**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

****

**BÁO CÁO ĐỒ ÁN CUỐI KÌ**

**KHAI PHÁ DỮ LIỆU**

Tên đề tài:

**SỬ DỤNG RAPIDMINER VÀ ÁP DỤNG THUẬT TOÁN TẬP THÔ CHO BÀI TOÁN KHAI PHÁ DỮ LIỆU**

**Giảng viên hướng dẫn:**

* **ThS. Trịnh Minh Tuấn**
* **GV. Nguyễn Hồ Duy Trí**

**Lớp: IS252.K13**

**Sinh viên thực hiện: Nhóm 11**

* **Nguyễn Ngọc Công – 15520069**
* **Nguyễn Xuân Hội - 15520268**

*TP. Hồ Chí Minh, tháng 12 năm 2019*

# MỤC LỤC

[MỤC LỤC 2](#_Toc27286551)

[LỜI CÁM ƠN 4](#_Toc27286552)

[NHẬN XÉT CỦA GIẢNG VIÊN 5](#_Toc27286553)

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 6](#_Toc27286554)

[1. Lý do chọn đề tài 6](#_Toc27286555)

[2. Dữ liệu 6](#_Toc27286556)

[3. Mô tả bài toán 8](#_Toc27286557)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 8](#_Toc27286558)

[1. Tập thô 8](#_Toc27286559)

[2. Hệ thông tin 8](#_Toc27286560)

[2.1. Bảng dữ liệu 8](#_Toc27286561)

[2.2. Bảng quyết định 9](#_Toc27286562)

[CHƯƠNG 3: ỨNG DỤNG 10](#_Toc27286563)

[1. Tiền xử lý dữ liệu 10](#_Toc27286564)

[1.1. Xử lý dữ liệu bị thiếu 10](#_Toc27286565)

[1.2. Thực hiện khử nhiễu bằng kỹ thuật Bining và làm trơn smoothing 10](#_Toc27286566)

[1.3. Kết quả 12](#_Toc27286567)

[2. Thuật toán khai thác dữ liệu 12](#_Toc27286568)

[3. Kết quả đạt được 12](#_Toc27286569)

[4. Chương trình ứng dụng 12](#_Toc27286570)

[CHƯƠNG 4: KẾT LUẬN 12](#_Toc27286571)

[1. Ưu điểm 12](#_Toc27286572)

[2. Hạn chế 12](#_Toc27286573)

[3. Phát triển 12](#_Toc27286574)

[Bảng phân công công việc của các thành viên trong nhóm 12](#_Toc27286575)

[Bảng đánh giá chéo 12](#_Toc27286576)

[Tài liệu tham khảo 12](#_Toc27286577)

# LỜI CÁM ƠN

Đầu tiên, nhóm em xin gởi lời cảm ơn chân thành đến tập thể quý Thầy Cô Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM và đặc biệt là quý Thầy Cô khoa Hệ thống thông tin đã giúp chúng em có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này.

Đặc biệt, xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới Thầy ThS. Trịnh Minh Tuấn (Giảng viên lý thuyết môn Khai thác dữ liệu) và Anh Nguyễn Hồ Duy Trí (Giảng viên hướng dẫn thực hành môn Khai phá dữ liệu). Các Thầy Cô đã trực tiếp hướng dẫn tận tình, sửa chữa và đóng góp nhiều ý kiến quý báo giúp nhóm tác giả hoàn thành tốt báo cáo môn học của mình.

Trong thời gian một học kỳ thực hiện đề tài, chúng em đã vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới. Từ đó, nhóm tác giả vận dụng tối đa những gì đã thu thập được để hoàn thành một báo cáo đồ án tốt nhất. Tuy nhiên, trong quá trình thực hiện, nhóm tác giả không tránh khỏi những thiếu sót. Chính vì vậy, nhóm tác giả rất mong nhận được những sự góp ý từ phía các Thầy Cô nhằm hoàn thiện những kiến thức mà nhóm tác giả đã học tập và là hành trang để nhóm tác giả thực hiện tiếp các đề tài khác trong tương lai.

Xin chân thành cảm ơn các quý Thầy Cô!

Nhóm sinh viên thực hiện

Nguyễn Ngọc Công

Nguyễn Xuân Hội

# NHẬN XÉT CỦA GIẢNG VIÊN

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

## Lý do chọn đề tài

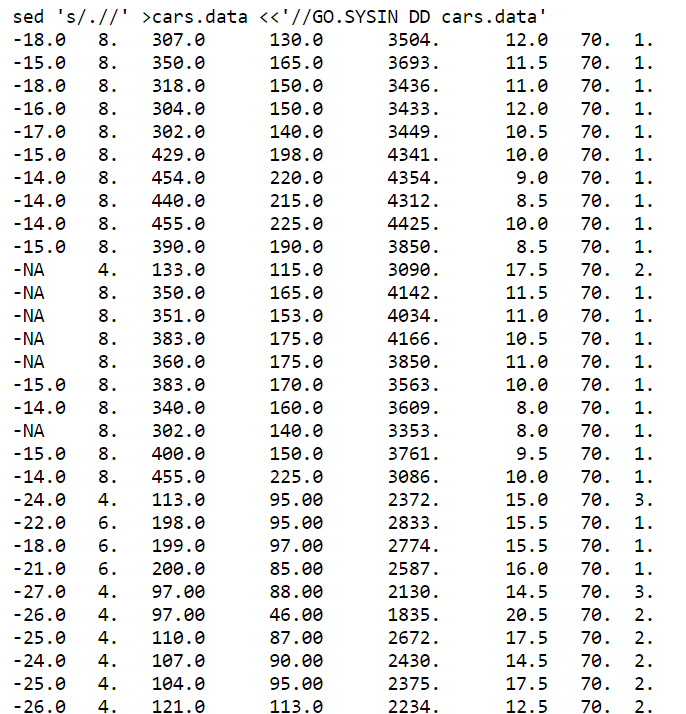
Sau những kiến thức tiếp thu từ những tiết lý thuyết và thực hành được giảng dạy tại lớp nhóm chúng tôi đã quyết định áp dụng những kiến thức này để kiểm chứng lý thuyết cho những dữ liệu đầu vào và kết quả sau quá trình sử dụng công cụ RapidMiner. Ngoài ra nhóm chúng tôi đã phát triển một thuật toán áp dụng lý thuyết tập thô cho bài toán khai phá dữ liệu để từ đó hiểu rõ hơn về những vấn đề này.

Tóm tắt những phần thực hiện trong đề tài:

* Tiền xữ lý dữ liệu
* Sử dụng công cụ RapidMiner
* Phát triển thuật toán áp dụng lý thuyết tập thô
* Hoàn thành báo cáo: cơ sở lý thuyết, kết quả đạt được, kết luận, bảng phân công công việc, …

## Dữ liệu

Bộ dữ liệu của nhóm chúng tôi sử dụng được công bố từ Hiệp hội Thống kê Hoa Kỳ (ASA) vào năm 1983, được thu thập bởi Ernesto Ramos và David Donoho. Dữ liệu bao gồm 406 dòng và 8 thuộc tính.



Hình Dữ liệu đầu vào của bài toán khai phá dữ liệu

**Mô tả các thuộc tính:**

* MPG: Miles per gallon – Số dặm mà xe chạy được trên mỗi galon xăng
* Cylinders: Số xi-lanh của xe
* Engine displacement: Dung tích động cơ (cu. Inches)
* Horsepower: Mã lực
* Vehicle weight: Trọng lượng xe (lbs)
* Acceleration: Thời gian để xe đạt được vận tốc từ 0 đến 60 dặm ( giây)
* Model year: Năm của dòng xe
* Origin of car: Nguồn gốc xe (1. American, 2. European, 3. Japanese)

## Mô tả bài toán

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## Tập thô

Lý thuyết tập thô (rough set theory) lần đầu tiên được đề xuất bởi Z. Pawlak và nhanh chóng được xem như một công cụ xử lý các thông tin mơ hồ và không chắc chắn. Phương pháp này đóng vai trò hết sức quan trọng trong lĩnh vực trí tuệ nhận tạo và các ngành khoa học khác liên quan đến nhận thức, đặc biệt là lĩnh vực máy học, thu nhận tri thức, phân tích quyết định, phát hiện và khám phá tri thức từ cơ sở dữ liệu, các hệ chuyên gia, các hệ hỗ trợ quyết định, lập luận dựa trên quy nạp và nhận dạng.

Lý thuyết tập thô dựa trên giả thiết rằng để định nghĩa một tập hợp, chúng ta cần phải có thông tin về mọi đối tượng trong tập vũ trụ. Ví dụ, nếu các đối tượng là những bệnh nhân bị một bệnh nhất định thì các triệu chứng của bệnh tạo thành thông tin về bệnh nhân. Như vậy tập thô có quan điểm hoàn toàn khác với quan điểm truyền thống của tập hợp, trong đó mọi tập hợp đều được định nghĩa duy nhất bởi các phần tử của nó mà không cần biết bất kỳ thông tin nào về các phần tử của tập hợp. Rõ ràng, có thể tồn tại một số đối tượng giống nhau ở một số thông tin nào đó, và ta nói chúng có quan hệ bất khả phân biệt với nhau. Đây chính là quan hệ mấu chốt và là điểm xuất phát của lý thuyết tập thô : biên giới của tập thô là không rõ ràng, và để xác định nó chúng ta phải đi xấp xỉ nó bằng các tập hợp khác nhằm mục đích cuối cùng là trả lời được (tất nhiên càng chính xác càng tốt) rằng một đối tượng nào đó có thuộc tập hợp hay không. Lý thuyết tập thô với cách tiếp cận như vậy đã được ứng dụng trong rất nhiều lĩnh vực của đời sống xã hội.

## Hệ thông tin

### Bảng dữ liệu

Một tập dữ liệu có thể biểu diễn dưới dạng một bảng, trên đó mỗi dòng biểu diễn thông tin ứng với một đối tượng, mỗi cột biểu diễn một thuộc tính có thể đo được của đối tượng. Bảng này được gọi là một hệ thống thông tin.

Hệ thống thông tin là một cặp S = (U,A), trong đó U là tập hữu hạn khác rỗng các đối tượng gọi là tập vũ trụ còn A là tập hữu hạn khác rỗng các thuộc tính.

Với mỗi đối tượng u ∈ U và thuộc tính a ∈ A , ta ký hiệu u(a) là giá trị của đối tượng u tại thuộc tính a. Nếu gọi Ia là tập tất cả các giá trị của thuộc tính a, thì u(a) ∈ Ia với mọi u ∈ U. Bây giờ, nếu đặt B = {b1, b2, …, bk} là tập con của tập thuộc tính A, ta ký hiệu bộ các giá trị u(bi) bởi u(B). Như vậy, nếu u và v là 2 đối tượng u(B)=v(B) nếu u(bi)=v(bi), với mọi i=1,2,…,k.

Ví dụ: Một hệ thống thông tin bao gồm 8 đối tượng U={u1,u2,u3,u4,u5,u6,u7,u8}, tập thuộc tính A={Giới tính, Cân nặng }, và miền giá trị cho từng thuộc tính là IGiới tính = {Nam, Nữ}, ICân nặng = {Thấp, Trung Bình, Cao}.

|  |  |  |
| --- | --- | --- |
|  | **Giới tính** | **Cân nặng (kg)** |
| 1 | Nam | 92 |
| 2 | Nam | 65 |
| 3 | Nữ | 45 |
| 4 | Nam | 55 |
| 5 | Nam | 75 |
| 6 | Nữ | 50 |
| 7 | Nữ | 52 |
| 8 | Nữ | 42 |

*Bảng 1: Bảng dữ liệu*

### Bảng quyết định

Bảng quyết định là một hệ thống thông tin có dạng T=(U,A),với U là tập các đối tượng và A là tập các thuộc tính, trong đó tập thuộc tính A được chia thành 2 tập thuộc tính rời nhau là C và D, C được gọi là tập thuộc tính điều kiện và D là tập thuộc tính quyết định. Tức là T = (U, C ∪ D).

Ví dụ: Bảng sau đây là một bảng quyết định. Bảng này có 8 đối tượng như trong bảng 1, nhưng có thêm thuộc tính quyết định (Kết quả). Trong bài toán phân lớp thì thuộc tính quyết định chính là lớp của đối tượng cần xếp lớp. Trong ví dụ này thuộc tính quyết định Quyết định có 2 giá trị là Mua, Không mua.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Giới tính** | **Cân nặng (kg)** | **Quyết định** |
| 1 | Nam | 92 | Mua |
| 2 | Nam | 65 | Không |
| 3 | Nữ | 45 | Không |
| 4 | Nam | 55 | Không |
| 5 | Nam | 75 | Mua |
| 6 | Nữ | 60 | Mua |
| 7 | Nữ | 52 | Mua |
| 8 | Nữ | 42 | Không |

*Bảng 2: Bảng quyết định*

# CHƯƠNG 3: ỨNG DỤNG

## Tiền xử lý dữ liệu

### Xử lý dữ liệu bị thiếu

* Đối với thuộc tính có dữ liệu là số:
  + Những thuộc tính phù hợp: MPG, Engine displacement, Horsepower, Vehicle weight, Acceleration.
  + Tính giá trị trung bình trên tất cả các giá trị có được của thuộc tính trên.
  + Dùng giá trị trung bình này để điền vào dữ liệu thiếu và làm tương tự với các thuộc tính dữ liệu số còn lại.
* Đối với các thuộc tính còn lại:
  + Những thuộc tính còn lại: Cylinders, Model year, Origin of car.
  + Dùng giá trị trung bình này để điền vào dữ liệu thiếu và làm tương tự với các thuộc tính dữ liệu số còn lại.

### Thực hiện khử nhiễu bằng kỹ thuật Bining và làm trơn smoothing

Thực hiện khử nhiễu và làm trơn với những thuộc tính có kiểu dữ liệu là số: MPG, Engine displacement, Horsepower, Vehicle weight.

Sau khi dùng kỹ thuạt Bining, làm trơn và rời rạc hóa dữ liệu, ta được bảng kết quả như sau:

1. MPG

|  |  |
| --- | --- |
| Nhãn khái niệm | Giá trị (dặm trên gallon xăng) |
| Low | MPG < 20 |
| Medium | 20 <= MPG <= 30 |
| High | MPG > 30 |

Bảng Chia giỏ cho thuộc tính MPG

1. Engine displacement

|  |  |
| --- | --- |
| Nhãn khái niệm | Giá trị (cu. inches) |
| Low | Engine displacement < 200 |
| Medium | 200 <= Engine displacement <= 300 |
| High | Engine displacement >300 |

Bảng Chia giỏ cho thuộc tính Engine displacement

1. Horsepower

|  |  |
| --- | --- |
| Nhãn khái niệm | Giá trị |
| Low | Horsepower < 100 |
| Medium | 100 <= Horsepower <= 150 |
| High | Horsepower > 150 |

Bảng Chia giỏ cho thuộc tính Horsepower

1. Vehicle weight

|  |  |
| --- | --- |
| Nhãn khái niệm | Giá trị (lbs) |
| Low | Vehicle weight < 3000 |
| Medium | 3000 <= Vehicle weight <= 4000 |
| High | Vehicle weight > 4000 |

Bảng Chia giỏ cho thuộc tính Vehicle weight

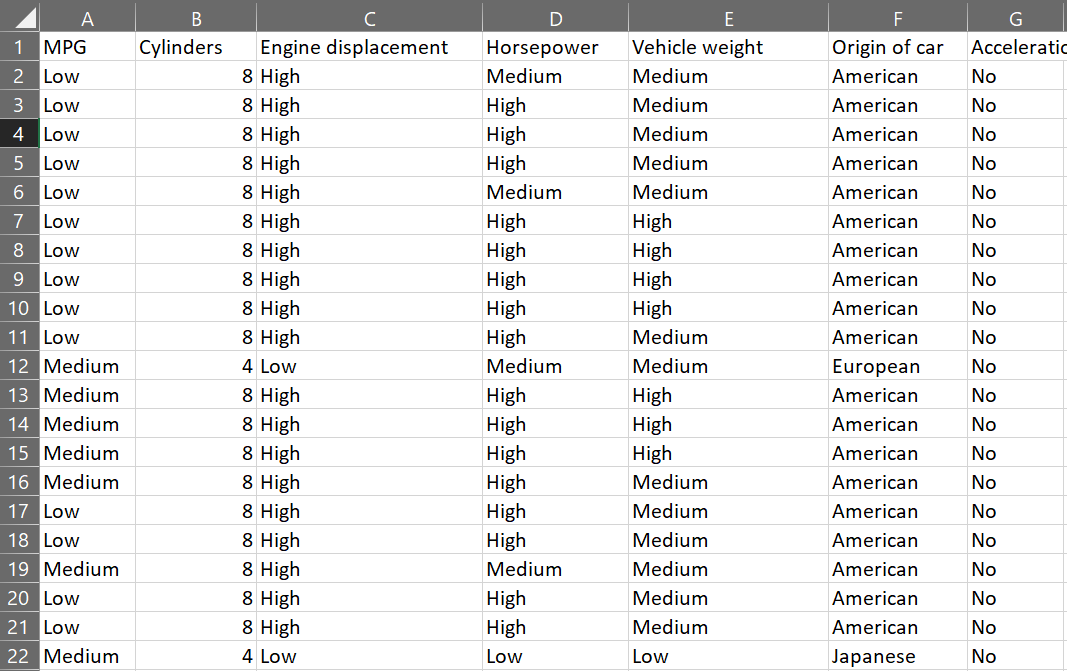
1. Acceleration

|  |  |
| --- | --- |
| Nhãn khái niệm | Giá trị (giây) |
| Yes | Acceleration < 20 |
| No | 20<= Acceleration <= 30 |

Bảng Chia giỏ cho thuộc tính Acceleration

### Kết quả

Kết quả của quá trình tiền xữ liệu được lưu dưới dạng file CSV bao gồm 7 cột trong đó có Acceleration là thuộc tính quyết định. Trong file CSV kết quả sẽ không còn dữ liệu thiếu sót và tất cả dữ liệu thể hiện những thông tin về xe o tô khá rõ ràng.



Hình Kết quả của quá trình tiền xữ lý dữ liệu

## Sử dụng công cụ RapidMiner

### Áp dụng lý thuyết 1

- Hội viết -

### Áp dụng lý thuyết 2

- Hội viết -

## Thuật toán khai thác dữ liệu

Thuật toán áp dụng

Sơ đồ luồng dữ liệu

Dữ liệu đầu vào

Xử lý

Kết quả

## Kết quả đạt được

Sau quá trình áp dụng kiến thức vào thực tiễn …

Kết quả đạt được khi sử dụng công cụ RapidMiner:

* --- Hội viết -----

Kết quả đạt được khi phát triển thuật toán áp dụng lý thuyết tập thô:

* --- Công viết -----

## Chương trình ứng dụng

Github lưu trữ

# CHƯƠNG 4: KẾT LUẬN

## Ưu điểm

Đề tài hoàn thành có một số ưu điểm sau đây:

* Công cụ RapidMiner
* Thuật toán được phát triển đã cho ra kết quả như mong đợi, hiệu suất thuật toán nhanh, dễ triển khai.

## Hạn chế

Vì thời gian là có giới hạn nên đề tài của nhóm chúng tôi có những hạn chế sau đây:

* Công cụ RapidMiner
* Thuật toán còn sơ khai, chưa ứng dụng được vài tất cả dữ liệu.

## Phát triển

Để thuật toán tối ưu hơn chúng tôi phải cải tiến hơn về việc hoạt động tốt khi bộ dữ liệu lớn hơn. Và độ chính xác của thuật toán cũng là một vấn đề quan trọng khi tiếp tục phát triển đề tài.

# Bảng phân công công việc của các thành viên trong nhóm(% tham gia)

|  |  |  |
| --- | --- | --- |
|  | Nguyễn Ngọc Công | Nguyễn Xuân Hội |
| Tìm kiếm dữ liệu | 70% | 30% |
| Tìm hiểu những thuật toán để triển khai | 30% | 70% |
| Viết thuật toán | 80% | 20% |
| Sử dụng công cụ RapidMiner | 20% | 80% |
| Viết báo cáo | 50% | 50% |

Bảng phân chia công việc

# Bảng đánh giá chéo

# Tài liệu tham khảo

<http://lib.stat.cmu.edu/datasets/cars.desc>

<http://lib.stat.cmu.edu/datasets/cars.data>