

TRƯỜNG ĐẠI HỌC ĐIỆN LỰC
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN
KHAI PHÁ DỮ LIỆU

ĐỀ TÀI:

ÁP DỤNG THUẬT TOÁN K-MEANS VÀO BÀI TOÁN
PHÂN CỤM HOA IRIS

Sinh viên thực hiện : NGUYỄN TRỌNG HUY
ĐỖ NGUYỄN THIÊN KHIÊM
Giảng viên hướng dẫn : TS. NGUYỄN THỊ THANH TÂN
Ngành : CÔNG NGHỆ THÔNG TIN
Chuyên ngành : CÔNG NGHỆ PHẦN MỀM
Lớp : D13CNPM5
Khóa : 2018-2023

Hà Nội, tháng 3 năm 2021

PHIẾU CHẤM ĐIỂM

STT	Họ tên	Phân công công việc	Điểm	Chữ ký
1	Nguyễn Trọng Huy MSV-18810310375			
2	Đỗ Nguyễn Thiện Khiêm MSV-18810310442			

Phiếu chấm điểm của giảng viên

Giảng viên chấm điểm	Chữ ký	Ghi chú
Giảng viên chấm 1:		
Giảng viên chấm 2:		

MỤC LỤC

LỜI CẢM ƠN	1
LỜI MỞ ĐẦU	2
CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU	3
1.1. Khái niệm về khai phá dữ liệu	3
1.2. Các bước của quá trình khai phá dữ liệu	5
1.3. Khai phá dữ liệu và công nghệ cơ sở dữ liệu	6
1.4. Các tác vụ của khai phá dữ liệu	6
1.5. Kiến trúc của một hệ thống khai phá dữ liệu	8
1.6. Ý nghĩa và vai trò của khai phá dữ liệu	9
1.7. Ứng dụng của khai phá dữ liệu	10
1.8. Kết luận chương 1	10
CHƯƠNG 2: TỔNG QUAN VỀ THUẬT TOÁN K-MEANS.....	11
2.1. Giới thiệu thuật toán phân cụm trong học máy	11
2.2. Thuật toán K-Means	11
2.2.1. Các bước chính thực hiện thuật toán K-Means	12
2.2.2. Ví dụ minh họa thuật toán K-Means	12
2.3. Kết luận chương 2	17
CHƯƠNG 3: ÁP DỤNG THUẬT TOÁN K-MEANS VÀO BÀI TOÁN PHÂN CỤM HOA IRIS	18
3.1. Mô tả bài toán	18
3.2. Mô tả tập dữ liệu	18
3.3. Áp dụng thuật toán K-means vào bài toán	21
KẾT LUẬN	29
TÀI LIỆU THAM KHẢO.....	30

MỤC LỤC HÌNH ẢNH

Hình 1. 1: Các lĩnh vực trong khai phá dữ liệu.....	5
Hình 1. 2: Các bước của khai phá dữ liệu.....	5
Hình 1. 3: Các tác vụ trong khai phá dữ liệu	8
Hình 1. 4: Kiến trúc của một hệ thống khai phá dữ liệu.....	9

LỜI CẢM ƠN

Thực tế thì không có sự thành công nào mà không gắn liền với học tập và thực hành. Kèm theo đó chính là sự hỗ trợ, sự giúp đỡ từ giảng viên hướng dẫn và sự tìm tòi, học hỏi của bản thân. Trong suốt quá trình học tập ở giảng đường Đại học đã đến nay, em đã nhận được rất nhiều sự quan tâm, giúp đỡ của thầy cô, gia đình và bạn bè.

Với lòng biết ơn sâu sắc nhất, em xin gửi đến thầy cô ở Khoa Công Nghệ Thông Tin- trường Đại Học Điện Lực đã truyền đạt vốn kiến thức quý báu cho chúng em trong suốt thời gian học tập tại trường. Và đặc biệt, trong kỳ này, em được tiếp cận với môn học rất hữu ích đối với sinh viên ngành Công Nghệ Thông Tin. Đó là môn: ***“Khai phá dữ liệu”***.

Chúng em xin chân thành cảm ơn cô *Nguyễn Thị Thanh Tân* đã tận tâm hướng dẫn chúng em qua từng buổi học trên lớp cũng như những buổi nói chuyện, thảo luận về môn học. Trong thời gian được học tập và thực hành dưới sự hướng dẫn của cô, em không những thu được rất nhiều kiến thức bổ ích, mà còn được truyền sự say mê và thích thú đối với bộ môn ***“Khai phá dữ liệu”***. Nếu không có những lời hướng dẫn, dạy bảo của cô thì chúng em nghĩ báo cáo này rất khó có thể hoàn thành được.

Mặc dù đã rất cố gắng hoàn thiện báo cáo với tất cả sự nỗ lực. Tuy nhiên, do thời gian có hạn mà đây lại là bước đầu tiên đi vào thực tế, và vốn kiến thức còn hạn chế, nhiều bỡ ngỡ, nên báo cáo Quản lý dự án xây dựng phần mềm ***“Áp dụng thuật toán K-Means vào bài toán phân cụm hoa Iris”*** chắc chắn sẽ không thể tránh khỏi những thiếu sót. Em rất mong nhận được sự quan tâm, thông cảm và những đóng góp quý báu của các thầy cô và các bạn để báo cáo này được hoàn thiện hơn.

Chúng em xin trân trọng cảm ơn quý thầy cô giáo!

LỜI MỞ ĐẦU

Trong ngành khoa học dữ liệu, ta thường nghĩ đến việc sử dụng dữ liệu để huấn luyện mô hình và tạo các dự đoán đối với dữ liệu mới. Đó gọi là học có giám sát (supervised learning). Tuy nhiên, nhiều lúc chúng ta không muốn tạo các dự đoán mà muốn tìm điểm các tương đồng trong dữ liệu và phân dữ liệu vào các nhóm khác nhau. Đó gọi là học không giám sát (unsupervised learning).

Phân cụm là một trong những dạng học không giám sát được ứng dụng nhiều nhất. Trong bài viết này, chúng ta sẽ tìm hiểu về K-means Clustering, một thuật toán được sử dụng phổ biến vì sự đơn giản trong tính toán và tốc độ xử lý nhanh.

Chính vì thế, chúng em xin được làm đề tài này để trình bày về các cơ sở lý thuyết tra cứu phân cụm hoa Iris sử dụng thuật toán K-Means. Tên đề tài: “**Áp dụng thuật toán K-Means vào bài toán phân cụm hoa Iris**”.

CHƯƠNG 1: TỔNG QUAN VỀ KHÁI PHÁ DỮ LIỆU

1.1. Khái niệm về khai phá dữ liệu

Khai phá dữ liệu được định nghĩa là quá trình trích xuất các thông tin có giá trị tiềm ẩn bên trong lượng lớn dữ liệu được lưu trữ trong các cơ sở dữ liệu, kho dữ liệu. Cụ thể hơn đó là tiến trình trích lọc, sản sinh những tri thức hay những mẫu tiềm ẩn, chưa biết nhưng hữu ích từ các cơ sở dữ liệu lớn. Đồng thời là tiến trình khái quát các sự kiện rời rạc trong dữ liệu thành các tri thức mang tính khái quát, tính qui luật hỗ trợ tích cực cho các tiến trình ra quyết định. Hiện nay, ngoài thuật ngữ khai phá dữ liệu, người ta còn dùng một số thuật ngữ khác có ý nghĩa tương tự như: Khai phá tri thức từ CSDL (Knowledge mining from database), trích lọc dữ liệu (Knowledge extraction), phân tích dữ liệu/mẫu (data/pattern analysis), khảo cổ dữ liệu (data archaeology), nạo vét dữ liệu (data dredging). Nhiều người coi khai phá dữ liệu và một số thuật ngữ thông dụng khác là khám phá tri thức trong CSDL (Knowledge Discovery in Databases- KDD) là như nhau. Tuy nhiên trên thực tế khai phá dữ liệu chỉ là một bước thiết yếu trong quá trình Khám phá tri thức trong CSDL.

Sau đây là một số định nghĩa mang tính mô tả của nhiều tác giả về khai phá dữ liệu.

Định nghĩa của Ferruzza: “Khai phá dữ liệu là tập hợp các phương pháp được dùng trong tiến trình khám phá tri thức để chỉ ra sự khác biệt các mối quan hệ và các mẫu chưa biết bên trong dữ liệu”

Định nghĩa của Parsaye: “Khai phá dữ liệu là quá trình trợ giúp quyết định, trong đó chúng ta tìm kiếm các mẫu thông tin chưa biết và bất ngờ trong CSDL lớn”

Định nghĩa của Fayyad: “Khai phá tri thức là một quá trình không tầm thường nhận ra những mẫu dữ liệu có giá trị, mới, hữu ích, tiềm năng và có thể hiểu được”.

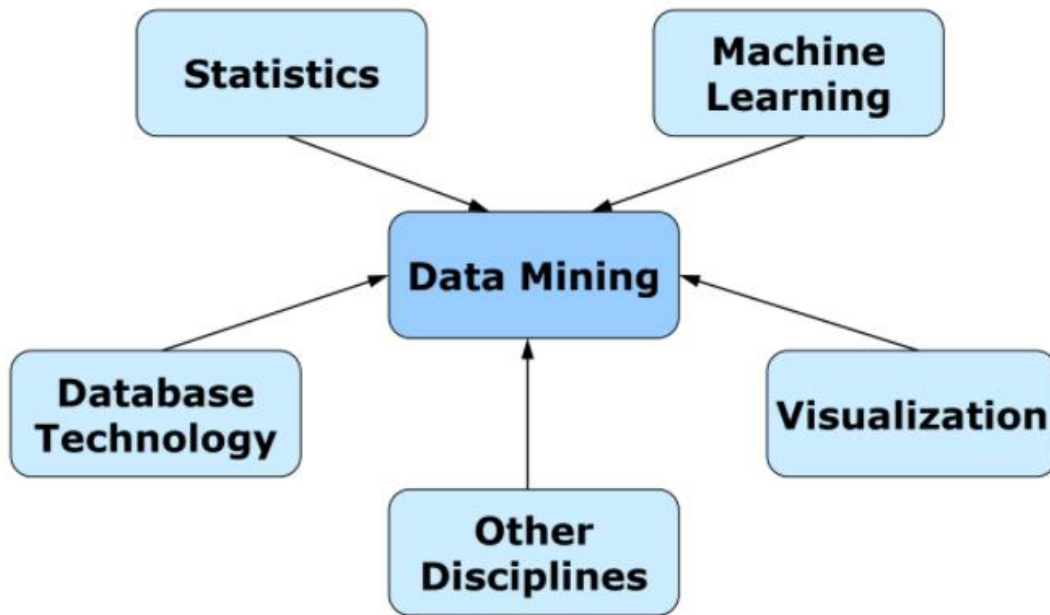
Lượng lớn dữ liệu sẵn có để khai phá

- Bất kỳ loại dữ liệu được lưu trữ hay tạm thời, có cấu trúc hay bán cấu trúc hay phi cấu trúc.
- Dữ liệu được lưu trữ:

- + Các tập tin truyền thống (flat files)
- + Các cơ sở dữ liệu quan hệ hay quan hệ đối tượng
- + Các cơ sở dữ liệu giao tác hay kho dữ liệu
- + Các cơ sở dữ liệu hướng ứng dụng
- + Các kho thông tin: the World Wide Web, ...
- Dữ liệu tạm thời: các dòng dữ liệu (data streams)

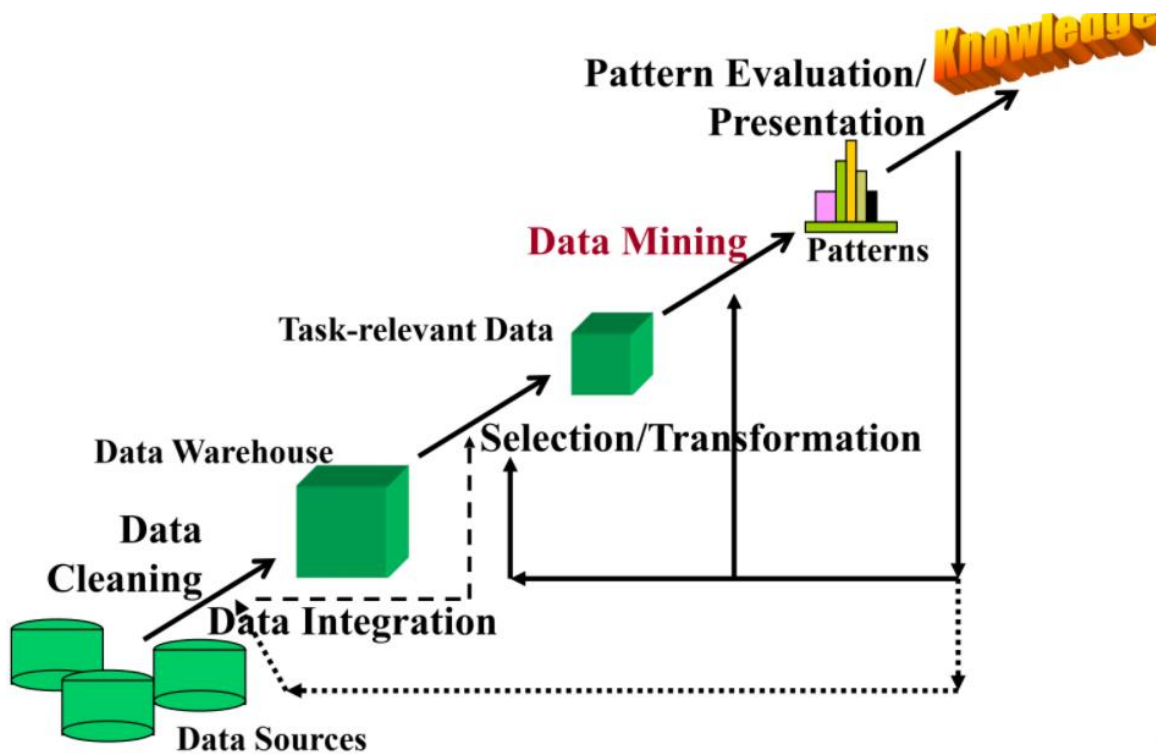
Tri thức đạt được từ quá trình khai phá:

- Mô tả lớp/ khái niệm
 - Mẫu thường xuyên, các môi quan hệ kết hợp, tương quan
 - Mô hình phân loại, dự đoán
 - Mô hình gom cụm/ phân cụm
 - Các phần tử biên
 - Xu hướng hay mức độ thường xuyên của các đối tượng có hành vi thay đổi theo thời gian
 - Tri thức đạt được có thể có tính mô tả hay dự đoán tùy thuộc vào quá trình khai phá cụ thể: mô tả (Descriptive) và dự đoán (Predictive).
 - Tri thức đạt được có thể có cấu trúc, bán cấu trúc hoặc phi cấu trúc.
 - Tri thức đạt được có thể được/ không được người dùng quan tâm
- ⇒ Các độ đo đánh giá tri thức đạt được.
- Tri thức đạt được có thể được dùng trong việc hỗ trợ ra quyết định, điều khiển quy trình, quản lý thông tin, xử lý truy vấn, ...



Hình 1. 1: Các lĩnh vực trong khai phá dữ liệu

1.2. Các bước của quá trình khai phá dữ liệu



Hình 1. 2: Các bước của khai phá dữ liệu

Quá trình khám phá tri thức là một chuỗi lặp gồm các bước:

- Data cleaning (làm sạch dữ liệu)
- Data integration (tích hợp dữ liệu)
- Data selection (chọn lựa dữ liệu)
- Data transformation (biến đổi dữ liệu)
- Data mining (khai phá dữ liệu)
- Pattern evaluation (đánh giá mẫu)
- Knowledge presentation (biểu diễn tri thức)

Được thực thi với:

- Data sources (các nguồn dữ liệu)
- Data warehouse (kho dữ liệu)
- Task-relevant data (dữ liệu cụ thể sẽ được khai phá)
- Patterns (mẫu kết quả từ khai phá dữ liệu)
- Knowledge (tri thức đạt được)

1.3. Khai phá dữ liệu và công nghệ cơ sở dữ liệu

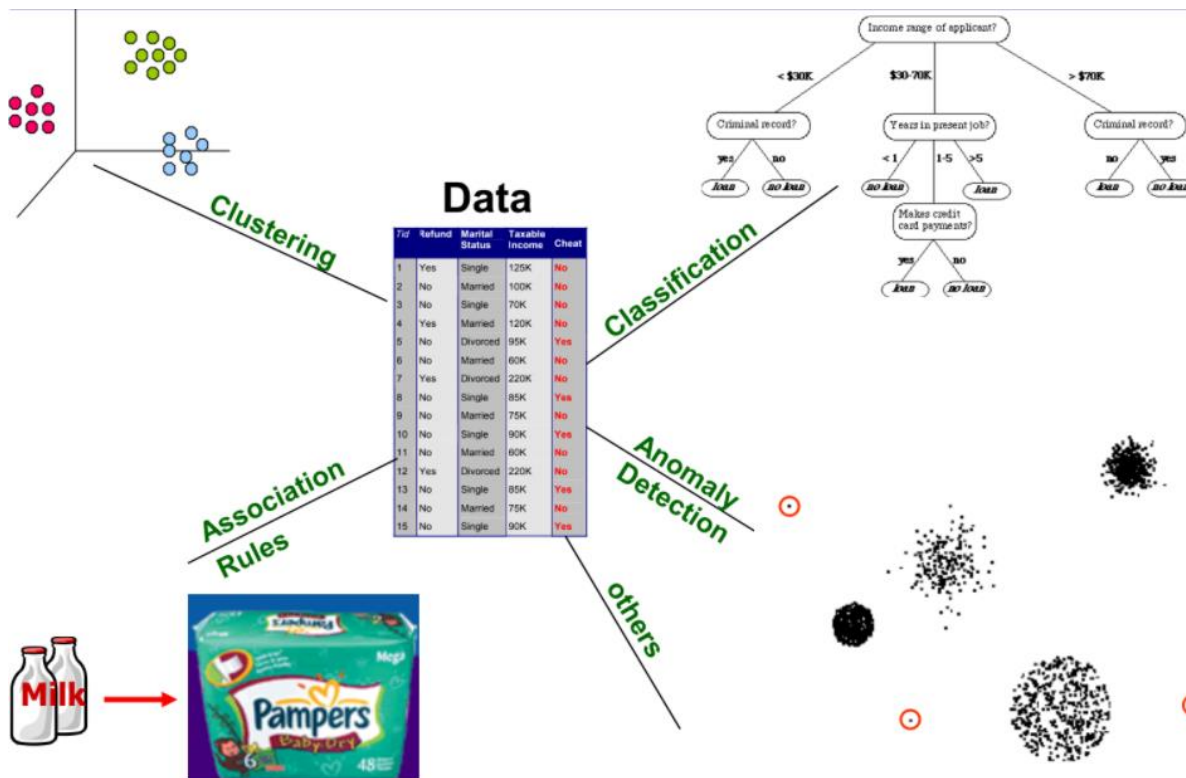
- Công nghệ cơ sở dữ liệu cho việc quản lý dữ liệu được khai phá.
- Dữ liệu rất lớn, có thể vượt qua khả năng bộ nhớ chính (main memory).
- Dữ liệu được thu thập theo thời gian
- Các hệ cơ sở dữ liệu có khả năng xử lý hiệu quả lượng lớn dữ liệu với các cơ chế phân trang (paging) và hoán chuyển (swapping) dữ liệu vào/ ra bộ nhớ chính.
- Các hệ cơ sở dữ liệu hiện tại có khả năng xử lý nhiều loại dữ liệu phức tạp.
- Các chức năng khác: xử lý đồng thời, bảo mật, hiệu năng, tối ưu hóa, ... của các hệ cơ sở dữ liệu đã được phát triển tốt.
- Các hệ cơ sở dữ liệu (DBMS) hỗ trợ khai phá dữ liệu:
 - + Oracle Data Mining
 - + Các công cụ khai phá của Microsoft: MS SQL Server 2019, ...
 - + IBM

1.4. Các tác vụ của khai phá dữ liệu

Các tác vụ khai phá dữ liệu có thể được phân thành hai loại: miêu tả và dự báo hay các đặc tính chung của dữ liệu trong CSDL hiện có. Các kỹ thuật này gồm có: phân cụm (clustering), tóm tắt (summerization), trực quan hoá (visualiztion), phân tích sự

phát triển và độ lệch (Evolution and deviation analyst), phân tích luật kết hợp (association rules) ...

- Khai phá mô tả lớp các đặc tính chung của dữ liệu trong cơ sở dữ liệu. Kỹ thuật khai phá dữ liệu mô tả: Có nhiệm vụ mô tả về các tính chất hay các đặc tính chung của dữ liệu trong CSDL hiện có. Các kỹ thuật này gồm có: phân cụm (clustering), tóm tắt (summerization), trực quan hoá (visualiztion), phân tích sự phát triển và độ lệch (Evolution and deviation analyst), phân tích luật kết hợp (associationrules) ...
- Khai phá luật kết hợp/ tương quan
 - + Giải thuật Apriori
- Phân loại dữ liệu
 - + Giải thuật phân loại với cây quyết định
 - + Giải thuật phân loại với mạng Bayes
- Dự đoán thực hiện việc suy luận trên dữ liệu hiện thời để đưa ra các dự báo. Kỹ thuật khai phá dữ liệu dự đoán: Có nhiệm vụ đưa ra các đoán dựa vào các suy diễn trên dữ liệu hiện thời. Các kỹ thuật này gồm có: Phân lớp (classification), hồi quy (regression)...
- Gom cụm dữ liệu
 - + Giải thuật gom cụm k-means
 - + Giải thuật gom cụm phân cấp nhóm
- Phân tích xu hướng
- Phân tích độ lệnh và phần tử biên
- Phân tích độ tương tự



Hình 1. 3: Các tác vụ trong khai phá dữ liệu

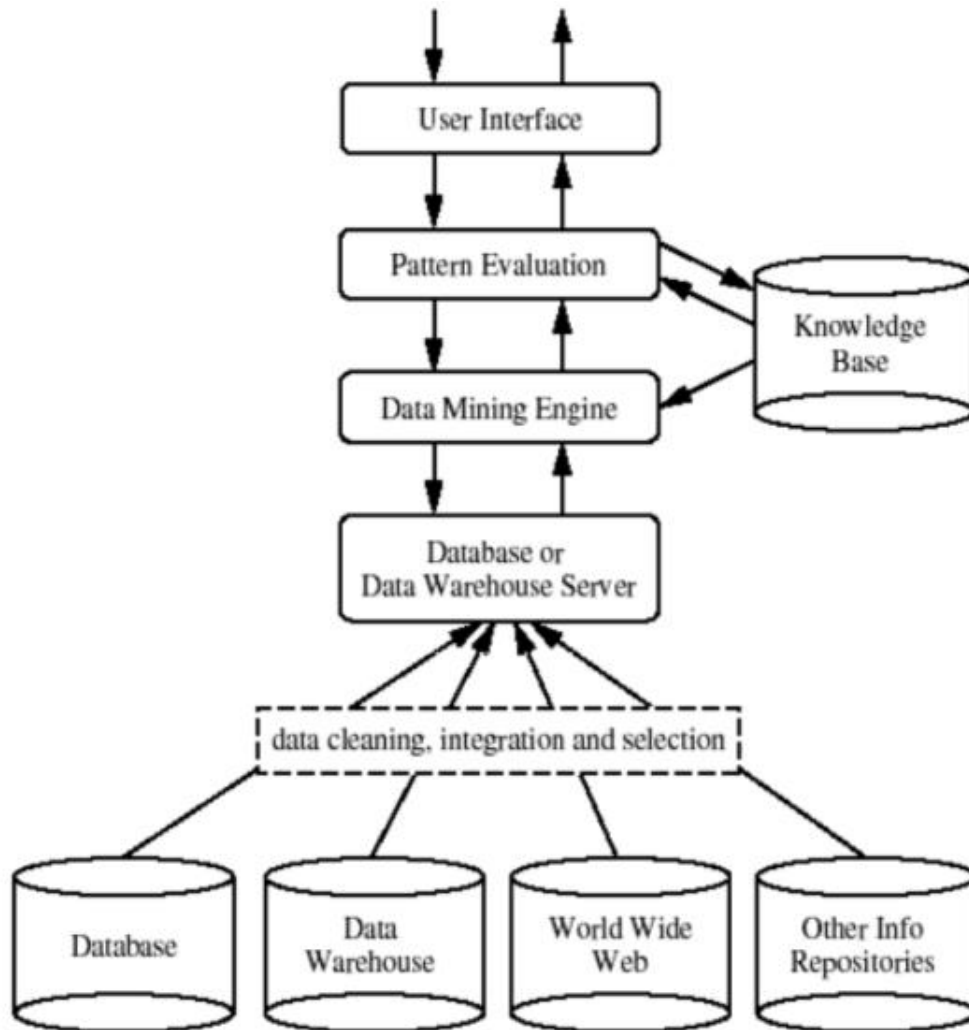
1.5. Kiến trúc của một hệ thống khai phá dữ liệu

- Các thành phần chính:

- + Database, data warehouse, World Wide Web và information repositories: Thành phần này là cá nguồn dữ liệu/ thông tin sẽ được khai phá. Trong những tình huống cụ thể, thành phần này là nguồn nhập (input) của các kỹ thuật tích hợp và làm sạch dữ liệu.
- + Database hay data warehouse server: Thành phần chịu trách nhiệm chuẩn bị dữ liệu thích hợp cho các yêu cầu khai phá dữ liệu.
- + Knowledge base: Thành phần chứa tri thức miền, được dùng để hướng dẫn quá trình tìm kiếm, đánh giá các mẫu kết quả được tìm thấy. Tri thức miền có thể là các phân cấp khái niệm, niềm tin của người sử dụng, các ràng buộc hay các ngưỡng giá trị, siêu dữ liệu, ...
- + Pattern evaluation module: Thành phần làm việc với các độ đo hỗ trợ tìm kiếm và đánh giá các mẫu sao cho mẫu được tìm thấy là những mẫu được quan tâm

bởi người sử dụng. Thành phần này có thể được tích hợp vào thành phần Data mining engine.

+ User interface: Thành phần hỗ trợ sự tương tác giữa người sử dụng và hệ thống khai phá dữ liệu.



Hình 1. 4: Kiến trúc của một hệ thống khai phá dữ liệu

1.6. Ý nghĩa và vai trò của khai phá dữ liệu

- Công nghệ hiện đại trong lĩnh vực quản lý thông tin:
 - + Hiện diện khắp nơi và có tính ảnh hưởng trong nhiều khía cạnh của đời sống hàng ngày.
 - + Được áp dụng trong nhiều ứng dụng thuộc nhiều lĩnh vực khác nhau.
 - + Hỗ trợ các nhà khoa học, giáo dục học, kinh tế học, doanh nghiệp, khách hàng, ...

1.7. Ứng dụng của khai phá dữ liệu

- Trong kinh doanh (business).
- Trong tài chính (finance) và tiếp thị bán hàng (sale marketing).
- Trong bảo hiểm (insurance)
- Trong khoa học (science) và y sinh học (biomedicine).
- Trong điều khiển (control) và viễn thông (telecommunication).

1.8. Kết luận chương 1

Thông qua những ý kiến trên, ta thấy rằng khai phá dữ liệu là một phần tất yếu trong cuộc sống. Nhờ có nó mà đất nước ngày càng phát triển, giảm thiểu nhân công mà thay vào đó là sử dụng hệ thống máy móc để làm việc. Nó giúp ích rất nhiều trong cuộc sống hiện đại ngày nay.

CHƯƠNG 2: TÌM HIỂU VỀ THUẬT TOÁN K-MEANS

2.1. Giới thiệu thuật toán phân cụm trong học máy

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 qui 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

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).

2.2. Thuật toán K-Means

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.

2.2.1. Các bước chính thực hiện thuật toán K-Means

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

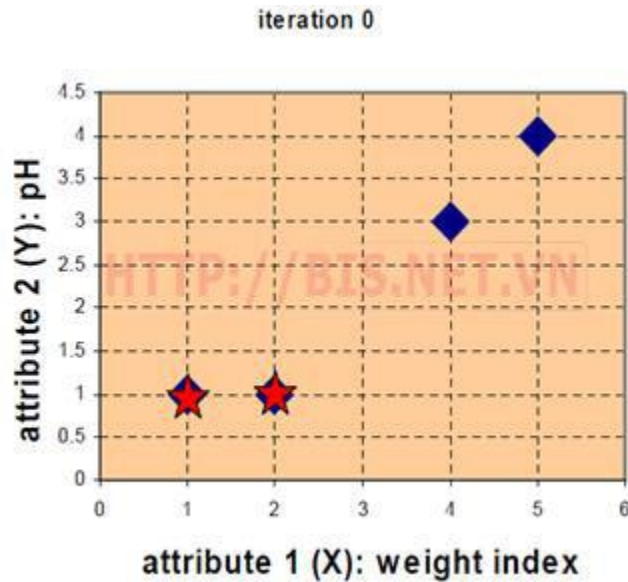
- B1.** Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.
- B2.** Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean)
- B3.** Nhóm các đối tượng vào nhóm gần nhất
- B4.** Xác định lại tâm mới cho các nhóm
- B5.** Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng

2.2.2. 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)

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{array}{l} c_1 = (1,1) \text{ group-1} \\ c_2 = (2,1) \text{ group-2} \end{array}$$

	A	B	C	D	
	1	2	4	5	X
	1	1	3	4	Y

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 (c1) và hàng thứ 2 trong ma trận khoảng cách biểu diễn khoảng cách của các đối tượng đến tâm của nhóm thứ 2 (c2).

Ví dụ, khoảng cách từ loại thuốc C=(4,3) đến tâm c1(1,1) là 3.61 và đến tâm c2(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

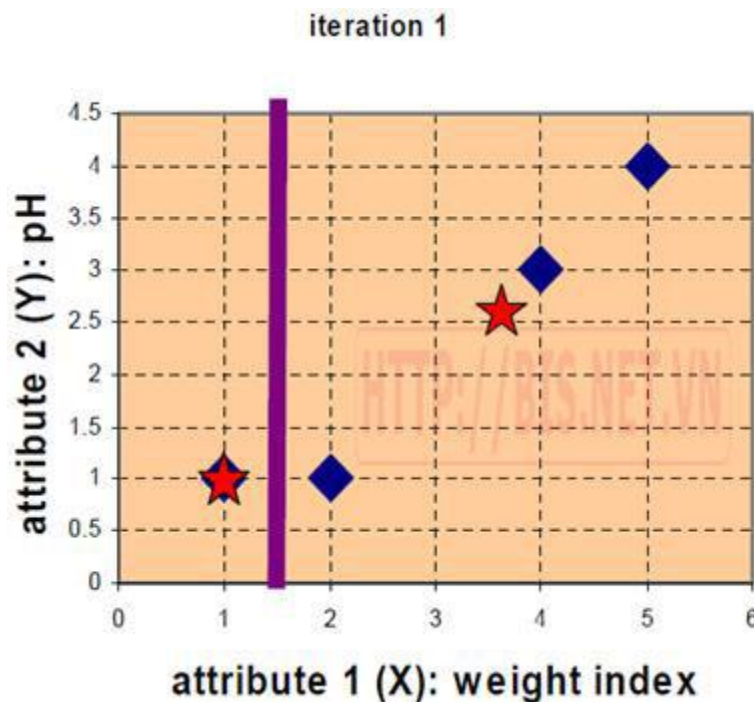
$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} \text{group-1} \\ \text{group-2} \end{matrix}$$

$A \quad B \quad C \quad D$

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

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

<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>

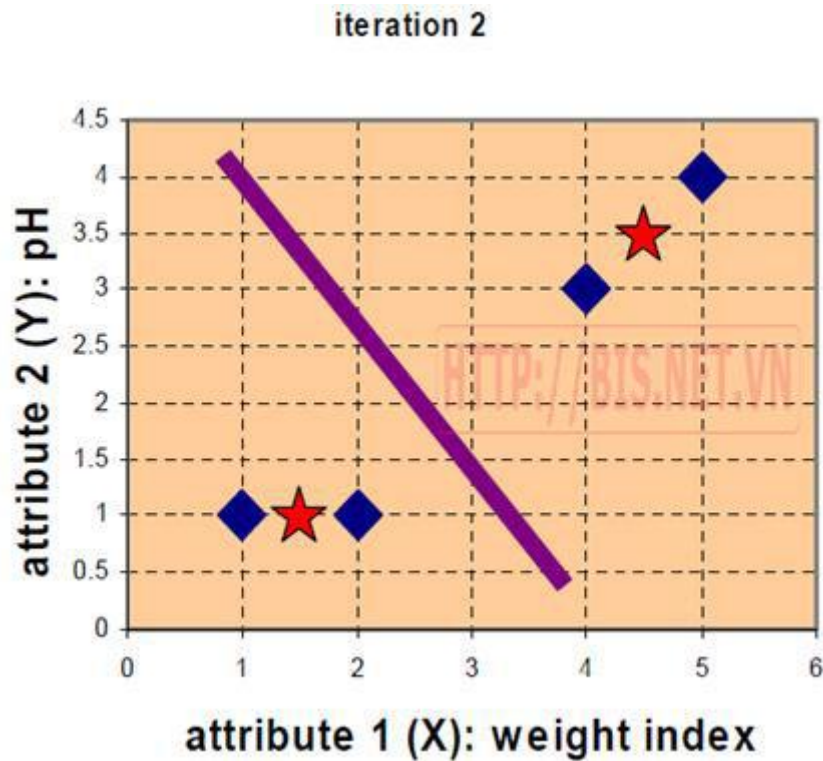
Bước 7. Nhóm các đối tượng vào nhóm

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

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

Bước 8. Tính lại tâm cho nhóm mới

$$\mathbf{c}_1 = \left(\frac{1+2}{2}, \frac{1+1}{2} \right) = (1\frac{1}{2}, 1) \quad \mathbf{c}_2 = \left(\frac{4+5}{2}, \frac{3+4}{2} \right) = (4\frac{1}{2}, 3\frac{1}{2})$$



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

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{array}{l} c_1 = (1\frac{1}{2}, 1) \text{ group-1} \\ c_2 = (4\frac{1}{2}, 3\frac{1}{2}) \text{ group-2} \end{array}$$

	A	B	C	D	
	1	2	4	5	X
	1	1	3	4	Y

Bước 9. Nhóm các đối tượng vào nhóm

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

	A	B	C	D
	1	2	4	5
	1	1	3	4

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

2.3. Kết luận chương 2

Thuật toán K-Means có ưu điểm là đơn giản, dễ hiểu và cài đặt. Tuy nhiên, một số hạn chế của K-Means là hiệu quả của thuật toán phụ thuộc vào việc chọn số nhóm K (phải xác định trước) và chi phí cho thực hiện vòng lặp tính toán khoảng cách lớn khi số cụm K và dữ liệu phân cụm lớn.

CHƯƠNG 3: ÁP DỤNG THUẬT TOÁN K-MEANS VÀO BÀI TOÁN PHÂN CỤM HOA IRIS

3.1. Mô tả bài toán

Trên thế giới có nhiều loại hoa Iris, cho đến nay, người ta có thể phân biệt các loài hoa này bằng mắt thường một cách dễ dàng, nhưng chuyện đó rất mất công sức và thời gian nếu lượng mẫu vật quá lớn. Nhờ nghiên cứu kỹ lưỡng, hiện nay người ta đã có thể tìm ra các đặc trưng cụ thể của từng loài hoa, mà theo đó có thể dễ dàng dùng để phân biệt các loại hoa này. Hơn thế nữa, các dữ liệu này hoàn toàn có thể được xử lý đưa vào trong máy tính để phân tích, nghĩa là giờ đây, ta có thể giúp máy tính hiểu được bộ dữ liệu và phân biệt được các loài hoa này khi có đủ dữ liệu của chúng được đưa vào. Nội dung báo cáo này chính là thuật toán có thể giúp máy tính có thể phân biệt được các loại hoa Iris (ở đây là 3 loại) khi đưa vào bộ dữ liệu phù hợp.

3.2. Mô tả tập dữ liệu

Bộ dữ liệu hoa Iris là một bộ dữ liệu đa biến được giới thiệu bởi nhà thống kê và nhà sinh vật học người Anh Ronald Fisher trong bài báo năm 1936 của ông. Việc sử dụng nhiều phép đo trong các vấn đề phân loại. Đôi khi nó được gọi là tập dữ liệu Iris của Anderson vì Edgar Anderson đã thu thập dữ liệu để định lượng sự biến đổi hình thái của hoa Iris của ba loài liên quan.

Bộ dữ liệu bao gồm 50 mẫu từ mỗi ba loài Iris (Iris Setosa, Iris virginica và Iris Versolor). Bốn đặc điểm được đo từ mỗi mẫu: chiều dài và chiều rộng của cánh hoa và cánh hoa, tính bằng centimet.

Bộ dữ liệu được lấy trên trang kaggle.

Link tải: <https://www.kaggle.com/arshid/iris-flower-dataset>

> **IRIS.csv** (4.51 KB)

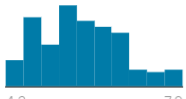



[Detail](#)

[Compact](#)

[Column](#)

About this file

The dataset is a CSV file which contains a set of 150 records under 5 attributes - Petal Length, Petal Width, Sepal Length, Sepal width and Class(Species)

# sepal_length	# sepal_width	# petal_length	# petal_width	▲ species
 4.3 7.9	 2 4.4	 1 6.9	 0.1 2.5	3 unique values
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3	1.4	0.1	Iris-setosa

Bộ dữ liệu sau khi tải về được lưu dưới dạng file csv:

A1									
	A	B	C	D	E	F	G	H	I
1	Id	SepalLengt	SepalWidth	PetalLengt	PetalWidth	Species			
2	1	5.1	3.5	1.4	0.2	Iris-setosa			
3	2	4.9	3	1.4	0.2	Iris-setosa			
4	3	4.7	3.2	1.3	0.2	Iris-setosa			
5	4	4.6	3.1	1.5	0.2	Iris-setosa			
6	5	5	3.6	1.4	0.2	Iris-setosa			
7	6	5.4	3.9	1.7	0.4	Iris-setosa			
8	7	4.6	3.4	1.4	0.3	Iris-setosa			
9	8	5	3.4	1.5	0.2	Iris-setosa			
10	9	4.4	2.9	1.4	0.2	Iris-setosa			
11	10	4.9	3.1	1.5	0.1	Iris-setosa			
12	11	5.4	3.7	1.5	0.2	Iris-setosa			
13	12	4.8	3.4	1.6	0.2	Iris-setosa			
14	13	4.8	3	1.4	0.1	Iris-setosa			
15	14	4.3	3	1.1	0.1	Iris-setosa			
16	15	5.8	4	1.2	0.2	Iris-setosa			
17	16	5.7	4.4	1.5	0.4	Iris-setosa			
18	17	5.4	3.9	1.3	0.4	Iris-setosa			
19	18	5.1	3.5	1.4	0.3	Iris-setosa			
20	19	5.7	3.8	1.7	0.3	Iris-setosa			
21	20	5.1	3.8	1.5	0.3	Iris-setosa			
22	21	5.4	3.4	1.7	0.2	Iris-setosa			
23	22	5.1	3.7	1.5	0.4	Iris-setosa			
24	23	4.6	3.6	1	0.2	Iris-setosa			
25	24	5.1	3.3	1.7	0.5	Iris-setosa			
26	25	4.8	3.4	1.9	0.2	Iris-setosa			
27	26	5	3	1.6	0.2	Iris-setosa			
28	27	5	3.4	1.6	0.4	Iris-setosa			
29	28	5.2	3.5	1.5	0.2	Iris-setosa			
30	29	5.2	3.4	1.4	0.2	Iris-setosa			
31	30	4.7	3.2	1.6	0.2	Iris-setosa			
32	31	4.8	3.1	1.6	0.2	Iris-setosa			

datasets_Hoa_Iris

Ready

3.3. Áp dụng thuật toán K-means vào bài toán

Chương trình được viết bằng ngôn ngữ python 3, trên nền tảng anaconda 3.7 và viết trên Spyder 4. Các thư viện cần thiết sử dụng trong chương trình bao gồm:

- Pandas: dùng để đọc file csv
- numpy: xử lý mảng và toán học
- scipy: dùng để gọi hàm cdist, một hàm tính khoảng cách

Các thư viện được thêm vào qua các lệnh:

```
import pandas as pd
```

```
import numpy as np
```

```
from scipy.spatial.distance import cdist
```

Dữ liệu đưa vào chương trình sẽ được đọc trực tiếp qua file csv ở trên. Vì cột đầu tiên của bộ dữ liệu(Id) là không cần thiết nên ta có thể bỏ đi. Việc đọc file và xử lý dữ liệu thừa được thực hiện qua câu lệnh:

```
datas = pd.read_csv('G:\Datas\HocTap\MachineLearning\BaoCao\datasets_Hoa_Iris.csv').drop('Id', axis = 1).values
```

Vì ở đây chỉ có 3 loài hoa, nên ta sẽ chia làm 3 cụm. Các hàm cần thiết sử dụng trong chương trình:

- Hàm Random_cluster(train, n_cluster): chọn ra ngẫu nhiên số cụm = n_cluster(trong bài này là 3), các cụm được chọn là các mẫu đã có trong tập dữ liệu.
- Hàm Choose_cluster(train, centers): tính khoảng cách từ các điểm dữ liệu đến các tâm cụm, chọn ra tâm cụm gần nó nhất để trả về(giá trị trả về là index của tâm cụm trong mảng các tâm cụm, ở đây có 3 tâm cụm nên trả về 1 trong 3 giá trị 0, 1, 2 đồng thời là tên cụm luôn). Đồng thời giá trị đó chính là nhãn cho điểm dữ liệu.
- Hàm Update_centers(train, labels, n_cluster): tìm lại tâm cụm mới.
- Hàm Check_update(centers, new_centers): kiểm tra xem các điểm dữ liệu sau khi cập nhật tâm cụm có thay đổi với lần trước không, nếu không có nghĩa thuật toán đã tìm được tâm cụm cần tìm

- Hàm Kmeans (init_centres, init_labels, train, n_cluster): Hàm chính của chương trình, chạy các hàm khác để tìm ra lời giải, trả về tọa độ các tâm cụm, số lần lặp để tìm ra tâm cụm, và tâm cụm tương ứng của từng điểm dữ liệu

Code thuật toán:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
# thư viện dùng để tính khoảng cách giữa các cặp điểm trong 2 tập hợp một
cách hiệu quả
from scipy.spatial.distance import cdist

datas = pd.read_csv("Iris.csv", delimiter=";").drop('Id', axis=1).values
train_full = datas
train = datas[:, :4]
# do tập dữ liệu có nên được chia làm 3 cụm.
n_cluster = 3

# Bước 1: Chọn ngẫu nhiên 3 tâm cho 3 cụm. Mỗi cụm sẽ đại diện bằng một tâm
cụm
# Kết quả trả về là của 3 tâm cụm là
# [[57 38 17 3]
#  [54 39 13 4]
#  [51 37 15 4]]
def Random_cluster(train, n_cluster):
    return train[np.random.choice(train.shape[0], n_cluster, replace=False)]

# Bước 2: Tính khoảng cách giữa các đối tượng đến 3 tâm cụm
#   Loại hoa 1   Loại hoa 2   Loại hoa 3
#   51         70         63
#   35         32         33
#   14         47         60
#   2         14         25
# Khoảng cách từ loại Loại hoa 1(51, 35, 14, 2) đến tâm cụm (57,38,17,3) là:
7.41
# Khoảng cách từ loại Loại hoa 1(51, 35, 14, 2) đến tâm (54,39,13,4) là: 5.48
def Choose_cluster(train, centers):
    # Tính toán khoảng cách theo cặp và trung tâm
    D = cdist(train, centers)
    # Chỉ mục trả về của trung tâm gần nhất
    return np.argmin(D, axis= 1)

#Bước 3: Tính lại các tọa độ tâm cho các nhóm mới dựa vào tọa độ các đối
tượng trong nhóm bằng đến khi nào k có sự thay đổi nữa thì chuyển sang bước
tiếp theo
def Update_centers(train, labels, n_cluster):
    centers = np.zeros((n_cluster, train.shape[1]))
    for k in range(n_cluster):
        # thu thập các điểm được giao cho cụm thứ k
        Xk = train[labels == k, :]
        # Lấy trung bình
        centers[k,:] = np.mean(Xk, axis= 0)
```

```

    return centers

# điều kiện dừng của thuật toán: khi không có sự thay đổi của các đối tượng
def Check_update(centers, new_centers):
    return (set([tuple(a) for a in centers]) == set([tuple(a) for a in
new_centers]))

def kmeans_visualize(train, centers, labels, n_cluster, title):
    plt.xlabel('x')
    plt.ylabel('y')
    plt.title(title)
    plt_colors = ['b', 'g', 'r', 'c', 'm', 'y', 'k', 'w']
    for i in range(n_cluster):
        data = train[labels == i]
        plt.plot(data[:, 0], data[:, 1], plt_colors[i] + '^', markersize=4,
                label='cluster_' + str(i)) # Vẽ cụm i lên đồ thị
        plt.plot(centers[i][0], centers[i][1], plt_colors[i + 4] + 'o',
markersize=10,
                label='center_' + str(i)) # Vẽ tâm cụm i lên đồ thị
    plt.legend() # Hiện bảng chú thích
    plt.show()

# Sử dụng thuật toán Kmeans
def Kmeans(init_centres, init_labels, train, n_cluster):
    centers = init_centres
    labels = init_labels
    times = 0
    while True:
        labels = Choose_cluster(train, centers)
        # kmeans_visualize(train, centers, labels, n_cluster, 'Assigned label for
data at time = ' + str(times + 1))
        new_centers = Update_centers(train, labels, n_cluster)
        if Check_update(centers, new_centers):
            break
        centers = new_centers
        kmeans_visualize(train, centers, labels, n_cluster, 'Update center
position at time = ' + str(times + 1))
        times += 1
    return (centers, labels, times)

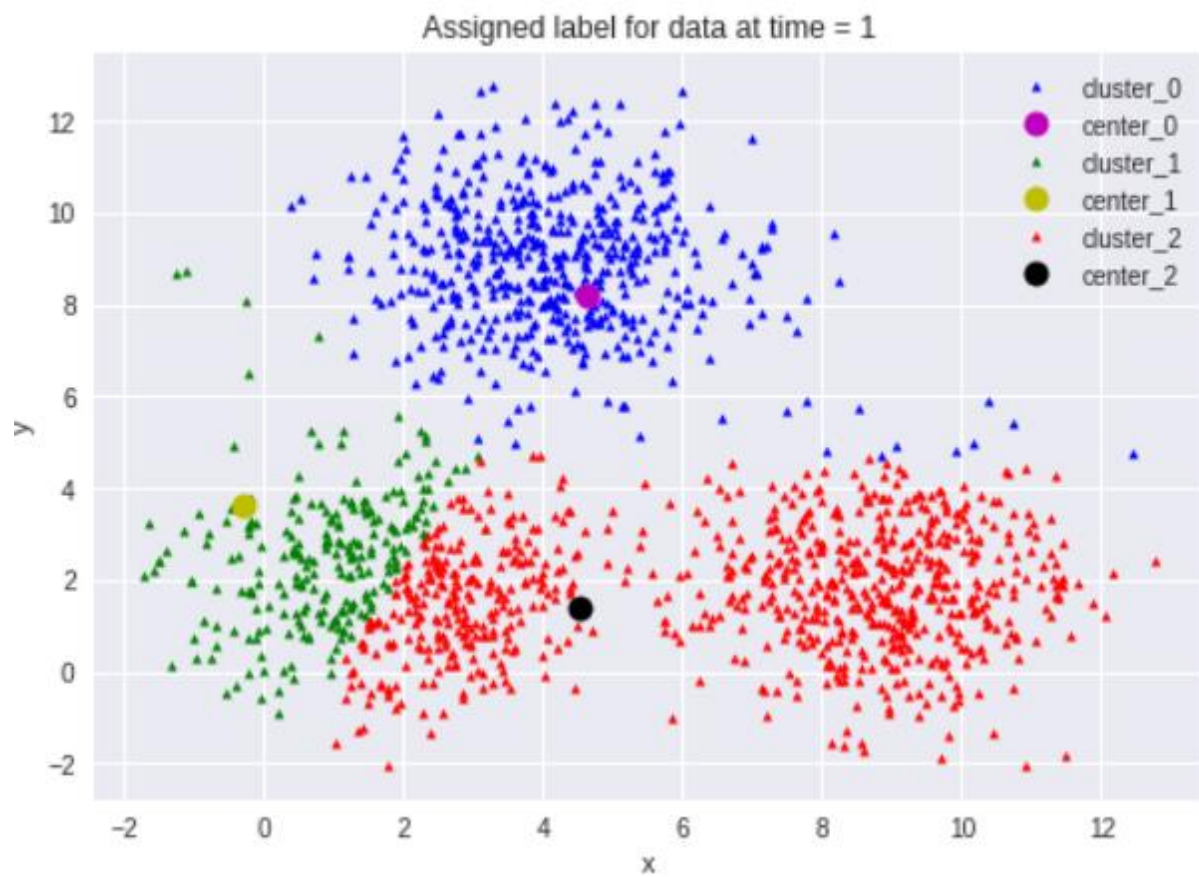
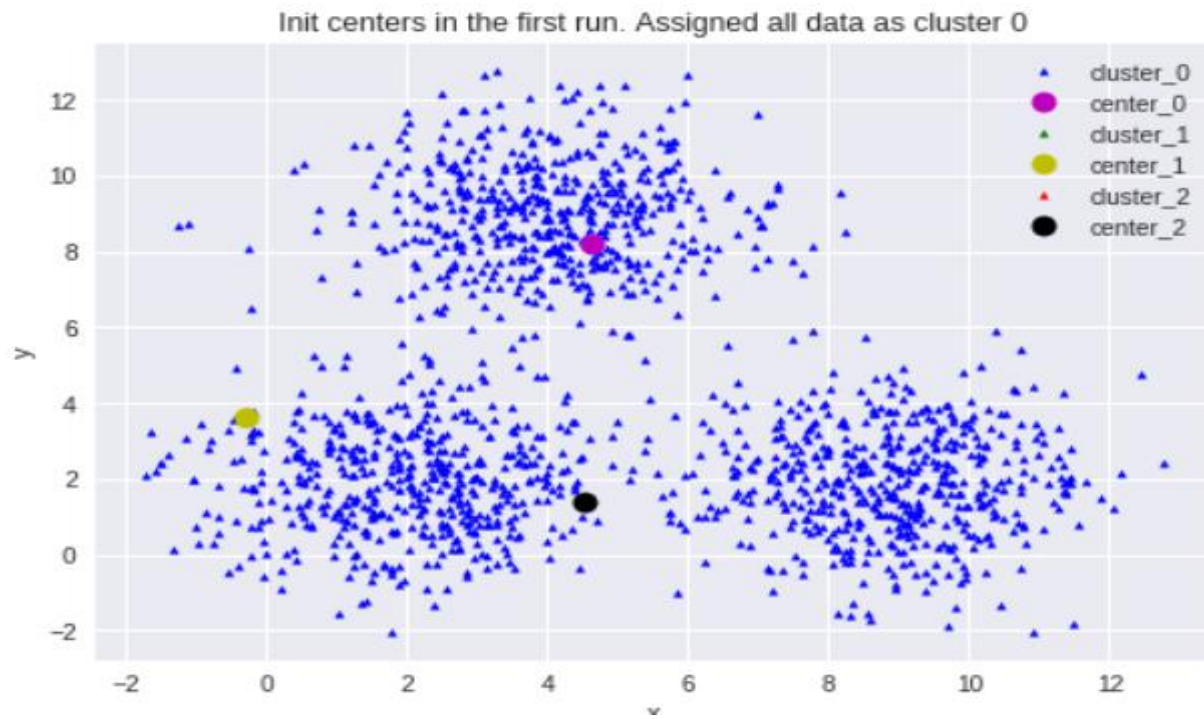
if __name__ == "__main__":
    init_centers = Random_cluster(train, n_cluster)
    init_labels = np.zeros(train.shape[0])
    # kmeans_visualize(train, init_centers, init_labels, n_cluster,
# 'Init centers in the first run. Assigned all data as
cluster 0')
    centers, labels, times = Kmeans(init_centers, init_labels, train,
n_cluster)

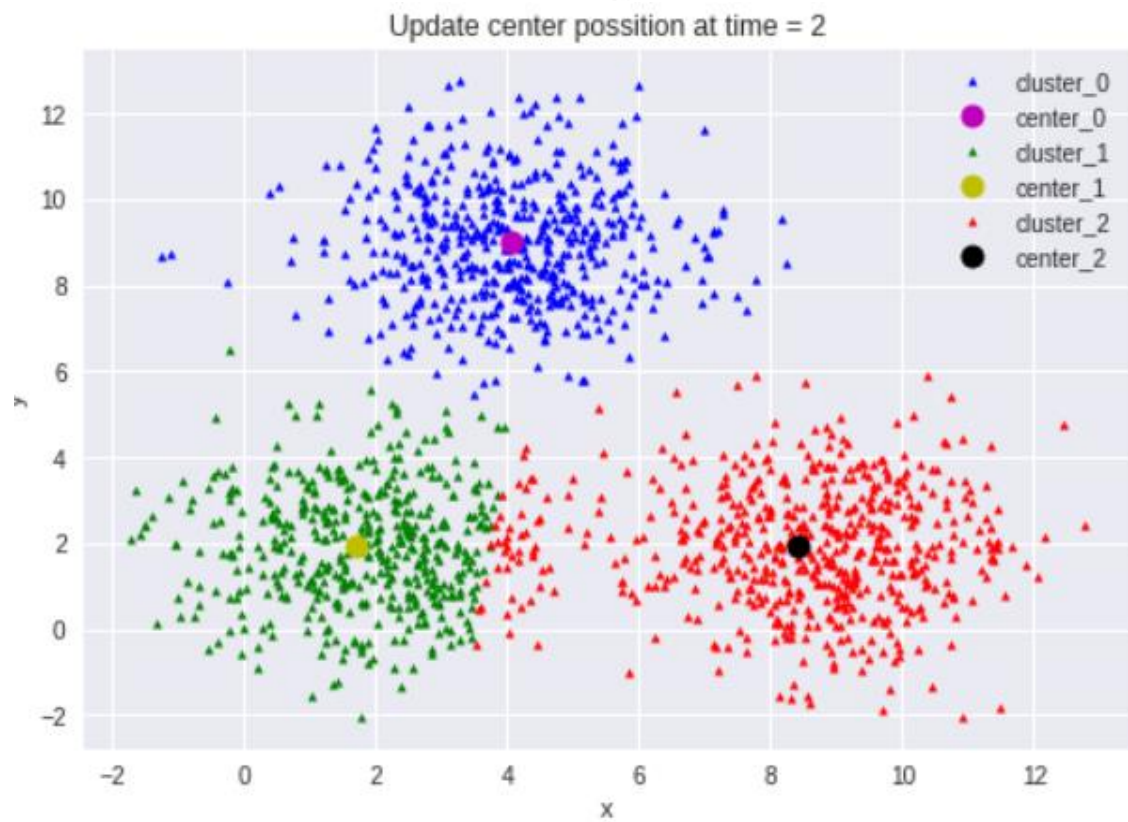
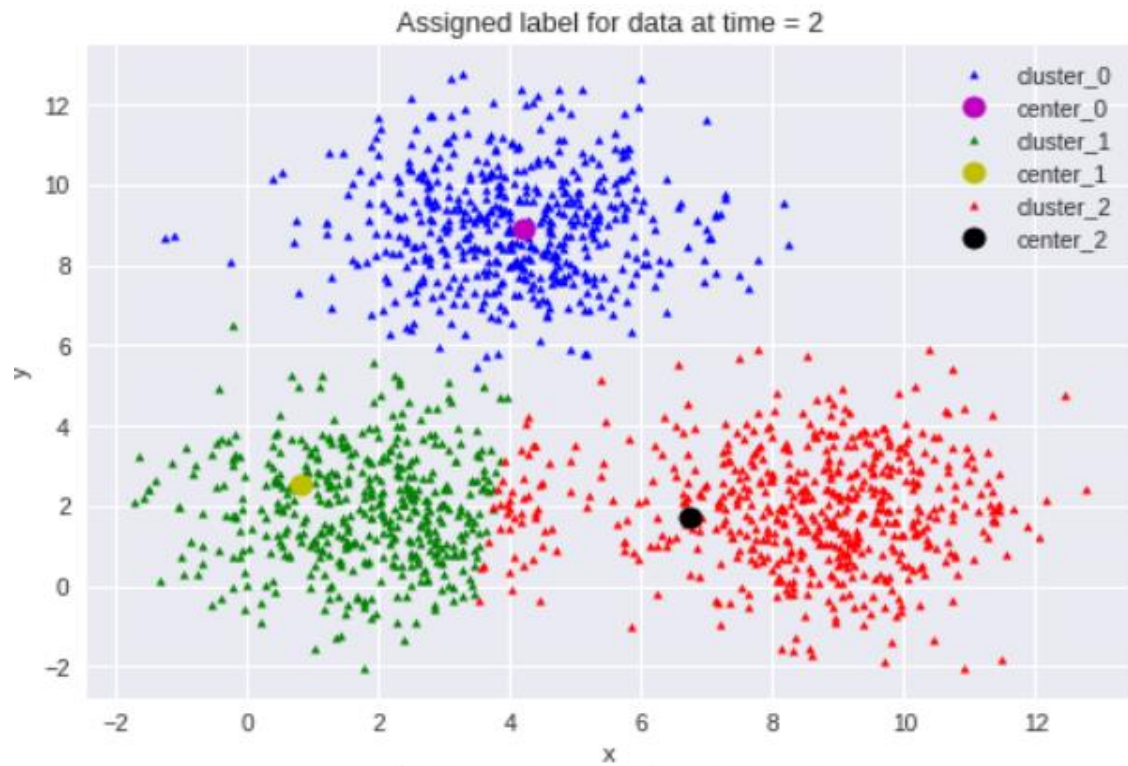
    print('Done! Kmeans has converged after', times, 'times')

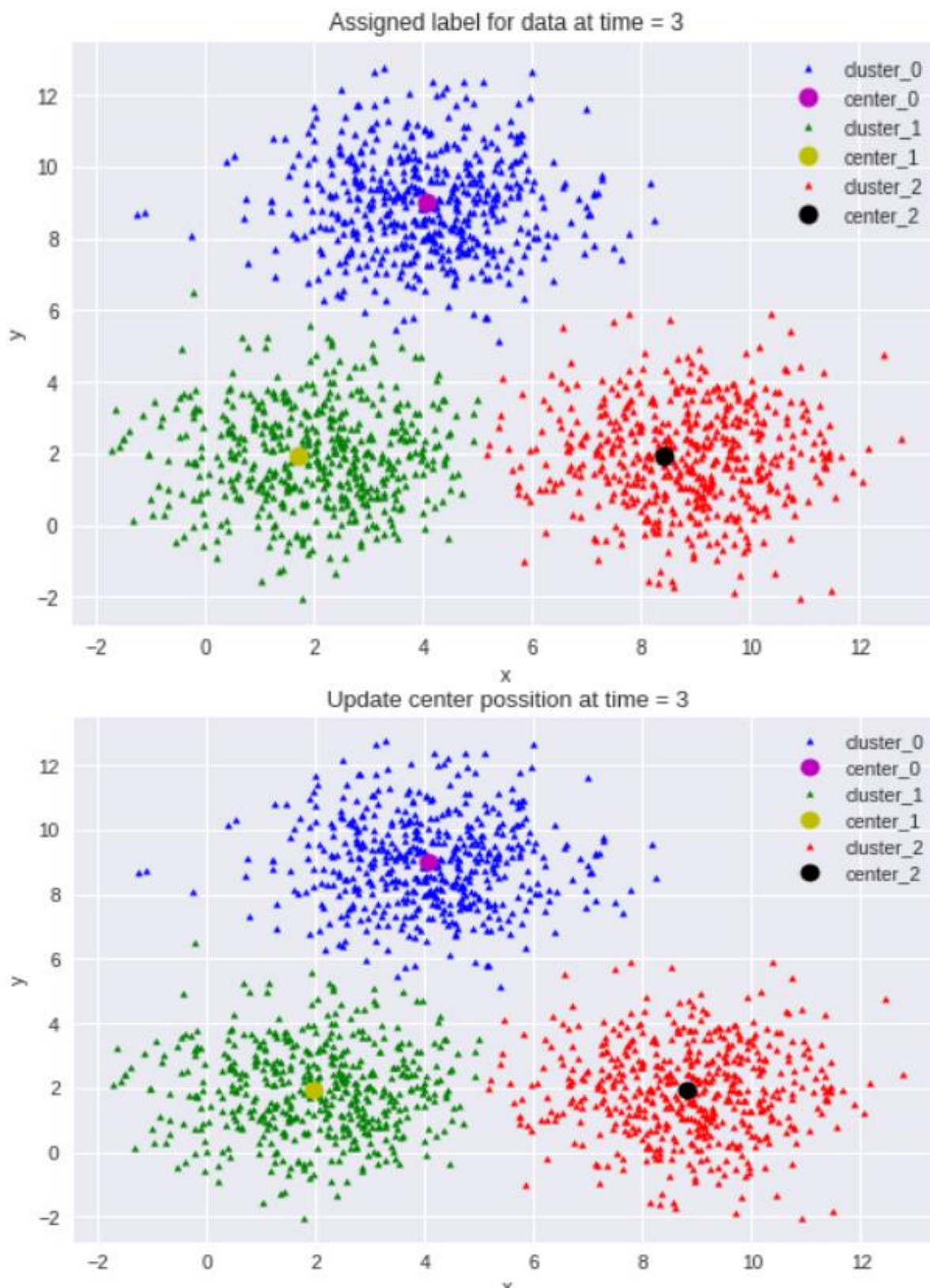
```

2.4. Đánh giá kết quả

Dưới đây là kết quả chạy thuật toán bằng Spyder:







Hình 3.1 : Kết quả hình ảnh sau khi chạy chương trình

Nhận xét:

Thuật toán chạy khá ổn với bộ dữ liệu, đặc biệt với loại hoa setosa với tỷ lệ chính xác 100%, với loại hoa versicolor thì tỉ lệ này thấp hơn một chút: 96%. Đặc biệt với loại hoa virginica, thuật toán tỏ ra không được hiệu quả lắm khi nhầm lẫn khá nhiều với loại hoa phía trên, độ chính xác đạt 72%.

Nguyên nhân có thể là do từ môi trường thực tế, 2 loài hoa này có thể có các dữ liệu tương đương nhau nhiều hơn, cũng có thể do thuật toán chưa tương thích được với bộ dữ liệu này, cần phải xử lý dữ liệu nhiều hơn hoặc điều chỉnh thuật toán cho phù hợp.

Bộ dữ liệu cũng có giới hạn số lượng chỉ 150 mẫu thử, khá ít cho 1 chương trình học máy, nên khó tránh khỏi việc sai sót, việc có thể đạt được kết quả như trên cũng có thể coi là không tồi.

KẾT LUẬN

Bài báo cáo được xây dựng với mục đích giúp mọi người có thể tìm hiểu cũng như có thêm hiểu biết về thuật toán Kmeans nói chung, và cụ thể có thể có 1 ứng dụng thực tế về phân loại loài hoa Iris trong tự nhiên với bộ dữ liệu chuẩn. Bài báo cáo vẫn còn nhiều thiếu sót và còn có thể phát triển thêm để kết quả có thể được cải thiện hơn như:

- Tiền xử lý dữ liệu tốt hơn
- Lấy thêm nhiều dữ liệu hơn cho hệ thống học
- Cải thiện thuật toán, thử với các hàm tính khoảng cách khác, ...

Trên đây là bài báo cáo của em, do thời gian và lượng kiến thức có hạn, việc sai sót là khó tránh khỏi, mong thầy cô có thể bỏ qua cho.

Một lần nữa em xin cảm ơn thầy Ngô Trường Giang đã nhiệt tình chỉ dạy em trong quá trình học tập môn nhập môn học máy để em có thể hoàn thành báo cáo này.

TÀI LIỆU THAM KHẢO

[1]: Tài liệu của Cô Nguyễn Thị Thanh Tân

[2]: <https://machinelearningcoban.com>

[3]: <http://bis.net.vn/forums/t/374.aspx>

[4]: <https://codetudau.com/de-dang-hieu-phuong-phap-k-means-qua-hinh-ve/index.html>

[5]: <https://nguyenvanhieu.vn/thuat-toan-phan-cum-k-means/?fbclid=IwAR1FDytz3YdcP5VEife5NF3JmbAFFPM6bgBJk1ZoDegE0qFmOMfhBXXAa7c>