BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
PHÂN HIỆU TẠI TP. HỒ CHÍ MINH
BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN TỐT NGHIỆP

ĐỀ TÀI: NGHIÊN CỨU THUẬT TOÁN K-MEANS VÀ ỨNG DỤNG PHÂN CỤM SINH VIÊN BỘ MÔN CÔNG NGHỆ THÔNG TIN

Giảng viên hướng dẫn: TRẦN PHONG NHÃ

Sinh viên thực hiện: NGUYỄN VŨ THÁI

Lớp : CÔNG NGHỆ THÔNG TIN

Khoá : 57

Tp. Hồ Chí Minh, tháng 08 năm 2020

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH

CỘNG HÒA XÃ HỘI CHỦ NGHĨAVIỆT NAM
Độc lập – Tự do – Hạnh phúc

NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP

BỘ MÔN: CÔNG NGHỆ THÔNG TIN

-------***-------

Mã sinh viên:	5751071039	**Họ tên sinh viên:**	Nguyễn Vũ Thái
Khóa:	57	**Lớp:**	CQ.57.CNTT

❖ **Tên đề tài**

NGHIÊN CỨU THUẬT TOÁN K-MEANS VÀ ỨNG DỤNG PHÂN CỤM SINH VIÊN BỘ MÔN CÔNG NGHỆ THÔNG TIN.

❖ **Mục đích, yêu cầu**

- Mục đích:
  - Hiểu rõ về phân cụm, thuật toán K-Means và cài đặt thuật toán vào hệ thống.
  - Xây dựng ứng dụng hỗ trợ giúp phân loại sinh viên nhằm nắm bắt được tình hình học tập sau đó đưa ra hướng giải quyết hoặc hướng phát triển phù hợp.
- Yêu cầu:
  - Nghiên cứu về phân cụm dữ liệu và thuật toán K-Means.
  - Đánh giá thuật toán.
  - Áp dụng và đánh giá thuật toán vào bài toán phân cụm sinh viên.

❖ **Nội dung và phạm vi đề tài**

- Nội dung đề tài:
  - Tổng quan về trí tuệ nhân tạo và Machine Learning.
  - Tổng quan về khai phá dữ liệu.
  - Tổng quan về C# và .NET Framework, DevExpress.
  - Phân cụm dữ liệu và thuật toán K-Means.
  - Áp dụng và cài đặt thuật toán vào bài toán phân loại sinh viên dựa vào điểm trung bình.
- Phạm vi đề tài:

Nghiên cứu thuật toán K-Means và áp dụng vào phân cụm sinh viên trên cơ sở dữ liệu của trường đại học.

❖ **Công nghệ, công cụ và ngôn ngữ lập trình**

- Công cụ lập trình: Visual Studio 2019.
- Công nghệ sử dụng: .NET Framework, DevExpress.
- Ngôn ngữ lập trình: C#.

❖ **Các kết quả chính dự kiến sẽ đạt được**

- Quyển báo cáo đề tài tốt nghiệp
- Hiểu được thuật toán K-Means.
- Áp dụng thuật toán vào bài toán phân cụm điểm sinh viên Bộ môn Công nghệ thông tin.

❖ **Giáo viên và cán bộ hướng dẫn**

Họ tên: TRẦN PHONG NHÃ

Đơn vị công tác: Bộ môn Công nghệ Thông tin – Trường Đại học Giao thông Vận tải Phân hiệu tại thành phố Hồ Chí Minh.

Điện thoại: 0906761014 Email: tpnha@utc2.edu.vn

Ngày … tháng … năm 2020 **Đã giao nhiệm vụ TKTN**

Trưởng BM Công nghệ Thông tin **Giảng viên hướng dẫn**

ThS. Trần Phong Nhã

Đã nhận nghiệm vụ TKTN

Sinh viên: Nguyễn Vũ Thái Ký tên:

Điện thoại: 0961509754 Email: nvthai1602@gmail.com

LỜI CẢM ƠN

Lời nói đầu tiên, em xin gửi tới Quý Thầy Cô Bộ môn Công Nghệ Thông Tin – Trường Đại học Giao thông Vận tải Phân hiệu tại thành phố Hồ Chí Minh lời chúc sức khỏe và lòng biết ơn sâu sắc, chân thành nhất.

Em xin chân thành gửi lời cảm ơn đến Quý Thầy Cô đã giúp đỡ cũng như tạo điều kiện cho em hoàn thành đồ án với đề tài "Nghiên cứu thuật toán K-Means và ứng dụng trong phân cụm sinh viên Bộ môn Công nghệ thông tin". Đặc biệt, em xin chân thành cảm ơn thầy Trần Phong Nhã, người đã tận tình giúp đỡ, hướng dẫn, cung cấp cho em những kiến thức, kỹ năng cơ bản cần có để nghiên cứu và hoàn thành đề tài này.

Mặc dù đã cố gắng trong quá trình nghiên cứu, do kiến thức còn hạn chế nên vẫn còn nhiều thiếu sót. Vì vậy, em rất mong nhận được sự đóng góp ý của Quý Thầy Cô giảng viên bộ môn để đề tài của em được hoàn thiện hơn.

Lời sau cùng, em kính chúc Quý Thầy Cô Bộ môn Công Nghệ Thông Tin và đặc biệt là thầy Trần Phong Nhã thật dồi dào sức khỏe, gặt hái được nhiều thành công trong cuộc sống cũng như trong sự nghiệp giảng dạy.

Em xin chân thành cảm ơn!

Tp. Hồ Chí Minh, ngày … tháng … năm 2020

Sinh viên thực hiện

Nguyễn Vũ Thái

NHẬN XÉT CỦA GIÁO VIÊN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp. Hồ Chí Minh, ngày … tháng … năm 2020

Giáo viên hướng dẫn

Trần Phong Nhã

MỤC LỤC

DANH MỤC THUẬT NGỮ

STT	THUẬT NGỮ TIẾNG ANH	Ý NGHĨA TIẾNG VIỆT	TỪ VIẾT TẮT	GHI CHÚ
1	Information Technology	Công Nghệ Thông Tin	CNTT	
2	Stupid Pointless Annoying Messages	Thư rác	SPAM	
3	Artificial Intelligence	Trí tuệ nhân tạo	AI	
4	Machine Learning	Học máy	ML	
5	Deep Learning	Học sâu	DL	
6	Knowledge Discovery in Database	Khai phá tri thức	KDD	
7	Application Programming Interface	Giao diện lập trình ứng dụng	API	
8	Database	Cơ sở dữ liệu	CSDL	
9	Data Mining	Khai phá dữ liệu	KPDL	
10	Structured Query Language	Ngôn ngữ truy vấn dữ liệu có cấu trúc	SQL	
11	Association rules	Luật kết hợp		
12	Classification	Phân lớp		
13	Clustering	Phân cụm		
14	Regression	Hồi quy		

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH

MỞ ĐẦU

1. Lý do chọn đề tài

Trong những năm gần đây, sự phát triển không ngừng của ngành công nghệ thông tin và các lĩnh vực liên quan, dẫn đến hệ quả là khối lượng thông tin lưu trữ ngày càng lớn. Sự bùng nổ về dữ liệu dẫn đến yêu cầu cấp thiết là cần có những kỹ thuật và công cụ mới để tự động chuyển đổi dữ liệu khổng lồ kia thành các tri thức có ích. Từ đó, các kỹ thuật khai phá dữ liệu đã trở thành một lĩnh vực không thể thiếu của nền công nghệ thông tin thế giới hiện nay nói chung và Việt Nam nói riêng. Khai phá dữ liệu đang được áp dụng một cách rộng rãi trong nhiều lĩnh vực kinh doanh và đời sống khác nhau như: marketing, tài chính, ngân hàng và bảo hiểm, khoa học, y tế…

Ngành giáo dục nói chung và các trường đại học nói riêng, với lượng dữ liệu điểm khá lớn nên:

- Việc nhìn tổng quát về kết quả học tập của các học sinh ở một khối hay toàn trường trong một học kỳ sẽ mất nhiều thời gian để thống kê, tính toán và có thể xảy ra sai sót.
- Để có thể dễ dàng hơn trong việc quản lý những sinh viên có kết quả học tập chưa tốt qua đó đưa ra những giải pháp cho sinh viên có thể cải thiện việc học tập.
- Dựa vào điểm của các môn học của sinh viên từ các kì học trước qua đó giúp cho sinh viên năm ba có thể lựa chọn nghành học phù hợp với bản thân và dễ dàng kiếm việc sau khi ra trường.
- Dựa vào kết quả phân cụm có thể khảo sát mở lớp cho các môn học có học phần tiên quyết đã học trước đó của sinh viên một cách dễ dàng và chính xác.

Với tầm quan trọng của giáo dục nhất là trong thời đại của cuộc cách mạng khoa học – công nghệ hiện đại cùng sự phát triển không ngừng của Trí tuệ nhân tạo hiện nay và với những lý do trên em xin chọn đề tài "NGHIÊN CỨU THUẬT TOÁN K-MEANS VÀ ỨNG DỤNG PHÂN CỤM SINH VIÊN BỘ MÔN CÔNG NGHỆ THÔNG TIN" làm đề tài đồ án tốt nghiệp.

2. Mục đích nghiên cứu

- Nghiên cứu các vấn đề cơ bản về phân cụm dữ liệu, các thuật toán liên quan đến phân cụm. Phân tích và triển khai áp dụng thuật toán K-Means.
- Phân tích thực trạng và nhu cầu ứng dụng công nghệ thông tin vào xử lý dữ liệu điểm trong trường đại học. Đề ra giải pháp ứng dụng công nghệ thông tin vào việc phân cụm sinh viên dựa vào dữ liệu điểm.
- Cài đặt và đánh giá thuật toán K-Means.
- Áp dụng cơ sở lý thuyết nền tảng để xây dựng và triển khai ứng dụng.

3. Đối tượng và phạm vi nghiên cứu

- Tìm hiểu thuật toán K-Means để phân cụm sinh viên dựa trên dữ liệu của trường đại học đã có.
- Cài đặt và thử nghiệm với dữ liệu của trường đại học.

4. Phương pháp nghiên cứu

Phương pháp nghiên cứu lý thuyết:

- Phân tích và tổng hợp các tài liệu về khai phá dữ liệu, sử dụng thuật toán K-Means trong phân cụm dữ liệu.

Phương pháp thực nghiệm:

- Phân tích, tìm ra giải pháp và vận dụng lý thuyết, các thuật toán có liên quan để trợ giúp việc lập trình, xây dựng ứng dụng.
- Ứng dụng kết hợp kỹ thuật phân cụm dữ liệu để phân cụm sinh viên.

5. Cấu trúc báo cáo đồ án tốt nghiệp

Cấu trúc đồ án được chia thành các chương như sau:

Mở đầu: Giới thiệu tổng quan về đề tài đồ án tốt nghiệp.

Chương 1: Cơ sở lý thuyết

- Tổng quan về Machine Learning.
- Tổng quan về khám phá tri thức, khai phá dữ liệu.
- Tổng quan về phân cụm dữ liệu và các thuật toán liên quan.
- Tổng quan ngôn ngữ C# và .Net Framework, DevExpress.

Chương 2: Phân tích thuật toán K-Means

- Giới thiệu và tiến hành phân tích thuật toán K-Means.
- Ví dụ minh họa và nhận xét về thuật toán.

Chương 3: Thử nghiệm và đánh giá

- Cài đặt thuật toán.
- Đánh giá thuật toán đối với bài toán phân cụm sinh viên.
- Đưa ra kết quả đạt được, những thứ còn tồn tại.
- Hướng phát triển về thuật toán cho ứng dụng.

Kết luận và kiến nghị

Tài liệu tham khảo

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

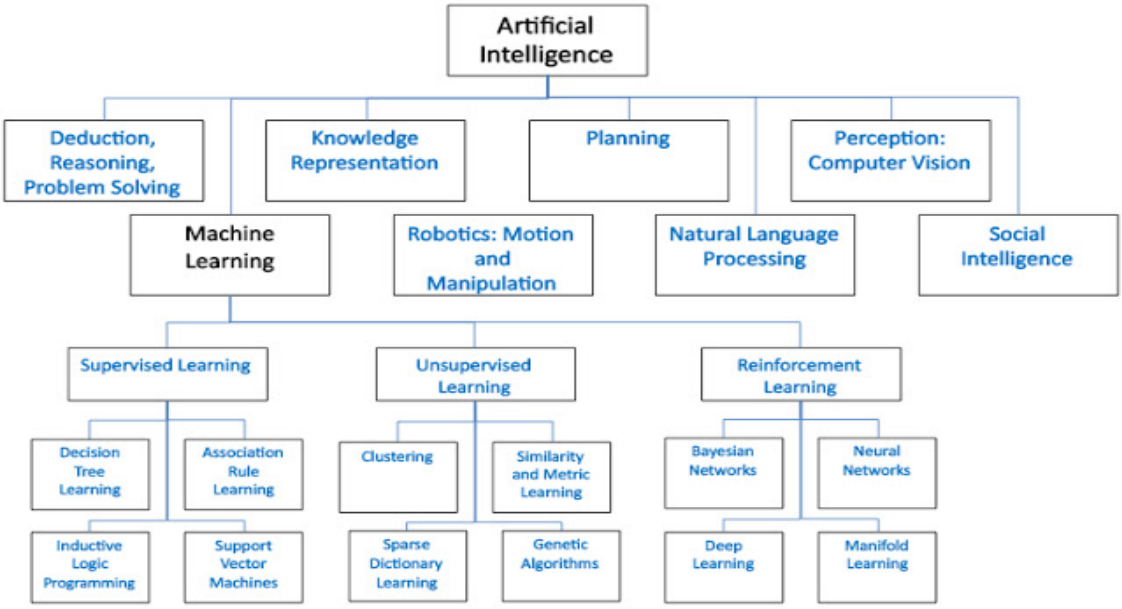
1.1. Tổng quan về Machine Learning

1.1.1. Giới thiệu về Machine Learning

Những năm gần đây, AI hay Trí tuệ nhân tạo đã đạt được nhiều thành tựu rực rỡ trong nhiều lĩnh vực: Thị giác máy tính (computer vision), xử lý ngôn ngữ tự nhiên (natural language processing), hệ thống khuyến nghị (recommendation system). Với tốc độ phát triển vô cùng nhanh chóng nhờ vào những tiến bộ trong ngành khoa học dữ liệu (Data Science) và những siêu máy tính có tốc độ tính toán cực kì nhanh chóng, AI đã và đang giúp cho cuộc sống của con người ngày một tốt đẹp hơn [3].

Xe tự lái của Google và Tesla, hệ thống tự nhận diện khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix… chỉ là một vài trong rất nhiều những ứng dụng của trí tuệ nhân tạo.

Và Machine Learning chính là một tập con trong trí tuệ nhân tạo. Nó là một lĩnh vực nhỏ trong ngành khoa học máy tính, có khả năng tự học hỏi dựa tên dữ liệu được đưa vào mà không cần phải lập trình cụ thể [10].



Hình 1.1 Các nhánh của Machine Learning trong Trí tuệ nhân tạo

a) Định nghĩa Machine Learning

Machine Learning là một thuật toán có khả năng học tập từ dữ liệu, có nghĩa là chương trình máy tính sẽ học hỏi từ kinh nghiệm E từ các tác vụ T, với kết quả được đo bằng hiệu suất P. Nếu hiệu suất của nó áp dụng trên tác vụ T khi được đánh giá bởi P và cải thiện theo kinh nghiệm E [12].

Ví dụ 1: Giả sử như bạn muốn máy tính xác định một tin nhắn có phải là SPAM hay không?

- Tác vụ T: Xác định 1 tin nhắn có phải SPAM hay không?
- Kinh nghiệm E: Xem lại những tin nhắn đánh dấu là SPAM xem có những đặc tính gì để có thể xác định nó là SPAM.
- Độ đo P: Là phần trăm số tin nhắn SPAM được phân loại đúng.

b) Sự hữu ích của Machine Learning

Từ lâu đã có nhiều thuật toán ML nổi tiếng nhưng khả năng tự động áp dụng các phép tính phức tạp vào Big Data, lặp đi lặp lại với tốc độ nhanh hơn, chỉ mới phát triển gần đây.

Các ứng dụng của ML đã trở nên quá quen thuộc như:

- Xe tự lái, giảm thiểu tai nạn của Google? Chính là bản chất của ML.
- Các ưu đãi Recommendation Online như của Amazon & Netflix? Ứng dụng của Machine Learning trong cuộc sống hằng ngày.
- Các ưu đãi Recommendation Online như của Amazon & Netflix? Ứng dụng của Machine Learning trong cuộc sống hằng ngày.
- Nhận diện lừa đảo? Một trong những nhu cầu sử dụng hiển nhiên ngày nay.

c) Đối tượng sử dụng

Hầu hết mọi ngành công nghiệp đang làm việc với hàm lượng lớn dữ liệu đều nhận ra tầm quan trọng của công nghệ ML. Những cái nhìn sâu sắc từ nguồn dữ liệu này, sẽ giúp các tổ chức vận hành hiệu quả hơn hoặc tạo được lợi thế cạnh tranh so với các đối thủ.

- Các dịch vụ tài chính

Ngân hàng và những doanh nghiệp hoạt động trong lĩnh vực tài chính sử dụng công nghệ ML với 2 mục đích chính: xác định insights trong dữ liệu và ngăn chặn

lừa đảo. Insights sẽ biết được các cơ hội đầu tư hoặc thông báo đến nhà đầu tư thời điểm giao dịch hợp lý. Khai phá dữ liệu cũng có thể tìm được những khách hàng đang có hồ sơ rủi ro cao hoặc sử dụng giám sát mạng để chỉ rõ những tín hiệu lừa đảo.

- Chính phủ

Các tổ chức chính phủ hoạt động về an ninh cộng đồng hoặc tiện ích xã hội sở hữu rất nhiều nguồn dữ liệu có thể khai thác insights. Ví dụ, khi phân tích dữ liệu cảm biến, chính phủ sẽ tăng mức độ hiệu quả của dịch vụ và tiết kiệm chi phí. ML còn hỗ trợ phát hiện gian lận và giảm thiểu khả năng trộm cắp danh tính.

- Chăm sóc sức khỏe

ML là 1 xu hướng phát triển nhanh chóng trong ngành chăm sóc sức khỏe, nhờ vào sự ra đời của các thiết bị và máy cảm ứng đeo được sử dụng dữ liệu để đánh giá tình hình sức khỏe của bệnh nhân trong thời gian thực. Công nghệ ML còn giúp các chuyên gia y tế xác định những xu hướng hoặc tín hiệu để cải thiện khả năng điều trị, chẩn đoán bệnh.

- Tiếp thị và bán hàng

Dựa trên hành vi mua hàng trước đây, các trang website sử dụng ML phân tích lịch sử mua hàng, từ đó giới thiệu những vật dụng mà bạn có thể sẽ quan tâm và yêu thích. Khả năng tiếp nhận dữ liệu, phân tích và sử dụng những dữ liệu đó để cá nhân hóa trải nghiệm mua sắm hoặc thực hiện chiến dịch Marketing chính là tương tai của ngành bán lẻ [10, 11].

1.1.2. Phân nhóm các thuật toán Machine Learning

Các thuật toán ML thường được chia làm 4 nhóm:

- Supervise learning (Học có giám sát).
- Unsupervised learning (Học không giám sát).
- Semi-supervised lerning (Học bán giám sát).
- Reinforcement learning (Học cũng cố).

a. Supervised Learning (Học có giám sát).

Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới dựa trên các cặp (*input, outcome*) đã biết từ trước. Cặp dữ liệu này còn được gọi là

(*data, label*), tức (*dữ liệu, nhãn*). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

Ví dụ: trong nhận dạng chữ viết tay, ta có ảnh của hàng nghìn ví dụ của mỗi chữ số được viết bởi nhiều người khác nhau. Chúng ta đưa các bức ảnh này vào trong một thuật toán và chỉ cho nó biết mỗi bức ảnh tương ứng với chữ số nào. Sau khi thuật toán tạo ra một mô hình, tức một hàm số mà đầu vào là một bức ảnh và đầu ra là một chữ số, khi nhận được một bức ảnh mới mà mô hình chưa nhìn thấy bao giờ, nó sẽ dự đoán bức ảnh đó chứa chữ số nào.

Ví dụ này khá giống với cách học của con người khi còn nhỏ. Ta đưa bảng chữ cái cho một đứa trẻ và chỉ cho chúng đây là chữ A, đây là chữ B. Sau một vài lần được dạy thì trẻ có thể nhận biết được đâu là chữ A, đâu là chữ B trong một cuốn sách mà chúng chưa nhìn thấy bao giờ [10, 12].

Thuật toán supervised learning còn được tiếp tục chia nhỏ ra thành hai loại chính:

Classification (Phân loại)

Một bài toán được gọi là *classification* nếu các *label* của *input data* được chia thành một số hữu hạn nhóm.

Ví dụ: Gmail xác định xem một email có phải là spam hay không; các hãng tín dụng xác định xem một khách hàng có khả năng thanh toán nợ hay không. Ba ví dụ phía trên được chia vào loại này.

Regression (Hồi quy)

Nếu *label* không được chia thành các nhóm mà là một giá trị thực cụ thể.

Ví dụ: một căn nhà rộng x m2, có y phòng ngủ và cách trung tâm thành phố z km sẽ có giá là bao nhiêu?

b. Unsupervised Learning (Học không giám sát).

Trong thuật toán này, chúng ta không biết được *outcome* hay *nhãn* mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm hoặc giảm số chiều của dữ liệu để thuận tiện trong việc lưu trữ và tính toán. Một cách toán học, Unsupervised learning là khi chúng ta chỉ có dữ liệu vào X mà không biết *nhãn* Y tương ứng.

Những thuật toán loại này được gọi là Unsupervised learning vì không giống như Supervised learning, chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào. Giống như khi ta học, không có thầy cô giáo nào chỉ cho ta biết đó là chữ A hay chữ B. Cụm *không giám sát* được đặt tên theo nghĩa này.

Các bài toán Unsupervised learning được tiếp tục chia nhỏ thành hai loại:

Clustering (Phân nhóm)

Một bài toán phân nhóm toàn bộ dữ liệu XX thành các nhóm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm.

Ví dụ: phân nhóm khách hàng dựa trên hành vi mua hàng. Điều này cũng giống như việc ta đưa cho một đứa trẻ rất nhiều mảnh ghép với các hình thù và màu sắc khác nhau, ví dụ tam giác, vuông, tròn với màu xanh và đỏ, sau đó yêu cầu trẻ phân chúng thành từng nhóm. Mặc dù không cho trẻ biết mảnh nào tương ứng với hình nào hoặc màu nào, nhiều khả năng chúng vẫn có thể phân loại các mảnh ghép theo màu hoặc hình dạng.

Association rules (Luật kết hợp)

Là bài toán khi chúng ta muốn khám phá ra một quy luật dựa trên nhiều dữ liệu cho trước.

Ví dụ: những khách hàng nam mua quần áo thường có xu hướng mua thêm đồng hồ hoặc thắt lưng; những khán giả xem phim Spider Man thường có xu hướng xem thêm phim Bat Man, dựa vào đó tạo ra một hệ thống gợi ý khách hàng, thúc đẩy nhu cầu mua sắm.

c. Semi-Supervised Learning (Học bán giám sát).

Các bài toán khi chúng ta có một lượng lớn dữ liệu XX nhưng chỉ một phần trong chúng được gán nhãn được gọi là Semi-Supervised Learning. Những bài toán thuộc nhóm này nằm giữa hai nhóm được nêu bên trên.

Một ví dụ điển hình của nhóm này là chỉ có một phần ảnh hoặc văn bản được gán nhãn (ví dụ bức ảnh về người, động vật hoặc các văn bản khoa học, chính trị) và phần lớn các bức ảnh, văn bản khác chưa được gán nhãn được thu thập từ internet. Thực tế cho thấy rất nhiều các bài toán ML thuộc vào nhóm này vì việc thu thập dữ

liệu có nhãn tốn rất nhiều thời gian và có chi phí cao. Rất nhiều loại dữ liệu thậm chí cần phải có chuyên gia mới gán nhãn được (ảnh y học). Ngược lại, dữ liệu chưa có nhãn có thể được thu thập với chi phí thấp từ internet [10].

d. Reinforcement Learning (Học củng cố)

Reinforcement learning là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao. Hiện tại, Reinforcement Learning chủ yếu được áp dụng vào Lý Thuyết Trò Chơi, các thuật toán cần xác định nước đi tiếp theo để đạt được điểm số cao nhất.

Ví dụ: AlphaGo gần đây nổi tiếng với việc chơi cờ vây thắng cả con người. Cờ vây được xem là có độ phức tạp cực kỳ cao với tổng số nước đi là xấp xỉ 1076110761, so với cờ vua là 1012010120 và tổng số nguyên tử trong toàn vũ trụ là khoảng 10801080.

Vì vậy, thuật toán phải chọn ra 1 nước đi tối ưu trong số hàng nhiều tỉ tỉ lựa chọn, và tất nhiên, không thể áp dụng thuật toán tương tự như IBM Deep.

Về cơ bản, AlphaGo bao gồm các thuật toán thuộc cả Supervised learning và Reinforcement Learning. Trong phần Supervised Learning, dữ liệu từ các ván cờ do con người chơi với nhau được đưa vào để huấn luyện. Tuy nhiên, mục đích cuối cùng của AlphaGo không phải là chơi như con người mà phải thậm chí thắng cả con người.

Vì vậy, sau khi *học* xong các ván cờ của con người, AlphaGo tự chơi với chính nó với hàng triệu ván chơi để tìm ra các nước đi mới tối ưu hơn. Thuật toán trong phần tự chơi này được xếp vào loại Reinforcement learning [10].

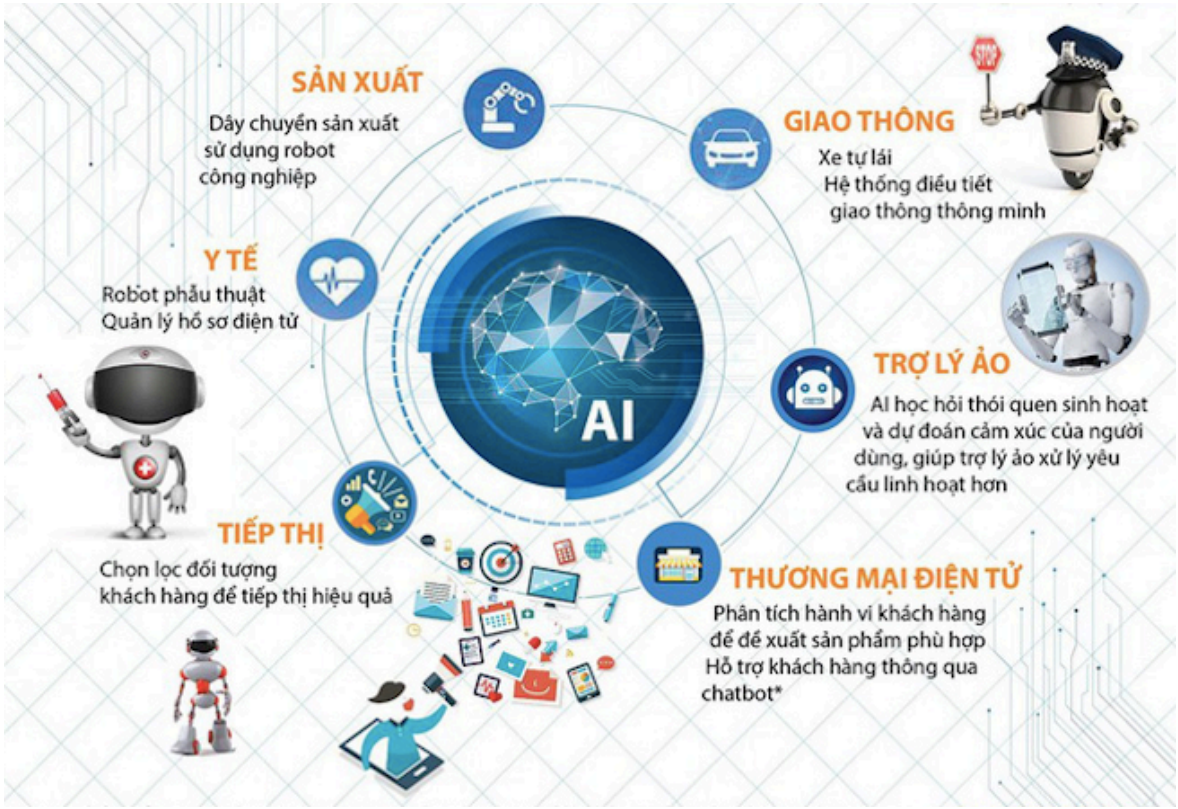
1.1.3. Các ứng dụng của Machine Learning

Có rất nhiều ứng dụng mang tính thực tế cao của máy học mà khó có thể kể hết được. Những ứng dụng dưới đây là những ứng dụng phổ biến và được chọn lọc theo góc nhìn cá nhân.

- Nhận diện và phát hiện khuôn mặt: Nhận diện và phát hiện khuôn mặt là ứng dụng khá thú vị của máy học và được áp dụng khá nhiều vào đời sống. Tiêu biểu là tính năng phát hiện khuôn mặt ở máy chụp ảnh. Ứng dụng được phát triển thêm thành phát hiện chớp mắt, phát hiện cười....

- Xe tự lái: Xe tự lái mặc dù phát triển từ đầu thập niên 90 nhưng cho tới nay vẫn còn là vấn đề được nhiều người quan tâm. Các hãng lớn như Google, Tesla, NVDIA đang nỗ lực để tạo ra một cỗ máy có thể hoàn toàn tự động lái xe và giảm thiểu tai nạn cho con người.

- Nhận dạng giọng nói: Các trợ lý ảo như Siri, Conrtana hay Google Now là ví dụ điển hình cho nhận dạng giọng nói. Một ví dụ khác nữa là tính năng dịch thuật trực tuyến của Youtube. Với ứng dụng của DL, khả năng dịch thuật chín xác ngôn ngữ từ các video Youtube đang ngày một phát triển vượt bậc [11].



Hình 1.2 Ứng dụng của Trí tuệ nhân tạo

1.2. Khai phá tri thức và quá trình khai phá tri thức

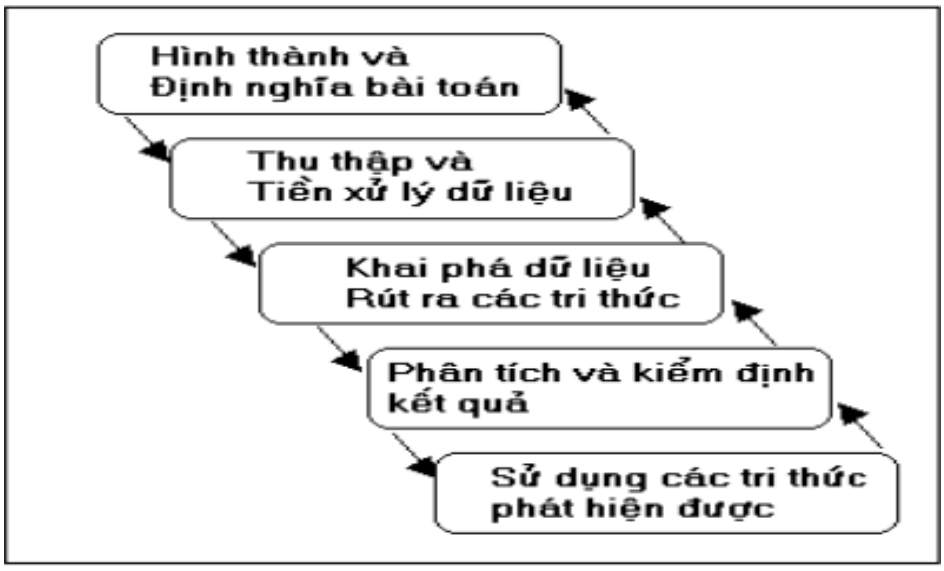
1.2.1. Khai phá tri thức

Khám phá tri thức hay phát hiện tri thức trong cơ sở dữ liệu là một quy trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: Phân tích, tổng hợp, hợp thức, khả ích và có thể hiểu được.

Khai phá dữ liệu là một bước trong quá trình khám phá tri thức, gồm các thuật toán khai thác dữ liệu chuyên dùng dưới một số qui định về hiệu quả tính toán chấp nhận được để tìm ra các mẫu hoặc các mô hình trong dữ liệu. Nói cách khác, mục tiêu của khai phá dữ liệu là tìm kiếm các mẫu hoặc mô hình tồn tại trong cơ sở dữ liệu nhưng ẩn trong khối lượng lớn dữ liệu [9, 14].

1.2.2. Quá trình khai phá tri thức

Quy trình phát hiện tri thức thường tuân theo các bước sau:



Hình 1.3 Quá trình khai phá tri thức

- **Bước thứ nhất:** Hình thành, xác định và định nghĩa bài toán. Là tìm hiểu lĩnh vực ứng dụng từ đó hình thành bài toán, xác định các nhiệm vụ cần phải hoàn thành. Bước này sẽ quyết định cho việc rút ra được các tri thức hữu ích và cho phép chọn các phương pháp khai phá dữ liệu thích hợp với mục đích ứng dụng và bản chất của dữ liệu.

- **Bước thứ hai:** Thu thập và tiền xử lý dữ liệu. Là thu thập và xử lý thô, còn được gọi là tiền xử lý dữ liệu nhằm loại bỏ nhiễu (làm sạch dữ liệu), xử lý việc thiếu dữ liệu (làm giàu dữ liệu), biến đổi dữ liệu và rút gọn dữ liệu nếu cần thiết, bước này thường chiếm nhiều thời gian nhất trong toàn bộ qui trình phát hiện tri thức. Do dữ liệu được lấy từ nhiều nguồn khác nhau, không đồng nhất, có thể gây ra các nhầm lẫn. Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và rời rạc hoá.

- **Bước thứ ba:** Khai phá dữ liệu, rút ra các tri thức. Là khai phá dữ liệu, hay nói cách khác là trích ra các mẫu hoặc/và các mô hình ẩn dưới các dữ liệu. Giai đoạn này rất quan trọng, bao gồm các công đoạn như: chức năng, nhiệm vụ và mục đích của khai phá dữ liệu, dùng phương pháp khai phá nào? Thông thường, các bài toán khai phá dữ liệu bao gồm: các bài toán mang tính mô tả - đưa ra tính chất chung nhất của dữ liệu, các bài toán dự báo - bao gồm cả việc phát hiện các suy diễn dựa trên dữ liệu hiện có. Tùy theo bài toán xác định được mà ta lựa chọn các phương pháp khai phá dữ liệu cho phù hợp

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

- **Bước thứ năm:** Sử dụng các tri thức phát hiện được. Là hiểu tri thức đã tìm được, đặc biệt là làm sáng tỏ các mô tả và dự đoán.

Các bước trên có thể lặp đi lặp lại một số lần, kết quả thu được có thể được lấy trung bình trên tất cả các lần thực hiện. Các kết quả của quá trình phát hiện tri thức có thể được đưa và ứng dụng trong các lĩnh vực khác nhau. Do các kết quả có thể là các dự đoán hoặc các mô tả nên chúng có thể được đưa vào các hệ thống hỗ trợ ra quyết định nhằm tự động hoá quá trình này [4, 14].

1.3. Tổng quan về khai phá dữ liệu

1.3.1. Khai phá dữ liệu

Khai phá dữ liệu là một giai đoạn quan trọng trong quá trình Khai phá tri thức. Về bản chất nó là giai đoạn duy nhất tìm ra được thông tin mới. Việc khai phá dữ liệu còn được coi như là việc khai phá tri thức từ dữ liệu, trích lọc tri thức, phân tích dữ liệu mẫu, khảo cứu dữ liệu [2].

Khai phá dữ liệu được định nghĩa là quá trình trích lọc các thông tin có giá trị ẩn trong lượng lớn dữ liệu được lưu trữ trong các CSDL hoặc các kho dữ liệu. Khai phá dữ liệu còn được xem là quá trình tìm kiếm, khám phá ở nhiều góc độ để tìm ra mối tương quan, các mối liên hệ dưới nhiều góc độ khác nhau nhằm tìm ra các mẫu

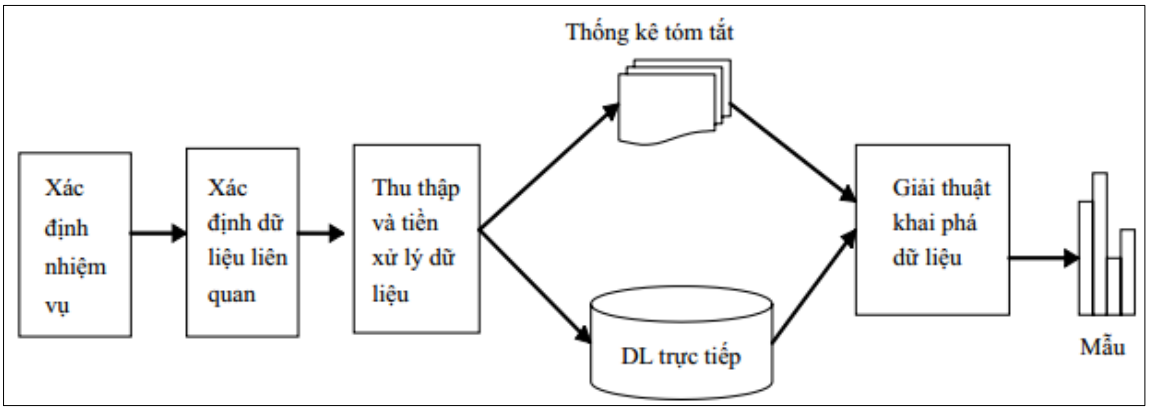
hay các mô hình tồn tại bên trong cơ sở dữ liệu đang bị che khuất. Để rút trích các mẫu, mô hình tiềm ẩn có tính "tri thức" ta phải tìm và áp dụng các phương pháp, kỹ thuật khai phá sao cho các kỹ thuật và phương pháp này phải phù hợp với tính chất, đặc trưng của dữ liệu và mục đích sử dụng. Tuy khai phá dữ liệu chỉ là một bước trong quá trình khám phá tri thức nhưng nó là bước tiên quyết, quan trọng và ảnh hưởng đến toàn bộ quá trình [5, 14].

1.3.2. Mục tiêu của khai phá dữ liệu

Qua những nội dung đã trình bày ở trên, ta có thể hiểu một cách sơ lược rằng khai phá dữ liệu là quá trình tìm kiếm thông tin hữu ích, tiềm ẩn và mang tính dự báo trong các cơ sở dữ liệu lớn. Việc khai phá dữ liệu nhằm các mục đích chính như sau:

- Khai thác những thông tin tiềm ẩn mang tính dự đoán từ những cơ sở dữ liệu lớn dựa trên các công cụ khai phá dữ liệu nhằm dự đoán những xu hướng trong tương lai nhằm giúp các đối tượng cần tri thức khai phá như: các tổ chức, doanh nghiệp, nhà nghiên cứu, … để hỗ trợ việc đưa ra quyết định kịp thời, được định hướng trên những tri thức khám phá mang lại.
- Thực hiện phân tích xử lý, tính toán dữ liệu một cách tự động cho mỗi quá trình xử lý dữ liệu để tìm ra tri thức [14].

1.3.3. Quá trình khai phá dữ liệu



Hình 1.4 Quá trình khai phá dữ liệu

- Quá trình xử lý KPDL bắt đầu bằng cách xác định chính xác vấn đề cần giải quyết.
- Sau đó sẽ xác định các dữ liệu liên quan dùng để xây dựng giải pháp.

- Bước tiếp theo là thu thập các dữ liệu có liên quan và xử lý chung thành dạng sao cho giải thuật KPDL có thể hiểu được. Về lý thuyết thì có vẻ rất đơn giản nhưng khi thực hiện thì đây thực sự là một quá trình rất khó khăn, gặp phải rất nhiều vướng mắc như: các dữ liệu phải được sao ra nhiều bản (nếu được chiết xuất vào các tệp), quản lý tập các tệp dữ liệu, phải lặp đi lặp lại nhiều lần toàn bộ quá trình (nếu mô hình dữ liệu thay đổi), …
- Bước tiếp theo là chọn thuật toán KPDL thích hợp và thực hiện việc KPDL để tìm được các mẫu (pattern) có ý nghĩa dưới dạng biểu diễn tương ứng với các ý nghĩa đó (thường được biểu diễn dưới dạng các luật xếp loại, cây quyết định, luật sản xuất, biểu thức hồi quy, …).
- Đặc điểm của mẫu phải là mới (ít nhất là đối với hệ thống đó). Độ mới có thể được đo tương ứng với độ thay đổi trong dữ liệu (bằng cách so sánh các giá trị hiện tại với các giá trị trước đó hoặc các giá trị mong muốn), hoặc bằng tri thức (mối liên hệ giữa phương pháp tìm mới và phương pháp cũ như thế nào). Thường thì độ mới của mẫu được đánh giá bằng một hàm logic hoặc một hàm đo độ mới, độ bất ngờ của mẫu. Ngoài ra, mẫu còn phải có khả năng sử dụng tiềm tàng. các mẫu này sau khi được xử lý và diễn giải phải dẫn đến những hành động có ích nào đó được đánh giá bằng một hàm lợi ích. Mẫu khai thác được phải có giá trị đối với các dữ liệu mới với độ chính xác nào đó [4, 14].

1.3.4. Các phương pháp khai phá dữ liệu

Với hai mục đích khai phá dữ liệu là Mô tả và Dự đoán, người ta thường sử dụng các phương pháp sau cho khai phá dữ liệu:

- Phân cụm (*Clustering*)
- Luật kết hợp (*Association rules*)
- Phân lớp (*Classification*)
- Hồi qui (*Regression*)
- Trực quan hóa (*Visualiztion*)
- Tổng hợp (*Summarization*)
- Mô hình ràng buộc (*Dependency modeling*)

- Biểu diễn mô hình (*Model Evaluation*)
- Phân tích sự phát triển và độ lệch (*Evolution and deviation analyst*)
- Phương pháp tìm kiếm (*Search Method*)

Có nhiều phương pháp khai phá dữ liệu được nghiên cứu ở trên, trong đó có ba phương pháp được các nhà nghiên cứu sử dụng nhiều nhất đó là: Luật kết hợp, Phân lớp dữ liệu và Phân cụm dữ liệu [14].

1.4. Tổng quan về phân cụm dữ liệu và các thuật toán liên quan

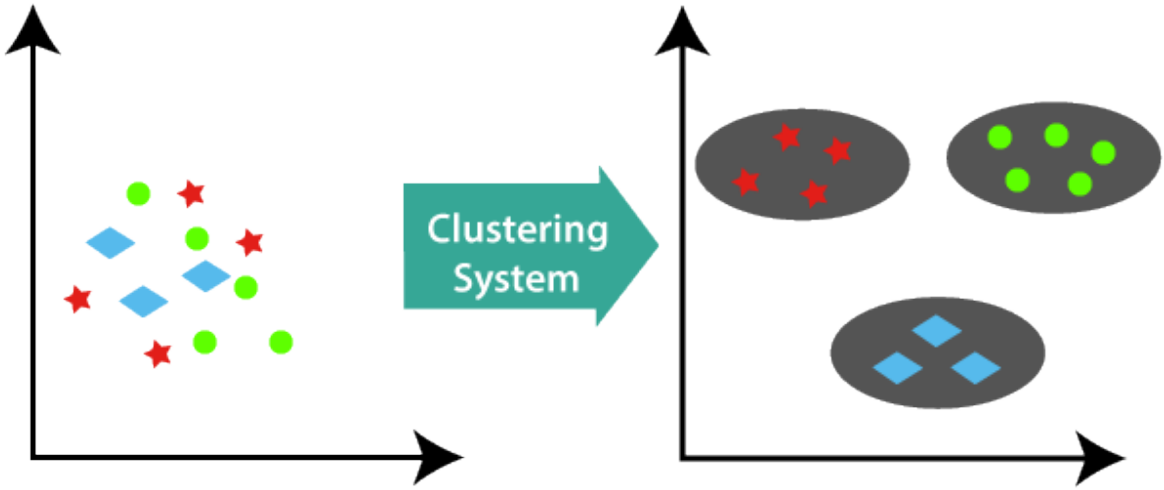
1.4.1. Giới thiệu

Phân cụm dữ liệu là một kỹ thuật trong Khai phá dữ liệu nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn và quan trọng trong tập dữ liệu lớn để từ đó cung cấp thông tin, tri thức cho việc ra quyết định.

Phân cụm dữ liệu là sự phân chia một cơ sở dữ liệu lớn thành các nhóm dữ liệu nhỏ hơn trong từng nhóm các đối tượng sẽ mang tính chất tương tự như nhau. Trong mỗi nhóm, một số chi tiết có thể không quan tâm đến để đổi lấy dữ liệu đơn giản hóa. Hay ta có thể hiểu "Phân cụm dữ liệu là quá trình tổ chức các đối tượng thành từng nhóm mà các đối tượng ở mỗi nhóm đều tương tự nhau theo một tính chất nào đó, những đối tượng không tương tự tính chất sẽ ở nhóm khác" [14].

Bài toán phân cụm là một nhánh ứng dụng chính của lĩnh vực học không giám sát không giống như phân lớp dữ liệu, phân cụm dữ liệu không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì thế, có thể coi phân cụm dữ liệu là một cách học bằng quan sát, trong khi phân lớp dữ liệu là học bằng ví dụ, ngoài ra phân cụm dữ liệu còn có thể được sử dụng như một bước tiền xử lí cho các thuật toán khai phá dữ liệu khác như là phân loại và mô tả đặc điểm, có tác dụng trong việc phát hiện ra các cụm.

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



Hình 1.5 Minh họa phân cụm dữ liệu

Trong trường hợp này, chúng ta dễ dàng xác định được ba cụm dựa vào các dữ liệu đã cho. Các tiêu chí "tương tự" để phân cụm trong trường hợp này là khoảng cách: hai hoặc nhiều đối tượng thuộc nhóm của chúng được "đóng gói" theo một khoảng cách nhất định. Điều này được gọi là phân cụm dựa trên khoảng cách.

Một kiểu khác của phân cụm dữ liệu là phân cụm dữ liệu vào khái niệm hai hay nhiều đối tượng thuộc cùng nhóm nếu có một định nghĩa khái niệm chung cho tất cả các đối tượng trong đó. Nói cách khác, đối tượng của nhóm phải phù hợp nới nhau theo miêu tả các khái niệm đã được định nghĩa, không phải theo những biện pháp đơn giản tương tự [14].

1.4.2. Các mục tiêu của phân cụm dữ liệu

Mục tiêu của phân cụm dữ liệu là để xác định các nhóm nội tại bên trong một bộ dữ liệu không có nhãn. Nhưng để có thể quyết định được cái gì tạo thành một cụm tốt. Nhưng làm thế nào để quyết định cái gì đã tạo nên một phân cụm dữ liệu tốt? Nó có thể được hiển thị rằng không có tiêu chuẩn tuyệt đối "tốt nhất" mà sẽ là độc lập với mục đích cuối cùng của phân cụm dữ liệu. Do đó, mà người sử dụng phải cung cấp tiêu chuẩn, theo cách như vậy mà kết quả của phân cụm dữ liệu sẽ phù hợp với nhu cầu của họ cần.

Theo các nghiên cứu đến thời điểm hiện nay thì chưa có một phương pháp phân cụm tổng quát nào có thể giải quyết trọn vẹn cho tất cả các dạng cấu trúc cơ sở dữ liệu. Hơn nữa, đối với các phương pháp phân cụm cần có cách thức biểu diễn cấu trúc của cơ sở dữ liệu, với mỗi cách thức biểu diễn khác nhau sẽ có tương ứng một thuật toán phân cụm phù hợp [5].

Vì vậy phân cụm dữ liệu vẫn đang là một vấn đề khó và mờ, vì phải giải quyết vấn đề cơ bản một cách trọn vẹn và phù hợp với nhiều dạng dữ liệu khác nhau, đặc biệt là đối với dữ liệu hỗn hợp đang ngày càng tăng trong các hệ quản trị cơ sở dữ liệu và đây cũng là một trong những thách thức lớn trong lĩnh vực khai phá dữ liệu.

1.4.3. Một số thuộc tính

a) Các kiểu dữ liệu trong phân cụm

Trong phân cụm, các đối tượng dữ liệu thường được diễn tả dưới dạng các thuộc tính. Các thuộc tính này là các tham số cho giải quyết vấn đề phân cụm và sự lựa chọn chúng có tác động đáng kể đến kết quả phân cụm. Phân loại các kiểu thuộc tính khác nhau là vấn đề cần giải quyết đối với hầu hết các tập dữ liệu nhằm cung cấp các phương tiện thuận lợi để nhận dạng sự khác nhau của các phần tử dữ liệu. Có hai đặc trưng để phân loại: kích thước miền và hệ đo.

Cho một CSDL D chứa n đối tượng trong không gian k chiều; x, y, z là các đối tượng thuộc D: $x = (x_1, x_2, ..., x_k)$; $y = (y_l, y_2, ..., y_k)$; $z = (z_l, z_2, ..., z_k)$ Trong đó x_i, yi, z_i với i = 1, k là các đặc trưng hoặc thuộc tính tương ứng của các đối tượng x, y, z; như vậy sẽ có các kiểu dữ liệu sau:

Phân loại kiểu dữ liệu dựa trên kích thước miền

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

Thuộc tính rời rạc: Nếu miền giá trị của nó là tập hữu hạn, đếm được (ví dụ, các thuộc tính số, ...); trường hợp đặc biệt của thuộc tính rời rạc là thuộc tính nhị phân mà miền giá trị chỉ có hai phần tử (ví dụ: Yes/no, True/False, On/Off...).

Phân loại kiểu dữ liệu dựa trên hệ đo

Thuộc tính định danh: Là dạng thuộc tính khái quát hóa của thuộc tính nhị phân, trong đó mềi n giá trị là rời rạc không phân biệt thứ tự và có nhiều hơn hai phần tử. Nếu x và y là hai đối tượng thuộc tính thì chỉ có thể xác định là $x \neq y$ hoặc $x = y$.

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

Thuộc tính khoảng: Để đo các giá trị theo xấp xỉ tuyến tính, với thuộc tính khoảng có thể xác định một thuộc tính là đứng trước hoặc đứng sau thuộc tính khác với một khoảng là bao nhiêu. Nếu $x_i > y_i$ thì có thể nói x cách y một khoảng $x_i – y_i$ tương ứng với thuộc tính thứ i.

Thuộc tính tỉ lệ: Là thuộc tính khoảng nhưng được xác định một cách tương đối so với điểm mốc đầy ý nghĩa [9].

1.4.4. Một số kỹ thuật phân cụm dữ liệu

Các kỹ thuật có rất nhiều cách tiếp cận và các ứng dụng trong thực tế, nhưng chung quy lại thì nó đều hướng đến hai mục tiêu đó là chất lượng của các cụm tìm được và tốc độ thực hiện thuật toán.

Phương pháp phân cụm theo phân hoạch

Ý tưởng chính của kỹ thuật này là phân hoạch một tập hợp dữ liệu có n phần tử cho trước thành k nhóm dữ liệu sao mỗi phần tử dữ liệu chỉ thuộc về một nhóm dữ liệu có tối thiểu ít nhất một phần tử dữ liệu. Số các cụm được thiết lập là các đặc trưng được lựa chọn trước.

Phương pháp này là tốt cho việc tìm các cụm hình cầu trong không gian Euclidean. Ngoài ra, phương pháp này cũng thuộc vào khoảng cách cơ bản giữa các điểm để lựa chọn các điểm dữ liệu nào có quan hệ là gần nhau với mỗi điểm khác và các điểm dữ liệu nào không có quan hệ hoặc có quan hệ là xa nhau so với mỗi điểm khác.

Có rất nhiều thuật toán phân hoạch như: K-Means (MacQueen 1967), K-Medoids (Kaufman và Rousseew 1987), PAM (partition Around Medoids), CLARA (Clustering Large Applications), CLARANS (Clustering Large Applications based on Randomized Search), CLASA (Clustering Large Applications based on Simulated Annealing).

Phương pháp phân cụm theo phân cấp

Phương pháp này xây dựng một phân cấp trên cơ sở các đối tượng dữ liệu đang xem xét. Nghĩa là sắp xếp một tập dữ diệu đã cho thành một cấu trúc có dạng hình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy. Cây phân cụm có thể được xây dựng theo hai phương pháp sau: hòa nhập nhóm, thường được gọi là tiếp cận từ dưới lên và phân chia nhóm, thường được gọi là tiếp cận từ trên xuống.

Các thuật toán điển hình của phương pháp phân cụm phân cấp đó là: ANGNES (Agglomerative Nesting), DIANA (Divisive Analysis), BIRCH (Balanced Interative Reducing and Clustering using Hierarchies), CURE (Clustering Using Representatives), ROCK, Chameleon, …

Phương pháp phân cụm theo mật độ

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

Chúng ta có các thuật toán phân cụm dựa trên mật độ như: DBSCAN (KDD'96), DENCLUE (KDD' 98), OPTICS, …

Phương pháp phân cụm trên lưới

Kỹ thuật phân cụm dựa trên mật độ không thích hợp với dữ liệu nhiều chiều, để giải quyết cho đòi hỏi này, người ta đã sử dụng phương pháp phân cụm dựa trên lưới. Đây là phương pháp dựa trên cấu trúc dữ liệu lưới để phân cụm dữ liệu, phương pháp này chủ yếu tập trung áp dụng cho dữ liệu không gian. Ví dụ như dữ liệu được biểu diễn dưới dạng cấu trúc hình học của đối tượng trong không gian cùng với các quan hệ, các thuộc tính, các hoạt động của chúng [9, 14].

1.4.4. Ứng dụng của phân cụm dữ liệu

Phân cụm dữ liệu được sử dụng trong một lượng lớn các ứng dụng cho một loạt các chủ để, các lĩnh vực khác nhau như phân đoạn ảnh, nhận dạng đối tượng, ký tự và các chuyên nghành cổ điển như tâm lý học, kinh doanh, … Một số ứng dụng cơ bản của phân cụm dữ liệu bao gồm:

- Thương mại: tìm kiếm nhóm các khách hàng quan trọng dựa vào các thuộc tính đặc trưng tương đồng và những đặc tả của họ trong các bản ghi mua bán của cơ sở dữ liệu.
- Sinh học: phân loại động, thực vật qua các chức năng gen tương đồng của chúng.
- Thư viện: phân loại các cụm sách có nội dung và ý nghĩa tương đồng nhau để cung cấp cho độc giả, cũng như đặt hàng với nhà cung cấp.
- Bảo hiểm: nhận dạng nhóm tham gia bảo hiểm có chi phí yêu cầu bồi thường trung bình cao, xác định gian lận trong bảo hiểm thông qua các mẫu cá biệt.
- Quy hoạch đô thị: nhận dạng các nhóm nhà theo kiểu, vị trí địa lí, giá trị... nhằm cung cấp thông tin cho quy hoạch đô thị.
- Nghiên cứu địa chấn: phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho việc nhận dạng các vùng nguy hiểm [9].

1.4.5. Các yêu cầu và những vấn đề còn tồn tại

a) Các yêu cầu về thuật toán của phân cụm dữ liệu:

Theo các nghiên cứu cho thấy hiện nay chưa có một phương pháp phân cụm tổng quát nào có thể giải quyết trọn vẹn cho tất cả các dạng cấu trúc cơ sở dữ liệu. Hơn nữa, các phương pháp phân cụm cần có cách thức biểu diễn cấu trúc của các cơ sở dữ liệu, với mỗi cách thức biểu diễn khác nhau sẽ có tương ứng thuật toán phân cụm phù hợp.

Vì vậy, phân cụm dữ liệu vẫn đang là một vấn đề khó và mờ vì phải giải quyết nhiều vấn đề cơ bản một cách trọn vẹn và phù hợp với nhiều dạng dữ liệu khác nhau,

đặc biệt là với kho dữ liệu hỗn hợp đang ngày càng tăng và đây là một trong những thách thức lớn trong lĩnh vực khai phá dữ liệu.

Vậy phân cụm dữ liệu là một thách thức trong lĩnh vực nghiên cứu vì những ứng dụng tiềm năng của chúng được đưa ra ngay chính trong những yêu cầu đặc biệt của chúng. Do đặc thù của cơ sở dữ liệu là lớn, phức tạp và có dữ liệu nhiễu nên những thuật toán phân cụm được áp dụng phải thỏa mãn những yêu cầu sau:

- Thuật toán phải xử lý và áp dụng được với cơ sở dữ liệu nhiều nhiễu, phức tạp gồm các dữ liệu không gian, dữ liệu số, kiểu nhị phân, dữ liệu định danh, hạng mục, thích nghi với kiểu dữ liệu hỗn hợp.
- Thuật toán phải có khả năng xác định được với những cụm với hình dáng bất kỳ bao gồm cả những cụm có hình dạng lồng vào nhau, cụm có hình dạng lõm, hình cầu, hình que, …
- Tối thiểu lượng tri thức cần cho xác định các tham số đầu vào. Do các giá trị đầu vào thường ảnh hưởng rất lớn đến thuật toán phân cụm và rất phức tạp để xác định các giá trị vào thích hợp đối với các cơ sở dữ liệu lớn.
- Thuật toán phải thực hiện với mọi thứ tự đầu vào dữ liệu. Nói cách khác kết quả của thuật toán nên độc lập với dữ liệu đầu vào (cùng một tập dữ liệu, khi đưa vào xử lý cho thuật toán phân cụm dữ liệu với các thứ tự vào của các đối tượng dữ liệu ở các lần thực hiện khác nhau thì không ảnh hưởng lớn đến kết quả phân cụm).
- Thuật toán không đòi hỏi tri thức về cơ sở dữ liệu từ người dùng.
- Thuật toán phải làm việc được với cơ sở dữ liệu chứa nhiều lớp đối tượng dữ liệu phức tạp và có tính chất khác nhau.
- Thuật toán phải dễ hiểu, dễ cài đặt và khả thi: Người sử dụng có thể chờ đợi những kết quả phân cụm dễ hiểu, dễ lý giải và dễ sử dụng. Nghĩa là, sự phân cụm có thể cần được giải thích ý nghĩa và ứng dụng rõ ràng. Việc nghiên cứu cách để chọn một ứng dụng đạt được mục tiêu rất quan trọng có thể gây ảnh hưởng tới sự lựa chọn các phương pháp phân cụm.

b) Những vấn đề còn tồn tại:

Có một số vấn đề với phân cụm dữ liệu:

- Kỹ thuật phân cụm hiện nay không trình bày được tất cả các yêu cầu đầy đủ và đồng thời.
- Giao dịch với số lượng lớn các mẫu và số lượng lớn các mẫu tin của dữ liệu có thể gặp vấn đề phức tạp về thời gian.
- Hiệu quả của phương pháp phụ thuộc vào định nghĩa của khoảng cách (đối với phân cụm dữ liệu dựa trên khoảng cách). Nếu không tồn tại một thước đo khoảng cách rõ ràng chúng ta phải tự xác định, một điều mà không thật sự dễ dàng chút nào, nhất là trong không gian đa chiều [9, 14].

Kết quả của thuật toán phân cụm dữ liệu có thể được giải thích theo nhiều cách khác nhau.

1.5. Tổng quan về C# và .Net Framework, DevExpress

1.5.1. Ngôn ngữ lập trình C#

C# là một ngôn ngữ lập trình đơn giản, được phát triển bởi đội ngũ kỹ sư của Microsoft vào năm 2000, trong đó người dẫn đầu là Anders Hejlsberg và Scott Wiltamuth.

C# là ngôn ngữ lập trình hiện đại, hướng đối tượng và nó được xây dựng trên nền tảng của hai ngôn ngữ mạnh nhất là C++ và Java.

C# với sự hỗ trợ mạnh mẽ của .NET Framework giúp cho việc tạo một ứng dụng Windows Forms hay WPF (Windows Presentation Foundation), . . .trở nên rất dễ dàng [1].

C# là ngôn ngữ đơn giản

Ngôn ngữ C# đơn giản vì nó dựa trên nên tảng C++ và Java. Nếu chúng ta thân thiện với C và C++ hoặc thậm chí là Java, chúng ta sẽ thấy C# khá giống về diện mạo, cú pháp, biểu thức, toán tử, và những chức năng khác được lấy trực tiếp từ ngôn ngữ C và C++, nhưng nó được đã được cải tiến để làm cho ngôn ngữ đơn giản hơn [18].

C# là ngôn ngữ hiện đại

Ngôn ngữ C# chứa những đặc tính như là xử lý ngoại lệ, thu gom bộ nhớ tự động, những kiểu dữ liệu mở rộng, và bảo mật mã nguồn đó là những đặc tính của một ngôn ngữ hiện đại.

C# là ngôn ngữ hướng đối tượng

Những đặc điểm chính của ngôn ngữ hướng đối tượng là sự đóng gói, sự kế thừa, và đa hình. C# hỗ trợ tất cả những đặc tính trên.

C# là ngôn ngữ có ít từ khóa

C# là ngôn ngữ sử dụng giới hạn những từ khóa. Phần lớn các từ khóa được sử dụng để mô tả thông tin. Chúng ta có thể nghĩ rằng một ngôn ngữ có nhiều từ khóa thì sẽ mạnh hơn. Điều này không phải sự thật, ít nhất là trong trường hợp ngôn ngữ C#, chúng ta có thể tìm thấy rằng ngôn ngữ này có thể được sử dụng để làm bất cứ nhiệm vụ nào.

C# là ngôn ngữ mạnh mẽ và mềm dẻo

Ngôn ngữ C# chỉ bị giới hạn ở chính bởi bản thân hay là trí tưởng tượng của chúng ta. Ngôn ngữ này không đặt những ràng buộc lên những việc có thể làm. C# được sử dụng cho nhiều các dự án khác nhau như là tạo ra ứng dụng xử lý văn bản, ứng dụng đồ họa, bản tính, hay thậm chí những trình biên dịch cho các ngôn ngữ khác.

C# là ngôn ngữ hướng module

Mã nguồn C# có thể được viết trong những phần được gọi là những lớp, những lớp này chứa các phương thức thành viên của nó. Những lớp và những phương thức có thể được sử dụng lại trong ứng dụng hay các chương trình khác. Bằng cách truyền các mẫu thông tin đến những lớp hay phương thức chúng ta có thể tạo ra những mã nguồn dùng lại có hiệu quả [18].

1.5.2. .Net Framework

.NET framework – trong thuật ngữ lập trình có nghĩa là một tập hợp API – là giao diện lập trình ứng dụng .NET Framework là một nền tảng lập trình và cũng là một nền tảng thực thi ứng dụng chủ yếu trên hệ điều hành Microsoft Windows được phát triển bởi Microsoft. NET Framework được thiết kế như là môi trường tích hợp để đơn giản hóa việc phát triển và thực thi các ứng dụng có thể chạy trên nền tảng Windows. Nhiều công cụ được tạo ra để xây dựng ứng dụng .Net và IDE được phát triển và hỗ trợ bởi chính Microsoft Visual Studio [15].

Sự hợp nhất thông qua các chuẩn Internet công cộng

Để có thể giao tiếp tốt với khách hàng, các đối tác trong kinh doanh được phân chia phụ thuộc vào từng khu vực địa lý khác nhau. Hoặc tất cả các ứng dụng trong tương lai và những giải pháp phát triển thì luôn cần được hỗ trợ cho các chuẩn internet được tích hợp chặt chẽ với các giao thức mà không bắt buộc người dùng phải hiểu rõ về cơ sở hạ tầng của nó.

Khả năng biến đổi được thông qua một kiến trúc ghép nối lỏng

Đa số các hệ thống lớn có tầm cỡ thế giới được xây dựng trên những kiến trúc không đồng bộ dựa trên nền thông điệp – Message based. Những dự án được xây dựng ứng dụng trên một kiến trúc như vậy thì thường rất phức tạp. .NET Framework được xây dựng để mang lại những lợi thế về năng suất kiến trúc theo lối ghép nối chặt cùng khả năng biến đổi được và vận hành nhanh chóng với lối kiến trúc ghép nối lỏng.

Hỗ trợ nhiều ngôn ngữ

Các chuyên gia thường sử dụng những ngôn ngữ khác nhau vì mỗi ngôn ngữ có những ưu điểm riêng. .NET Framework cho phép các ứng dụng được viết trong nhiều ngôn ngữ lập trình và có thể tích hợp chúng với nhau một cách chặt chẽ. Ngoài ra, khi sử dụng .NET Framework người dùng có thể tận dụng những lợi ích của kỹ năng phát triển sẵn có.

Nâng cao năng suất cho các nhà phát triển

Số lượng chuyên viên lập trình các ứng dụng không nhiều nên họ phải làm việc trong nhiều giờ mới có thể hoàn thành công việc. Khi sử dụng .NET Framework có sẵn, thì bạn có thể loại bỏ những khâu lập trình không cần thiết và chỉ tập trung vào viết các logic doanh nghiệp. Vì ưu điểm của .NET Framework là tiết kiệm thời gian thực hiện các giao dịch tự động và dễ dàng sử dụng trong việc quản lý bộ nhớ một cách tự động hiệu quả.

Bảo vệ những sự đầu tư thông qua việc bảo mật đã được cải tiến

Một trong những vấn đề quan trọng nhất liên quan đến Internet đó chính là bảo mật thông tin. Kiến trúc bảo mật của .NET Framework được thiết kế từ dưới lên nhằm

đảm bảo cho ứng dụng và dữ liệu được bảo vệ thông qua một mô hình bảo mật an toàn và tinh vi.

Tận dụng những dịch vụ của hệ điều hành

Windows cung cấp cho bất cứ một nền tảng nào số lượng đa dạng các dịch vụ có sẵn như: Truy cập dữ liệu, bảo mật tích hợp, giao diện tương tác người dùng. .NET Framework đã tận dụng lợi ích này để hướng người dùng theo các sử dụng dễ dàng nhất [13, 15].

1.5.3. DevExpress

Đối với những lập trình viên .NET thì DevEpress là một công cụ hết sức hữu dụng, cung cấp rất nhiều control trong Visual Studio. DevExpress không chỉ giúp thiết kế winform hay website đẹp hơn mà còn giúp cho việc lập trình được dễ dàng hơn, ta có thể thấy rõ nhất là trong việc tương tác với cơ sở dữ liệu.

DevExpress được ra mắt lần đầu tiên vào năm 2011 và được đông đảo lập trình viên .NET sử dụng. Từ đó đến này đã trải qua rất nhiều phiên bản với nhiều nâng cấp đáng kể. Với DevExpress ta có thể tự tạo cho mình một bộ Office riêng chỉ trong vòng một vài tiếng [17].

Thành phần của DevExpress

- WinForms Controls: Cung cấp các control cho WinForms.
- ASP.NET Controls: Cung cấp các control cho WebForms.
- WPF Controls: Cung cấp các control cho WPF.
- Silverlight Controls: Cung cấp các control cho Silverlight.
- XtraCharts: Control cung cấp các loại biểu đồ.
- XtraReports: Cung cấp các control tạo báo cáo.
- XPO: Cung cấp môi trường làm việc với database.
- XAF: Một công nghệ mới giúp việc phát triển phần mềm một cách nhanh chóng [16].

Ưu điểm

- Hạn chế xuất hiện nhiều form riêng lẻ
- Có thể tự co giãn form bên trong form chính theo kích thước của form chính thay đổi
- Cung cấp rất nhiều UI đẹp cho Winform, Web
- Hỗ trợ rất nhiều Control hữu dụng
- Giúp việc lập trình trở nên nhanh, dễ dàng hơn, dễ quản lý.

Nhược điểm

- Giá bản quyền cao
- Cài đặt nặng
- Bộ thư viện khá nặng và tốn thời gian khi load chương trình lần đầu [16].

CHƯƠNG 2: PHÂN TÍCH THUẬT TOÁN K-MEANS

2.1. Tổng quan thuật toán K-Means

Thuật toán K-Means do MacQueen giới thiệu trong tài liệu "J. Some Methods for Classification and Analysis of Multivariate Observations" năm 1967 [8].

K-Means là một thuật toán phân cụm đơn giản thuộc loại học không giám sát (tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm k-means là phân chia một bộ dữ liệu thành các cụm khác nhau.

Là thuật toán gom cụm dữ liệu theo phương pháp phân hoạch. Một trong những thuật toán đơn giản và tốt, sử dụng Heuristic hội tụ nhanh để đạt được tối ưu, nên được biết như một thuật toán hiệu quả trong việc gom cụm dữ liệu lớn.

K-Means là một thuật toán dùng trong các bài toán phân loại/ nhóm n đối tượng thành K nhóm dựa trên thuộc tính của đối tượng (k và n là số nguyên dương).

Về nguyên lý, có n đối tượng, mỗi đối tượng có m thuộc tính, ta phân chia được các đối tượng thành k nhóm dựa trên các thuộc tính của đối tượng bằng việc áp dụng thuật toán này.

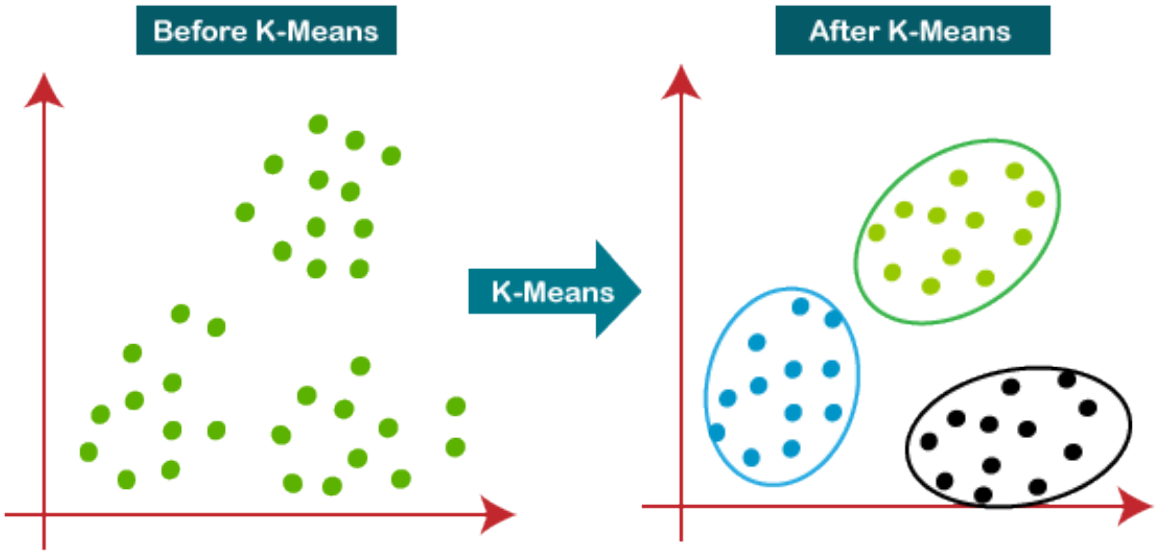
Coi mỗi thuộc tính của đối tượng (đối tượng có m thuộc tính) như một tọa độ của không gian m chiều và biểu diễn đối tượng như một điểm của không gian m chiều.

$a_i = (x_{i1}, x_{i2}, \dots x_{im})$

a_i (i = 1 … n) – đối tượng thứ i

x_{ij} (i = 1 … n, j = 1 … m) – thuộc tính thứ j của đối tượng i

Phần tử trung tâm của nhóm được xác định bằng giá trị trung bình các phần tử trong nhóm [19].



Hình 2.1 Mô phỏng phân cụm với thuật toán K-Means

2.1.1. Giới thiệu thuật toán

2.1.2. Một số khái niệm dùng trong thuật toán

Ma trận phân hoạch

Ma trận phân hoạch là ma trận biểu hiện cho sự phụ thuộc của một điểm vào một cụm nào đó.

	X_1	X_2	X_3	X_4	X_5
C_1	1	1	0	0	0
C_2	0	0	1	0	1
C_3	0	0	0	1	0

Bảng 2.1 Minh hoạ về ma trận phân hoạch

Trong ma trận phân hoạch trên, ta thấy ứng với mỗi cột chỉ có một dòng có giá trị 1 điều đó thể hiện mỗi đối tượng trong một thời điểm chỉ thuộc về một cụm.

Sau mỗi lượt gán cụm cho các đối tượng, ma trận phân hoạch lại được cập nhật. Trong suốt quá trình cập nhật, một đối tượng vẫn chỉ thuộc một cụm duy nhất.

Phần tử trọng tâm

K phần tử trọng tâm (k nhóm) ban đầu được chọn ngẫu nhiên, sau mỗi lần nhóm các đối tượng vào các cụm, trọng tâm được tính toán lại.

Xét cụm dữ liệu C_j gồm m đối tượng thuộc cụm:

$$C_j = \{r_1, r_2, r_3, \dots, r_m\} \quad (1)$$

Mỗi đối tượng có n thuộc tính:

$$r_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{in}) \ (1 \le \text{i} \le \text{m}) \quad (2)$$

Trọng tâm cụm là đối tượng m_j được xác định như sau:

$$m_j = \left(\frac{1}{m}\sum_{i=1}^{m} x_{i1}, \frac{1}{m}\sum_{i=1}^{m} x_{i2}, \dots, \frac{1}{m}\sum_{i=1}^{m} x_{in}\right) \quad (3)$$

Ví dụ: Cho cụm $C_1 = \{r_1, r_2, r_3\}$ với $r_1 = (1, 2, 1)$, $r_2 = (1, 3, 2)$,

$$r_3 = (1, 1, 3)$$

Trọng tâm cụm là:

$$m_j = \left(\frac{1+1+1}{3}, \frac{2+3+1}{3}, \frac{1+2+3}{3}\right) = (1, 2, 2)$$

Thuộc tính khoảng cách

Là khoảng cách từ một đối tượng bất kỳ đến trọng tâm nào đó. Việc xác định khoảng cách có ý nghĩa trong việc xác định đối tượng đang xét thuộc cụm nào. Một đối tượng thuộc một cụm khi khoảng cách từ điểm đó đến trọng tâm của cụm đó là nhỏ nhất.

$a_i = (x_{i1}, x_{i2}, \dots x_{im})$ i = 1 … n – đối tượng thứ i cần phân loại.

$c_j = (x_{j1}, x_{j2}, \dots x_{jm})$ j = 1 … k – phần tử trung tâm nhóm j, c_j được tính khoảng cách dựa trên công thức:

$$\text{d}_{ij} = \sqrt{\sum_{s=1}^{m} \left(x_{is} - x_{js}\right)^2} \quad (4)$$

d_{ij} – khoảng cách Euclidean từ ai đến c_j.

x_{is} – thuộc tính thứ s của đối tượng a_i.

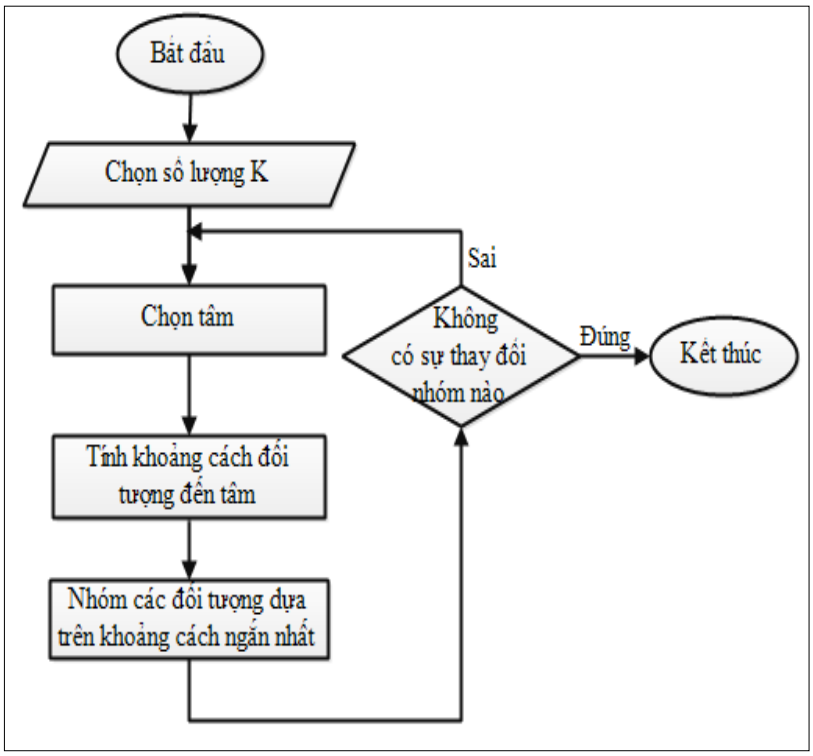
X_{js} – thuộc tính thứ s của phần tử trung tâm c_j.

2.1.3. Mô tả thuật toán

Ý tưởng

Với mục tiêu chia tập gồm n đối tượng của dữ liệu thành k cụm (k≤ n, k là số nguyên dương) sao cho các đối tượng trong cùng một cụm có khoảng cách bé, còn các đối tượng khác vùng thì có khoảng cách lớn hơn nhiều.

Thuật toán K-Means được mô tả như sau:



Hình 2.2 Sơ đồ khối thuật toán K-Means

Các bước thực hiện cơ bản của thuật toán

- ***Đầu vào:*** Số cụm k và các trọng tâm cụm $\{m_j\}$ $k_j = 1$
- ***Đầu ra:*** Các cụm C[i] ($1 \leq i \leq k$) và hàm tiêu chuẩn E đạt giá trị tối thiểu.

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

Bước 1: Khởi tạo

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

Bước 2: Tính khoảng cách

Đối với mỗi điểm Xi ($1 \leq i \leq n$), tính 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.

Bước 3: Gán (hoặc gán lại) từng điểm vào cụm có trọng tâm gần nhất với điểm đang xét nhất.

Bước 4: 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.

Điều kiện dừng:

Quay lại bước 2, 3, 4 cho đến khi các trọng tâm của cụm không thay đổi.

Thuật toán K-Means trên được chứng minh là hội tụ và có độ phức tạp tính toán là: $O\ ((3nkd)\tau T^{flop})$

Trong đó:

- n là số đối tượng dữ liệu
- k là số cụm dữ liệu
- d là số chiều
- τ là số vòng lặp
- T^{flop} là thời gian để thực hiện một phép tính cơ sở như phép nhân, chia, …

Giải thuật hội tụ: Không còn sự phân chia lại các dối tượng giữa các cụm, hay trọng tâm các cụm là không đổi. Lúc đó tổng các khoảng cách nội tại từ các đối tượng thuộc cụm đến trọng tâm cụm là cực tiểu:

$$C_j = \sum_{j=1}^{k} \sum_{r_i \in C_j} (r_i, m_j) \rightarrow \min \qquad (5)$$

Như vậy, do K-Means phân tích phân cụm đơn giản nên có thể áp dụng đối với tập dữ liệu lớn. Chất lượng phân cụm dữ liệu của thuật toán 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 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 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 [6, 7].

2.1.4. Ví dụ về thuật toán

Cho tập dữ liệu gồm 10 sinh viên, mỗi sinh viên gồm ba thuộc tính điểm X_1, X_2, X3 tương ứng với 3 môn học. Sử dụng thuật toán K-Means để phân các đối tượng thành 3 cụm (k = 3).

Mã SV	X_1	X_2	X_3
A01	9.7	7.9	7.6
A02	9.9	7.2	8
A03	6.6	4.9	5.8
A04	6.4	6.3	6.2
A05	5.7	8.3	7.3
A06	4.3	6.3	5.4
A07	8.2	7.6	6.6
A08	4	6.6	4.4
A09	6.5	6.5	7.7
A10	5.6	4.8	4.8

Bảng 2.2 Tập dữ liệu ví dụ về thuật toán K-Means

Ta có n = 10, k = 3

- Ta dùng công thức Euclidean để tính khoảng cách.
- Chia 10 điểm trên làm 3 cụm sao cho khoảng cách từ mỗi điểm đến trọng tâm của mỗi nhóm là gần nhất.

Bước 1: Khởi tạo tâm cụm:

A01 là tâm cụm C_1(9.7, 7.9, 7.6), A06 là tâm cụm C_2(4.3, 6.3, 5.4), A10 là tâm cụm C_3 (5.6, 4.8, 4.8).

Bước 2: Tính khoảng cách các đối tượng đến trọng tâm:

Khoảng cách của đối tượng A lần lượt đến C_1, C_2, C_3.

$$d(A01,\ C_1) = \sqrt{(9.7-9.7)^2 + (7.9-7.9)^2 + (7.6-7.6)^2} = 0$$

$$d(A01,\ C_2) = \sqrt{(9.7-4.3.)^2 + (7.9-6.3)^2 + (7.6-5.4)^2} = 6.0464$$

$$d(A01,\ C_3) = \sqrt{(9.7-5.6)^2 + (7.9-4.8)^2 + (7.6-4.8)^2} = 5.8532$$

Tính tương tự với các đối tượng còn lại, ta có bảng sau:

				Khoảng cách Euclidean		
Mã SV	X_1	X_2	X_3	**C_1(9.7, 7.9, 7.6)**	**C_2(4.3, 6.3, 5.4)**	**C_3(5.6, 4.8, 4.8)**
A01	9.7	7.9	7.6	**0**	6.0464	5.8532
A02	9.9	7.2	8	**0.8306**	6.2394	5.8728
A03	6.6	4.9	5.8	4.6744	2.7211	**1.4177**
A04	6.4	6.3	6.2	3.9256	2.2472	**2.2023**
A05	5.7	8.3	7.3	4.0311	**3.0935**	4.3023
A06	4.3	6.3	5.4	6.046	**0**	2.0736
A07	8.2	7.6	6.6	**1.8276**	4.2825	4.2237
A08	4	6.6	4.4	6.6648	**1.8063**	2.4413
A09	6.5	6.5	7.7	3.4943	**3.1890**	3.4799
A10	5.6	4.8	4.8	3.8532	2.7036	**0**

Bước 3: Gán (hoặc gán lại) từng điểm vào cụm có trọng tâm gần nhất với điểm đang xét nhất.

- d (A01, C_1) < d (A01, C_2) và d (A01, C_1) < d (A01, C_3)→ A01 thuộc cụm C_1
- d (A02, C_1) < d (A02, C_2) và d (A02, C_1) < d (A02, C_3)→ A02 thuộc cụm C_1
- d (A03, C_3) < d (A03, C_1) và d (A03, C_3) < d (A03, C_2)→ A03 thuộc cụm C_3
- d (A04, C_3) < d (A04, C_1) và d (A04, C_3) < d (A04, C_2)→ A04 thuộc cụm C_3
- d (A05, C_2) < d (A02, C_1) và d (A05, C_2) < d (A05, C_3)→ A05 thuộc cụm C_2
- d (A06, C_2) < d (A06, C_1) và d (A06, C_2) < d (A06, C_3)→ A06 thuộc cụm C_2
- d (A07, C_1) < d (A07, C_2) và d (A07, C_1) < d (A07, C_3)→ A07 thuộc cụm C_1
- d (A08, C_2) < d (A08, C_1) và d (A08, C_2) < d (A08, C_3)→ A08 thuộc cụm C_2
- d (A09, C_2) < d (A09, C_1) và d (A09, C_2) < d (A09, C_3)→ A09 thuộc cụm C_2

- d (A10, C_3) < d (A10, C_1) và d (A10, C_3) < d (A10, C_2)→ A10 thuộc cụm C_3

Được thể hiện qua bảng sau:

	LẦN 1		
	Cụm 1	Cụm 2	Cụm 3
Mã SV	Gần C_1	Gần C_2	Gần C_3
A01	x		
A02	x		
A03			z
A04			z
A05		y	
A06		y	
A07	x		
A08		y	
A09		y	
A10			z

⇨ Cụm C1, C2, C3 lần lượt gồm các phần tử là:

Cụm 1 gồm có (A01, A02, A07)

Cụm 2 gồm có (A05, A06, A08, A09)

Cụm 3 gồm có (A3, A4, A10)

Bước 4: Tính lại trọng tâm các cụm

$$C_1 = \left(\frac{9.7 + 9.9 + 8.2}{3} ; \frac{7.9 + 7.2 + 7.6}{3}; \frac{7.6 + 8 + 6.6}{3}\right) = (9.3 ; 7.6; 7.4)$$

$$C_2 = \left(\frac{5.7 + 4.3 + 4+6.5}{4} ; \frac{8.3 + 6.3 + 6.6 +6.5}{4}; \frac{7.3 + 5.4 + 4.4+7.7}{3}\right) = (5.1 ; 6.9; 6.2)$$

$$C_3 = \left(\frac{6.6 + 6.4 + 5.6}{3} ; \frac{4.9 + 6.3 + 4.8}{3}; \frac{5.8 + 6.2 + 4.8}{3}\right) = (6.2 ; 5.3; 5.6)$$

*Làm tương tự **bước 2** trên, ta có bảng sau:*

				Khoảng cách Euclidean		
Mã SV	X_1	X_2	X_3	**C_1(9.3, 7.6,7.4)**	**C_2(5.1, 6.9, 6.2)**	**C_3(6.2, 5.3, 5.6)**
A01	9.7	7.9	7.6	**0.5821**	4.8828	4.7806
A02	9.9	7.2	8	**0.9463**	5.1104	4.7903
A03	6.6	4.9	5.8	4.0967	2.5369	**0.6204**
A04	6.4	6.3	6.2	3.3559	1.4199	**1.1579**
A05	5.7	8.3	7.3	3.6427	**1.8524**	3.4584
A06	4.3	6.3	5.4	5.5021	**1.3081**	2.1426
A07	8.2	7.6	6.6	**1.3337**	3.1735	3.1864
A08	4	6.6	4.4	6.1378	**2.1474**	2.8094
A09	6.5	6.5	7.7	2.9803	**2.0788**	2.4226
A10	5.6	4.8	4.8	5.2782	2.5887	**1.1317**

Bước 3: Gán (hoặc gán lại) từng điểm vào cụm có trọng tâm gần nhất với điểm đang xét nhất., ta được bảng sau:

	LẦN 2		
	Cụm 1	Cụm 2	Cụm 3
Mã SV	Gần C_1	Gần C_2	Gần C_3
A01	x		
A02	x		
A03			z
A04			z
A05		y	
A06		y	
A07	x		
A08		y	
A09		y	
A10			z

⇨ Cụm C1, C2, C3 lần lượt gồm các phần tử là:

Cụm 1 gồm có (A01, A02, A07)

Cụm 2 gồm có (A05, A06, A08, A09)

Cụm 3 gồm có (A3, A4, A10)

Bước 4: Tính lại trọng tâm các cụm

	X_1	X_2	X_3
C_1	9.3	7.6	7.4
C_2	5.1	6.9	6.2
C_3	6.2	5.3	5.6

Ta có:

Cụm 1, trọng tâm C_1 (9.3; 7.6; 7.4)

Cụm 2, trọng tâm C_2 (5.1; 6.9; 7.4)

Cụm 3, trọng tâm C_3 (6.2; 5.3; 5.6)

Vì không có sự thay đổi giữa các đối tượng trong hai cụm và tọa độ tâm không đổi nên ta dừng thuật toán lại.

Vậy Cụm 1 gồm A01, A02, A07; Cụm 2 gồm A05, A06, A08, A09; Cụm 3 gồm A03, A04, A10.

Kết luận:

Cụm 1 gồm có A01, A02, A07.

Cụm 2 gồm có A05, A06, A08, A09.

Cụm 3 gồm có A03, A04, A10.

Ta thấy rằng cụm 1 gồm có 3 sinh viên A01, A02, A07 với điểm 3 môn tương ứng như sau:

Mã SV	X1	X2	X3
A01	9.7	7.9	7.6
A02	9.9	7.2	8
A07	8.2	7.6	6.6

Qua bảng trên ta thấy:

- Những sinh viên ở cụm này đều có điểm số tương đối cao qua đó Ban giám hiệu nhà trường hay Trưởng khoa có thể có những đề xuất cho những sinh viên đó học bồi dưỡng thêm để tham gia những cuộc thi của trường, thành phố hay tạo ra những sản phẩm phần mềm có giá trị cao nhằm phục vụ cho những hoạt động thực tiễn của trường.
- Giả sử như điểm của 3 môn X1, X2, X3 lần lượt là điểm của 3 môn Tin học đại cương, Lập trình hướng đối tượng, Lập trình nâng cao. Ba môn học này là những học phần chuyên sâu thể hiện khối kiến thức đặc thù của chuyên nghành "Công nghệ phần mềm". Với kết quả phân cụm trên có thể thấy những sinh viên này hoàn toàn có thể theo học chuyên nghành "Công nghệ phần mềm". Vì vậy thông qua kết quả phân cụm này Ban giám hiệu hay trưởng Khoa có thể gợi ý, đề xuất cho sinh viên chọn chuyên nghành học phù hợp với khả năng và tiến hành đăng ký những môn học phù hợp cho những học kì tiếp theo.

Ta thấy rằng cụm 3 gồm 3 sinh viên A03, A04, A10 với điểm 3 môn học tương ứng như sau:

Mã SV	X1	X2	X3
A03	6.6	4.9	5.8
A04	6.4	6.3	6.2
A10	5.6	4.8	4.8

Qua bảng trên ta thấy rằng:

- Những sinh viên ở cụm này đều có điểm số tương đôi không cao qua đó Ban giám hiệu nhà trường hay trưởng Khoa có thể có những giải pháp cụ thể nhằm cải thiện điểm số cho những sinh viên này.
- Giả sử như điểm của 3 môn X1, X2, X3 lần lượt là điểm của 3 môn Nhập môn hệ quản trị cơ sở dữ liệu, Cơ sở dữ liệu, Phân tích thiết kế hướng đối tượng. Ba môn học này là những học phần chuyên sâu thể hiện khối kiến thức đặc thù của chuyên nghành "Hệ thống thông tin". Với kết quả phân cụm trên có thể

thấy những sinh viên này nên cân nhắc có nên theo học chuyên sâu chuyên nghành "Hệ thống thông tin" hay không? Vì vậy, thông qua kết quả phân cụm Lãnh đạo nhà trường, khoa có thể đề xuất chuyên nghành phù hợp với sinh viên và những môn học phù hợp với những học kì tiếp theo.

2.2. Đặc điểm của thuật toán

Chất lượng của thuật toán K-Means phụ thuộc nhiều vào các tham số đầu vào như: số cụm K, và K vector trọng tâm khởi tạo ban đầu. Trong trường hợp các vector 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 sẽ 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 nào để chọn tham số đầu vào, giải pháp thường được sử dụng nhất là thử nghiệm các giá trị đầu vào, giải pháp thường được sử dụng nhất là thử nghiệm với giá trị đầu vào K khác nhau rồi sau đó chọn giải pháp tốt nhất.

Ưu điểm

- Thực hiện tương đối nhanh.
- K-Means phù hợp với các cụm có dạng hình cầu.
- Có khả năng mở rộng và dễ dàng sửa đổi với những dữ liệu mới.
- Bảo đảm hội tụ sau một bước lặp hữu hạn.
- Số cụm luôn ổn định (k cụm cho trước).
- Luôn có ít nhất một điểm trong cụm.
- Các cụm được tách biệt rõ ràng không có hiện tượng một đối tượng xuất hiện trong nhiều cụm dữ liệu.

Nhược điểm

- Không đảm bảo được tối ưu toàn cục và kết quả đầu ra phụ thuộc vào nhiều việc chọn k điểm khởi đầu.
- Cần xác định trước số cụm.
- Khó xác định được thực sự số cụm có trong không gian dữ liệu. Do đó có thể phải thử với các số k khác nhau.

- Khó phát hiện các loại cụm có hình dạng phức tạp, nhất là các dạng cụm không lồi.
- Nhạy cảm với nhiễu và các phần tử ngoại lai.
- Chỉ có thể áp dụng khi tính được trọng tâm.

Cải tiến thuật toán K-means

Thay vì chọn số điểm (k) làm trọng tâm, chúng ta không chọn số điểm (k) làm trọng tâm cho số cụm mà sẽ tăng số cụm từ 1 lên k cụm bằng cách đưa trung tâm cụm mới vào cụm có mức độ biến dạng lớn nhất và tính lại trọng tâm các cụm. Với thuật toán K-Means bắt đầu bằng cách chọn k cụm và chọn ngẫu nhiên k điểm làm trung tâm cụm, hoặc chọn phân hoạch ngẫu nhiên k cụm và tính trọng tâm của từng cụm này. Việc chọn ngẫu nhiên k điểm làm trung tâm cụm như đã nói ở trên có thể cho ra các kết quả khác nhau tùy vào chọn k điểm này [6].

Kết luận

Thuật toán K-Means là thuật toán điển hình trong bài toán phân cụm dữ liệu. Mặc dù có nhiều khuyết điểm, nhưng thuật toán K-Means lại thường được sử dụng để gom cụm tập dữ liệu lớn do tính toán đơn giản và Heuristic hội tụ nhanh để đạt được tối ưu nhất. Do phù hợp với không gian dữ liệu mà các cụm dạng hình cầu, nên cần loại bỏ các mẫu cá biệt trước khi chạy thuật toán.

2.3. Ứng dụng

Thuật toán K-Means thường được sử dụng để tìm ra các nhóm mà không được gắn nhãn một cách rõ ràng trong tập dữ liệu. Điều này thường có ý nghĩa trong việc xác nhận tính đúng của các giả thiết về những kiểu nhóm đang tồn tại hay chỉ ra những nhóm chưa biết trong tập dữ liệu phức tạp. Một vài ví dụ như sau:

Một công ty vận chuyển muốn mở chuỗi các trung tâm giao nhận hàng trong thành phố. Trong tình huống đó cần đối mặt với các vấn đề như sau:

- Họ cần phải phân tích để biết khu vực mà có nhiều đơn đặt hàng thường xuyên.

- Họ cần biết bao nhiêu trung tâm nên được mở để có thể đảm bảo giao nhận hiệu quả trong một khu vực.
- Họ cần tìm ra vị trí thích hợp để mở trung tâm trong các khu vực nhằm đảm bảo tối ưu khoảng cách giữa trung tâm và khách hàng của họ.

Phân tích thông tin tội phạm có liên quan tới nghiện ma túy ở Việt Nam. Nguồn dữ liệu bao gồm các loại hình phạm tội do nhiều loại thuốc khác nhau gây ra, bao gồm Heroin, Cocaine cho tới các loại gây nghiện trong toa bác sĩ, đặc biệt là với trẻ vị thành niên. Tỉ lệ phạm tội do lạm dụng thuốc có thể giảm nhờ việc xây dựng các trung tâm cai nghiện tại chỗ trong những khu vực chịu tác động lớn bởi loại hình tội phạm này. Với nguồn dữ liệu được cho, các mục tiêu khác nhau có thể được định ra [7, 8].

Ví dụ như:

- Phân loại tội phạm dựa trên nhóm tuổi.
- Phân tích dữ liệu để xác định hình thức trung tâm cai nghiện cần xây dựng. Tìm ra số lượng trung tâm cai nghiện cần xây dựng để đạt hiệu quả trong việc giảm tỉ lệ tội phạm do nghiện thuốc.

Ngoài ra còn được ứng dụng vào phân khúc thị trường, thống kê địa lý, gom nhóm hình ảnh, hoặc dùng thuật toán để tiền xử lý tạo ra dữ liệu dùng cho các phương pháp thuật toán khác.

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

3.1. Giới thiệu bài toán

Giả sử sau khi có điểm cuối học kì, cuối năm học … bạn nhìn bảng điểm của khoa Công nghệ thông tin … và muốn đưa ra một vài kết luận nào đó về tình hình học tập, bạn muốn khảo sát điểm của sinh viên để mở môn học cho phù hợp hay những bạn cần tư vấn cho sinh viên của khoa mình lựa chọn chuyên nghành phù hợp với năng lục của sinh viên. Bạn sẽ làm thế nào?

Thứ nhất, tính điểm trung bình. Đây là một trong những cách đơn giản nhất. Dựa vào điểm trung bình có thể đưa ra vài nhận xét về tình hình học tập. Giả sử học kỳ này điểm trung bình môn Tin học đại cương của lớp CQ.57.CNTT là 7.2, của lớp CQ.58.CNTT là 7.6, như vậy có thể nhận xét một cách "nóng vội" rằng học kỳ này sinh viên lớp CQ.58.CNTT học Tin học đại cương tốt hơn lớp CQ.57.CNTT và cả 2 lớp đều học khá môn Tin học đại cương.

Tất nhiên những nhận xét này có độ tin cậy và chính xác chưa cao vì có thể có những sinh viên của lớp điểm rất cao, ngược lại cũng có điểm rất thấp. Rõ ràng tính trung bình rất đơn giản và "mạnh mẽ" (từ điểm của hàng trăm sinh viên một lớp, tức là hàng trăm mẫu dữ liệu khác nhau, chuyển thành một điểm trung bình duy nhất, và có thể dựa vào điểm trung bình để nhận xét về điểm của hàng trăm sinh viên), nhưng vì thế mà nó làm mất thông tin về phân bố của dữ liệu.

Thứ hai, tính kì vọng và phương sai. Một người học lớp thống kê cơ bản sẽ biết cách tính kì vọng và phương sai của dữ liệu. Trong trường hợp này thì kì vọng chính là điểm trung bình cộng. Dựa vào kì vọng và phương sai, ta sẽ có thể đưa ra những nhận xét sâu sắc và ít "nóng vội" hơn. Chẳng hạn nếu lớp CQ.57.CNTT có điểm kì vọng môn Tin học đại cương là 7.2 và phương sai 2.5, lớp CQ.58.CNTT có điểm kì vọng 7.6 nhưng phương sai 1.5, thì có thể kết luận là nhìn chung lớp CQ.57.CNTT học Tin học đại cương tốt hơn lớp CQ.58.CNTT nhưng lớp CQ.58.CNTT học "đều" môn Tin học đại cương hơn lớp CQ.57.CNTT (nghĩa là không có chênh lệch quá lớn giữa các sinh viên)…và có thể có thêm rất nhiều kết luận khác nữa.

Thứ ba, đếm số lượng điểm trong ngưỡng nào đó. Chẳng hạn ta có thể nói: trong 100 sinh viên thì có 80 sinh viên đạt điểm trên 5.0 và 20 sinh viên dưới 5.0. Như vậy có thể hiểu khối lớp có 4/5 sinh viên trên trung bình. Một cách chi tiết hơn, ta có thể nói: trong 80 sinh viên trên trung bình thì có 20 sinh viên là trên 8.0, 20 sinh viên là từ 6.5 đến cận 8.0, và 40 sinh viên là từ 5.0 đến cận 6.0. Như vậy có thể hiểu sâu sắc hơn rằng khối lớp có 1/5 sinh viên giỏi, 2/5 sinh viên khá v.v….

Những lý do vừa nêu trên cũng đã cho chúng ta thấy được rằng việc phân tích, đánh giá kết quả học tập của sinh viên không phải là một việc đơn giản. Nó đòi hỏi giảng viên, các nhà quản lý giáo dục có một sự đầu tư, nghiên cứu, tìm tòi và sáng tạo … nhằm đưa ra được các phân tích, đánh giá đúng đắn nhất, chính xác nhất về kết quả học tập của sinh viên từ đó đề ra định hướng, hoạch định cho nhà trường trong việc: đầu tư bồi dưỡng giáo viên bộ môn còn yếu, phát hiện sinh viên giỏi để bồi dưỡng, sinh viên kém để định hướng học tập, có kế hoạch tăng giờ, tăng tiết, định hướng nghề nghiệp cho sinh viên dựa trên sở thích, định hướng ngành học phù hợp với năng lực, khảo sát mở lớp học phần phù hợp v.v…

Đầu vào:

- Tập dữ liệu gồm 250 sinh viên Khoa Công nghệ thông tin trường Đại học Giao Thông Vận Tải phân hiệu tại Thành Phố Hồ Chí Minh đã qua các bước tiền xử lý. Mỗi sinh viên gồm các thuộc tính Mã sinh viên, Họ, Tên, Tên lớp và điểm các môn học.
- Danh sách các môn học được chọn để phân cụm.
- Số cụm k cần phân cụm.

Đầu ra: Ứng dụng có khả năng phân cụm chính xác điểm các sinh viên trong tập dữ liệu và file Excel chi tiết ứng với mỗi cụm.

3.2. Tập dữ liệu sử dụng

Trong bài nghiên cứu này, em đánh giá hiệu năng của thuật toán trên tập dữ liệu thực về điểm của sinh viên khoa Công nghệ thông tin trường Đại học Giao Thông Vận Tải phân hiệu tại Thành Phố Hồ Chí Minh. Tập dữ liệu mẫu Diemsinhvien.xlsx gồm 250 mẫu dữ liệu với 14 thuộc tính mô tả thông tin và điểm của sinh viên.

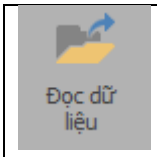
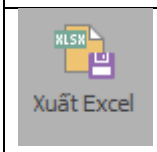
Thứ tự	Thuộc tính	Tên thuộc tính
1	f_masv	Mã sinh viên
2	f_ho	Họ
3	f_ten	Tên
4	f_lop	Tên lớp
5	f_tinhocdaicuong	Điểm Tin hoc đại cương
6	f_laptrinhhuongdoituong	Điểm Lập trình hướng đối tượng
7	f_laptrinhnangcao	Điểm Lập trình nâng cao
8	f_cautrucdulieugiaithuat	Điểm Cấu trúc dữ liệu và giải thuật
9	f_cosodulieu	Điểm Cơ sở dữ liệu
10	f_hequantricosodulieu	Điểm Hệ quản trị cơ sở dữ liệu
11	f_phantichthietkehethong	Điểm Phân tích thiết kế hệ thống
12	f_phantichthietkehdt	Điểm Phân tích thiết kế hướng đối tượng
13	f_congnghejava	Điểm Công nghệ Java
14	f_congngheoracle	Điểm Công nghệ Oracle

Bảng 3.1 Các thuộc tính của tập dữ liệu

3.3. Môi trường thử nghiệm

- Công nghệ: .NET Framework 4.6, DevExpress 19.1.
- Ngôn ngữ lập trình: C#.
- Công cụ: Visual Studio 2019, Microsoft Excel 2019.

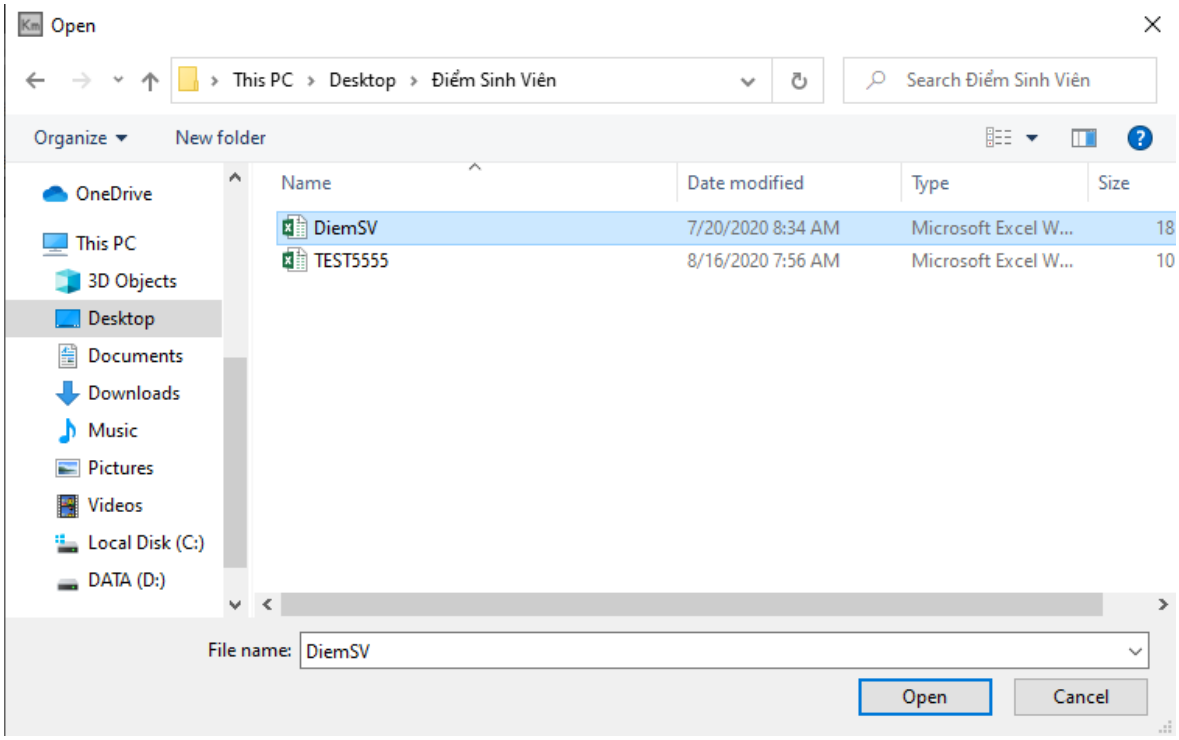
3.4. Giao diện chương trình

	Dùng mở tập tin Excel định dạng (*.xlsx) chứa danh sách sinh viên cần phân tích.
	Khởi động chức năng phân tích danh sách sinh viên thành từng cụm sau khi đã mở tập tin Excel.
	Xuất bảng thông số các cụm và danh sách chi tiết sinh viên thuộc từng cụm sang 1 tập tin Excel mới.
	Cho thấy quá trình chạy từng bước chi tiết của thuật toán
	Giới thiệu về phần mềm.
	Thoát phần mềm.

Bảng 3.2 Các chức năng của chương trình

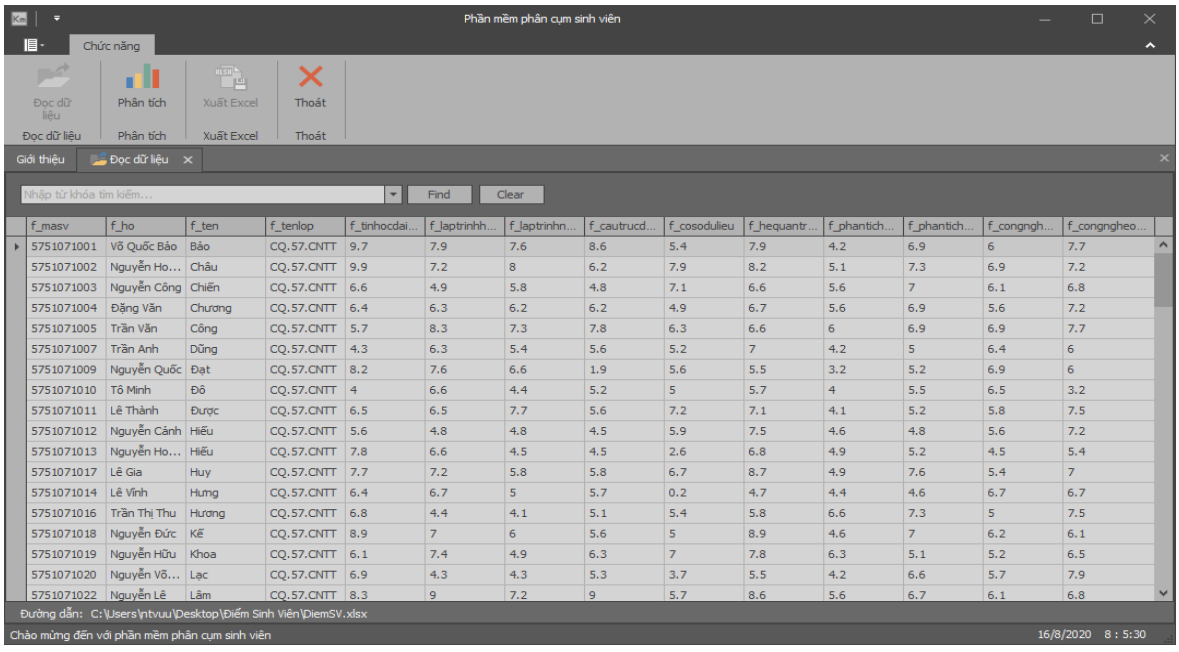
3.4.1. Đọc dữ liệu:

Để đọc một tập tin Excel để phân tích, bạn có thể click vào nút “Đọc dữ liệu” hoặc tổ hợp phím “Ctrl+O” để mở tập tin.



Hình 3. 1 Chọn file cần phân tích

Danh sách sinh viên hiển thị trên giao diện phần mềm.



Hình 3.2 Giao diện hiển thị dữ liệu

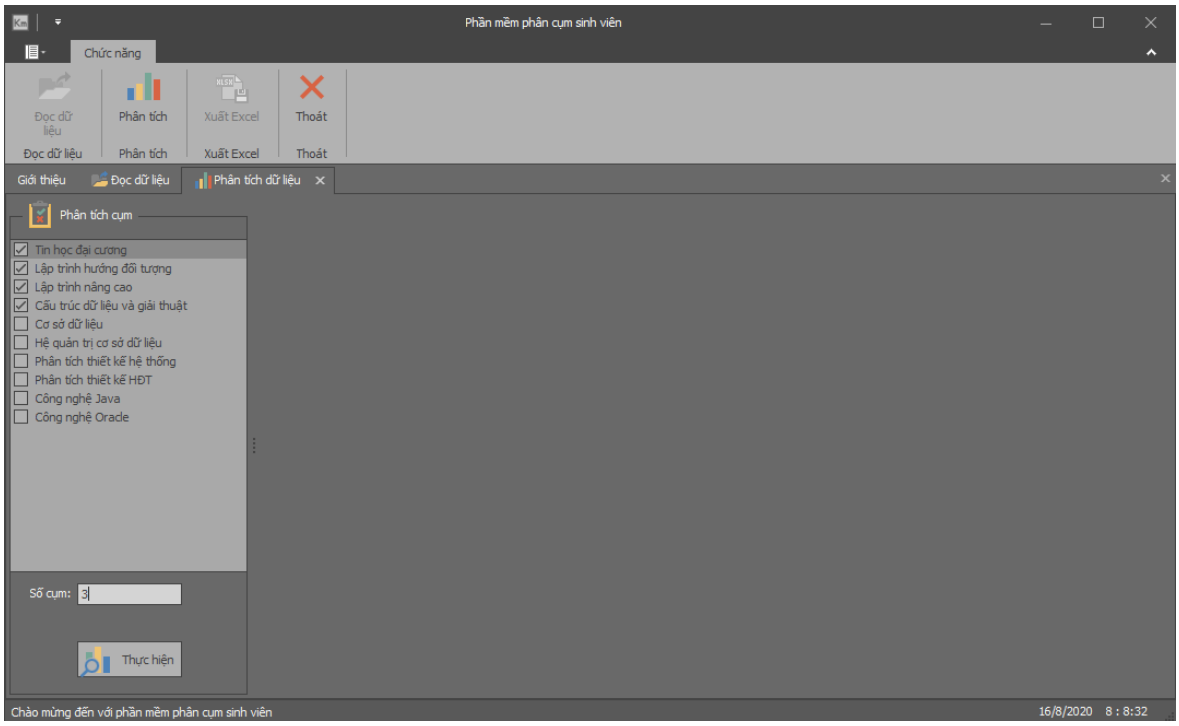
***Lưu ý:**

1. Tập tin Excel phải là định dạng từ Excel 2007 trở lên (*.xlsx)

2. Các cột thông tin phải theo một định dạng chuẩn (cột mã sinh viên là ‘f_masv’, cột tên là ‘f_ten’, cột điểm môn Tin học đại cương là ‘f_tinhocdaicuong’.

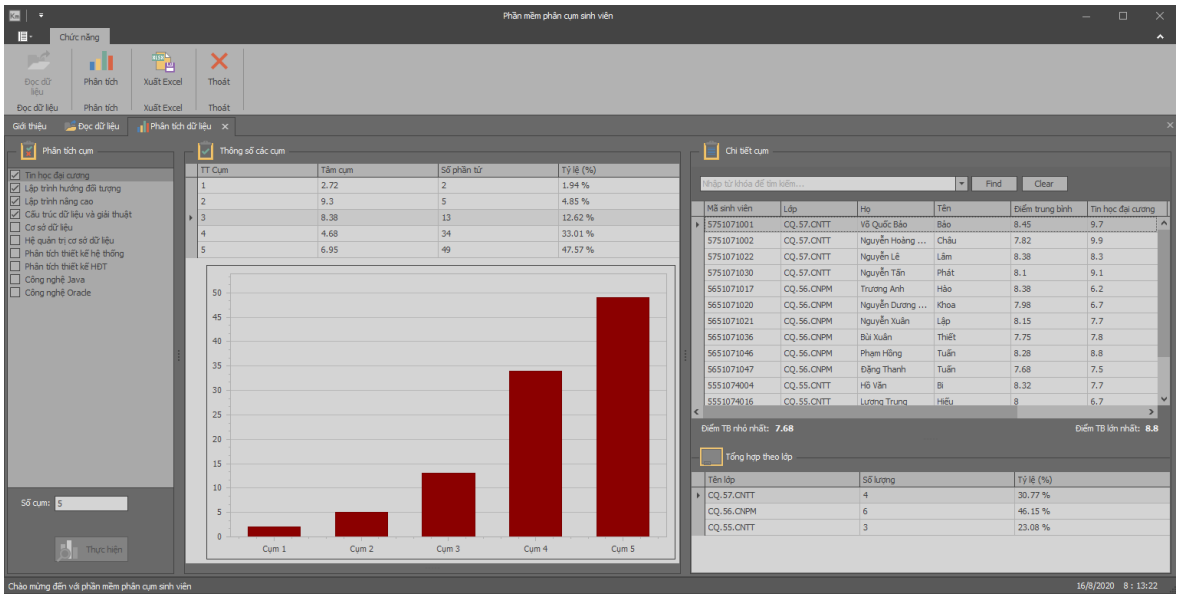
3.4.2. Phân tích

Sau khi đã mở tập tin Excel, bạn có thể chọn vào nút “Phân tích” để khởi động cửa sổ phân tích.

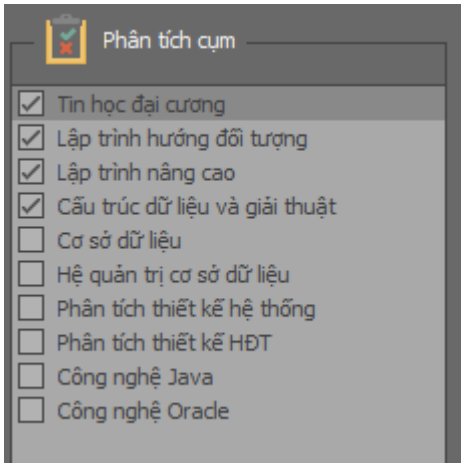


Hình 3.3 Giao diện ban đầu chức năng phân tích

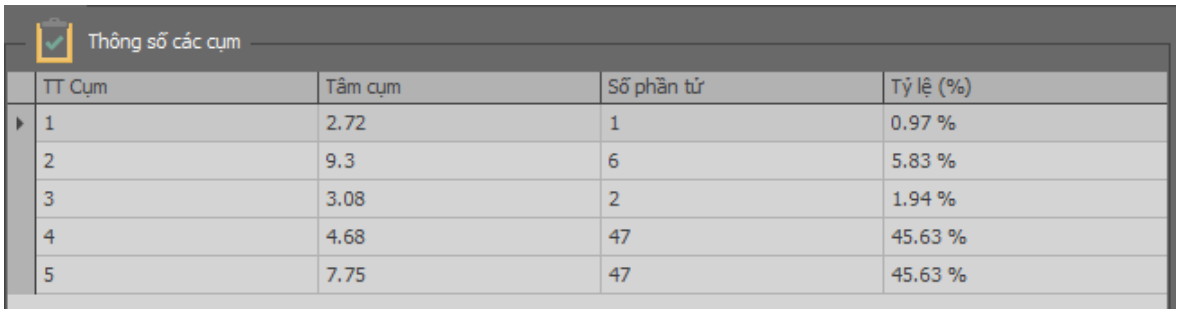
Để phân tích dữ liệu, bạn cần nhập vào số cụm và chọn các môn học cần phân tích và nhấn nút “Thực hiện” để bắt đầu phân tích dữ liệu.



Hình 3.4 Giao diện sau khi thực hiện phân tích

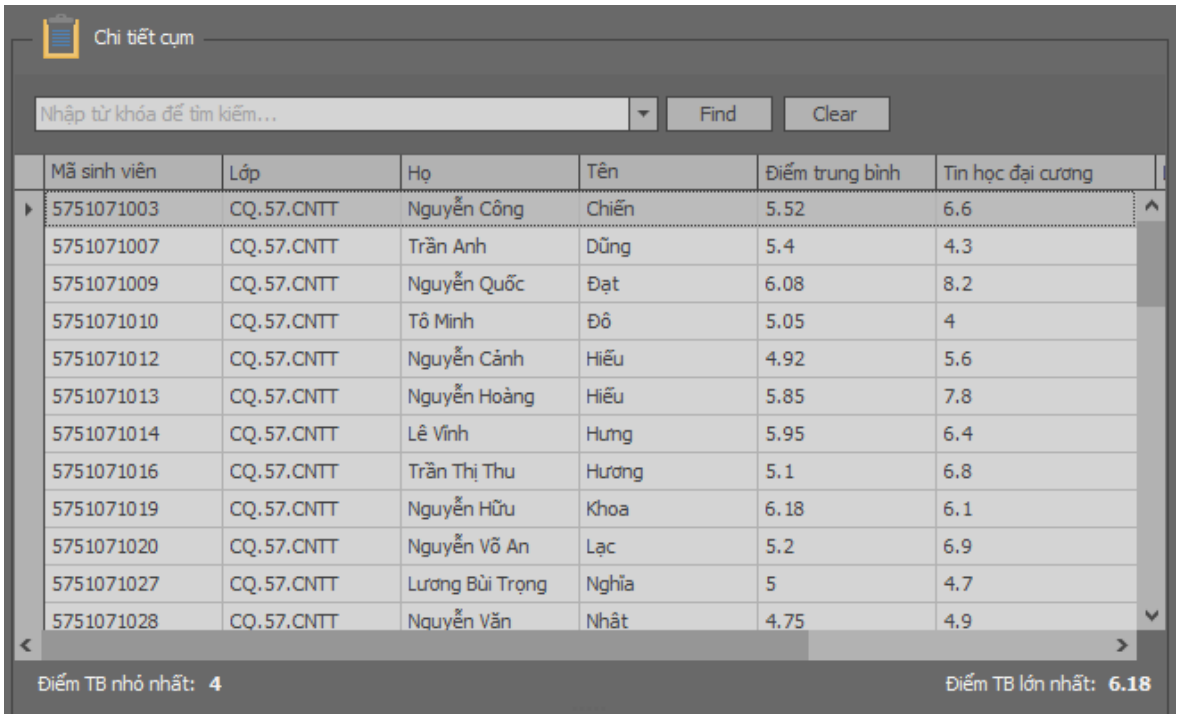


Hình 3.5 Hình chọn số cụm và các môn học cần phân tích

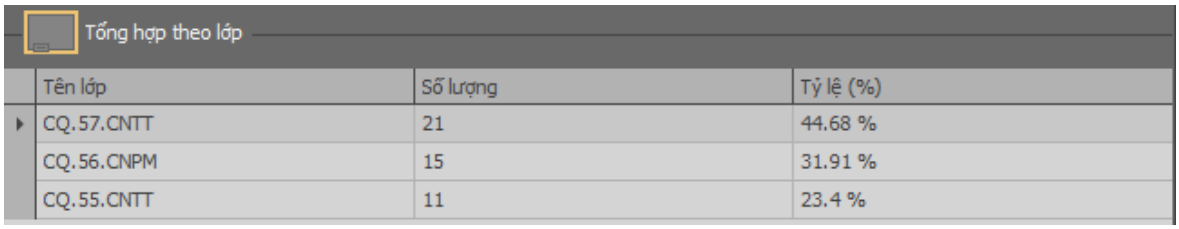


Hình 3.6 Thông số các cụm sau khi phân tích

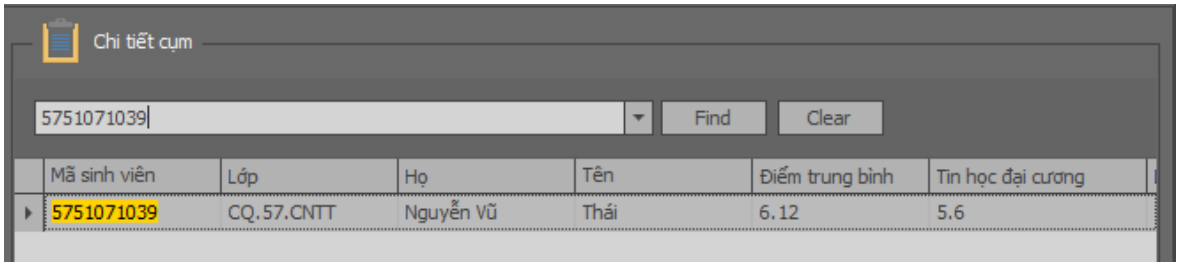
Cửa sổ thông số các cụm hiển thị các cụm sau khi phân tích, các thông số bao gồm: Thứ tự các cụm, tâm cụm, số lượng phần tử(sinh viên) trong cụm, tỷ lệ phần trăm số lượng phần tử so với tổng số lượng sinh viên trong danh sách.



Hình 3.7 Danh sách chi tiết các sinh viên thuộc mỗi cụm

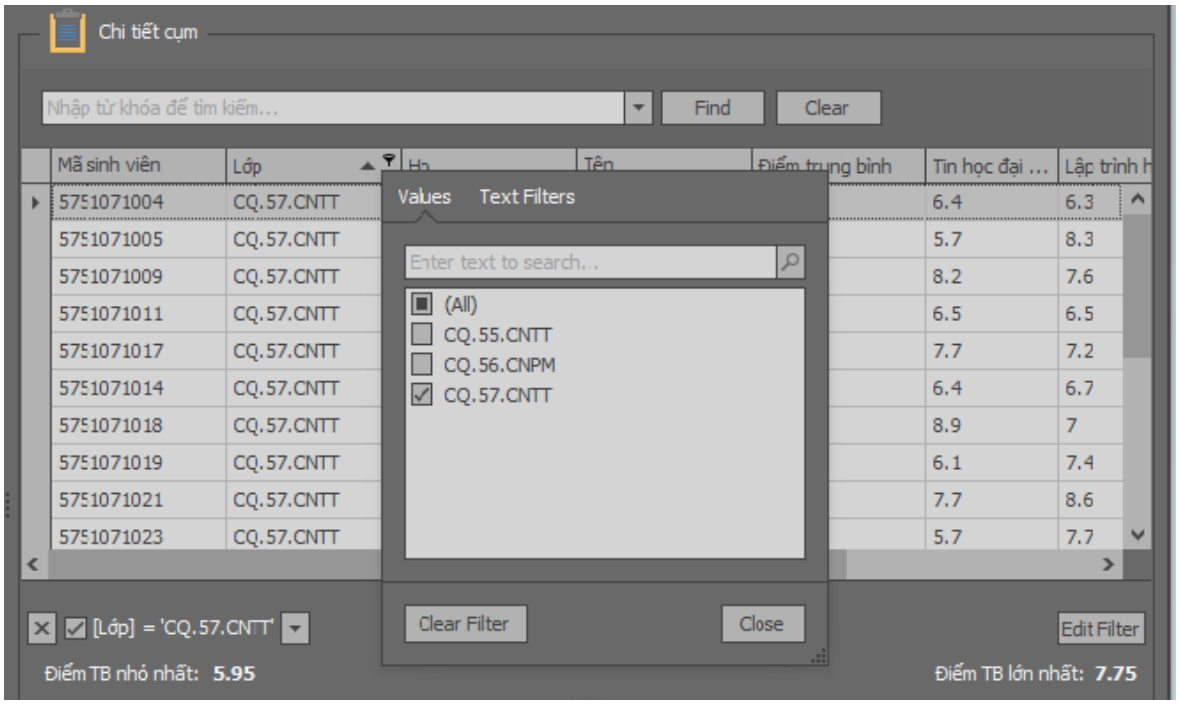


Hình 3.8 Danh sách các lớp có trong chi tiết cụm



Hình 3.9 Tìm kiếm sinh viên có trong chi tiết cụm

Chức năng tìm kiếm tìm kiếm một mã số sinh viên, điểm, tên, lớp bất kỳ.



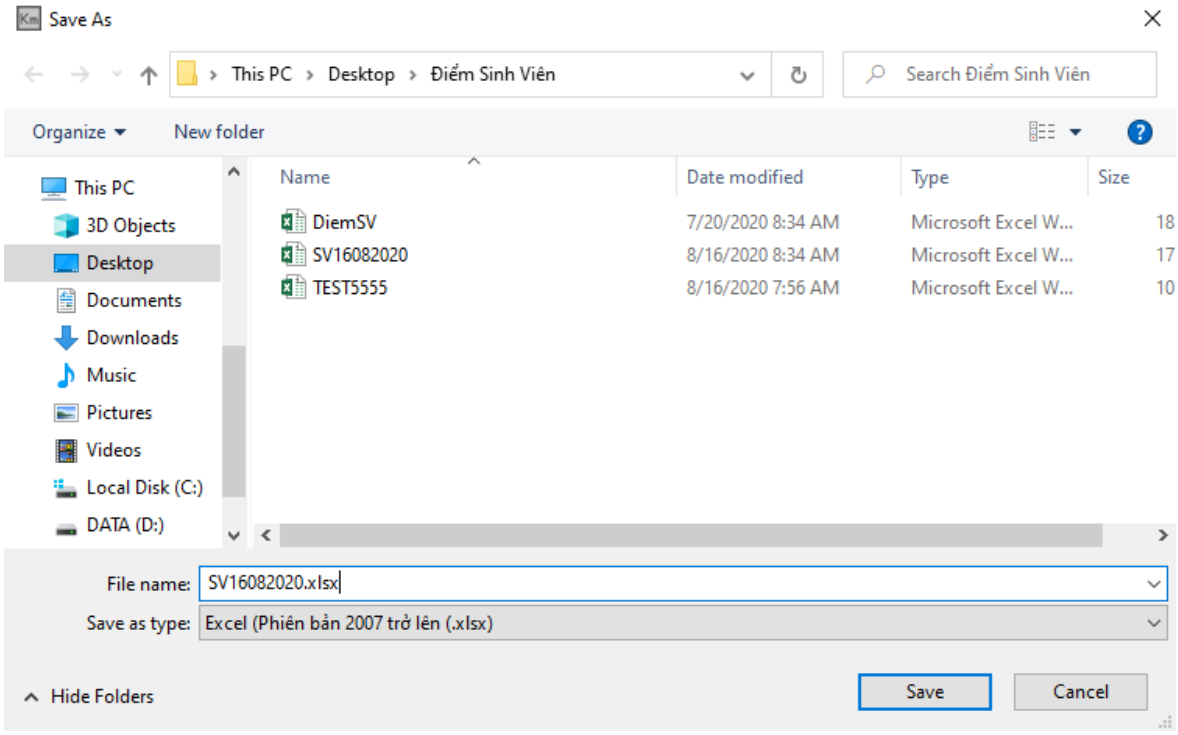
Hình 3.10 Lọc sinh viên theo lớp

***Lưu ý:**

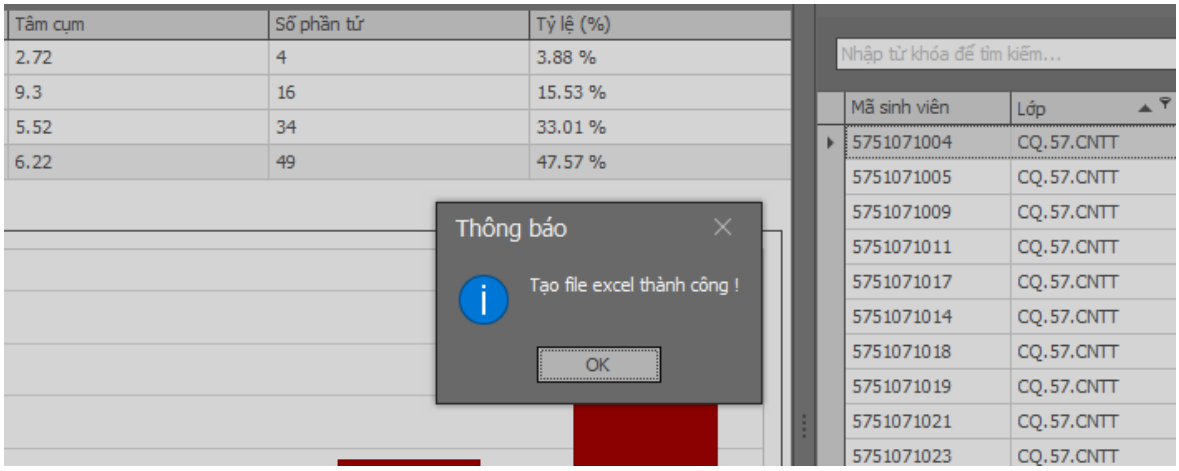
1. Số cụm nhập vào phải là số và lớn hơn 0.
2. Khi chọn các môn học, phải chọn các môn có trong tập tin Excel.

3.4.3. Xuất Excel

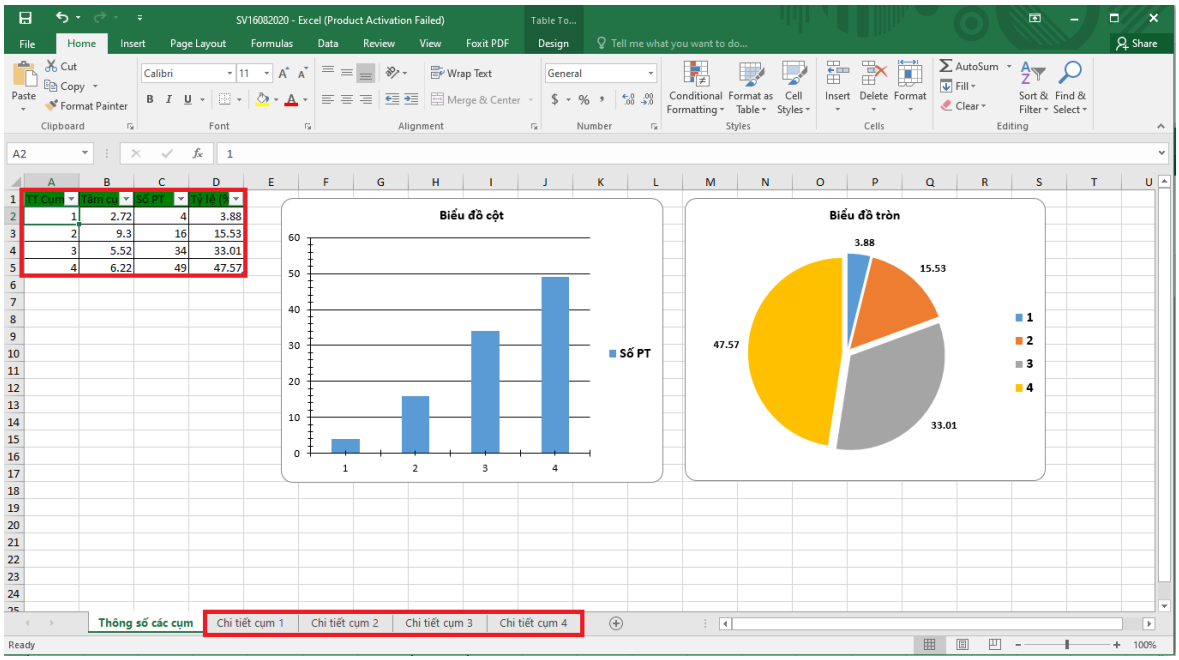
Sau khi phân tích, nếu bạn có nhu cầu xuất các bảng trong cửa sổ phân tích thì bạn có thể chọn nút "Xuất Excel" trên thanh công cụ hoặc nhấn tổ hợp phím "Ctrl+S" để lưu các bảng ra định dạng Excel.



Hình 3.11 Cửa sổ chọn đường dẫn lưu tập tin Excel

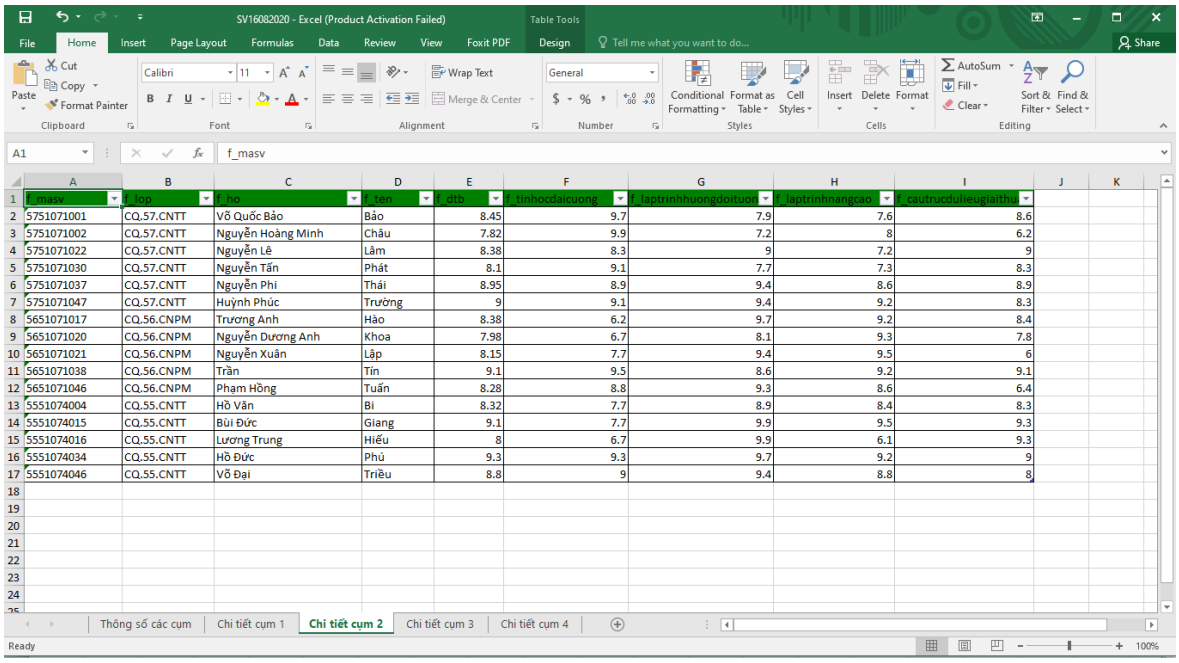


Hình 3.12 Thông báo lưu tập tin thành công



Hình 3.13 Kết quả sau khi xuất Excel

Cấu trúc file được tạo như sau: Trang đầu tiên chứa "Thông số các cụm" và biểu đồ. Các trang tiếp theo là chi tiết các cụm.

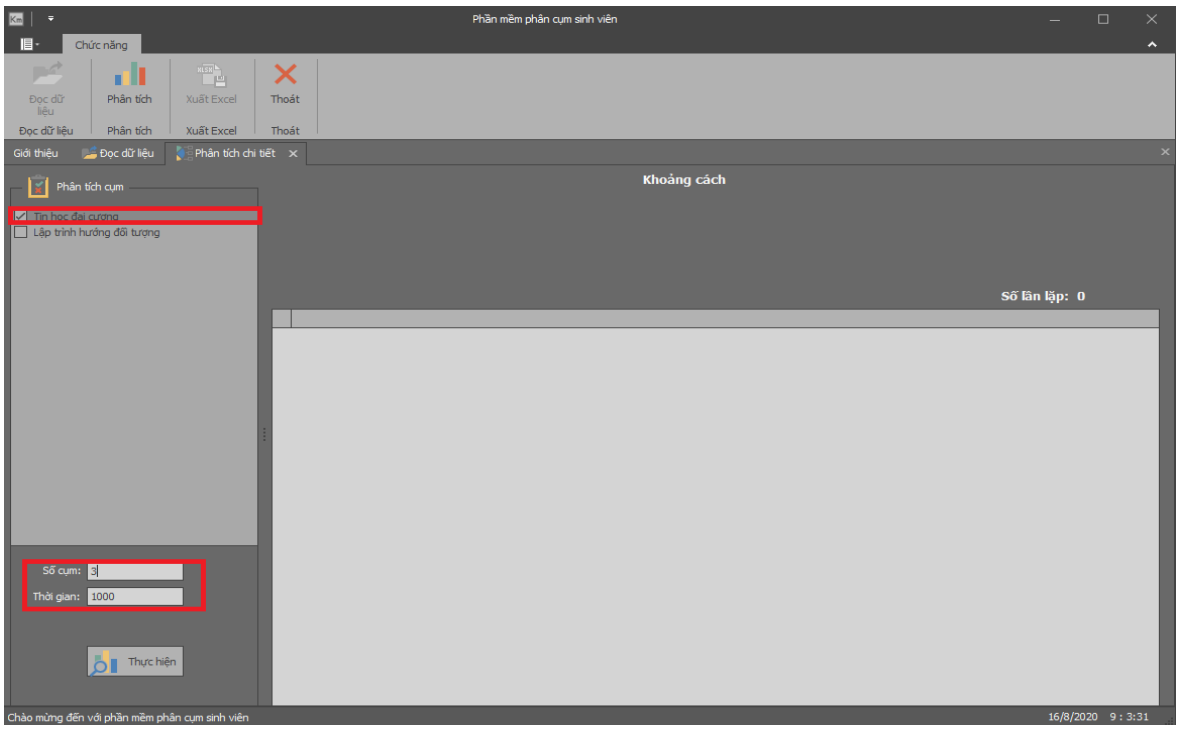


Hình 3.14 Chi tiết các cụm trong file Excel

3.4.5. Phân tích chi tiết

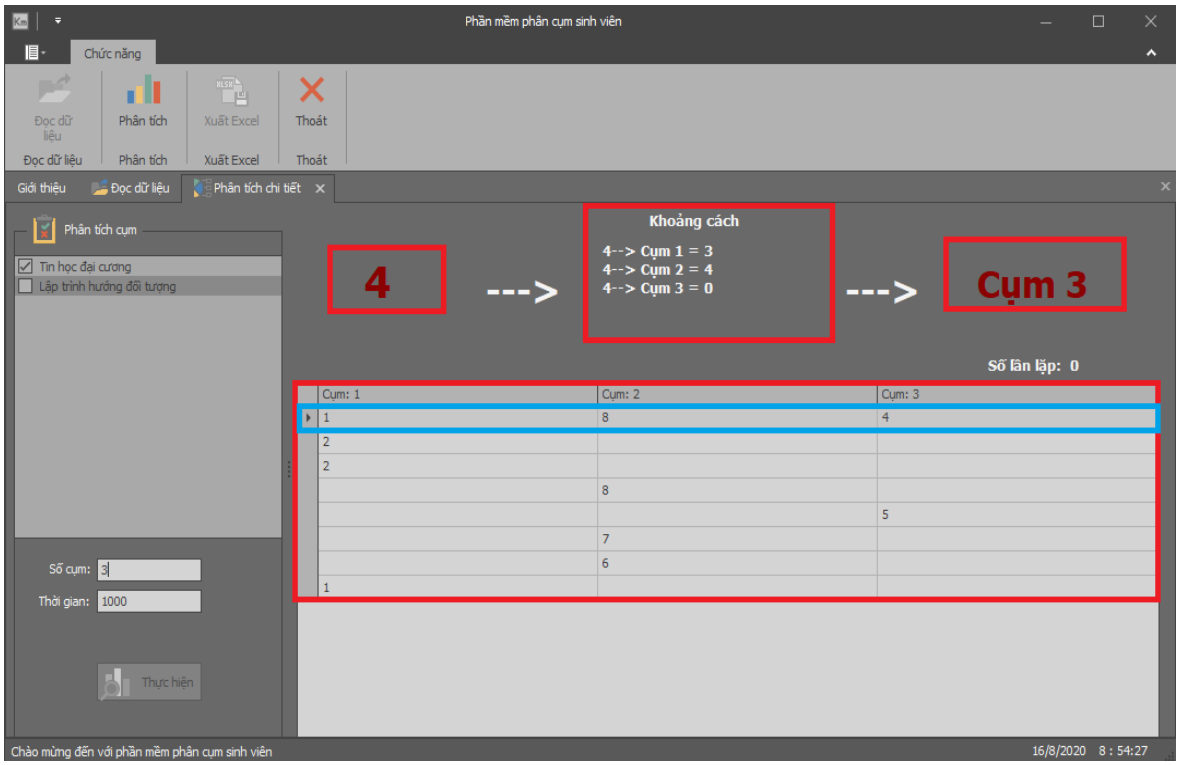
Phân tích chi tiết là chức năng cho thấy rõ các bước chạy của thuật toán. Trong quá trình đang "Phân tích" hoặc "Phân tích chi tiết" người dùng phải đợi quá trình

hoàn tất (hoặc đóng ứng dụng rồi khởi động lại) mới thực hiện thao tác khác.



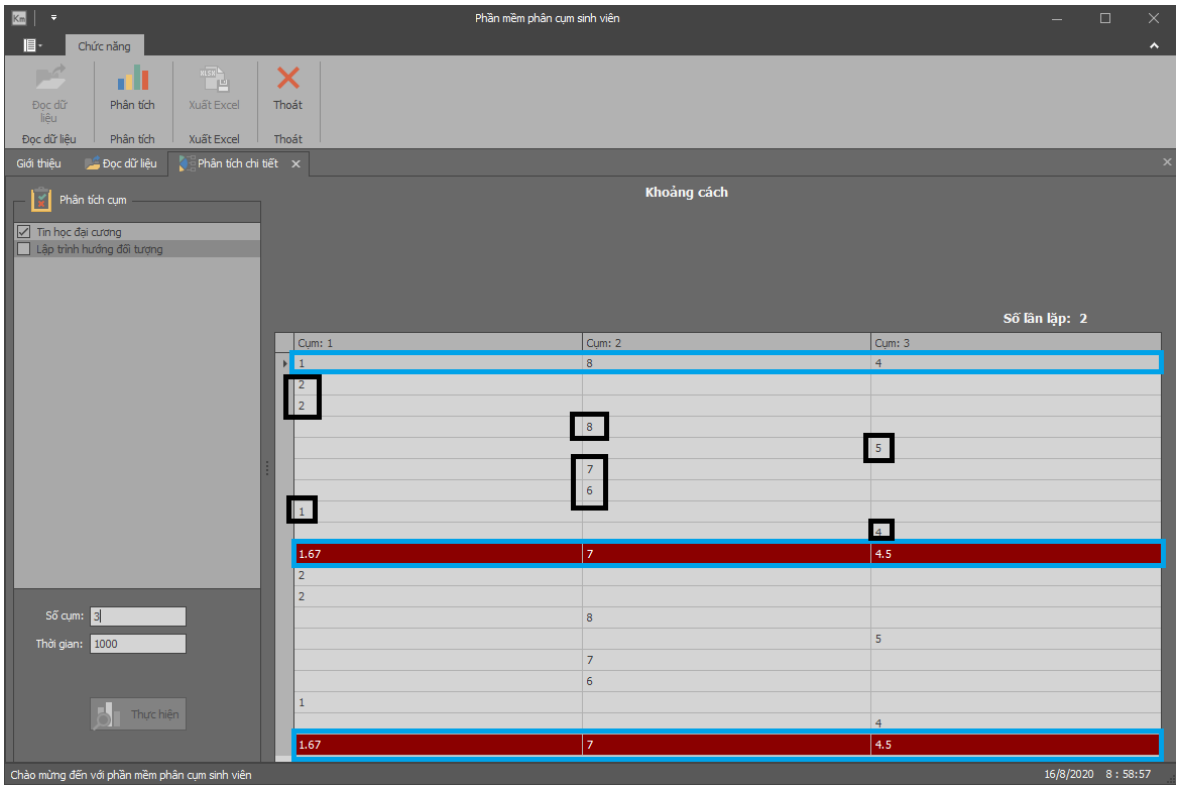
Hình 3.15 Giao diện phân tích cụm chi tiết

Sau khi nhấn chọn nút "Phân tích chi tiết" ứng dụng sẽ mở ra tab như hình trên. Thực hiện chọn môn học và nhập vào "số cụm" và "thời gian" (tốc độ thực thi – nhập: 500 hoặc 1000 hoặc 1500) và nhấn nút "Thực hiện".



Hình 3.16 Giao diện phân tích chi tiết khi thực hiện

Ô vuông đầu tiên là vùng xuất hiện của điểm trung bình các môn học đã chọn của các sinh viên. Ô vuông thứ hai là tính khoảng cách từ điểm đó đến các tâm cụm. Ô vuông thứ ba là tên cụm mà điểm đó thuộc về. Dòng đầu tiên trong Gridview được có viền xanh là dòng các tâm cụm ban đầu.



Hình 3.17 Giao diện phân tích chi tiết khi thực hiện xong

Sau khi gán các giá trị vào các tâm cụm thì sẽ tính lại các tâm cụm mới (các dòng được border màu xanh nền đỏ). Tâm cụm mới được tính bằng trung bình cộng các giá trị trong cột tương ứng với tâm cụm đó. Khi hai dòng cập nhật tâm cụm liên tiếp nhau có các giá trị từng cột bằng nhau thì thuật toán sẽ dừng.

3.5. Kết quả và đánh giá

Với những gì chương trình mang lại, ít nhiều cũng giúp được nhà trường đánh giá được kết quả học tập của sinh viên dựa trên điểm trung bình của các môn học hỗ trợ ra quyết định mở chuyên ngành và tư vấn cho sinh viên có nên theo một chuyên ngành nào đó dựa trên kết quả học tập của một số môn học nền tảng của chuyên ngành đó hoặc có thể khảo sát mở lớp học phần phù hợp với mỗi chuyên nghành.

KẾT LUẬN VÀ KIẾN NGHỊ

1. Kết quả đạt được

Trong quá trình nghiên cứu và hoàn thành đồ án tốt nghiệp với đề tài "Nghiên cứu thuật toán K-Means và ứng dụng phân cụm sinh viên Bộ môn Công nghệ thông tin", em đã đạt được kết quả như sau:

Về mặt khoa học

- Nắm bắt được các kiến thức cơ bản về Học máy, phát hiện tri thức, khai phá dữ liệu dựa trên kỹ thuật phân cụm phân hoạch dữ liệu.
- Hiểu rõ các quy trình và phương pháp phân cụm phân hoạch dữ liệu từ mô hình thực tế đến những bài toán cụ thể.

Về mặt ứng dụng

Từ những kết quả về mặt kiến thức đã đạt được ở trên, em đã xây dựng thành công ứng dụng phân loại điểm của sinh viên bằng thuật toán K-Means. Ứng dụng có các chức năng: trả về số cụm (do người dùng yêu cầu) và danh sách các đối tượng trong từng cụm, mỗi đối tượng chỉ thuộc một cụm duy nhất, trực quan hóa kết quả qua biểu đồ, in ra file Excel. Nhằm giúp cho lãnh đạo Khoa Công nghệ thông tin có thể dựa vào để phân tích và đưa ra các biện pháp kịp thời và chính xác trong việc quản lý sinh viên.

Về mặt con người

Qua quá trình làm đồ án mặc dù thời gian không quá nhiều nhưng em cũng đã học hỏi, rèn luyện thêm cho bản thân một số kỹ năng về tìm kiếm, nghiên cứu tài liệu, phân tích bài toán, cách nhìn nhận và xử lý vấn đề, làm việc nhóm, lập trình, rèn luyện tính kiên nhẫn và cách trình bày văn bản hợp lý hơn … rất hữu ích cho bản thân em trong công việc và cuộc sống sau này.

Như vậy, em đã hoàn thành cơ bản những mục tiêu được đặt ra ban đầu với đề tài nghiên cứu này.

2. Tồn tại

Bên cạnh những khía cạnh đạt được, do thời gian thực hiện có hạn cùng với trình độ kiến thức còn nhiều hạn chế nên đã còn những thiếu sót như:

- Một số chức năng còn chưa khắc phục được.
- Tập dữ liệu mẫu được sử dụng còn hạn chế.

3. Hướng phát triển

Trong tương lai, nếu có điều kiện đồ án của em sẽ được phát triển theo các hướng sau:

- Phát triển bài toán với số dữ liệu lớn hơn, bao quát hơn.
- Tiếp tục nghiên cứu các phương pháp, các cách tiếp cận mới về phân cụm dữ liệu: phân cụm thống kê, phân cụm khái niệm, phân cụm mờ, … tìm kiếm, so sánh và lựa chọn thuật toán tối ưu nhất để giải quyết bài toán đã đưa ra.

Vì thời gian thực hiện đề tài có hạn nên trong quá trình làm việc, nghiên cứu không thể tránh khỏi những thiếu sót, rất mong nhận được sự đóng góp ý kiến của Quý Thầy Cô.

TÀI LIỆU THAM KHẢO

[1]. Mark J. Price, *C# 8.0 with .NET core 3.0 – Modern Cross-Platform Development*, Packt Publishing, 2019.

[2]. Charu C. Aggarwal, *Data Mining: The Textbook*, Springer, 2015.

[3]. David Natingga, *Data Science Algorithms in a Week*: *Top 7 algorithms for computing, data analysis, and machine learning*, Birmingham, 2017.

[4]. Võ Thị Ngọc Châu, *Giáo trình Khai phá dữ liệu*, Đại học Bách Khoa Thành phố Hồ Chí Minh.

[5]. Nguyễn Vương Thịnh, *Bài giảng môn học Khai phá dữ liệu*, Đại học Hàng Hải Việt Nam.

[6]. Nguyễn Văn Lễ, Mạnh Thiên Lý, Nguyễn Thị Định, Nguyễn Thị Thanh Thủy, *Cải tiến thuật toán K-Means và ứng dụng hỗ trợ sinh viên chọn chuyên nghành theo học chế tín chỉ*, Tạp chí Khoa học công nghệ và Thực phẩm, Trường Đại học Công nghiệp Thực phẩm TP.HCM, 2018.

[7]. Nguyễn Thị Hữu Phương, Nguyễn Trường Xuân, Đặng Văn Đức, *Sử dụng thuật toán K – Means trong bài toán phân loại đám mây điểm LiDAR*, Tạp chí Khoa Học & Công Nghệ, Trường Đại học Mỏ - Địa chất, 2017

[8]. Nguyễn Văn Huân, Phạm Việt Bình, Trương Mạnh Hà, Vũ Xuân Nam, Đoàn Mạnh Hồng, *Cải tiến thuật toán K – Means và ứng dụng phân cụm dữ liệu tự động* 61 (12/2): 102 - 106, Tạp chí Khoa Học & Công Nghệ, Trường Đại học Thái Nguyên.

[9]. Trần Hùng Cường, Ngô Đức Vinh, *Tổng quan về phát hiện tri thức và khai phá dữ liệu,* Tạp chí Khoa Học & Công Nghệ, Số 5.2011, Trường Đại học Công nghiệp Hà Nội, 2011.

[10]. Giới thiệu về Machine Learning, link: https://machinelearningcoban.com/2016/12/26/introduce/, truy cập vào ngày 14/08/2020.

[11]. Machine Learning là gì? Ứng dụng của nó trong doanh nghiệp sản xuất, link:

https://ifactory.com.vn/machine-learning-la-gi-ung-dung-cua-no-trong-doanh-nghiep-san-xuat/, truy cập vào ngày 14/08/2020.

[12]. Machine Learning cho người mới bắt đầu, link: https://viblo.asia/p/machine-learning-cho-nguoi-moi-bat-dau-part-1-3Q75wpyGKWb, truy cập vào ngày 14/08/2020.

[13]. Tầm quan trọng của .Net Framework, link: http://bugnetproject.com/net-framework-la-gi-tai-sao-no-quan-trong-trong-moi- may-tinh/, truy cập vào ngày 14/07/2020.

[14]. Tổng quan về khai phá dữ liệu, link: https://www.academia.edu/19660657/khai_phá_dữ_liệu, truy cập vào ngày 14/08/2020.

[15]. Tổng quan về .Net Framework, link: https://vi.wikipedia.org/wiki/.NET_Framework, truy cập vào ngày 14/08/2020.

[16]. Tìm hiểu về DevExpress – UI Control cho .NET Framework, link: https://viblo.asia/p/tim-hieu-ve-devexpress-ui-control-cho-net-framework-RnB5pBLJZPG, truy cập ngày 15/08/2020.

[17]. DevExpress – Sự lựa chọn tuyệt vời cho Winforms Control, link: https://techtalk.vn/devexpress-su-lua-chon-tuyet-voi-cho-winforms-control.html, truy cập ngày 15/08/2020.

[18]. Tổng quan ngôn ngữ C#, link: https://voer.edu.vn/c/ngon-ngu-c/cf37fa1e/383e2f05, truy cập vào ngày: 16/08/2020.

[19]. Ứng dụng thuật toán K-Means, link: https://kipalog.com/posts/Thuat-toan-Kmean-va-ung-dung, truy cập vào ngày 16/08/2020.

[20]. K-Means Clustering, link: https://machinelearningcoban.com/2017/01/01/kmeans/, truy cập vào ngày 16/08/2020.