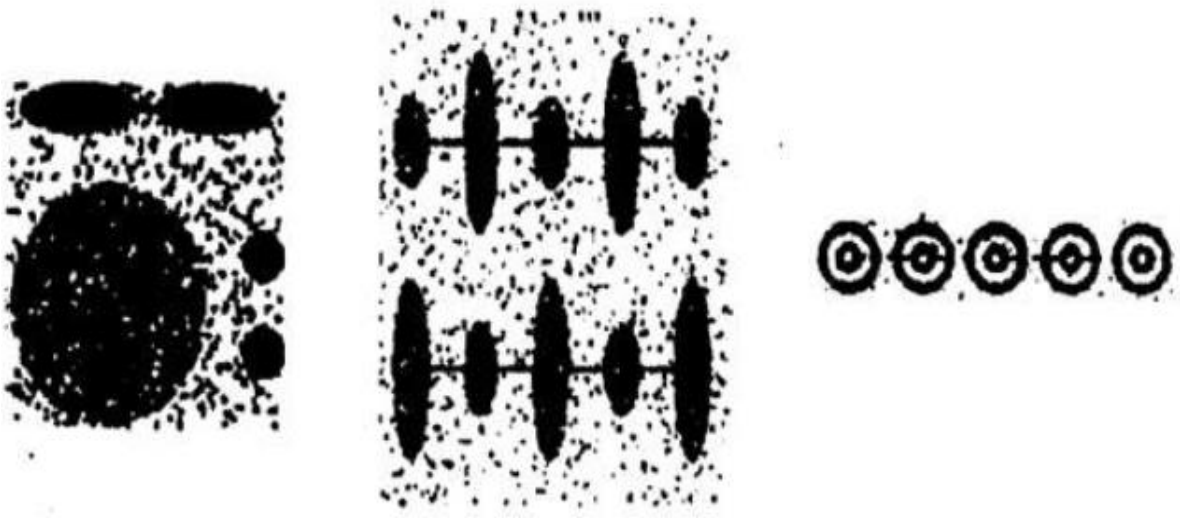
- BIRCH có một hạn chế: Thuật toán này có thể không xử lý tốt nếu các cụm không có hình dạng cầu, bởi vì nó sử dụng khái niệm bán kính hoặc đường kính để kiểm soát ranh giới các cụm và chất lượng của các cụm được khám phá không được tốt.

- BIRCH không thích hợp với dữ liệu đa chiều.

2.1.2 Thuật toán CURE

Thuật toán CURE (Clustering Using Representatives) là thuật toán sử dụng chiến lược dưới lên (Bottom up) của kỹ thuật phân cụm phân cấp. CURE sử dụng nhiều đối tượng để diễn tả cho mỗi cụm dữ liệu. Các đối tượng đại diện cho cụm này ban đầu được lựa chọn rải rác đều ở các vị trí khác nhau, sau đó chúng được di chuyển bằng cách co lại theo một tỉ lệ nhất định. Tại mỗi bước của thuật toán, hai cụm có cặp đối tượng đại diện gần nhất (đối tượng thuộc về mỗi cụm) sẽ được trộn lại thành một cụm.

Việc co lại các cụm có tác dụng làm giảm tác động của các phần tử ngoại lai. Như vậy, thuật toán này có khả năng xử lý tốt trong trường hợp có các phần tử ngoại lại và làm cho hiệu quả với những hình dạng không phải là hình cầu và kích thước độ rộng biến đổi.



Hình 2. 5 Cụm dữ liệu khai phá bởi thuật toán CURE

Các bước cơ bản để thực hiện thuật toán CURE:

Bước 1: Chọn một mẫu ngẫu nhiên từ tập dữ liệu ban đầu.

Bước 2: Phân hoạch mẫu này thành nhiều nhóm dữ liệu có kích thước bằng nhau.

Bước 3: Phân cụm các điểm của mỗi nhóm: Thực hiện phân cụm dữ liệu cho các nhóm cho đến khi mỗi nhóm này được phân thành n'/(pq) cụm (q>1).

Bước 4: Loại bỏ các thành phần ngoại lệ: Đầu tiên, khi một cụm được hình thành Cho đến khi số lượng cụm giảm xuống một phần của số lượng cụm ban đầu. Sau đó, khi các giá trị ngoại lệ được lấy mẫu với giai đoạn khởi tạo mẫu thuật toán tự động xóa các nhóm con khỏi dữ liệu.

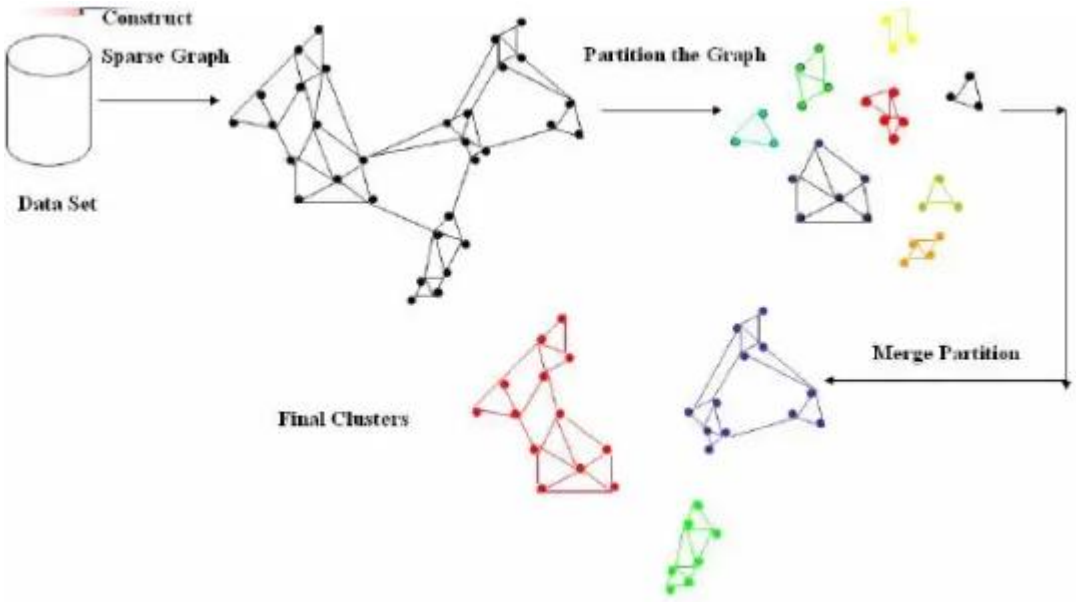
Bước 5: Phân cụm cụm không gian: Các đối tượng đại diện cho các cụm di Di chuyển về phía trung tâm của cụm. Tức là nó sẽ được thay thế bằng một đối tượng gần đó Trung tâm hơn.

Bước 6: Đánh dấu dữ liệu với các nhãn tương ứng.

2.1.3 Thuật toán CHAMELEON

Phương pháp Chameleon một cách tiếp cận khác trong việc phân cụm được phát triển bởi Karypis, Han và Kumar năm 1999, sử dụng mô hình động trong phân cụm phân cấp.

Khi xử lý phân cụm, hai cụm được hợp nhất nếu kết nối và độ chặt chẽ giữa hai cụm có liên quan chặt chẽ đến khả năng kết nối và tính nhỏ gọn vốn có của các đối tượng trong cụm. Các quy trình tích hợp dựa trên mô hình động tạo điều kiện thuận lợi cho việc phát hiện các cụm đồng nhất, tự nhiên. Có giá trị cho tất cả các kiểu dữ liệu miễn là hàm tương tự được chỉ định.



Hình 2. 6 Khái quát thuật toán CHAMELEON

2.2 Thuật toán phân cụm dữ liệu dựa vào mật độ

2.2.1 Thuật toán DBSCAN

Thuật toán DBSCAN (**D**ensity - **B**ased **S**patial **C**lustering of **A**pplications with **N**oise) được Ester, P. Kriegel và J. Sander đề xuất năm 1996.

Một thuật toán đi tìm các đối tượng mà có số đối tượng láng giềng lớn hơn một ngưỡng tối thiểu. Một cụm được xác bởi một tập hợp tất cả các đối tượng được kết nối chặt chẽ với các đối tượng liền kề.

Thuật toán DBSCAN dựa trên khái niệm mật độ có thể được áp dụng cho các tập dữ liệu không gian đa chiều lớn. Dưới đây là một số định nghĩa và bổ đề được sử dụng trong thuật toán DBSCAN.

Các bước cơ bản để thực hiện thuật toán DBSCAN:

Bước 1: Chọn một đối tượng p tuỳ ý.

Bước 2: Lấy tất cả các đối tượng mật độ đến được từ p với Eps và MinPts.

Bước 3: Nếu p là điểm nhân thì tạo ra một cụm theo Eps và MinPts.

Bước 4: Nếu p là một điểm biên, không có điểm nào là mật độ đến được mật độ từ p và DBSCAN sẽ đi thăm điểm tiếp theo của tập dữ liệu.

Bước 5: Quá trình tiếp tục cho đến khi tất cả các đối tượng được xử lý.

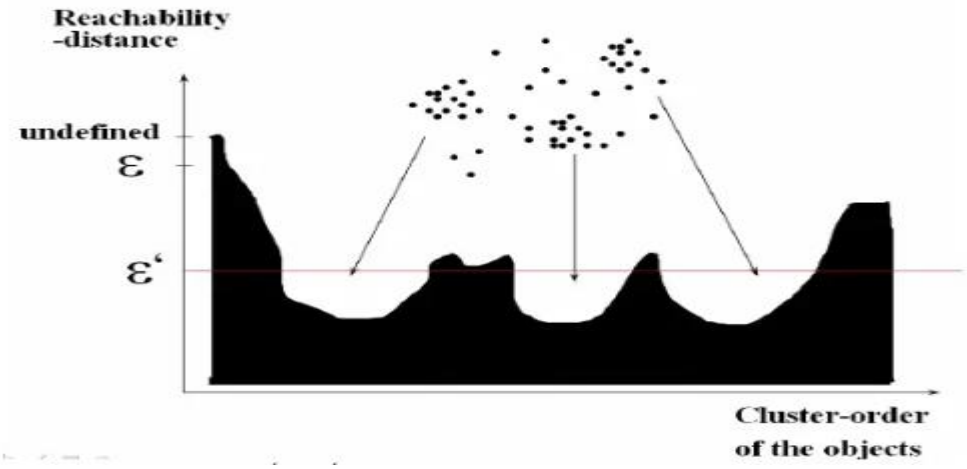
Hình 2. 7 Hình dạng các cụm được khám phá bởi DBSCAN

2.2.2 Thuật toán OPTICS

Thuật toán OPTICS (**O**rdering **P**oints **T**o **I**dentify the **C**lustering **S**tructure) do Ankerst, Breunig, Kriegel và Sander đề xuất năm 1999.

Là thuật toán mở rộng cho thuật toán DBSCAN, bằng cách giảm bớt các tham số đầu vào. Thuật toán thực hiện tính toán và sắp xếp các đối tượng theo thứ tự tăng dần nhằm tự động phân cụm và phân tích cụm tương tác hơn là đưa ra phân cụm một tập dữ liệu rõ ràng.

OPTICS diễn tả cấu trúc dữ liệu phân cụm dựa trên mật độ chứa thông tin tương đương với phân cụm dựa trên mật độ với một dãy các tham số đầu vào. OPTICS xem xét bán kính tối thiểu nhằm xác định các láng giềng phù hợp với thuật toán.



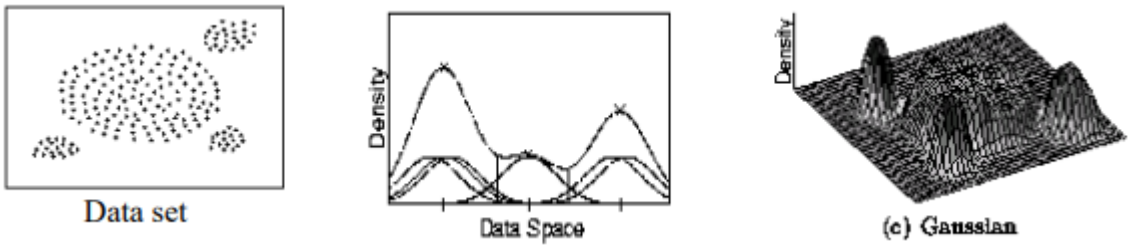
Hình 2. 8 Sắp xếp các cụm trong OPTICS phụ thuộc vào ε

2.2.3 Thuật toán DENCLUE

Là thuật toán phân cụm dữ liệu dựa trên một tập các hàm phân phối mật độ. Thuật toán DENCLUE (DENsity - Based CLUstEring) do Hinneburg và Keim đề xuất vào năm 1998.

Ý tưởng của thuật toán DENCLUE:

- Tác động của một đối tượng đối với các đối tượng lân cận của nó được xác định bởi hàm ảnh hưởng.
- Mật độ tổng thể của không gian dữ liệu được mô hình phân tích như là tổng của tất cả các chức năng hành động của đối tượng.
- Các cụm được xác định bởi các đối tượng mật độ cao trong đó mật độ cao là các điểm cực đại của hàm mật độ toàn cục.



Hình 2. 9 DENCLUE với hàm phân phối Gaussian

2.3 Thuật toán phân cụm dữ liệu dựa vào lưới

2.3.1 Thuật toán STING

STING là kỹ thuật phân cụm đa phân giải dựa trên lưới, trong đó vùng không gian dữ liệu được phân rã thành số hữu hạn các cells chữ nhật, điều này có ý nghĩa là các cells lưới được hình thành từ các cells lưới con để thực hiện phân cụm. C

Có nhiều mức của các cells chữ nhật tương ứng với các mức khác nhau của phân giải trong cấu trúc lưới, và các cells này hình thành cấu trúc phân cấp : mỗi cells ở mức cao được phân hoạch thành các số các cells nhỏ ởmức thấp hơn tiếp theo trong cấu trúc phân cấp.

Các điểm dữ liệu được nạp từ CSDL, giá trị của các tham số thống kê cho các thuộc tính của đối tượng dữ liệu trong mỗi ô lưới được tính toán từ dữ liệu và lưu trữ thông qua các tham số thống kê ở các cell mức thấp hơn (điều này giống với cây CF).

Các giá trị của các tham số thống kê gồm :

- Số trung bình – mean
- Số tối đa – max
- Số tối thiểu – min
- Số đếm –count
- Độ lệch chuẩn –s, …

STING có khả năng mở rộng cao , nhưng do sử dụng phương pháp đa phân giải nên nó phụ thuộc chặt chẽ vào trọng tâm của mức thấp nhất.

Các bước cơ bản để thực hiện thuật toán STING:

Bước 1: Xác định tầng để bắt đầu

Bước 2: Với mỗi cái của tầng này, tính toán khoảng tin cậy (hoặc ước lượngkhoảng) của xác suất mà cells này liên quan tới truy vấn

Bước 3: Từ khoảng tin cậy của tính toán trên,gán nhãn cho là có liên quan hoặc không liên quan.

Bước 4: Nếu lớp này là lớp cuối cùng, chuyển sang Bước 6. Nếu khác thì chuyển sang Bước 5.

Bước 5: Duyệt xuống dưới của cấu trúc cây phân cấp một mức. Chuyển sang Bước 2 cho các cells mà hình thành các cells liên quan của lớp có mức cao hơn.

Bước 6: Nếu đặc tả được câu truy vấn, chuyển sang bước 8. Nếu không thì chuyển sang bước 7.

Bước 7: Truy lục lại dữ liệu vào trong các cells liên quan và thực hiện xử lý. Trả lại kết quả phù hợp yêu cầu của truy vấn. Chuyển sang Bước 9.

Bước 8: Tìm thấy các miền có các cells liên quan. Trả lại miền mà phù hợp với yêu cầu của truy vấn. Chuyển sang bước 9.

Bước 9: Dừng.

2.3.2 Thuật toán CLIQUE

Trong không gian đa chiều, các cụm có thể tồn tại trong tập con của cácchiều hay còn gọi là không gian con. Thuật toán CLIQUE là thuật toán hữu ích cho phân cụm dữ liệu không gian đa chiều trong các phân cụm dữ liệu lớn thành các không gian con.

Thuật toán CLIQUE bao gồm các bước :

- Cho n là tập lớn của các điểm dữ liệu đa chiều, không gian dữ liệu thường là không giống nhau bởi các điểm dữ liệu. Phương pháp này xác định những vùng gần, thưa và đặc, trong không gian dữ liệu nhất định, bằng cách đó phát hiện ra toàn thể phân bố mẫu của tập dữ liệu.

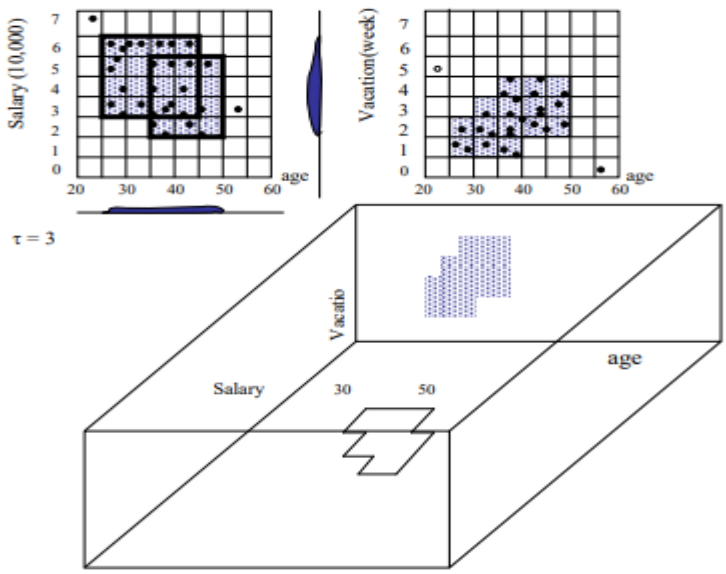
- Một đơn vị là dày đặc nếu phần nhỏ của tất cả các điểm dữ liệu chứa trong nó vượt quá tham số mẫu đưa vào. Trong thuật toán CLIQUE, cụm được định nghĩa là tập tối đa liên thông các đơn vị dày đặc.

Các đặc trưng của CLINQUE:

- Tự động tìm kiếm không gian con của không gian đa chiều, sao chomật độ đặc của các cụm tồn tại trong không gian con.

- Mẫn cảm với thứ tự của dữ liệu vào và không phù hợp với bất kỳ quytắc phân bố dữ liệu nào.

- Phương pháp này tỷ lệ tuyến tính với kích thước vào và có tính biến đổi tốt khi số chiều của dữ liệu tăng.



Hình 2. 10 Quá trình nhận dạng các ô của CLIQUE

CLIQUE có khả năng áp dụng tốt đối với dữ liệu đa chiều, nhưng nó lại rất nhạy cảm với thứ tự của dữ liệu vào, độ phức tạp tính toán của CLIQUE là O(n).

2.4 Thuật toán phân cụm dữ liệu dựa vào mô hình

2.4.1 Thuật toán EM

Thuật toán EM (Expectation - Maximization) được đề xuất vào 1958 bởi Hartley và được nghiên cứu đầy đủ bởi Dempster, Laird và Rubin công bố năm 1977.

Phương pháp tối đa hóa kì vọng (EM) là thuật toán gom nhóm (clustering) dữ liệu (như k-means) được dùng trong tác vụ khám phá tri thức (knowledge discovery).

Trong thống kê, thuật toán EM lặp (iterate) và tối ưu hóa (optimize) khả năng (likelihood) nhìn thấy dữ liệu quan sát (seeing observed data) thông qua việc ước lượng

tham số (parameters estimation) cho mô hình thống kê (statistical model) cho các biến không quan sát được (unobserved variables).

Thuật toán gồm 2 bước xử lý: Đánh giá dữ liệu chưa được gán nhãn (bước E) và đánh giá các tham số của mô hình, khả năng lớn nhất có thể xẩy ra (bước M).

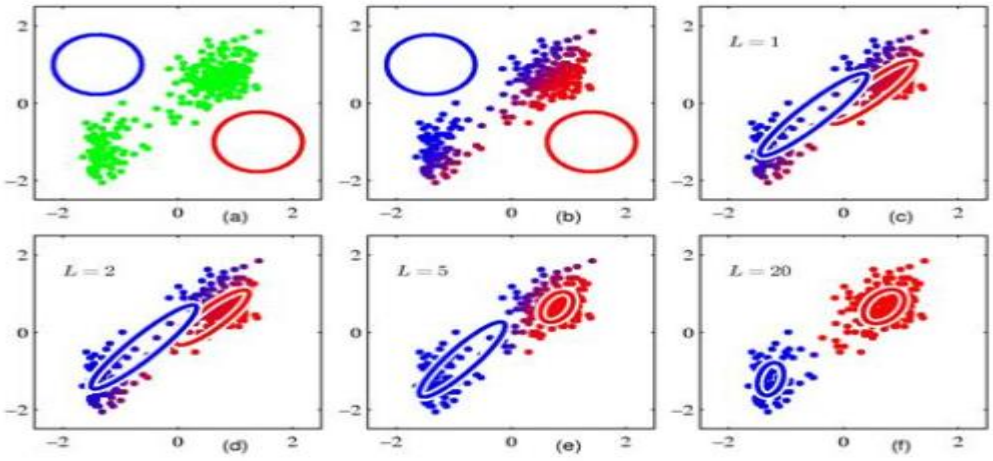
Cụ thể thuật toán EM ở bước lặp thứ t thực hiện các công việc sau:

1. **Bước E**: Tính toán để xác định giá trị của các biến chỉ thị dựa trên mô hình hiện tại và dữ liệu:

$$z_{ij}^{(t)} = E\left(z_{ij}\middle|x\right) = \Pr\left(z_{ij} = 1\middle|x\right) = \frac{f_j(x_i)\pi_j^{(t)}}{\sum_g^k 1 f g^{(x_i)\pi} g}$$

2. **Bước M**: Đánh giá xác xuất π

$$\pi_j^{(t+1)} = \sum_{i=1}^{n} z_{ij}^{(t)}/n$$



Hình 2. 11 Expectation Maximization

2.4.2 Thuật toán COBWEB

COBWEB được đề xuất bởi Fisher năm 1987. Các đối tượng đầu vào của thuật toán được mô tả bởi cặp thuộc tính giá trị, nó thực hiện phân cụm phân cấp bằng cách tạo cây phân lớp, các cấu trúc cây khác nhau.

Thuật toán này sử dụng công cụ đánh giá Heuristic được gọi là công cụ phân loại CU (Category untility) để quản lý cấu trúc cây. Từ đó cấu trúc cây được hình thành dựa trên phép đo độ tương tự mà phân loại tương tự và phi tương tự, cả hai có thể mô tả phân chia giá trị thuộc tính giữa các nút trong lớp

Các bước chính của thuật toán:

Bước 1: Khởi tạo cây bắt đầu bằng một nút rỗng.

Bước 2: Sau khi thêm vào từng nút một và cập nhật lại cây cho phù hợp tại mỗi thời điểm.

Bước 3: Cập nhật cây bắt đầu từ lá bên phải trong mỗi trường hợp, sau đó cấu trúc lại cây.

Bước 4: Quyết định cập nhật dựa trên sự phân hoạch và các hàm tiêu chuẩn phân loại.

2.5 Thuật toán phân cụm dữ liệu dựa vào cụm phân hoạch

2.5.1 Thuật toán PAM

Thuật toán PAM là thuật toán mở rộng của thuật toán K-means nhằm có khả năng xử lý hiệu quả đối với dữ liệu nhiễu hoặc phần tử ngoại lai, PAM sử dụng các đối tượng medoid để biểu diễn cho các cụm dữ liệu, một đối tượng medoid là đối tượng đặt tại vị trí trung tâm nhất bên trong mỗi cụm

Thuật toán PAM được ápd ụng cho dữ liệu không gian. Để xác định các medoid, PAM được áp dụngcho dữ liệu không gian. Để xác định các medoid, PAM bắt đầu bằng cách lựachon k đối tượng medoid bất kỳ. Sau mỗi bước thực hiện , PAM cố gắng hoánchuyển giữa đối tượng Medoid Om và một đối tượng Op, không phải là medoid, miễn là sự hoán chuyển này nhằm cải tiến chất lượng của phân cụm,quá trình này kết thúc khi chất lượng phân cụm không thay đổi.

Các bước cơ bản để thực hiện thuật toán PAM:

Bước 1: Chọn k đối tượng medoid bất kỳ.

Bước 2: Tính TCmp cho tất cả các cặp đối tượng Om, Op. Trong đó, Om là đối tượng medoid và Op là đối tượng không phải medoid.

Bước 3: Chọn cặp đối tượng Om và Op. Tính MinOm, MinOp, TCmp, nếu TCmp là âm thay thế Om bởi Op và quay lại Bước 2. Nếu TCmp dương, chuyển sang Bước 4.

Bước 4: Với mỗi đối tượng không phải medoid, xác định đối tượng medoid tương tự với nó nhất đồng thời gán nhãn cụm cho chúng.

2.5.2 Thuật toán K-MEANS

K-means là thuật toán gom cụm theo phương pháp phân hoạch và đã được sử dụng rộng rãi. Cho tập các đối tượng, mục tiêu gom cụm hay phân mảnh là chia tập đối tượng này thành nhiều nhóm hay "cụm" sao cho các đối tượng trong một cụm có khuynh hướng tương tự nhau hơn so với đối tượng khác nhóm.

Nói cách khác, các thuật toán gom cụm đặt các điểm tương tự trong cùng một cụm trong khi các điểm không tương tự đặt trong nhóm khác.

Thuật toán *k*-means là thuật toán gom cụm lặp đơn giản. Nó phân mảnh tập dữ liệu cho trước thành *k* cụm, giá trị *k* do người dùng xác định. Thuật toán dễ thực hiện, thi hành nhanh, dễ thích nghi và phổ biến trong thực tế. Đây là một trong những thuật toán kinh điển trong khai thác dữ liệu.

Thuật toán k-means áp dụng cho các đối tượng được biểu diễn bởi các điểm trong không gian vectơd chiều U {x | i 1,..., N} i = = , với $di\ x\ \in R$ biểu thị đối tượng (hay điểmdữ liệu) thứi. Thuật toán k-means gom cụm toàn bộ các điểm dữ liệu trong U thành *k* cụm C ={C1, C2, … , C_k},sao cho mỗi điểm dữ liệu xi nằm trong một cụm duy nhất.

Trong các thuật toán gom cụm, các điểm được nhóm theo khái niệm độ gần hay độ tương tự. Với *k*-means, phép đo mặc định cho độ tương tư là khoảng cách Euclide. Đặc biệt, có thể thấy *k*-means cố gắng cực tiểu hóa hàm giá trị không âm sau:

$$Cost = \sum_{j=1}^{n} (argmin_j \| x_i - c_j \|_2^2$$

Thuật toán k-means bao gồm các bước cơ bản như sau:

Đầu vào: Số cụm k và hàm E

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - m_i|^2$$

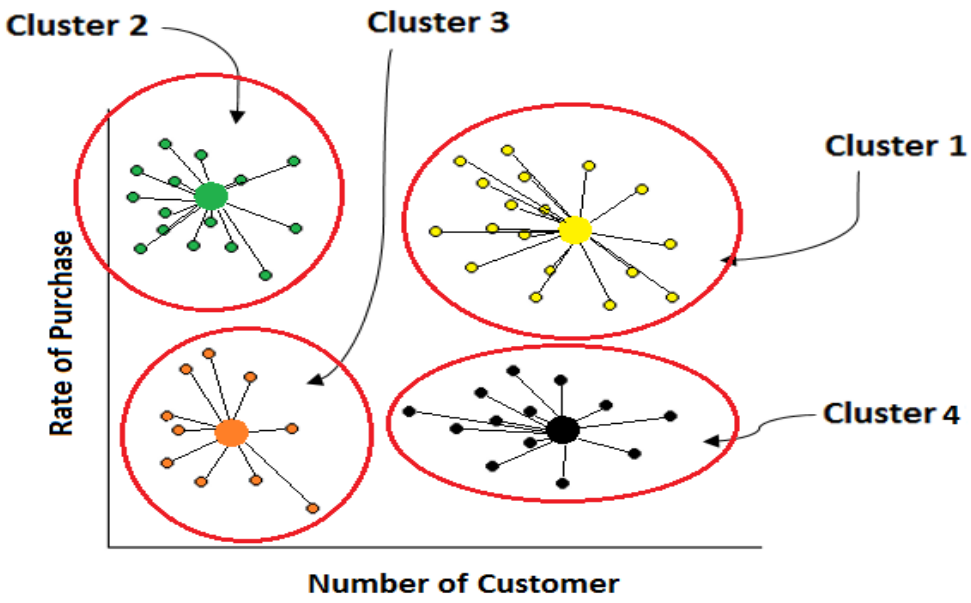
Đầu ra: Các cụm $C[i](1 \leq i \leq k)$ với hàm tiêu chuẩn E đạt giá trị tối thiểu.

Bước 1: Khởi tạo: Chọn k đối tượng mj (j=1...k) là trọng tâm ban đầu của k cụm từ tập dữ liệu (việc lựa chọn này có thể là ngẫu nhiên hoặc theo kinh nghiệm).

Bước 2: Tính toán khoảng cách: Đối với mỗi đối tượng Xi ($1 \leq i \leq n$) , tính toán khoảng cách từ nó tới mỗi trọng tâm mj với j=1,..,k, sau đó tìm trọng tâm gần nhất đối với mỗi đối tượng.

Bước 3: Cập nhật lại trọng tâm: Đối với mỗi j=1,..,k, cập nhật trọng tâm cụm mj bằng cách xác định trung bình cộng của các vector đối tượng dữ liệu.

Bước 4: Điều kiện dừng: Lặp các Bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi.



Hình 2. 12 Ví dụ về K-Means

CHƯƠNG 3 – PHƯƠNG PHÁP XÁC ĐỊNH K PHÙ HỢP TRONG THUẬT TOÁN K-MEANS

3.1 Elbow Method

Phương pháp Elbow là một trong những phương pháp phổ biến nhất để xác định giá trị tối ưu của K.

Sử dụng phương pháp Elbow để tìm số cụm K phù hợp bằng cách trực quan đồ thị. Phương pháp Elbow được sử dụng để xác định số lượng cụm tối ưu trong gom cụm K-Means. Phương pháp Elbow vẽ biểu đồ giá trị của hàm chi phí được tạo ra bởi các giá trị khác nhau của K cụm. Nếu K tăng, độ nghiêng trung bình sẽ giảm, mỗi cụm sẽ có ít cá thể cấu thành hơn và các thể hiện sẽ gần với trung tâm tương ứng của chúng hơn.

Tuy nhiên, sự cải thiện về độ nghiêng trung bình sẽ giảm khi K tăng. Giá trị của K mà tại đó sự cải thiện về độ nghiêng giảm nhiều nhất được gọi là giá trị khuỷu, tại đó, chúng ta nên dừng việc chia dữ liệu thành các cụm xa hơm

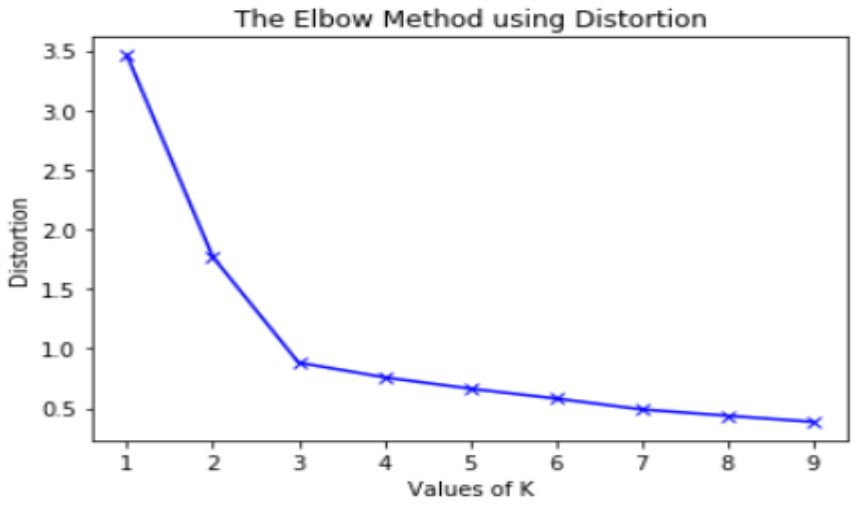
Công thức: $$\boldsymbol{WCSS = \sum_{P_i \text{ in Cluster } K} distance(Pi\ C_K)^2}$$

Nó thực hiện phân cụm K-means trên một tập dữ liệu nhất định cho các giá trị K khác nhau (phạm vi từ 1-10).

Đối với mỗi giá trị của K, hãy tính giá trị WCSS.

Vẽ đồ thị một đường cong giữa các giá trị WCSS được tính toán và số lượng các cụm K.

Điểm uốn cong hoặc một điểm của đồ thị có dạng như một khủyu tay thì điểm đó được coi là điểm tốt nhất của K.



Hình 3. 1 Đồ thị kết quả hệ số k cụm từ thực nghiệm dữ liệu với phương pháp Elbow

3.2 SILHOUETTE

Phương pháp này có thể được sử dụng để kiểm tra chất lượng phân cụm bằng cách đo khoảng cách giữa các cụm. Về cơ bản, nó cung cấp một cách để đánh giá các thông số như số lượng cụm bằng cách cho điểm Silhouette. Điểm số này là một số liệu đo lường mức độ gần của mỗi điểm trong một cụm với các điểm trong các cụm lân cận.

Phân tích điểm Silhouette:

Điểm có phạm vi là [-1, 1] như sau:

- **Điểm +1**: Điểm gần +1 cho biết mẫu ở xa cụm lân cận.
- **Điểm 0**: Điểm 0 cho biết mẫu nằm trên hoặc rất gần ranh giới quyết định giữa hai cụm lân cận.
- **Điểm -1**: Điểm âm chỉ ra rằng các mẫu đã được chỉ định vào các cụm sai.

Tính toán điểm Silhouette:

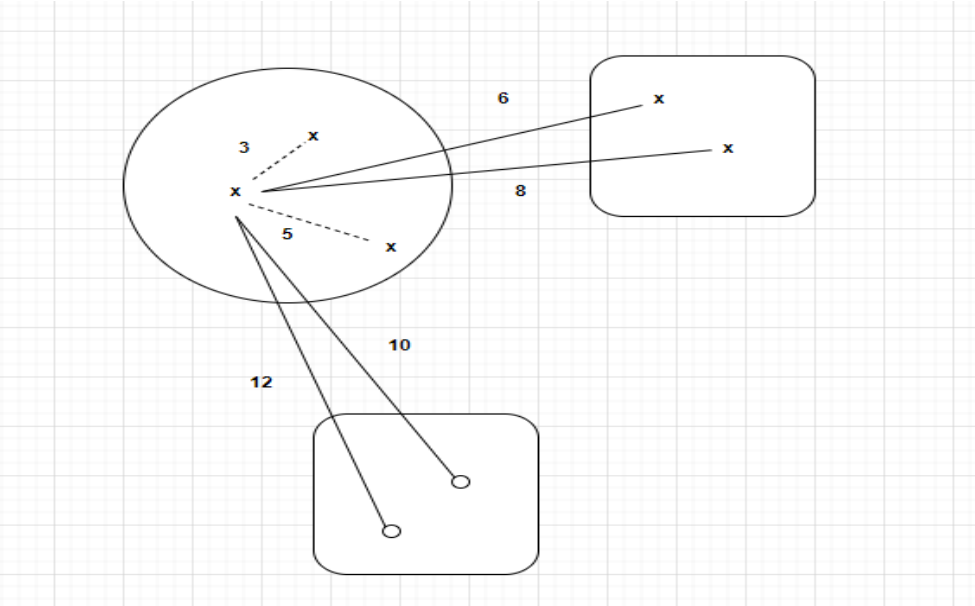
Hệ số Silhouette được xác định cho mỗi cụm và bao gồm hai điểm:

Ở đây **a** là khoảng cách trung bình giữa một mẫu và tất cả các điểm khác trong cùng một cụm.

Còn **b** là khoảng cách trung bình giữa mẫu và tất cả các điểm khác trong cụm gần nhất tiếp theo.

Điểm Silhouette có thể được tính bằng cách sử dụng công thức sau:

$$Silhouette = \frac{(b - a)}{\max(a, b)}$$



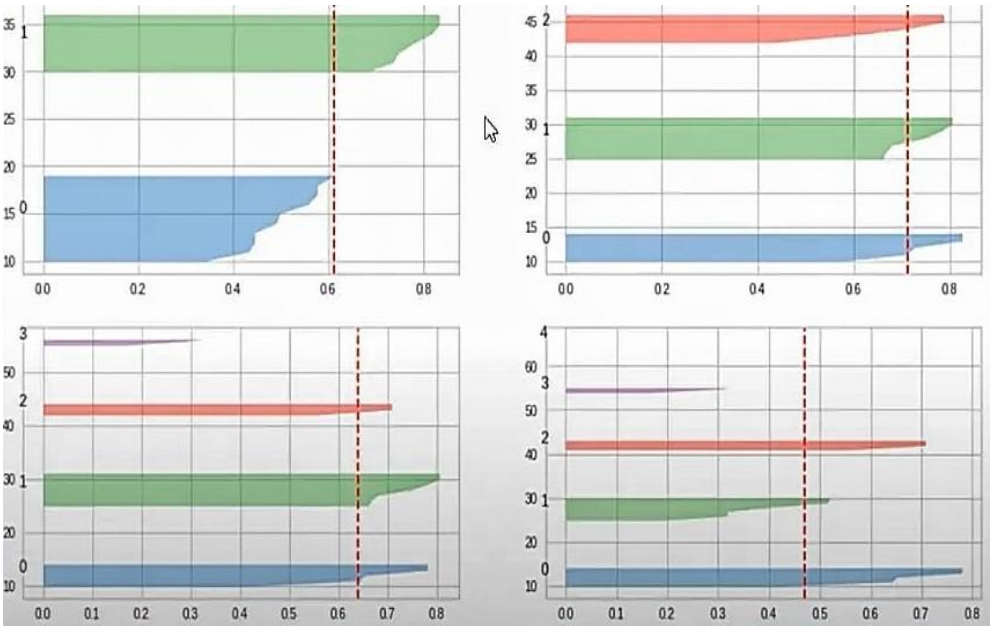
Hình 3. 2 Ví dụ về Silhouette

$$a = \frac{3+5}{2} = 4$$

$$b = \min\left(\frac{8+6}{2}; \frac{12+10}{2}\right) = 7$$

Kết luận:

$$\boldsymbol{S = \frac{7-4}{7} = \frac{3}{7}}$$



Hình 3. 3 Biểu đồ mức độ phân bố dữ liệu

Ưu điểm:

- Đơn giản, dễ hiểu, tương đối hiệu quả.
- Các đối tượng tự động gán vào nhóm.

Nhược điểm:

- Cần xác định số nhóm K trước
- Gặp vấn đề khi các nhóm có cùng kích thước, mật độ khác nhau.

TÀI LIỆU THAM KHẢO

1. Bài 4: K-means Clustering (2017). Available at: https://machinelearningcoban.com/2017/01/01/kmeans/ (Accessed: 18 April 2022).

2. "Clustering Algorithms: From Start To State Of The Art". 2022. Toptal Engineering Blog. https://www.toptal.com/machine-learning/clustering-algorithms.

3. Nguyen, Quy. 2021. "Thuật Toán Phân Cụm K-Means". Quy'S Blog. https://ndquy.github.io/posts/thuat-toan-phan-cum-kmeans/.

4. "Clustering In Machine Learning - Geeksforgeeks". 2018. Geeksforgeeks. https://www.geeksforgeeks.org/clustering-in-machine-learning/.

5. "8 Clustering Algorithms In Machine Learning That All Data Scientists Should Know". 2020. Freecodecamp.Org. https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/.

6. "Clustering With Machine Learning — A Comprehensive Guide | Rocketloop". 2022. Rocketloop. https://rocketloop.de/en/blog/clustering-machine-learning-comprehensive-guide/.