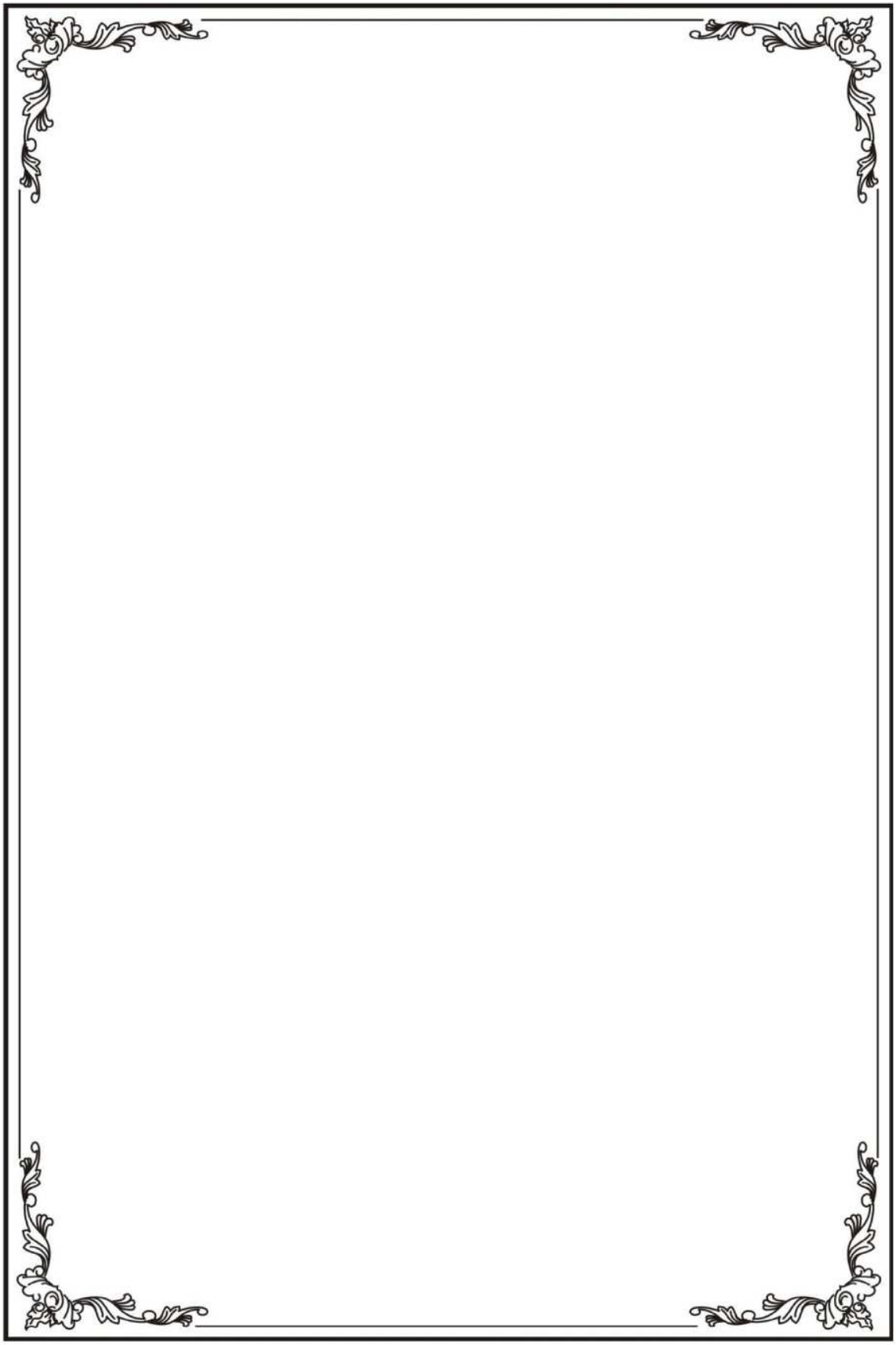


TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC
KHAI PHÁ DỮ LIỆU

ĐỀ TÀI:

KHAI PHÁ DỮ LIỆU TỈ LỆ MẮC VÀ TỬ VONG DO VIRUS COVID-19 BẰNG THUẬT TOÁN PHÂN CỤM K-MEANS

Sinh viên thực hiện: ĐOÀN THỊ HÒA

VŨ THỊ MINH THƯƠNG

TRỊNH THỊ HỒNG

Giảng viên hướng dẫn : VŨ VĂN ĐỊNH

Khoa : CÔNG NGHỆ THÔNG TIN

Chuyên ngành : HT THƯƠNG MẠI ĐIỆN TỬ

Lớp : D13HTTMĐT1

Khóa : 2018-2023

Hà Nội, tháng 02 năm 2021

PHIẾU CHẤM ĐIỂM

Sinh viên thực hiện :

Họ và tên	Chữ ký	Điểm
Đoàn Thị Hòa		
Vũ Thị Minh Thương		
Trịnh Thị Hồng		

Giảng viên chấm :

Họ và tên	Chữ ký	Ghi chú
Giảng viên 1:		

Giảng viên 2:		

MỤC LỤC

LỜI CẢM ƠN 1

TÓM TẮT 2

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 3

1.1 Đặt vấn đề 3

1.2 Cơ sở hình thành đề tài 4

1.3 Mục tiêu đề tài 5

1.4 Đối tượng và phương pháp nghiên cứu 5

1.5 Ý nghĩa đề tài 5

1.5.1 Ý nghĩa khoa học 5

1.5.2 Ý nghĩa thực tiễn 6

1.6 Bố cục đề tài 6

CHƯƠNG 2: KHAI PHÁ DỮ LIỆU 7

2.1 Tổng quan về kỹ thuật Khai phá dữ liệu(Data mining) 7

2.1.1 Khái niệm về khai phá dữ liệu 7

2.1.2 Quy trình khai phá dữ liệu 8

2.1.3 Ứng dụng của khai phá dữ liệu 11

2.2 Tổng quan về hệ hỗ trợ ra quyết định 11

2.3 Phân cụm dữ liệu và ứng dụng 12

2.3.1 Mục đích của phân cụm dữ liệu 12

2.3.2 Các bước cơ bản để phân cụm 13

2.3.3 Các loại đặc trưng 15

2.3.4 Các ứng dụng của phân cụm 16

2.3.5 Phân loại các thuật toán phân cụm 18
2.4 Cơ sở dữ liệu Y khoa 20
2.4.1 Sơ lược về Đại dịch covid-19 20
2.4.2 Sự lây truyền 21
2.4.3 Dấu hiệu và triệu chứng 22
CHƯƠNG 3: KỸ THUẬT PHÂN CỤM VÀ THUẬT TOÁN K-MENAS 23
3.1 Giới thiệu về kỹ thuật phân cụm trong Khai phá dữ liệu 23
3.2 Thuật Toán K-Means 24
3.3 Áp dụng và sử dụng thuật toán K-means vào bộ dataset Covid-19 29
CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ 31
4.1 Xây dựng mô hình bằng Weka 31
KẾT LUẬN 41
TÀI LIỆU THAM KHẢO 42

DANH MỤC HÌNH ẢNH

Hình 2. 1 Knowledge Discovery in Databases 10
Hình 2. 2 Sơ đồ hệ hỗ trợ quyết định 12
Hình 2. 3 Các bước trong quá trình phân cụm 15

Hình 3. 1 Các kỹ thuật phân cụm……..……..23
Hình 3. 2 Mô tả thuật toán K-Means 24
Hình 3. 3 Tập dữ liệu Covid-19 sau khi phân cụm 30

Hình 4. 1 Nhập dữ liệu vào Weka.....31
Hình 4. 2 Dữ liệu đưa vào được phân đoạn – tiền xử lý 32
Hình 4. 3 Các thuộc tính bộ dữ liệu tỷ lệ nguwoif chết và nhiễm virus trên 1 triệu người 33
Hình 4. 4 Đầu ra phân lớp 34
Hình 4. 5 Đầu ra phân cum bằng K-means vói tất cả thuộc tính 35
Hình 4. 6 Biểu đồ tỷ lê các cụm theo toàn bộ thuộc tính trên toàn bộ dữ liệu 36
Hình 4. 7 Đầu ra phân cum bằng K-means vói thuộc tính quốc gia và tỷ lệ người chết 37
Hình 4. 8 Biểu đồ tỷ lê các cụm theo thuộc tính quốc gia và người chết trên toàn bộ dữ liệu 38
Hình 4. 9 Đầu ra phân cum bằng K-means vói thuộc tính quốc gia và tỷ lệ người mắc bệnh 39
Hình 4. 10 Biểu đồ tỷ lê các cụm theo thuộc tính quốc gia và người chết trên toàn bộ dữ liệu. 40

DANH MỤC BẢNG BIỂU

Bảng 2. 1 Triệu chứng và tỉ lệ mắc bệnh 22

Bảng 4. 1 Bảng phân tích dữ liệu đầu ra với tất cả các thuộc tính.....35
Bảng 4. 2 Bảng phân tích dữ liệu đầu ra với thuộc tính Quốc gia và tỷ lệ người chết 37
Bảng 4. 3 Bảng phân tích dữ liệu đầu ra với thuộc tính Quốc gia và tỷ lệ người chết 39

LỜI CẢM ƠN

Qua bài tập lớn này, chúng em xin gửi lời cảm ơn tới thầy cô khoa công nghệ thông tin, đặc biệt là thầy Vũ Văn Định đã cho chúng em có cơ hội được tìm hiểu một góc kiến thức mới, hay và bổ ích cùng với đó là sự tận tâm dạy dỗ chúng em, giúp chúng em có thể hoàn thiện đề tài này. Trong quá trình tìm hiểu và hoàn thiện, đề tài sẽ không thể tránh khỏi những sai sót, khuyết điểm. Vì vậy, nhóm thực hiện chúng em hy vọng nhận được sự đánh giá và đóng góp nhiệt tình từ phía thầy và các bạn để bài của nhóm chúng em được hoàn thiện hơn.

Qua bài tập lớn này, chúng em xin cảm ơn các bạn bè lớp D13HTTMDT1 đã giúp đỡ chúng em trong quá trình học tập và làm bài tập lớn, đã chia sẻ kinh nghiệm kiến thức của các bạn đã tạo nên nền tảng kiến thức cho chúng em.

Cuối cùng, chúng em xin gửi lời cảm ơn gia đình đặc biệt là cha mẹ đã tạo điều kiện tốt nhất cho con có đủ khả năng thực hiện bài tập lớn này, trang trải học phí, đông viên tinh thần cho em để học tập trong môi trường đại học tuyệt vời này.

Chúng em xin chân thành cảm ơn!

Nhóm sinh viên thực hiện

ĐOÀN THỊ HÒA

VŨ THỊ MINH THƯƠNG

TRỊNH THỊ HỒNG

TÓM TẮT

Ngành y tế và giáo dục luôn là vấn đề sống còn của bất kỳ quốc gia nào trên thế giới. Trong những năm gần đây, chính phủ Việt nam đặc biệt đầu tư cho hai ngành mũi nhọn này thông qua các chính sách , nguồn vốn dành cho trang thiết bị hạ tầng và nghiên cứu khoa học. Trong lĩnh vực kho học, càng ngày càng có nhiều công trình khoa học trong y tế. Tuy nhiên các nghiên cứu khoa học về ứng dụng công nghệ thông tin để giải quyết bài toán về y tế là không nhiều. Do sự nguy hiểm và tình hình lây lan diễn biến phức tạm của đại dịch Covid-19 xảy ra trên toàn thế giới, vậy nên chúng e làm đề tài sử dụng môn học khai phá dữ liệu để xác định đánh gía tỷ lệ mắc bệnh và tử vong của người dân trên 200 quốc gia và vũng lãnh thổ để cho thấy sự nguy hiểm và nhóm các nước bị ảnh hưởng nhiều nhất.

Nghiên cứu tiến hành theo 4 bước chính:

(1) Tìm hiểu nghiệp vụ y tế liên quan đến virus corona.

(2) Thu nhập và tiền xử lý dữ liệu.

(3) Tìm hiểu bài toán phân cụm trong khai phá dữ liệu, lựa chọn thuật toán phù hợp với yêu cầu bài toán đặt ra và dữ liệu thu nhập được.

(4) Hiện thực chương trình máy tính và đánh giá ý nghĩa thực tiễn.

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1.1 Đặt vấn đề

Ứng dụng công nghệ thông tin vào việc lưu trữ và xử lý thông tin ngày nay được áp dụng hầu hết trong lĩnh vực, điều này đã tạo ra một lượng lớn dữ liệu được lưu trữ với kích thước tăng lên không ngừng. Đây chính là điều kiện tốt cho việc khai thác kho dữ liệu để đem lại tri thức có ích với các công cụ truy vấn, lập việc khai thác kho dữ liệu để đem lại tri thức có ích với các công cụ truy vấn, lập bẳng biểu và khai phá dữ liệu.

Khai phá dữ liệu là một kỹ thuật dựa trên nền tảng của nhiều lý thuyết như xác xuất, thống kê, máy học nhằm tìm kiếm các tri thức tiềm ẩn trong các kho dữ liệu có kích thước lớn mà người dùng khó có thể nhận biết bằng những kỹ thuật thông thường. Nguồn dữ liệu y khoa rất lớn, nếu áp dụng khai phá dữ liệu trong lĩnh vực này sẽ mang lại nhiều ý nghĩa cho ngành y tế. Nó sẽ cung cấp nững thông tin quý giá nhằm hỗ trợ trong việc chuẩn đoán và điều trị sớm giúp bệnh nhân thoát đưuọc nhiều căn bệnh hiểm nghèo.

Trong lĩnh vực y khoa Việt Nam, hiện nay các tuyến y tế phường, xã, vùng sâu, vùng xa còn thiếu nhân lực y tế có trình độ chuyên môn và thiếu các trang thiết bị cần thiết trong chuẩn đoán bệnh. Vì vậy xây dựng hệ thống chuẩn đoán rất cần thiết cho ngành y tế hiện nay ở Việt Nam. Hệ hỗ trợ sẽ kết hợp với cán bộ y tế giúp chuẩn đoán sớm một số bệnh phát hiện sớm được những bệnh nguy hiểm và giảm gánh nặng kinh tế cho gia đình bệnh nhân và xã hội. Để minh chứng cho những lợi ích mà việc chuẩn đoán mang lại, đề tài chọn bộ dữ liệu về virus corona để thử nghiệm và đánh giá.

Ứng dụng kỹ thuật phân cụm dữ liệu trong khai phá dữ liệu nhằm xây dựng hệ thống đánh giá là một trong những hướng nghiên cứu chính của đề tài. Sau khi phân tích một số thuật toán cũng như đặc điểm của dữ liệu thu nhập được về visrus

covid-19 , đề tài đề xuất ứng dụng mô hình phân cụm và thuật toán K-Means để tìm ra qui luật tìm ẩn trong dữ liệu.

1.2 Cơ sở hình thành đề tài

Theo thống kê năm 2019 từ tổ chức Y tế Thế giới(WHO),ra tuyên bố gọi "COVID-19" là "Đại dịch toàn cầu. Khởi nguồn vào tháng 12 năm 2019 với tâm dịch đầu tiên được ghi nhận tại thành phố Vũ Hán thuộc miền Trung Trung Quốc, bắt nguồn từ một nhóm người mắc viêm phổi không rõ nguyên nhân Ca COVID-19 tử vong đầu tiên được ghi nhận vào ngày 9 tháng 1 năm 2020 tại Vũ Hán. Theo dõi 17 bệnh nhân tử vong đầu tiên ở Trung Quốc thống kê đến ngày 22 tháng 1 năm 2020, thời gian bắt đầu mắc COVID-19 đến khi tử vong nằm trong khoảng 6 đến 41 ngày, với số trung vị là 14 ngày. Theo đài Trung ương Trung Quốc NHC, tính đến ngày 2 tháng 2 năm 2020, phần lớn ca tử vong (trên tổng số 490 ca) có độ tuổi cao – khoảng 80% ca là người có độ tuổi lớn hơn 60, và 75% trong số họ có bệnh lý nền như bệnh tim mạch và đái tháo đường.

Ca tử vong so với SARS-CoV-2 ngoài Trung Quốc đầu tiên là tại Philippines vào ngày 1 tháng 2,và ca tử vong đầu tiên ngoài châu Á (tại Pháp) là vào ngày 15 tháng 2 năm 2020. Tính đến ngày 24 tháng 2 năm 2020, ngoài lãnh thổ Trung Quốc đại lục, hơn chục người đã tử vong tại Iran, Hàn Quốc và Ý. Sau đó thêm các ca tử vong do coronavirus cũng được báo cáo tại Bắc Mỹ, Úc, San Marino, Tây Ban Nha, Iraq, và Anh Quốc và có thể cả CHDCND Triều Tiên.

Số ca tử vong trên toàn cầu do hoặc có liên quan tới COVID-19 đã vượt qua con số 10.000 người vào ngày 20 tháng 3 năm 2020, và hơn 207.008 Tính đến ngày 27 tháng 4 năm 2020. Vì vậy xây dựng hệ thống đánh giá tỉ lệ mắc bệnh và tỉ lệ chết để phát hiện sớm những nguy cơ dịch bệnh là vấn đề quan tâm nhất của gia đình và xã hội. Đề tài áp dụng

Môn khai phá dữ liệu xây dựng đánh giá các tỷ lệ với bộ dữ liệu thu thập được từ trong nước và ngoài nước

1.3 Mục tiêu đề tài

Đề tài tập chung vào nghiên cứu kỹ thuật phân cụm trong khai phá dữu liệu, từ đó nắm bắt được những giải thuật làm tiền đề cho nghiên cứu và xây dựng ứng dụng cụ thể. Sau khi phân tích đặc điểm của dữu liệu thu nhập đưuọc và lựa chọn giải thuật phù hợp với dữ liệu, việc xây dựng và đánh giá chất lượng, độ hiệu quả của hệ thống cũng là mục tiêu chính của đề tài.

1.4 Đối tượng và phương pháp nghiên cứu

Đề tài tập chung vào nghiên cứu kỹ thuật phân cụm trong khai phá dữ liệu cụ thể là nghiên cứu thuật toán k-means để áp dụng vào việc phân tích cơ sở dữ liệu tỷ lệ mắc và chết của covid-19 . thu nhập dữ liệu mắc bệnh va chết vi covid-19 từ các tình nguyện viện trên 200 quốc gia và vùng lãng thổ khác nhau .Sử dụng phương pháp và nghiên cứu hồi cứu với sự hỗ trợ chuyên môn của các bác sĩ chuyên khoa, đề tài tiến hành nghiên cứu trên cơ sở thuật toán phân cụm trong khai phá dữ liệu.

1.5 Ý nghĩa đề tài

1.5.1 Ý nghĩa khoa học

Với sự trợ giúp của máy tính, đề tài đóng góp một biện pháp thực hiện hỗ trợ các cán bộ y tế đánh giá bệnh cho bệnh nhân. Kết quả, Kinh nghiệm thu đưuọc khi thực hiện đề tài này sẽ giúp các cán bộ y tế phát hiện sớm bệnh cho bệnh nhân, đồng thời mong muốn những người đang công tác trong lĩnh vực y khoa và Khoa học máy tính ngồi lại với nhau để tìm ra những giải pháp tốt hơn trong vấn đề điều trị bệnh bằng cách kết hợp giữa 2 lịnh vực y học và khoa học máy tính.

1.5.2 Ý nghĩa thực tiễn

Đánh giá tỷ lệ nhiễm , chết do virus và phát hiện bệnh là cả một quá trình, đòi hỏi các cán bộ y tế không những phải thật vững chuyên môn mà còn có đầy đủ các trang thiết bị y tế mới có thể chuẩn đoán chính xác bệnh cho bệnh nhân. Nếu

chuẩn đoán sai bệnh sẽ đưa đến điều trị sai, không phát hiện sớm bệnh cho bệnh nhân,…

1.6 Bố cục đề tài

Đề tài được chia thành các phần:

Chương 1: Tổng quan đề tài

Chương 2: Khai phá dữ liệu

Chương 3: Kỹ thuật phân cụm và sử dụng thuật toán K-means

Chương 4: Thực nghiệm và đánh giá

CHƯƠNG 2: KHAI PHÁ DỮ LIỆU

2.1 Tổng quan về kỹ thuật Khai phá dữ liệu (Data mining)

2.1.1 Khái niệm về khai phá dữ liệu

Khai phá dữ liệu (data mining) là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính. Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp.Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến. Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu" hoặc KDD.

Khai phá dữ liệu là một bước của quá trình khai thác tri thức (Knowledge Discovery Process), bao gồm:

• Xác định vấn đề và không gian dữ liệu để giải quyết vấn đề (Problem understanding and data understanding).

• Chuẩn bị dữ liệu (Data preparation), bao gồm các quá trình làm sạch dữ liệu (data cleaning), tích hợp dữ liệu (data integration), chọn dữ liệu (data selection), biến đổi dữ liệu (data transformation).

• Khai thác dữ liệu (Data mining): xác định nhiệm vụ khai thác dữ liệu và lựa chọn kỹ thuật khai thác dữ liệu. Kết quả cho ta một nguồn tri thức thô.

• Đánh giá (Evaluation): dựa trên một số tiêu chí tiến hành kiểm tra và lọc nguồn tri thức thu được.

• Triển khai (Deployment).

Quá trình khai thác tri thức không chỉ là một quá trình tuần tự từ bước đầu tiên đến bước cuối cùng mà là một quá trình lặp và có quay trở lại các bước đã qua.

2.1.2 Quy trình khai phá dữ liệu

2.1.2.1 Nghiên cứu lĩnh vực

Ta cần nghiên cứu lĩnh vực cần sử dụng Data mining để xác định được những tri thức ta cần chất lọc, từ đó định hướng để tránh tốn thời gian cho những tri thức không cần thiết.

2.1.2.2 Tạo tập tin dữ liễu đầu vào

Ta xây dựng tập tin để lưu trữ các dữ liệu đầu vào để máy tính có thể lưu trữ và xử lý.

2.1.2.3 Tiền xử lý, làm sạch, mã hóa

Ở bước này ta tiến hành bỏ bớt những dữ liệu rườm rà, không cần thiết, tinh chỉnh lại cấu trúc của dữ liệu và mã hóa chúng để tiện cho quá trình xử lý .

2.1.2.4 Rút gọn chiều

Thông thường một tập dữ liệu có chiều khá lớn sẽ sinh ra một lượng dự liệu khổng lồ,ví dụ với n chiều ta sẽ có 2^n nguyên tổ hợp .Do đó , đây là một bước quan trọng giúp giảm đáng kể hao tổn hề tài nguyên trong quá trình xử lý tri thức.Thông thường ta sẽ dùng Rough set (http://en.wikipedia.org/wiki/Rough_set) để giảm số chiều.

2.1.2.5 Chọn tác vụ khai thác dữ liệu

Để đạt được mục đích ta cần, ta chọn được tác vụ khai thác dữ liệu sao cho phù hợp.Thông thường có các tác vụ sau:

- Đặc trưng(feature)
- Phân biệt(discrimination)
- Kết hợp(association)

• Phân lớp(classification)

• Gom cụm(clusterity)

• Xu thế(trend analysis)

• Phân tích độ lệch

• Phân tích độ hiếm

2.1.2.6 Khai thác dữ liệu: Tìm kiếm tri thức

Sau khi tiến hành các bước trên thì đây là bước chính của cả quá trình , ta sẽ tiến hành khai thác và tìm kiếm tri thức.

2.1.2.7 Đánh giá mẫu tìm được

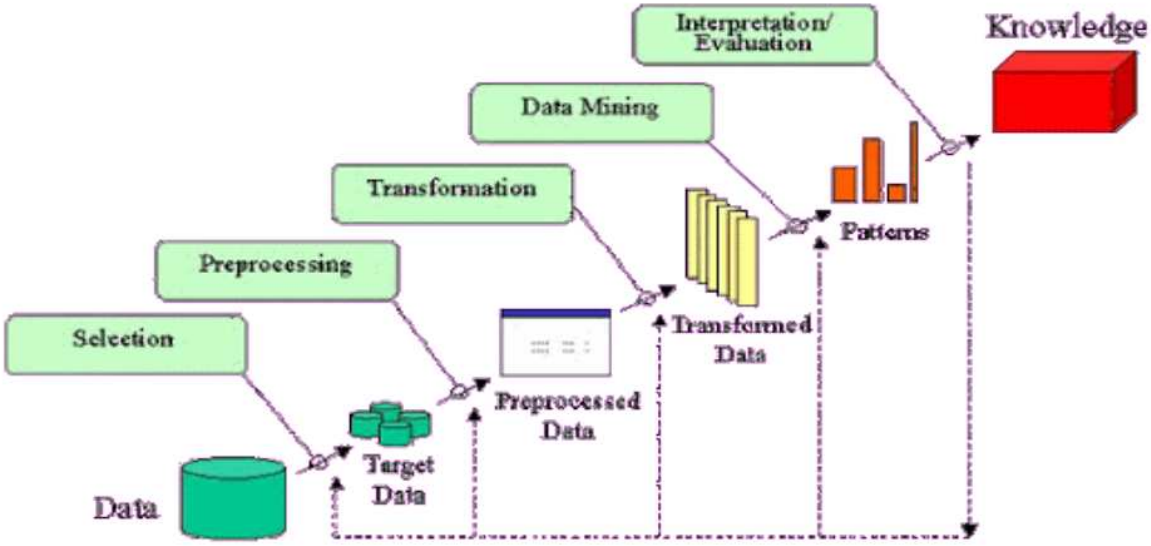
Ta cần đánh giá lại trong các tri thức tìm được , ta sẽ sử dụng được những tri thức nào , những tri thức nào dư thừa,không cần biết.

2.1.2.8 Biểu diễn tri thức

Ta biểu diễn tri thức vừa thu nhập được dưới dạng ngôn ngữ tự nhiên và hình thức sao cho người dùng có thể hiểu được những tri thức đó.

2.1.2.9 Sử dụng các tri thức vừa khám phá

Ta có thể tham khảo tiến trình KDD(Knowledge Discovery in Databases) để hiểu rõ hơn về khai phá dữ liệu:



Hình 2. 1 Knowledge Discovery in Databases

Chuẩn bị dữ liệu (data preparation), bao gồm các quá trình làm sạch dữ liệu (data cleaning), tích hợp dữu liệu (data integration), chọn dữ liệu (data selection), biến đổi dữ liệu (data transformation).

Khai thác dữu liệu (data mining): xác định nhiệm vụ khai thác dữ liệu và lựa chọn kỹ thuật khai thác dữu liệu. Kết quả cho ta một nguồn tri thức thô.

Đánh giá (evaluation): dựa trên một tiêu chí tiến hành kiểm tra và lọc nguồn tri thức thu được.

Triển khai (deployment).

Quá trình khai thác tri thức không chỉ là một quá trình tuần tự từ bước đầu tiên đến bước cuối cùng mà là một quá trình lặp và có quay trở lại các bước đã qua.

2.1.3 Ứng dụng của khai phá dữ liệu

Kinh tế - ứng dụng trong kinh doanh, tài chính, tiếp thị bán hàng, bảo hiểm, thương mại, ngân hàng,.. Đưa ra các bản báo cáo giàu thông tin, phân tích rủi ro trước khi đưa ra các chiến lược kinh doanh, sản xuất, phân loại khách hàng từ đó phân định ra thị trường, thị phần:…

Khoa học: Thiên văn học - dự đoán đường đi các thiên thể, hành tinh,...; Công nghệ sinh học – tìm ra các gen mới, cây con giống mới,…

Web: các công cụ tìm kiếm.

2.2 Tổng quan về hệ hỗ trợ ra quyết định

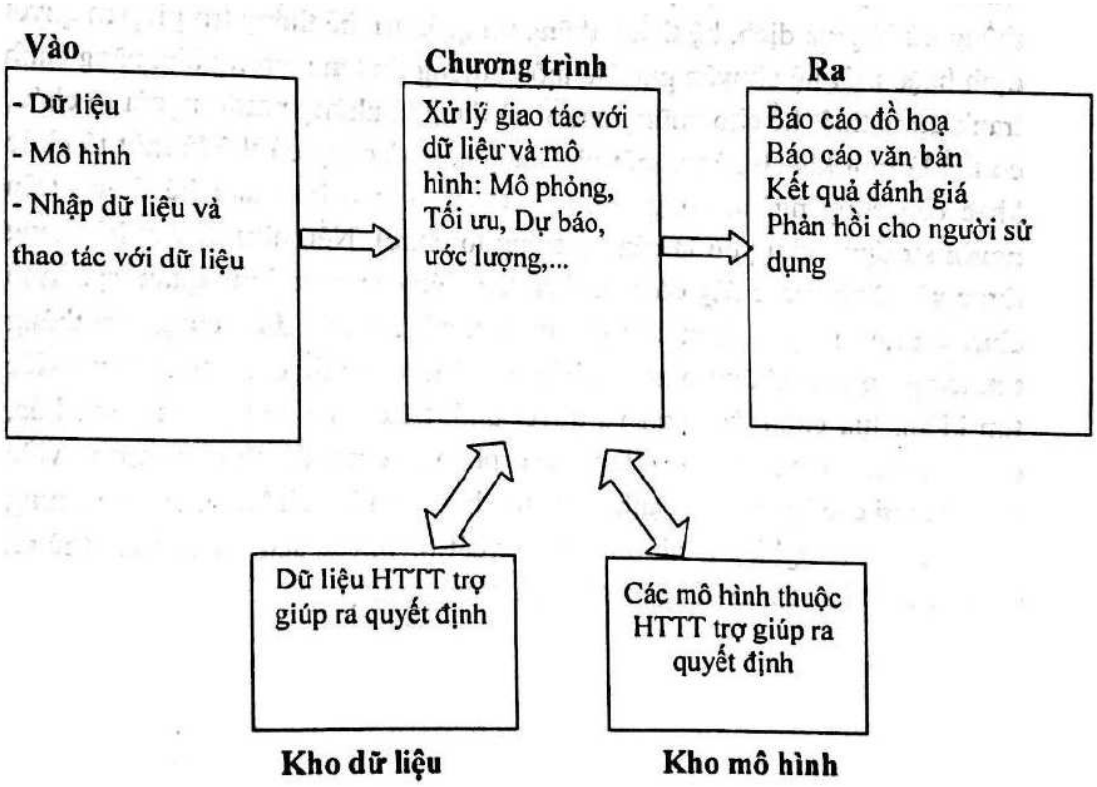
Hệ hỗ trợ ra quyết định là một hệ thống thuộc hệ thống thông tin, có nhiệm vụ cung cấp các thông tin hỗ trợ cho việc ra quyết định để tham khảo và giải quyết vấn đề. Hệ hỗ trợ ra quyết định có thể dùng cho cá nhân hay tổ chức và có thể hỗ trợ gián tiếp hoặc trực tiếp.

Trong lĩnh vực y tế, hệ hỗ trợ ra quyết định dựa vào tri thức đã học sẽ cung cấp thông tin chẩn đoán bệnh cho nhân viên y tế. Thông tin này được trích lọc để cung cấp một cách thông minh có giá trị cho quá trình chuẩn đoán, theo dõi và điều trị bệnh hiệu quả hơn, từ đó ta thấy một số lợi ích của hệ hỗ trợ ra quyết định trong y tế như sau:

• Tăng cường chất lượng chuẩn đoán, chăm sóc bệnh nhân.

• Giảm nguy cơ sai sót để tránh các tình huống nguy hiểm cho bệnh nhân.

• Tăng cường hiệu quả ứng dụng công nghệ thông tin vào lĩnh vực y tế để giảm bớt những thủ tục giấy tờ không cần thiết.



Hình 2. 2 Sơ đồ hệ hỗ trợ quyết định

2.3 Phân cụm dữ liệu và ứng dụng

2.3.1 Mục đích của phân cụm dữ liệu

Phân loại là một trong những hành vi nguyên thuỷ nhất của con người nhằm nắm giữ lượng thông tin khổng lồ họ nhận được hằng ngày vì sự xử lý mọi thông tin như một thực thể đơn lẻ là không thể. Phân cụm dữ liệu nhằm mục đích chính là khai phá cấu trúc của mẫu dữ liệu để thành lập các nhóm dữ liệu từ tập dữ liệu lớn, theo đó, cho phép người ta đi sâu vào phân tích và nghiên cứu cho từng cụm dữ liệu này nhằm khai phá và tìm kiếm các thông tin tiềm ẩn, hữu ích phục vụ cho ra quyết định.

Một vài ví dụ về ý nghĩa thực tiễn của phân cụm dữ liệu như sau :

- Khám phá ra các vị trí địa lý thuận lợi cho việc xây dựng các kho hàng phục vụ mua bàn hàng của một công ty thương mại

- Xác định các cụm ảnh như ảnh của các loài động vật như loài thú, chim,... trong tập CSDL ảnh về động vật nhằm phục vụ cho việc tìm kiếm ảnh

- Xác định các nhóm người bệnh nhằm cung cấp thông tin cho việc phân phối các thuốc điều trị trong y tế

- Xác định nhóm các khách hàng trong CSDL ngân hàng có vốn các đầu tư vào bất động sản cao...

Như vậy, phân cụm dữ liệu là một phương pháp xử lý thông tin quan trọng và phổ biến, nó nhằm khám phá mối liên hệ giữa các mẫu dữ liệu bằng cách tổ chức chúng thành các cụm tương tự.

Tiếp theo, giả sử rằng tất cả các dạng dữ liệu được biểu diễn bởi khái niệm đặc trưng, các đặc trưng hình thành nên vector đặc trưng l- chiều. Thuật ngữ phân cụm được hiểu là phân cụm dữ liệu.

2.3.2 Các bước cơ bản để phân cụm

Chọn lựa đặc trưng : Các đặc trưng phải được chọn lựa một cách hợp lý để có thể "mã hoá" nhiều nhất thông tin liên quan đến công việc quan tâm. Mục tiêu chính là phải giảm thiểu sự dư thừa thông tin giữa các đặc trưng. Các đặc trưng cần được tiền xử lý trước khi dùng trong các bước sau.

Chọn độ đo gần gũi : Đây là một độ đo chỉ ra mức độ tương tự hay không tương tự giữa hai vector đặc trưng. Phải đảm bảo rằng tất cả các vector đặc trưng góp phần như nhau trong việc tính toán độ đo gần gũi và không có đặc trưng nào át hẳn đặc trưng nào. Điều này được đảm nhận bởi quá trình tiền xử lý.

Tiêu chuẩn phân cụm : Điều này phụ thuộc vào sự giải thích của chuyên gia cho thuật ngữ "dễ nhận thấy" dựa vào loại của các cụm được chuyên gia cho rằng đang ẩn dấu dưới tập dữ liệu. Chẳng hạn, một cụm loại chặt (compact)

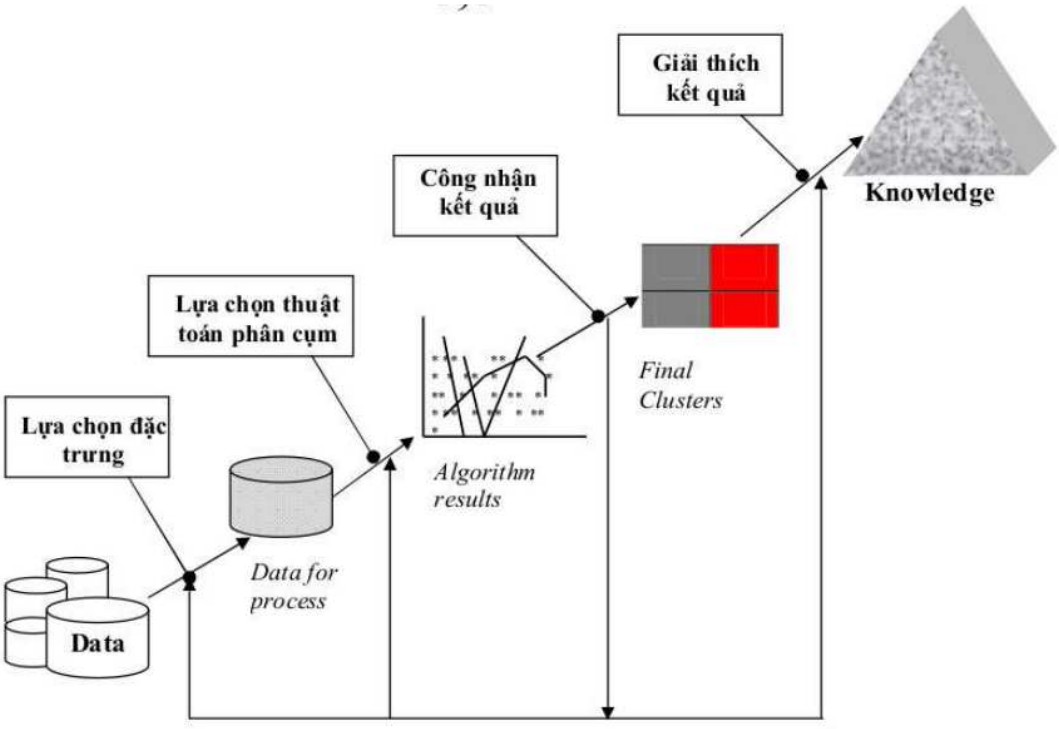
của các vector đặc trưng trong không gian l-chiều có thể dễ nhận thấy theo một tiêu chuẩn, trong khi một cụm loại "dài và mỏng" lại có thể đươc dễ nhận thấy bởi một tiêu chuẩn khác. Tiêu chuẩn phân loại có thể được diễn đạt bởi hàm chi phí hay một vài loại quy tắc khác.

Thuật toán phân loại : Cần lựa chọn một sơ đồ thuật toán riêng biệt nhằm làm sáng tỏ cấu trúc cụm của tập dữ liệu.

Công nhận kết quả : Khi đã có kết quả phân loại thì ta phải kiểm tra tính đúng đắn của nó. Điều này thường được thực hiện bởi việc dùng các kiểm định phù hợp.

Giải thích kết quả : Trong nhiều trường hợp, chuyên gia trong lĩnh vực ứng dụng phải kết hợp kết quả phân loại với bằng chứng thực nghiệm và phân tích để đưa ra các kết luận đúng đắn. Trong một số trường hợp, nên có cả bước khuynh hướng phân cụm; trong bước này có các kiểm định khác nhau để chỉ ra một dữ liệu có hay không một cấu trúc phân cụm. Ví dụ như tập dữ liệu của ta có thể hoàn toàn ngẫu nhiên vì vậy mọi cố gắng phân cụm đều vô nghĩa.

Các lựa chọn khác nhau của các đặc trưng, độ đo gần gũi, tiêu chuẩn phân cụm có thể dẫn tới các kết quả phân cụm khác nhau. Do đó, việc lựa chọn một cách hợp lý nhất hoàn toàn dựa vào kiến thức và kinh nghiệm của chuyên gia. Tính chủ quan (của chuyên gia) là một thực tế mà ta phải chấp nhận.



Hình 2. 3 Các bước trong quá trình phân cụm

2.3.3 Các loại đặc trưng

Có bốn loại đặc trưng, đó là:

Các đặc trưng danh nghĩa (nominal): Gồm các đặc trưng mà các giá trị của nó mã hoá các trạng thái. Chẳng hạn cho một đặc trưng là giới tính của một người thì các giá trị có thể của nó là 1 ứng với nam và 0 ứng với nữ. Rõ ràng là bất kỳ sự so sánh về lượng nào giữa các giá trị loại này đều là vô nghĩa.

Các đặc trưng thứ tự (ordinal): Là các đặc trưng mà các giá trị của nó có thể sắp một cách có ý nghĩa. Ví dụ về một đặc trưng thể hiện sự hoàn thành khoá học của một sinh viên. Giả sử các giá trị có thể là 4, 3, 2, 1 tương ứng với các ý nghĩa: ”xuất sắc”, “rất tốt“, “tốt“, “không tốt“. Các giá trị này được sắp xếp theo một thứ tự có ý nghĩa nhưng sự so sánh giữa hai giá trị liên tiếp là không

quan trọng lắm về lượng.

Các đặc trưng đo theo khoảng (interval –scaled): Với một đặc trưng cụ thể nếu sự khác biệt giữa hai giá trị là có ý nghĩa về mặt số lượng thì ta có đặc trưng đo theo khoảng (còn gọi là thang khoảng). Ví dụ về đặc trưng nhiệt độ, nếu từ 10-15 độ thì được coi là rét đậm, còn nếu dưới 10 độ thì được coi là rét hại, vì vậy mỗi khoảng nhiệt độ mang một ý nghĩa riêng.

Các đặc trưng đo theo tỷ lệ (ratio-scaled): Cũng với ví dụ nhiệt độ ở trên ta không thể coi tỷ lệ giữa nhiệt độ Hà Nội 10 độ với nhiệt độ Matxcơva 1 độ mang ý nghĩa rằng Hà Nội nóng gấp mười lần Matxcơva. Trong khi đó, một người nặng 100 kg được coi là nặng gấp hai lần một người nặng 50 kg. Đặc trưng cân nặng là một đặc trưng đo theo tỷ lệ (thang tỷ lệ).

2.3.4 Các ứng dụng của phân cụm

Phân cụm là một công cụ quan trọng trong một số ứng dụng. Sau đây là một số ứng dụng của nó:

Giảm dữ liệu: Giả sử ta có một lượng lớn dữ liệu (N). Phân cụm sẽ nhóm các dữ liệu này thành m cụm dữ liệu dễ nhận thấy và m << N. Sau đó xử lý mỗi cụm như một đối tượng đơn.

Rút ra các giả thuyết : Các giả thuyết này có liên quan đến tính tự nhiên của dữ liệu và phải được kiểm tra bởi việc dùng một số tập dữ liệu khác.

Kiểm định giả thuyết : Ta sẽ phân cụm để xét xem có tồn tại một tập dữ liệu nào đó trong tập dữ liệu thoả mãn các giả thuyết đã cho hay không. Chẳng hạn xem xét giả thuyết sau đây: “Các công ty lớn đầu tư ra nước ngoài “. Để kiểm tra, ta áp dụng kỹ thuật phân cụm với một tập đại diện lớn các công ty. Giả sử rằng mỗi công ty được đặc trưng bởi tầm vóc, các hoạt động ở nước ngoài và khả năng hoàn thành các dự án. Nếu sau khi phân cụm, một cụm các

công ty được hình thành gồm các công ty lớn và có vốn đầu tư ra nước ngoài (không quan tâm đến khả năng hoàn thành các dự án) thì giả thuyết đó được củng cố bởi kỹ thuật phân cụm đã thực hiện.

Dự đoán dựa trên các cụm : Đầu tiên ta sẽ phân cụm một tập dữ liệu thành các cụm mang đặc điểm của các dạng mà nó chứa. Sau đó, khi có một dạng mới chưa biết ta sẽ xác định xem nó sẽ có khả năng thuộc về cụm nào nhất và dự đoán được một số đặc điểm của dạng này nhờ các đặc trưng chung của cả cụm.

Cụ thể hơn, phân cụm dữ liệu đã được áp dụng cho một số ứng dụng điển hình trong các lĩnh vực sau [13] :

Thương mại : Trong thương mại, phân cụm có thể giúp các thương nhân khám phá ra các nhóm khách hàng quan trọng có các đặc trưng tương đồng nhau và đặc tả họ từ các mẫu mua bán trong cơ sở dữ liệu khách hàng.
Sinh học : Trong sinh học, phân cụm được sử dụng để xác định các loại sinh vật, phân loại các Gen với chức năng tương đồng và thu được các cấu trúc trong các mẫu.

Phân tích dữ liệu không gian : Do sự đồ sộ của dữ liệu không gian như dữ liệu thu được từ các hình ảnh chụp từ vệ tinh các thiết bị y học hoặc hệ thống thông tin địa lý (GIS), ...làm cho người dùng rất khó để kiểm tra các dữ liệu không gian một cách chi tiết. Phân cụm có thể trợ giúp người dùng tự động phân tích và xử lý các dữ liệu không gian như nhận dạng và chiết xuất các đặc tính hoặc các mẫu dữ liệu quan tâm có thể tồn tại trong cơ sở dữ liệu không gian.

Lập quy hoạch đô thị : Nhận dạng các nhóm nhà theo kiểu và vị trí địa

lý,...nhằm cung cấp thông tin cho quy hoạch đô thị.

Nghiên cứu trái đất : Phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho nhận dạng các vùng nguy hiểm.

Địa lý : Phân lớp các động vật và thực vật và đưa ra đặc trưng của chúng.

Web Mining : Phân cụm có thể khám phá các nhóm tài liệu quan trọng, có nhiều ý nghĩa trong môi trường Web. Các lớp tài liệu này trợ giúp cho việc khám phá tri thức từ dữ liệu,...

2.3.5 Phân loại các thuật toán phân cụm

Các thuật toán phân cụm có thể được xem như các sơ đồ cung cấp cho ta các cụm "dễ nhận thấy" bởi việc chỉ xem xét một phần của tập chứa tất cả các cách phân cụm của X. Kết quả phân cụm phụ thuộc vào thuật toán và tiêu chuẩn phân cụm.

Như vậy, một thuật toán phân cụm là một chức năng học cố gắng tìm ra các đặc trưng riêng biệt của các cụm ẩn dấu dưới tập dữ liệu. Có nhiều cách để phân loại các thuật toán phân cụm, sau đây là một cách phân loại:

Các thuật toán phân cụm tuần tự (Sequential Algorithms):

Các thuật toán này sinh ra một cách phân cụm duy nhất, chúng là các phương pháp trực tiếp và nhanh. Trong hầu hết các thuật toán thuộc loại này, tất cả các vector đặc trưng tham gia vào thuật toán một hoặc vài lần (không hơn 6 lần). Kết quả cuối cùng thường phụ thuộc vào thứ tự các vector tham gia vào thuật toán.

Những sơ đồ loại này có khuynh hướng sinh ra các cụm có hình dạng chặt siêu cầu hoặc siêu elipxoit tuỳ theo độ đo được dùng.

Các thuật toán phân cụm phân cấp (Hierarchical Algorithms)

- Các thuật toán tích tụ (Agglomerative Algorithms): Chúng sinh ra một dãy cách phân cụm mà số cụm, m, giảm dần ở mỗi bước. Cách phân cụm ở mỗi bước là kết quả của cách phân cụm ở bước trước đó bằng việc trộn hai cụm vào một. Các đại

diện chính của loại này là thuật toán liên kết đơn (phù hợp với các cụm dài và mỏng) và thuật toán liên kết đầy đủ (phù hợp với các cụm chặt). Các thuật toán tích tụ thường dựa trên lý thuyết đồ thị và lý thuyết ma trận.

- Các thuật toán phân rã (Divise Algorithms): Sinh ra một dãy cách phân cụm mà số cụm, m, tăng dần ở mỗi bước. Cách phân cụm ở mỗi bước là kết quả cách phân cụm ở bước trước đó bằng việc chia đôi một cụm đơn.

- Các thuật toán phân cụm dựa trên việc tối ưu hoá hàm chi phí: Hàm chi phí J đo độ "dễ nhận thấy" của các cách phân cụm. Thường thì số các cụm, m, là cố định. Thuật toán sẽ dùng các khái niệm về phép tính vi phân và sinh ra các cách phân cụm liên tiếp trong khi cố gắng tối ưu hoá J. Thuật toán sẽ dừng khi một tối ưu địa phương được xác định. Các thuật toán loại này cũng được gọi là các sơ đồ tối ưu hoá hàm lặp. Chúng được phân tiếp như sau:

- Các thuật toán phân cụm chặt hay rõ: Vector thuộc hoàn toàn vào một cụm cụ thể. Việc đưa một vector về các cụm cụ thể được thực hiện một cách tối ưu theo tiêu chuẩn phân cụm tối ưu.

- Các thuật toán phân cụm theo các hàm xác suất: Dựa vào lý thuyết phân lớp Bayes và mỗi vector x được phân về cụm thứ i nếu p(C i | x) là lớn nhất (xác suất để x được phân đúng vào cụm C i).

- Các thuật toán phân cụm mờ: Các vector thuộc về một cụm nào đó với một độ chắc chắn.

- Các thuật toán phân cụm theo khả năng : Trong trường hợp này ta đo khả năng một vector đặc trưng thuộc về một cụm nào đó.

- Các thuật toán phát hiện biên phân tách : Các thuật toán này cố gắng đặt các biên phân tách một cách tối ưu giữa các cụm.

Các thuật toán khác

- Các thuật toán phân cụm nhánh và cận : Các thuật toán này cung cấp cho ta các cách phân cụm tối ưu toàn cục mà không phải xét tới tất cả các cách phân cụm có thể, với m cố định và một tiêu chuẩn phân cụm định trước. Nhưng đòi hỏi rất nhiều tính toán.

- Các thuật toán phân cụm di truyền : Sử dụng dân số ban đầu của các cách phân cụm có thể và sinh ra các số dân mới một cách lặp đi lặp lại. Số dân mới này nhìn chung chứa các cách phân cụm tốt hơn so với thế hệ trước, theo một tiêu chuẩn đã định trước.

- Phương pháp thư giãn ngẫu nhiên : Đảm bảo rằng với các điều kiện chắc chắn, độ hội tụ theo xác suất tới cách phân cụm tối ưu toàn cục nhưng tốn nhiều thời gian tính toán.

- Thuật toán phân cụm tìm khe : Xem mỗi vector đặc trưng như là một biến ngẫu nhiên x. Chúng dựa trên một giả định được công nhận rộng rãi rằng vùng phân bố của x nơi có nhiều vector tương ứng với vùng mật độ cao của hàm mật độ xác suất (probability density function), vì vậy việc ước lượng các hàm mật độ xác suất sẽ làm rõ các khu vực nơi các cụm hình thành.

- Thuật toán học cạnh tranh: Không dùng các hàm chi phí, chúng tạo ra vài cách phân cụm và các cách này hội tụ tới cách dễ nhận thấy nhất. Các đại diện tiêu biểu của loại này là sơ đồ học cạnh tranh cơ bản và thuật toán học lỗ rò.

- Các thuật toán dựa trên kỹ thuật biến đổi hình thái học : Cố gắng đạt được sự phân chia tốt hơn giữa các cụm.

2.4 Cơ sở dữ liệu Y khoa

2.4.1 Sơ lược về Đại dịch covid-19

Đại dịch COVID-19 là một đại dịch bệnh truyền nhiễm với tác nhân là virus SARS-CoV-2, đang diễn ra trên phạm vi toàn cầu.Khởi nguồn vào tháng 12 năm 2019 với tâm dịch đầu tiên được ghi nhận tại thành phố Vũ Hán thuộc miền Trung Trung Quốc, bắt nguồn từ một nhóm người mắc viêm phổi không rõ nguyên

nhân. Các nhà khoa học Trung Quốc đã tiến hành nghiên cứu và phân lập được một chủng loại corona virus mới, được Tổ chức Y tế Thế giới lúc đó tạm thời gọi là 2019-nCoV, có trình tự gen giống với SARS-CoV trước đây với mức tương đồng lên tới 79,5%.

Virus corona chủ yếu ảnh hưởng đến đường hô hấp dưới (cũng có các triệu chứng ở đường hô hấp trên nhưng ít gặp hơn) và dẫn đến một loạt các triệu chứng được mô tả giống như cúm,bao gồm sốt, ho, khó thở, đau cơ và mệt mỏi, với sự phát triển cao hơn nữa sẽ dẫn đến viêm phổi, hội chứng suy hô hấp cấp tính, nhiễm trùng huyết, sốc nhiễm trùng và có thể gây tử vong. Các phản ứng y tế đối với căn bệnh này thường là cố gắng kiểm soát các triệu chứng lâm sàng vì hiện tại chưa tìm thấy phương pháp điều trị hiệu quả nào.

2.4.2 Sự lây truyền

Virus corona chủng mới chủ yếu lây lan qua các giọt bắn trong không khí khi một cá nhân bị nhiễm bệnh ho hoặc hắt hơi trong phạm vi khoảng 3 foot (0,91 m) đến 6 foot (1,8 m).Trong số 41 trường hợp ban đầu, hai phần ba có tiền sử tiếp xúc với Chợ bán buôn hải sản Hoa Nam. Tháng 5 năm 2020, một nghiên cứu tại Đại học Hong Kong - Trung Quốc cũng cho biết virus này cũng lây qua mắt cao gấp 100 lần so với SARS

Hệ số lây nhiễm cơ bản R_0

Khả năng lây lan virus giữa người với người khá đa dạng, có người mắc nhưng không truyền virus, có người lại có khả năng truyền bệnh cho nhiều người. Hệ số lây nhiễm cơ bản R_0 (cũng được gọi là *hệ số sinh sản cơ bản* hoặc *hệ số sinh sản cơ sở*) chỉ ra khả năng truyền virus từ người sang người, được ước tính là từ 2 đến 4 (R_0=2÷4). Con số này có ý nghĩa: trong quần thể người, một người mới nhiễm có khả năng truyền virus cho bao nhiêu người khác và khiến họ mắc bệnh. Như vậy, theo như các báo cáo hiện nay, một người mắc chủng coronavirus này có thể lan truyền cho 4 người khác.

2.4.3 Dấu hiệu và triệu chứng

Các triệu chứng của COVID-19. Có báo cáo virus tồn tại trong cơ thể người nhưng không gây triệu chứng. Các triệu chứng được báo cáo gồm sốt trong 90% trường hợp mắc bệnh, mệt mỏi và ho khan trong 80% trường hợp, 20% bị khó thở và suy hô hấp chiếm 15%. X-quang ngực đã tiết lộ các dấu hiệu ở cả hai phổi. Dấu hiệu sống nhìn chung là ổn định vào thời điểm nhập viện của những bệnh nhân. Các xét nghiệm máu thường cho thấy số lượng bạch cầu thấp (giảm bạch cầu và giảm bach cầu lympho). Nhiều bệnh nhân còn có thể gặp các biểu hiện ngoài da, đặc biệt là ở các ngón chân

Tuy nhiên, theo Trung tâm Kiểm soát và Phòng ngừa dịch bệnh Mỹ (CDC), 25% số người bệnh có thể không có triệu chứng gì hoặc triệu chứng không rõ ràng.

Chuẩn đoán mắc bệnh.

Triệu chứng	Tỷ lệ
Sốt	83–99%
Ho	59–82%
Mất vị giác	40–84%
Mệt mỏi	44–70%
Khó thở	31–40%
Ho có đờm	28–33%
Đau và nhức cơ	11–35%

Bảng 2. 1 Triệu chứng và tỉ lệ mắc bệnh

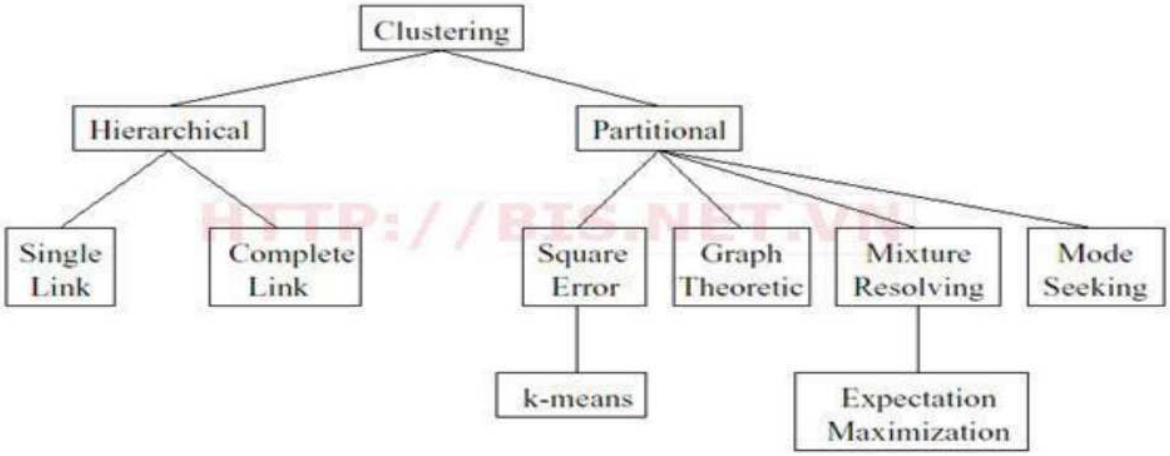
CHƯƠNG 3: KỸ THUẬT PHÂN CỤM VÀ THUẬT TOÁN K-MENAS

3.1 Giới thiệu về kỹ thuật phân cụm trong Khai phá dữ liệu

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 của của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: data reduction, "natural clusters", "useful" clusters, outlier detection

Các kỹ thuật phân cụm được phân loại như sau (xem hình)

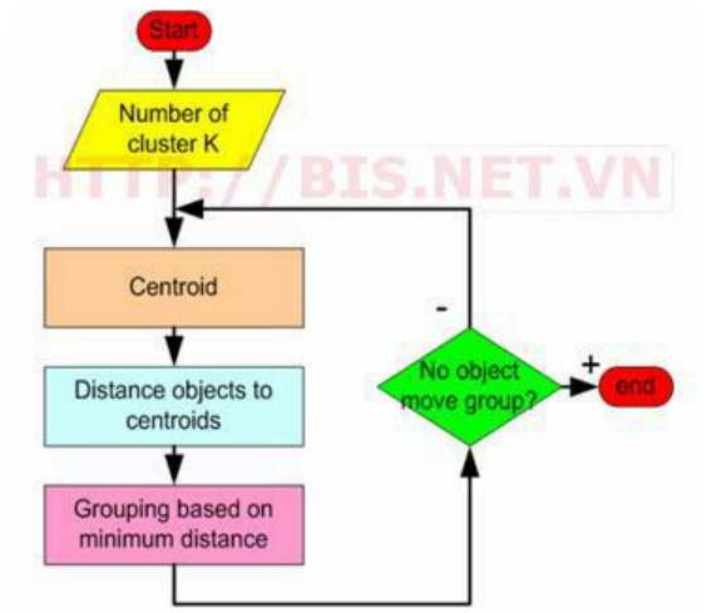


Hình 3. 1 Các kỹ thuật phân cụm

3.2 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 đinh trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất.

Thuật toán K-Means được mô tả như sau



Hình 3. 2 Mô tả thuật toán K-Means

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

1. Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.
2. Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean)
3. Nhóm các đối tượng vào nhóm gần nhất
4. Xác định lại tâm mới cho các nhóm

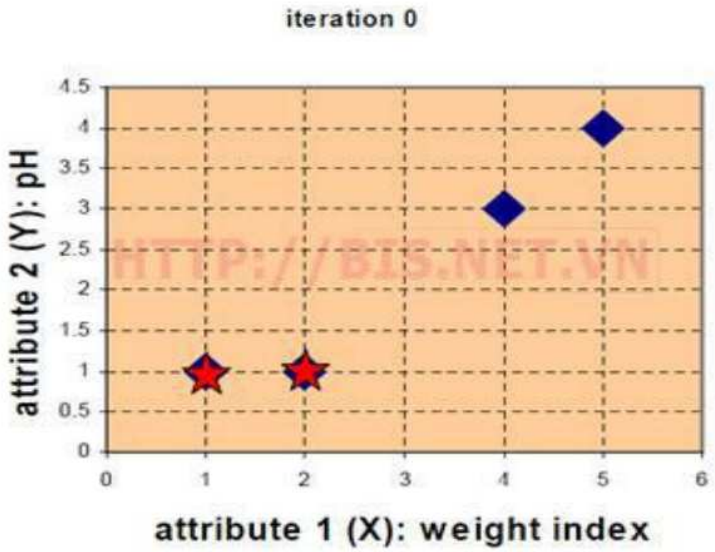
5. Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng

Ví dụ minh họa thuật toán K-Mean:

Giả sử ta có 4 loại thuốc A,B,C,D, mỗi loại thuộc 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 c1(1,1)) và B là tâm của nhóm thứ 2 (tạo độ tâm nhóm thứ hai c2 (2,1)).



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

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (2,1) & group-2 \end{matrix}$$

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

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

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 (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:

$$\mathbf{c}_1 = (1,1) \quad \sqrt{(4-1)^2 + (3-1)^2} = 3.61$$

$$\mathbf{c}_2 = (2,1) \quad \sqrt{(4-2)^2 + (3-1)^2} = 2.83$$

Bước 3. Nhóm các đối tượng vào nhóm gần nhất

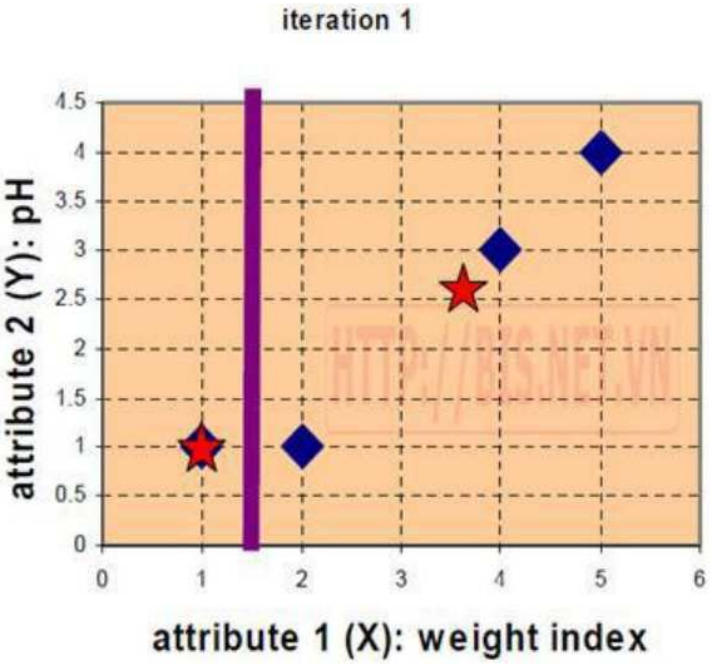
$$\mathbf{G}^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group-1 \\ group-2 \end{matrix}$$

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

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 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, c1(1,1). Tâm nhóm 2 được tính như sau:

$$\mathbf{c}_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = \left(\frac{11}{3}, \frac{8}{3}\right).$$



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

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

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

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

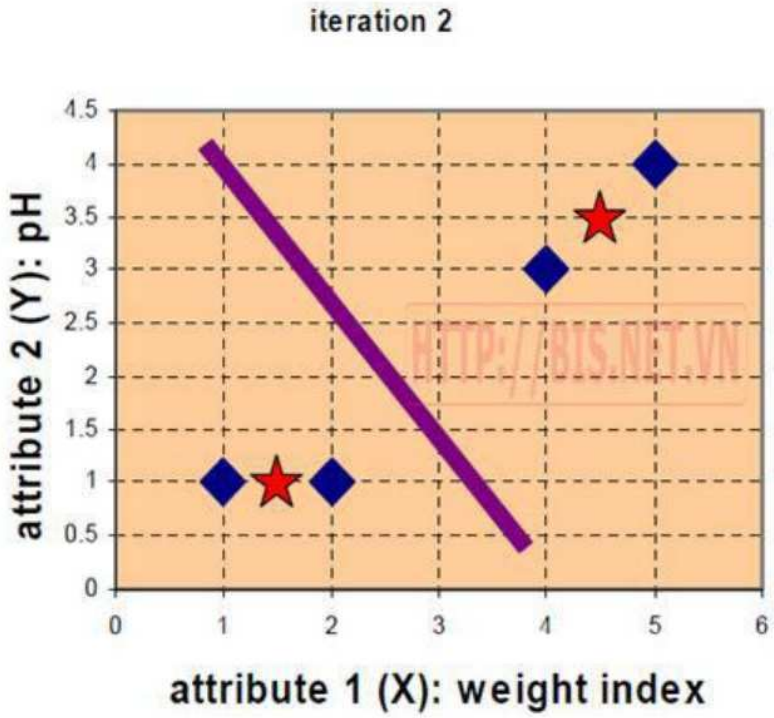
Bước 7. Nhóm các đối tượng vào nhóm

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

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

Bước 8. Tính lại tâm cho nhóm mới

$$\mathbf{c}_1 = (\frac{1+2}{2}, \frac{1+1}{2}) = (1\tfrac{1}{2}, 1) \qquad \mathbf{c}_2 = (\frac{4+5}{2}, \frac{3+4}{2}) = (4\tfrac{1}{2}, 3\tfrac{1}{2})$$



Bước 9. Tính lại khoảng cách từ các đối tượng đến tâm mới

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

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

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

Bước 10. Nhóm các đối tượng vào nhóm

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

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

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

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.

3.3 Áp dụng và sử dụng thuật toán K-means vào bộ dataset Covid-19

Tập dữ liệu covid-19 bao gồm dữ liệu của 200 quốc gia gồm những nước đã có người chết vì virus covid-19 và các nước chưa có người chết vì covid-19. Tập dữ liệu bao gồm các thuộc tính như sau:

1. Quốc Gia
2. Tỉ lệ người nhiễm trên 1 triệu người.
3. Tỷ lệ người chết trên 1 triệu người.
4. Tổng số người kiểm tra của tất cả các nước
5. Tỷ lệ kiểm tra trên 1 triệu người

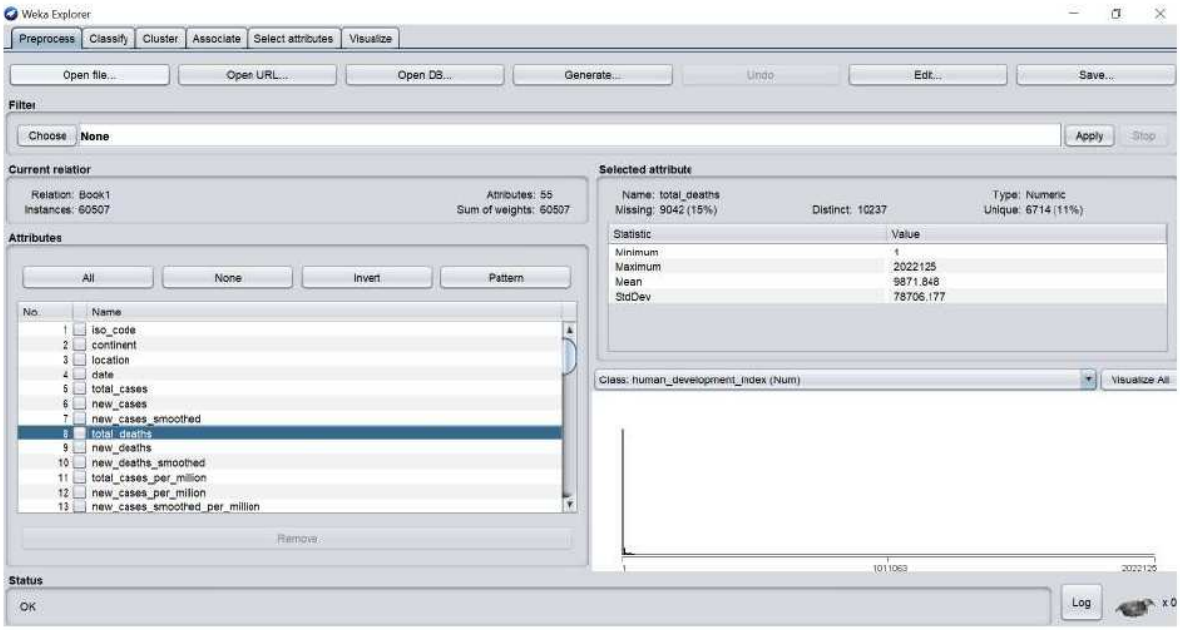
Đây là một bài toán phân cụm và chúng ta có thể sử dụng các phương pháp phân cụm khác như k-Medians, Expectation Maximization (EM) để phân loại cũng cho kết quả khá tốt. Chúng ta có thể hình dung tập dữ liệu này thông qua biểu diễn dưới dạng file CSV như sau, các cột từ 1 đến 5 tương ứng với các chỉ số nếu trên

XXXXXXX

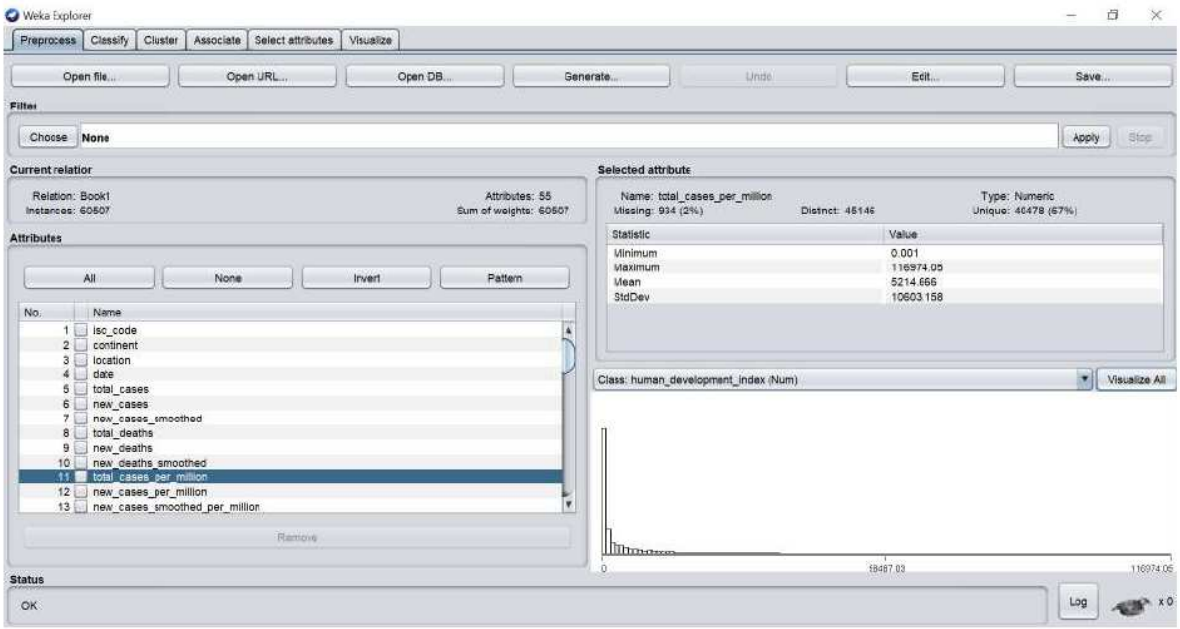
Hình 3. 3 Tập dữ liệu Covid-19 sau khi phân cụm

CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ

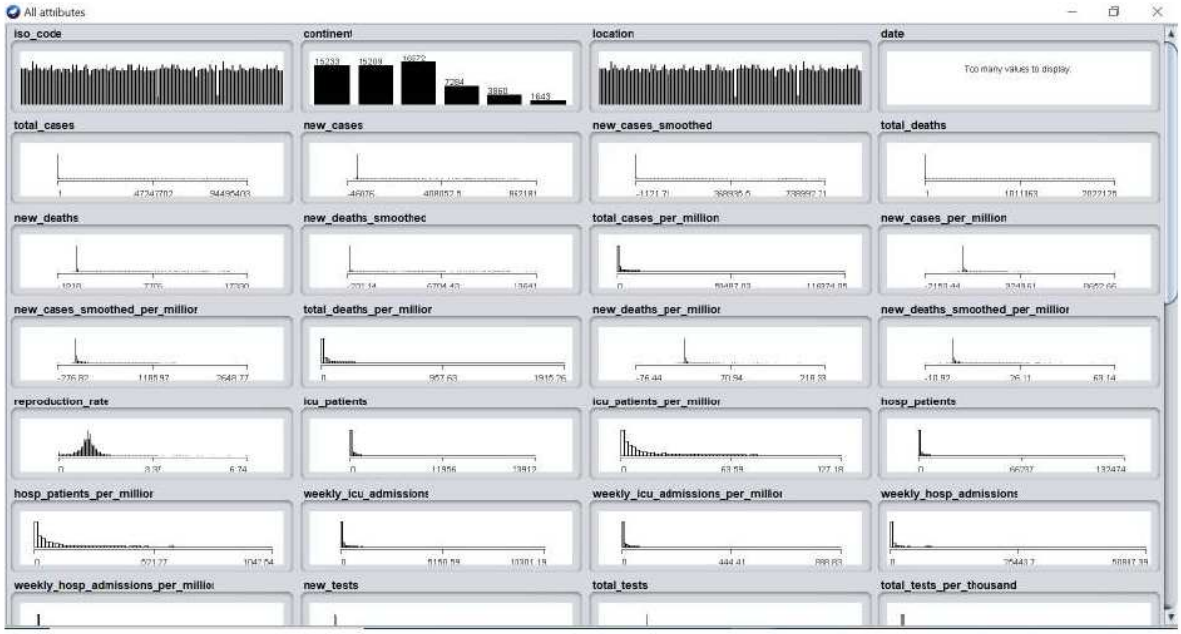
4.1 Xây dựng mô hình bằng Weka



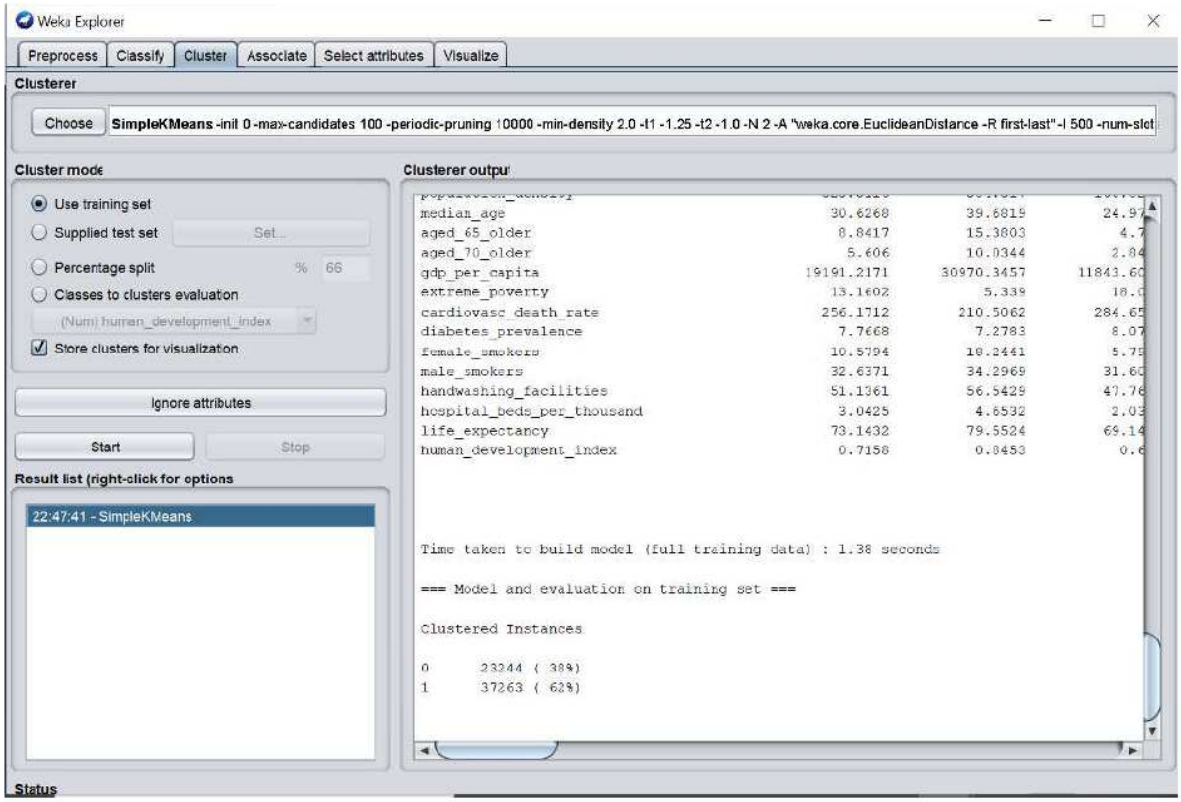
Hình 4. 1 Nhập dữ liệu vào Weka



Hình 4. 2 Dữ liệu đưa vào được phân đoạn – tiền xử lý



Hình 4. 3 Các thuộc tính bộ dữ liệu tỷ lệ người chết và nhiễm virus trên 1 triệu người

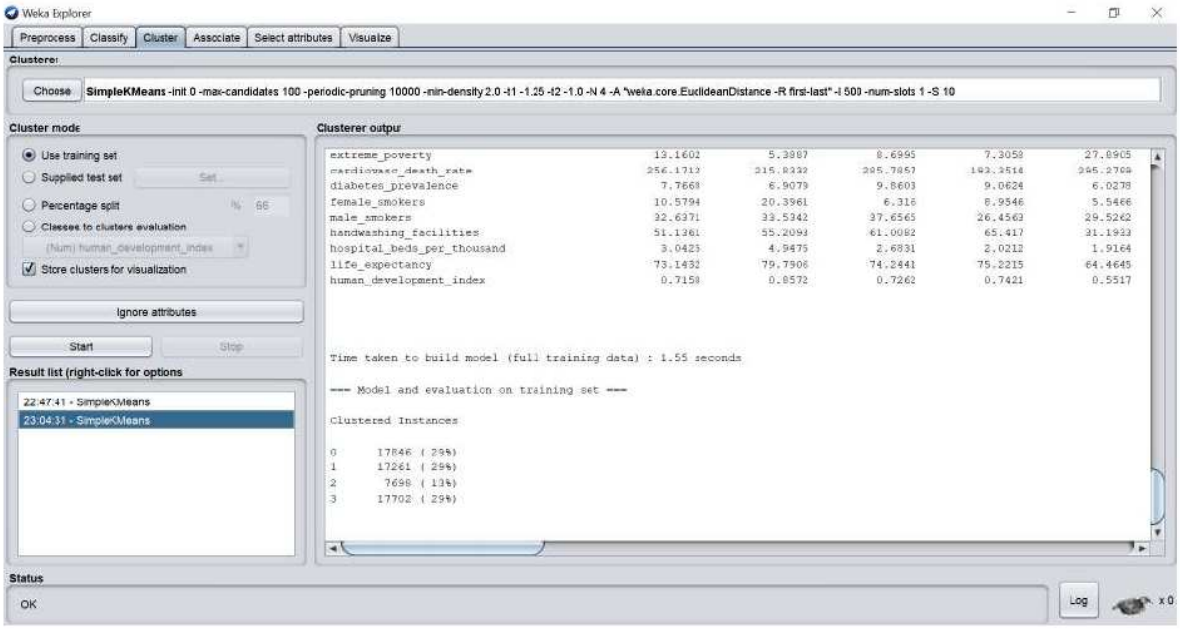


Hình 4. 4 Đầu ra phân lớp

• Đầu ra phân lớp:

Các thuộc tính:

1. Quốc Gia
2. Tỉ lệ người nhiễm trên 1 triệu người.
3. Tỷ lệ người chết trên 1 triệu người.
4. Tổng số người kiểm tra của tất cả các nước
5. Tỷ lệ kiểm tra trên 1 triệu người

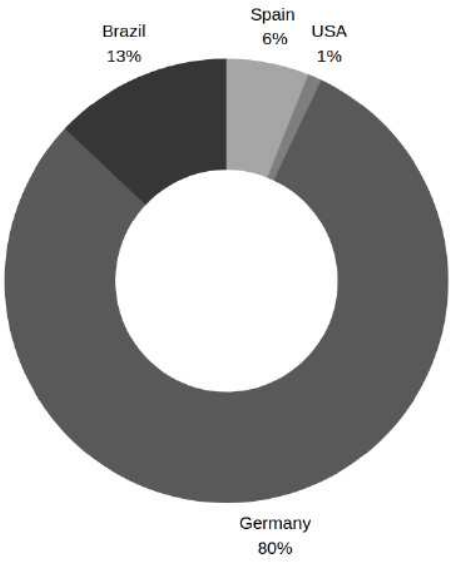


Hình 4. 5 Đầu ra phân cụm bằng K-means với tất cả thuộc tính

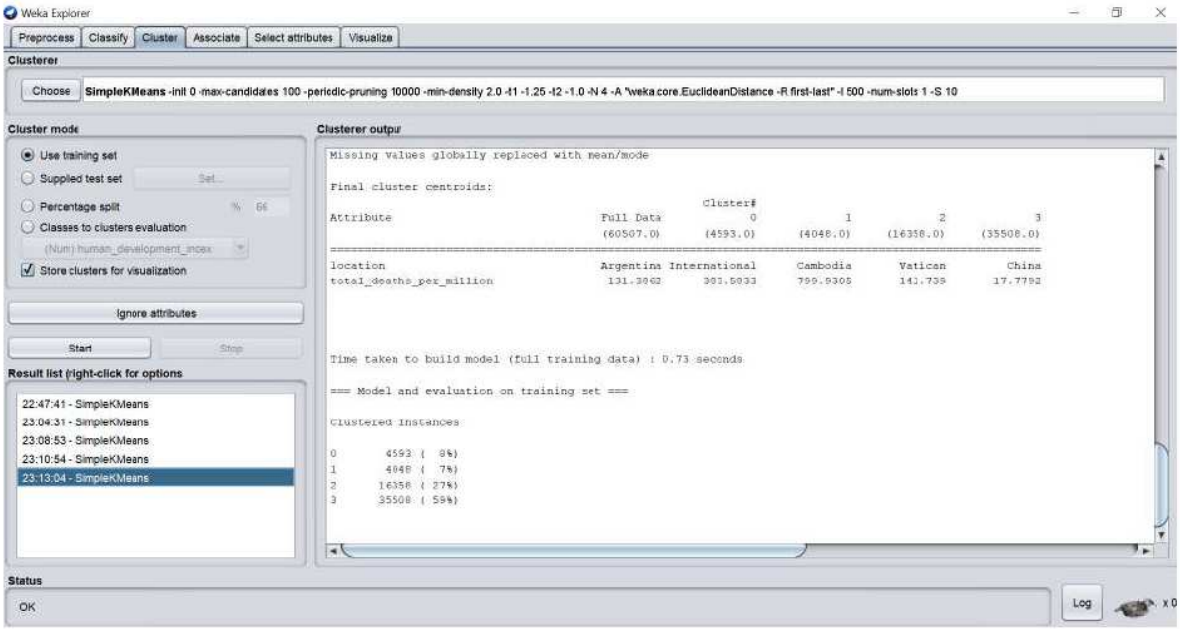
Bảng phân tích dữ liệu:

	0	1	2	3
Tâm là Quốc Gia	Spain	USA	Germany	Brazil
Tỷ lệ che phủ toàn bộ dữ liệu	6%	1%	80%	13%

Bàng 4. 1 Bảng phân tích dữ liệu đầu ra với tất cả cac thuộc tính



Hình 4. 6 Biểu đồ tỷ lệ các cụm theo toàn bộ thuộc tính trên toàn bộ dữ liệu

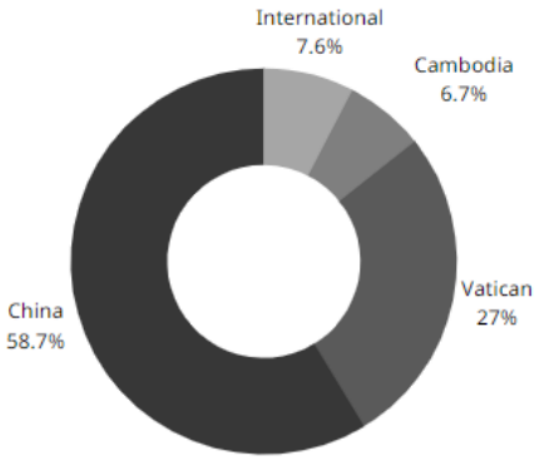


Hình 4. 7 Đầu ra phân cụm bằng K-means với thuộc tính quốc gia và tỷ lệ người chết

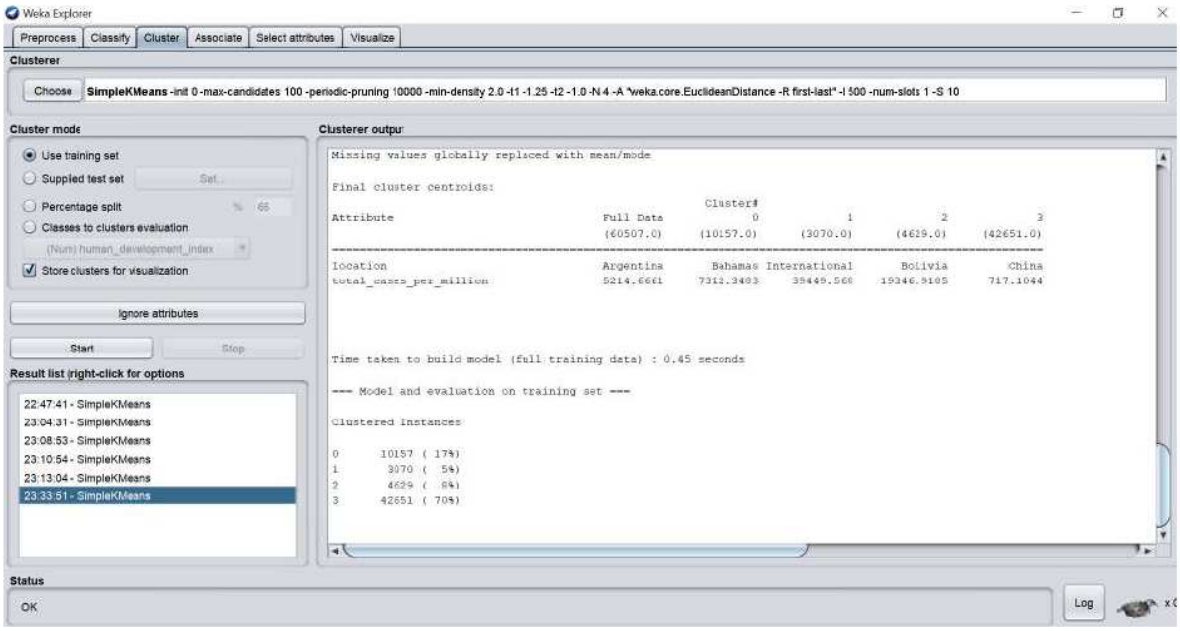
Bảng phân tích dữ liệu.

Cluster	0	1	2	3
Tâm là Quốc Gia	International	Cambodia	Vatican	China
Tỷ lệ che phủ toàn bộ dữ liệu	8 %	7%	27%	59%

Bàng 4. 2 Bảng phân tích dữ liệu đầu ra với thuộc tính Quốc gia và tỷ lệ người chết



Hình 4. 8 Biểu đồ tỷ lệ các cụm theo thuộc tính quốc gia và người chết trên toàn bộ dữ liệu

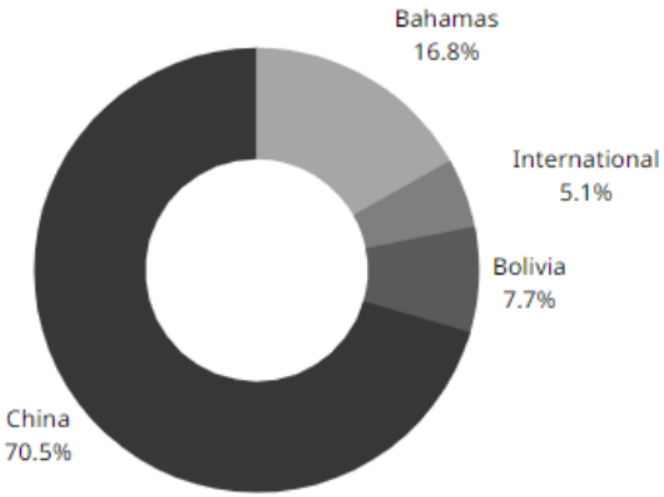


Hình 4. 9 Đầu ra phân cụm bằng K-means với thuộc tính quốc gia và tỷ lệ người mắc bệnh

Bảng phân tích dữ liệu:

Cluster	0	1	2	3
Tâm là Quốc Gia	Bahamas	International	Bolivia	China
Tỷ lệ che phủ toàn bộ dữ liệu	17%	5%	8%	70%

Bàng 4. 3 Bảng phân tích dữ liệu đầu ra với thuộc tính Quốc gia và tỷ lệ người chết



Hình 4. 10 Biểu đồ tỷ lệ các cụm theo thuộc tính quốc gia và người chết trên toàn bộ dữ liệu.

KẾT LUẬN

Sau thời gian thực hiện, chúng em đã thực hiện được một số kết quả sau:

- Tìm hiểu được về khai phá dữ liệu
- Vai trò của khai phá dữ liệu
- Tìm hiểu về thuật toán K-Means
- Tìm hiểu về K-means giải quyết bài toán phân cụm người mắc bệnh và chết trên từng quốc gia

Chúng em đã tìm hiểu lý thuyết xác suất đến thuật toán K-means . Tuy độ chính xác còn chưa cao do bản chất của phương pháp cũng như tập dữ liệu chưa đủ lớn mong thầy cô giúp đỡ để bài toán của chúng em được hoàn thiện hơn.

TÀI LIỆU THAM KHẢO

[1]. Các tài liệu tham khảo của thầy Vũ Văn Định

[2]. https://tecktalk.vn

[3]. https://machinelearningcoban.com

[4]. https://tailieu.vn

[5]. https://bigdatauni.com