MỤC LỤC

DANH MỤC HÌNH MINH HỌA

LỜI CẢM ƠN

Trước hết em xin chân thành cảm ơn thầy Ngô Trường Giang là giáo viên hướng dẫn em trong quá tình làm đồ án. Thầy đã giúp em rất nhiều và đã cung cấp cho em nhiều tài liệu quan trọng phục vụ cho quá trình tìm hiểu về đề tài "Tìm hiểu một số phương pháp phân cụm dữ liệu và ứng dụng".

Thứ hai, em xin chân thành cảm ơn các thầy cô trong bộ môn công nghệ thông tin đã chỉ bảo em trong quá trình học và rèn luyện trong 4 năm học vừa qua. Đồng thời em cảm ơn các bạn sinh viên lớp CT1002 đã gắn bó với em trong quá trình rèn luyện tại trường.

Cuối cùng em xin chân thành cảm ơn ban giám hiệu trường Đại Học Dân Lập Hải Phòng đã tạo điều kiện cho em có kiến thức, thư viện của trường là nơi mà sinh viên trong trường có thể thu thập tài liệu trợ giúp cho bài giảng trên lớp. Đồng thời các thầy cô trong trường giảng dạy cho sinh viên kinh nghiệm cuộc sống. Với kiến thức và kinh nghiệm đó sẽ giúp cho em trong công việc và cuộc sống sau này.

Em xin chân thành cảm ơn!

Hải Phòng, ngày tháng năm 2010

Sinh viên

VŨ MINH ĐÔNG

CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

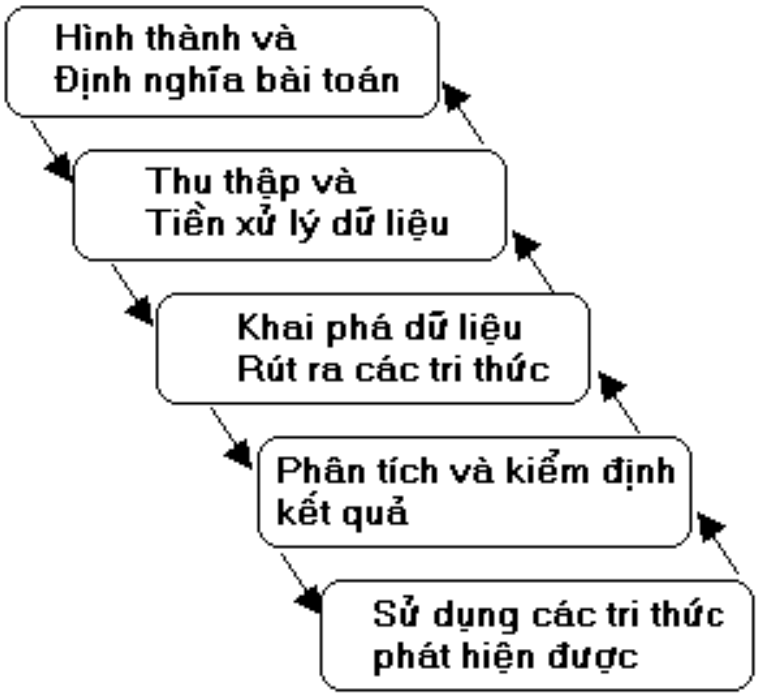
1.1 Giới thiệu về khám phá tri thức

Nếu cho rằng các điện tử và các sóng điện tử chính là bản chất của công nghệ điện tử truyền thống thì dữ liệu, thông tin và tri thức hiện đang là tiêu điểm của một lĩnh vực mới trong nghiên cứu và ứng dụng về phát hiện tri thức (Knowledge Discovery) và khai phá dữ liệu (Data Mining).

Thông thường chúng ta coi dữ liệu như một dãy các bit, hoặc các số và các ký hiệu, hoặc các "đối tượng" với một ý nghĩa nào đó khi được gửi cho một chương trình dưới một dạng nhất định. Chúng ta sử dụng các bit để đo lường các thông tin và xem nó như là các dữ liệu đã được lọc bỏ các dư thừa, được rút gọn tới mức tối thiểu để đặc trưng một cách cơ bản cho dữ liệu. Chúng ta có thể xem tri thức như là các thông tin tích hợp, bao gồm các sự kiện và các mối quan hệ giữa chúng. Các mối quan hệ này có thể được hiểu ra, có thể được phát hiện, hoặc có thể được học. Nói cách khác, tri thức có thể được coi là dữ liệu có độ trừu tượng và tổ chức cao.

Phát hiện tri thức trong các cơ sở dữ liệu là một qui trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: hợp thức, mới, khả ích, và có thể hiểu được. Còn khai thác dữ liệu là một bước trong qui trình phát hiện tri thức gồm có các thuật toán khai thác dữ liệu chuyên dùng dưới một số qui định về hiệu quả tính toán chấp nhận được để tìm ra các mẫu hoặc các mô hình trong dữ liệu. Nói một cách khác, mục đích của phát hiện tri thức và khai phá dữ liệu chính là tìm ra các mẫu hoặc các mô hình đang tồn tại trong các cơ sở dữ liệu nhưng vẫn còn bị che khuất bởi hàng núi dữ liệu.

Quy trình phát hiện tri thức:



Hình 1. 1: Quy trình phát hiện tri thức

Bước thứ nhất: là tìm hiểu lĩnh vực ứng dụng và hình thành bài toán, bước này sẽ quyết định cho việc rút ra được các tri thức hữu ích và cho phép chọn các phương pháp khai phá dữ liệu thích hợp với mục đích ứng dụng và bản chất của dữ liệu.

Bước thứ hai: là thu thập và xử lý thô, còn được gọi là tiền xử lý dữ liệu nhằm loại bỏ nhiễu, xử lý việc thiếu dữ liệu, biến đổi dữ liệu và rút gọn dữ liệu nếu cần thiết, bước này thường chiếm nhiều thời gian nhất trong toàn bộ qui trình phát hiện tri thức.

Bước thứ ba: là khai phá dữ liệu, hay nói cách khác là trích ra các mẫu hoặc các mô hình ẩn dưới các dữ liệu.

Bước thứ tư: là hiểu tri thức đã tìm được, đặc biệt là làm sáng tỏ các mô tả và dự đoán. Các bước trên có thể lặp đi lặp lại một số lần, kết quả thu được có thể được lấy trung bình trên tất cả các lần thực hiện.

1.2 Khai phá dữ liệu và các khái niệm liên quan

Khai phá dữ liệu như là một qui trình phân tích được thiết kế để thăm dò một lượng cực lớn các dữ liệu nhằm phát hiện ra các mẫu thích hợp hoặc các mối quan hệ mang tính hệ thống giữa các biến và sau đó sẽ hợp thức hoá các kết quả tìm được bằng cách áp dụng các mẫu đã phát hiện được cho các tập con mới của dữ liệu. Qui trình này bao gồm ba giai đoạn cơ bản: thăm dò, xây dựng mô hình hoặc định nghĩa mẫu, hợp thức, kiểm chứng.

1.2.1 Khái niệm khai phá dữ liệu

Do sự phát triển mạnh mẽ của khai phá dữ liệu (Data mining) về phạm vi các lĩnh vực ứng dụng trong thực tế và các phương pháp tìm kiếm, lên có rất nhiều các khái niệm khác nhau về khai phá dữ liệu. Trong bài này em xin nêu ra một định nghĩa ngắn gọn như sau:

Khai phá dữ liệu là quá trình khám phá các tri thức mới và các tri thức có ích ở dạng tiềm năng trong nguồn dữ liệu đã có.

1.2.2 Các phương pháp khai phá dữ liệu

Với hai đích chính của khai phá dữ liệu là: dự đoán (Prediction) và mô tả (Description), người ta thường sử dụng các phương pháp sau cho khai phá dữ liệu:

- Phân lớp (Classfication)
- Hồi qui (*Regression*)
- Trực quan hóa (*Visualiztion*)
- Phân cụm (*Clustering*)
- Tổng hợp (Summarization)
- Mô hình ràng buộc (*Dependency modeling*)
- Biểu diễn mô hình (*Model Evaluation*)
- Phân tích sự phát triển và độ lệch *(Evolution and deviation analyst*)
- Luận kết hợp (Associantion rules)
- Phương pháp tìm kiếm (*Search Method*)

1.2.3 Các lĩnh vực ứng dụng trong thực tiễn

- Phân tích dữ liệu và hỗ trợ ra quyết định.
- Phân lớp văn bản, tóm tắt văn bản, phân lớp các trang Web và phân cụm ảnh màu.
- Chuẩn đoán triệu chứng, phương pháp trong điều trị y học.
- Tìm kiếm, đối sánh các hệ Gene và thông tin di truyền trong sinh học.
- Phân tích tình hình tài chính, thị trường, dự báo giá cổ phiếu trong tài chính, thị trường và chứng khoán.
- Bảo hiểm …

1.2.4 Các hướng tiếp cận cơ bản và kỹ thuật áp dụng trong khai phá dữ liệu

Các kỹ thuật khai phá dữ liệu thường được chia thành 2 nhóm chính:

- Kỹ thuật khai phá dữ liệu mô tả: có nhiệm vụ mô tả về các tính chất hoặc các đặc tính chung của dữ liệu trong CSDL hiện có. Các kỹ thuật này gồm có: phân cụm (Clustering), tổng hợp (Summerization), trực quan hóa (Visualiztion), phân tích sự phát triển và độ lệch (Evolution and deviation analyst), luận kết hợp (Associantion rules)
- Kỹ thuật khai phá dữ liệu dự đoán: có nhiệm vụ đưa ra các dự đoán vào các suy diễn trên dữ liệu hiện thời. Các kỹ thuật này gồm có: phân lớp (Classification), hồi quy (Regression). . .

Sau đây em xin được giới thiệu 3 phương pháp thông dụng nhất là: phân cụm dữ liệu, phân lớp dữ liệu và khai phá luận kết hợp.

- *Phân lớp dữ liệu*: Mục tiêu của phương pháp phân lớp dữ liệu là dự đoán nhãn lớp cho các mẫu dữ liệu. Quá trình phân lớp dữ liệu thường gồm 2 bước: xây dựng mô hình và sử dụng mô hình để phân lớp dữ liệu.

Bước 1: một mô hình sẽ được xây dựng dựa trên việc phân tích các mẫu dữ liệu sẵn có. Mỗi mẫu tương ứng với một lớp, được quyết định bởi một thuộc tính gọi là thuộc tính lớp. Các mẫu dữ liệu này còn được gọi là tập dữ liệu huấn luyện (Training dataset). Các nhãn lớp của tập dữ liệu huấn luyện đều phải được xác định trước khi xây dựng mô hình vì vậy phương pháp này còn được gọi là học có thầy (Supervised learning) khác với phân cụm dữ liệu là học không có thầy (Unsupervised learning).

Bước 2: sử dụng mô hình để phân lớp dữ liệu. Trước hết chúng ta phải tính độ chính xác của mô hình. Nếu độ chính xác là chấp nhận được, mô hình sẽ được sử dụng để dự đoán nhãn lớp cho các mẫu dữ liệu khác trong tương lai.

- *Phân cụm dữ liệu*: Mục tiêu chính của phân cụm dữ liệu là nhóm các đối tượng tương tự nhau trong tập dữ liệu vào các cụm sao cho các đối tượng thuộc cùng một lớp là tương đồng còn các đối tượng thuộc các cụm khác nhau sẽ không tương đồng. Trong phương pháp này bạn sẽ không thể biết kết quả các cụm thu được sẽ như thế nào khi bắt đầu quá trình. Vì vậy, thông thường cần có một chuyên gia về lĩnh vực đó để đánh giá các cụm thu được. Phân cụm dữ liệu còn là bước tiền xử lý cho các thuật toán khai phá dữ liệu khác.
- *Khai phá luận kết hợp*: Mục tiêu của phương pháp này là phát hiện đưa ra các mối liên hệ giữa các giá trị dữ liệu trong CSDL. Mẫu đầu ra của giải thuật khai phá dữ liệu là tập luận kết hợp tìm được.

CHƯƠNG 2: PHÂN CỤM DỮ LIỆU VÀ CÁC TIẾP CẬN

2.1 Khái niệm chung

Khai phá dữ liệu (Datamining) là quá trình trích xuất các thông tin có giá trị tiềm ẩn bên trong tập dữ liệu lớn được lưu trữ trong các cơ sở dữ liệu, kho dữ liệu. Người ta định nghĩa [1]:

"Phân cụm dữ liệu là một kỹ thuật trong Data Mining, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn, quan trọng trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho việc ra quyết định "

Như vậy phân cụm dữ liệu là quá trình chia một tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong một cụm "tương tự" (Similar) với nhau và các phần tử trong các cụm khác nhau sẽ "phi tương tự" (Dissimilar) với nhau. Số các cụm dữ liệu được phân ở đây có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định.

2.2 Các kiểu dữ liệu và độ đo tương tự

2.2.1 Các kiểu dữ liệu

Cho một một cơ sở dữ liệu D chứa n đối tượng trong không gian k chiều trong đó x, y, z là các đối tượng thuộc D: $x = (x_1, x_2, \ldots, x_k)$; $y = (y_1, y_2, \ldots, y_k)$; $z = (z_1, z_2, \ldots, z_k)$, trong đó x_i, y_i, z_i với $i = \overline{1,k}$ là các đặc trưng hoặc các thuộc tính tương ứng của các đối tượng x, y, z.

a) Phân loại theo kích thước miền

- Thuộc tính liên tục (*Continnuous Attribute*): nếu miền giá trị của nó là vô hạn không đếm được.
- Thuộc tính rời rạc (*DiscretteAttribute*): nếu miền giá trị của nó là tập hữu hạn, đếm được.

- Lớp các thuộc tính nhị phân: là trường hợp đặc biệt của thuộc tính rời rạc mà miền giá trị của nó chỉ có hai phần tử được diễn tả như: *Yes / No* hoặc *False / True*, …

b) Phân loại dựa theo hệ đo

Giả sử rằng chúng ta có hai đối tượng x, y và các thuộc tính x_i, y_i tương ứng với thuộc tính thứ i của chúng. Chúng ta có các lớp kiểu dữ liệu như sau:

- Thuộc tính định danh (*Nominal Scale*): đây là dạng thuộc tính khái quát hóa của thuộc tính nhị phân, trong đó miền giá trị là rời rạc không phân biệt thứ tự và có nhiều hơn hai phần tử: nghĩa là nếu x và y là hai đối tượng thuộc tính thì chỉ có thể xác định là x # y hoặc x = y.
- Thuộc tính có thứ tự (*Ordinal Scale*): là thuộc tính định danh có thêm tính thứ tự, nhưng chúng không được định lượng. Nếu x và y là hai thuộc tính thứ tự thì ta có thể xác định là x # y hoặc x = y hoặc x > y hoặc x <y
- Thuộc tính khoảng (*Interval Scale*): Với thuộc tính khoảng, chúng ta có thể xác định một thuộc tính là đứng trước hoặc đứng sau thuộc tính khác với một khoảng là bao nhiêu. Nếu $x_i > y_i$ thì ta nói x cách y một khoảng $x_i - y_i$ tương ứng với thuộc tính thứ i.
- Thuộc tính tỉ lệ (*Ratio Scale*): là thuộc tính khoảng nhưng được xác định một cách tương đối so với điểm mốc, thí dụ như thuộc tính chiều cao hoặc cân nặng lấy điểm 0 làm mốc.

Trong các thuộc tính dữ liệu trình bày ở trên, thuộc tính định danh và thuộc tính có thứ tự gọi chung là thuộc tính hạng mục (Categorical), thuộc tính khoảng và thuộc tính tỉ lệ được gọi là thuộc tính số (Numeric).

2.2.2 Độ đo tương tự và phi tương tự

Để phân cụm, người ta phải đi tìm cách thích hợp để xác định "khoảng cách" giữa các đối tượng, hay là phép đo tương tự dữ liệu. Đây là các hàm để đo sự giống nhau giữa các cặp đối tượng dữ liệu, thông thường các hàm này hoặc là để tính độ tương tự (*Similar*) hoặc là tính độ phi tương tự (*Dissimilar*) giữa các đối tượng dữ liệu.

Tất cả các độ đo dưới đây được xác định trong không gian metric. Một không gian metric là một tập trong đó có xác định các "khoảng cách" giữa từng cặp phần tử, với những tính chất thông thường của khoảng cách hình học. Nghĩa là, một tập X (các phần tử của nó có thể là những đối tượng bất kỳ) các đối tượng dữ liệu trong cơ sở dữ liệu D như đã đề cập ở trên được gọi là một không gian metric nếu:

- Với mỗi cặp phần tử x, y thuộc X đều có xác định, theo một quy tắc nào đó, một số thực δ(x, y), được gọi là khoảng cách giữa x và y.
- Quy tắc nói trên thoả mãn hệ tính chất sau: *(i)* δ(x, y) > 0 nếu x ≠ y ; (*ii)* δ(x, y)=0 nếu x =y; *(iii)* δ(x, y) = δ(y, x) với mọi x, y; *(iv)* δ(x, y) ≤ δ(x, z)+δ(z, y).

Hàm δ(x, y) được gọi là một metric của không gian. Các phần tử của X được gọi là các điểm của không gian này.

♦ **Thuộc tính khoảng:**

Sau khi chuẩn hóa, độ đo phi tương tự của hai đối tượng dữ liệu x, y được xác định bằng các matrix khoảng cách như sau:

- Khoảng cách *Minskowski*: d(x, y) = $(\sum_{i=1}^{n} |x_i - y_i|^q)^{1/q}$ trong đó q là số tự nhiên dương.

- Khoảng cách *Euclide*: d(x, y) = $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$, đây là trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp q =2.

- Khoảng cách *Mahattan*: d(x, y) = $\sum_{i=1}^{n} |x_i - y_i|$, đây là trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp q =1.

- Khoảng cách cực đại: d(x, y) = $Max_{i=1}^{n} |x_i - y_i|$, đây là trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp $q \to \infty$.

♦ **Thuộc tính nhị phân:**

- α là tổng số các thuộc tính có giá trị là 1 trong x, y.
- β là tổng số các thuộc tính có giá trị là 1 trong x và 0 trong y.
- γ là tổng số các thuộc tính có giá trị là 0 trong x và 1 trong y.
- δ là tổng số các thuộc tính có giá trị là 0 trong x và y.
- τ = α + β + γ + δ.

Các phép đo độ tương đồng đối với dữ liệu thuộc tính nhị phân được định nghĩa như sau:

Hệ số đối sánh đơn giản: d(x, y) = $\frac{\alpha + \delta}{\tau}$, ở đây cả hai đối tượng x và y có vai trò như nhau, nghĩa là chúng đối xứng và có cùng trọng số.

Hệ số Jacard: d(x, y) = $\frac{\alpha}{\alpha + \beta + \gamma}$, (bỏ qua số các đối sánh giữa 0-0). Công thức tính này được sử dụng trong trường hợp mà trọng số của các thuộc tính có giá trị 1 của đối tượng dữ liệu có cao hơn nhiều so với các thuộc tính có giá trị 0, như vậy các thuộc tính nhị phân ở đây là không đối xứng.

♦ **Thuộc tính định danh:**

Độ đo phi tương tự giữa hai đối tượng x và y được định nghĩa như sau:

$$d(x, y) = \frac{p - m}{p}$$

Trong đó m là số thuộc tính đối sánh tương ứng trùng nhau và p là tổng số các thuộc tính.

♦ **Thuộc tính có thứ tự:**

Giả sử i là thuộc tính thứ tự có M_i giá trị (M_i kích thước miền giá trị):

Các trạng thái M_i được sắp thứ tự như sau: [1… M_i], chúng ta có thể thay thế mỗi giá trị của thuộc tính bằng giá trị cùng loại r_i với $r_i \in \{1\ldots M_i\}$. Mỗi một thuộc tính có thứ tự có các miền giá trị khác nhau, vì vậy chúng ta chuyển đổi chúng về cùng miền giá trị [0, 1] bằng cách thực hiện phép biến đổi sau cho mỗi thuộc tính:

$$Z_i^{(i)} = \frac{r_i^{(i)} - 1}{M_i - 1}$$

Sử dụng công thức tính độ phi tương tự của các thuộc tính khoảng đối với các giá trị $Z_i^{(i)}$, đây cũng chính là độ phi tương tự của thuộc tính có thứ tự.

♦ **Thuộc tính tỉ lệ:**

Có nhiều cách khác nhau để tính độ tương tự giữa các thuộc tính tỉ lệ. Một trong những số đó là sử dụng công thức tính logarit cho mỗi thuộc tính. Hoặc loại bỏ đơn vị đo của các thuộc tính dữ liệu bằng cách chuẩn hóa chúng, hoặc gán trọng số cho mỗi thuộc tính giá trị trung bình độ lệch chuẩn. Với mỗi thuộc tính dữ liệu đã được gán trọng số tương ứng w_i ($1 \le i \le k$), độ tương đồng dữ liệu được xác định như sau:

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

2.3 Các kỹ thuật tiếp cận trong phân cụm dữ liệu

Các kỹ thuật phân cụm có rất nhiều cách tiếp cận và ứng dụng trong thực tế, nó hướng tới hai mục tiêu chung đó là chất lượng của các cụm khám phá được và tốc độ thực hiện của thuật toán. Hiện nay, các kỹ thuật phân cụm có thể phân loại theo các cách tiếp cận chính sau.

2.3.1 Phương pháp phân cụm phân hoạch

Phương pháp phân cụm phân hoạch nhằm phân một tập dữ liệu có n phần tử cho trước thành k nhóm dữ liệu sao cho: mỗi phần tử dữ liệu chỉ thuộc về một nhóm dữ liệu và mỗi nhóm dữ liệu có tối thiểu ít nhất một phần tử dữ liệu. Các thuật toán phân hoạch dữ liệu có độ phức tạp rất lớn khi xác định nghiệm tối ưu toàn cục cho vấn đề PCDL, do nó phải tìm kiếm tất cả các cách phân hoạch có thể được. Chính vì vậy, trên thực tế người ta thường đi tìm giải pháp tối ưu cục bộ cho vấn đề này bằng cách sử dụng một hàm tiêu chuẩn để đánh giá chất lượng của các cụm cũng như để hướng dẫn cho quá trình tìm kiếm phân hoạch dữ liệu. Với chiến lược này, thông thường người ta bắt đầu khởi tạo một phân hoạch ban đầu cho tập dữ liệu theo phép ngẫu nhiên hoặc theo heuristic và liên tục tinh chỉnh nó cho đến khi thu được một phân hoạch mong muốn, thoả mãn ràng buộc cho trước. Các thuật toán phân cụm phân hoạch cố gắng cải tiến tiêu chuẩn phân cụm, bằng cách tính các giá trị đo độ tương tự giữa các đối tượng dữ liệu và sắp xếp các giá trị này, sau đó thuật toán lựa chọn một giá trị trong dãy sắp xếp sao cho hàm tiêu chuẩn đạt giá trị tối thiểu. Như vậy, 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. Một số thuật toán phân cụm phân hoạch điển hình như K-means, PAM, CLARA, CLARANS, …sẽ được trình bày chi tiết ở chương sau.

2.3.2 Phương pháp phân cụm phân cấp

Phương pháp này xây dựng một phân cấp trên cơ sở các đối tượng dữ liệu đang xem xét. Nghĩa là sắp xếp một tập dữ liệu đã cho thành một cấu trúc

có dạng hình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy có hai cách tiếp cận phổ biến của kỹ thuật này đó là:

- Hòa nhập nhóm: thường được gọi là tiếp cận Bottom-Up: Phương pháp này bắt đầu với mỗi đối tượng được khởi tạo tương ứng với các cụm riêng biệt, sau đó tiến hành nhóm các đối tượng theo một độ đo tương tự (như khoảng cách giữa hai trung tâm của hai nhóm), quá trình này được thực hiện cho đến khi tất cả các nhóm được hòa nhập vào một nhóm (mức cao nhất của cây phân cấp) hoặc cho đến khi các điều kiện kết thúc thỏa mãn. Như vậy, cách tiếp cận này sử dụng chiến lược ăn tham trong quá trình phân cụm.
- Phân chia nhóm: thường được gọi là tiếp cận 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.

Một số thuật toán phân cụm phân cấp điển hình như CURE, BIRCH, …sẽ được trình bày chi tiết ở trong chương sau.

Thực tế áp dụng, có nhiều trường hợp người ta kết hợp cả hai phương pháp phân cụm phân hoạch và phương phân cụm phân cấp, nghĩa là kết quả thu được của phương pháp phân cấp có thể cải tiến thông quan bước phân cụm phân hoạch. Phân cụm phân hoạch và phân cụm phân cấp là hai phương pháp PCDL cổ điển, hiện nay đã có nhiều thuật toán cải tiến dựa trên hai phương pháp này đã được áp dụng phổ biến trong khai phá dữ liệu.

2.3.3 Phương pháp phân cụm dựa trên mật độ

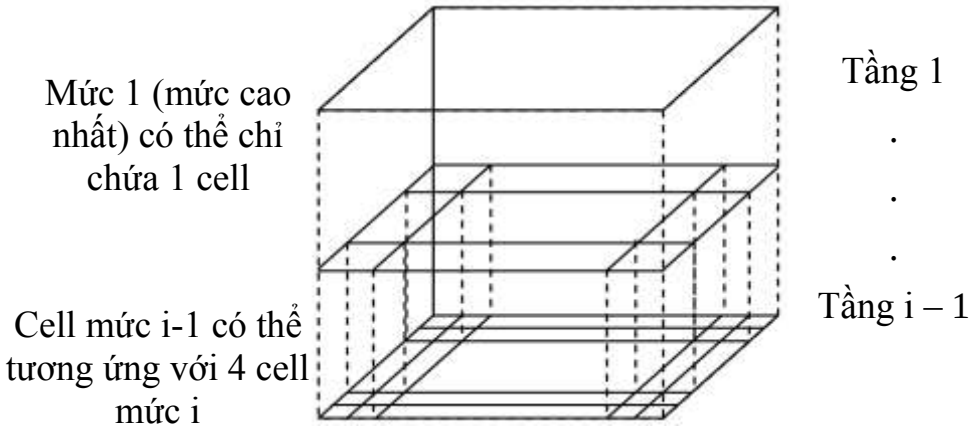
Phương pháp này nhóm các đối tượng theo hàm mật độ xác định. Mật độ được định nghĩa như là số các đối tượng lân cận của một đối tượng dữ liệu theo một ngưỡng nào đó. Trong cách tiếp cận này, khi một cụm dữ liệu đã

xác định thì nó tiếp tục được phát triển thêm các đối tượng dữ liệu mới miễn là số các đối tượng lân cận của các đối tượng này phải lớn hơn một ngưỡng đã được xác định trước. Phương pháp phân cụm dựa vào mật độ của các đối tượng để xác định các cụm dữ liệu có thể phát hiện ra các cụm dữ liệu với hình thù bất kỳ. Kỹ thuật này có thể khắc phục được các phân tử ngoại lai hoặc giá trị nhiễu rất tốt, tuy vậy việc xác định các tham số mật độ của thuật toán rất khó khăn, trong khi các tham số này lại có tác động rất lớn đến kết quả phân cụm dữ liệu. Một số thuật toán PCDL dựa trên mật độ điển hình như DBSCAN, OPTICS, . . . sẽ được trình bày chi tiết trong chương tiếp theo.

2.3.4 Phương pháp phân cụm dựa trên lưới

Kỹ thuật phân cụm dựa trên mật độ không thích hợp với dữ liệu nhiều chiều, để giải quyết cho đòi hỏi này, người ta đã dử dụng phương pháp phân cụm dựa trên lưới. Đây là phương pháp dựa trên cấu trúc dữ liệu lưới để PCDL, phương pháp này chủ yếu tập trung áp dụng cho lớp dữ liệu không gian. Thí dụ như dữ liệu được biểu diễn dưới dạng cấu trúc hình học của đối tượng trong không gian cùng với các quan hệ, các thuộc tính, các hoạt động của chúng. Mục tiêu của phương pháp này là lượng hoá tập dữ liệu thành các ô (cell), các cell này tạo thành cấu trúc dữ liệu lưới, sau đó các thao tác PCDL làm việc với các đối tượng trong từng cell này. Cách tiếp cận dựa trên lưới này không di chuyển các đối tượng trong các cell mà xây dựng nhiều mức phân cấp của nhóm các đối tượng trong một cell. Trong ngữ cảnh này, phương pháp này gần giống với phương pháp phân cụm phân cấp nhưng chỉ có điều chúng không trộn các cell. Do vậy các cụm không dựa trên độ đo khoảng cách (hay còn gọi là độ đo tương tự đối với các dữ liệu không gian) mà nó được quyết định bởi một tham số xác định trước. Ưu điểm của phương pháp PCDL dựa trên lưới là thời gian xử lý 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 cell trong mỗi chiều của không gian

lưới. Một thí dụ về cấu trúc dữ liệu lưới chứa các cell trong không gian như hình 6 sau:



Hình 2. 1: Mô hình cấu trúc dữ liệu lưới

Một số thuật toán PCDL dựa trên cấu trúc lưới điển hình STING, WaveCluster. .

2.3.5 Phương pháp phân cụm dựa trên mô hình

Phương pháp này cố gắng khám phá các phép xấp xỉ tốt của các tham số mô hình sao cho khớp với dữ liệu một cách tốt nhất. Chúng có thể sử dụng chiến lược phân cụm phân hoạch hoặc chiến lược phân cụm phân cấp, dựa trên cấu trúc hoặc mô hình mà chúng giả định về tập dữ liệu và cách mà chúng tinh chỉnh các mô hình này để nhận dạng ra các phân hoạch.

Phương pháp PCDL dựa trên mô hình cố gắng khớp giữa dữ liệu với mô hình toán học, nó dựa trên giả định rằng dữ liệu được tạo ra bằng hỗn hợp phân phối xác suất cơ bản. Các thuật toán phân cụm dựa trên mô hình có hai tiếp cận chính: mô hình thống kê và mạng Nơron. Phương pháp này gần giống với phương pháp dựa trên mật độ, bởi vì chúng phát triển các cụm riêng biệt nhằm cải tiến các mô hình đã được xác định trước đó, nhưng đôi khi nó không bắt đầu với một số cụm cố định và không sử dụng cùng một khái niệm mật độ cho các cụm.

2.3.6 Phương pháp phân cụm có dữ liệu ràng buộc

Sự phát triển của PCDL không gian trên CSDL lớn đã cung cấp nhiều công cụ tiện lợi cho việc phân tích thông tin địa lí, tuy nhiên hầu hết các thuật toán này cung cấp rất ít cách thức cho người dùng để xác định các ràng buộc trong thế giới thực cần phải được thỏa mãn trong quá trình phân cụm. Để PCDL không gian hiệu quả hơn, các nghiên cứu bổ sung cần được thực hiện để cung cấp cho người dùng khả năng kết hợp các ràng buộc trong thuật toán phân cụm.

Hiện nay, các phương pháp phân cụm trên đã, đang được phát triển và áp dụng nhiều trong các lĩnh vực khác nhau và đã có một số nhánh nghiên cứu được phát triển trên cơ sở của các phương pháp đó như:

- *Phân cụm thống kê*: Dựa trên các khái niệm phân tích thống kê, nhánh nghiên cứu này sử dụng các độ đo tương tự để phân hoạch các đối tượng, nhưng chúng chỉ áp dụng cho các dữ liệu có thuộc tính số.
- *Phân cụm khái niệm*: Kỹ thuật này được phát triển áp dụng cho dữ liệu hạng mục chúng phân cụm các đối tượng theo các khái niệm mà chúng xử lí.
- *Phân cụm mờ*: Sử dụng kỹ thuật mờ để PCDL, các thuật toán thuộc loại này chia ra lược đồ phân cụm thích hợp với tất cả các hoạt động đời sống hàng ngày, chúng chỉ xử lí các dữ liệu thực hiện không chắc chắn.
- *Phân cụm mạng Kohonen*: Loại phân cụm này dựa trên khái niệm của các mạng Nơron. Mạng Kohonen có tầng Nơron vào và các tầng Nơron ra. Mỗi Nơron của tầng vào tương ứng với mỗi thuộc tính của bản ghi, mỗi một Nơron vào kết nối với tất cả các Nơron của tầng ra. Mỗi liên kết được gắn liền với một trọng số nhằm xác định vị trí của Nơron ra tương ứng.

Các kỹ thuật PCDL trình bày ở trên đã được sử dụng rộng rãi trong thực tế, thế nhưng hầu hết chúng chỉ nhằm áp dụng cho tập dữ liệu với

cùng một kiểu thuộc tính. Vì vậy, việc PCDL trên tập dữ liệu có kiểu hỗn hợp là một vấn đề đặt ra trong khai phá dữ liệu trong giai đoạn hiện nay.

2.4 Các ứng dụng phân cụm dữ liệu

Phân cụm dữ liệu có rất nhiều ứng dụng trong các lĩnh vực khác nhau:

- *Thương mại*: Giúp các doanh nhân khám pha ra các nhóm khách hàng quan trọng để đưa ra các mục tiêu tiếp thị.
- *Sinh học*: Xác định các loài sinh vật, phân loại các Gen với chức năng tương đồng và thu được các cấu trúc trong các mẫu.
- *Lập quy hoặch đô thị*: Nhận dạng các nhóm nhà theo kiểu và vị trí địa lý, nhằm cung cấp thông tin cho quy hoạch đô thị.
- *Thư viện*: Phân loại các cụm sách có nội dung và ý nghĩa tương đồng nhau để cung cấp cho độc giả.
- *Bảo hiểm*: Nhận dạng nhóm tham gia bảo hiểm có chi phí bồi thường cao, nhận dạng gian lận thương mại.
- *Nghiên cứu trái đất*: Phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho nhân dạng các vùng nguy hiểm.
- *World Wide Web*: Có thể khám phá các nhóm tài liệu quan trọng, có nhiều ý nghĩa trong môi trường web. Các lớp tài liệu này trợ giúp cho việc khai phá dữ liệu từ dữ liệu.

CHƯƠNG 3: MỘT SỐ THUẬT TOÁN CƠ BẢN TRONG PHÂN CỤM DỮ LIỆU

3.1 Các thuật toán phân cụm phân hoạch

3.1.1 Thuật toán K-means

Thuật toán phân hoạch K-means do MacQueen đề xuất trong lĩnh vực thống kê năm 1967. Thuật toán dựa trên độ đo khoảng cách của các đối tượng dữ liệu trong cụm. Trong thực tế, nó đo khoảng cách tới giá trị trung bình của các dữ liệu trong cụm. Nó được xem như là trung tâm của cụm. Như vậy, nó cần khởi tạo một tập trung tâm các trung tâm cụm ban đầu và thông qua đó nó lặp lại các bước gồm gán mỗi đối tượng tới cụm mà trung tâm gần và tính toán lại trung tâm của mỗi cụm trên cơ sở gán mới cho các đối tượng. Quá trình lặp này dừng khi các trung tâm cụm hội tụ.

Mục đích của thuật toán K-means là sinh *k* cụm dữ liệu $\{C_1, C_2, \ldots, C_k\}$ từ một tập dữ liệu chứa *n* đối tượng trong không gian *d* chiều $X_i = \{ x_{i1}, x_{i2}, \ldots, x_{id} \}$, $i = 1 \div n$, sao cho hàm tiêu chuẩn:

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

đạt giá trị tối thiểu, trong đó m_i là trọng tâm của cụm C_i, D là khoảng cách giữa hai đối tượng.

Thuật toán K-means bao gồm các bước sau:

Input: Số cụm k và các trọng tâm cụm $m_j \overset{k}{_{j=1}}$

Output: Các cụm C_i (i = $\overline{1,k}$) và hàm tiêu chuẩn E đạt giá trị tối thiểu

Begin

Bước 1: Khởi tạo:

Chọn k trọng tâm $m_j \overset{k}{_{j=1}}$ ban đầu trong không gian R^d (d là số chiều của dữ liệu). Việc lựa chọn này có thể là ngẫu nhiên hoặc theo kinh nghiệm.

Bước 2: Tính toán khoảng cách:

Đối với mỗi điểm X_i ($1 \le i \le n$), tính khoảng cách của nó tới mỗi trọng tâm m_j: j = $\overline{1,k}$. Và sau đó tìm trọng tâm gần nhất đối với mỗi điểm.

Bước 3: Cập nhật lại trọng tâm:

Đối với mỗi j = $\overline{1,k}$, cập nhật trọng tâm cụm m_j bằng cách xác định trung bình cộng các vectơ đối tượng dữ liệu.

Điều kiện dừng:

Lặp lại các bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi.

End.

Nhận xét:

- Độ phức tạp của thuật toán là $O((3nkd)\tau T^{flop})$ với n là số đối tượng dữ liệu đưa vào, k là số cụm dữ liệu, d là số chiều, τ số vòng lặp, T^{flop} là thời gian để thực hiện một phép tính cơ sở như phép tính nhân, chia, …
- Do K-means phân tích cụm đơn giản nên có thể áp dụng đối với tập dữ liệu lớn. Tuy nhiên, nhược điểm của K-means là chỉ áp dụng với dữ liệu có thuộc tính số và khám phá ra các cụm có dạng hình cầu, K-means còn rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu.

3.1.2 Thuật toán K-Medoids

Thuật toán K-Medoids có khả năng khắc phục được nhiễu bằng cách chọn đối tượng ở gần tâm cụm nhất làm đại diện cho cụm đó (medoid). Thuật toán K-Medoids được thực hiện qua các bước sau:

Bước 1: Chọn K đối tượng bất kỳ trong N đối tượng ban đầu làm các medoid ban đầu

Bước 2: Lặp cho tới khi hội tụ

- Gán mỗi đối tượng còn lại vào cụm có medoid gần nhất với nó.
- Thay thế medoid hiện tại bằng một đối tượng không phải là medoid sao cho chất lượng phân cụm được cải thiện (chất lượng được đánh giá sử dụng hàm chi phí, hàm tính độ phi tương tự giữa một đối tượng và medoid của cụm chứa đối tượng đó).

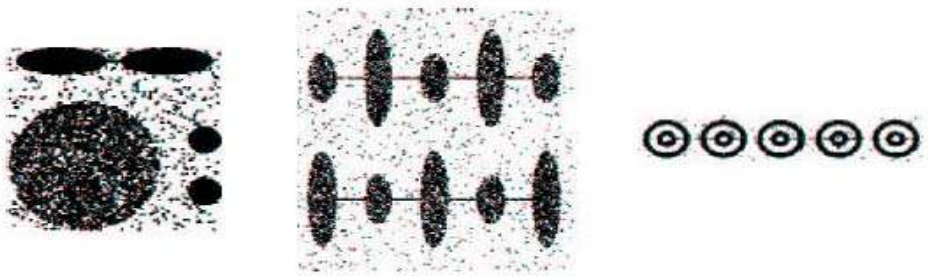
K-medoid tỏ ra hiệu quả hơn K-means trong trường hợp dữ liệu có nhiễu hoặc đối tượng ngoại lai (Outlier). Nhưng so với K-means thì K-Medoid có độ phức tập tính toán cao hơn. Cả hai thuật toán trên đều có nhược điểm chung là số lượng cụm k được cung cấp bởi người dùng.

Ngoài thuật toán K-means và K-Medoid, phân cụm phân hoạch còn bao gồm một số thuật toán khác như: thuật toán PAM, thuật toán CLARA, …

3.2 Thuật toán phân cụm phân cấp

Trong khi hầu hết các thuật toán thực hiện phân cụm với các cụm hình cầu và kích thước tương tự, như vậy là không hiệu quả khi xuất hiện các phần tử ngoại lai. Thuật toán CURE khắc phục được vấn đề này và tốt hơn với các phần tử ngoại lai. Thuật toán này định nghĩa một số cố định các điểm đại diện nằm rải rác trong toàn bộ không gian dữ liệu và được chọn để mô tả các cụm được hình thành. Các điểm này được tạo ra nhờ lựa chọn các đối tượng nằm rải rác cho cụm và sau đó "co lại" hoặc di chuyển chúng về trung tâm cụm bằng nhân tố co cụm. Quá trình này được lặp lại và như vậy trong quá trình này, có thể đo tỉ lệ gia tăng của cụm. Tại mỗi bước của thuật toán, hai cụm có cặp các điểm đại diện gần nhau (mỗi điểm trong cặp thuộc về mỗi cụm khác nhau) được hòa nhập.

Như vậy, có nhiều hơn một điểm đại diện mỗi cụm cho phép CURE khám phá được các cụm có hình dạng không phải là hình cầu. 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 lai và làm cho hiệu quả với những hình dạng không phải là hình cầu và kích thước độ rộng biến đổi. Hơn nữa, nó tỉ lệ tốt với cơ sở dữ liệu lớn mà không làm giảm chất lượng phân cụm.



Hình 3. 1: Các cụm dữ liệu được khám phá bởi CURE

Để xử lý được các cơ sở dữ liệu lớn, CURE sử dụng mẫu ngẫu nhiên và phân hoạch, một mẫu là được xác định ngẫu nhiên trước khi được phân hoạch và sau đó tiến hành phân cụm trên mỗi phân hoạch, như vậy mỗi phân hoạch

là từng phần đã được phân cụm, các cụm thu được lại được phân cụm lần thứ hai để thu được các cụm con mong muốn, nhưng mẫu ngẫu nhiên không nhất thiết đưa ra một mô tả cho toàn bộ tập dữ liệu.

Thuật toán CURE được thực hiện qua các bước cơ bản sau:

Chọn một mẫu ngẫu nhiên S từ tập dữ liệu ban đầu.

Phân hoạch mẫu S thành các nhóm dữ liệu có kích thước bằng nhau: Ý tưởng chính ở đây là phân hoạch mẫu thành p nhóm dữ liệu bằng nhau, kích thước của mỗi phân hoạch là n'/p (n' là kích thước của mẫu).

Phân cụm các điểm của mỗi nhóm: thực hiện PCDL cho các nhóm cho đến khi mỗi nhóm được phân thành n'/pq cụm (với q>1).

Loại bỏ các phần tử ngoại lai: trước hết, khi các cụm được hình thành cho đến khi số các cụm giảm xuống một phần so với số các cụm ban đầu. Sau đó, trong trường hợp các phần tử ngoại lai được lấy mẫu cùng với quá trình pha khởi tạo mẫu dữ liệu, thuật toán sẽ tự động loại bỏ các nhóm nhỏ.

Phân cụm các cụm không gian: các đối tượng đại diện cho các cụm di chuyển về hướng trung tâm cụm, nghĩa là chúng được thay thế bởi các đối tượng gần trung tâm hơn.

Đánh dấu dữ liệu với các nhãn tương ứng.

Độ phức tạp tính toán của thuật toán CURE là $O(n^2\log(n))$. CURE là thuật toán tin cậy trong việc khám phá ra các cụm có hình thù bất kỳ và có thể áp dụng tốt đối với dữ liệu có phần tử ngoại lai và trên các tập dữ liệu hai chiều. Tuy nhiên, nó lại rất nhạy cảm với tham số như số các đối tượng đại diện, tỉ lệ co của các phần tử đại diện.

Ngoài thuật toán CURE ra, phân cụm phân cấp còn bao gồm một số thuật toán khác như: thuật toán BIRCH, thuật toán AGNES, thuật toán DIANA, thuật toán ROCK, thuật toán CHANMELEON.

3.3 Thuật toán COP-Kmeans

Thuật toán COP-Kmeans là một thuật toán phân cụm dữ liệu nửa giám sát, với phương pháp tiếp cận dựa trên tìm kiếm. Trong thuật toán COP-Kmeans (Wagstaff đề xuất năm 2001), các thông tin bổ trợ được cung cấp dưới dạng một tập các ràng buộc must-link và cannot-link.

Trong đó :

- Must-link: hai đối tượng dữ liệu phải cùng nằm trong một cụm.
- Cannot-link: hai đối tượng dữ liệu phải khác cụm với nhau.

Các rằng buộc này được áp dụng vào trong suốt quá trình phân cụm. Nhằm điều hướng quá trình phân cụm để đạt được kết quả phân cụm theo ý muốn. Thuật toán COP-Kmeans được thực hiện như sau:

Input:

- Tập các đối tượn dữ liệu $X = \{x_1,\ldots,x_n\}$, $x_1 \in R^d$
- Số lượng cụm: K
- Tập ràng buộc must-link và cannot-link

Output:

- K phân hoạch tách rời: $X_h \ _{h=1}^{K}$ của X sao cho hàm mục tiêu được tối ưu.

Các bước thự hiện:

1. Khởi tạo các cụm: các tâm ban đầu được chọn ngẫu nhiên sao cho không vi phạm ràng buộc đã cho.
2. Lặp cho tới khi hội tụ

   - Gán cụm: gán mỗi đối tượng dữ liệu vào trong cụm gần nhất sao cho không vi phạm ràng buộc.

   - Ước lượng tâm: cập nhật lại tâm là trung bình của tất cả đối tượng nằm trong cụm của tâm đó.

   - $t \leftarrow t + 1$.

CHƯƠNG 4: ỨNG DỤNG THUẬT TOÁN K-MEANS CHO PHÂN ĐOẠN ẢNH

4.1 Tổng quan về phân vùng ảnh

Phân vùng ảnh là bước then chốt trong xử lý ảnh. Giai đoạn này nhằm phân tích ảnh thành những thành phần có cùng tính chất nào đó dựa theo biên hay các vùng liên thông. Tiêu chuẩn để xác định vùng liên thông có thể là cùng mức xám, cùng mầu hay cùng độ nhám, v. v.

Nếu phân vùng dựa trên các miền liên thông, ta gọi là kỹ thuật phân vùng dựa theo miền đồng nhất. Nếu ta phân vùng dựa vào biên gọi là kỹ thuật phân vùng biên. Ngoài ra, còn có các kỹ thuật khác như phân vùng dựa vào biên độ, phân vùng theo kết cấu (Texture Segmentation).

Mục đích của phân tích ảnh là để có một miêu tả tổng hợp về nhiều phần tử khác nhau cấu tạo nên ảnh thô (brut image). Vì lượng thông tin chứa trong ảnh là rất lớn, trong khi đó đa số ứng dụng chỉ cần một số thông tin đặc trưng nào đó, do vậy cần có một quá trình giảm lượng thông tin khổng lồ ấy. Quá trình này bao gồm phân vùng ảnh và trích chọn đặc tính chủ yếu. Các kỹ thuật dùng cho quá trình này sẽ được đề cập tới ở phần sau.

4.1.1 Phân vùng ảnh theo ngưỡng biên độ

Đặc tính đơn giản nhất và có hữu ích của ảnh đó là biên độ của các tính chất vật lý của ảnh như: độ phản xạ, độ truyền sáng, màu sắc hoặc đáp ứng đa phổ. Thí dụ, trong ảnh X-quang, biên độ mức xám biểu diễn đặc tính bão hòa của các phần hấp thụ của cơ thể làm cho ta có khả năng phân biệt xương với các phần mềm, tế bào lành với các tế bào bị nhiễm bệnh, v. v.

Kỹ thuật phân ngưỡng theo biên độ rất có ích đối với ảnh nhị phân như văn bản in, đồ họa, ảnh màu hay ảnh X-quang. Việc chọn ngưỡng trong kỹ

thuật này là bước rất quan trọng. Người ta thường tiến hành theo các bước chung sau:

- Xem xét lược đồ xám của ảnh để xác định các đỉnh và các khe. Nếu ảnh có dạng rắn lượn (nhiều đỉnh và nhiều khe), các khe có thể sử dụng để chọn ngưỡng.
- Chọn ngưỡng t sao cho một phần xác định trước η của toàn bộ số mẫu là thấp hơn t.
- Điều chỉnh ngưỡng dựa trên xem xét lược đồ xám của các điểm lân cận.
- Chọn ngưỡng như xem xét lược đồ xám của những điểm thỏa tiêu chuẩn chọn. Thí dụ với ảnh có độ tương phản thấp, lược đồ của những điểm có biên độ Laplace g(m, n) lớn hơn giá trị t định trước (sao cho từ 5% đến 10% số điểm ảnh với gradient lớn nhất sẽ coi như biên) sẽ cho phép xác định các đặc tính ảnh lưỡng cực tốt hơn ảnh gốc.

Khi có một mô hình phân lớp xác suất, việc xác định ngưỡng dựa vào tiêu chuẩn nhằm cực tiểu xác suất của sai số hoặc một số tính chất khác theo luật Bayes.

4.1.2 Phân vùng ảnh theo miền đồng nhất

Kỹ thuật phân vùng ảnh thành các miền đồng nhất dựa vào các tính chất quan trọng nào đó của miền. Việc lựa chọn các tính chất của miền sẽ xác định tiêu chuẩn phân vùng. Ở đây cũng cần phải xác định rõ tính đồng nhất của một miền của ảnh vì đó là điểm chủ yếu xác định tính hiệu quả của việc phân vùng. Các tiêu chuẩn hay được dùng là sự thuần nhất về mức xám, màu sắc đối với ảnh màu, kết cấu sợi và chuyển động.

Thí dụ, trong ứng dụng ảnh về hàng không, việc phân vùng theo màu cho phép phân biệt thảm thực vật: cánh đồng màu xanh hay vàng, rừng xanh thẫm, đường màu xám, mái nhà đỏ, v. v.

Đối với ảnh chuyển động, người ta tiến hành trừ hai ảnh quan sát được tại hai thời điểm khác nhau. Trong trường hợp này phần ảnh không thay đổi sẽ nhận giá trị không, những phần thay đổi sẽ nhận giá trị dương hay âm tương ứng với thay đổi hay dịch chuyển.

Các phương pháp thực hiện đó là:

- **Phương pháp tách cây tứ phân**

Phương pháp này kiểm tra tính hợp thức của tiêu chuẩn một cách tổng thể trên miền lớn của ảnh. Nếu thỏa mãn tiêu chuẩn việc phân đoạn coi như kết thúc. Trong trường hợp ngược lại, ta chia miền đang xét thành bốn miền nhỏ hơn. Với mỗi miền nhỏ, ta lại áp dụng một cách đệ quy phương pháp trên cho đến khi tất cả các miền đều thỏa mãn.

- **Phương pháp cục bộ hay phân vùng bởi hợp**

Ý tưởng của phương pháp này là xem xét ảnh từ các miền nhỏ nhất rồi hợp chúng lại nếu thỏa mãn tiêu chuẩn được một miền đồng nhất lớn hơn. Ta lại tiếp tục với các miền thu được cho tới khi không thể hợp được nữa. Số miền còn lại cho ta kết quả phân đoạn. Như vậy, miền nhỏ nhất của bước xuất phát là điểm ảnh.

Điều quan trọng của phương pháp này là nguyên lý hợp hai vùng. Việc hợp hai vùng thực hiện theo nguyên tắc sau:

- Hai vùng phải đáp ứng tiêu chuẩn, thí dụ như cùng màu hoặc cùng mức xám.
- Chúng phải kế cận nhau.

- **Phương pháp tổng hợp**

Hai phương pháp vừa xét ở trên có một số nhược điểm. Phương pháp tách sẽ tạo nên một cấu trúc phân cấp và thiết lập mối quan hệ giữa các vùng. Tuy nhiên nó thực hiện việc chia quá chi tiết.

Phương pháp hợp cho phép làm giảm số miền liên thông xuống tối thiểu, nhưng cấu trúc hàng ngang dàn trải, không cho ta thấy mối quan hệ giữa các miền.

Chính vì vậy người ta nghĩ đến việc phối hợp cả hai phương pháp. Trước tiên, dùng phương pháp tách để tạo nên cây tứ phân, phân đoạn theo hướng từ gốc tới lá. Tiếp theo, tiến hành duyệt cây theo chiều ngược lại và hợp các vùng có cùng tiêu chuẩn. Với phương pháp này ta thu được một miêu tả cấu trúc của ảnh với các miền liên thông có kích thước tối đa.

Các bước chính bao gồm:

- Kiểm tra tiêu chuẩn đông nhất.
- Hợp vùng.

4.1.3 Phân vùng dựa theo đường biên

Biên là một trong những đặc trưng quan trọng của ảnh. Cũng vì thế mà trong nhiều ứng dụng, người ta sử dụng các phân đoạn dựa theo biên. Việc phân đoạn ảnh dựa vào biên được tiến hành qua một số bước như sau:

- Phát hiện và làm nổi biên.
- Làm mảnh biên.
- Nhị phân hóa đường biên.
- Mô tả biên.

4.1.4 Phân đoạn dựa theo kết cấu bề mặt

Kết cấu là thuật ngữ phản ánh sự lặp lại của các phần tử sợi (texel) cơ bản. Sự lặp lại này có thể ngẫu nhiên hay có tính chu kì hoặc gần như có chu kì. Một texel chứa rất nhiều điểm ảnh. Trong phân tích ảnh, kết cấu được phân làm hai loại chính:

- Thống kê.
- Cấu trúc.

Khi đối tượng xuất hiện trên một nền có tính kết cấu cao, việc phân đoạn dựa vào tính kết cấu trở nên khá quan trọng. Nguyên nhân là vì kết cấu sợi thường chứa mật độ cao các gờ (edge) và làm cho phân đoạn dựa vào biên trở nên kém hiệu quả, trừ phi ta loại tính kết cấu. Việc phân đoạn dựa vào miền đồng nhất cũng có thể áp dụng cho các đặc trưng kết cấu và có thể dùng để phân đoạn các miền có tính kết cấu.

4.2 Thuật toán K-means cho phân đoạn ảnh

Tầm quan trọng và những khó khăn của việc gom nhóm các đối tượng mang tính tri giác của con người từ lâu đã được nghiên cứu nhiều trong các lĩnh vực của thị giác máy tinh đặc biệt trong lĩnh vực của xử lí ảnh. Và phân đoạn ảnh đã có những ứng dụng mạnh mẽ và rộng rãi trong các bài toán phân tích và hiểu ảnh tự động, nhưng nó cũng là một bài toán khó mà đến bây giờ các nhà khoa học vẫn chưa giải quyết được một cách hoàn toàn thấu đáo. Làm thế nào để phân chia một ảnh thành các tập con. Những cách khả thi để có thể làm được điều đó. Đó là những câu hỏi mà người ta đã đặt ra từ lâu và mong muốn tìm được câu trả lời.

Trong khoảng 30 năm trở lại đây đã có rất nhiều các thuật toán được đề xuất để giải quyết bài toán phân đoạn ảnh. Các thuật toán hầu hết đều dựa vào hai thuộc tính quan trọng của mỗi điểm ảnh so với các điểm lân cận của nó, đó là: sự khác (dissimilarity) và giống nhau (similarity). Các phương pháp dựa trên sự giống nhau của các điểm ảnh được gọi là phương pháp miền (region-based methods), còn các phương pháp dựa trên sự khác nhau của các điểm ảnh được gọi là các phương pháp biên (boundary-based methods). Trong bài báo cáo này em xin phép được trình bày thuật toán K-means để giải quyết bài toán phân đoạn ảnh.

4.2.1 Mô tả bài toán

Input:

- Ảnh có kích thước m*n.

- Số cụm (k) muốn phân đoạn.

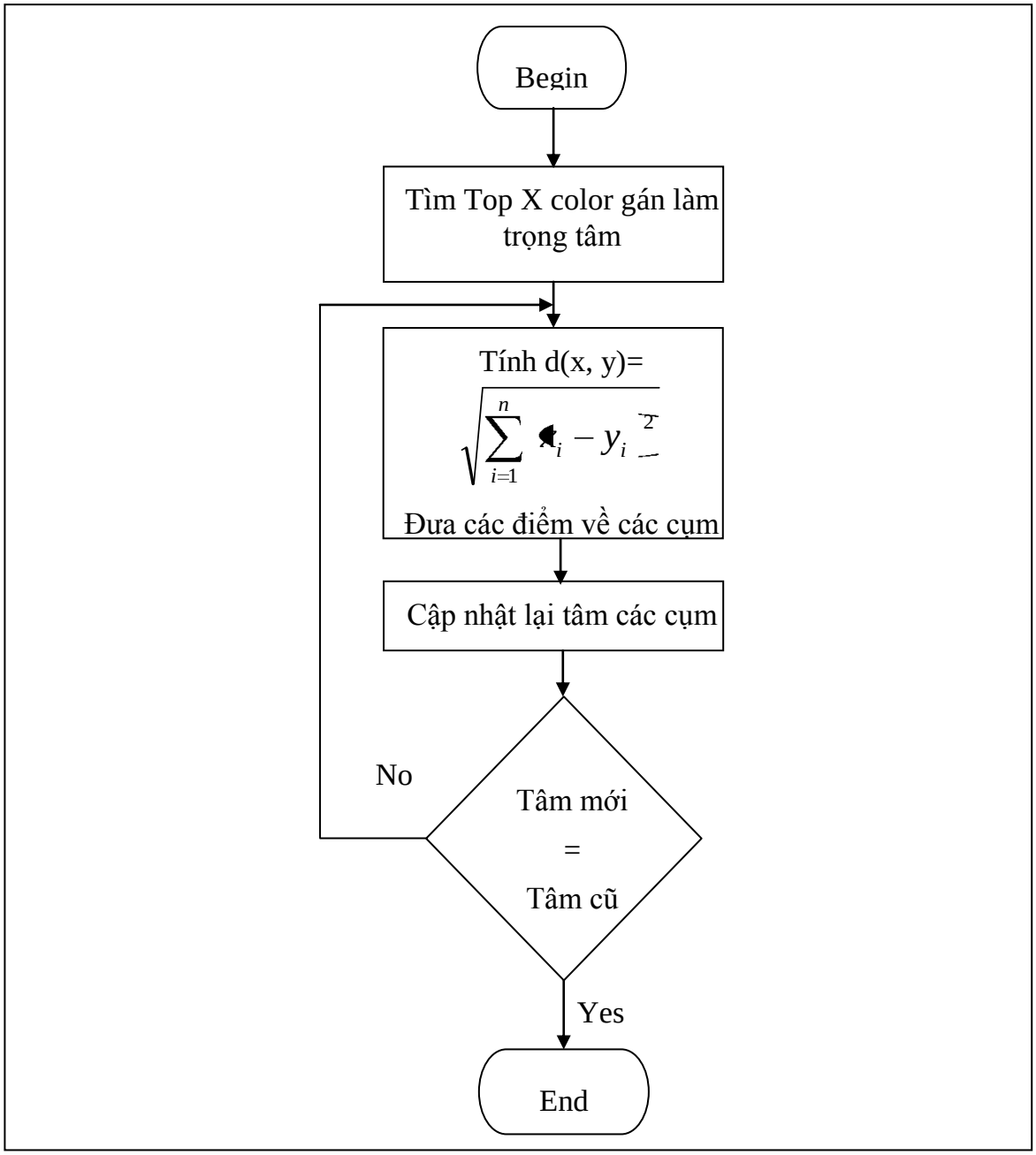
Output:

- Ảnh được phân thành k đoạn có màu sắc tương đồng nhau.

4.2.2 Các bước thực hiện chính trong thuật toán

Thuận toán sẽ dựa vào số lượng cụm mong muốn, trọng tâm các cụm mà tính toán khoảng cách giữa các điểm với các trọng tâm cụm. Sau đó gán các điểm tới cụm mà nó có khoảng cách tới trọng tâm cụm đó là nhỏ nhất, cập nhật lại trọng tâm cụm. Kết quả thu được sau khi tâm các cụm là không đổi.

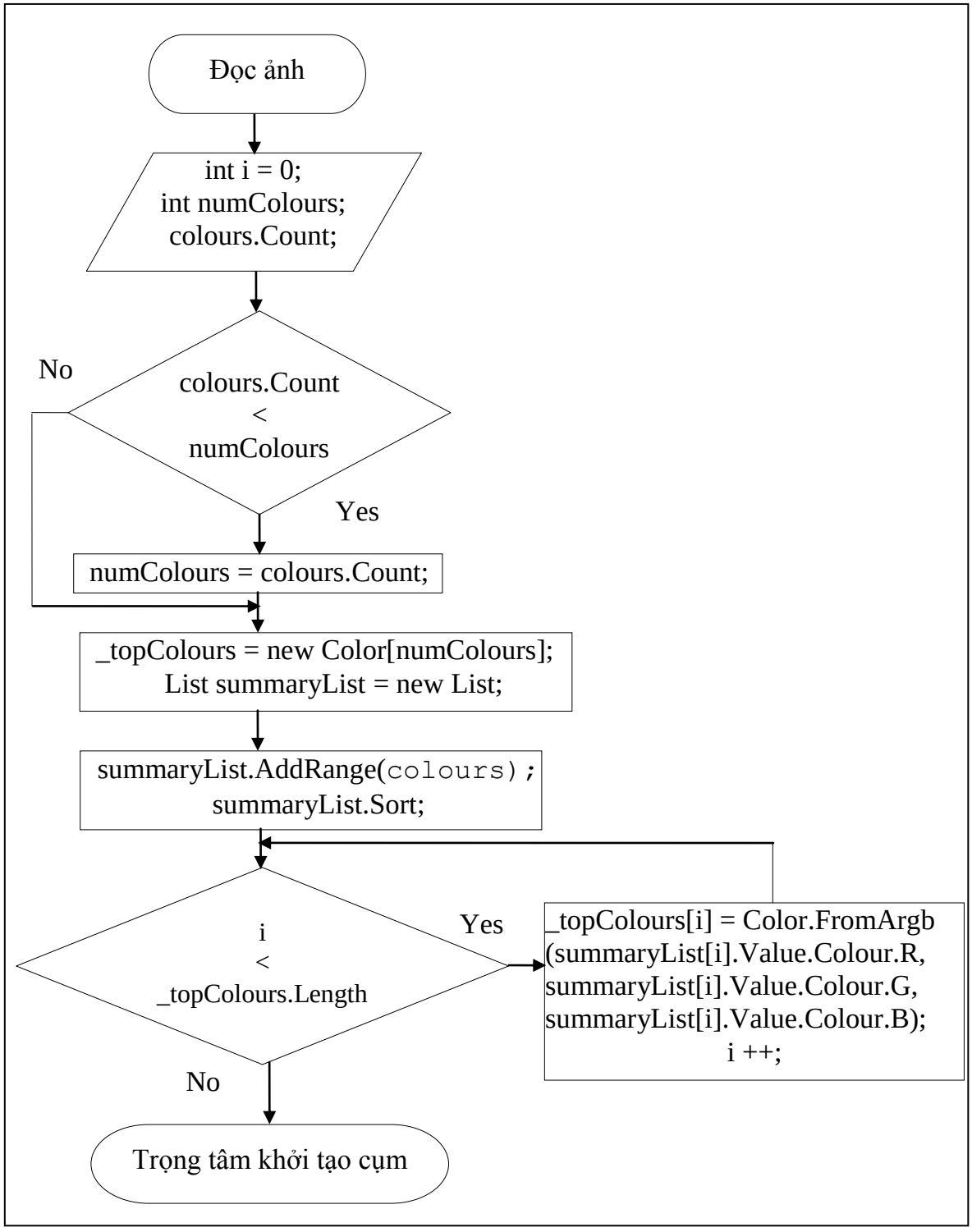
Lưu đồ tổng quát của thuật toán:



Hình 4. 1: Thuật toán K-means.

4.2.2.1 Tìm kiếm Top X color

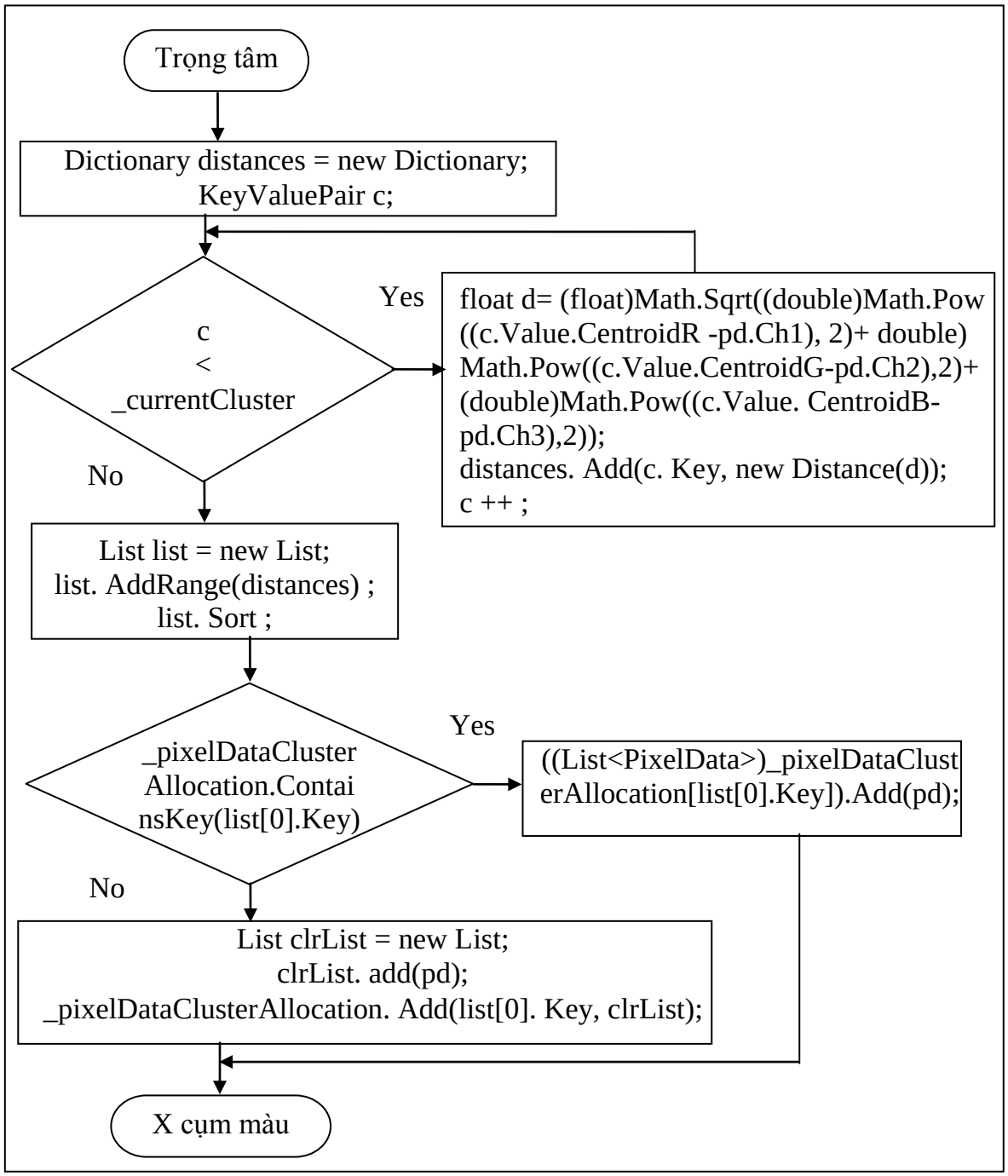
Đầu tiên ta so sánh số màu thực tế có trong ảnh và số cụm màu, nếu số màu thực tế nhỏ hơn số cụm màu thì ta nhận số cụm màu chính là số màu thực tế. Tạo danh sách chứa các loại màu, sau đó sắp xếp chúng theo thứ tự giảm dần. Lấy X phần tử đầu tiên của danh sách.



Hình 4. 2: Tìm kiếm Top X color.

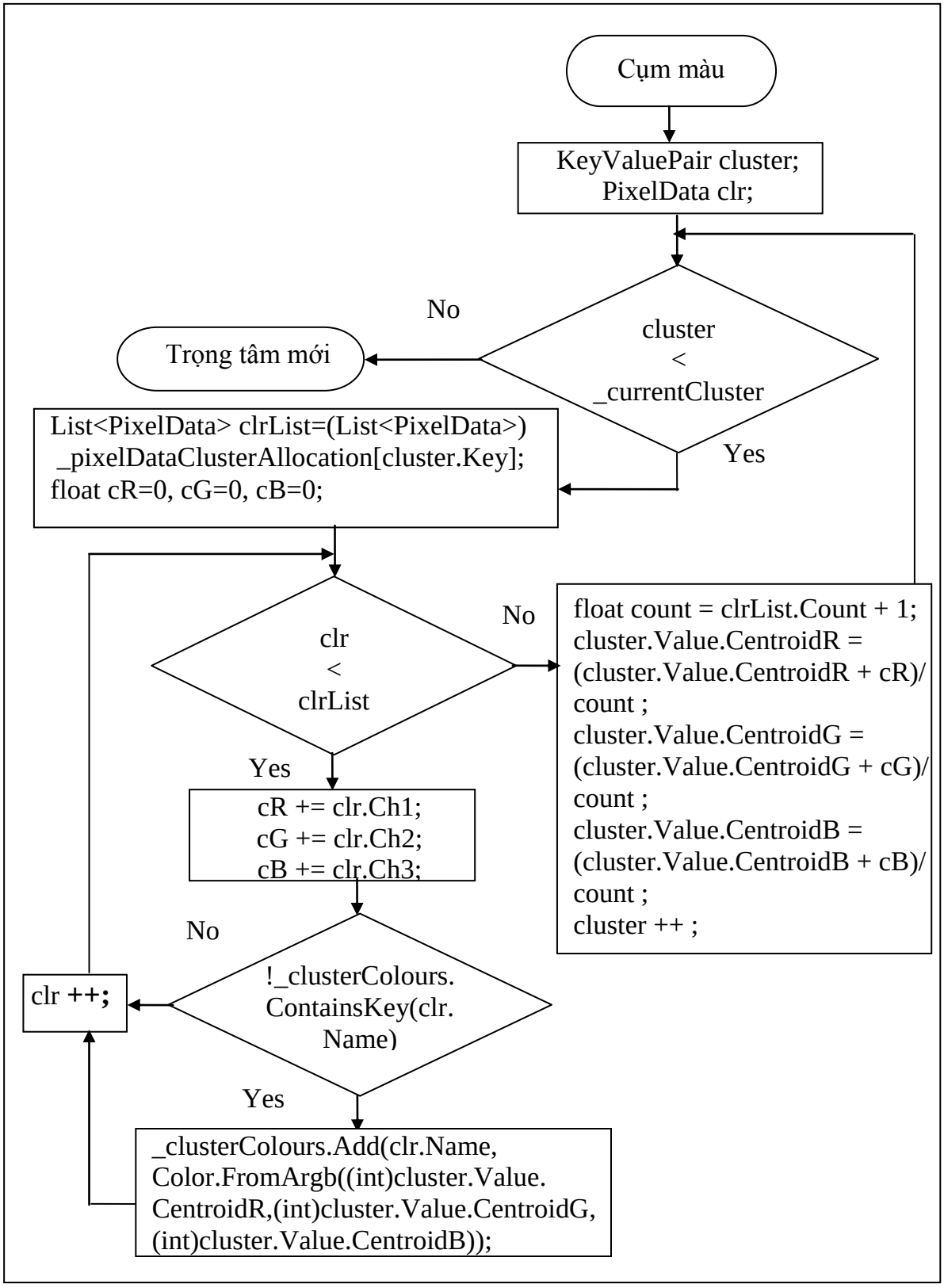
4.2.2.2 Tính khoảng cách và phân cụm

Sử dụng thuật toán Euclide tính khoảng cách màu của các điểm với các tâm cụm. Dựa vào khoảng cách đó đưa các điểm vào cụm mà khoảng cách của nó tới tâm cụm là nhỏ nhất.



Hình 4. 3: Phân cụm.

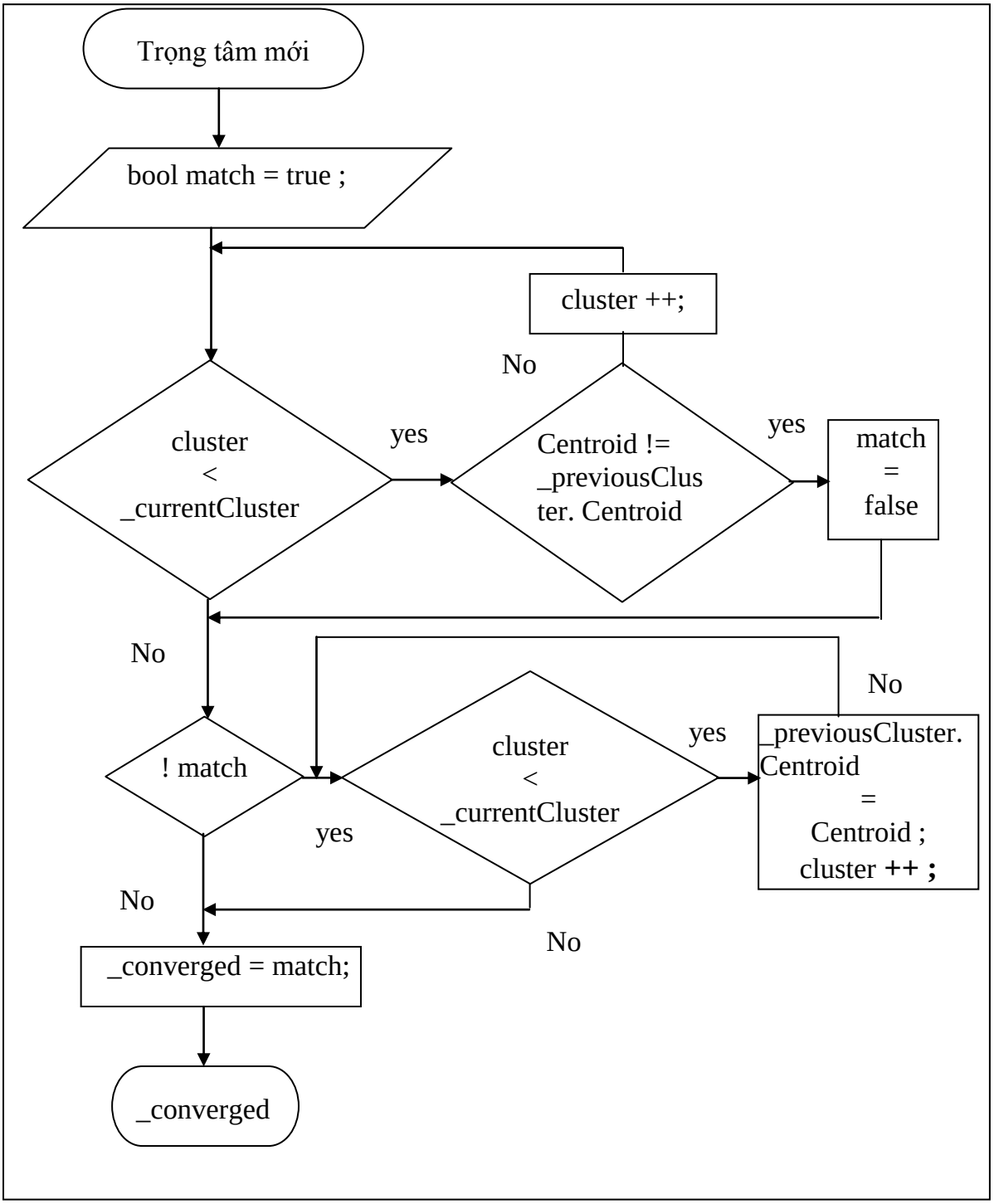
4.2.2.3 Tính lại trọng tâm cụm



Hình 4. 4: Tính trọng tâm mới.

4.2.2.4 Kiểm tra hội tụ

Để kiểm tra tính hội tụ của dữ liệu chúng ta kiểm tra trọng tâm hiện tại vừa tính được với trọng tâm trước đó của cụm.



Hình 4. 5: Kiểm tra hội tụ.

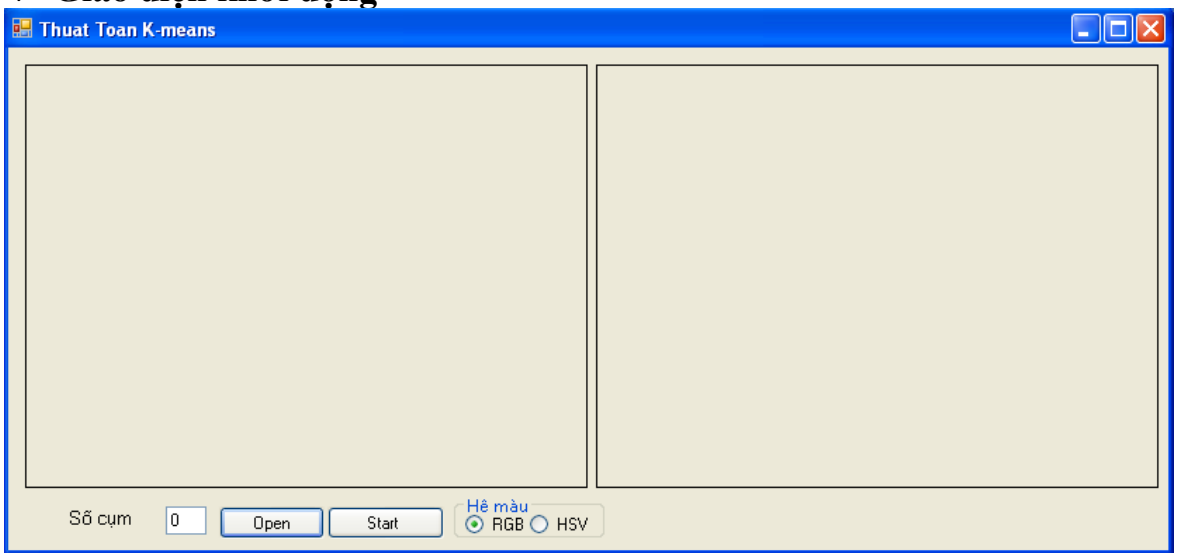
4.2.3 Kết quả thực nghiệm

4.2.3.1 Môi trường cài đặt.

Chương trình được lập trình với ngôn ngữ C#, cài đặt và chạy thử nghiệm trên môi trường hệ điều hành Windows XP.

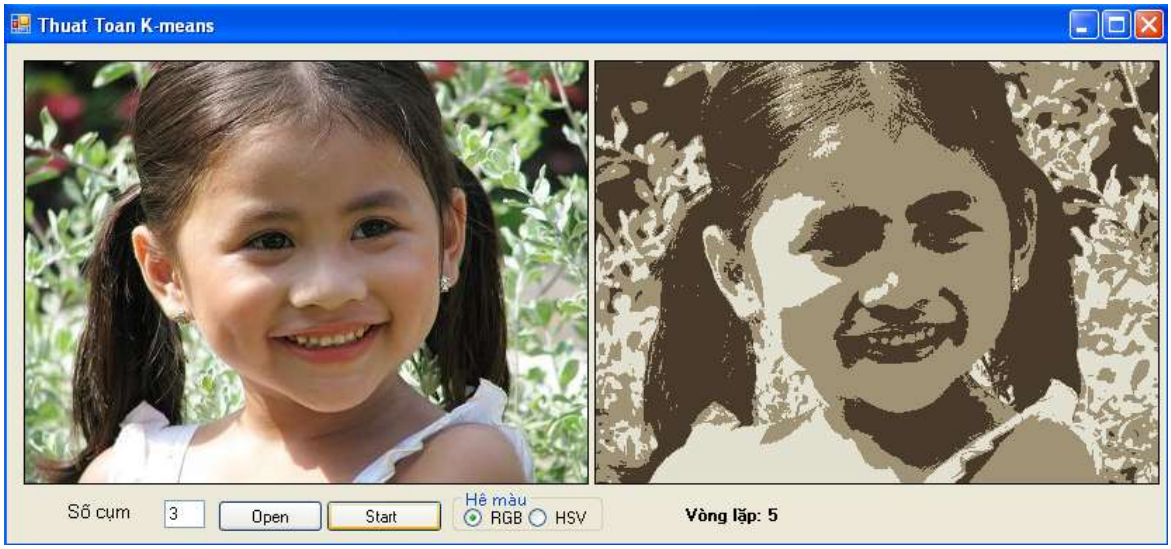
4.2.3.2 Một số giao diện.

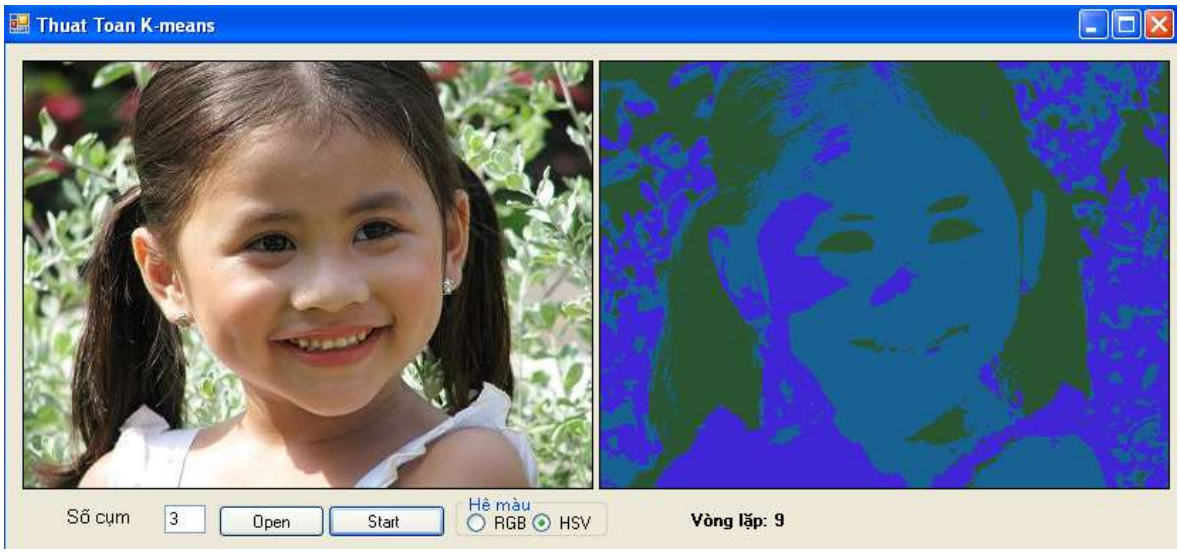
♦ **Giao diện khởi động**



♦ **Đưa dữ liệu vào**

- **Quá trình xử lý dữ liệu.**





- **Kết quả phân cụm.**

KẾT LUẬN

Trong quá trình nghiên cứu, tìm hiểu và hoàn thành đề tài đồ án tốt nghiệp "Tìm hiểu một số phuơng pháp phân cụm dữ liệu và ứng dụng", em đã thu nhận được thêm những kiến thức và em cũng nhận thấy phân cụm dữ liệu trong khai phá dữ liệu là một lĩnh vực nghiên cứu rộng lớn, còn nhiều điều mà chúng ta cần khám phá. Trong đề tài em đã cố gắng tập trung tìm hiểu và nghiên cứu tổng quan khai phá dữ liệu, phân cụm dữ liệu và một số thuật toán của nó, tổng quan về phân vùng ảnh. Cài đặt thử nghiệm thuật toán k-means với ứng dụng là phân đoạn ảnh.

Do thời gian thực hiện hạn chế nên em mới chỉ tìm hiểu đươc một số kỹ thuật cơ bản trong phân cụm dữ liệu, cài đặt thử nghiệm với thuật toán K-means. Nhưng còn một số các kỹ thuật em vẫn chưa tìm hiểu, khai thác và ứng dụng cho các bài toán … Trong thời gian tới em sẽ cố gắng tiếp tục nghiên cứu, tìm hiểu thêm một sô kỹ thuật phân cụm và nhất là có thể tìm hiểu và phát triển các kỹ thuật phân đoạn ảnh để có thể xử lý với ảnh động.

Sinh viên

VŨ MINH ĐÔNG

TÀI LIỆU THAM KHẢO

Tài liệu tham khảo tiếng Việt

[1.] Nhập môn xử lý ảnh, **Lương Mạnh Bá và Nguyễn Thanh Thủy**, nhà xuất bản Khoa học Kỹ thuật, 1999.

[2.] Giáo trình xử lý ảnh, **Ngô Quốc Tạo**, lớp CHCLC – ĐH Công Nghệ ĐHQG Hà Nội năm 2001- 2002.

[3.] Bài giảng môn Data Mining, **Ngô Quốc Tạo**, lớp CHK5 – ĐH Thái Nguyên 2006 – 2008.

[4.] Thuật toán phân cụm dữ liệu nửa giám sát, **Lưu Tuấn Lâm** – Đồ án tốt nghiệp ĐHDL Hải Phòng.

Tài liệu tham khảo tiếng Anh

[5.] Discovering Knowledge in Data: An Introduction to Data Mining, **Daniel T. Larose**, ISBN 0-471-66657-2 CopyrightC 2005 John Wiley & Sons, Inc.

[6.] In Proc. 1996 Int. Conf. Data Mining and Knowledge Discovery (KDD-96), **A. Arning, R. Agrawal and P. Raghavan. Alinear method for deviation detection in larger databases,** Portland, Oregon, August 1996.

[7.] http://www.wikipedia.org