

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG



BÁO CÁO

Quản Trị Dự Án

**Đề tài : Nghiên cứu bài toán phân đoạn ảnh
và ứng dụng trong xử lý ảnh vệ tinh**

GVHD: Thầy Phạm Văn Hải

Nhóm sinh viên thực hiện

Trần Văn Thành MSSV 20133561

Đặng Văn Thuận MSSV 20133828

Trần Văn Hiếu MSSV 20151369

Hà Nội, tháng 5 năm 2018

Mục Lục

LỜI MỞ ĐẦU.....	3
1. Tổng quan.....	5
2. Phân đoạn hình ảnh bằng phân cụm.....	5
3. Các kỹ thuật phân cụm.....	6
3.1. K-means clustering	8
3.2. Đánh giá thuật toán K-Means.....	14
3.3. Phân đoạn ảnh màu	21
4. Kết quả thực nghiệm.....	22
5. Kết luận.....	33
6. Tài liệu tham khảo	35

Danh mục bảng

Bảng 1: Kết quả sau phân đoạn với $K = 4$, $\text{winSize} = 5$	25
Bảng 2: Kết quả sau phân đoạn khi thay đổi $K = 2 \rightarrow 6$, $\text{winSize} = 5$	26
Bảng 3: So sánh các phương pháp phân vùng ảnh.....	34

Danh mục hình ảnh

Hình 1: Các kỹ thuật phân cụm.....	7
Hình 2: Các thiết lập để xác định ranh giới các cụm ban đầu.....	9
Hình 3: Tính toán trọng tâm của các cụm mới.....	10
Hình 4: Các bước thuật toán K-Means.	11
Hình 5: Bài toán với 3 clusters.....	14
Hình 6: Ví dụ minh họa thuật toán K-Means.....	16
Hình 7: Ảnh đầu vào (cột trái) và ảnh sau phân đoạn (cột phải).	24
Hình 8: Ảnh đầu vào (2a), ảnh sau phân đoạn (2b-2f) với số cụm $K=2 \rightarrow 6$ và $\text{winSize} = 5$	26
Hình 9: Ảnh vệ tinh Hà Nội.....	27
Hình 10: Ảnh lũ lụt.	28
Hình 11: Ảnh phá rừng.	29
Hình 12: Ảnh đảo Trường Sa.....	30
Hình 13: Xác định kích thước khối U.....	31
Hình 14: Giao diện chương trình.	32

LỜI MỞ ĐẦU

Trong những năm gần đây, cùng với sự phát triển vượt bậc của khoa học máy tính, thì xử lý ảnh là một lĩnh vực đang được quan tâm. Nó là một ngành khoa học còn tương đối mới mẻ so với nhiều ngành khoa học khác, nó cũng là đối tượng nghiên cứu lĩnh vực thị giác máy tính, là quá trình biến đổi ảnh từ một ảnh ban đầu sang một ảnh mới. Nhờ có công nghệ số hóa hiện đại, ngày nay con người có thể xử lý tín hiệu nhiều chiều thông qua nhiều hệ thống khác nhau, từ những mạch số đơn giản cho đến những mạch song song cao cấp. Xử lý ảnh đã được nghiên cứu mạnh mẽ và đã có nhiều ứng dụng trong thực tế : như trong Y học, xử lý ảnh để phát hiện và nhận dạng khối u, cải thiện ảnh X quang, nhận dạng đường biên mạch máu từ những hình ảnh chụp bằng tia X, hay việc xử lý ảnh vệ tinh để tính toán tốc độ đô thị hóa, tốc độ chặt phá rừng, diện tích che phủ rừng...

Xử lý ảnh có liên quan đến nhiều ngành khác như: hệ thống thông tin, lý thuyết thông tin, lý thuyết thống kê, trí tuệ nhân tạo, nhận dạng...

Xử lý ảnh cũng đã tạo ra được rất nhiều ứng dụng hữu ích trong thực tế như bài toán nhận dạng vân tay, chữ viết, giọng nói...

Để phân tích các đối tượng trong ảnh, chúng ta cần phải phân biệt được các đối tượng cần quan tâm với phần còn lại của bức ảnh. Những đối tượng này có thể tìm ra được nhờ các kỹ thuật phân đoạn ảnh, theo nghĩa tách tiền cảnh ra khỏi hậu cảnh trong ảnh. Chúng ta cần phải hiểu được là:

- Không có kỹ thuật phân đoạn ảnh nào là vạn năng, theo nghĩa có thể đáp ứng cho mọi ảnh.
- Không có kỹ thuật phân đoạn ảnh nào là hoàn hảo.

Có thể hiểu phân vùng là quá trình chia ảnh thành nhiều vùng, mỗi vùng chứa một đối tượng hay nhóm đối tượng cùng kiểu. Chẳng hạn, một đối tượng có thể là một ký tự trên một trang văn bản hoặc một đoạn thẳng trong một bản vẽ kỹ thuật hoặc một nhóm các đối tượng có thể biểu diễn một từ hay đoạn thẳng tiếp xúc nhau.

Phân vùng ảnh còn là một thao tác ở mức thấp trong toàn bộ quá trình xử lý ảnh. Quá trình này thực hiện 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 vùng ảnh luôn đóng 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...

Mục đích của dự án

Xử lý ảnh vệ tinh để tính toán được tốc độ đô thị hóa, hay tốc độ sỏi mòn rừng, tốc độ chặt phá rừng trên 1 diện tích, theo phương pháp truyền thống từ xưa hay dùng các bản vẽ, dựa vào bản đồ và các chuyến đi thực nghiệm, nhưng không phải nơi đâu trên trái đất con người cũng có thể đặt chân đến để kiểm tra đo đạc. Vì vậy ngày nay với các công cụ khoa học hiện đại, từ vệ tinh con người có thể thu được các hình ảnh từ bất cứ nơi đâu trên trái đất. Qua các ảnh vệ tinh chụp được một khu rừng, sau khi xử lý phân đoạn ảnh (sử dụng thuật toán K-means) thành các vùng khác nhau ta tính được tỷ lệ phần trăm diện tích của các vùng trong tấm ảnh qua các năm, từ đó kết luận được tốc độ đô thị hóa hay tốc độ che phủ rừng, tốc độ chặt phá rừng... thay đổi qua các thời kỳ.

Xử lý ảnh trong y tế xác định vị trí kích thước khối U.

Từ những lý do trên, nhóm chúng em đã quyết định chọn đề tài “Nghiên cứu bài toán phân đoạn ảnh và ứng dụng trong xử lý ảnh vệ tinh” cho bài nghiên cứu của mình.

Chúng em xin chân thành cảm ơn thầy Phạm Văn Hải – giảng viên Bộ Môn Hệ Thống Thông Tin, Viện CNTT&TT, Trường ĐH BKHN, thầy đã nhiệt tình hướng dẫn, tạo điều kiện thuận lợi nhất cho nhóm chúng em trong quá trình nghiên cứu và thực hiện đề tài này.

Nhóm sinh viên thực hiện

Trần Văn Thành - 20133561

Đặng Văn Thuần - 20133828

Trần Văn Hiếu - 20151369

Hà Nội, tháng 5, năm 2018

mailto:Email: thanhtranvan.hust@gmail.com

1. Tổng quan

Phân đoạn ảnh là chủ đề nghiên cứu chính cho nhiều nghiên cứu về xử lý ảnh. Mục đích rõ ràng và nhiều ứng dụng vô tận: hầu hết các vấn đề phân tích hình ảnh và thị giác máy tính đòi hỏi một giai đoạn phân đoạn để phát hiện các đối tượng hoặc phân chia hình ảnh thành các vùng có thể coi là đồng nhất theo một tiêu chuẩn nhất định, chẳng hạn như màu sắc, kết cấu, văn bản... Kết quả của việc phân đoạn ảnh là một tập các vùng chung bao trùm toàn bộ hình ảnh, hay một tập các đường nét được trích xuất từ hình ảnh. Mỗi một điểm ảnh trong tập điểm ảnh trong một vùng là tương tự nhau với sự lưu ý về một vài tính chất hoặc thuộc tính toán chẳng hạn như màu sắc, cường độ và kết cấu... Kỹ thuật phân đoạn hình ảnh được phân thành 3 nhóm: Clustering, edge detection, region growing. Một số thuật toán phân cụm phổ biến như là K-means thì thường được dùng trong phân đoạn hình ảnh. Phân đoạn hình ảnh đề cập đến quá trình phân vùng một hình ảnh kỹ thuật số thành nhiều cụm khác nhau (các tập pixels).

Có nhiều phương pháp phân đoạn hình ảnh khác nhau. Ngưỡng biểu đồ giả định rằng hình ảnh bao gồm các vùng có khác nhau màu xám (hoặc màu) dao động, và tách nó ra thành một số mức, mỗi mức tương ứng với một vùng. Các thuật toán đường biên tiếp cận dựa trên sử dụng các toán phát hiện đường biên như Sobel, Laplacian... Khu vực kết quả có thể được kết nối, do đó đường biên cần phải được tham gia. Phương pháp tiếp cận dựa trên vùng được dựa trên sự giống nhau của dữ liệu hình ảnh vùng. Một số các phương pháp tiếp cận được sử dụng rộng rãi hơn trong thể loại này là: Ngưỡng, gom cụm, vùng tăng trưởng, chia nhỏ và sáp nhập. Gom nhóm hay gom cụm là việc tìm kiếm các nhóm riêng biệt trong không gian đặc trưng. Người ta cho rằng các nhóm này có cấu trúc khác nhau và có thể được phân biệt rõ ràng. Nhiệm vụ phân cụm tách các dữ liệu vào số lượng các phân vùng, đó là khối lượng trong không gian đặc trưng n -chiều.

2. Phân đoạn hình ảnh bằng phân cụm

Clustering là một kỹ thuật phân lớp. Cho một vector của N phép đo mô tả mỗi pixel hoặc nhóm các pixel (ví dụ như vùng) trong một hình ảnh tương tự như các vector đo lường và do đó phân nhóm của chúng trong không gian đo lường N -chiều giống nhau của các pixel tương ứng hoặc các nhóm pixel tương ứng. Vì vậy, phân nhóm trong không gian đo lường có thể là một chỉ số tương tự của các vùng hình ảnh, và có thể được sử dụng cho mục đích phân đoạn. Vector của các phép đo mô tả một số

tính năng hình ảnh hữu ích và do đó cũng được biết đến như là một vector tính năng. Giống nhau giữa các khu vực hình ảnh hoặc nhóm các pixel ước lượng (các khoảng cách phân nhỏ) trong không gian đặc trưng. Các phương pháp phân nhóm là một số trong những kỹ thuật phân chia nhỏ dữ liệu đầu tiên được phát triển.

Hầu hết các thuật toán phân cụm phổ biến bị hai nhược điểm chính. Thứ nhất, số các cụm được xác định trước, mà làm cho họ không đủ để xử lý hàng loạt các cơ sở dữ liệu hình ảnh rất lớn. Thứ hai, các cụm được đại diện bởi trọng tâm của chúng và được xây dựng bằng cách sử dụng một khoảng cách Euclidean do đó thường gây một hình dạng cụm hyperspheric, mà làm cho chúng không thể nắm bắt được cấu trúc thực sự của dữ liệu. Điều này đặc biệt đúng trong trường hợp của việc phân cụm nhóm màu được tùy tiện định hình.

3. Các kỹ thuật phân cụm

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp Unsupervised Learning trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các quy trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (Dissimilar) nhau.

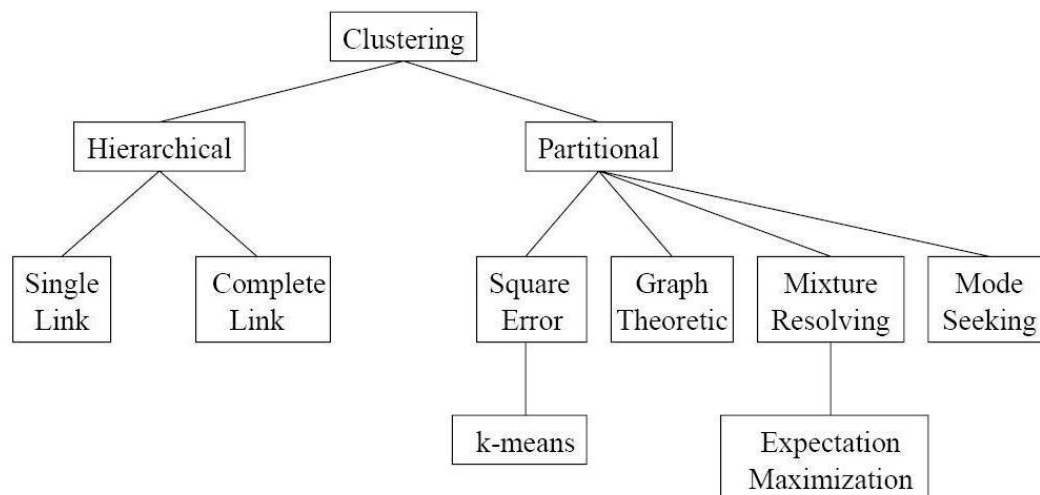
Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán phân cụm (Clustering Algorithms) đều sinh ra các cụm (clusters). Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh hiệu quả của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: “*data reduction*”, “*natural clusters*”, “*useful clusters*”, “*outlier detection*”.

Phân cụm là một kỹ thuật khai phá dữ liệu được sử dụng trong phân tích số liệu thống kê, khai phá dữ liệu, nhận dạng mẫu, phân tích hình ảnh ... Các phương pháp phân nhóm khác nhau bao gồm cụm phân cấp, trong đó xây dựng một hệ thống phân cấp của các cụm từ các thành phần riêng lẻ. Bởi vì tính đơn giản và hiệu quả của nó, phương pháp tiếp cận phân nhóm là một trong những kỹ thuật đầu tiên được sử dụng cho sự phân đoạn hình ảnh tự nhiên (kết cấu). Trong phân vùng nhóm; mục tiêu là để tạo ra một tập các cụm phân vùng dữ liệu trong các nhóm tương tự nhau. Các phương pháp khác của việc phân nhóm là khoảng cách dựa theo đó nếu hai hoặc nhiều đối tượng thuộc cùng một nhóm gần nhau theo một khoảng cách cho trước, nên nó được

gọi là khoảng cách được dựa theo cụm. Trong đề tài này, em đã sử dụng phương pháp tiếp cận phân nhóm K-means để thực hiện phân vùng ảnh bằng cách sử dụng phần mềm Matlab. Một phương pháp phân nhóm tốt sẽ tạo ra các cụm chất lượng cao. Chất lượng của các kết quả phân nhóm phụ thuộc vào cả hai biện pháp tương tự được sử dụng bởi các phương pháp và thực hiện của nó. Chất lượng của một phương pháp phân cụm cũng được đo bằng khả năng của nó để khám phá một số hoặc tất cả các mô hình ẩn. Phân đoạn hình ảnh là cơ sở phân tích hình ảnh, những kiến thức về phân đoạn hình ảnh là một phần rất quan trọng và vấn đề lâu đời nhất và khó khăn nhất của xử lý ảnh. Phân cụm có nghĩa là phân loại và phân biệt những thứ được cung cấp với các tính chất tương tự nhau. Kỹ thuật Clustering phân loại các điểm ảnh với các đặc điểm giống nhau thành một cụm, do đó tạo thành các cụm khác nhau theo sự gắn kết giữa các điểm ảnh trong một cụm.

🌈 Kỹ thuật phân cụm có thể áp dụng trong rất nhiều lĩnh vực như:

- Marketing: Xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng giá trị, phân loại và dự đoán hành vi khách hàng,...) sử dụng sản phẩm hay dịch vụ của công ty để giúp công ty có chiến lược kinh doanh hiệu quả hơn;
- Biology: Phân nhóm động vật và thực vật dựa vào các thuộc tính của chúng;
- Libraries: Theo dõi độc giả, sách, dự đoán nhu cầu của độc giả...;
- Insurance, Finance: Phân nhóm các đối tượng sử dụng bảo hiểm và các dịch vụ tài chính, dự đoán xu hướng (trend) của khách hàng, phát hiện gian lận tài chính (identifying frauds);
- WWW: Phân loại tài liệu (document classification); phân loại người dùng web (clustering weblog);...



Hình 1: Các kỹ thuật phân cụm.

3.1. K-means clustering

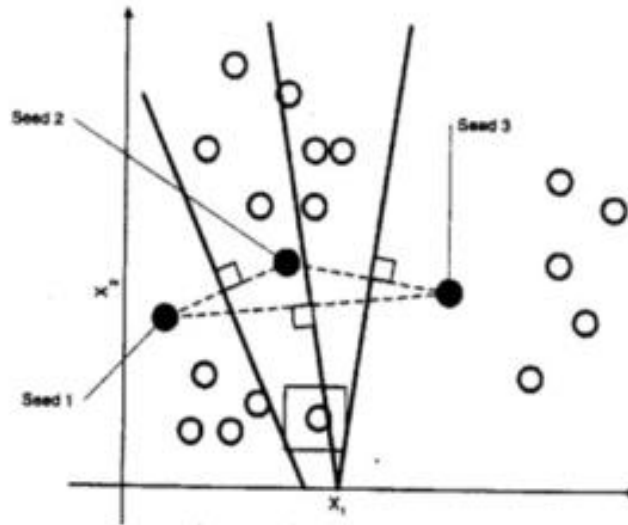
K-means là các thuật toán phân nhóm được sử dụng để xác định các nhóm quang phổ tự nhiên xuất hiện trong một tập dữ liệu. Thuật toán K-means là một thuật toán phân nhóm không được giám sát, phân loại các dữ liệu đầu vào thành nhiều lớp khác nhau dựa vào khoảng cách của chúng mỗi khác. Thuật toán giả định rằng các tính năng dữ liệu tạo thành một không gian vector và cố gắng để tìm cụm tự nhiên trong chúng. Tập dữ liệu được phân chia thành K cụm và các điểm dữ liệu được phân bổ ngẫu nhiên vào các cụm, kết quả là trong các cụm có cùng một số các điểm dữ liệu. Phân cụm là một cách để nhóm các đối tượng riêng biệt. K-means là phân cụm xử lý từng đối tượng là có một vị trí trong không gian. Nó tìm thấy phân vùng sao cho các đối tượng trong mỗi cụm được càng gần nhau càng tốt, và càng xa các đối tượng trong các cụm khác càng tốt. K-means là phân cụm yêu cầu bạn xác định số lượng các cụm được phân vùng và định lượng khoảng cách hai đối tượng với nhau. K-means được sử dụng để nhóm các đối tượng thành ba cụm bằng cách sử dụng số liệu khoảng cách Euclide. Phương pháp phổ biến nhất cho phân vùng ảnh là phân nhóm K-means.

K-means là thuật toán phân cụm mà định nghĩa các cụm bởi trung tâm của cá phần tử. Phương pháp này dựa trên độ đo khoảng cách của các đối tượng dữ liệu trong cụm, nó được xem như là trung tâm cụm. Như vậy nó cần khởi tạo một tập 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 này dừng lại khi các tâm cụm hội tụ.

Đối với mỗi điểm dữ liệu, khoảng cách từ các điểm dữ liệu đến mỗi cụm được tính toán. Nếu các điểm dữ liệu là gần nhất với nhóm riêng của mình, để nó lại nhóm riêng của nó. Nếu điểm dữ liệu không được gần nhất với cụm riêng của mình, di chuyển nó vào cụm gần nhất. Các bước trên được lặp đi lặp lại cho đến khi một đường chuyển hoàn chỉnh thông qua tất cả các điểm dữ liệu và không có điểm dữ liệu di chuyển từ một nhóm này đến nhóm khác. Tại thời điểm này các cụm ổn định và quá trình phân cụm kết thúc. Sự lựa chọn của phân vùng ban đầu ảnh hưởng rất lớn đến các cụm kết quả cuối cùng, cụm liên đới và khoảng cách cụm nội bộ và sự gắn kết với nhau. Tiếp cận thuật toán K-means là lặp đi lặp lại, tính toán chuyên sâu và do đó chỉ áp dụng cho những vùng con của hình ảnh hơn là những hình ảnh đầy đủ và có thể được giải quyết bằng bằng thuật toán huấn luyện không được giám sát.

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác định trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất.



Hình 2: Các thiết lập để xác định ranh giới các cụm ban đầu.

Thuật toán K-means lấy tham số đầu vào là K và phân chia một tập n đối tượng vào trong K cụm để cho kết quả độ tương đồng trong cụm là cao trong khi độ tương đồng ngoài cụm là thấp. Độ tương đồng cụm được đo khi đánh giá giá trị trung bình của các đối tượng trong cụm, nó có thể được quan sát như là “trọng tâm” của cụm. Giải thuật xử lý như sau: trước tiên nó lựa chọn ngẫu nhiên K đối tượng, mỗi đối tượng đại diện cho một trung bình cụm hay tâm cụm dữ liệu. Đối với những đối tượng còn lại, mỗi đối tượng sẽ được ấn định vào một cụm mà nó giống nhất dựa trên khoảng cách giữa đối tượng và trung bình cụm. Sau đó sẽ tính lại trung bình cụm mới cho mỗi cụm. Xử lý này sẽ được lặp lại cho tới khi hàm tiêu chuẩn hội tụ. Bình phương sai số thường dùng làm hàm tiêu chuẩn hội tụ, định nghĩa như sau. Mục đích của thuật toán K-means là sinh K cụm dữ liệu $\{C_1, C_2, \dots, C_k\}$ từ một tập dữ liệu chứa n đối tượng trong không gian d chiều $X_i = \{x_{i1}, x_{i2}, \dots, x_{id}\}$, $i=1 \div n$, sao cho hàm tiêu chuẩn:

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

' $|x - m_i|$ ' là khoảng cách Euclide giữa điểm dữ liệu x và tâm m_i .

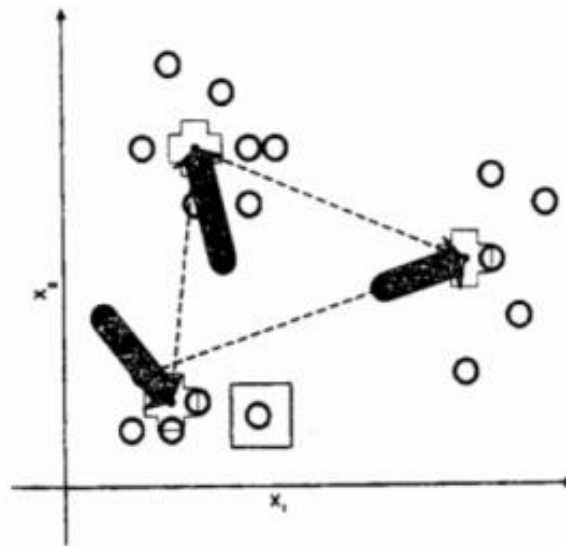
' C_i ' là tập điểm dữ liệu trong cụm thứ i^{th} cluster.

$\{m_1, m_2, m_3, \dots, m_k\}$ m_i là tâm cụm C_i .

$\{C_1, C_2, C_3, \dots, C_k\}$ là tập các cụm.

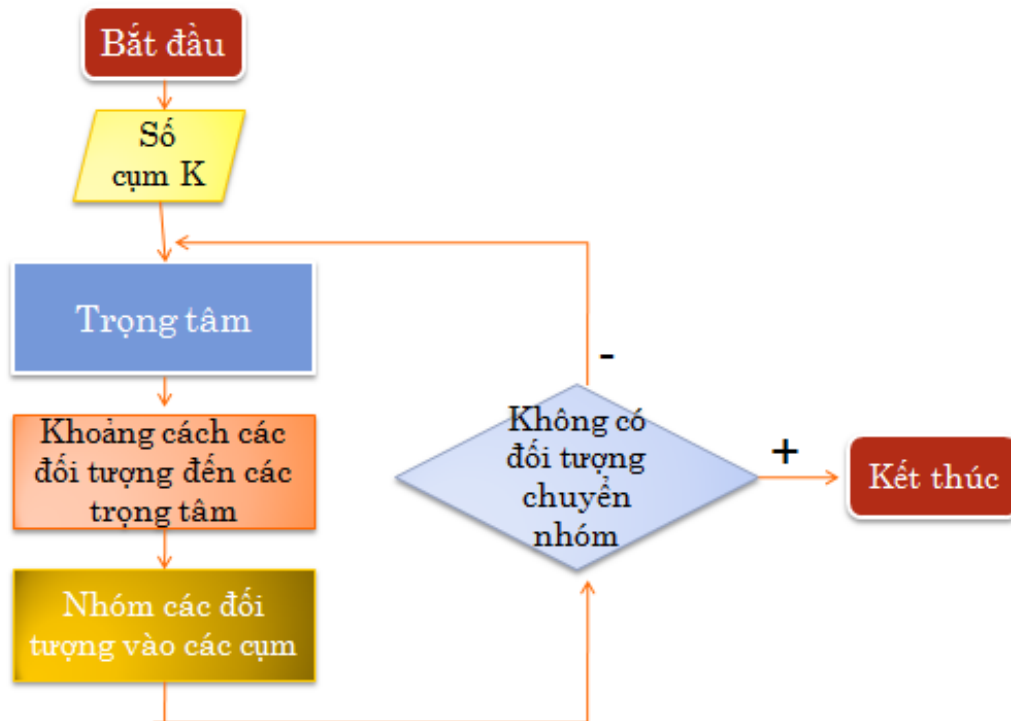
'k' là số lượng các cụm, tâm cụm.

Trọng tâm của cụm là một vector, trong đó giá trị của mỗi phần tử của nó là trung bình cộng của các thành phần tương ứng của các đối tượng vector dữ liệu trong cụm đang xét. Tham số đầu vào của thuật toán là K cụm, tham số đầu ra của thuật toán là các trọng tâm của các cụm dữ liệu. Độ đo khoảng cách giữa các đối tượng dữ liệu là khoảng cách Euclide vì đây là mô hình khoảng cách nên dễ lấy đạo hàm và xác định các cực trị tối thiểu. Hàm tiêu chuẩn và độ đo khoảng cách có thể được xác định cụ thể hơn tùy ý vào ứng dụng hoặc quan điểm của người dùng.



Hình 3: Tính toán trọng tâm của các cụm mới.

Thuật toán K-Means được mô tả như sau:



Hình 4: Các bước thuật toán K-Means.

Thuật toán K-Means thực hiện qua các bước chính sau:

Input : Tập dữ liệu X , số cụm K và hàm $E = \sum_{i=1}^k \sum_{x \in C_i} |x - m_i|^2$

Output : Các cụm C_1 ($1 \leq i \leq k$) tách rời với hàm tiêu chuẩn E đạt GTNN

Begin

1. Khởi tạo

Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.

2. Tính toán khoảng cách

Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean):

$$D_{j=1}^k \sqrt{\sum_{i=1}^n (x_i - m_j)^2}$$

Đối với mỗi điểm $x_i (1 \leq i \leq n)$, tính toán khoảng cách của nó tới mỗi trọng tâm $m_j (1 \leq j \leq k)$. Sau đó tìm trọng tâm gần nhất đối với mỗi điểm và nhóm chúng vào các nhóm gần nhất.

3. Cập nhật lại trọng tâm:

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

$$v_j = (1/c_j) \sum_{i=1}^{c_j} x_i$$

ở đây, ' c_j ' đại diện cho số các điểm dữ liệu trong cụm thứ j^{th} .

4. Gán lại các điểm gần trung tâm nhóm mới

Nhóm các đối tượng vào nhóm gần nhất dựa trên trọng tâm của nhóm.

Điều kiện dừng: Thực hiện lại bước 2 và 3 cho đến khi không có sự thay đổi nhóm nào của các đối tượng

End.

K-means biểu diễn các cụm bởi các trọng tâm của các đối tượng trong cụm đó. Thuật toán K-means chi tiết được trình bày:

BEGIN

1. Nhập n đối tượng dữ liệu
2. Nhập k cụm dữ liệu
3. $MSE = +\infty$;
4. For I = 1 to k do $m_i = X_{i+(i-1)*[n/k]}$; % khởi tạo k trọng tâm
5. Do {
6. OldMSE = MSE;
7. MSE' = 0;
8. For j = 1 to k do
9. { $m'[j] = 0$; $n'[j] = 0$ }
10. End for
11. For I = 1 to n do
12. For j = 1 to k do

```

13.           $D^2(x[i]; m[j]);$  % khoảng cách Euclide bình phương
14.      End for
15.      % Tìm trọng tâm gần nhất  $m[h]$  tới  $X[i]$ 
16.           $m'[h] = m'[h] + X[i]$  ;  $n'[h] = n'[h] + 1$ ;
17.           $MSE' = MSE' + D^2(x[i]; m[j]);$ 
18.      End for
19.       $n[j] = \max(n'[j], 1)$ ;  $m[j] = m'[j]/n[j]$ ;
20.       $MSE' = MSE'$ ;
21.      } while( $MSE' \leq OldMSE$ );

```

END

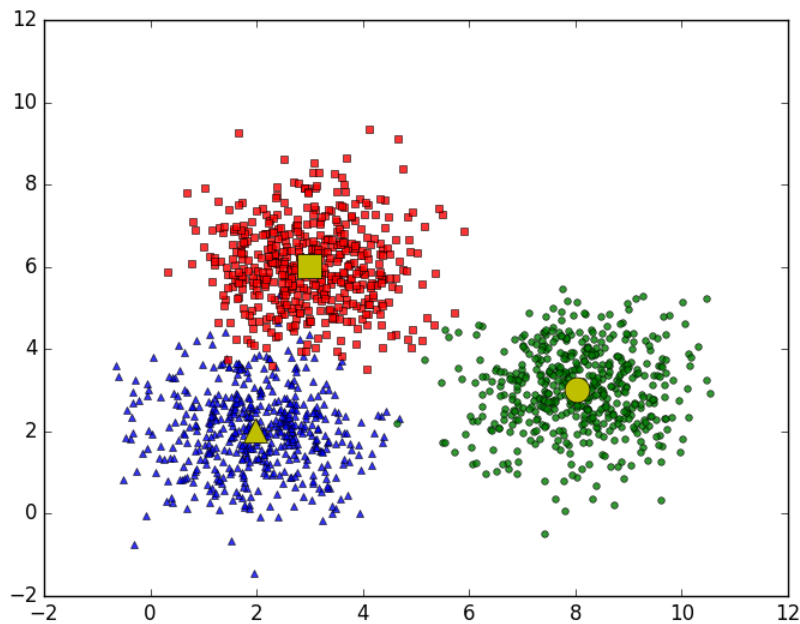
Trong đó :

- MSE : Sai số bình phương trung bình hay là hàm tiêu chuẩn.
- $D^2(x[i]; m[j])$: khoảng cách Euclide từ đối tượng thứ i tới trọng tâm j .
- $OldMSE$ $m'[j]$, $n'[j]$: biến tạm lưu giá trị cho trạng thái trung gian cho các biến tương ứng.

Chất lượng thuật toán K-means phụ thuộc nhiều vào các tham số đầu vào như : số cụm k , và k trọng tâm khởi tạo ban đầu. Trong trường hợp các trọng tâm khởi tạo ban đầu mà quá lệch so với các trọng tâm cụm tự nhiên thì kết quả phân cụm K-means là rất thấp, nghĩa là các cụm dữ liệu được khám phá rất lệch so với thực tế. Trên thực tế chưa có một giải pháp nào để chọn tham số đầu vào, giải pháp thường được sử dụng nhất là thử nghiệm với các giá trị đầu vào k khác nhau rồi sau đó chọn giải pháp tốt nhất.

Thuật toán K-means được chứng minh là hội tụ và có độ phức tạp tính toán là $O(tkn)$ với t là số lần lặp, k là số cụm, n là số đối tượng của tập dữ liệu vào. Thông thường $k \ll n$ và $t \ll n$ thường kết thúc tại một điểm tối ưu cục bộ.

Tuy nhiên, nhược điểm của K-means là còn rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu. Hơn nữa, chất lượng phân cụm dữ liệu của thuật toán K-means phụ thuộc nhiều vào các tham số đầu vào như: số cụm K và K trọng tâm khởi tạo ban đầu. Trong trường hợp các trọng tâm khởi tạo ban đầu mà quá lệch so với các trọng tâm cụm tự nhiên thì kết quả phân cụm của K-means là rất thấp, nghĩa là các cụm dữ liệu được khám phá rất lệch so với các cụm trong thực tế. Trên thực tế chưa có một giải pháp tối ưu nào để chọn các tham số đầu vào, giải pháp thường được sử dụng nhất là thử nghiệm với các giá trị đầu vào K khác nhau rồi sau đó chọn giải pháp tốt nhất.



Hình 5: Bài toán với 3 clusters.

Link ảnh goo.gl/x1vqyh

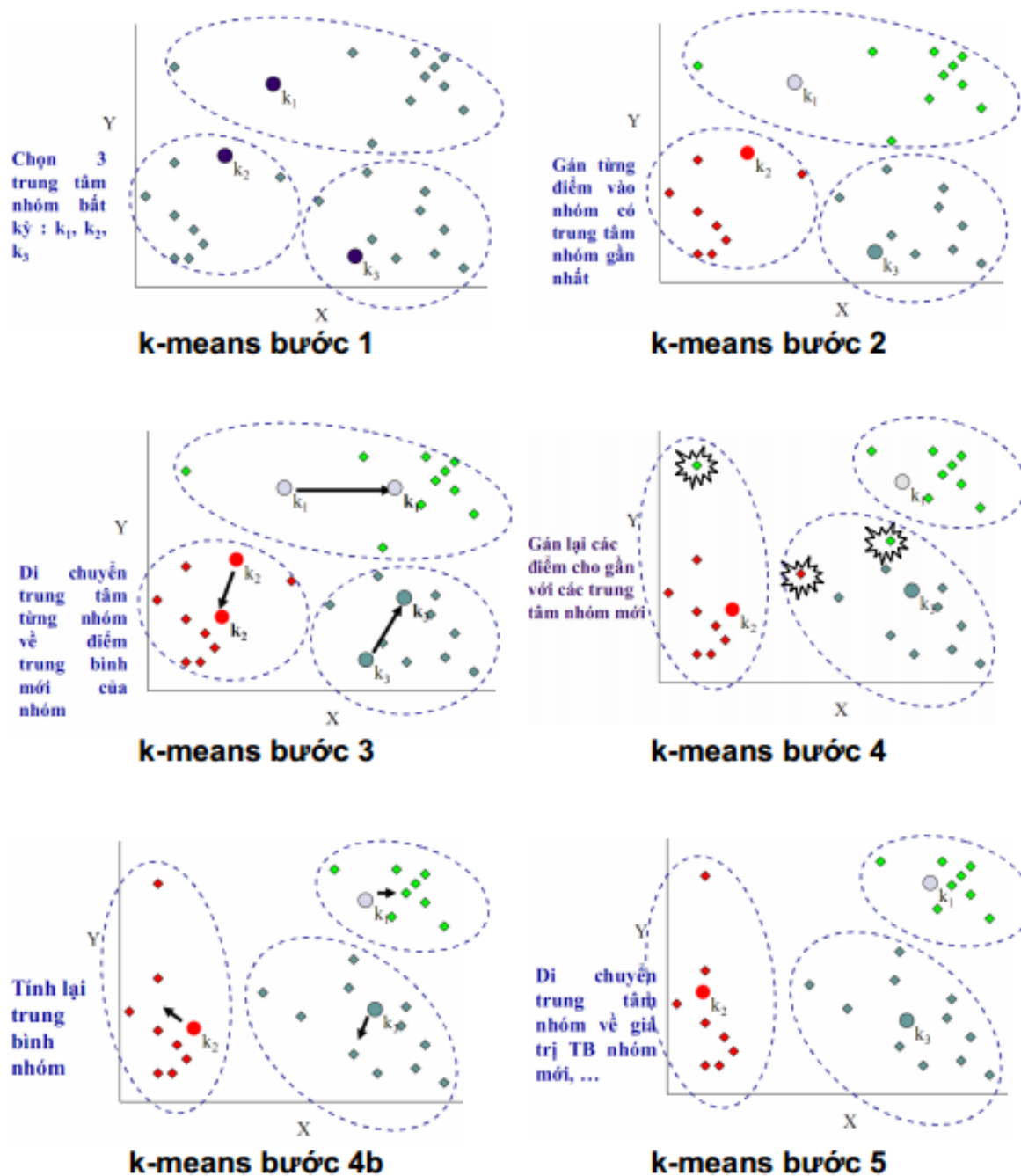
- X :tập dữ liệu, C_i : cụm thứ i
- $X = C_1 \cup C_2 \dots \cup C_k$
- $C_i \cap C_j = \emptyset, \forall 1 \leq i, j \leq k$: số cụm

3.2. Đánh giá thuật toán K-Means

- Ưu điểm :
 - ✓ K-means là có độ phức tạp tính toán $O(tkn)$, t : số lần lặp
 - ✓ Có khả năng mở rộng, có thể dễ dàng sửa đổi với những dữ liệu mới.
 - ✓ Bảo đảm hội tụ sau 1 số bước lặp hữu hạn.
 - ✓ Luôn có ít nhất 1 điểm dữ liệu trong 1 cụm dữ liệu, luôn có K cụm dữ liệu.
 - ✓ Các cụm không phân cấp và không bị chồng chéo dữ liệu lên nhau.
 - ✓ Mọi thành viên của 1 cụm là gần với chính cụm đó hơn bất cứ 1 cụm nào khác.
 - ✓ K-means phân tích phân cụm đơn giản nên có thể áp dụng đối với tập dữ liệu lớn.
- Nhược điểm :

- Không khắc phục được nhiều và giá trị k phải được cho bởi người dùng.
 - Không có khả năng tìm ra các cụm không lỗi hoặc các cụm có hình dạng phức tạp.
 - Khó khăn trong việc xác định các trọng tâm cụm ban đầu
 - Chọn ngẫu nhiên các trung tâm cụm lúc khởi tạo
 - Độ hội tụ của thuật toán phụ thuộc vào việc khởi tạo các vector trung tâm cụm
 - Khó để chọn ra được số lượng cụm tối ưu ngay từ đầu, mà phải qua nhiều lần thử để tìm ra được số lượng cụm tối ưu.
 - Rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu.
 - Không phải lúc nào mỗi đối tượng cũng chỉ thuộc về 1 cụm, chỉ phù hợp với đường biên giữa các cụm rõ.
 - Chỉ thích hợp áp dụng với dữ liệu có thuộc tính số và khám ra các cụm có dạng hình cầu.
- ❖ **Ví dụ :** Giả sử có một tập đối tượng được định vị trong hệ trục tọa độ X, Y . Cho $k = 3$ tức người dùng cần phân các đối tượng vào trong 3 cụm.

Theo giải thuật, ta chọn ngẫu nhiên 3 trung tâm cụm ban đầu (**Hình kmeans bước 1**). Sau đó, mỗi đối tượng được phân vào trong các cụm đã chọn dựa trên tâm cụm gần nhất (**Hình k-means bước 2**). Cập nhật lại các tâm (Hình k-means bước 3). Đó là giá trị trung bình của mỗi cụm được tính toán lại dựa trên các đối tượng trong cụm. Tùy theo các tâm mới này, các đối tượng được phân bố lại vào trong các cụm dựa trên tâm cụm gần nhất (Hình k-means bước 4).

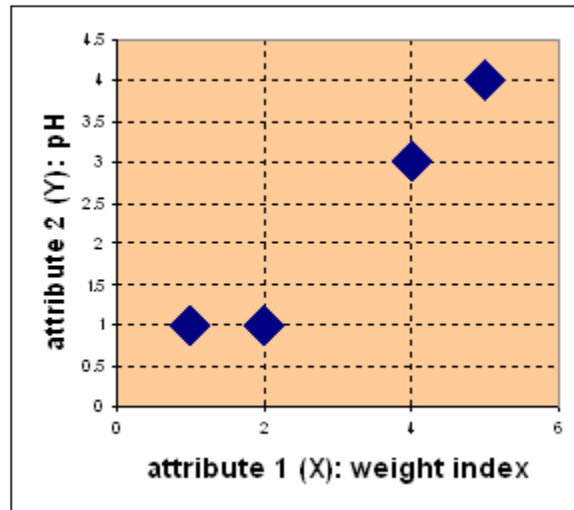


Hình 6: Ví dụ minh họa thuật toán K-Means.

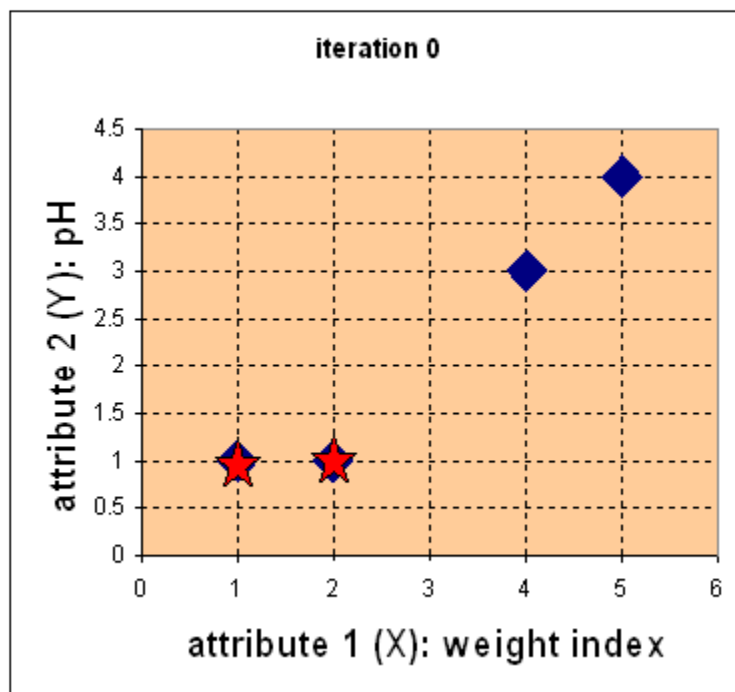
Ví dụ minh họa thuật toán K-Means:

Giả sử ta có 4 loại thuốc A, B, C, D, mỗi loại thuốc được biểu diễn bởi 2 đặc trưng X và Y như sau. Mục đích của ta là nhóm các thuốc đã cho vào 2 nhóm ($K=2$) dựa vào các đặc trưng của chúng.

Object	Feature 1 (X): weight index	Feature 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4



Bước 1. Khởi tạo tâm (centroid) cho 2 nhóm. Giả sử ta chọn A là tâm của nhóm thứ nhất (tọa độ tâm nhóm thứ nhất $c_1(1,1)$) và B là tâm của nhóm thứ 2 (tọa độ tâm nhóm thứ hai $c_2(2,1)$).



Bước 2. Tính khoảng cách từ các đối tượng đến tâm của các nhóm (Khoảng cách Euclidean)

$$D^0 = \begin{matrix} & \begin{matrix} A & B & C & D \end{matrix} \\ \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} & \begin{matrix} c_1 = (1,1) \text{ group-1} \\ c_2 = (2,1) \text{ group-2} \end{matrix} \end{matrix}$$

$$\begin{matrix} \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & \begin{matrix} X \\ Y \end{matrix} \end{matrix}$$

Mỗi cột trong ma trận khoảng cách (D) là một đối tượng (cột thứ nhất tương ứng với đối tượng A, cột thứ 2 tương ứng với đối tượng B,...). Hàng thứ nhất trong ma trận khoảng cách biểu diễn khoảng cách giữa các đối tượng đến tâm của nhóm thứ nhất (c_1) và hàng thứ 2 trong ma trận khoảng cách biểu diễn khoảng cách của các đối tượng đến tâm của nhóm thứ 2 (c_2).

Ví dụ, khoảng cách từ loại thuốc C=(4,3) đến tâm $c_1(1,1)$ là 3.61 và đến tâm $c_2(2,1)$ là 2.83 được tính như sau:

$$c_1 = (1,1) \quad \sqrt{(4-1)^2 + (3-1)^2} = 3.61$$

$$c_2 = (2,1) \quad \sqrt{(4-2)^2 + (3-1)^2} = 2.83$$

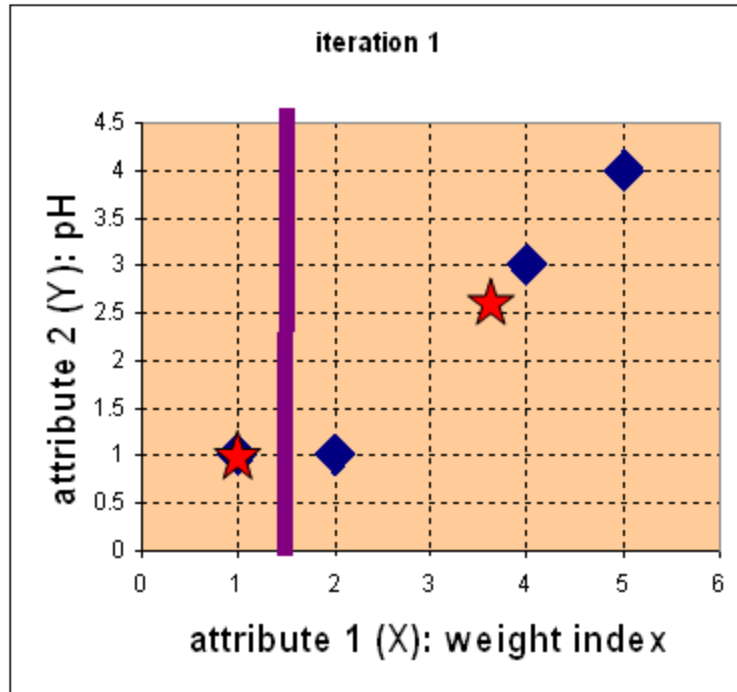
Bước 3. Nhóm các đối tượng vào nhóm gần nhất

$$G^0 = \begin{matrix} & \begin{matrix} A & B & C & D \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} & \begin{matrix} \text{group-1} \\ \text{group-2} \end{matrix} \end{matrix}$$

Ta thấy rằng nhóm 1 sau vòng lặp thứ nhất gồm có 1 đối tượng A và nhóm 2 gồm các đối tượng còn lại B,C,D.

Bước 4. Tính lại tọa độ các tâm cho các nhóm mới dựa vào tọa độ của các đối tượng trong nhóm. Nhóm 1 chỉ có 1 đối tượng A nên tâm nhóm 1 vẫn không đổi, $c_1(1,1)$. Tâm nhóm 2 được tính như sau:

$$c_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3} \right) = \left(\frac{11}{3}, \frac{8}{3} \right)$$



Bước 5. Tính lại khoảng cách từ các đối tượng đến tâm mới:

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{array}{l} \mathbf{c}_1 = (1,1) \text{ group - 1} \\ \mathbf{c}_2 = (\frac{11}{3}, \frac{8}{3}) \text{ group - 2} \end{array}$$

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	
	1	2	4	5	<i>X</i>
	1	1	3	4	<i>Y</i>

Bước 6. Nhóm các đối tượng vào nhóm:

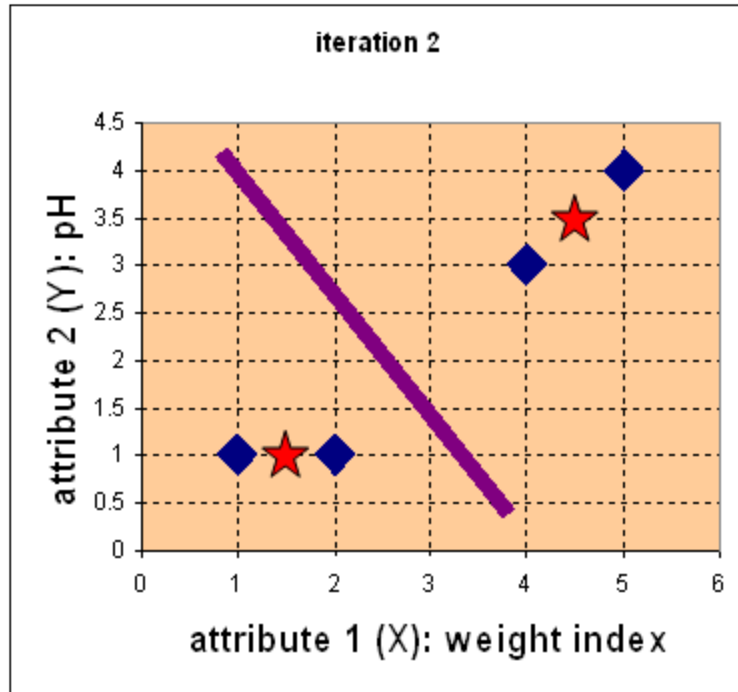
$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{array}{l} \text{group - 1} \\ \text{group - 2} \end{array}$$

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
--	----------	----------	----------	----------

Bước 7. Tính lại tâm cho nhóm mới:

$$\mathbf{c}_1 = (\frac{1+2}{2}, \frac{1+1}{2}) = (1\frac{1}{2}, 1)$$

$$\mathbf{c}_2 = (\frac{4+5}{2}, \frac{3+4}{2}) = (4\frac{1}{2}, 3\frac{1}{2})$$



Bước 8. Tính lại khoảng cách từ các đối tượng đến tâm mới:

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1\frac{1}{2}, 1) & \text{group - 1} \\ \mathbf{c}_2 = (4\frac{1}{2}, 3\frac{1}{2}) & \text{group - 2} \end{matrix}$$

$$\begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & \begin{matrix} X \\ Y \end{matrix} \end{matrix}$$

Bước 9. Nhóm các đối tượng vào nhóm:

$$G^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} \text{group - 1} \\ \text{group - 2} \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

Ta thấy $G_2 = G_1$ (Không có sự thay đổi nhóm nào của các đối tượng) nên thuật toán dừng và kết quả phân nhóm như sau:

Object	Feature 1 (X): weight index	Feature 2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

3.3. Phân đoạn ảnh màu

Mục đích cơ bản là phân đoạn màu sắc một cách tự động bằng cách sử dụng $L \times a \times b$ không gian màu và phân cụm K-means. Toàn bộ quá trình có thể được tóm tắt trong các bước sau:

Bước 1: Đọc hình ảnh

Bước 2: Chuyển hình ảnh từ không gian màu RGB sang không gian màu $L \times a \times b$. Không gian $L \times a \times b$ bao gồm một '*' L' lớp sáng, lớp kết tủa màu 'a*' và lớp kết tủa màu 'b*'. Tất cả các thông tin màu sắc nằm trong các lớp '*' a' và 'b *'. Chúng ta có thể đo lường sự khác biệt giữa hai màu sắc bằng cách sử dụng khoảng cách tiêu chuẩn Euclide.

Bước 3: Phân loại màu sắc trong không gian " $a \times b$ " sử dụng K-means clustering.

Bước 4: Đánh nhãn mỗi pixel trong hình ảnh sử dụng kết quả từ K-means. Đối với tất cả các đối tượng trong đầu vào, K-means trả về một chỉ số tương ứng với một cluster. Ghi nhãn mỗi điểm ảnh trong hình ảnh với chỉ số cluster của nó.

Bước 5: Tạo hình ảnh những hình ảnh mà phân đoạn bởi màu sắc. Sử dụng các nhãn điểm ảnh, chúng ta phải chia đối tượng trong hình ảnh bằng màu sắc, sẽ cho kết quả bằng ba hình ảnh khác nhau.

Bước 6: Phân khúc hạt nhân vào một hình ảnh riêng biệt. Sau đó, lập trình xác định các chỉ số của cụm có chứa các đối tượng màu xanh bởi vì K-means sẽ không trả lại cùng một giá trị cluster_idx mỗi lần khác nhau. Chúng ta có thể làm điều này bằng cách sử dụng các giá trị trung tâm cụm, chứa 'a*' và 'b *' giá trị cho mỗi cụm.

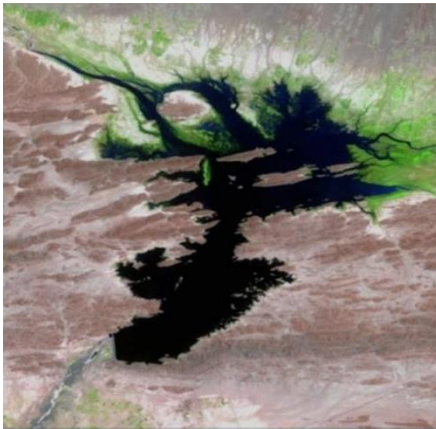
Giả sử màu sắc bề mặt của các đối tượng trong ảnh là một thuộc tính không đổi và màu sắc đó được ánh xạ vào một không gian 2 chiều và màu. Khi đó áp dụng giải thuật phân cụm K-means cho việc xác định các cụm màu, mỗi cụm màu có tập các điểm ảnh tương tự nhau. Sau khi phân đoạn ảnh, mỗi điểm ảnh chỉ thuộc về một vùng duy nhất. Những vùng duy nhất này thông thường sẽ tương ứng với toàn bộ hay từng phần của các đối tượng thật sự có trong ảnh.

4. Kết quả thực nghiệm

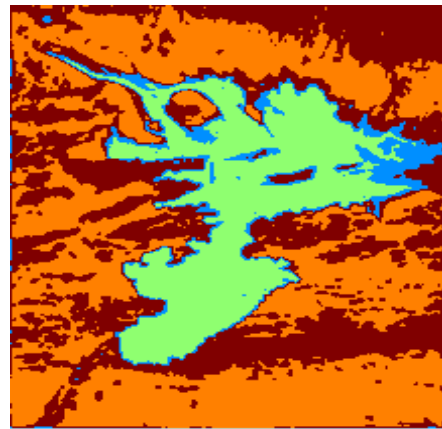
Bộ dữ liệu ảnh màu vệ tinh, mỗi ảnh có dung lượng khoảng là 140 Kb được lấy từ trang <https://earthengine.google.com/timelapse/>.

Chương trình được viết bằng ngôn ngữ Matlab v8.3.0.532 (R2014a) và được thực hiện trên máy tính có vi bộ vi xử lý CPU Intel core i5-3320M-2.6Ghz và bộ nhớ 8GB RAM.

✚ Kết quả thử nghiệm 1 số hình ảnh với số cụm $K = 4$, winSize = 5



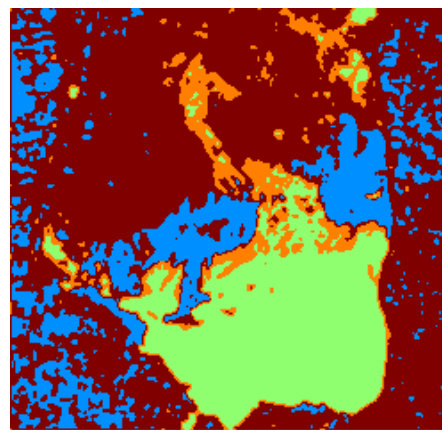
1(a)



1(b)



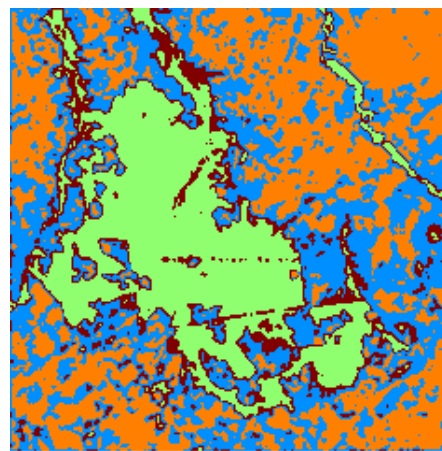
1(c)



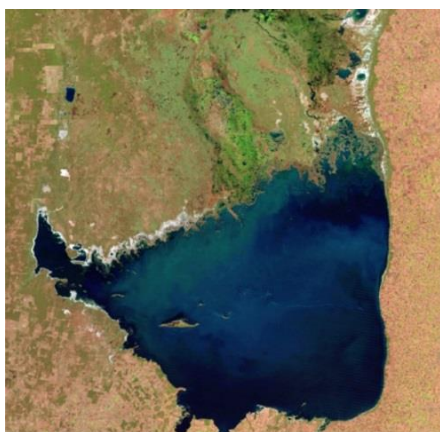
1(d)



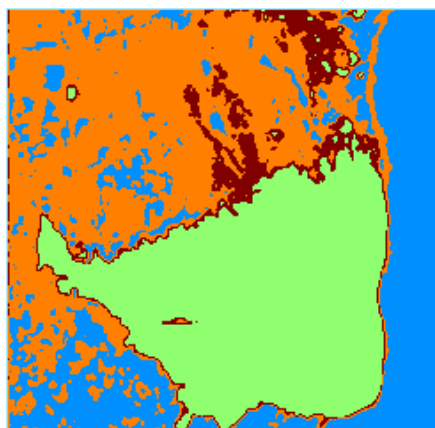
1(e)



1(f)



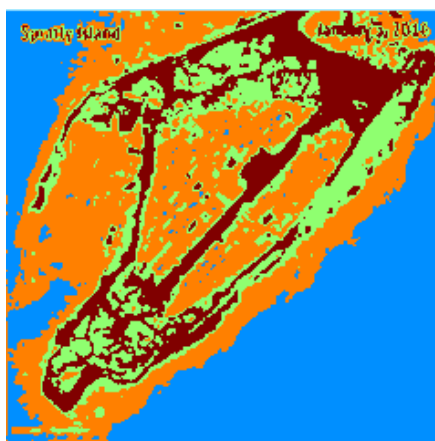
1(g)



1(h)



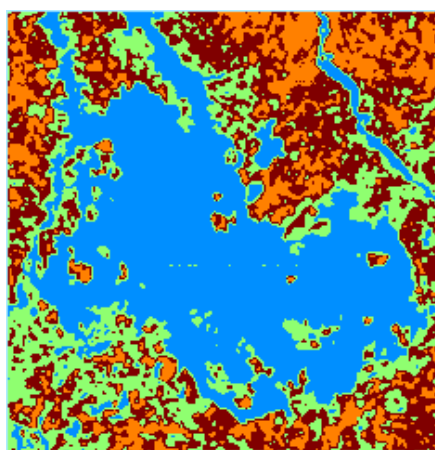
1(i)



1(k)



1(l)



1(m)

Hình 7: Ảnh đầu vào (cột trái) và ảnh sau phân đoạn (cột phải).

Bảng 1: Kết quả sau phân đoạn với $K = 4$, winSize = 5

Ảnh gốc	Số cụm K	winSize	Thời gian chạy(s)
1(a)	4	5	2.16458
1(c)	4	5	2.00838
1(e)	4	5	2.06011
1(g)	4	5	1.58454
1(i)	4	5	3.12197
1(l)	4	5	1.70087

🌈 Khi thay đổi số cụm K

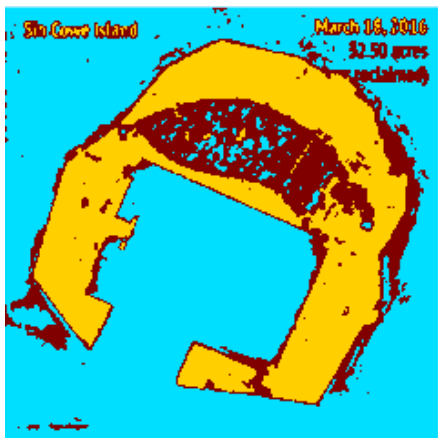
Với cùng 1 ảnh đầu vào, số cụm do người dùng chọn là $K = 2, 3, 4, 5, 6...$ được điều chỉnh sao cho phù hợp với các đối tượng trong ảnh ban đầu.



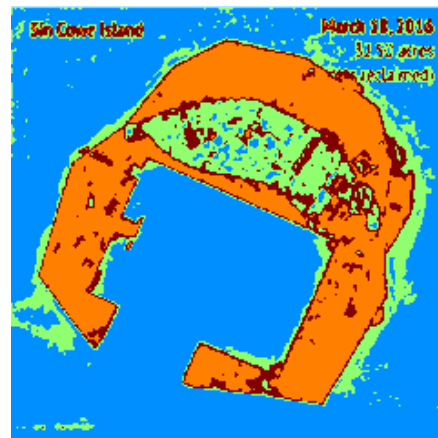
2(a)



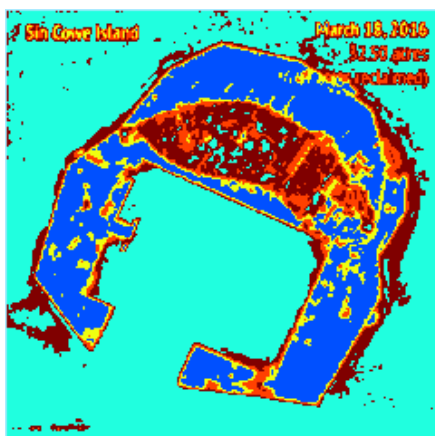
2(b)



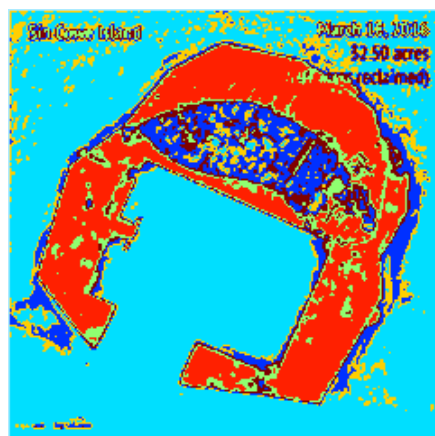
2(c)



2(d)



2(e)



2(f)

Hình 8: Ảnh đầu vào (2a), ảnh sau phân đoạn (2b-2f) với số cụm $K=2 \rightarrow 6$ và $\text{winSize} = 5$.

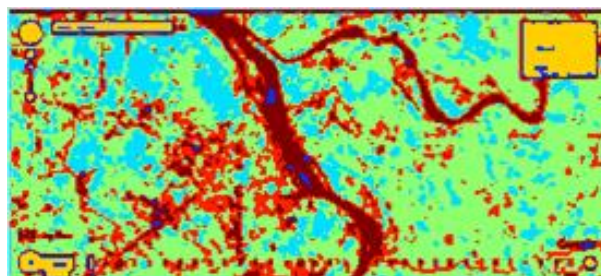
Bảng 2: Kết quả sau phân đoạn khi thay đổi $K = 2 \rightarrow 6$, $\text{winSize} = 5$

Số cụm K	winSize	Thời gian chạy(s)
2	5	3.07647
3	5	3.22492
4	5	3.51832
5	5	3.74447
6	5	4.08005

- Một số kết quả xử lý ảnh vệ tinh
- Hình ảnh Hà Nội



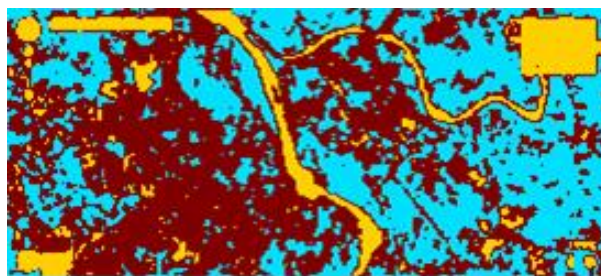
Hà Nội – 1984



Đô thị = 17.9805% K = 6



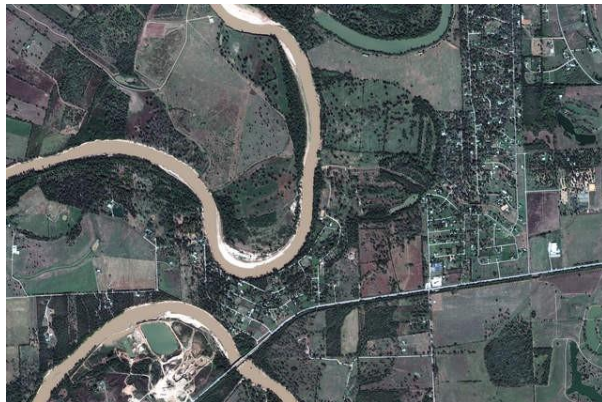
Hà Nội – 2016



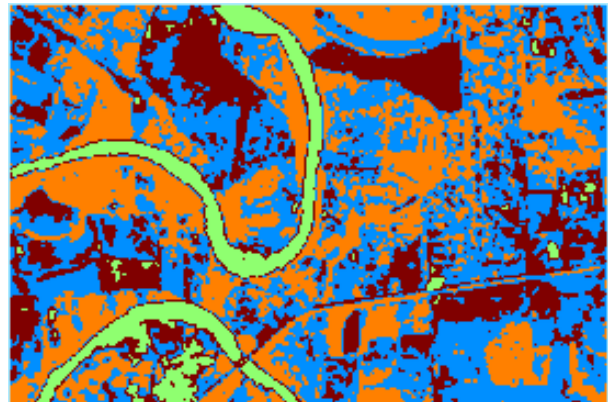
Đô thị = 51.4025% K = 3

Hình 9: Ảnh vệ tinh Hà Nội.

- Hình ảnh lũ lụt



Trước lũ



Nước = 6.76083% K = 5



Khi lũ



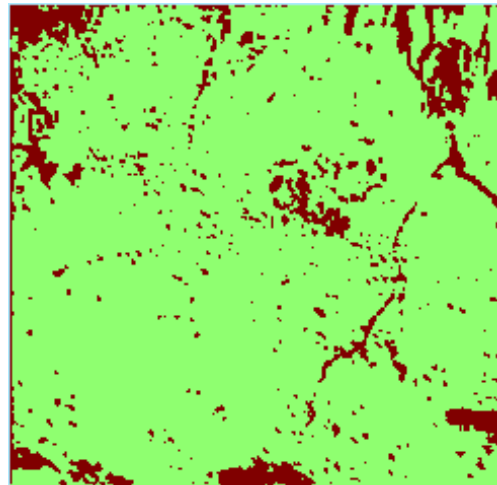
Nước = 56.5563% K = 2

Hình 10: Ảnh lũ lụt.

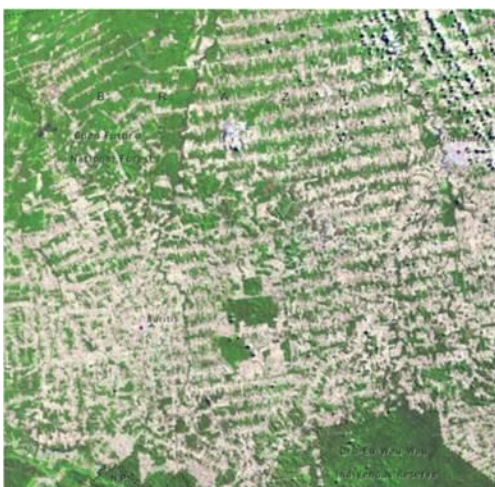
- Hình ảnh phá rừng



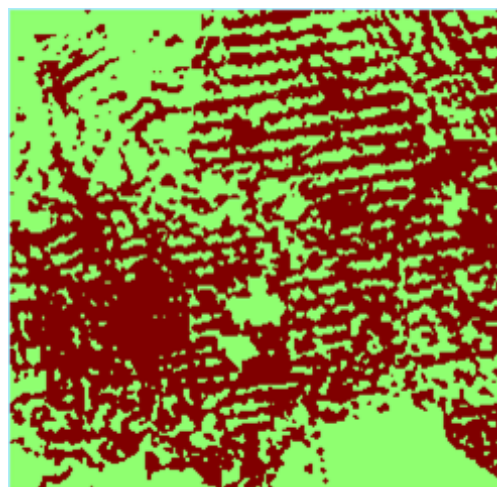
1975



Cây = 87.8043% K = 2



2009



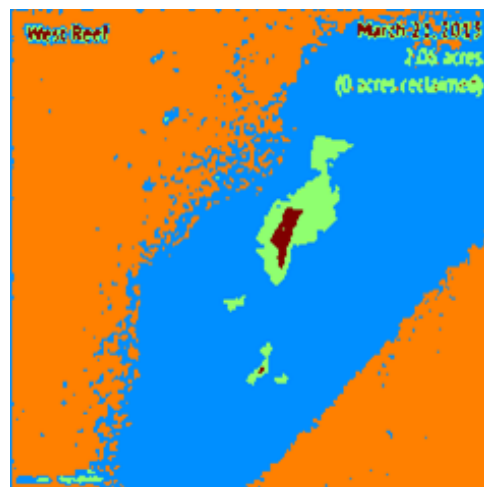
Cây = 47.3041% K = 2

Hình 11: Ảnh phá rừng.

- Hình ảnh đảo Đá tây



2013



Đất = 0.808158% K = 4



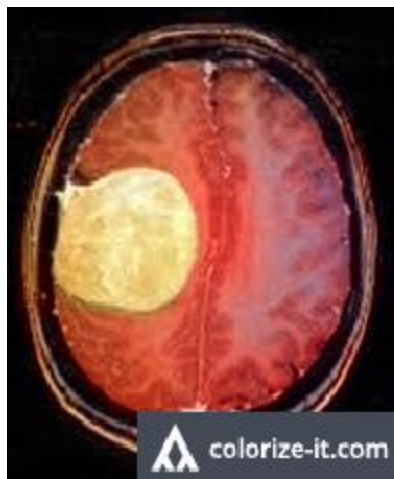
2016



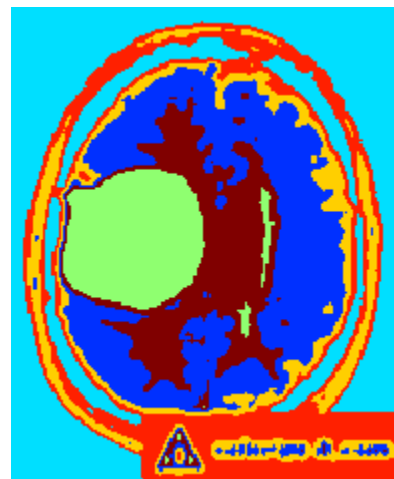
Đất = 17.6693% K = 2

Hình 12: Ảnh đảo Trường Sa.

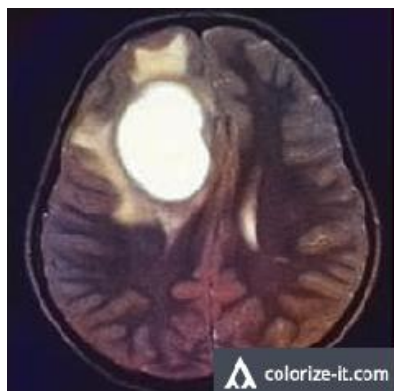
- ✚ Kết quả xử lý ảnh y tế
- Xác định kích thước khối U



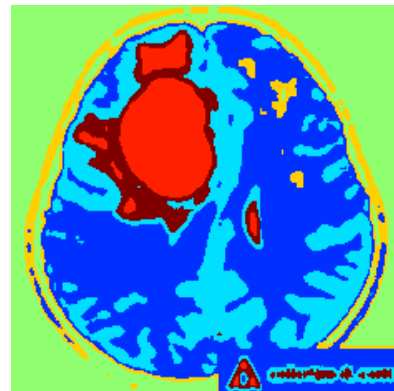
Tumor 1



$$U = 8.8889\% \quad K = 6$$




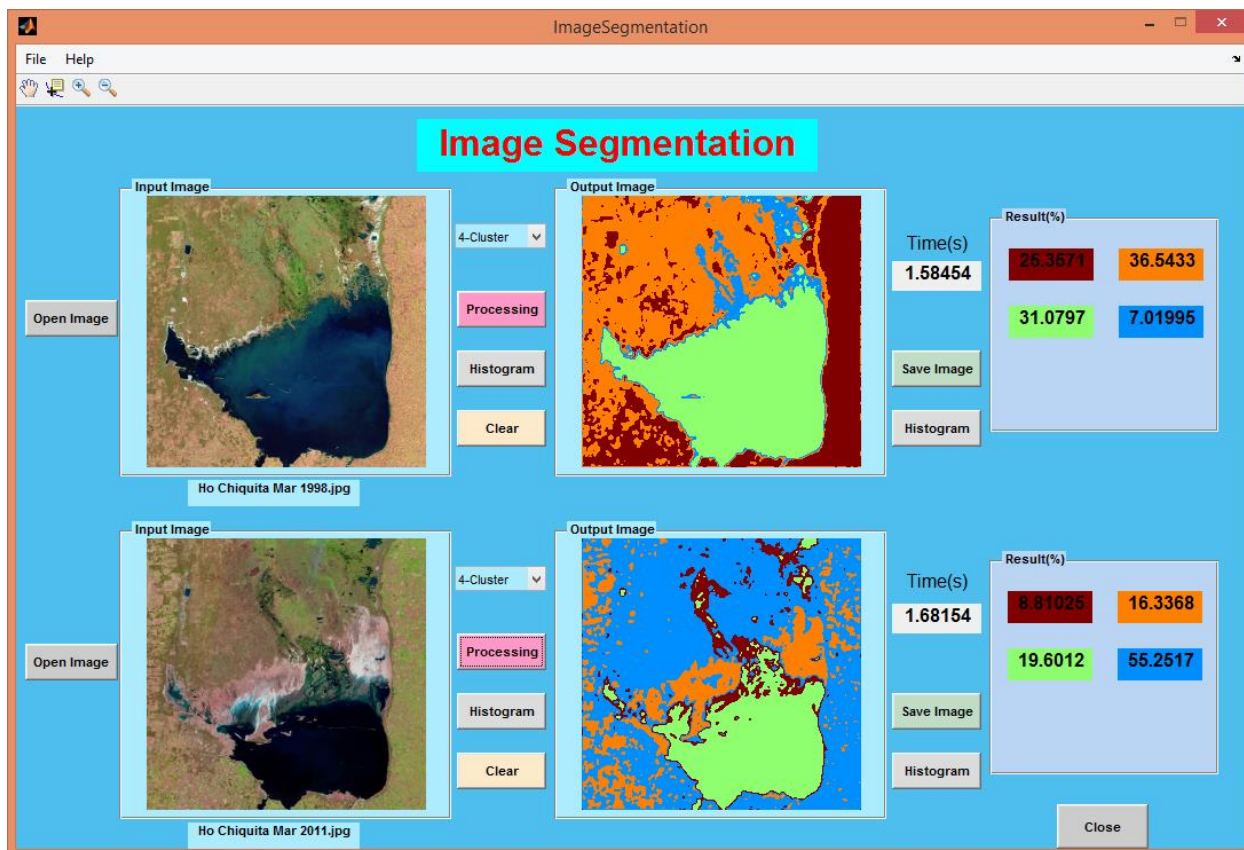
Tumor 2



$$U = 6.49719\% \quad K = 6$$

Hình 13: Xác định kích thước khối U.

 Giao diện chương trình:



Hình 14: Giao diện chương trình.

1. **Open Image** : Mở thư mục và chọn 1 ảnh màu RGB với định dạng cho phép.
2. **K-Cluster** : Chọn tham số cluster (số cụm trong thuật toán K-means) $K=2,3,4,5,6$.
3. **Processing** : Nút xử lý ảnh, sau khi nhấn nút này ảnh đầu ra, các thông tin kết quả ảnh hiện lên.
4. **Clear** : Xóa ảnh và các thông tin kết quả xử lý ảnh khỏi màn hình.
5. **Save Image** : Save as ảnh sau khi phân vùng (lưu dạng đen trắng).
6. **Histogram** : Hiện đồ thị histogram của ảnh trước và sau khi xử lý.
7. **Toolbar** : Gồm các tính năng zoom in, zoom out ảnh, Data cursor xem chi tiết từng pixel trên ảnh, Pan để kéo ảnh di chuyển ngang dọc.
8. **Menubar** : File(Open mở 1 ảnh, Quit thoát chương trình).
9. **Help** : About(thông tin về phần mềm).
10. **Close** : Đóng chương trình.

5. Kết luận

Sử dụng phân đoạn hình ảnh dựa trên màu sắc, nó có thể làm giảm được chi phí tính toán và tránh được tính toán đặc trưng cho từng pixel trong hình ảnh. Mặc dầu, màu sắc không thường xuyên được sử dụng cho phân đoạn ảnh. Nó cho thấy được sự phân biệt rõ ràng của các vùng đại diện trong ảnh. Loại phân đoạn ảnh này có thể được dùng để lập bản đồ về sự thay đổi độ che phủ của đất qua các thời kỳ, phát triển một hình ảnh chính xác và đáng tin cậy hơn mà có thể được sử dụng trong xác định vị trí khối u, nhận dạng khuôn mặt, nhận dạng vân tay và trong việc định vị một đối tượng rõ ràng từ một hình ảnh vệ tinh. Ưu điểm của thuật toán K-means là đơn giản và khá hiệu quả. Nó hoạt động tốt khi các cụm không phân biệt được một cách rõ ràng chẳng hạn như hình ảnh trên website...

- ✓ Giảm được chi phí tính toán.
- ✓ Phân biệt rõ ràng các vùng đại diện trong ảnh.
- ✓ Có thể dùng để lập bản đồ về sự thay đổi độ che phủ của đất qua các thời kỳ.
- ✓ Xác định kích thước khối u trong y tế.

So sánh các phương pháp phân vùng ảnh

Bảng 3: So sánh các phương pháp phân vùng ảnh

Phương pháp phân vùng	Ưu điểm	Nhược điểm
Featured-based techniques(tính năng kỹ thuật)		
Clustering(cụm)	<ul style="list-style-type: none"> Phân loại không cần giám sát Tồn tại các kinh nghiệm cải tiến (heuristic) và hữu hạn 	<ul style="list-style-type: none"> Không quan tâm đến các thông tin trong không gian ảnh Có vấn đề trong việc xác định số lượng các cụm ban đầu (Clustering) Khó khăn trong việc điều chỉnh các cụm (Clustering) sao cho phù hợp với cá vùng trong ảnh
Adaptive Clustering	<ul style="list-style-type: none"> Sở hữu tính liên tục trong không gian và tính thích nghi cục bộ với các vùng ảnh Sử dụng các ràng buộc 	<ul style="list-style-type: none"> Cực đại hóa một xác xuất hậu điều kiện có thể bị sai do các cực trị địa phương Hội tụ chậm
Histogram thresholding	<ul style="list-style-type: none"> Không cần biết trước bất kỳ thông tin nào từ ảnh Các giải thuật nhanh và dễ dàng cài đặt 	<ul style="list-style-type: none"> 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

6. Tài liệu tham khảo

Thuật toán K-Means với bài toán phân cụm dữ liệu. (2010). Retrieved from <http://bis.net.vn/forums/t/374.aspx>

Bá, L. M., & Thủy, N. T. (2008). *Nhập Môn Xử Lý Ảnh*. Hà Nội: ĐHBKHN.

P, Ganesan; V.Rajini, Dr;. (2013, May No.5). Application of Modified K-Means Clustering Algorithm for Satellite Image. *International Journal of Advanced Research in Computer Science*, Volume 4(Special Issue), 41-44.

Tiep, V. H. (2018). *Machine Learning Co Ban*.