LỜI CẢM ƠN

Trong suốt thời gian học tập, hoàn thành bài đồ án tốt nghiệp em đã may mắn được các thầy cô chỉ bảo, dìu dắt và được gia đình, bạn bè quan tâm, động viên.

Trước tiên em xin được bày tỏ lòng biết ơn chân thành nhất tới PGS TS Ngô Quốc Tạo, người đã định hướng và nhiệt tình chỉ bảo, hướng dẫn em trong suốt quá trình thực hiện bài đồ án tốt nghiệp này.

Em cũng xin gửi lời cảm ơn tới các thầy cô trong ngành hệ thống thông tin nói riêng và trường đại học Dân Lập Hải Phòng nói chung đã dạy bảo, cung cấp những kiến thức quý báu cho em trong suốt quá trình nghiên cứu và học tập tại trường.

Em cũng xin gửi lời cảm ơn tới gia đình, bạn bè những người luôn cổ vũ, quan tâm và giúp đỡ em trong suốt thời gian học tập cũng như thời gian làm đồ án tốt nghiệp.

Do thời gian và kiến thức có hạn nên không tránh khỏi những thiếu sót nhất định. Em rất mong nhận được sự đóng góp quý báu của thầy cô và các bạn!

Em xin chân thành cảm ơn!

Hải Phòng, tháng 11 năm 2013

Sinh viên

Bùi Trung Thành

MỤC LỤC

LỜI CẢM ƠN 1
LỜI NÓI ĐẦU 4
CHƯƠNG I: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU 7
1.1. Giới thiệu về khám phá tri thức 7
1.2. Khai phá dữ liệu và các khái niệm liên quan 9
1.2.1. Khái niệm khai phá dữ liệu 9
1.2.2. Các bước trong quá trình khai phá dữ liệu 10
1.2.3. Các thành phần trong khai phá dữ liệu 11
1.2.4. Các hướng tiếp cận và kỹ thuật áp dụng trong khai phá dữ liệu 12
1.2.5. Ứng dụng của khai phá dữ liệu 13
CHƯƠNG IIPHÂN CỤM DỮ LIỆU VÀ CÁCTHUẬT TOÁN PHÂN CỤM DỮ LIỆU 14
2.1. Phân cụm dữ liệu 14
2.1.1. Định nghĩa về phân cụm dữ liệu 14
2.1.2. Một số ví dụ về phân cụm dữ liệu 15
2.2. Một số kiểu dữ liệu trong phân cụm 17
2.2.1. Kiểu dữ liệu dựa trên kích thước miền 18
2.2.2. Kiểu dữ liệu dựa trên hệ đo 18
2.3. Phép đo độ tương tự và khoảng cách đối với các kiểu dữ liệu 20
2.3.1. Khái niệm tương tự và phi tương tự 20
2.3.2. Độ đo khoảng cách 21
2.4. Các hướng tiếp cận của bài toán phân cụm dữ liệu 24
2.4.1. Phương pháp phân cụm phân hoạch 24
2.4.2. Phương pháp phân cụm phân cấp 24
2.4.3. Phương pháp phân cụm dựa trên mật độ 26
2.4.4. Phương pháp phân cụm dựa trên lưới 29
2.4.5. Phương pháp phân cụm dựa trên mô hình 30
2.4.6. Phương pháp phân cụm dựa trên dữ liệu ràng buộc 30
2.5. Một số thuật toán phân cụm dữ liệu 30
2.5.1. Các thuật toán phân cụm phân hoạch 30
2.5.2. Thuật toán phân cụm phân cấp 32
2.5.3. Thuật toán COP – Kmeans 33

CHƯƠNG III: ỨNG DỤNG THUẬT TOÁN K - MEANS TRONG PHÂN ĐOẠN ẢNH 35
3.1. Tổng quan về phân vùng ảnh 35
3.2. Các hướng tiếp cận phân đoạn ảnh 36
3.2.1. Các phương pháp dựa trên không gian đặc trưng 36
3.2.2. Các phương pháp dựa trên không gian ảnh 37
3.2.3. Các phương pháp dựa trên mô hình vật lý 38
3.3. Một số phương pháp phân đoạn cụ thể 41
3.3.1. Phương pháp phân đoạn yếu của B.G. Prasad 41
3.3.2. Phương pháp phân đoạn dựa trên ngưỡng cục bộ thích nghi 46
3.3.3. Phân đoạn sơ khởi bằng Watershed 47
3.3.4. Trộn các vùng 50
3.4. Thuật toán k-means cho phân đoạn ảnh 53
3.4.1. Mô tả bài toán 54
3.4.2. Các bước thực hiện chính trong thuật toán 54
3.4.3. Kết quả thực nghiệm 58
3.4.4. Ưu, nhược điểm của thuật toán k – means 59
KẾT LUẬN 61
TÀI LIỆU THAM KHẢO 62

DANH MỤC HÌNH

Hình 1: Quy trình phát hiện tri thức 8
Hình 2: Các bước trong khai phá dữ liệu 10
Hình 3: Hai phương pháp tiếp cận phân cấp 25
Hình 4: p là một điểm hạt nhân với bán kính Eps 1cm và ngưỡng trù mật là min Pts là 3. Khoảng cách được dùng là khoảng cách Euclide trong không gian hình học hai chiều, q là một điểm liên thông mật độ trực tiếp từ p. 27
Hình 5: q là một điểm liên thông mật độ từ p 27
Hình 6: p và q là hai điểm có kết nối mật độ 28
Hình 7: Những cụm dữ liệu được khám phá bởi CURE 32
Hình 8: ví dụ phân đoạn ảnh bằng phương pháp phân đoạn yếu 42
Hình 9:(a) Ảnh gốc. (b) Kết quả phân đoạn bằng ngưỡng toàn cục 100. 52
Hình 10: (a) Ảnh gốc (b) Sau khi áp dụng giải thuật watershed. 53
Hình 11: Vùng sáng elip hiển thị khác nhau khi do nền khác nhau. 53
Hình 12: Thuật toán k - means 56
Hình 13: Tìm kiếm top x color 57
Hình 14: Giao diện chính của chương trình 59
Hình 15: Chọn ảnh đầu vào 59
Hình 16:Kết quả của quá trình phân cụm ảnh 59

LỜI NÓI ĐẦU

Trong những năm gần đây sự phát triển mạnh mẽ của CNTT đã làm cho khả năng thuthập và lưu trữ thông tin của các hệ thống thông tin tăng lên nhanh chóng. Bên cạnh đó, việc tin học hóa một cách ồ ạt làm cho hoạt động sản xuất kinh doanh cũng như nhiều lĩnh vực khác đã tạo ra một lượng dữ liệu khổng lồ. Hàng triệu cơ sở dữ liệu (CSDL) đã được sử dụng cho các hoạt động sản xuất, kinh doanh….Trong đó, có nhiều CSDL lên tới hàng nghìn Gigabyte, thậm chí lên mức Terabyte.

Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết, cần có công cụ mới, hiện đại để có thể chuyển đổi lượng dữ liệu khổng lồ này thành các tri thức có ích. Từ đó, khái niệm "khai phá dữ liệu" đã ra đời, nó đã trở thành lĩnh vực thời sự của nền CNTT của thế giới nói chung và Viêt Nam nói riêng. Khai phá dữ liệu đang được ứng dụng rất rộng rãi trong nhiều lĩnh vực của đời sống: Marketing, ngân hàng, bảo hiểm, y tế, khoa học, internet….

Các kỹ thuật khai phá dữ liệu được chia thành 2 nhóm chính: kỹ thuật khai phá dữ liệu mô tả và kỹ thuật khai phá dữ liệu dự đoán.

Bài báo cáo đồ án tốt nghiệp này em xin trình bày vấn đề "Phân cụm cứng", một trong những vấn đề cơ bản của khai phá dữ liệu.

Bài báo cáo được trình bày trong 3 chương:

- Chương 1: Trình bày tổng quan về Khai phá dữ liệu; Phân cụm dữ liệu;Ứng dụng trong đời sống.
- Chương 2: Phương pháp phân cụm cứng trong phân đoạn ảnh.
- Chương 3: Xây dựng chương trình demo.

Kết luận: Tóm tắt những vấn đề tìm hiểu được trong bài, các vấn đề liên quan và đưa ra hướng phát triển trong tương lai.

TÓM TẮT ĐỀ TÀI

Bài báo cáo đồ án tốt nghiệp của em, nghiên cứu về " phương pháp phân cụm cứng trong phân đoạn ảnh". Nội dung nghiên cứu gồm 3 chương như sau:

CHƯƠNG I: Tổng quan về khai phá dữ liệu

Chương này nghiên cứu tổng quan về khai phá dữ liệu và khám phá tri thức. Quy trình khám phá tri thức; khai phá dữ liệu, nhiệm vụ của khai phá dữ liệu, cách hướng tiếp cận và kĩ thuật áp dụng trong khai phá dữ liệu, cũng như là ứng dụng của khai phá dữ liệu trong thực tế

CHƯƠNG II: Phân cụm dữ liệu và các thuật tóan phân cụm dữ liệu

Chương này nghiên cứu về phân cụm dữ liệu; một số kiểu dữ liệu; các độ đo khoảng cách; các hướng tiếp cận phân cụm dữ liệu và một số thuật tóan phân cụm dữ liệu.

CHƯƠNG III: Ứng dụng thuật tóan k-means trong phân đoạn ảnh

Chương này nghiên cứu tổng quan về phân đoạn ảnh; các phương pháp phân đoạn ảnh; một số thuật tóan phân đoạn ảnh; nghiên cứu thuật tóan k-means trong phân đoạn ảnh và giao diện chương trình cài đặt mô phỏng thuật toán k-means trong phân đoạn ảnh.

CHƯƠNG I: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

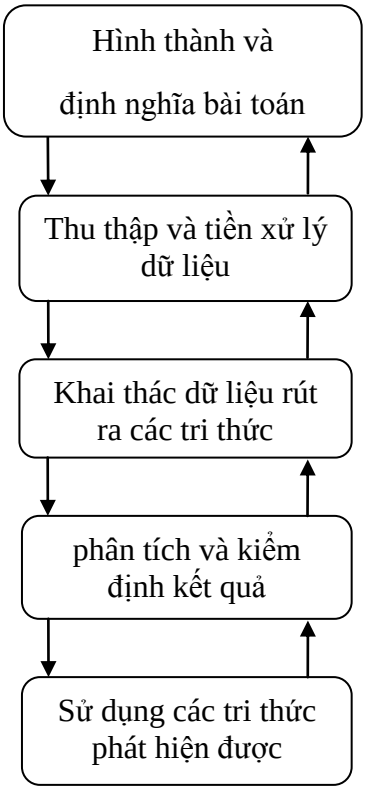
1.1. Giới thiệu về khám phá tri thức

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

Thông thường chúng ta coi dữ liệu là một dãy các bit, hoặc các số và các kí hiệu, hoặc "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 dưa 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 thông tin và các mối quan hệ. 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 cơ sở dữ liệu là quy 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 phá dữ liệu là một bước trong quy trình khám phá tri thức, gồm các thuật toán khai phá dữ liệu chuyên dùng dưới một số quy 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 khám phá tri thức như sau:



Hình 1: Quy trình phát hiện tri thức

- **Bước 1**: Tìm hiểu lĩnh vực ứng dụng và hình thành bài toán, bước này sẽ quyết định cho việc rút ra 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 2**: Thu thập và xử lý thô, được gọi là tiền xử lý dữ liệu để loại bỏ nhiễu, xử lý việc thiếu dữ liệu, biến đổi dữ liệu và rút gọn dữ liệu cần thiết, bước này thường chiếm thời gian nhất trong toàn bộ quy trình của khám phá tri thức.

- **Bước 3**: Là khai phá dữ liệu hay nói cách khác là trích ra các mẫu hoặc các mô hình ẩn dưới các dữ liệu.

- **Bước 4**: 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 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 quy 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 hóa các kết quả tìm được bằng cách áp dụng các mẫu đã phát hiện cho các tập con mới của dữ liệu. Quy trình này gồm 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

Khoảng hơn một thập kỷ trở lại đây, lượng thông tin được lưu trữ trên các thiết bị điện tử không nhừng tăng lên. Sự tích lũy dữ liệu này xảy ra với một tốc độ bùng nổ.Câu hỏi đặt ra là chúng ta có thể khai thác gì từ "núi" dữ liệu khổng lồ ấy? Và từ đó khái niệm "khai phá dữ liệu " đã ra đời.

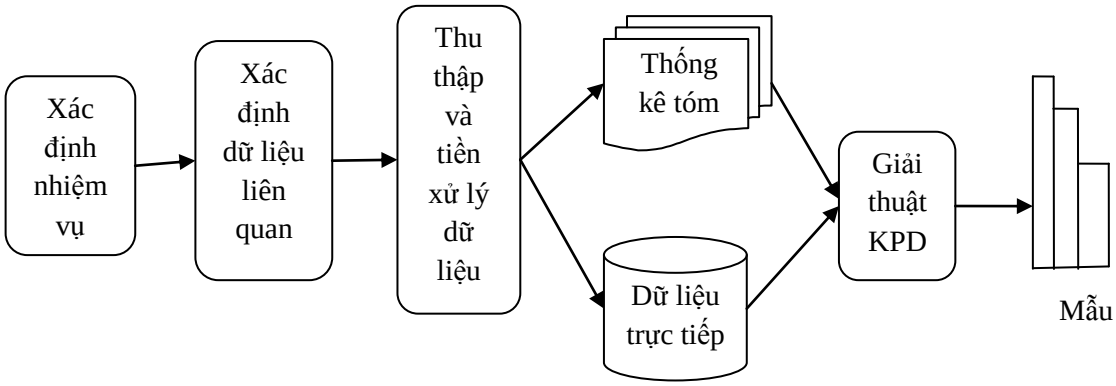
Khai phá dữ liệu được dùng để mô tả quá trình phát hiện ra tri thức trong CSDL. Quá trình này kết xuất ra các tri thức tiềm ẩn từ dữ liệu giúp cho việc dự báo trong kinh doanh, các hoạt động sản xuất, …Khai phá dữ liệu làm giảm chi phí về thời gian so với phương pháp truyền thống trước kia.Vậy "khai phá dữ liệu là gì"?

Khai phá dữ liệu là quá trình trợ giúp quyết định, trong đó chúng ta khám phá các mẫu thông tin có ích, chưa biết và bất ngờ trong CSDL lớn.

Khai phá dữ liệu là một bước chính quan trọng và mang tính quyết định trong quá trình KDD.

1.2.2. Các bước trong quá trình khai phá dữ liệu

Quá trình khai phá dữ liệu gồm các bước như sau:



Hình 2: Các bước trong khai phá dữ liệu

- Xác định nhiệm vụ: Xác định chính xác các vấn đề cần giải quyết.
- Xác định các dữ liệu liên quan dùng để xây dựng giải pháp giải quyết nhiệm vụ bài toán.
- Thu thập các dữ liệu có liên quan và xử lý chúng thành dạng sao cho giải thuật khai phá dữ liệu có thể hiểu được.
- Chọn thuật toán khai phá dữ liệu thích hợp và thực hiện việc khai phá nhằm tìm được các mẫu có ý nghĩa dưới dạng biểu diễn tương ứng với các ý nghĩa đó. Đặc điểm của mẫu phải là mới (ít nhất là đối với hệ thống đó). Độ mới có thể được đo tương ứng với độ thay đổi trong dữ liệu (bằng cách so sánh các giá trị hiện tại với các giá trị trước đó hoặc các giá trị mong muốn), hoặc bằng tri thức (mối liên hệ giữa phương pháp tìm mới và phương pháp cũ như thế nào). Thường thì độ mới của mẫu được đánh giá bằng một hàm logic hoặc một hàm đo độ mới, độ bất ngờ của mẫu. Ngoài ra, mẫu còn phải có khả năng sử dụng tiềm tàng. Các mẫu này sau khi được xử lý và diễn giải phải dẫn đến những hành động có ích nào đó được đánh giá bằng một hàm lợi ích. Ví dụ như trong dữ liệu các khoản vay, hàm lợi ích đánh giá khả năng tăng lợi nhuận từ các khoản

vay. Mẫu khai thác được phải có giá trị đối với các dữ liệu mới với độ chính xác nào đó.

1.2.3. Các thành phần trong khai phá dữ liệu

Giải thuật khai phá dữ liệu bao gồm 3 thành phần chính như sau: biểu diễn mô hình, kiểm định mô hình và phương pháp tìm kiếm.

- Biểu diễn mô hình: Mô hình được biểu diễn bằng một ngôn ngữ sao cho có thể khai phá được. Nếu mô hình có sự mô tả hạn chế thì sẽ không thể học được hoặc sẽ không thể có các mẫu tạo ra. Nếu diễn tả mô hình càng lớn thì càng làm tăng mức độ nguy hiểm do bị học quá nhiều và làm giảm đi khả năng dự đoán các dữ liệu chưa biết. Hơn nữa, việc tìm kiếm sẽ càng trở nên phức tạp hơn và việc giải thích mô hình cũng khoa khăn hơn.
- Kiểm định mô hình: Đánh giá xem một mẫu có đáp ứng được các tiêu chuẩn của quá trình phát hiện tri thức hay không. Việc đánh giá mô hình được thực hiện thông qua kiểm tra dữ liệu, đối với nhiệm vụ dự đoán thì việc đánh giá mô hình ngoài kiểm tra dữ liệu còn dựa trên độ chính xác dự đoán mà việc đánh giá độ chính xác dự đoán dựa trên đánh giá chéo.
- Tìm kiếm mô hình: Bao gồm tìm kiếm theo số và tìm kiếm theo mô hình. Cụ thể như sau:

    - Tìm kiếm theo số:Giải thuật cần tìm các tham số để tối ưu hoá các tiêu chuẩn đánh giá mô hình với các dữ liệu quan sát được và với một miêu tả mô hình đã định.
    - Tìm kiếm mô hình: Quá trình này xảy ra giống như một vòng lặp qua phương pháp tìm kiếm tham số. Khi miêu tả, mô hình bị thay đổi tạo nên một họ các mô hình, với mỗi một miêu tả mô hình phương pháp tìm kiếm tham số được áp dụng để đánh giá chất lượng mô hình. Các phương pháp tìm kiếm mô hình thường sử dụng các kỹ thuật tìm kiếm heuristic bởi kích thước của không gian các mô hình có thể ngăn cản các tìm kiếm tổng thể.

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

Khai phá dữ liệu là một chuyên ngành rất rộng và có rất nhiềuhướng nghiên cứu (bài toán) khác nhau. Tuy nhiên, chúng đượctiếp cận theo các hướng chính như sau:

- **Phân lớp và dự đoán (Học có giám sát)**: Phân lớp dữ liệu là việc xây dựng một mô hình mà có thể phân cácđối tượng thành những lớp để dự đoán giá trị bị mất tại một sốthuộc tính của dữ liệu hay tiên đoán giá trị của dữ liệu sẽ xuất hiệntrong tương lai.
- **Phân cụm**: Phân cụm dữ liệu là kỹ thuật khai phá dữ liệu tương tự như phân lớp dữ liệu. Tuy nhiên, phân cụm dữ liệu là quá trình học không giám sát, là quá trình nhóm những đối tượng vào các lớp tương ứng để sao cho các đối tượng trong một nhóm là tương đương nhau, chúng khác so với các đối tượng của nhóm khác.
- **Luật kết hợp**: Là quá trình khám phá các tập giá trị thuộc tính xuất hiện phổ biến trong các đối tượng dữ liệu. Từ tập phổ biến có thể tạo ra các luật kết hợp giữa các giá trị thuộc tính trong tập các đối tượng.
- **Khai phá chuỗi theo thời gian**:Phân tích chuỗi được sử dụng để tìm mẫu trong tập rời rạc. Chuỗi được tạo thành từ tập các giá trị rời rạc. Phân tích chuỗi theo thời gian và khai phá luật kết hợp là tương tự nhau nhưng có thêm tính thứ tự và thời gian.
- **Phân tích ngoại lệ**: Phân tích ngoại lệ cũng là một dạng của phân cụm, nó tập trung vào các trường hợp rất khác biệt so với các trường hợp khác. Đôi khi nó thể hiện những lỗi trong dữ liệu hoặc thể hiện phần thú vị nhất trong dữ liệu đó.
- **Hồi quy**: Phương pháp hồi quy được sử dụng để đưa ra các dự báo dựa trên các dữ liệu đang tồn tại bằng cách áp dụng các công thức. Một hàm sẽ được học ra từ bộ dữ liệu hiện có bằng cách sử dụng các kỹ thuật hồi quy và tuyến tính từ việc thống kê. Sau đó, dữ liệu mới sẽ căn cứ vào hàm này để đưa ra những dự đoán.

1.2.5. Ứng dụng của khai phá dữ liệu

Hiện nay, kỹ thuật khai phá dữ liệu đang được ứng dụng một cách rộng rãi trong rất nhiều lĩnh vực kinh doanh và đời sống khác nhau như marketing, tài chính, ngân hàng và bảo hiểm, khoa học, y tế, an ninh, internet,…

- Y học và chăm sóc sức khỏe: Chuẩn đoán bệnh trong y tế dựa trên kết quả xét nghiệm…
- Tài chính và thị trường chứng khoán: Áp dụng vào phân tích các thẻ tín dụng tiêu biểu của khách hàng, phân đoạn tài khoản nhận được, phân tích đầu tư tài chính cũng như chứng khoán, giấy chứng nhận và các quỹ tình thương, đánh giá tài chính, phát hiện kẻ gian… Dự báo giá của các loại cổ phiếu trong thị trường chứng khoán…
- Bảo hiểm: Áp dụng vào việc phân tích mức độ rủi ro xảy ra đối với từng loại hàng hóa, dịch vụ hay chiến lược tìm kiếm khách hàng mua bảo hiểm…
- Quá trình sản xuất: Các ứng dụng giải quyết sự tối ưu của các nguồn tài nguyên như máy móc, nhân sự và nguyên vật liệu, thiết kế tối ưu trong quá trình sản xuất, bố trí phân xưởng và thiết kế sản phẩm, chẳng hạn như quá trình tự động dựa vào yêu cầu khách hàng…
- Thiên văn học: Quan sát chú trọng tới việc thu thập và phân tích dữ liệu, sử dụng các nguyên tắc cơ bản của vật lý. Thiên văn học lý thuyết định hướng theo sự phát triển các mô hình máy tính hay mô hình phân tích để miêu tả các vật thể và hiện tượng thiên văn. Hai lĩnh vực bổ sung lẫn cho nhau, thiên văn học lý thuyết tìm cách giải thích các kết quả quan sát, và việc quan sát lại thường được dùng để xác nhận các kết quả lý thuyết.
- Thể thao, giải trí
- Viễn thông
- Máy tìm kiếm
- Quảng cáo: Phân tích, trích trọn những đặc trưng…

CHƯƠNG II
PHÂN CỤM DỮ LIỆU VÀ CÁCTHUẬT TOÁN PHÂN CỤM DỮ LIỆU

2.1. Phân cụm dữ liệu

Phân cụm dữ liệu là một trong những hướng nghiên cứu trọng tâm củalĩnh vực khai phá dữ liệu (Data Mining) và lĩnh vực khám phá tri thức.

2.1.1. Định nghĩa về phân cụm dữ liệu

Chúng ta thấy rằng, mục đích của phân cụm là nhóm các đối tượng vào các cụm sao cho các đối tượng trong cùng một cụm có tính tương đồng cao và độ bất tương đồng giữa các cụm 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 phân chia một tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong một cụm "tương tự" với nhau và các phần tử trong các cụm khác nhau sẽ "phi tương tự" 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 của phương pháp phân cụm.

Sau khi xác định các đặc tính của dữ liệu, người ta đi tìm cách thích hợp để xác định khoảng cách giữa các đối tượng, hay là phép đo tương tự dữ liệu. Đây chính 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. Giá trị của hàm tính độ đo tương tự càng lớn thì sự giống nhau giữa các đối tượng dữ liệu càng lớn và ngược lại, còn hàm tính độ phi tương tự thì tỉ lệ nghịch với độ tương tự.

Trong quá trình phân cụm dữ liệu thì vấn đề trở ngại lớn nhất đó là nhiễu (noise). Nhiễu xuất hiện do quá trình thu thập thông tin, dữ liệu thiếu chính xác hoặc không đầy đủ. Vì vậy chúng ta phải khử nhiễu trong quá trình phân cụm dữ liệu.

Các bước chính trong quá trình phân cụm dữ liệu:

- Xây dựng hàm tính độ tương tự.
- Xây dựng các tiêu chuẩn phân cụm.
- Xây dựng mô hình cho cấu trúc cụm dữ liệu
- Xây dựng thuật toán phân cụm và các xác lập các điều kiện khởi tạo.
- Xây dựng các thủ tục biểu diễn và đánh giá kết quả phân cụm

Phân cụm dữ liệu là bài toán thuộc vào lĩnh vực học máy không giám sát và đang được ứng dụng rộng rãi để khai thác thông tin từ dữ liệu

2.1.2. Một số ví dụ về phân cụm dữ liệu

Phân cụm dữ liệu có thể được ứng dụng trong nhiều lĩnh vực của cuộc sống ví dụ như:

- Thương mại: Tìm kiếm nhóm các khách hàng quan trọng có đặc trưng tương đồng và những đặc tả họ từ các bản ghi mua bán trong cơ sở dữ liệu khác hàng;

- Phân cụm dữ liệu phục vụ cho biểu diễn dữ liệu gene: Phân cụm là một trong những phân tích được sử dụng thường xuyên nhất trong biểu diễn dữ liệu gene. Dữ liệu biểu diễn gene là một tập hợp các phép đo được lấy từ **DNA microarray** là một tấm thủy tinh hoặc nhựa trên đó có gắn các đoạn DNA thành các hàng siêu nhỏ. Một tập hợp dữ liệu biểu diễn gene có thể được biểu diễn thành một ma trận giá trị thực

Dữ liệu biểu diễn gene sẽ được phân cụm theo 2 cách. Cách thứ nhất là nhóm các mẫu gene giống nhau ví dụ như gom cụm dòng của ma trận D. Cách thứ 2 là nhóm các mẫu khác nhau trên các hồ sơ tương ứng, ví dụ như gom các cột của ma trận D.

- Phân cụm dữ liệu phục vụ trong sức khỏe tâm lý: Phân cụm dữ liệu áp dụng trong nhiều lĩnh vực sức khỏe, tâm lý, bao gồm cả việc thúc đẩy và duy trì sức khỏe, cải thiện cho hệ thống chăm sóc sức khỏe và công tác

phòng chống bệnh tật và người khuyết tật. Trong sự phát triển của hệ thống chăm sóc sức khỏe, phân cụm dữ liệu được sử dụng để xác định các nhóm của người dân mà có thể được hưởng lợi từ các dịch vụ cụ thể. Trong thúc đẩy y tế, nhóm phân tích được lựa chọn để nhằm mục tiêu vào nhóm sẽ có khả năng mang lại lợi ích cho sức khỏe cụ thể từ các chiến dịch quảng cáo và tạo điều kiện thuận lợi cho sự phát triển của quảng cáo. Ngoài ra, phân cụm dữ liệu còn được sử dụng để xác định các nhóm dân cư bị rủi ro do phát triển y tế và các điều kiện những người có nguy cơ nghèo.

- Phân cụm dữ liệu trong hoạt động nghiên cứu thị trường: Trong nghiên cứu thị trường phân cụm dữ liệu được sử dụng để phân đoạn thị trường và xác định mục tiêu thị trường. Trong phân đoạn thị trường, phân cụm dữ liệu được dùng để phân chia thị trường thành những cụm mang ý nghĩa. Chẳng hạn như chia đối tượng nam giới từ 21 – 30 tuổi và nam giới ngoài 51 tuổi, đối tượng nam giới ngoài 51 tuổi thường không có xu hướng mua những sản phẩm mới.

- Phân cụm dữ liệu trong hoạt động phân đoạn ảnh: Phân đoạn ảnh là việc phân tích mức xám hay mầu của ảnh thành lát đồng nhất. Trong phân đoạn ảnh phân cụm dữ liệu thường được dùng để phát hiện biên của đối tượng trong ảnh.

Vấn đề phân cụm dữ liệu được quan tâm một cách rộng rãi, mặc dù chưa có định nghĩa đồng bộ về phân cụm dữ liệu. Nói một cách đại khái, phân cụm dữ liệu nghĩa là ta cho một tập dữ liệu và một phương pháp tương tự, chúng ta nhóm dữ liệu lại chẳng hạn như điểm dữ liệu trong cùng một nhóm giống nhau và điểm dữ liệu trong các nhóm khác nhau về sự không đồng dạng. Rõ ràng là vấn đề này được bắt gặp trong nhiều ứng dụng, chẳng hạn như khai phá văn bản, biểu diễn gene, phân loại khách hàng, xử lí ảnh...

2.2. Một số kiểu dữ liệu trong phân cụm

Trong phân cụm các đối tượng dữ liệu thường được diễn tả dưới dạng các đặc tính (hay còn gọi là thuộc tính). Các thuộc tính này là các tham số để giải quyết vấn đề phân cụm và lựa chọn chúng có tác động đáng kể đến kết qủa phân cụm. Phân loại các thuộc tính khác nhau là vấn đề cần giải quyết đối với hầu hết các tập dữ liệu nhằm cung cấp các phương tiện thuận lợi để nhận dạng sự khác nhau của các phần tử dữ liệu. Các thuật toán phân cụm thường sử dụng một trong hai cấu trúc dữ liệu sau:

1. Ma trận dữ liệu: Là mảng n hàng, p cột trong đó p là số thuộc tính của đối tượng, các phần tử trong mỗi hàng chỉ giá trịthuộc tính tương ứng của đối tượng đó. Mảng được cho như sau:

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

2. Ma trận phi tương tự: Là ma trận n hàng, n cột, phần tử d(i,j) chứa khoảng cách hay độ khác biệt giữa đối tượng i,j; d(i,j) là một số không âm trong đó nếu d(i,j) xấp xỉ bằng 0 thì đối tượng i và j khá gần nhau, nếu d(i,j) càng lớn thì 2 đối tượng i và j khá khác nhau. Do đó d(i,j)=d(j,i)=0 nên ta biểu diễn ma trận này như sau:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Phần lớn các thuật toán phân cụm dữ liệu sử dụng cấu trúc phi tương tự. Do vậy, nếu dữ liệu cần phân cụm được tổ chức dưới dạng ma trận dữ liệu thì phải biến đổi về dạng ma trận phi tương tự trước khi tiến hành phân cụm dữ liệu.

Có 2 đặc trưng để phân loại: kích thước miền và hệ đo. Cho một cơ sở dữ liệu D chứa n đối tượng trong không gian k chiều; x, y, z là các đối tượng thuộc D, với $x=(x_1, x_2,...x_k)$; $y=(y_1, y_2,...y_k)$; $z=(z_1, z_2,...z_k)$; trong đó x_i, y_i, z_i với i=1...k là các đặc trưng hoặc các thuộc tính tương ứng của các đối tượng x, y, z. Như vậy nó sẽ có các kiểu dữ liệu sau:

2.2.1. Kiểu dữ liệu dựa trên kích thước miền

- **Thuộc tính liên tục:** Nếu miền giá trị của nó là vô hạn không đếm được, nghĩa là giữa 2 giá trị tồn tại vô số giá trị khác (ví dụ các thuộc tính màu, cường độ, nhiệt độ, âm thanh).

- **Thuộc tính rời rạc:** Nếu miền giá trị của nó là tập vô hạn đếm được (ví dụ là các thuộc tính số, ...) trường hợp đặc biệt của thuộc tính rời rạc là thuộc tính nhị phân mà miền giá trị chỉ có 2 phần tử (yes/no, true/false, on/off).

2.2.2. Kiểu dữ liệu dựa trên hệ đo

- **Thuộc tính định danh:** Là dạng thuộc tính khái quát hóa của thuộc tính nhị phân, trong đó có miền giá trị là rời rạc không phân biệt thứ tự và có nhiều hơn 2 phần tử. Nếu x và y là 2 đối tượng thuộc tính thì chỉ có thể xác định x=y hay x<>y.

- **Thuộc tính có thứ tự:** Là thuộc tính định danh nhưng có thêm tính thứ tự nhưng chúng không được định lượng. Nếu x và y là 2 thuộc tính thứ tự thì có thể xác định là x=y, x<>y, x>y, x<y.

- **Thuộc tính khoảng:** Để đo các giá trị theo xấp xỉ tuyến tính, với thuộc tính khoảng có thể xác định một thuộc tính là đứng trước hay đứng sau thuộc tính khác với khoảng là bao nhiêu. Nếu

$x_i > y_i$ thì có thể nói x cách y 1 khoảng là $x_i - y_i$ tương ứng với thuộc tính thứ i.

Việc chọn lựa đơn vị đo cho các thuộc tính cũng ảnh hưởng đến chất lượng phân cụm. Nếu đơn vị đo của các thuộc tính càng được chia nhỏ thì khoảng cách xác định của thuộc tính đó càng lớn và ảnh hưởng nhiều hơn đến kết quả phân cụm. Để tránh phụ thuộc vào việc lựa chọn đơn vị đo, thì dữ liệu cần được chuẩn hóa. Việc chuẩn hóa sẽ gán cho tất cả các thuộc tính 1 trọng số bằng nhau.Tuy nhiên trong nhiều trường hợp người sử dụng có thể thay đổi trọng số cho các thuộc tính ưu tiên.

Để chuẩn hóa các độ đo, 1 cách làm phổ biến là biến đổi các thuộc tính về dạng không có đơn vị đo. Giả sử đối với thuộc tính f ta thực hiện như sau:

+ Tính độ lệnh trung bình:

$$S_f = \frac{1}{n}\left(\left|x_{1f} - m_f\right| + \left|x_{2f} - m_f\right| + \ldots + \left|x_{nf} - m_f\right|\right)$$

Trong đó $x_{1f}...x_{nf}$ là các giá trị thuộc tính f của n phần tử dữ liệu và m_f là giá trị trung bình của f, được cho như sau:

$$m_f = \frac{1}{n}\left(x_{1f} + x_{2f} + \ldots + x_{nf}\right)$$

+ Độ đo được chuẩn hóa:

$$z_{if} = \frac{x_{if} - m_f}{S_f}$$

- **Thuộc tính nhị phân**: Là thuộc tính có 2 giá trị là 0 và 1.
- **Thuộc tính tỷ lệ**: Là thuộc tính khoảng nhưng được xác định một cách tương đối so với điểm mốc.

Trong các thuộc tính được trình bày ở trên thuộc tính định danh và thuộc tính thứ tự gọi chung là thuộc tính hạng mục; còn thuộc tính khoảng cách và thuộc tính tỉ lệ được gọi là thuộc tính số.

2.3. Phép đo độ tương tự và khoảng cách đối với các kiểu dữ liệu

2.3.1. Khái niệm tương tự và phi tương tự

Khi các đặc tính của dữ liệu được xác định, người ta 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 phitương tự (Dissimilar)* giữa các đối tượng dữ liệu. Giá trị của hàm tính độ đo tương tự càng lớn thì sự giống nhau giữa đối tượng càng lớn và ngược lại, còn hàm tính độ phi tương tự tỉ lệ nghịch với hàm tính độ tương tự. Độ tương tự hoặc độ phi tương tự có nhiều cách để xác định, chúng thường được đo bằng khoảng cách giữa các đối tượng. Tất cả các cách đo độ tương tự đều phụ thuộc vào kiểu thuộc tính mà chúng ta phân tích. Thí dụ, đối với thuộc tính hạng mục (*Categorical*) người ta không sử dụng độ đo khoảng cách mà sử dụng một hướng hình học của dữ liệu.

Tất cả các độ đo dưới đây được xác định trong không đo gian metric. Bất kỳ một metric nào cũng là một độ đo, nhưng điều ngược lại không đúng. Để tránh sự nhầm lẫn, thuật ngữ độ đo ở đây đề cập đến hàm tính độ *tương tự* hoặc hàm tính *độ phi tương tự*. 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 CSDL 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 d(x, y), được gọi là khoảng cách giữa x và y.
- Quy tắc trên thoả mãn hệ tính chất sau:
  i. d(x,y) > 0 nếu x≠ y;
  ii. d(x, y) = 0 nếu x= y;
  iii. d(x, y) =d(y, x) với mọi x, y;
  iv. d(x, y) ≤d(x, z) + d (z, y);

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

2.3.2. Độ đo khoảng cách

✓ ***Thuộc tính khoảng***: Sau khi chuẩn hoá, độ đo phi tương tự của hai đối tượng dữ liệu x, y được xác định bằng các metric khoảng cách như sau:

- Khoảng cách Minskowski: $d\ x,y\ =(\sum_{i=1}^{n}|x_i-y_i|^{\ q})^{1/q}$, với q là 1 số nguyên dương.
- Khoảng cách Euclide: $d\ x,y\ =\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$, (Trường hợp đặc biệt của Minskowski trong trường hợp q=2).
- Khoảng cách Manhattan: $d\ x,y\ =\sum_{i=1}^{n}|x_i-y_i|$, (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 của khoảng cách Minskowski trong trường hợp $q\to\infty$)

✓ ***Thuộc tính nhị phân:*** Trước hết chúng ta có xây dựng bản tham số sau:

	y:1	y:0	
x:1	A	β	α+β
x:0	Γ	δ	γ+δ
	α+γ	β+δ	τ

Trong đó: τ=α+β+γ+δ, các đối tượng x, y mà tất cả các thuộc tính của nó đều là nhị phân biểu thị bằng 0 và 1. Bảng trên cho ta các thông tin sau:

- α là tổng số các thuộc tính có giá trị là 1 trong cả 2 đối tượng x,y.

- β là tổng số các giá trị thuộc tính có giá trị là 1 trong x và 0 trong y.

- γ là tổng số các giá trị thuộc tính có giá trị 0 trong x và 1 trong y.

- δ là tổng số các giá trị thuộc tính có giá trị 0 trong x và y.

Các phép đo độ tương tự 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 2 đối tượng x và y có vai trò như nhau, nghĩa là chúng đối xứng và có trọng số.

- Hệ số Jacard: $d\ x,y\ = \frac{\alpha}{\alpha+\beta+\gamma}$, tham số này 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 giá trị 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 thuộc tính đối sánh tương ứng trùng nhau và p là tổng số các thuộc tính.

✓ **Thuộc tính có thứ tự:**Phép đo độ phi tương tự giữa các đối tượng dữ liệu với thuộc tính thứ tự được thực hiện như sau, ở đây ta giả sửi là thuộc tính thứ tự có$M_i$giá trị ($M_i$kích thước miền giá trị): Các trạng thái M_iđược sắp thứ tự như sau: $[1\ldots M_i]$, ta có thể thay thế mỗi giá trị của thuộc tính bằng giá trị cùng loại r_i, với $r_i \in 1...M_i$. Mỗi thuộc tính thứ tự có miền giá trị khác nhau, vì vậy ta chuyển đổi chúng về miền giá trị [0,1] băng cách

thực hiện phép biến đổi sao cho mỗi thuộc tính: $z_i^{(j)} = \frac{r_i^{(j)} - 1}{M_i - 1}$, với i= 1…$M_i$.Sử dụng công thức tính độ phi tương tự của thuộc tính khoảng cách đối với các giá trị $z_i^{(j)}$, đây cũng chính là độ phi tương tự của thuộc tính có giá trị.

- ✓ **Thuộc tính tỉ lệ:** Có nhiều cách khác nhau để tính độ tương tự giữa cácthuộc tính tỉ lệ. Một trong những số đó là sử dụng công thức tính logarit cho mỗithuộc tính x_i, thí dụ$q_i = \log(x_i)$, lúc nàyq_i đóng vai trò như thuộc tính khoảng. Phép biến đổi logarit này thích hợp trong trường hợp các giá trị của thuộc tính là số mũ.

Trong thực tế, khi tính độ đo tương tự dữ liệu, người ta chỉ xem xét một phần các thuộc tính đặc trưng đối với các kiểu dữ liệu hoặcđánh trọng số chocho tất cả các thuộc tính dữ liệu. Trong một số trường hợp, người ta loại bỏ đơnvị đo của các thuộc tính dữ liệu bằng cách chuẩn hoá chúng hoặc gán trọng số cho mỗi thuộc tính giá trị trung bình, độ lệch chuẩn. Các trọng số này có thể sửdụng trong các độ đo khoảng cách trên, thí dụ với mỗi thuộc tính dữ liệu đã được gán trọng số tương ứngw_i ($1 \le i \le k$), độ tương tựdữ liệu được xác địnhnhư sau: $d(x,y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}$.

Người ta có thể chuyển đổi giữa các mô hình cho các kiểu dữ liệu trên, thídụ dữ liệu kiểu hạng mục có thể chuyển đổi thành dữ liệu nhị phân và ngược lại. Nhưnggiải pháp này rất tốt kém về chi phí tính toán, cần phải cân nhắc khi ápdụng cách thức này.

Tuỳ từng trường hợp dữ liệu cụ thể mà người ta sử dụng các mô hình tính độ tương tự khác nhau. Việc xác định độ tương tựdữ liệu thích hợp, chính xác, đảm bảo khách quan là rất quan trọng vàgóp phần xây dựng thuật toán PCDL có hiệu quả cao trong việc đảm bảo chất lượng cũng như chi phí tính toán của thuật toán.

2.4. Các hướng tiếp cận của bài toán phân cụm dữ liệu

Các phương pháp phân cụm được chia thành các nhóm: phương pháp phân hoạch, phương pháp phân cấp, phương pháp dựa trên mật độ, phương pháp dựa trên lưới, phân cụm dựa trên mô hình, phân cụm dựa trên ràng buộc.

2.4.1. Phương pháp phân cụm phân hoạch

Phương pháp phân cụm phân hoạchnhằm phân một tập dữ liệu có n phần tử cho trước trong cơ sở dữ liệu D thành K nhóm dữ liệu sao cho:

- Mỗi cụm chứa ít nhất một đối tượng.
- Mỗi đối tượng thuộc về một cụm duy nhất.
- K là số cụm đã được cho trước.

Các thuật toán phân hoạchdữ 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 hoặch có thể được.

Một số thuật toán phân cụm phân hoạch điển hình như: K - MEANS, PAM, CLARA, CLARANS ....

2.4.2. Phương pháp phân cụm phân cấp

Phân cụm dữ liệu phân cấp 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. Cây phân cấp có thể được xây dựng theo 2 phương pháp tổng quát: Phương pháp trên xuống (Topdown) và phương pháp dưới lên (Bottom up).

Đây là các phương pháp tạo phân cấp cụm chứ không tạo phân hoạch các đối tượng. Phương pháp này không càn phải xác định số cụm từ đầu. Số cụm sẽ do khoảng cách giữa các cụm hoặc điều kiện dừng quyết định. Tiêu chuẩn gom cụm thường được xác định bởi ma trận khoảng cách. Phân cấp cụm thường được biểu diễn dưới dạng đồ thị dạng cây các cụm. Lá của cây biểu diễn đối tượng riêng lẻ, nút trong biểu diễn các cụm.

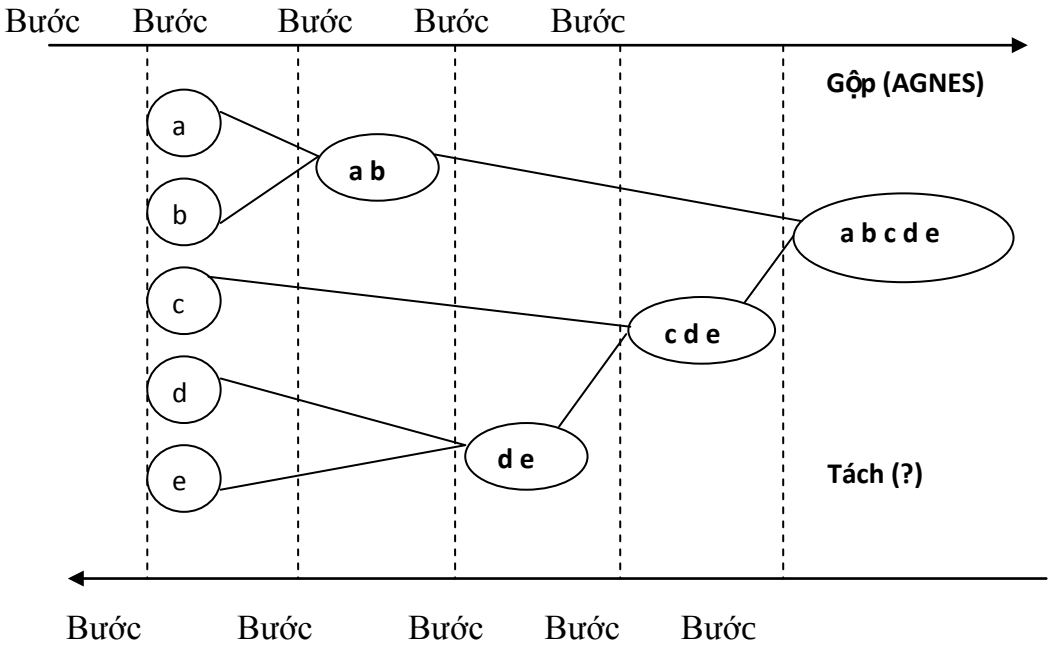
Các phương pháp tiếp cận để gom cụm phân cấp gồm:

Gộp:

- Xuất phát mỗi đối tượng và tạo một cụm chứa nó.
- Nếu hai cụm đủ gần nhau sẽ được gộp lại thành một cụm duy nhất.
- Lặp lại bước 2 đến khi chỉ còn một cụm duy nhất là toàn bộ không gian.

Tách:

- Xuất phát từ cụm duy nhất là toàn bộ không gian
- Chọn cụm có độ phân biệt cao nhất để tách đôi. Bước này sẽ áp dụng các phương pháp phân hoạch đối với cụm đã chọn.
- Lặp lại bước 2 đến khi mỗi đối tượng thuộc một cụm hoặc đạt điều kiện dừng.



Hình 3: Hai phương pháp tiếp cận phân cấp

Các khoảng cách giữa các cụm thường dùng là:

- *Khoảng cách nhỏ nhất*: Khoảng cách này thường được gọi là khoảng cách liên kết đơn hoặc khoảng cách láng giềng gần nhất. Đây là loại khoảng cách phù hợp để phát hiện các cụm có dạng chuỗi hơn là dạng khối.

$$d(C_i, C_j) = \min_{x \in Ci,\, y \in Cj}\{d(x,y)\}$$

- *Khoảng cách lớn nhất*: Khoảng cách này còn được gọi là khoảng cách liên kết hoàn toàn hoặc khoảng cách láng giềng xa nhất. Đây là loại khoảng cách phù hợp nhất để phát hiện các cụm có dạng khối hơn là dạng chuỗi.

$$d(C_i,C_j) = \max_{x \in Ci,\, y \in Cj}\{d(x,y)\}$$

- *Khoảng cách trung bình:*

$$d(C_i,C_j) = \text{avg}_{x \in Cj,\, y \in Cj}\{d(x,y)\}$$

- *Khoảng cách trọng tâm:* Khoảng cách giữa hai trọng tâm của hai cụm được chọn làm khoảng cách của hai cụm đó. Khoảng cách phù hợp để phát hiện các cụm có dạng khối và tốc độ tính toán nhanh do chỉ quan tâm đến trọng tâm nên giảm khối lượng tính toán.

Một số thuật toán phân cụm phân cấp điển hình như: CURE, BIRCH,…

2.4.3. 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.

Các kí hiệu và khái niệm:

1. p, q, o là các điểm dữ liệu bất kỳ (các đối tượng).

2. Với Eps dương cho trước, tập hợp $NEps(p) = \{q \mid d(q,p) \le Eps\}$ được gọi là lân cận bán kính Eps của p.

3. p được gọi là điểm hạt nhân nếu thoả: $|\ NEps(p)| \ge \min Pts$

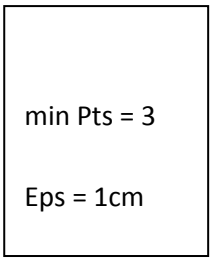
Trong đó min Pts: số nguyên dương cho trước, min Pts là ngưỡng tối thiểu để coi một điểm là trù mật. Từ đây khi nói một điểm là hạt nhân thì ta hiểu nó gắn với một bán kính và một ngưỡng trù mật nhất định

4. p được gọi là điểm biên nếu nó không phải là điểm nhân.

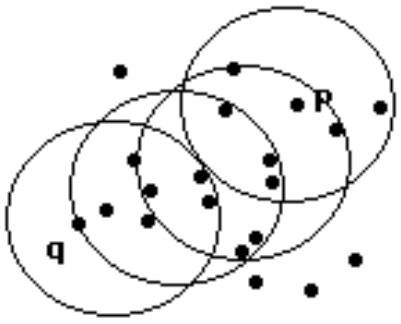
5. q được gọi là *đi tới được trực tiếp theo mật độ* từ p nếu p là một điểm nhân và q thuộc lân cận của p.

6. p_n được gọi là *đi tới được theo mật độ* từ p1 nếu tồn tại một dãy các điểm pi (i = 2, …n) sao cho pi liên thông mật độ trực tiếp từ p_{i+1}

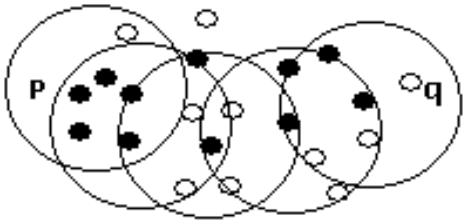
7. p và q được gọi là có *kết nối theo mật độ* nều tồn tại điểm o sao cho cả p và q đều liên thông mật độ từ o



Hình 4: p là một điểm hạt nhân với bán kính Eps 1cm và ngưỡng trù mật là min Pts là 3. Khoảng cách được dùng là khoảng cách Euclide trong không gian hình học hai chiều, q là một điểm liên thông mật độ trực tiếp từ p.



Hình 5: q là một điểm liên thông mật độ từ p



Hình 6: p và q là hai điểm có kết nối mật độ

Ý tưởng của các thuật toán dựa trên mật độ: Một cụm là một tập tối đại các điểm có kết nối mật độ.

Thuật toán DBSCAN có các bước sau:

1. Chọn một điểm p bất kỳ thuộc không gian dữ liệu D

2. Tìm tập P gồm tất cả các điểm liên thông mật độ từ p với ngưỡng bán kính Eps và ngưỡng mật độ Pts

3. Nếu p là một điểm hạt nhân thì

    a. P chính là một cụm cần tìm

    b. D = D \ P (loại P ra khỏi D)

4. Quay lại bước 1 cho đến khi tất cả các điểm trong D đều đã được xét.

5. Các điểm đã xét nhưng không thuộc cụm nào thì chính là các mẫu cá biệt.

Ưu điểm của DBSCAN là tìm được các cụm từ có hình dạng bất kỳ do nhiễu hoặc mẫu khác biệt gây ra.

Khuyết điểm của DBSCAN là khó chọn được các ngưỡng EPs và min Pts tốt. Do đó kết quả gom cụm không tốt khi mật độ trong các cụm tự nhiên là chênh lệch nhau nhiều. Một điểm yếu nữa là không phù hợp cho yêu cầu phân cấp cụm mà chỉ đáp ứng nhu cầu phân hoạch.

Bán kính lân cận và ngưỡng trù mật là các tham số quyết định đến kết quả gom cụm.Để có kết quả gom cụm tốt ta có thể thử với một số bộ tham số và chọn

ra kết quả tối ưu. Để tạo cây phân cấp cụm thì có thể áp dụng chiến lược phân giải tăng dần như sau:

1. Đầu tiên chọn bán kính lân cận và ngưỡng trù mật độ thô (Eps lớn và min Pts nhỏ);

2. Chọn cụm có độ phân biệt lớn nhất (thông qua ma trận phân biệt của cụm hoặc một tiêu chí đánh giá tuỳ thuộc nhu cầu ứng dụng). Cụm được chọn ở bước này sẽ tạo thành một nút của cây phân cấp;

3. Phân hoạch cụm được chọn bằng thuật toán DBSCAN;

4. Nếu tất cả các cụm tạo ra được đều có độ phân biệt nội tại đủ thấp hoặc đã đặt được số cụm cần thiết thì dừng. Các cụm còn lại tại thời điểm kết thúc thuật toán tạo thành các nút lá của cây phân cấp.

5. Giảm bán kính lân cận và tăng ngưỡng trù mật. Mức độ điều chỉnh tuỳ thuộc bản chất dữ liệu và nhu cầu gom cụm.

6. Quay lại bước 2.

Đặc điểm của phương pháp tạo cây phân cấp cụm dựa trên thuật toán DBSCAN có thể tạo cây đa phân.

Một số thuật toán PCDL dựa trên mật độ điển hình như: DPSCAN, OPTIS, DENCLUE….

2.4.4. Phương pháp phân cụm dựa trên lưới

Kỹ thuật phân cụm trên mật độ không thích hợp với dữ liệu nhiều chiều, để giải quyết cho vấn đề này, người ta sử 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.Ý tưởng: Dùng các cấu trúc dữ liệu dạng lưới với nhiều cấp độ phân giải. Những ô lưới có mật độ cao sẽ tạo thành những cụm. Phương pháp này rất phù hợp với các phân tích trong gom cụm ứng dụng trong không gian (phân loại sao, thiên hà…)

Một số thuật toán PCDL dựa trên cấu trúc lưới điển hình như sau: STING, WAVECLUSTER, CLIQUE…

2.4.5. Phương pháp phân cụm dựa trên mô hình

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 sác xuất cơ bản.Ý tưởng của các phương pháp này là: Dự liệu phát sinh từ một sự kết hợp nào đó của các phân phối xác suất ẩn. Có hai phương pháp tiếp cận chính:

- Tiếp cận thống kê (phương pháp COBWEB, CLASSIT, AutoClass).
- Tiếp cận mạng nơron học cạnh tranh, bản đồ tự cấu trúc SOM.

2.4.6. Phương pháp phân cụm dựa trên dữ liệu ràng buộc

Sự phát triển của phân cụm dữ liệu không gian trên cơ sở dữ liệu lớn đã cung cấp nhiều công cụ tiện lợi cho việc phân tích thông tin địa lý, 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 dữ liệu. Để phân cụm dữ liệu 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.

Tóm lại, các kỹ thuật phân cụm dữ liệu 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.

2.5. Một số thuật toán phân cụm dữ liệu

2.5.1. Các thuật toán phân cụm phân hoạch

- **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 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 cụm. Như vậy nó cần khởi tạo 1 tập trung tâm các trung tâm cụm ban đầu và thông qua đó nó lặp lại các bước gồm gán mỗi đối tượng tới các 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 k – means là sinh ra k cụm $\{C_1, C_2, \ldots C_k\}$, từ một tập dữ liệu chứa n đối tượng trong không gian d chiều $X_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ ($i = \overline{1..n}$), sao cho hàm tiêu chuẩn: $E = \sum_{i=1}^{k} \sum_{x \in C_i} D^2(x - m_i)$ đạtgiá trị tối thiểu. Trong đó: m_i là trọng tâm của cụm C_i, D là khoảng cách giữa hai đối tượng. Do k – means phân tích cụm đơn giản nên có thể áp dụng đối với dữ liệu lớn. Nhược điểm của nó 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.

- **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. Thuật toán được thực hiện qua các bước sau:

+ Chọn k đối tượng bất kỳ trong n đối tượng ban đầu làm các medoids ban đầu.

+ Lặp cho tới khi hội tụ: Gán mỗi đối tượng còn lại vào cụm có medoids gần nhất với nó. Thay thế medoids còn lại bằng một đối tượng không phải là medoids sao cho chất lượng phân cụm được cải thiện.

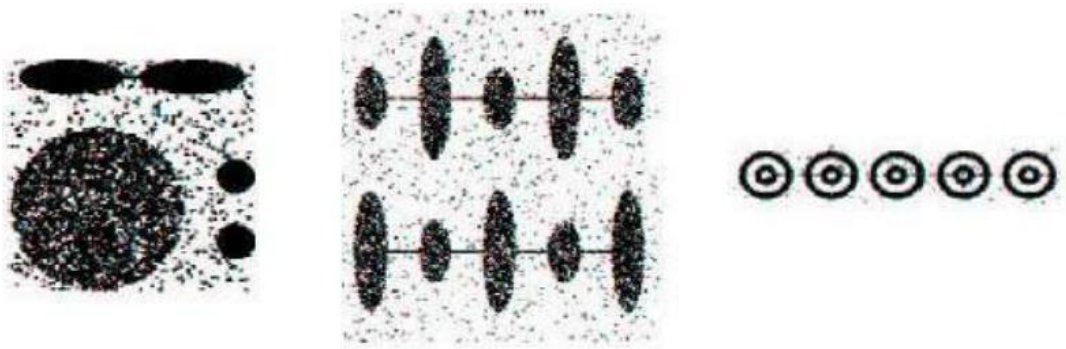
K – medoids tỏ ra hiệu quả hơn k – means trong trường hợp dữ liệu có nhiễu hoặc đối tượng ngoại lai. Nhưng so với k – means thì k – medoids có độ phức tạp tính toán cao hơn. Cả hai thuật toán đều có nhược điểm chung là số lượng k cụm cần được người dùng cung cấp ngay từ đầu.

Ngoài ra còn có các thuật toán phân cụm phân hoạch khác: PAM, CLARA

2.5.2. Thuật toán phân cụm phân cấp

Thuật toán phân cụm phân cấp khắc phục được nhược điểm của các thuật toán phân cụm khác là chỉ hiệu quả với các cụm có dạng cầu và kích thước tương tự và không hiệu quả với đối tượng phần tử ngoại lai. Thuật toán CURE khắc phục được những điều này. Thuật toán này định nghĩa một số cố định các điểm đại diện nằm rải rác trong toàn bộ không gian dữ liệu và chọn để mô tả các cụm được hình thành. Các điểm này được tạo ra nhờ lựa chọn các đối tượng nằm 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 thế trong quá trình này, có thể đo tỷ lệ gia tăng của cụm. Tại mỗi bước của thuật toán 2 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 1 điểm đại diện cho mỗi cụm cho phép CURE khám phá các cụm không phải là dạng hình cầu. Việc co lại các cụm có tác dụng làm giảm tác động của 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 7: Những cụm dữ liệu được khám phá bởi CURE

Để xử lý dữ liệu lớn thuật toán 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 ngẫu nhiên không nhất thiết đưa ra một mô tả cho toàn bộ dữ liệu.

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

- Chọn ngẫu nhiên S từ tập ban đầu
- Phân hoạch S thành các nhóm dữ liệu có kích thước bằng nhau
- Phân cụm các điểm của mỗi nhóm
- Loại bỏ các phần tử ngoại lai: Trước hết, khi các cụm được hình thành cho đến khi các cụm giảm xuống một phần so với các cụm ban đầu. Sau đó, trong trường hợp các phần tử ngoại lai được lấy mẫu cùng với quá trình pha khởi tạo dữ liệu mẫu, 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 bằng các đối tượng gần trung tâm hơn.
- Đánh dấu dữ liệu với các nhãn tương ứng.

Độ phức tạp của thuật toán CURE là $O(n^2\log(n))$. Là thuật toán tin cậy trong việc khám phá ra các cụm có hình thù bất kỳ và có thể áp dụng tốt với các đối tượng 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 ra còn có một số thuật toán phân cụm phân cấp khác như: Thuật toán BIRCH, thuật toán AGNES, thuật toán DIANA, thuật toán ROCK.

2.5.3. Thuật toán COP – Kmeans

Là một thuật toán phân cụm dữ liệu nửa giám sát (phân cụm dựa trên lưới), với phương pháp tiếp cận dựa trên tìm kiếm. Trong thuật toán COP – Kmeans (được 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 nằm khác cụm với nhau.

Các ràng 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 tượng dữ liệu $X = \{X_1,...,X_n\}$ với $X_1 \subseteq R^d$, số lượng cụm K, tập ràng buộc must – link và cannot – link.
- Output: K phân hoạch tách rời sao cho hàm mục tiêu đạt giá trị tối ưu.
    - Bước 1: Khởi tạo các cụm, các tâm ban đầu được chọn ngẫu nhiên sao cho không vi phạm ràng buộc đã cho.
    - 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 các đối tượng nằm trong cụm của tâm đó.
    - $t \leftarrow t+1$

CHƯƠNG III:
ỨNG DỤNG THUẬT TOÁN K - MEANS TRONG PHÂN ĐOẠN ẢNH

3.1. Tổng quan về phân vùng ảnh

Phân đoạn ảnh là một thao tác ở mức thấp, là bước then chốt trong toàn bộ quá trình xử lý ảnh. Quá trình này thực hiện việc phân vùng ảnh thành các vùng rời rạc và đồng nhất với nhau hay nói cách khác là xác định các biên của các vùng ảnh đó. Các vùng ảnh đồng nhất này thông thường sẽ tương ứng với toàn bộ hay từng phần của các đối tượng thật sự bên trong ảnh. Vì thế, trong hầu hết các ứng dụng của lĩnh vực xử lý ảnh (image processing), thị giác máy tính, phân đoạn ảnh luôn đóng một vai trò cơ bản và thường là bước tiền xử lý đầu tiên trong toàn bộ quá trình trước khi thực hiện các thao tác khác ở mức cao hơn như nhận dạng đối tượng, biểu diễn đối tượng, nén ảnh dựa trên đối tượng, hay truy vấn ảnh dựa vào nội dung … Vào những thời gian đầu, các phương pháp phân vùng ảnh được đưa ra chủ yếu làm việc trên các ảnh mức xám do các hạn chế về phương tiện thu thập và lưu trữ. Ngày nay, cùng với sự phát triển về các phương tiện thu nhận và biểu diễn ảnh , các ảnh màu đã hầu như thay thế hoàn toàn các ảnh mức xám trong việc biểu diễn và lưu trữ thông tin do các ưu thế vượt trội hơn hẳn so với ảnh mức xám. Do đó, các kỹ thuật, thuật giải mới thực hiện việc phân vùng ảnh trên các loại ảnh màu liên tục được phát triển để đáp ứng các nhu cầu mới. Các thuật giải, kỹ thuật này thường được phát triển dựa trên nền tảng các thuật giải phân vùng ảnh mức xám đã có sẵn.

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

Mục đích của phân tích ảnh là để có nhiều mô tả tổng hợp về nhiều phần tử khác nhau cấu tạo nên ảnh thô. Vì lượng thông tin chứa trong ảnh rất lớn,

trong khi đó đa số ứng dụng chỉ cần có 1 số thông tin đặc trưng nào đó, do đó cần có một quá trình giảm lượng thông tin khổng lồ ấy.

3.2. Các hướng tiếp cận phân đoạn ảnh

Phân đoạn ảnh là chia ảnh thành các vùng không trùng lắp. Mỗi vùng gồm một nhóm pixel liên thông và đồng nhất theo một tiêu chí nào đó. Tiêu chí này phụ thuộc vào mục tiêu của quá trình phân đoạn. Ví dụ như đồng nhất về màu sắc, mức xám, kết cấu, độ sâu của các layer… Sau khi phân đoạn mỗi pixel chỉ thuộc về một vùng duy nhất. Để đánh giá chất lượng của quá trình phân đoạn là rất khó. Vì vậy trước khi phân đoạn ảnh cần xác định rõ mục tiêu của quá trình phân đoạn là gì. Xét một cách tổng quát, ta có thể chia các hướng tiếp cận phân đoạn ảnh thành ba nhóm chính như sau:

- Các kỹ thuật phân đoạn ảnh dựa trên không gian đặc trưng.
- Các kỹ thuật dựa trên không gian ảnh.
- Các kỹ thuật dựa trên các mô hình vật lý.

3.2.1. Các phương pháp dựa trên không gian đặc trưng

Nếu chúng ta giả định màu sắc bề mặt của các đối tượng trong ảnh là một thuộc tính bất biến và các màu sắc đó được ánh xạ vào một không gian màu nào đó, vậy thì chúng ta sẽ có một cái nhìn đối với mỗi đối tượng trong ảnh như là một *cụm* (cluster) các điểm trong không gian màu đó. Mức độ phân tán của các điểm trong trong một cụm được xác định chủ yếu bởi sự khác biệt về màu sắc. Một cách khác, thay vì ánh xạ các pixel trong ảnh vào một không gian màu cụ thể, ta xây dựng một *histogram* dựa trên các đặc trưng màu dạng *ad-hoc* cho ảnh đó (ví dụ như Hue), và thông thường, các đối tượng trong ảnh sẽ xuất hiện như các giá trị đỉnh trong histogram đó. Do đó, việc phân vùng các đối tượng trong ảnh tương ứng với việc xác định các cụm – đối với cách biểu diễn thứ nhất – hoặc xác định các vùng cực trị của histogram – đối với cách biểu diễn thứ hai.

Các phương pháp tiếp cận này chỉ làm việc trên một không gian màu xác định chẳng hạn phương pháp của Park áp dụng trên không gian màu RGB, còn phương pháp của Weeks và Hague thì áp dụng trên không gian màu HIS. Dựa trên không gian đặc trưng, ta có các phương pháp phân đoạn: phương pháp phân nhóm đối tượng không giám sát, phương pháp phân lớp trung bình-k thích nghi, phương pháp lấy ngưỡng histogram.

3.2.2. Các phương pháp dựa trên không gian ảnh

Hầu hết những phương pháp được đề cập trong phần trên đều hoạt động dựa trên các không gian đặc trưng của ảnh(thông thường là màu sắc). Do đó, các vùng ảnh kết quả là đồng nhất tương ứng với các đặc trưng đã chọn cho từng không gian. Tuy nhiên, không có gì đảm bảo rằng tất cả các vùng này thể hiển một sự cô đọng (compactness) về nội dung xét theo ý nghĩa không gian ảnh (ý nghĩa các vùng theo sự cảm nhận của hệ thần kinh con người). Mà đặc tính này là quan trọng thứ hai sau đặc tính về sự thuần nhất của các vùng ảnh. Do các phương pháp gom cụm cũng như xác định ngưỡng histogram đã nêu đều bỏ qua thông tin về vị trí của các pixel trong ảnh.

Trong các báo cáo khoa học về phân vùng ảnh mức xám, có khá nhiều kỹ thuật cố thực hiện việc thoả mãn cùng lúc cả hai tiêu chí về tính đồng nhất trong không gian đặc trưng của ảnh và tính cô đọng về nội dung ảnh. Tuỳ theo các kỹ thuật mà các thuật giải này áp dụng, chúng được phân thành các nhóm sau:

- Các thuật giải áp dụng kỹ thuật chia và trộn vùng.
- Các thuật giải áp dụng kỹ thuật tăng trưởng vùng.
- Các thuật giải áp dụng lý thuyết đồ thị.
- Các giải thuật áp dụng mạng neural.
- Các giải thuật dựa trên cạnh.

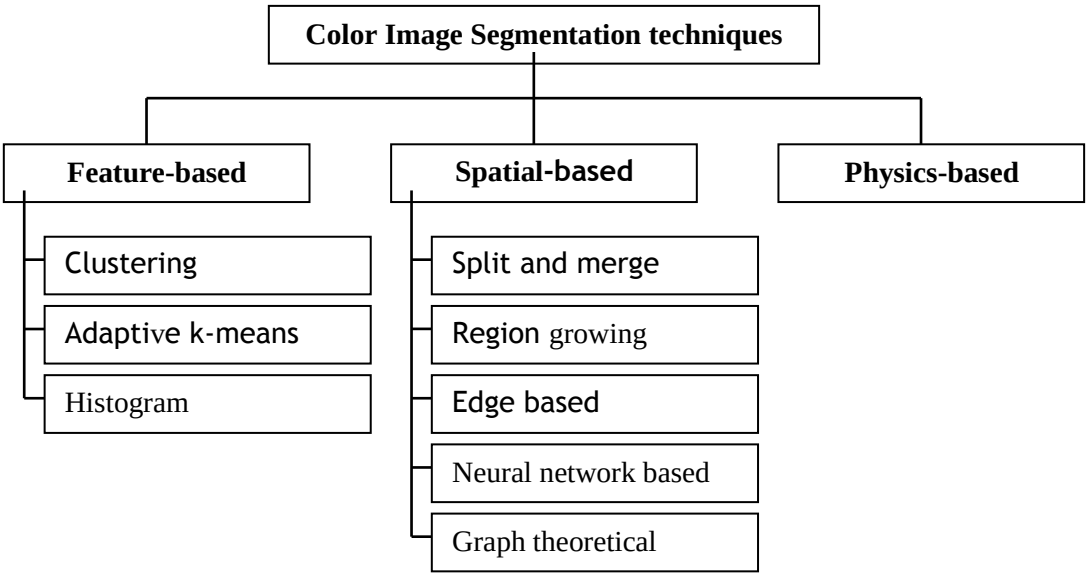
3.2.3. Các phương pháp dựa trên mô hình vật lý

Tất cả các giải thuật được xem xét qua, không ít thì nhiều ở mặt nào đó đều có khả năng phát sinh việc phân vùng lỗi trong các trường hợp cụ thể nếu như các đối tượng trong ảnh màu bị ảnh hưởng quá nhiều bởi các vùng sáng hoặc bóng mờ, các hiện tượng này làm cho các màu đồng nhất trong ảnh thay đổi nhiều hoặc ít một cách đột ngột. Và kết quả là các thuật giải này tạo ra các kết quả phân vùng quá mức mong muốn so với sự cảm nhận các đối tượng trong ảnh bằng mắt thường. Để giải quyết vấn đề này, các giải thuật phân vùng ảnh áp dụng các mô hình tương tác vật lý giữa bề mặt các đối tượng với ánh sáng đã được đề xuất. Các công cụ toán học mà các phương pháp này sử dụng thì không khác mấy so với các phương pháp đã trình bày ở trên, điểm khác biệt chính là việc áp dụng các mô hình vật lý để minh hoạ các thuộc tính phản chiếu ánh sáng trên bề mặt màu sắc của các đối tượng.

Cột mốc quan trọng trong lĩnh vực phân vùng ảnh màu dựa trên mô hình vật lý được Shafer đặt ra.Ông giới thiệu *mô hình phản xạ lưỡng sắc* cho các vật chất điện môi không đồng nhất. Dựa trên mô hình này, Klinker đã đặt ra một giải thuật đặt ra một số giả thiết quang học liên quan đến màu sắc, bóng sáng, bóng mờ của các đối tượng và cố gắng làm phù hợp chúng với hình dạng của các cụm. Hạn chế chính của giải thuật này là nó chỉ làm việc trên các vật chất điện môi không đồng nhất.Hai ông cùng tên Tsang đã áp dụng mô hình phản xạ lưỡng sắc trong không gian HSV để xác định các đường biên trong ảnh màu.

Healey đề xuất một *mô hình phản xạ đơn sắc* cho các vật chất kim loại. Các phương pháp đề cập trong phần này chỉ áp dụng cho hai loại vật chất là kim loại và điện môi không đồng nhất. Một thuật toán tổng quát và phức tạp hơn cũng được Maxwell và Shafer đề xuất trong.

Tóm lại, một cái nhìn tổng quan về các phương pháp phân đoạn ảnh như sau:



Mỗi phương pháp đều có những ưu nhược điểm nhất định:

Phương pháp phân vùng	Ưu điểm	Khuyết điểm
Featured-based techniques		
Clustering	▪ Phân loại không cần giám sát. ▪ Tồn tại các phương pháp heuristic và hữu hạn.	▪ Không quan tâm đến các thông tin trong không gian ảnh. ▪ Có vấn đề trong việc xác định số lượng các cụm ban đầu. ▪ Khó khăn trong việc điều chỉnh các cụm sao cho phù hợp với các vùng trong ảnh.
Adaptive Clustering	▪ Sở hữu tính liên tục trong không gian ảnh và tính thích nghi cục bộ đối với các vùng ảnh. ▪ Sử dụng các ràng buộc về không gian ảnh.	▪ Cực đại hoá một xác suất hậu điều kiện có thể bị sai do các cực trị địa phương. ▪ Hội tụ chậm.
Phương pháp phân vùng	**Ưu điểm**	**Khuyết điểm**
Histogram thresholding	▪ Không cần biết trước bất kỳ thông tin nào từ ảnh. ▪ Các giải thuật nhanh và dễ dàng cài đặt.	▪ Bỏ qua các thông tin về không gian ảnh. ▪ Lấy ngưỡng trong các histogram đa chiều là

		một quá trình phức tạp. ▪ Ảnh hưởng dễ dàng bởi nhiễu xuất hiện trong ảnh.
Spatial-based techniques		
Spit and Merge	▪ Sử dụng các thông tin về không gian ảnh là chính. ▪ Cho kết quả tốt với các ảnh chứa nhiều vùng màu đồng nhất.	▪ Định nghĩa mức độ đồng nhất về màu sắc có thể phức tạp và khó khăn. ▪ Quadtree có thể gây ra các kết quả không như mong muốn.
Region growing	▪ Các vùng ảnh đồng nhất và liên thông. ▪ Có một số thuật giải có tốc độ thực thi khá nhanh.	▪ Tốn kém chi phí sử dụng bộ nhớ và tính toán. ▪ Gặp khó khăn trong việc thu thập tập các điểm mầm và xác định các điều kiện đồng nhất đầy đủ. ▪ Chịu ảnh hưởng bởi các đặc tính tự nhiên của kỹ thuật này.
Graph theories	▪ Thể hiện tốt không gian ảnh bằng đồ thị. ▪ Một số thuật toán có tốc độ thực hiện nhanh.	▪ Một vài thuật giải mất khá nhiều thời gian thực hiện. ▪ Các đặc trưng cục bộ đôi khi được sử dụng nhiều hơn các đặc trưng toàn cục.
Neural networks	▪ Mức độ song song hoá cao và có tốc độ thực thi nhanh. ▪ Khả năng chống chịu tốt trước các thay đổi xấu. ▪ Một công cụ hữu hiệu cho các ứng dụng nhận dạng và xử lý ảnh y khoa.	▪ Màu sắc có thể làm tăng độ phức tạp của mạng. ▪ Quá trình học cần phải biết trước số lượng các phân lớp/cụm.
Edge-based	▪ Là phương pháp được hỗ trợ mạnh bởi các toán tử dò biên. ▪ Có hiệu năng tốt với các ứng dụng dò biên đối tượng theo đường cong.	▪ Khó khăn trong việc định nghĩa một hàm gradient cho các ảnh màu. ▪ Nhiễu hoặc các ảnh có độ tương phản kém ảnh hưởng xấu đến kết quả phân vùng.
Phương pháp phân vùng	**Ưu điểm**	**Khuyết điểm**

Physics-based techniques		
	▪ Khẳng định tính chắc chắn đối với các vùng bóng sáng/tối, và vùng bóng chuyển tiếp (diffuse hoặc shade) ▪ Phân vùng các đối tượng dựa vào thành phần vật liệu cấu tạo	▪ Bị giới hạn vào một số lượng nhất định các loại vật chất hình thành nên đối tượng. ▪ Khó khăn trong việc xác định vùng bóng sáng và bóng chuyển tiếp trong các ảnh thực. ▪ Một vài giải thuật đòi hỏi các thông tin về hình dạng đối tượng (không luôn luôn đáp ứng được). ▪ Chi phí tính toán khá cao.

Đối với bài toán truy vấn ảnh theo nội dung, bước tiền xử lý phân đoạn phải chú ý đến các thông tin tòan cục và cả cục bộ. Đồng thời đảm bảo tính liên tục trong không gian ảnh. Vì vậy, ở đây ta sẽ đi sâu vào các thuật toán phân đoạn: phương pháp phân đoạn yếu của B.G. Prasad áp dụng trong hệ thống truy vấn ảnh của ông; phương pháp phân đoạn trung bình-k thích nghi; phương pháp phân đoạn theo ngưỡng cục bộ thích nghi.

3.3. Một số phương pháp phân đoạn cụ thể

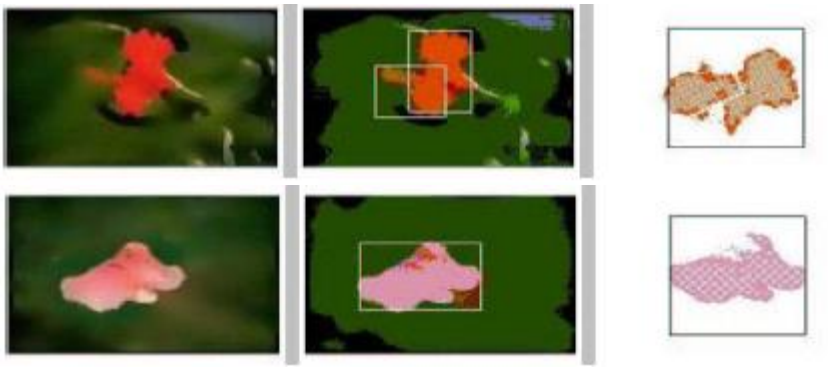
3.3.1. Phương pháp phân đoạn yếu của B.G. Prasad

Đây là phương pháp do B.G. Prasad đề xuất và được áp dụng tronh hệ thống truy vấn ảnh theo chỉ mục màu sắc hình dạng và vị trí của tác giả. Phương pháp này sử dụng sự lượng tử hóa màu trong không gian màu RGB, sử dụng 25 màu phân biệt (bằng giác quan) để phân đoạn ảnh dựa trên những màu trội. Vì 25 màu là đủ để phân biệt rõ tất cả các vùng màu trong cơ sở dữ liệu hình ảnh mà tác giả chọn.

Việc chọn số lượng màu phân biệt trong không gian màu giảm là một sự trao đổi giữa sự thể hiện và tốc độ đối với một ứng dụng riêng biệt. Với chỉ mục màu

hiệu quả, số lượng màu ít thì thích hợp và nó cũng làm giảm sự tính toán. Dưới đây là bảng gồm 25 màu (theo giác quan) được chọn từ bảng màu RGB chuẩn.

Ví dụ phân đoạn ảnh.



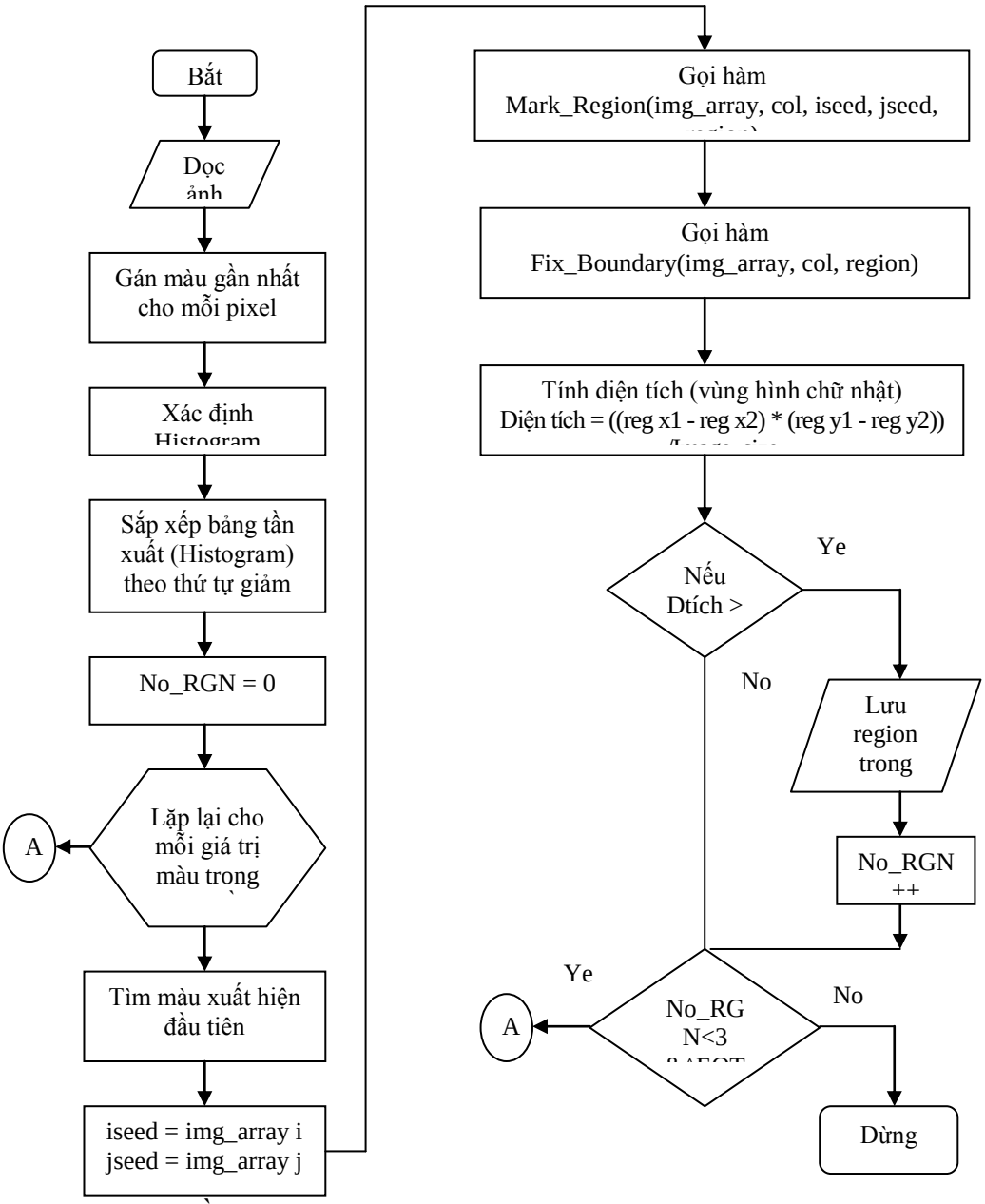
Hình 8: ví dụ phân đoạn ảnh bằng phương pháp phân đoạn yếu

Bảng tra màu

Thứ tự	**Màu**	**R**	**G**	**B**
1	Black	0	0	0
2	Sea green	0	182	0
3	Light green	0	255	170
4	Olive green	36	73	0
5	Aqua	36	146	170
6	Bright green	36	255	0
7	Blue	73	36	170
8	Green	73	146	0
9	Turquoise	73	219	170
10	Brown	109	36	0
11	Blue gray	109	109	170
12	Lime	109	219	0
13	Lavender	146	0	170
14	Plum	146	109	0
15	Teal	146	182	170
16	Dark red	182	0	0
17	Magenta	182	73	170
18	Yellow green	182	182	0
19	Flouro green	182	255	170
20	Red	219	73	0
21	Rose	219	146	170
22	Yellow	219	255	0
23	Pink	255	36	170
24	Orange	255	146	0
25	White	255	255	255

Không gian màu được chia đều và để tính điểm tương đồng giữa 2 vùng thì chúng ta sử dụng khoảng cách euclidean chuẩn. Như vậy, không gian màu RGB được chia thành những không gian con được gọi là phân loại màu (color category).

Phương pháp này phụ thuộc vào việc xác định các biên. Màu của một pixel (màu phân biệt được bằng giác quan) có thể được mô tả bằng loại màu trong vùng màu giảm tương ứng. Tiến trình phân đoạn và xử lý chọn vùng trội được trình bày bằng sơ đồ sau:



Sơ đồ xử lý chọn vùng và phân đoạn

Thủ tục này phân đoạn ảnh thành những vùng, bằng cách ánh xạ toàn bộ pixel của ảnh lên vùng màu tương ứng trong không gian màu đã được giảm, và sau đó nhóm các pixel cùng loại với nhau. Có nghĩa là: với mỗi pixel màu trên ảnh gốc, ta sẽ tìm được một màu trong 25 màu mà ta đã định nghĩa trước gần với nó nhất, nó sẽ được lưu lại và đó làm màu trong ảnh mới. Ở đây, ta sẽ sử dụng khoảng cách Euclidean để tìm màu kết quả.

Gọi p_r, p_g, p_b là những giá trị cường độ màu của pixel tương ứng với ba thành phần red, green, blue, và C_{iR}, C_{iG}, C_{iB} là những giá trị màu tương ứng với

nó trong bảng màu. Để tính khoảng cách màu C_d, ta sử dụng khoảng cách Euclidean như sau:

$$C_d = \min\left(\sqrt{(p_r - C_{iR})^2 + (p_g - C_{iG})^2 + (p_b - C_{iB})^2}\right), i = 1,...,25$$

Sau khi có được ảnh mới (ảnh đã được giảm thiểu màu), người ta sẽ đánh dấu vùng chọn trên ảnh này. Với mỗi vùng trội được chọn người ta sẽ vẽ ra một đường biên hình chữ nhật. Diện tích của hình chữ nhật biên được sử dụng để xác định diện tích vùng có màu trội được chuẩn hóa. Số lượng vùng hiện có, các thông tin của mỗi vùng như màu, diện tích được chuẩn hóa… được lưu trong metafile dành cho xử lý sau này. Thông tin này được sử dụng để xây dựng cây chỉ mục ảnh dành cho máy tìm kiếm (seach engine).

Giải thích sơ đồ (các bước) phân đoạn và dò đường biên:

(1) Đọc ảnh và tạo một mảng ảnh chứa các thành phần màu RGB của mỗi pixel trong ảnh.

(2) Với mỗi pixel trong ảnh, thực hiện:

(a) Tìm màu gần với màu của pixel trong ảnh gốc bằng cách sử dụng công thức tính C_d (với i=1,2,…,25)

$$C_d = \min\left(\sqrt{(p_r - C_{iR})^2 + (p_g - C_{iG})^2 + (p_b - C_{iB})^2}\right), i = 1,...,25$$

với p_r, p_g, p_b : 3 thành phần màu RGB của ảnh gốc

C_{iR}, C_{iG}, C_{iB} : 3 thành phần màu tại vị trí i trong bảng 25 màu.

Tính được C_d min ta sẽ có giá trị i, đây chính là vị trí của màu gần nhất cần tìm.

(b) Gán màu tìm được này cho pixel đang xét (ứng với C_d min).

(3) Tạo bảng tần xuất cho mỗi màu được gán (tạo histogram cho ảnh).

(4) Sắp xếp bảng tần xuất theo thứ tự giảm (để xác định những vùng màu trội).

(5) Lặp lại từ bước 6 đến 10 cho đến khi tìm được 3 vùng màu trội hoặc đến cuối bảng tần xuất.

(6) Tiếp tục quét điểm ảnh theo thứ tự, dừng lại ở pixel đầu tiên có cùng giá trị màu trong bảng tần xuất được sắp xếp.

(7) Gán vị trí pixel tìm được đó vào hai biến iseed, jseed tương ứng theo chiều ngang và chiều dọc của ảnh.

(8) Đánh dấu toàn bộ vùng (region) bằng cách sử dụng vùng lân cận 8 connected của pixel đó.

(9) Lấy tọa độ đường biên (x,y) của vùng được đánh dấu R và vẽ hình chữ nhật biên.

(10) Xác định kích thước chuẩn hóa s(R) của hình chữ nhật biên bằng công thức:

$s(R) = (|x_1 - x_2|*|y_1 - y_2|)/N$

Với (x_1, y_1) và (x_2, y_2)là tọa độ của tương ứng của 2 đỉnh đối nhau trong hình chữ nhật biên, N là kích thước của ảnh.

Theo thuật toán này, để tránh việc lấy lặp lại các vùng trội thì ở bước 6, sau khi chọn được pixel đầu tiên thỏa, chúng ta sẽ kiểm tra xem pixel này đã có nằm trong vùng đã xét nào hay chưa, nếu chưa thì sẽ chọn nó.

3.3.2. Phương pháp phân đoạn dựa trên ngưỡng cục bộ thích nghi

Số ngưỡng cục bộ và giá trị của chúng không được chỉ định trước mà được trích lọc thông qua quá trình kiểm tra các thông tin cục bộ. Giải thuật gồm các bước tuần tự như sau:

- Áp dụng giải thuật Watershed chia ảnh thành rất nhiều vùng con.
- Trộn các vùng và đồng thời phát hiện ngưỡng cục bộ. Ngưỡng được tính từ thông tin cục bộ của vùng và các vùng lân cận

Giải thuật này cho kết quả tương đối tin cậy trên nhiều loại ảnh khác nhau

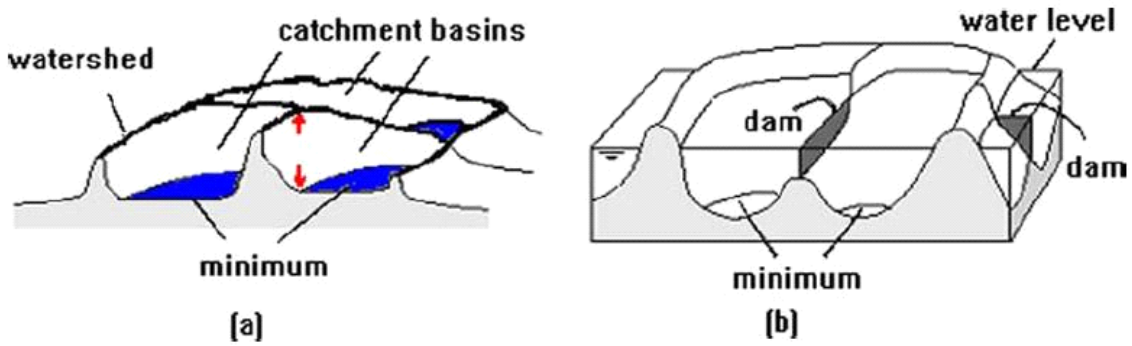
3.3.3. Phân đoạn sơ khởi bằng Watershed

Dữ liệu đầu vào của giải thuật Watershed là một ảnh xám. Vì vậy, trước tiên ta biến đổi ảnh đầu vào I thành ảnh xám. Sau đó, dùng giải thuật tìm cạnh Canny để lấy cường độ gradient, kí hiệu là I_G. Với ảnh gradient nhận được, ta hình liên tưởng đến một lược đồ địa hình, vùng có độ xám cao hơn là vùng trũng hơn và ngược lại. Tại mỗi pixel, việc đánh giá sẽ dựa vào giá trị mức xám của pixel đó.

Giải thuật định nghĩa hai thuật ngữ là vũng chứa nước (catchment basin) và đập ngăn nước (dams). Mỗi catchment basin được kết hợp với giá trị M nhỏ nhất. M là tập hợp các pixel liên thông mà một giọt nước rơi xuống từ pixel bất kì thuộc catchment basin này cứ rơi cho đến khi nó đạt được giá trị nhỏ nhất M. Trên đường rơi xuống, giọt nước chỉ đi qua những pixel thuộc về catchment basin này.

Dam thực chất là những đường phân nước, chúng tập hợp các pixel làm nhiệm vụ phân cách các catchment basin. Vì vậy, giọt nước rơi từ một bên của dams sẽ đạt trị nhỏ nhất của một catchment basin, trong khi đó giọt nước rơi từ cạnh khác của dam lại đạt trị nhỏ nhất trong catchment basin khác. Bạn xem minh họa cho catchment basin và dams trong hình 1 dưới đây.

Catchment basin và dam. Các pixel ở mức thấp và cao được thể hiện thông qua mũi tên đỏ hướng lên và hướng xuống.



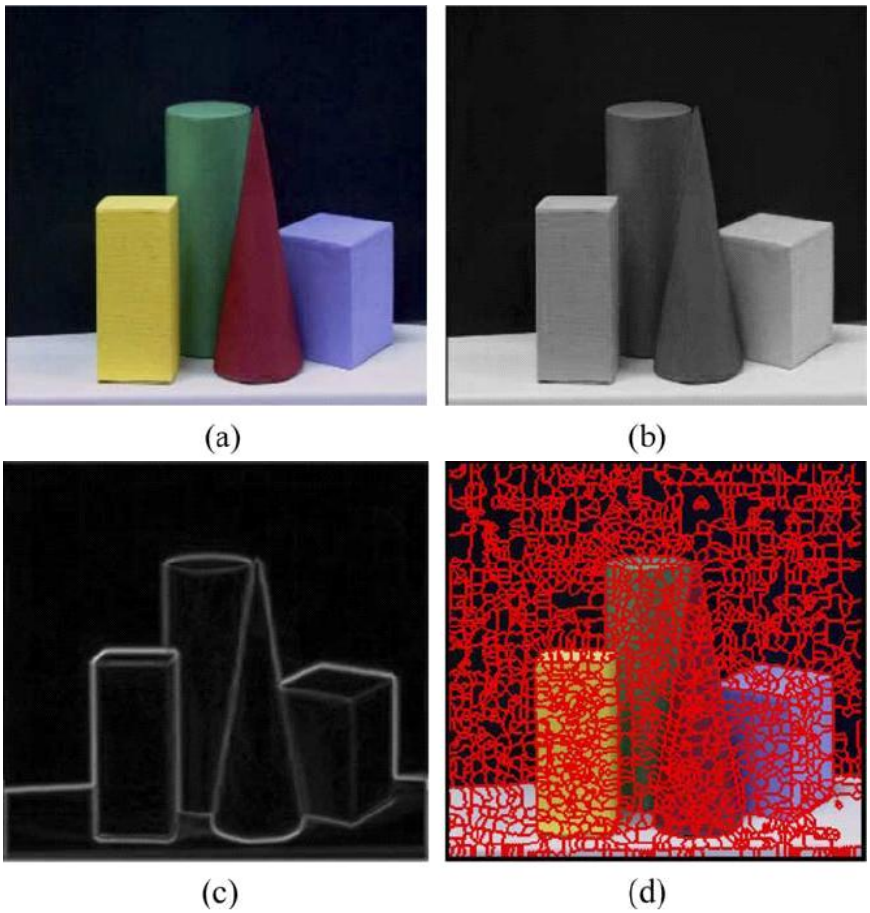
Áp dụng *giải thuật watershed, phiên bản của Vincent và Soille…* Phiên bản này mô phỏng việc ngâm nước dần dần bề mặt địa hình của ảnh từ vùng thấp nhất cho đến khi mọi pixel của ảnh đều được ngâm trong nước. Giải thuật gồm hai bước: *sắp thứ tự và làm ngập nước*.

Ở bước thứ nhất, ta sắp xếp các pixel theo thứ tự tăng dần của cường độ xám. Kế đến, trong bước làm ngập nước, giải thuật quét các pixel theo trình tự đã sắp xếp để xây dựng các catchment basin. Mỗi catchment basin có một nhãn phân biệt. Bạn hãy thử hình dung ta đem nhúng nước một bề mặt địa hình, bắt đầu tại điểm thấp nhất của mặt địa rồi cho nước dâng dần lên. Khi nước trong các vũng cạnh nhau có thể hoà vào nhau tại một điểm, tại đó ta xây dựng một đập chắn nước, rồi lại tiếp tục cho nước dâng lên. Quá trình xây đập chắn giữa các vũng và cho nước dâng cứ lặp đi lặp lại cho đến khi mọi điểm của bề mặt địa hình đều được ngâm nước.

Trở lại giải thuật, ta làm tương tự, tại một điểm mà nước trong các catchment basin có thể hoà vào nhau, ta xây dựng một đập chắn nước – dam. Cứ như thế, lặp quá trình cho nước dâng lên và xây dựng dam tại những điểm nước của các catchment basin có thể hoà lẫn vào nhau cho đến khi mọi điểm ảnh đều nằm trong nước. Khi đó, ta nhận được ảnh gồm vô số vùng con, mỗi vùng con tương ứng với một catchment basin, còn biên của mỗi vùng chính là dam. Bạn xem hình 2 minh họa quá phân ảnh ban đầu (a) thành vô số vùng con (d). Trước tiên ảnh gốc 2a được biến đổi thành ảnh xám 2b. Kế đến, áp dụng giải thuật tìm cạnh Canny trên ảnh xám gradient ở hình 2b, ta được ảnh 2c chỉ gồm các đường

nét. Đồng thời, áp dụng giải thuật watershed trên ảnh xám ta được hình 2d, chứa vô số vùng con.

Như vậy khi áp dụng giải thuật watershed vào ảnh I_G, ta nhận được ảnh kết quả gồm n vùng không trùng lắp. Do các vùng này sẽ được trộn trong giai đoạn trộn tiếp theo nên chúng tôi đặt đánh dấu chúng bằng kí hiệu R_i^{mi}, I = 1,…,n, mi = 1,…,Mi, với n là số lượng vùng và Mi là số lần trộn của R_i^{mi} trong quá trình trộn. R_i^0, i=1,…,n là tập các vùng khởi tạo, hay nói cách khác chúng là kết quả của giải thuật watershed trước khi quá trình trộn lặp của giai đoạn hai bắt đầu.



(a) Ảnh gốc ban đầu. (b) Ảnh xám. (c) Ảnh xám gradient sau khi đã áp dụng giải thuật tìm cạnh Canny. (d) Ảnh phân đoạn nhận được từ việc áp dụng giải thuật watershed

3.3.4. Trộn các vùng

- ***Đánh giá sự khác biệt giữa các vùng***

Để xác định trình tự trộn các vùng, ta xác định hàm thể hiện sự khác biệt giữa hai vùng lân cận $R_i^{m_i}$ và $R_j^{m_j}$, kí hiệu là $f(R_i^{m_i}, R_j^{m_j})$. Hàm thể hiện sự khác biệt dựa vào hai thành phần: *màu sắc và cạnh*.

- Đối với thành phần màu, giải thuật *sử dụng giá trị Huetrong không gian màu HSV* vì trị này ít bị ảnh hưởng bởi sự thay đổi nguồn chiếu sáng, ví dụ như hình dạng và bóng. Cụ thể, chúng tôi lấy trị trung bình của thành phần Hue trong vùng $R_i^{m_i}$, kí hiệu $\mu_h(R_i^{m_i})$.
- Thành phần thể hiện sự khác biệt còn lại là cạnh được biểu diễn bởi *cường độ gradient*. Cường độ gradient chính là trị của các pixel trong I_G. Chúng tôi qui định $\mu_G(R_i^{m_i}, R_j^{m_j})$ là trị gradient trung bình giữa hai vùng $R_i^{m_i}$ và $R_j^{m_j}$, được xác định dựa trên trị gradient của những pixel chung của hai vùng.

Gọi B_{ij} là tập pixel thuộc về ranh giới giữa hai vùng $R_i^{m_i}$ và $R_j^{m_j}$, $\mu_G(R_i^{m_i}, R_j^{m_j})$ được định nghĩa như sau:

$$\mu_G(R_i^{m_i}, R_j^{m_j}) = \frac{\sum_{(x,y)\in BIJ} I_G(x,y)}{|B_{ij}|} \quad (1)$$

Với $|B_{ij}|$ là số pixel của tập B_{ij}.

Ta có hàm thể hiện sự khác biệt dựa vào trị Hue và độ gradient được tính như sau:

$$f(R_i^{m_i}, R_j^{m_j}) = w_1 * d(\mu_h(R_i^{m_i}), \mu_h(R_j^{m_j})) + w_2 * \mu_G(R_i^{m_i}, R_j^{m_j}) \quad (2)$$

Trong đó, $d(\mu_h(R_i^{m_i}), \mu_h(R_j^{m_j}))$ là sự chênh lệch giữa trị trung bình của $R_i^{m_i}$ và $R_j^{m_j}$:

$$d(\mu_h(R_i^{mi}),\mu_h(R_j^{mj})) = \min\{|\mu_h(R_i^{mi}) - \mu_h(R_j^{mj})|, (360 - |\mu_h(R_i^{mi}) - \mu_h(R_j^{mj})|\}$$

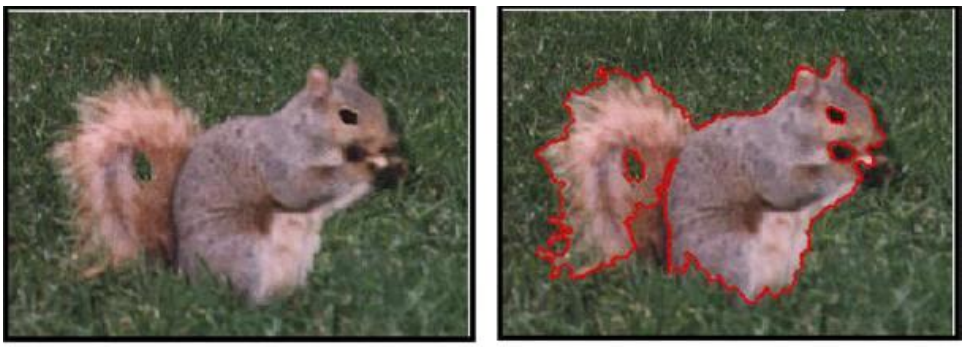
(3)

với w_1 và w_2 là *các* hệ số hằng đã định trước. Nếu hàm thể hiện sự khác biệt $f(R_i^{mi}, R_j^{mj})$ phụ thuộc chủ yếu vào trị Hue của màu sắc hơn là cường độ gradient thì $w_1 >> w_2$. Dựa vào kinh nghiệm thực tiễn trên các loại ảnh khác nhau, trị của w_1 và w_2 tương ứng là 0.8 và 0.2.

- ***Tìm ngưỡng cục bộ thích nghi***

Mặc dù phần mô tả quá trình trộn đã hoàn chỉnh nhưng ta vẫn chưa xác định được khi nào thì giải thuật dừng. Hay nói cách khác, ta vẫn chưa biết cách xác định vùng nào không trộn được và thời điểm nào thì không trộn. Như vậy, chúng ta cần có cơ chế tự động rút trích thông tin về ngưỡng cục bộ thông qua việc theo dõi sự thay đổi của mỗi vùng trong quá trình trộn. Các ngưỡng này sẽ cho biết có thể trộn một vùng hay không. Như thế, các ngưỡng này giúp hình thành phân vùng hoàn chỉnh cuối cùng.

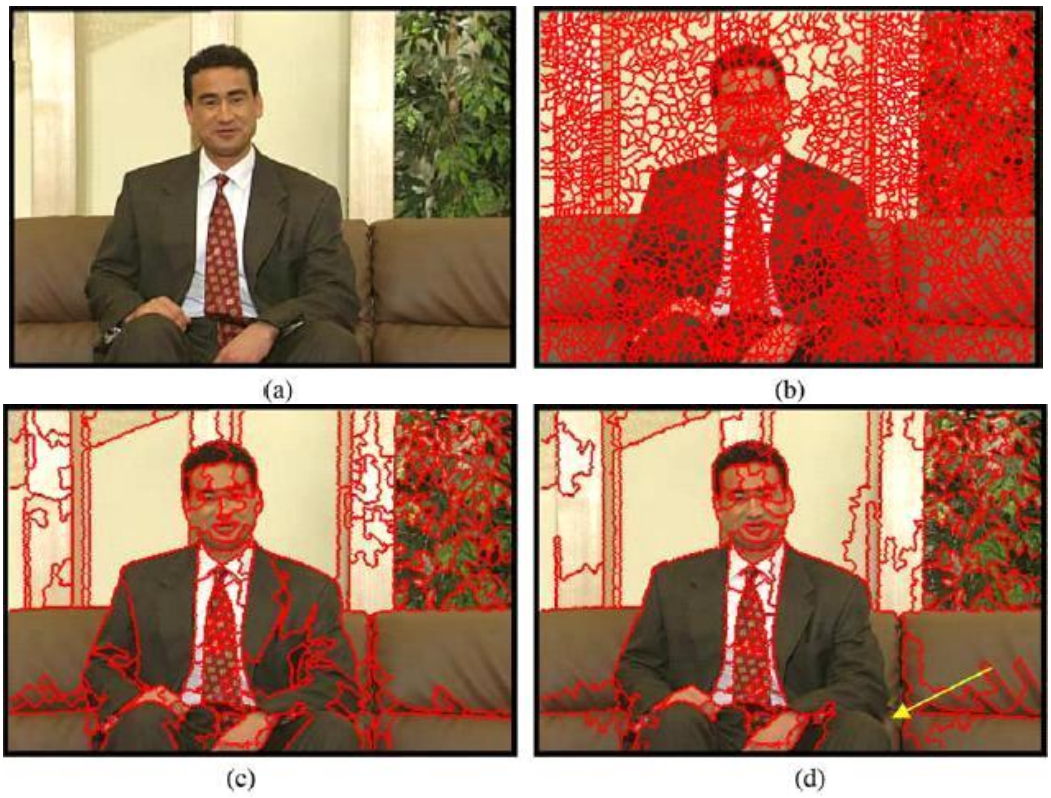
Như chúng ta đã biết quá trình phân đoạn là thao tác cục bộ, nên không phải mọi bước trộn cục bộ đều dừng đồng thời. Do đó việc sử dụng ngưỡng toàn cục là không đủ vì các vùng thường tách biệt với xung quanh nó bởi những ngưỡng khác nhau vào những lần xử lý khác nhau. Tuy nhiên trong một vài trường hợp thì ngưỡng toàn cục lại phù hợp. Ví dụ ở hình 6 mô tả một trường hợp ngoại lệ, chỉ dùng một ngưỡng toàn cục mà vẫn cho kết quả phân đoạn chính xác. Lý do là ảnh ví dụ chỉ chứa một đối tượng đồng nhất về màu sắc, đồng thời phần nền cũng có màu đồng nhất. Trong trường hợp này chỉ cần một ngưỡng cho quá trình trộn là đủ. Quá trình trộn sẽ dừng khi trọng số của các cạnh khảo sát lớn hơn ngưỡng chọn trước, cụ thể trong ví dụ này là 100. Bạn xem kết quả phân đoạn bằng ngưỡng trên ở hình 9b. Trong thực tế, các ảnh phân tích thường chứa nhiều hơn hai vùng nên rất khó phân đoạn nếu chỉ dùng một ngưỡng toàn cục.



Hình 9:(a) Ảnh gốc. (b) Kết quả phân đoạn bằng ngưỡng toàn cục 100.

Bạn sẽ cảm nhận được nhu cầu dùng ngưỡng cục bộ thay cho ngưỡng toàn cục khi xem hình 7. Ta có hình gốc 7a, hình 7b là kết quả của giải thuật watershed. Với ngưỡng toàn cục t = 20 ta được kết quả phân đoạn hình 7c, còn hình 7d là kết quả tương ứng với ngưỡng toàn cục t = 30. Trong hình 7.c, mọi vùng đều đồng nhất và có thể lớn hơn. Tuy nhiên, khi ngưỡng tăng lên 30 như ở hình 7d, các vùng nhìn bằng mắt thường là đồng nhất như mặt và ghế lại bị phân quá nhỏ. Trong khi đó, vùng chỉ ra bởi mũi tên vàng vẫn chưa đồng nhất. Để phân nó thành nhiều vùng đồng nhất thì ngưỡng phải nhỏ hơn 30, khi đó việc trộn hai vùng không đồng nhất là áo khoát của người đàn ông và cái ghế sẽ không được thực hiện.

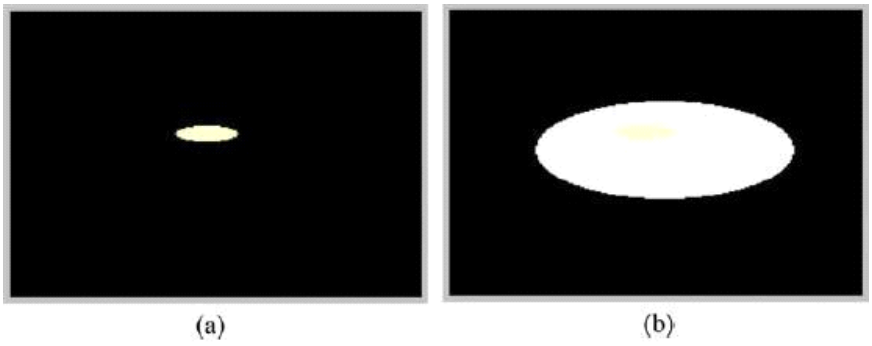
Chúng ta đã nhận biết được nhu cầu cần thiết tính ngưỡng cục bộ, nhưng tính ngưỡng thế nào và dựa vào yếu tố gì thì cần xem xét tiếp. Việc tính ngưỡng cục bộ phải dựa vào các thông tin cục bộ, liên quan đến vùng đang xét và những vùng lân cận xung quanh nó. Thế nhưng tại sao phải xét vùng lân cận? Ta phải xét các vùng lân cận vì một vùng thường bị ảnh hưởng bởi các vùng xung quanh nó. Bạn xem ví dụ hình 8 để thấy mối quan hệ khắng khít giữa một vùng và các vùng lân cận nó, cùng một vùng nhưng nếu đặt vào giữa những vùng lân cận khác nhau thì cảm nhận thị giác sẽ rất khác nhau. Trong hình 8a, đối tượng hình ellipse màu vàng nổi bật trên nền màu đen, khác hẳn với hình 8b, cũng đối tượng ellipse màu vàng này nhưng gần như hòa vào màu nền trắng xung quanh nó, rất khó nhận biết.



Hình 10: (a) Ảnh gốc (b) Sau khi áp dụng giải thuật watershed.

(c) Sau khi hoàn thành quá trình trộn dùng một ngưỡng toàn cục t=20.

(d) Sau khi trộn dùng một ngưỡng toàn cục t=30.



Hình 11: Vùng sáng elip hiển thị khác nhau khi do nền khác nhau.

3.4. Thuật toán k-means cho phân đoạn ảnh

Tầm quan trọng và những khó khăn của việc gom cụm 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 tính đặc biệt trong lĩnh vực xử lý ảnh. Và phân đoạn ảnh được ứng dụng rất mạnh mẽ trong các bài toán phân tích và hiểu ảnh tự động,

nhưng đó cũng là một bài toán khó mà tới bây giờ các nhà khoa học cũng chưa tìm ra cách giải 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 có câu trả lời.

Trong khoảng 30 trở lại đây đã có rất nhiều thuật toán được đề xuất để giải quyết vấn đề phân đoạn ảnh. Các thuật toán này hầu hết đều dựa vào hai thuộc tính quan trọng của mỗi điểm ảnh so với điểm lân cận của nó; đó là sự "*khác"* và *"giống nhau"*. Các phương pháp dựa trên sự giống nhau của các điểm ảnh gọi là phương pháp miền , còn các phương pháp dựa trên sự khác nhau của các điểm ảnh gọi là phương pháp biên. Trong đề tài này, em xin trình bày thuật toán k – means để giải quyết bài toán phân đoạn ảnh.

3.4.1. Mô tả bài toán

Input:+ Ảnh có kích thước m x n

+ Số cụm k muốn phân đoạn

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

3.4.2. Các bước thực hiện chính trong thuật toán

Thuật 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 lần lượt các điểm tới cụm mà có khoảng cách từ các điểm đó tới trọng tâm của cụm đó là nhỏ nhất, cập nhật lại trọng tâm cụm. Kết quả thu được sau khi tâm các cụm là không đổi.

- Các bước của thuật toán:Thuật toán k -means gồm 4 bước:

1. Chọn ngẫu nhiên k đối tượng làm trọng tâm ban đầu của k cụm

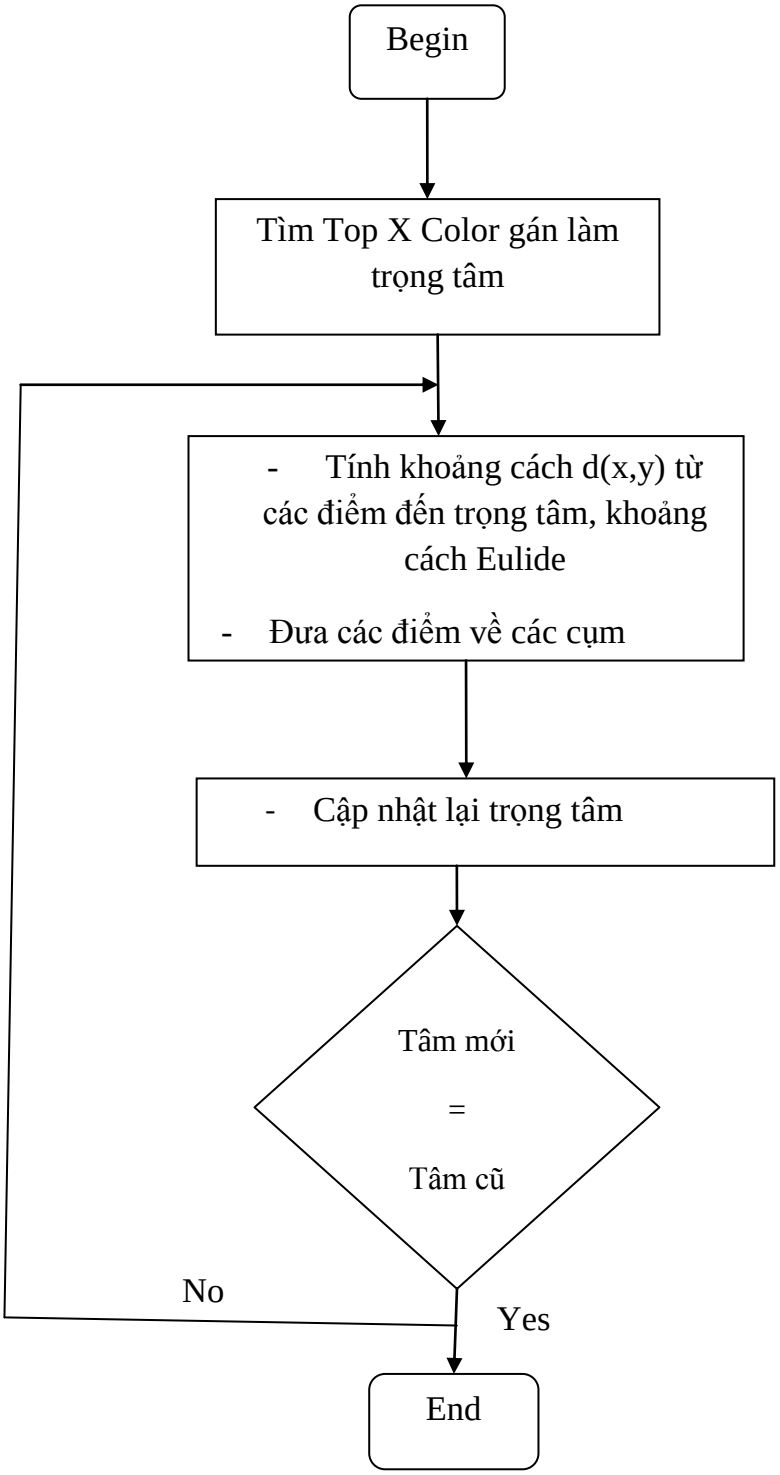
2. Gán (hoặc gán lại) từng đối tượng còn lại vào cụm có trọng tâm gần nó nhất.

Nếu không có phép gán lại nào thì dừng. Vì không có phép gán lại có nghĩa là các cụm đã ổn định và thuật toán không thể cải thiện làm giảm độ phân biệt hơn được nữa.

3. Tính lại trọng tâm cho từng cụm.

4. Quay lại bước 2.

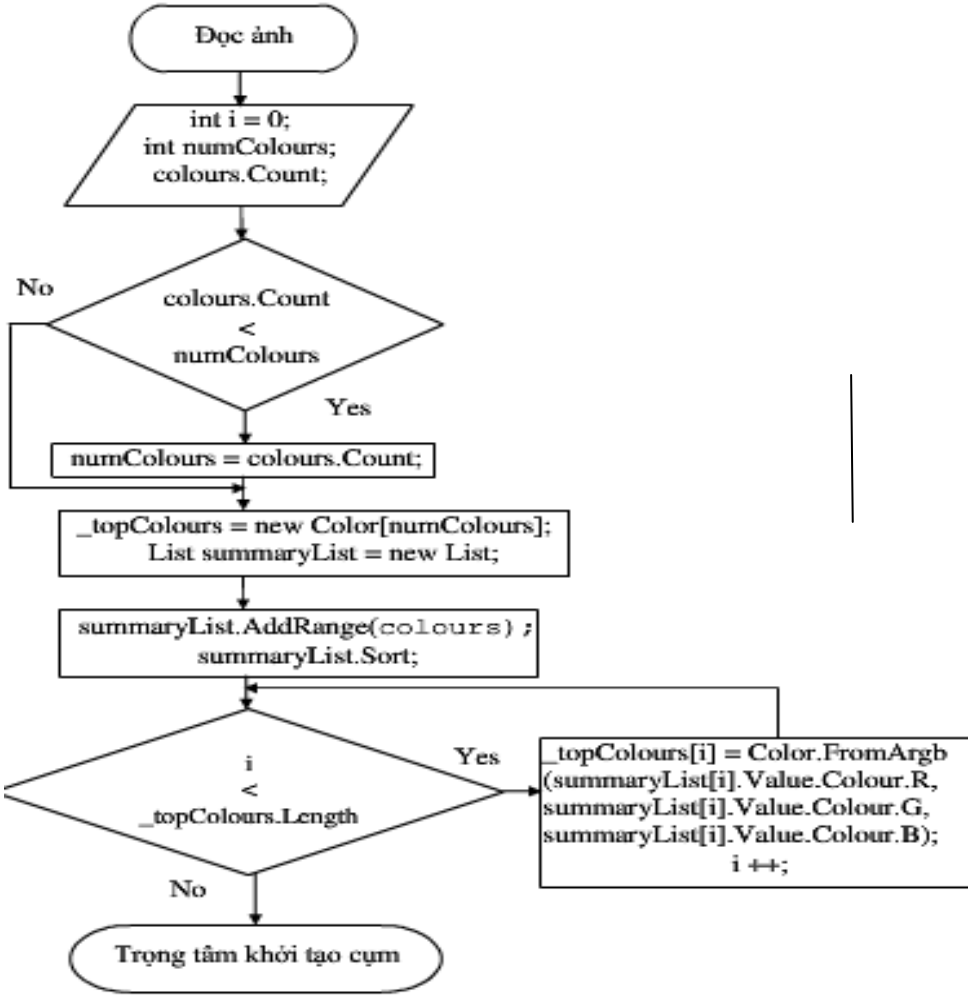
Lưu đồ tổng quát của thuật toán:



Hình 12: Thuật toán k - means

- **Tìm kiếm Top X Color**

Đầu tiên ta so sánh số màu thực tế 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 13: Tìm kiếm top x color

- **Tính khoảng cách và phân cụm:**

Dựa vào khoảng cách 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 nó tới tâm cụm là nhỏ nhất.

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2}$$

- **Tính lại trọng tâm**

Đối với mỗi cụm tính toán lại điểm trung tâm của nó dựa trên tất cả các điểm thuộc vào cụm đó.

- **Kiểm tra điều kiện hội tụ**

Quá trình phân cụm kết thúc nếu:

+ Không có (hoặc có không đáng kể) việc gán lại các điểm vào các cụm khác.

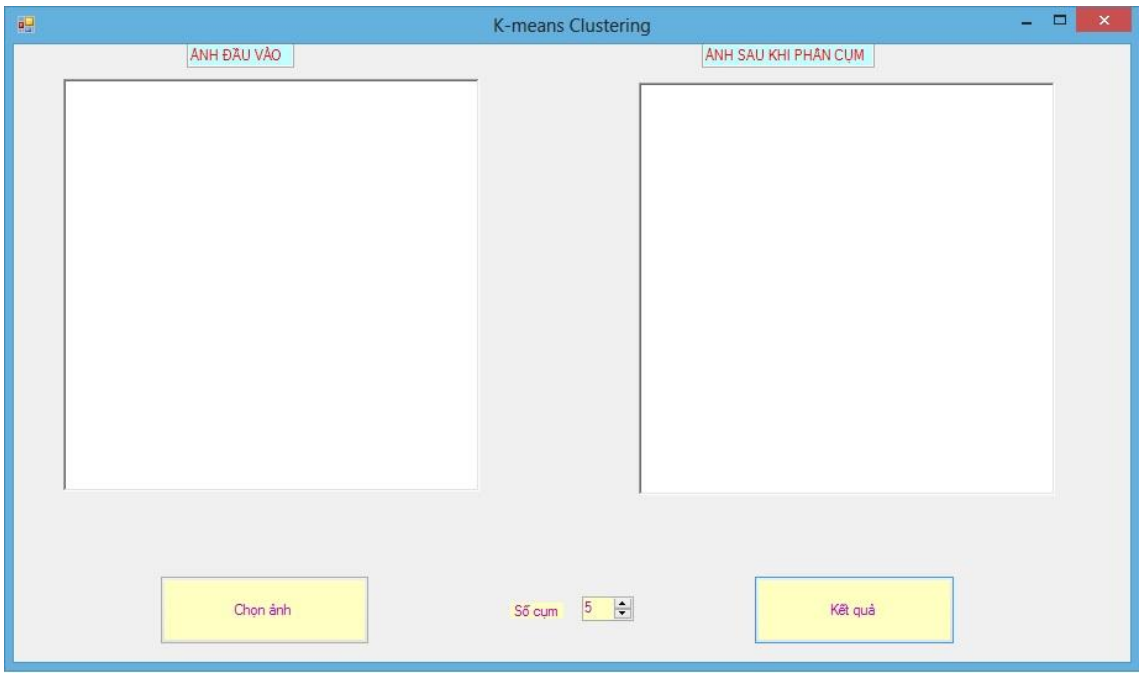
+ Không có (hoặc có không đáng kể) việc thay đổi trọng tâm của các cụm.

3.4.3. Kết quả thực nghiệm

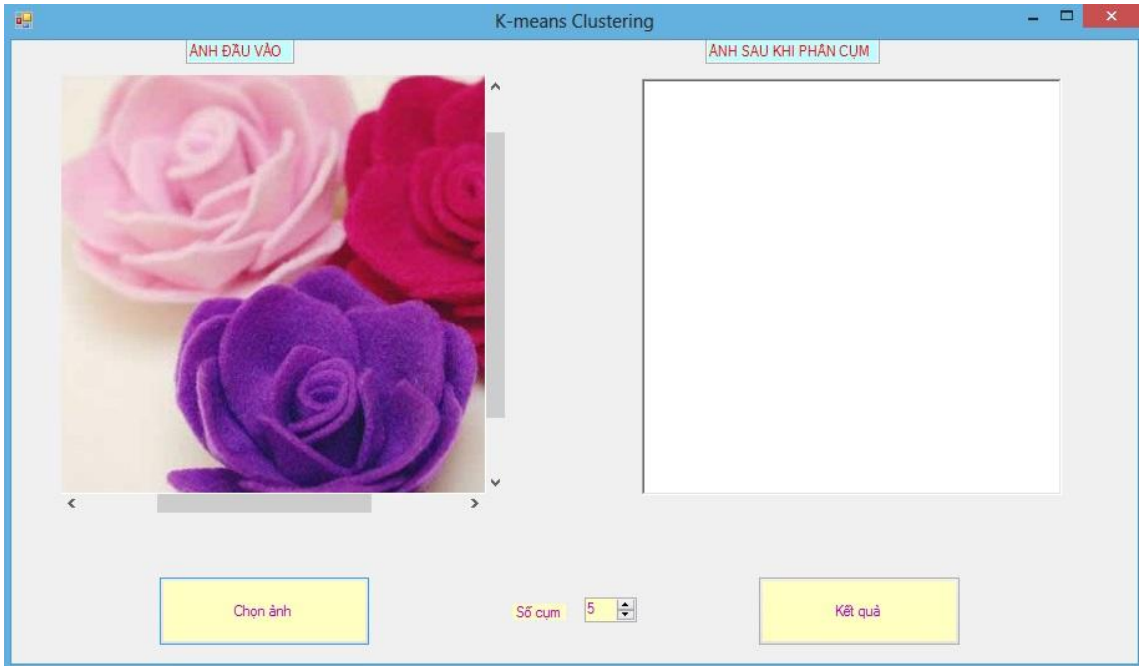
- **Môi trường cài đặt**

Chương trình được lập trình với ngôn ngữ C#, được cài đặt và chạy thử trên hệ điều hành Window.

- **Một số giao diện**



Hình 14: Giao diện chính của chương trình



Hình 15: Chọn ảnh đầu vào

Hình 16:Kết quả của quá trình phân cụm ảnh

3.4.4. Ưu, nhược điểm của thuật toán k – means

a)**Ưu điểm**

- Tương đối nhanh. Độ phức tạp của thuật toán là O (tkn), trong đó:

+ n: Số điểm trong không gian dữ liệu

+ k: Số cụm cần phân hoạch

+ t: Số lần lặp (t là khá nhỏ so với r)

- K-means phù hợp với các cụm có dạng hình cầu.

b) **Khuyết điểm**

+ Không đảm bảo đạt được tối ưu toàn cục và kết quả đầu ra phụ thuộc nhiều vào việc chọn k điểm khởi đầu. Do đó có thể phải chạy lại thuật toán với nhiều bộ khởi đầu khác nhau để có được kết quả đủ tốt. Trong thực tế, có thể áp dụng thuật giải di truyền để phát sinh các bộ khởi đầu.

+ Cần phải xác định trước số cụm.

+ Khó xác định số cụm thực sự mà không gian dữ liệu có. Do đó có thể phải thử với các giá trị k khác nhau.

+ Khó phát hiện các loại cụm có hình dạng phức tạp và nhất là các dạng cụm không lồi.

+ Không thể xử lý nhiễu và mẫu cá biệt.

+ Chỉ có thể áp dụng khi tính được trọng tâm.

KẾT LUẬN

A. Kết quả đạt được

Trong quá trình nghiên cứu và hoàn thành bài báo cáo đồ án tốt nghiệp "Phân cụm cứng trong phân đoạn ảnh", em đã thu nhận được thêm rất nhiều kiến thức và cũng nhận thấy phân cụm dữ liệu trong khai phá dữ liệu là một lĩnh vực rộng lớn và được ứng dụng rất mạnh mẽ. Hơn thế nữa nó còn rất nhiều vấn đề mà chúng ta cần khám phá. Trong đề tài của mình em đã cố gắng tìm hiểu và nghiên cứu tổng quan về khai phá dữ liệu, phân cụm dữ liệu và một số thuật toán của nó, tổng quan về phân đoạn ảnh. Cài đặt thử nghiệm thuật toán k – means và ứng dụng trong phân đoạn ảnh.

Do thời gian thực hiện hạn chế và kiến thức còn hạn chế nên em mới chỉ nghiên cứu được một số kỹ thuật cơ bản trong phân cụm dữ liệu, cài đặt thử nghiệm với thuật toán k – means. Còn một số kỹ thuật em vẫn chưa tìm hiểu, khai thác và ứng dụng vào các bài toán thực tế.

Mặc dù đã rất cố gắng, song do năng lực và trình độ có hạn nên trong quá trình thực hiện bài thực tập em đã không tránh khỏi những thiếu sót. Kính mong các thầy cô và các bạn quan tâm giúp đỡ chỉ bảo để chương trình của em một hoàn thiện hơn.

B. Hướng phát triển trong tương lai

Trong thời gian tới em sẽ cố gắng tiếp tục nghiên cứu, tìm hiểu thêm một số kĩ thuật phân cụm và nhất là có thể tìm hiểu và phát triển kỹ thuật phân đoạn ảnh để có thể xử lý với ảnh động.

TÀI LIỆU THAM KHẢO

[1] - Nguyễn Thị Ngọc, *Phân cụm dữ liệu dựa trên mật độ*, Đồ án tốt nghiệp đại học Ngành công nghệ Thông tin – ĐHDL Hải Phòng, 2008.

[2] - Trần Thị Quỳnh, *Thuật toán phân cụm dữ liệu nửa giám sát và giải thuật di truyền*, Đồ án tốt nghiệp đại học Ngành công nghệ Thông tin – ĐHDLHải Phòng, 2008.

[3] - Nguyễn. Lâm, *Thuật toán phân cụm dữ liệu nửa giám sát*,- Đồ án tốt nghiệp đại học Ngành công nghệ Thông tin – ĐHDL Hải Phòng, 2007.

[4] - Charles Elkan, *Department of Comput er Science and Engineering*, University of California, San Diego La jolla, CA 92093.

[5] - Andre w W. *Moore Associate Professoor School of Computer Science*, Carnegil Mellon University.

[6] - J.Han, M. Kamber and A.K.H. Tung, *Spatial Clustering Methods in Data Mining*, Sciences and Engineering Research Council of Canadia.