**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 LOẠI KẾT QUẢ HỌC TẬP CỦA HỌC SINH**

Giảng viên hướng dẫn : TRẦN PHONG NHÃ

Sinh viên thực hiện : LÝ THỊ THU THỦY

Lớp : CQ.56.CNPM

Khoá :56

Tp. Hồ Chí Minh, tháng 06 năm 2019

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH** Độc lập – Tự do – Hạnh phúc

# NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP

BỘ MÔN: **CÔNG NGHỆ THÔNG TIN**

-------\*\*\*-------

|  |  |  |  |
| --- | --- | --- | --- |
| **Mã sinh viên:** | 5651071037 | **Họ tên sinh viên:** | Lý Thị Thu Thủy |
| **Khóa:** | 56 | **Lớp:** | CQ.56.CNPM |

* **Tên đề tài**

NGHIÊN CỨU THUẬT TOÁN K-MEANS VÀ ỨNG DỤNG PHÂN LOẠI KẾT QUẢ HỌC TẬP CỦA HỌC SINH.

* **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 học sinh 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 học sinh.
* **Nội dung và phạm vi đề tài**
* Nội dung đề tài:
  + Tổng quan về C# và .NET Framework.
  + Tổng quan về khai phá dữ liệu
  + Phân cụm dữ liệu và thuật toán K-Means.
  + Áp dụng và cài đặt thuật toán vào bài toán phân loại học sinh dựa vào điểm trung bình.
* Phạm vi đề tài:

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

* **Công nghệ, công cụ và ngôn ngữ lập trình**
* Công cụ lập trình: Visual Studio 2017.
* Công nghệ sử dụng: .NET Framework.
* 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 học sinh.
* **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](mailto:tpnha@utc2.edu.vn)

|  |  |
| --- | --- |
| **Ngày … tháng … năm 2019**  **Trưởng BM Công nghệ Thông tin** | **Đã giao nhiệm vụ TKTN**  **Giảng viên hướng dẫn**  **ThS. Trần Phong Nhã** |

Đã nhận nghiệm vụ TKTN

Sinh viên: Lý Thị Thu Thủy Ký tên:

Điện thoại: 0785623959 Email: [lythithuthuy2008@gmail.com](mailto:lythithuthuy2008@gmail.com)

# LỜI CẢM ƠN

Lời đầu tiên cho phép em 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 loại kết quả học tập của học sinh”. 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 2019***

**Sinh viên thực hiện**

**Lý Thị Thu Thủy**

# NHẬN XÉT CỦA GIÁO VIÊN

***Tp. Hồ Chí Minh, ngày … tháng … năm 2019***

**Giáo viên hướng dẫn**

**Trần Phong Nhã**

MỤC LỤC

[NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP i](#_Toc11324071)

[LỜI CẢM ƠN i](#_Toc11324072)

[NHẬN XÉT CỦA GIÁO VIÊN ii](#_Toc11324073)

[DANH MỤC THUẬT NGỮ iii](#_Toc11324074)

[DANH MỤC BẢNG BIỂU iv](#_Toc11324075)

[DANH MỤC HÌNH ẢNH v](#_Toc11324076)

[MỞ ĐẦU 1](#_Toc11324077)

[1. Lý do chọn đề tài 1](#_Toc11324078)

[2. Mục đích nghiên cứu 2](#_Toc11324079)

[3. Đối tượng và phạm vi nghiên cứu 2](#_Toc11324080)

[4. Phương pháp nghiên cứu 2](#_Toc11324081)

[5. Cấu trúc báo cáo đồ án tốt nghiệp 2](#_Toc11324082)

[CHƯƠNG 1: CƠ SỞ LÝ THUYẾT 4](#_Toc11324083)

[1.1. Tổng quan về .NET Framework 4](#_Toc11324084)

[1.1.1 Giới thiệu 4](#_Toc11324085)

[1.1.2.Mục tiêu chính của .NET Framework 4](#_Toc11324086)

[1.1.3 Kiến trúc .NET Framework 5](#_Toc11324087)

[1.1.4.Quá trình phát triển của .NET Framework 7](#_Toc11324088)

[1.2. Tổng quan về ngôn ngữ C# 7](#_Toc11324089)

[1.2.1.Giới thiệu ngôn ngữ C# 7](#_Toc11324090)

[1.2.2.Đặc trưng của ngôn ngữ C# 8](#_Toc11324091)

[1.3. Tổng quan về khám phá tri thức và khai phá dữ liệu 9](#_Toc11324092)

[1.3.1.Định nghĩa về khám phá tri thức: 9](#_Toc11324093)

[1.3.2.Quá trình khám phá tri thức 9](#_Toc11324094)

[1.3.3.Khái niệm khai phá dữ liệu 11](#_Toc11324095)

[1.3.4.Quá trình khai phá dữ liệu 12](#_Toc11324096)

[1.3.5.Các phương pháp khai phá dữ liệu 13](#_Toc11324097)

[1.3.6.Ứng dụng, thách thức của khai phá dữ liệu 13](#_Toc11324098)

[1.4. Tổng quan về phân cụm dữ liệu 15](#_Toc11324099)

[1.4.1 Giới thiệu 15](#_Toc11324100)

[1.4.2.Các mục tiêu của phân cụm dữ liệu 16](#_Toc11324101)

[1.4.3 Một số thuộc tính 17](#_Toc11324102)

[1.4.4.Một số kỹ thuật phân cụm dữ liệu 18](#_Toc11324103)

[1.4.4.Ứng dụng của phân cụm dữ liệu 20](#_Toc11324104)

[1.4.5.Các yêu cầu và những vấn đề còn tồn tại 20](#_Toc11324105)

[CHƯƠNG 2: PHÂN TÍCH THUẬT TOÁN K-MEANS 23](#_Toc11324106)

[2.1. Tổng quan thuật toán K-Means 23](#_Toc11324107)

[2.1.1.Giới thiệu thuật toán 23](#_Toc11324108)

[2.1.2.Một số khái niệm dùng trong thuật toán 24](#_Toc11324109)

[2.1.3.Mô tả thuật toán 26](#_Toc11324110)

[2.1.4.Ví dụ về thuật toán 28](#_Toc11324111)

[2.2. Đặc điểm của thuật toán 31](#_Toc11324112)

[2.3. Ứng dụng 33](#_Toc11324113)

[2.4. Thuật toán khắc phục nhiễu dữ liệu 34](#_Toc11324114)

[CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ 36](#_Toc11324115)

[3.1. Giới thiệu bài toán 36](#_Toc11324116)

[3.2. Tập dữ liệu sử dụng 37](#_Toc11324117)

[3.3. Môi trường thử nghiệm 38](#_Toc11324118)

[3.4. Giao diện chương trình 38](#_Toc11324119)

[3.5. Kết quả và đánh giá 43](#_Toc11324120)

[KẾT LUẬN VÀ KIẾN NGHỊ 44](#_Toc11324121)

[TÀI LIỆU THAM KHẢO 46](#_Toc11324122)

# DANH MỤC THUẬT NGỮ

|  |  |  |  |
| --- | --- | --- | --- |
| Từ viết tắt | Mô tả | Ý nghĩa | Ghi chú |
| API | [Application Programming Interface](https://en.wikipedia.org/wiki/Application_programming_interface) | Giao diện lập trình ứng dụng |  |
| CLR | Common Language Runtime | Khối thực thi ngôn ngữ chung |  |
| CSDL | Database | Cơ sở dữ liệu |  |
| FCL | Framework Class Library | Thư viện lớp khung |  |
| IDE | Integrated Development Environment | Môi trường tích hợp phát triển |  |
| KDD | Knowledge Discovery in Database | Phát hiện tri thức |  |
| KPDL | Data mining | Khai phá dữ liệu |  |
| SQL | Structured Query Language | Ngôn ngữ truy vấn dữ liệu có cấu trúc |  |
|  | Association rules | Luật kết hợp |  |
|  | Classification | Phân lớp |  |
|  | Clustering | Phân cụm |  |
|  | Data Mining | Khai phá dữ liệu |  |
|  | Regression | Hồi quy |  |

# DANH MỤC BẢNG BIỂU

[Bảng 2.1 - Minh họa về ma trận phân hoạch 24](#_Toc11349898)

[Bảng 2.2 - Tập dữ liệu ví dụ thuật toán K-Means 28](#_Toc11349899)

[Bảng 3.1 - Bảng các thuộc tính của tập dữ liệu 38](#_Toc11349900)

# DANH MỤC HÌNH ẢNH

[Hình 1.1 - Kiến trúc .NET Framework 6](#_Toc11349014)

[Hình 1.2 - Quá trình khám phá tri thức 9](#_Toc11349015)

[Hình 1.3 - Quá trình khai phá dữ liệu 12](#_Toc11349016)

[Hình 1.4 - Minh họa phân cụm dữ liệu 16](#_Toc11349017)

[Hình 2.1 - Ví dụ mô phỏng phân cụm với thuật toán K-Means 24](#_Toc11349018)

[Hình 2.2 - Sơ đồ khối thuật toán K-Means 26](#_Toc11349019)

[Hình 3.1 - Giao diện chức năng Đọc dữ liệu 38](#_Toc11349020)

[Hình 3.2 - Cửa sổ chọn file cần phân tích 39](#_Toc11349021)

[Hình 3.3 - Giao diện hiển thị dữ liệu 39](#_Toc11349022)

[Hình 3.4 - Giao diện ban đầu chức năng Phân tích 40](#_Toc11349023)

[Hình 3.5 - Giao diện kết quả chọn môn Lý, Hóa, số cụm là 4 41](#_Toc11349024)

[Hình 3.6 - Hình chọn số cụm và các môn học cần phân tích 41](#_Toc11349025)

[Hình 3.7 - Bảng thông số cụm sau khi phân tích. 42](#_Toc11349026)

[Hình 3.8 - Danh sách chi tiết các học sinh thuộc 1 cụm 42](#_Toc11349027)

[Hình 3.9 - Danh sách các lớp có trong chi tiết cụm. 43](#_Toc11349028)

[Hình 3.10 - Giao diện mô tả kết quả vừa phân cụm 43](#_Toc11349029)

# MỞ ĐẦU

## 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 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 học sinh có học lực yếu – kém để đưa vào kế hoạch phụ đạo cũng như những học sinh có học lực khá – giỏi để đưa vào kế hoạch bồi dưỡng học sinh giỏi hoặc đề ra các kế hoạch, phương pháp giảng dạy phù hợp với từng học sinh nhằm nâng cao chất lượng giáo dục và hỗ trợ các hoạt động quản lý hiện nay hiệu quả hơn.

Với tầm quan trọng của giáo dục nhất là trong thời đại của cuộc cách mạng khoa học – công nghệ hiện đại cùng sự tiến đến nền văn minh trí tuệ hiện nay và với những lý do trên em xin chọn đề tài “NGHIÊN CỨU THUẬT TOÁN K-MEANS VÀ ỨNG DỤNG PHÂN LOẠI KẾT QUẢ HỌC TẬP CỦA HỌC SINH” làm đề tài đồ án tốt nghiệp.

## 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 giáo dục. Đề ra giải pháp ứng dụng công nghệ thông tin vào việc phân cụm học sinh 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.

## Đối tượng và phạm vi nghiên cứu

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

## 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 học sinh.

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

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.
* Tìm hiểu thuật toán cải tiến của K-means

Chương 3: Thử nghiệm và đánh giá

* Cài đặt thuật toán.
* Đánh giá thuật toán đối với bài toán phân cụm học sinh.
* 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

## Tổng quan về .NET Framework

### Giới thiệu

.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. [13]

### Mục tiêu chính của .NET Framework

Chính vì những mục tiêu sau đã làm cho .NET Framework trở nên quan trọng trong mỗi máy tính

**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 nǎng 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](http://docs.servicestack.net/what-is-a-message-based-web-service.html) – 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 nǎng 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](https://www.ibm.com/support/knowledgecenter/en/SS8PJ7_8.5.1/com.ibm.ccl.soa.deploy.domain.detail.doc/topics/dotnet_templates.html), 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**

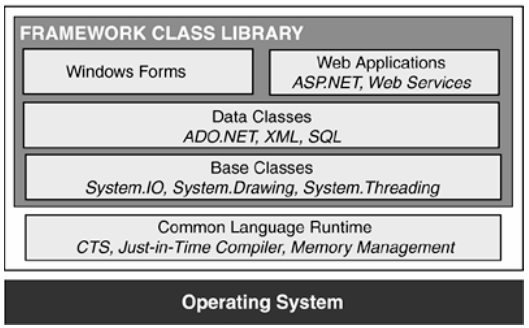
Một trong những vấn đề quan trọng nhất liên quan đến Internet đó chính là bảo mật thông tin. Kiến trúc bảo mật của .NET Framework được thiết kế từ dưới lên nhằm đảm bảo cho ứng dụng và dữ liệu được bảo vệ thông qua một mô hình bảo mật an toàn và tinh vi.

**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. [11]

### Kiến trúc .NET Framework

Kiến trúc của .NET Framwork được chia thành 2 phần chính: CLR và FCL



Hình 1.1 - Kiến trúc .NET Framework

* **CLR (Common Language Runtime)**

Bất kỳ chương trình nào viết bằng .NET cũng sử dụng một môi trường phần mềm được giới lập trình biết đến như CLR. CLR là một máy ảo có thể cung cấp một loạt các dịch vụ bao gồm:

**Bảo mật:** .NET có cơ chế bảo mật của riêng mình, gọi là Bảo mật truy nhập mã (CAS). CAS được xây dựng trên sự rõ ràng có liên quan đến một hợp ngữ cụ thể để quản lý quyền được chấp nhận lấy mã.

**Quản lý bộ nhớ:** CLR loại bỏ gánh nặng quản lý bộ nhớ cho nhà phát triển; nó tự mình xử lý việc quản lý bộ nhớ bằng cách phát hiện ra khi nào bộ nhớ có thể được giải phóng một cách an toàn.

**Hiệu suất:**Sau khi khởi động ứng dụng, .NET Framework biên dịch mã ngôn ngữ trung gian chung thành mã thực thi bằng cách sử dụng trình biên dịch "Just-in-time", và lưu trữ chương trình thực thi vào Native Image Cache. Nhờ kiểu lưu trữ này, các ứng dụng khởi chạy nhanh hơn, mặc dù lần đầu khởi chạy thường sẽ chậm hơn đôi chút.

Bộ thực thi bao gồm nhiều dịch vụ hỗ trợ, phát triển ứng dụng cũng như cải thiện tính đáng tin cậy của ứng dụng. [9]

* ***FCL (Framework Class Library)***

FCL là thư viện kiểu dữ liệu có thể tái sử dụng (gồm các class, structure, …) dành cho các ứng dụng thực thi trong .NET. Tất cả các ngôn ngữ hỗ trợ .NET Framework đều sử dụng thư viện lớp dùng chung này.

FCL cung cấp giao diện người dùng, kết nối cơ sở dữ liệu, phát triển ứng dụng web, khả năng truy cập dữ liệu, các thuật toán, truyền thông mạng và mã hoá. Cùng với các thư viện lớp (class libraries), nhiều mô hình ứng dụng cũng được sử dụng để tạo ra các ứng dụng web. [11]

### Quá trình phát triển của .NET Framework

* .NET Framework 1.0 (Năm 2002, Visual Studio .NET 2002)
* .NET Framework 1.1 (Năm 2003, Visual Studio .NET 2003)
* .NET Framework 2.0 (Năm 2005, Visual Studio .NET 2005)
* .NET Framework 3.0 (Năm 2006, Visual Studio 2005)
* .NET Framework 3.5 (Năm 2008, Visual Studio 2008)
* .NET Framework 4.0 (Năm 2010, Visual Studio 2010)
* .NET Framework 4.5 (Năm 2012, Visual Studio 2012)
* .NET Framewirk 4.6 (Năm 2015, Visual Studio 2015). [13]

## Tổng quan về ngôn ngữ C#

### Giới thiệu ngôn ngữ 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. [13]

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.

### Đặc trưng của ngôn ngữ C#

***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.

***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. [14]

## Tổng quan về khám phá tri thức và khai phá dữ liệu

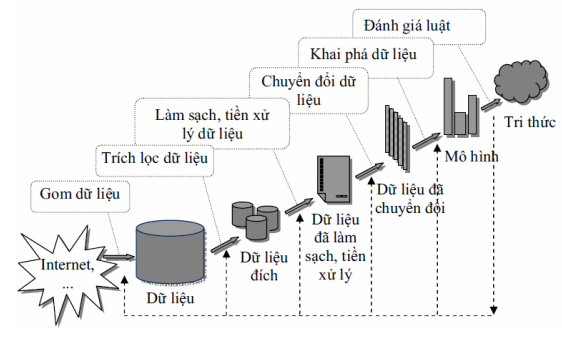
### Định nghĩa về khám phá tri thức:

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. [8]

### Quá trình khám phá tri thức

Quá trình khám phá tri thức tiến hành qua 6 giai đoạn như hình



Hình 1.2 - Quá trình khám phá tri thức

Bắt đầu của quá trình là kho dữ liệu thô và kết thúc với tri thức được chiết xuất ra. Về lý thuyết thì có vẻ rất đơn giản nhưng thực sự đây là một quá trình rất khó khăn gặp phải rất nhiều vướng mắc như: quản lý các tập dữ liệu, phải lặp đi lặp lại toàn bộ quá trình, ...

**Gom dữ liệu**: Tập hợp dữ liệu là bước đầu tiên trong quá trình khai phá dữ liệu. Đây là bước được khai thác trong một cơ sở dữ liệu, một kho dữ liệu và thậm chí các dữ liệu từ các nguồn ứng dụng Web.

**Trích lọc dữ liệu**: Ở giai đọan này dữ liệu được lựa chọn hoặc phân chia theo một số tiêu chuẩn nào đó phục vụ mục đích khai thác, ví dụ chọn tất cả những em học sinh có điểm Trung bình học kỳ lớn hơn 8.0 và có giới tính nữ.

**Làm sạch, tiền xử lý và chuẩn bị trước dữ liệu**: Giai đoạn thứ ba này là giai đoạn hay không được chú trọng, nhưng thực tế nó là một bước rất quan trọng trong quá trình khai phá dữ liệu. Một số lỗi thường mắc phải trong khi gom dữ liệu là tính không đủ chặt chẽ, logic. Vì vậy, dữ liệu thường chứa các giá trị vô nghĩa và không có khả năng kết nối dữ liệu.

***Ví dụ:*** Điểm Trung bình = 12.4. Giai đoạn này sẽ tiến hành xử lý những dạng dữ liệu không chặt chẽ nói trên. Những dữ liệu dạng này được xem như thông tin dư thừa, không có giá trị. Bởi vậy, đây là một quá trình rất quan trọng vì dữ liệu này nếu không được “làm sạch – tiền xử lý – chuẩn bị trước” thì sẽ gây nên những kết quả sai lệch nghiêm trọng.

**Chuyển đổi dữ liệu:** Tiếp theo là giai đoạn chuyển đổi dữ liệu, dữ liệu đưa ra có thể sử dụng và điều khiển được bởi việc tổ chức lại nó, tức là dữ liệu sẽ được chuyển đổi về dạng phù hợp cho việc khai phá bằng cách thực hiện các thao tác nhóm hoặc tập hợp.

**Khai phá dữ liệu:** Đây là bước mang tính tư duy trong khai phá dữ liệu. Trong giai đoạn này ta sử dụng các kỹ thuật nhằm phát hiện ra các tri thức tiềm ẩn trong dữ liệu. Một số kỹ thuật được sử dụng đó là: phân lớp, gom cụm, luật kết hợp, …

**Đánh giá các luật và biểu diễn tri thức**: Đây là giai đoạn cuối cùng trong tiến trình khám phá tri thức. Trong giai đoạn này, các mẫu dữ liệu được chiết xuất bởi các phần mềm khai phá dữ liệu. Không phải bất cứ mẫu nào cũng đều có ích, thậm chí còn bị sai lệch. Chính vì vậy, cần phải xác định và lựa chọn những tiêu chuẩn đánh giá sao cho sẽ chiết xuất ra các tri thức cần thiết.

Nếu khám phá tri thức là toàn bộ quá trình chiết xuất tri thức từ các cơ sở dữ liệu thì khai phá dữ liệu là giai đoạn chủ yếu của quá trình đó. Như trên đã trình bày, trong quá trình khám phá tri thức, giai đoạn khai phá dữ liệu được thực hiện sau các giai đoạn tinh lọc và tiền xử lý dữ liệu, tức là việc khai phá để tìm ra các mẫu có ý nghĩa được tiến hành trên tập dữ liệu có hy vọng là sẽ thích hợp với nhiệm vụ khai phá đó chứ không phải là khai phá hết dữ liệu với một thời gian đủ dài để lấy được một mẫu không thực sự có ích như khái niệm trong thống kê trước đây. Vì vậy, khai phá dữ liệu thường bao gồm việc thử tìm mô hình phù hợp với tập dữ liệu và tìm kiếm các mẫu từ tập dữ liệu theo mô hình đó. Chẳng hạn ta có mô hình là một luật kết hợp thì mẫu là các yếu tố tham gia cùng với các độ hỗ trợ (support) và độ tin cậy (confidence) trong các luật tương ứng. [7, 8]

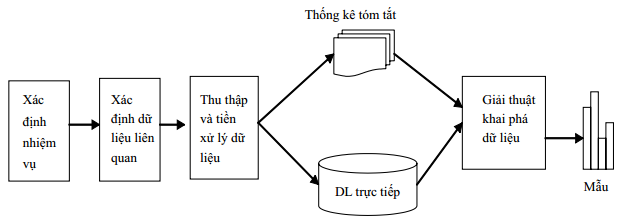
### Khái niệm khai phá dữ liệu

KPDL là một khái niệm ra đời vào những năm cuối của thập kỷ 80. Nó bao hàm một loạt các kỹ thuật nhằm phát hiện ra các thông tin có giá trị tiềm ẩn trong các tập dữ liệu lớn (các kho dữ liệu). Về bản chất, KPDL liên quan đến việc phân tích các dữ liệu và sử dụng các kỹ thuật để tìm ra các mẫu hình có tính chính quy (regularities) trong tập dữ liệu.

Năm 1989, Fayyad, Piatestsky-Shapiro và Smyth đã dùng khái niệm Khám phá tri thức trong CSDL để chỉ toàn bộ quá trình phát hiện các tri thức có ích từ các tập dữ liệu lớn. Trong đó, KPDL là một bước đặc biệt trong toàn bộ tiến trình, sử dụng các giải thuật đặc biệt để chiết xuất ra các mẫu (pattern) (hay các mô hình) từ dữ liệu.

Khai phá dữ liệu là một tiến trình sử dụng các công cụ phân tích dữ liệu khác nhau để khám phá ra các mẫu dưới nhiều góc độ khác nhau nhằm phát hiện ra các mối quan hệ giữa các dữ kiện, đối tượng bên trong CSDL, kết quả của việc khai phá là xác định các mẫu hay các mô hình đang tồn tại bên trong, nhưng chúng nằm ẩn khuất ở các CSDL. Để từ đó rút trích ra được các mẫu, các mô hình hay các thông tin và tri thức từ các CSDL. [12]

### Quá trình khai phá dữ liệu



Hình 1.3 - 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. [8]

### 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. [4]

### Ứng dụng, thách thức của khai phá dữ liệu

***Ứng dụng***

Mặc dù còn nhiều vấn đề mà KPDL cần phải tiếp tục nghiên cứu để giải quyết nhưng tiềm năng của nó đã được khẳng định bằng sự ra đời của rất nhiều ứng dụng. Các ứng dụng của KPDL trong khoa học cũng như được phát triển. Một số ứng dụng trong các lĩnh vực như:

**Ngân hàng**: Xây dựng mô hình dự báo rủi ro tín dụng; tìm kiếm tri thức, quy luật của thị trường chứng khoán và đầu tư bất động sản, …

**Thương mại điện tư**: công cụ tìm hiểu, định hướng, thúc đẩy, giao tiếp với khách hàng; phân tích khách hàng duyệt web; Phân tích hành vi mua sắm trên mạng và cho biết thông tin tiếp thị phù hợp vói loại khách hàng.

**Thiên văn học**: Hệ thống SKIcaT do JPl/Caltech phát triển được sử dụng cho các nhà thiên văn để tự động xác định các vì sao và các dải thiên hà trong một bản khảo sát lớn để có thể phân tích và phân loại (Fayyad, Djorgovski, & Weir).

**Sinh học phân tử**: Hệ thống tìm kiếm các mẫu trong cấu trúc phân tử (conklin, Fortier, và Glasgow 1993) và trong các dữ liệu gen (Holder, cook, và Djoko 1994). [12]

***Những thách thức***

* Khối lượng dữ liệu lớn và từ nhiều nguồn khác nhau: CSDL, Internet, các loại thiết bị thu nhận tín hiệu, các loại thiết bị thu nhận dạng, các loại thiết bị lưu trữ như băng từ, CD, … Số mẫu tin và số các thuộc tính quá lớn làm cho độ phức tạp và thời gian giải quyết bài toán tăng lên rất nhanh.
* Mô hình hay tri thức phát hiện được bị thay đổi theo thời gian tức là mô hình hay tri thức đó phụ thuộc vào thời điểm quan sát, lấy mẫu, thời điểm khai phá, kết quả đạt được sau khai phá cũng gây không ít khó khăn cho khai phá dữ liệu.
* Dữ liệu bị ảnh hưởng, bị nhiễu bởi tác động của môi trường bên ngoài, hay bộ dữ liệu không hoàn chỉnh làm cho dữ liệu không phản ánh trung thực, chính xác của các quy luật, tri thức mà ta tìm được.

Từ những vấn đề đặt ra ở trên nên tốc độ xử lý cần quan tâm trước nhất. Có hai phương hướng để giải quyết vấn đề này là nâng cao năng lực của phần cứng và cải tiến phần mềm. Tuy nhiên khi cải thiện năng lực của máy tính thì dữ liệu cũng tăng không ngừng, thậm chí còn tăng nhanh hơn gấp nhiều lần. Do vậy việc nghiên cứu đề xuất các thuật toán hiệu quả có khả năng làm việc trên khối dữ liệu lớn, và có độ phức tạp tính toán thấp là một hướng nghiên cứu đầy tiềm năng. Từ nhu cầu thực tế trên, gần đây đã xuất hiện nhiều ngành khoa học công nghệ hỗ trợ KPDL như tính toán song song, máy tính lượng tử, công nghệ nano, phát triển thuật toán… [12]

## Tổng quan về phân cụm dữ liệu

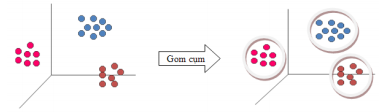
### Giới thiệu

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”.

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 1.4 - 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. [7]

### 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.

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.

### Một số thuộc tính

1. **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 = (x1, x2, ..., xk); y = (yl, y2, ..., yk); z = (zl, z2, ..., zk)Trong đó xi, yi, zi 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 ≠ y hoặc x = y.

Thuộc tính có thứ tự: Là thuộc tính định danh cóthêm 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 ≠ 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 xi > yi thì có thể nói x cách y một khoảng xi – yitương 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.

### 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.

Một số thuật toán điển hình của phương pháp phân cụm trên lưới như: STING, WaveCluster, …

**Phương pháp phân cụm dựa trên mô hình**

Điển hình trong phương pháp tiếp cận theo phân cụm dựa trên mô hình là các thuật toán như: EM, COBWEB, CLASSIT, AutoClass (Cheeseman and Stutz, 1996), … [10]

### 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.

### Các yêu cầu và những vấn đề còn tồn tại

1. **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.

1. **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.
* Kết quả của thuật toán phân cụm dữ liệu có thể được giải thích theo nhiều cách khác nhau. [7,10]

# CHƯƠNG 2: PHÂN TÍCH THUẬT TOÁN K-MEANS

## Tổng quan thuật toán K-Means

### Giới thiệu thuật toán

Thuật toán K-Means do MacQueen giới thiệu trong tài liệu “J. Some Methods for Classification and Analysis of Multivariate Observations” năm 1967. [6]

Là thuật toán gom cụm dữ liệu theo phương pháp phân hoạch. Là 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.

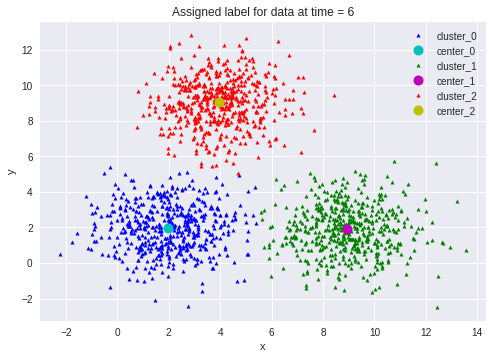
Coi mỗi thuộc tính của đối tượng (đối tượng có m thuộc tính) như một tọa độ của không gian m chiều và biểu diễn đối tượng như một điểm của không gian m chiều.

ai = (xi1, xi2, … xim)

ai (i = 1 … n) – đối tượng thứ i

xij (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. [16]



Hình 2.1 - Ví dụ mô phỏng phân cụm với thuật toán K-Means

### 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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X4 | X5 |
| C1 | 1 | 1 | 0 | 0 | 0 |
| C2 | 0 | 0 | 1 | 0 | 1 |
| C3 | 0 | 0 | 0 | 1 | 0 |

Bảng 2.1 - Minh họa 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.

Cluster i = {a1, a2, ...., at} – Cụm thứ i   
i = 1 … k,  k số cluster

j = 1 ... m, m số thuộc tính

t - số phần tử hiện có của nhóm thứ i

xsj - thuộc tính thứ j của phần tử s, s = 1 … t

cij - toạ độ thứ j của phần tử trung tâm nhóm i.

(1)

***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.

ai = (xi1, xi2, … xim) i = 1 … n – đối tượng thứ i cần phân loại.

cj = (xj1, xj2, … xjm) j = 1 … k – phần tử trung tâm nhóm j, cj được tính khoảng cách dựa trên công thức:

(2)

dij – khoảng cách Euclidean từ ai đến cj.

xis – thuộc tính thứ s của đối tượng ai.

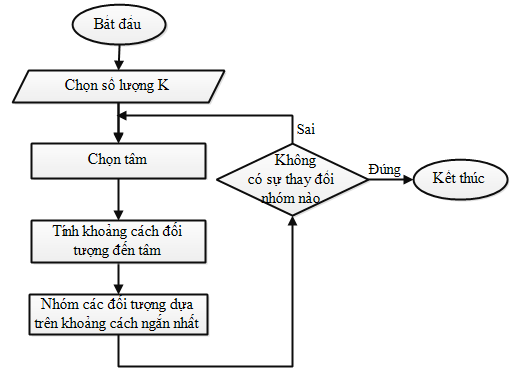
Xjs – thuộc tính thứ s của phần tử trung tâm cj.

### Mô tả thuật toán

***Ý tưởng***

Với mục tiêu chia tập gồm n đối tượng của cơ sở dữ liệu thành k cụm (k≤ n, k là số nguyên dương) sao cho các đối tượng trong cùng một vùng 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 {mj} kj = 1
* ***Đầu ra:*** Các cụm C[i] (1 ≤ i ≤ 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 {mj} kj = 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 ≤ i ≤ n), tính khoảng cách của nó tới mỗi trọng tâm mj (1 ≤ j ≤ 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 ≤ j ≤ k, cập nhật trọng tâm cụm mj 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:

*Trong đó:*

* n là số đối tượng dữ liệu
* k là số cụm dữ liệu
* d là số chiều
* là số vòng lặp
* là thời gian để thực hiện một phép tính cơ sở như phép nhân, chia, …

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. [16]

### Ví dụ về thuật toán

**Ví dụ 1:**

Cho tập dữ liệu gồm 6 học sinh, mỗi học sinh gồm hai thuộc tính điểm X1, X2. Sử dụng thuật toán K-Means để phân các đối tượng thành 2 cụm (k = 2)

|  |  |  |
| --- | --- | --- |
| Mã HS |  |  |
| A | 1 | 1 |
| B | 1,5 | 1,5 |
| C | 5 | 5 |
| D | 4 | 3 |
| E | 4 | 4 |
| F | 3,5 | 3 |

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

*Ta có* n = 6, k = 2

* Ta dùng công thức Euclidean để tính khoảng cách.
* Chia 6 điểm trên làm 2 cụm sao cho khoảng cách từ mỗi điểm đến trọng tâm của mỗi nhóm là gần nhất.

**Bước 1:** Khởi tạo tâm cụm:

A là tâm cụm (1, 1), D là tâm cụm (4, 3).

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

Khoảng cách của đối tượng A lần lượt đến C1, C2:

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | Khoảng cách Euclidean | |
| Mã HS | X1 | X2 | **(1, 1)** | **(4, 3)** |
| A | 1 | 1 | **0** | 3.61 |
| B | 1,5 | 1,5 | **0.71** | 2.92 |
| C | 5 | 5 | 5.66 | **2.24** |
| D | 4 | 3 | 3.61 | **0** |
| E | 4 | 4 | 4.24 | **1** |
| F | 3,5 | 3 | 3.2 | **0.5** |

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

* d (A, C1) < d (A, C2) 🡪 A thuộc cụm C1
* d (B, C1) < d (B, C2) 🡪 B thuộc cụm C1
* d (C, C2) < d (C, C1)🡪 C thuộc cụm C2
* d (D, C2) < d (D, C1)🡪 D thuộc cụm C2
* d (E, C2) < d (E, C1)🡪 E thuộc cụm C2
* d (F, C2) < d (F, C1)🡪 F thuộc cụm C2

Được thể hiện qua bảng sau:

|  |  |  |
| --- | --- | --- |
|  | LẦN 1 | |
|  | Cụm 1 | Cụm 2 |
| Mã HS | Gần C1 | Gần C2 |
| A | x |  |
| B | x |  |
| C |  | y |
| D |  | y |
| E |  | y |
| F |  | y |

* + - * Cụm C1, C2 lần lượt gồm các phần tử là:

Cụm 1 gồm có (A, B)

Cụm 2 gồm có (C, D, E, F)

**Bước 4**: Tính lại trọng tâm các cụm

=

=

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | Khoảng cách Euclidean | |
| Mã HS | X1 | X2 | **(1.25; 1.25)** | **(3,75 ; 4,125)** |
| A | 1 | 1 | **0,35** | 4,16 |
| B | 1,5 | 1,5 | **0,35** | 3,46 |
| C | 5 | 5 | 5,3 | **1,53** |
| D | 4 | 3 | 3,26 | **0,76** |
| E | 4 | 4 | 3,89 | **0,28** |
| F | 3,5 | 3 | 2,85 | **0,98** |

**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 |
| Mã HS | Gần C1 | Gần C2 |
| A | x |  |
| B | x |  |
| C |  | y |
| D |  | y |
| E |  | y |
| F |  | y |

* + - * Cụm C1, C2 lần lượt gồm các phần tử là:

Cụm 1 gồm có (B, D)

Cụm 2 gồm có (A, C, E, F)

**Bước 4**: Tính lại trọng tâm các cụm

|  |  |  |
| --- | --- | --- |
|  | X1 | X2 |
| C1 | 1.25 | 1.25 |
| C2 | 3,75 | 4,125 |

*Ta có:*

Cụm 1, trọng tâm C1 (1.25; 1.25)

Cụm 2, trọng tâm C2 (3,75; 4,125)

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 A, B; Cụm 2 gồm C, D, E, F.

***Kết luận:***

Cụm 1 gồm có A, B.

Cụm 2 gồm có C, D, E, F.

Với kết quả phân cụm trên sẽ giúp cho Ban Giám hiệu nhà trường có cái nhìn cụ thể hơn, chính xác hơn về học lực của các học sinh nhất là trong một khối hay toàn trường. Từ đó:

* Giúp cho Ban giám hiệu nhà trường sẽ điều chỉnh, đề ra kế hoạch giảng dạy mới phù hợp hơn với những học sinh này để có kết quả tốt hơn.
* Hỗ trợ Ban giám hiệu trong việc sắp xếp lớp học.
* Gợi ý cho việc chọn khối thi phù hợp.

## 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.

***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.

## 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. 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. [9]

## Thuật toán khắc phục nhiễu dữ liệu

***Thuật toán K-Medoids:***

Để khắc phục các trường hợp dữ liệu có nhiễu ta sử dụng thuật toán K-Medoids. Vì K-Medoids chịu ảnh hưởng ít hơn của nhiễu và các giá trị chênh lệnh so với giá trị trung bình.

Thuật toán K-Medoids tương tự thuật toán K-Means, mỗi cụm được đại diện bởi một trong số các đối tượng của cụm. Thông thường thì điểm gần vector trọng tâm sẽ được chọn làm điểm đại diện của cụm

***Ý tưởng:***

Để tìm ra K cụm với n đối tượng thì K – Medoids chọn ngẫu nhiên K đối tượng vào K cụm, coi mỗi đối tượng này là tâm của cụm. Phân bố các đối tượng còn lại vào cụm mà sự khác nhau của nó với đối tượng tâm của cụm là gần nhất.

Sau đó lặp lại quá trình: Thay đổi đối tượng tâm của mỗi cụm sao cho chất lượng của cụm được cải thiện. Chất lượng của cụm được lượng giá bởi một hàm đo sự khác nhau giữa một đối tượng và đối tượng tâm của cụm chứa nó. Quá trình lặp cho đến khi không còn sự thay đổi nào về phần tử cũng như hình dạng của các cụm.

Để chọn một đối tượng không là đối tượng tâm Orandom thay thế tốt cho một đối tượng tâm Oj thì mỗi đối tượng p xét theo 4 trường hợp sau đây:

*Trường hợp 1:* p đang phụ thuộc vào các cụm có tâm Oj (từ nay gọi là cụm Oj) được thay thế bởi Orandom và p gần nhất với Oi (i ≠ j) thì p được gán lại vào Oi.

*Trường hợp 2:* p đang thuộc vào Oj. Nếu Oj được thay thế bởi Orandom và p gần nhất với Orandom thì p được gán lại vào Orandom.

*Trường hợp 3:* p đang thuộc vào Oi (i ≠ j). Nếu Oj được thay thế bởi Orandom và p vẫn gần nhất với Oi thì không thay đổi gì cả. Tức là p vẫn thuộc Oi.

*Trường hợp 4:* p đang thuộc vào Oi (i ≠ j). Nếu Oj được thay thế bởi Orandom và p gần nhất với Orandom thì p được gán lại vào Orandom.

***Thuật toán K – Medoids được mô tả như sau:***

***Input****:* Số nguyên K và cơ sở dữ liệu gồm n đối tượng cần phân cụm.

***Output****:* Một tập gồm K cụm mà tổng giá trị của sự khác nhau của tất cả các đối tượng đến đối tượng tâm của nhóm chứa nó là nhỏ nhất.

***Các bước thực hiện thuật toán:***

***Bước 1:*** Chọn k đối tượng bất kì vào k cụm để làm tâm cụm.

***Bước 2:*** Lặp

***Bước 3:*** Gán mỗi đối tượng còn lại vào một cụm mà nó gần với đối tượng tâm cụm gần nhất.

***Bước 4:*** Chọn ngẫu nhiên một đối tượng không là đối tượng tâm, Orandom.

***Bước 5:*** Tính lại giá tri, S, đối với việc đổi Oj với Orandom.

***Bước 6:*** Nếu S < 0 thì đổi Oj với Orandom để tạo ra một tập với đối tượng tâm mới.

***Bước 7:*** Đến khi không có sự thay đổi nào nữa thì dừng.

***Nhận xét:*** Thuật toán K-Medoids mạnh hơn thuật toán K-Means trong các trường hợp dữ liệu có nhiễu vì K-Medoids chịu ảnh hưởng ít hơn của nhiễu và các giá trị chênh lệch so với giá trị trung bình. Tuy nhiên cả hai thuật toán này đều yêu cầu đưa vào số lượng cụm k.

# CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

## Giới thiệu bài toán

Ban giám hiệu của một trường Trung học Phổ thông nào đó. Sau một học kì hay một năm học, … khi nhìn bảng điểm của lớp, trường và muốn đưa ra một vài kết luận nào đó về tình hình học tập của học sinh trường mình phụ trách. Bạn sẽ làm thế nào nếu là một trong những thành viên của Ban giám hiệu?

**Đầu tiên**, chúng ta sẽ 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.

Ví dụ như học kỳ này điểm môn Toán của lớp 12A2 là 7.9, của lớp 12A4 là 7.6, như vậy có thể nhận xét ngay rằng học kỳ này học sinh lớp 12A2 học Toán tốt hơn lớp 12A4 và cả hai lớp đều học khá môn Toán.

Tuy nhiên theo em những nhận xét này chưa chính xác lắm vì có thể có những học sinh điểm rất cao, ngược lại cũng có điểm rất thấp. Rõ ràng tính trung bình rất đơn giản, nhưng từ điểm của hàng chục học sinh một lớp, tức là hàng chục 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 chục học sinh. Vì vậy nó đã 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 khả thi hơn. Chẳng hạn nếu lớp 12A2 có điểm kì vọng môn Toán là 7.9 và phương sai 2.5, lớp 12A4 có điểm kì vọng 7.6 nhưng phương sai 1.5, thì có kết luận là nhìn chung lớp 12A2 học Toán tốt hơn lớp 12A4 nhưng lớp 12A4 học đều môn Toán hơn lớp 12A2, … và có thể có nhiều kết luận khác nữa.

Với những trường hợp trên, chúng ta thấy rằng việc phân tích, đánh giá kết quả học tập học sinh không phải là chuyện đơn giản. Đòi hỏi Ban giám hiệu nhà trường, các nhà quản lý giáo dục có một sự đầu tư, nghiên cứu, tìm tòi, … nhằm đưa ra các phân tích, đánh giá đúng đắn nhất, chính xác nhất về kết quả học tập của học sinh từ đó đưa ra định hướng phát triển hơn cho nhà trường.

Ngoài các phương pháp truyền thống, trên thế giới đã và đang nghiên cứu, áp dụng rộng rãi các phương pháp có sử dụng đến phân cụm để gom nhóm các học sinh theo học lực. Một cách tổng quát, để phân cụm dữ liệu điểm của học sinh phụ thuộc chủ yếu vào các thuộc tính điểm và có nhiều phương pháp phân cụm như phương pháp phân cụm trên lưới, phân cụm dựa trên mô hình, phân cụm theo mật độ, … và để thực hiện phân cụm dữ liệu điểm em lựa chọn phương pháp phân cụm phân hoạch - một phương pháp phân cụm đơn giản và hiệu quả.

**Đầu vào**: Tập dữ liệu gồm 1000 học sinh trong học kỳ II trường THPT Long Trường. Mỗi học sinh gồm các thuộc tính Mã học sinh, Họ, Tên, Tên lớp, Điểm Toán, Điểm các môn học: Toán, Lý, Hóa, Sinh, Sử, Địa, Văn, Anh, Giáo dục công dân.

**Đầu ra**: Ứng dụng có khả năng phân cụm chính xác điểm các học sinh trong tập dữ liệu.

## 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 học kỳ II năm học 2018 - 2019 tại trường Trung học Phổ thông Long Trường. Tập dữ liệu mẫu Diemhocsinh.xlsx gồm 1000 mẫu dữ liệu với hơn 8 thuộc tính mô tả thông tin và điểm của học sinh.

|  |  |  |
| --- | --- | --- |
| Thứ tự | Thuộc tính | Tên thuộc tính |
| 1 | f\_mahs | Mã học sinh |
| 2 | f\_ho | Họ |
| 3 | f\_ten | Tên |
| 4 | f\_tenlop | Tên lớp |
| 5 | f\_toan | Điểm Toán |
| 6 | f\_ly | Điểm Lý |
| 7 | f\_hoa | Điểm Hóa |
| 8 | f\_sinh | Điểm Sinh |
| 9 | f\_su | Điểm Sử |
| 10 | f\_dia | Điểm Địa |
| 11 | f\_van | Điểm Văn |
| 12 | f\_anh | Điểm Anh |
| 13 | f\_gdcd | Điểm GDCD |

Bảng 3.1 - Bảng các thuộc tính của tập dữ liệu

## Môi trường thử nghiệm

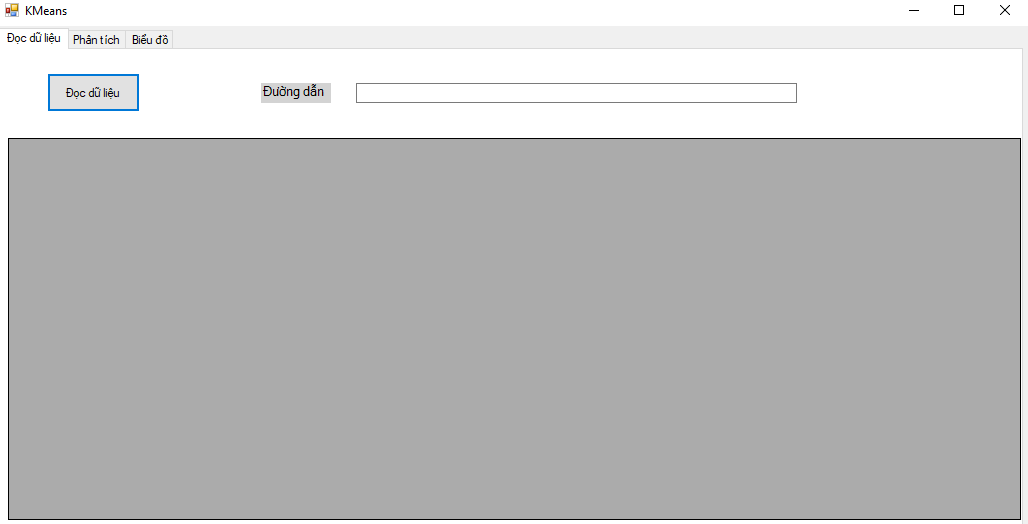
* Công nghệ: .NET Framework 4.6.
* Ngôn ngữ lập trình: C#.
* Công cụ: Visual Studio 2017, Microsoft Excel 2016.

## Giao diện chương trình

Chương trình gồm có các chức năng sau:

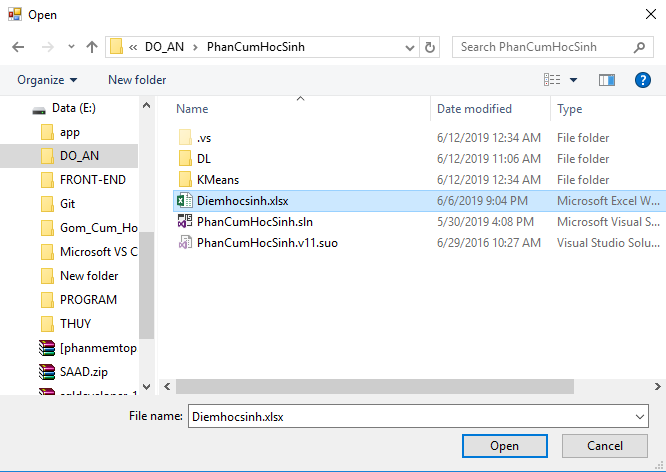
***Chức năng đọc dữ liệu:***

Giao diện đầu tiên khi chạy chương trình:



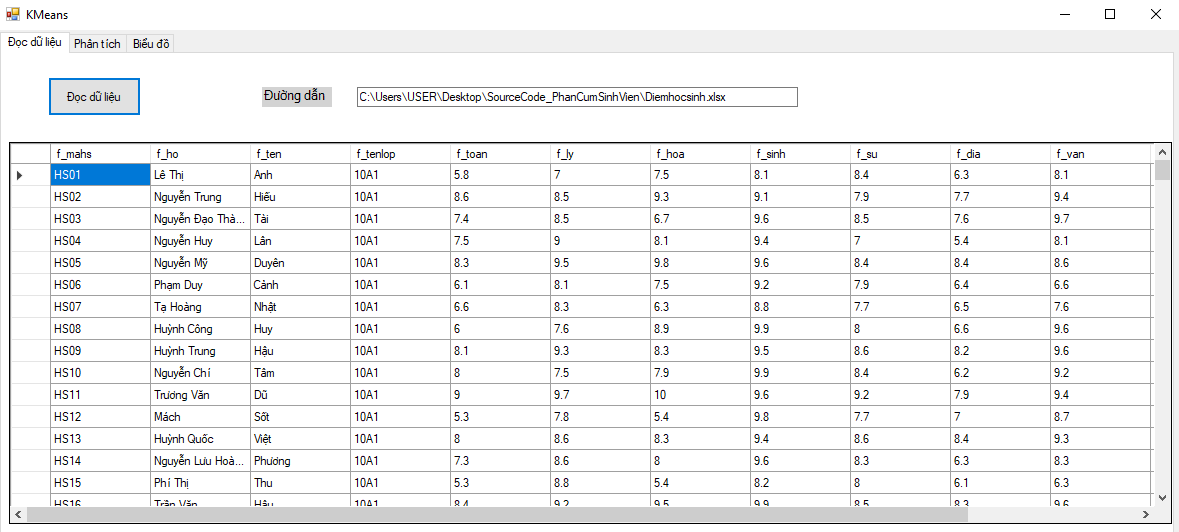
Hình 3.1 - Giao diện chức năng Đọc dữ liệu

Người dùng cần chọn file dữ liệu cần phân tích. Để đọc dữ liệu ta nhấp vào button “Đọc dữ liệu” sẽ hiện ra cửa sổ chọn file. Sau đó chọn file cần đọc và nhấp vào “Open” để mở file.



Hình 3.2 - Cửa sổ chọn file cần phân tích

Sau khi chọn file, dữ liệu sẽ được đọc và hiển thị như sau:



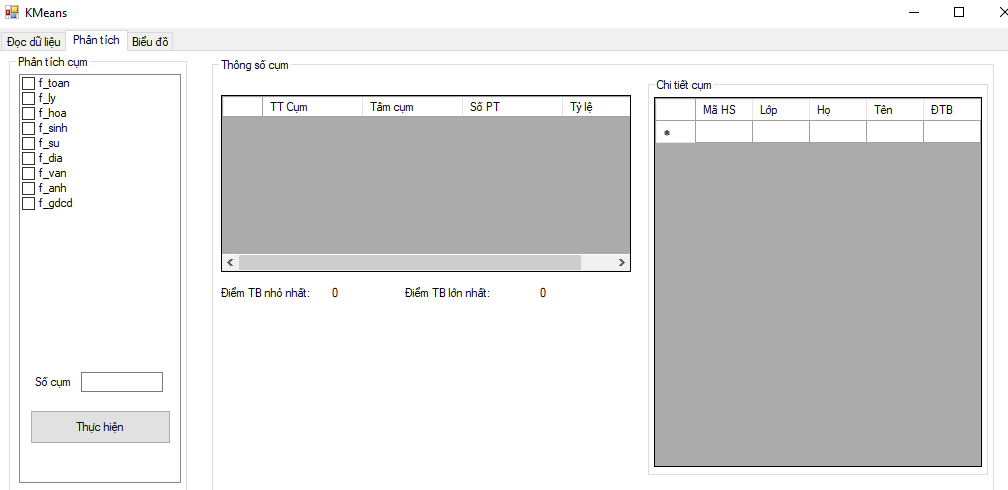
Hình 3.3 - Giao diện hiển thị dữ liệu

Lưu ý:

1. Tập tin Excel phải là định dạng từ Excel 2007 trở lên.
2. Các cột thông tin phải theo định dạng chuẩn (cột mã học sinh là “f\_mahs”, cột tên là “f\_ten”, cột điểm môn Toán là “f\_toan”, …

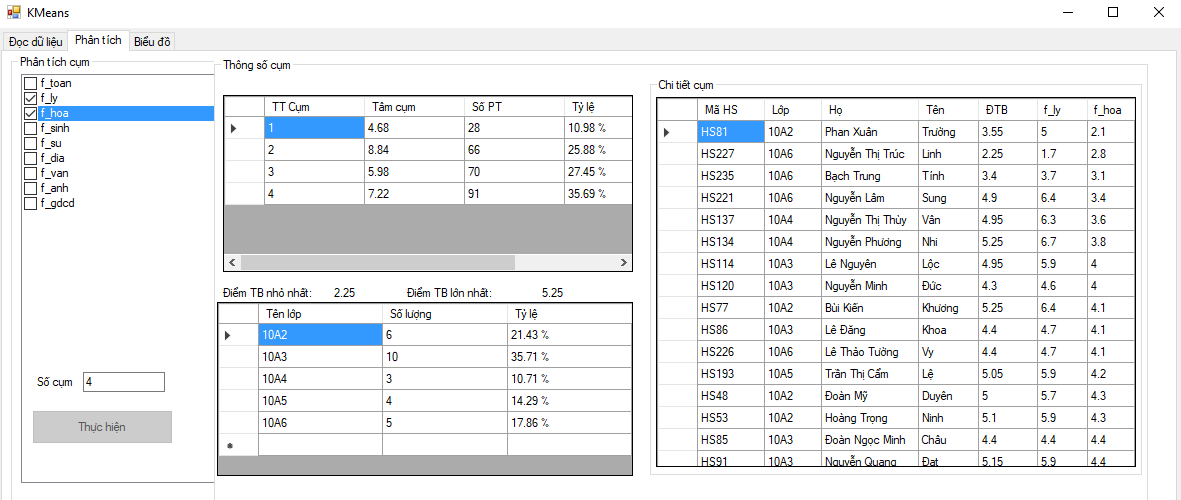
***Giao diện thực hiện phân cụm dữ liệu:***

Sau khi đọc được dữ liệu, bạn có thể chọn tab “Phân tích” để thực hiện chức năng phân tích. Để thực hiện, người dùng cần nhập vào số cụm, chọn các môn học và nhấn nút “Thực hiện” để phân tích dữ liệu.

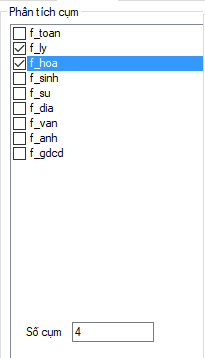


Hình 3.4 - Giao diện ban đầu chức năng Phân tích

* Phân cụm điểm trung bình các môn học cho phép chọn một hoặc nhiều môn học.
* Cho phép nhập số tâm cụm cần phân tích.
* Kết quả cho phép hiển thị thông số của các cụm.
* Hiển thị thông tin chi tiết của mỗi cụm: mã học sinh, lớp, họ và tên, điểm trung bình.

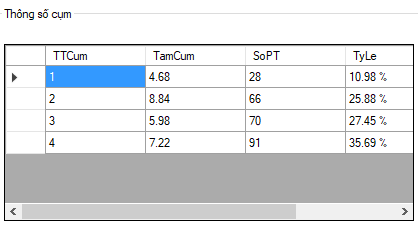


Hình 3.5 - Giao diện kết quả chọn môn Lý, Hóa, số cụm là 4



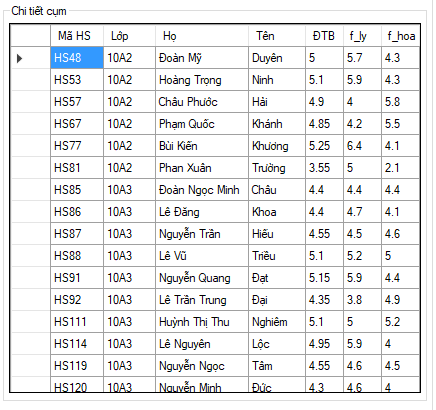
Hình 3.6 - Hình chọn số cụm và các môn học cần phân tích

(Do là ứng dụng demo nên số môn học còn khá hạn chế).

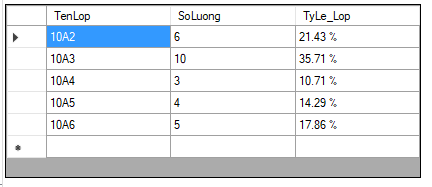


Hình 3.7 - Bảng thông số cụm sau khi phân tích.

Các thông số bao gồm: Thứ tự các cụm, tâm cụm, số lượng phần tử (học sinh), tỷ lệ phần trăm số lượng phần tử so với tổng số lượng học sinh trong danh sách.

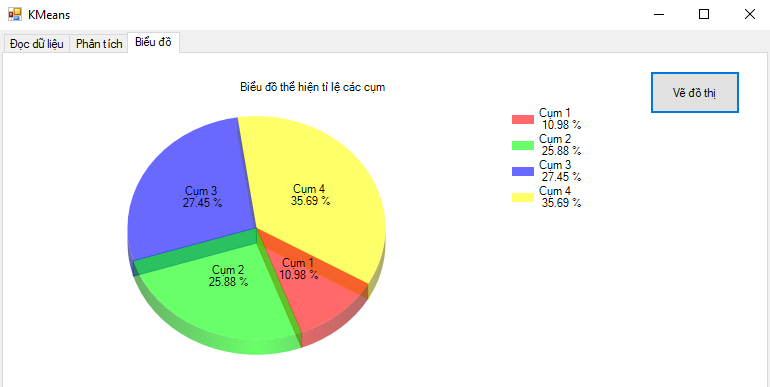


Hình 3.8 - Danh sách chi tiết các học sinh thuộc 1 cụm



Hình 3.9 - Danh sách các lớp có trong chi tiết cụm.

***Giao diện vẽ biểu đồ***



Hình 3.10 - Giao diện mô tả kết quả vừa phân cụm

## Kết quả và đánh giá

Ứng dụng phân cụm dữ liệu trong việc phân tích, đánh giá kết quả học tập của học sinh dựa trên điểm trung bình môn học, điểm trung bình học kỳ, bước đầu sẽ giúp ích cho Ban giám hiệu nhà trường, các nhà quản lý giáo dục có được một cái nhìn nhiều chiều hơn, đa dạng hơn về điểm số của học sinh từ đó thu được một số kết quả như: lựa chọn học sinh giỏi để bồi dưỡng, phát hiện học sinh yếu kém để phụ đạo, việc phân lớp, cũng như đề ra kế hoạch giảng dạy, tăng giờ tăng tiết, …

# 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 loại kết quả học tập của học sinh”, em đã đạt được kết quả như sau:

**Về mặt khoa học**

* Nắm bắt được các kiến thức cơ bản về 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.
* Nghiên cứu, tìm hiểu được giải pháp khắc phục nhiễu của thuật toán K-Means.

**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 kết quả học tập của học sinh 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. Nhằm giúp cho lãnh đạo Nhà trường có thể dựa vào để phân tích và đưa ra các biện pháp kịp thời và chính xác trong quản trị nhà trường.

**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.

1. **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.
* Giao diện ứng dụng chưa có tính thẩm mỹ cao.
* Tập dữ liệu mẫu được sử dụng còn hạn chế.

1. **Hướng phát triển**

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

* Tiếp tục nghiên cứu và hoàn thành chức năng phân cụm chi tiết để thể hiện rõ hơn các bước của quá trình phân cụm bằng thuật toán K-Means.
* Phát triển bài toán với số dữ liệu lớn hơn, bao quát hơn.
* Tìm hiểu cách sử dụng bộ công cụ DevExpress vào ứng dụng để giao diện trở nên đẹp 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. Andrew Troelsen Philip Japikse, *Pro C# 7 with .NET and .NET Core*, 2017.
2. Charu C. Aggarwal, *DATA MINING*, IBM T.J. Watson Research Center, 2015.
3. David Natingga, *Data Science Algorithms in a Week*: *Top 7 algorithms for computing, data analysis, and machine learning*, Birmingham- Mumbai, 2017.
4. Han Jiawei, *Data Mining: Concepts and Techniques 3nd Edition*, Morgan Kaufmann, 2011.
5. Jure Leskovec, Anand Rajaraman, Jeff Ullman, *Mining of Massive Datasets Chapter 7: Clustering,* Stanford University, 2011.
6. 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
7. 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.
8. 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.
9. Các chức năng của CLR, link:

<http://lifehack.vn/cong-nghe/phat-trien-ung-dung-web-net-nang-dong-de-thich-ung-va-hieu-qua/>, truy cập vào ngày 29/5/2019.

1. Kỹ thuật và ứng dụng của phân cụm, link:

<http://luanvan.net.vn/luan-van/luan-van-nghien-cuu-cac-ky-thuat-phan-cum-du-lieu-va-ung-dung-53110/> truy cập vào ngày 28/5/2019.

1. Tầm quan trọng của .Net Framework, link:

[http://bugnetproject.com/net-framework-la-gi-tai-sao-no-quan-trong-trong-moi- may-tinh/](http://bugnetproject.com/net-framework-la-gi-tai-sao-no-quan-trong-trong-moi-may-tinh/), truy cập vào ngày 26/5/2019

1. Tổng quan về khai phá dữ liệu, link:

https://www.academia.edu/19660657/khai\_phá\_dữ\_liệu, truy cập vào ngày 29/5/2019.

1. Tổng quan về .Net Framework, link:

<https://vi.wikipedia.org/wiki/.NET_Framework>, truy cập vào ngày 25/5/2019.

1. Tổng quan ngôn ngữ C#, link:

<https://voer.edu.vn/c/ngon-ngu-c/cf37fa1e/383e2f05>, truy cập vào ngày: 29/5/2019.

1. Ứng dụng thuật toán K-Means, link:

<https://kipalog.com/posts/Thuat-toan-Kmean-va-ung-dung>, truy cập vào ngày 28/5/2019.

1. Tìm hiểu về thuật toán K-Means, link:

<http://ungdung.khoa-hnvd.com/Hoc_thuat/KMeans.html>, truy cập vào ngày 29/5/2019.