HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

TRẦN ANH VIỆT

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP PHÂN TÍCH DỮ LIỆU TRÊN BẢNG QUYẾT ĐỊNH TRONG HỆ THỐNG DỮ LIỆU LỚN

Chuyên ngành: Hệ thống Thông tin

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2019

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **GS.TS VŨ ĐỨC THI**

Phản biện 1: TS. Nguyễn Duy Phương

Phản biện 2: PGS.TS Nguyễn Hải Châu

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

1. Lý do chọn đề tài

Các hệ thống dữ liệu lớn cũng như các phương pháp phân tích dữ liệu lớn đã được nhiều nhà khoa học quan tâm nghiên cứu. Hướng phân tích dữ liệu trên các bảng quyết định mà cụ thể là nghiên cứu các bài toán liên quan đến tập rút gọn trên bảng quyết định phát triển rất sôi động có nhiều ứng dụng trong thực tiễn.

Trong những năm gần đây, sự phát triển mạnh mẽ của công nghệ thông tin đã làm cho khả năng thu thập và lưu trữ thông tin của hệ thống thông tin tăng nhanh một cách nhanh chóng. Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết là cần có những kỹ thuật và công cụ mới để tự động chuyển đổi lượng 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 thời sự 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: Market tinh, tài chính ngân hàng và bảo hiểm, khoa học kinh tế…Rất nhiều tổ chức và công ty lớn trên thế giới đã áp dụng kỹ thuật khai phá dữ liệu vào các hoạt động sản xuất kinh doanh của mình và thu được nhiều lợi ích to lớn.

Trong lý thuyết tập thô, dữ liệu được biểu diễn thông qua một hệ thông tin $IS=(U,A)$ với U là tập các đối tượng và A là tập thuộc tính. Phương pháp tiếp cận chính của lý thuyết tập thô là dựa trên quan hệ không phân biệt được để đưa ra các tập xấp xỉ dưới và xấp xỉ trên của nó. Xấp xỉ dưới bao gồm các đối tượng chắc chắn thuộc tập đó, còn xấp xỉ trên chứa tất cả các đối tượng có khả năng thuộc về tập đó. Nếu tập xấp xỉ dưới bằng tập xấp xỉ trên thì tập đối tượng cần quan sát là tập rõ. Ngược lại là tập thô. Các tập xấp xỉ là cơ sở để đưa ra các kết luận từ tập dữ liệu. Bảng quyết định là hệ thông tin IS với tập thuộc tính A được chia thành hai tập con khác rỗng rời nhau C và D, lần lượt được gọi là tập thuộc tính điều kiện và tập thuộc tính quyết định. Nói cách khác, $DS=(U,C\cup D)$ với $C\cap D\not\varnothing$. Bảng quyết định là mô hình thường gặp trong thực tế, Khi mà giá trị dữ liệu tại

các thuộc tính điều kiện có thể cung cấp cho ta thông tin về giá trị của thuộc tính quyết định. Bảng quyết định là nhất quán khi phụ thuộc hàm $C \rightarrow D$ là đúng, trái lại là không nhất quán.

Rút gọn thuộc tính là ứng dụng quan trọng nhất trong lý thuyết tập thô. Mục tiêu của rút gọn thuộc tính là loại bỏ các thuộc tính dư thừa để tìm ra các thuộc tính cốt yếu và cần thiết trong cơ sở dữ liệu. Với bảng quyết định, rút gọn thuộc tính là tập con nhỏ nhất của tập thuộc tính điều kiện bảo toàn thông tin phân lớp của bảng quyết định. Đối với một bảng quyết định có nhiều tập rút gọn khác nhau tuy nhiên trong thực hành thường không đòi hỏi tìm tất cả các tập rút gọn mà chỉ cần tìm được một tập rút gọn tốt nhất theo một tiêu chuẩn đánh giá nào đó là đủ. Vì vậy, mỗi phương pháp rút gọn thuộc tính đều trình bày một thuật toán Heuristic tìm tập rút gọn. Các thuộc tính này giảm thiểu đáng kể khối lượng tính toán, nhờ đó có thể áp dụng đối với các bài toán có khối lượng dữ liệu lớn.

Cho bảng quyết định nhất quán $DS=(U,C \cup \{d\})$, tập thuộc tính $R \subseteq C$ được gọi là tập rút gọn của thuộc tính điều kiện C nếu R là tập tối thiểu thỏa mãn phụ thuộc hàm $R \rightarrow \{d\}$. Xét quan hệ r trên tập thuộc tính $R \subseteq C\{d\}$ được gọi là một tập tối thiểu của thuộc tính $\{d\}$ nếu R là tập thuộc tính tối thiểu thỏa mãn phụ thuộc hàm $R \rightarrow \{d\}$. Do đó, khái niệm tập rút gọn của bảng quyết định tương đương với tập tối thiểu của thuộc tính $\{d\}$ trên quan hệ, và một vài bài toán trên bảng quyết định liên quan đến tập rút gọn có thể được giải quyết bằng một số kết quả liên quan đến tập tối thiểu của một thuộc tính trong cơ sở dữ liệu quan hệ; bao gồm bài toán tìm tập tất cả các thuộc tính rút gọn, bài toán tìm họ tất cả các tập rút gọn, bài toán trích lọc tri thức dưới dạng các phụ thuộc hàm từ bảng quyết định, bài toán xây dựng bảng quyết định từ tập phụ thuộc hàm cho trước. Cho đến nay, hướng tiếp cận này chưa được nhiều tác giả quan tâm nghiên cứu.

Trên bảng quyết định nhất quán, vấn đề nhiên cứu đặt ra là xây dựng các thuật toán có ý nghĩa liên quan đến tập rút gọn sử dụng một số kết quả liên quan đến tập tối thiểu của một thuộc tính trong một cơ sở dữ liệu quan hệ.

2. Tổng quan về vấn đề nghiên cứu

Nhiều chính phủ quốc gia như Hoa Kỳ cũng đã rất quan tâm đến dữ liệu lớn. Trong tháng 3 năm 2012, chính quyền Obama đã công bố một khoản đầu tư 200 triệu USD để khởi động "Kế hoạch Nghiên cứu và Phát triển Big Data", mà đã là một sáng kiến phát triển khoa học và công nghệ chủ yếu thứ hai sau khi "xa lộ thông tin" bắt đầu vào năm 1993. Trong tháng 7 năm 2012 , dự án "Đẩy mạnh công nghệ thông tin Nhật Bản" được ban hành bởi Bộ Nội vụ và Truyền thông Nhật Bản chỉ ra rằng sự phát triển Big Data, nên có một chiến lược quốc gia và các công nghệ ứng dụng nên là trọng tâm. Trong tháng 7 năm 2012, Liên Hiệp Quốc đã đưa ra báo cáo *Big Data cho phát triển*, trong đó tóm tắt cách các chính phủ sử dụng Big Data để phục vụ tốt hơn và bảo vệ người dân của họ như thế nào.

Hiện nay, mặc dù tầm quan trọng của Big Data đã được thừa nhận rộng rãi. Xong vấn đề then chốt trong việc xử lí các hệ thống Big Data là nghiên cứu phát triển các phương pháp phân tích dữ liệu mà thực chất là khai phá các hệ thống dữ liệu lớn để phát hiện tri thức. Luận văn này nghiên cứu tìm hiểu một số phương pháp phân tích dữ liệu liên quan đến các tập rút gọn trên cấu trúc bảng quyết định sử dụng lí thuyết tập thô**.**

3. Mục đích nghiên cứu

Nghiên cứu và tìm hiểu một số nền tảng của hệ thống dữ liệu lớn. Tìm hiểu một số lĩnh vực phân tích tìm các giá trị của hệ thống dữ liệu lớn *(thực chất là khai phá dữ liệu tìm các tri thức)*.

Nghiên cứu và tìm hiểu một số thuật toán liên quan đến tập rút *gọn (tập thuộc tính rút gọn bảo toàn thông tin phân lớp của bảng quyết định)*. Trên cơ sở này tiến hành xây dựng phần mềm thử nghiệm.

4. Đối tượng và phạm vi nghiên cứu

Nghiên cứu và tìm hiểu các tài liệu liên quan đến hệ thống dữ liệu lớn. Phạm vi nghiên cứu tập trung vào các nền tảng của hệ thống dữ liệu lớn bao gồm những định nghĩa, các đặc trưng, sự phát triển của Big Data và những thách thức mà Big Data mang lại. Các phương pháp phân tích dữ liệu nói chung và phân tích dữ liệu trên các bảng quyết định liên quan đến các tập rút gọn dùng để phân lớp dữ liệu.

Các thuật toán cơ bản nhất liên quan đến tập rút gọn trên bảng quyết định nhất quán.

5. Phương pháp nghiên cứu

Ban đầu thu thập tài liệu Thu thập, tổng hợp các tư liệu, bài báo khoa học đã công bố, tham khảo, so sánh và phân tích để tìm ra vấn đề phù hợp phục vụ cho đề tài nghiên cứu; nghiên cứu tìm hiểu các nền tảng của hệ thống dữ liệu lớn, đặc biệt các phương pháp phân tích dữ liệu trên các bảng quyết định. Cuối cùng xây dựng một phần mềm thực nghiệm.

CHƯƠNG 1: NGHIÊN CỨU CÁC NỀN TẢNG CỦA HỆ THỐNG DỮ LIỆU LỚN

1. **Nghiên cứu một số nền tảng của hệ thống dữ liệu lớn (BigData)**
   - ***1.1 Định nghĩa mô tả và các đặc trưng của Dữ liệu lớn(BigData)***
   - ***1.2 Sự phát triển của BigData và các Công nghệ liên quan***
   - ***1.3 Các thách thức đối với BigData***
   - ***1.4 Các phương pháp tiền xử lý dữ liệu cho BigData***
   - ***1.5 Các hướng ứng dụng chính của BigData***
2. **Nghiên cứu một số lĩnh vực phân tích của Big Data**
3. **Kết luận chương**

CHƯƠNG 2: NGHIÊN CỨU MỘT SỐ CÁC PHƯƠNG PHÁP PHÂN TÍCH DỮ LIỆU TRÊN BẢNG QUYẾT ĐỊNH

2.1 Nghiên cứu khái quát hướng khai phá dữ liệu sử dụng lý thuyết tập thô

2.1.1 Những khái niệm cơ bản trong lý thuyết tập thô

2.1.2 Mô hình tập thô truyền thống

2.2 Nghiên cứu phân tích một số thuật toán liên quan đến tập rút gọn trong bảng quyết định rút gọn nhất quán:

2.2.1 Đặt vấn đề

2.2.2 Thuật toán tìm tất cả các thuộc tính rút gọn

2.2.3 Thuật toán tìm một tập rút gọn

2.2.4 Thuật toán tìm họ tất cả các tập rút gọn

2.2.5 Thuật toán tìm bảng quyết định không dư thừa

2.3 Kết luận chương

CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH THỬ NGHIỆM

3.1 Đặt vấn đề

3.2 Yêu cầu phần mềm nền tảng và cấu hình phần cứng máy PC

3.2.1 Yêu cầu phần mềm nền tảng

3.2.2 Cấu hình phần cứng máy PC

3.3 Giới thiệu chương trình và cách sử dụng

3.3.1 Cấu trúc chương trình

3.3.2 Giới thiệu chương trình

Sao chép thư mục chương trình vào thư mục bất kỳ trên ổ cứng máy PC. Chạy file **FindAllReductAttribute.exe** để mở chương trình. Giao diện chính của chương trình như sau:

Giao diện của chương trình chính

Hình 3.1 Giao diện chương trình tìm tất cả các tập rút gọn trên bảng quyết định nhất quán

Chương trình có 3 phần chính:

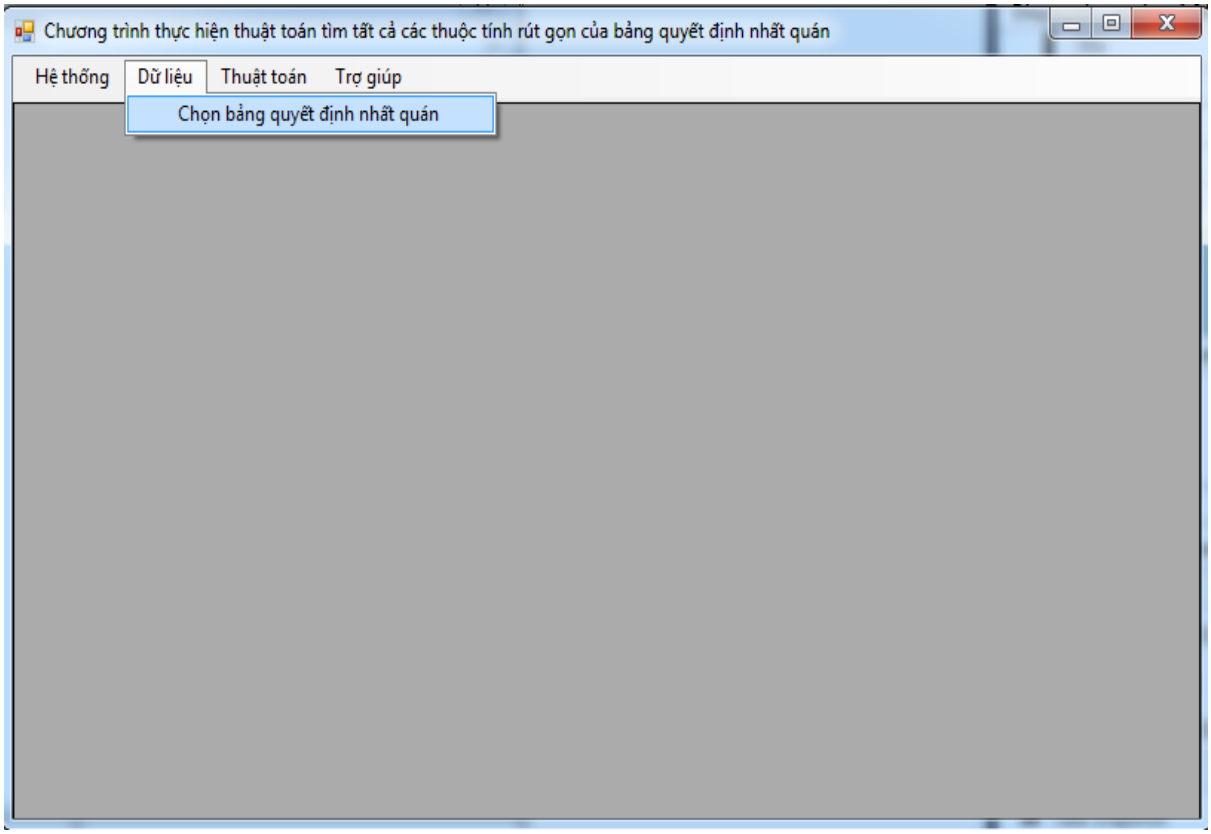
+ Phần 1: Các tab chức năng của chương trình *(Hệ thống / Dữ liệu / Thuật toán /Trợ giúp)*

+ Phần 2: Đầu vào chương trình *(Tab Dữ liệu)*

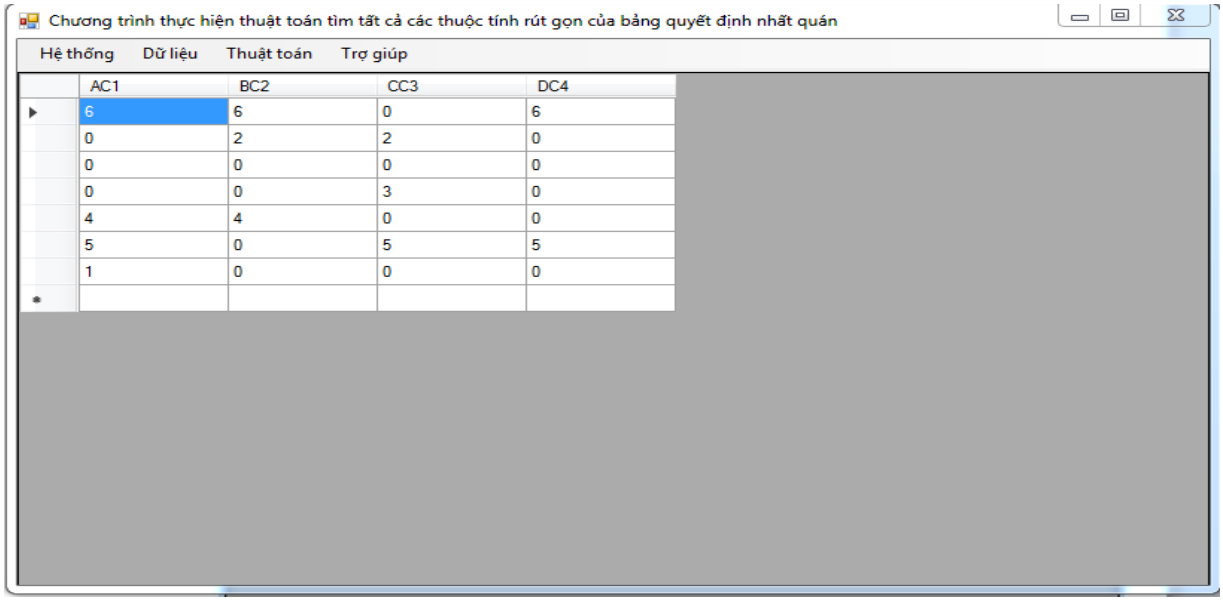
+ Phần 3: Thực hiện thuật toán *(Tab Thuật toán)*

Để thực hiện thuật toán, từ giao diện chương trình chính ta thực hiện theo các bước sau:

1. Chọn tab **"Dữ liệu"** từ giao diện chương trình chính để nhập dữ liệu đầu vào cho chương trình. Chương trình sẽ yêu cầu chọn file dữ liệu đầu vào để thực hiện cho bước tính toán tìm tập thuộc tính rút gọn ở bước sau.

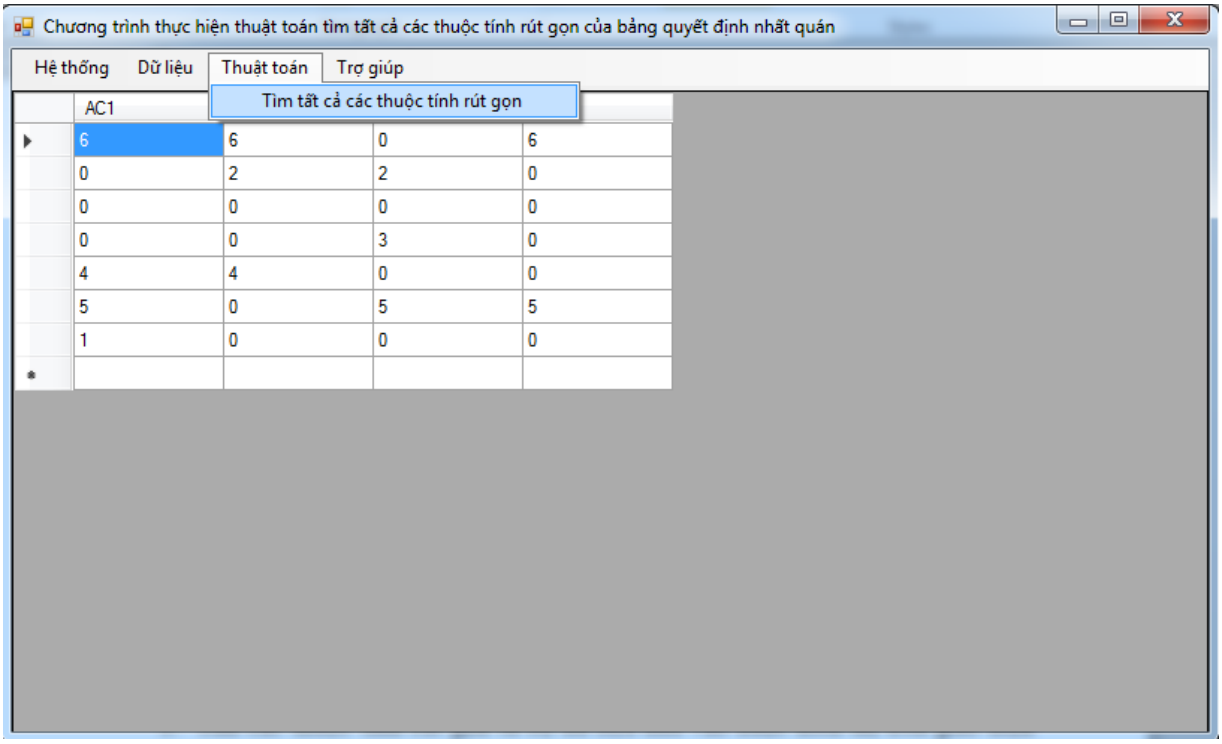


Hình 3.2 Chọn file dữ liệu đầu vào cho chương trình



Hình 3.3 Giao diện chương trình hiển thị dữ liệu đầu vào

2. Tìm các thuộc tính rút gọn từ bộ dữ liệu đầu vào được hiển thị trên giao diện chương trình. Chọn Tab "Thuật toán", sau đó chọn chức năng "Tìm tất cả các thuộc tính rút gọn" để thực hiện tìm các thuộc tính rút gọn.



Hình 3.4 Tìm tất cả các thuộc tính rút gọn

3.4 Thực hiện thuật toán với bộ dữ liệu Flu, EXAMPLE1, EXAMPLE

3.4.1 Bộ dữ liệu Flu

Cho bảng quyết định $DS=(U,C\cup\{c_3\},V,f)$

với U={$u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_8$}, C={c_1,c_2}.

Trong đó: C_1 – đau đầu, C_2 – Thân nhiệt, C_3 – Cảm cúm

U	**C_1**	**C_2**	**C_3**
U_1	Yes	Normal	No
U_2	Yes	High	Yes
U_3	Yes	Very High	Yes
U_4	No	Normal	No
U_5	No	High	No
U_6	No	Very High	Yes
U_7	No	High	Yes
U_8	No	Very High	No

Bảng 3.2 Triệu chứng cúm của bệnh nhân

Dữ liệu đầu vào:

- Bảng quyết định $DS=(U,C\cup\{C3\},V,f)$,
- Các đối tượng U ={$u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_8$},
- Các thuộc tính C ={ c_1,c_2,c_3,} .

Kết quả: Tập các thuộc tính rút gọn.

*** Thực hiện thuật toán:**

a) Kiểm tra xem bảng quyết định có nhất quán không

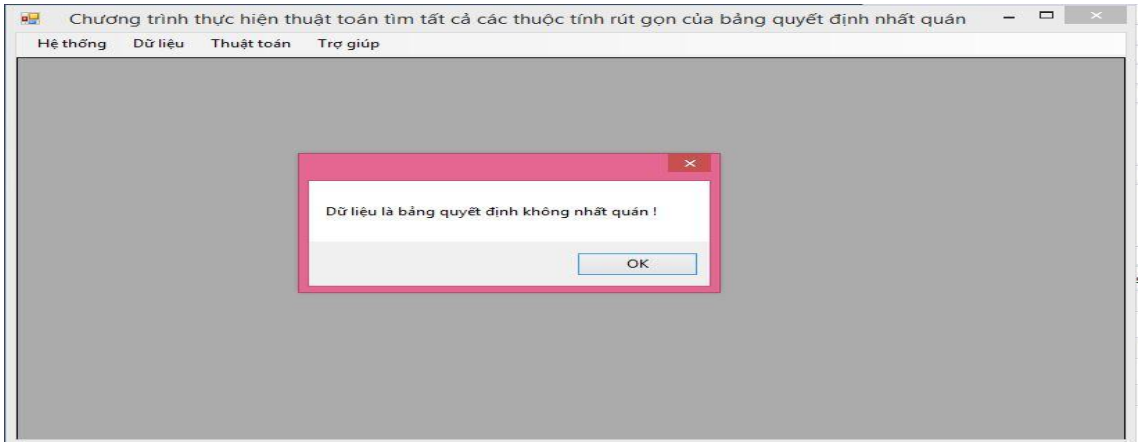
Xấp xỉ dưới của $\underline{B}$X (Yes) = {u2,u3},

Xấp xỉ dưới của $\underline{B}$X (No) = {u1,u4};

=>POSc(D)= $\bigcup_{X\subset U/D}$ ($\underline{B}$X) Không bằng U: Vậy bảng quyết định là không nhất quán.

b) Các bước thực hiện thuật toán

Vì bảng quyết định không nhất quán nên kết thúc thuật toán.



Hình 3.5 Kết quả của bộ dữ liệu Flu

3.4.2 Bộ dữ liệu "EXAMPLE1"

Xét bảng quyết định có các thuộc tính và các đối tượng như sau:

U	AC_1	BC_2	CC_3	DC_4
U_1	6	6	0	6
U_2	0	2	2	0
U3	0	0	0	0
U_4	0	0	3	0
U_5	0	4	0	4
U_6	5	0	5	5
U_7	1	0	0	1

Bảng 3.3 Bảng quyết định bộ dữ liệu Example1

Dữ liệu đầu vào:

- Bảng quyết định $DS=(U,C\cup\{DC_4\},V,f)$,

- Các đối tượng $U=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$,

- Các thuộc tính $C=\{AC_1,BC_2,CC_3,DC_4\}$.

Kết quả: Tập các thuộc tính rút gọn

*** Thực hiện thuật toán:**

a) Kiểm tra xem bảng quyết định có nhất quán không

Xấp xỉ dưới của BX (0) = {U3, U4, U2},

Xấp xỉ dưới của BX(1)= {U7},

Xấp xỉ dưới của $\underline{B}X$ (4) = {U5},

Xấp xỉ dưới của $\underline{B}X$ (5) = {U6},

Xấp xỉ dưới của $\underline{B}X$ (6) = {U1}.

=> $POSc(C_4)= \bigcup_{X \subset U/D} (\underline{B}X_i) = U$: Vậy bảng quyết định nhất quán.

b) Các bước thực hiện thuật toán

+ Bước 1: Hệ bằng nhau

ε_r = {{CC_3},{AC_1, DC_4},{ AC_1},{AC_1, BC_2, DC_4},{AC_1, CC_3},{ BC_2},{BC_2, CC_3}}.

+ Bước 2: Tập M_d = {{ AC_1, CC_3},{ BC_2, CC_3}}.

+ Bước 3: Tập $\bigcap_{K \in M_d} K$ = { CC_3},

$$\text{Tập } V = R - \bigcap_{K \in M_d} K = R - \bigcap_{K \in (\mathbf{K}_d^r)^{-1}} K = \{ AC_1, BC_2, DC_4 \}.$$

+ Bước 4: REAT(C) = V - {DC_4} = { AC_1, BC_2},

=> Hoàn thành thuật toán (thuộc tính rút gọn l { AC_1, BC_2}).

Hình 3.6 Kết quả khi thực hiện thuật toán với bộ dữ liệu Example1

3.4.3 Bộ dữ liệu "EXAMPLE"

Xét bảng quyết định có các thuộc tính và các đối tượng như sau:

U	*a*	*b*	*c*	*d*
u_1	6	6	0	6
u_2	0	2	2	0
u_3	0	0	0	0
u_4	0	0	3	0
u_5	4	4	0	0
u_6	5	0	5	5
u_7	1	0	0	0

Bảng 3.4 Bảng quyết định bộ dữ liệu Example

Dữ liệu đầu vào:

- Bảng quyết định $DS=(U,C\cup\{d\},V,f)$,

- Các đối tượng U =$\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$,

- Các thuộc tính $C=\{a,b,c,d\}$.

Kết quả: Tập các thuộc tính rút gọn

*** Thực hiện thuật toán:**

a) Kiểm tra xem bảng quyết định có nhất quán không

Xấp xỉ dưới của $\underline{B}X$ (0) = {u3,u4,u2,u7,u5},

Xấp xỉ dưới của $\underline{B}X$ (5) = {u6},

Xấp xỉ dưới của $\underline{B}X$ (6) = {u1},

=> POSc(D) = $\bigcup_{X\subset U/D}(\underline{B}X)$ = U: Vậy bảng quyết định là nhất quán.

b) Các bước thực hiện thuật toán

+ Bước 1: Hệ bằng nhau

ε_r = {{c},{a,d},{d},{a,b,d},{c,d},{b},{b,c,d},{b,d}}.

+ Bước 2: Tập M_d = {{c},{b}}.

+ Bước 3:Tập $\bigcap_{K\in M_d} K$ = {}.

Tập $V=R-\bigcap_{K\in M_d} K=R-\bigcap_{K\in(\mathbf{K}_d^r)^{-1}} K$ = {a,b,c,d}

+ Bước 4: REAT(C) = V - {d} = {a,b,c}

=> Hoàn thành thuật toán (*thuộc tính rút gọn là {a,b, c}*).

Hình 3.7 Kết quả tìm các tập rút gọn với bộ dữ liệu Example

3.5 Kết luận chương

Chương này đã hướng dẫn cài đặt chương trình và các kết quả thử nghiệm của chương trình tìm tập thuộc tính rút gọn. Đưa ra một số giao diện chính và cách sử dụng khi chạy chương trình…

KẾT LUẬN VÀ ĐỀ NGHỊ

1. Kết quả đạt được của luận văn

Khai phá dữ liệu là một trong những kỹ thuật quan trọng, mang tính thời sự không chỉ đối với Việt Nam mà của cả nền công nghệ thông tin toàn cầu hiện nay. Với sự bùng nổ thông tin dữ liệu toàn cầu, trong mọi mặt của đời sống xã hội cùng với sự phát triển và ứng dụng ngày càng rộng rãi của công nghệ thông tin trong mọi lĩnh vực đã khiến cho nhu cầu xử lý những khối dữ liệu khổng lồ để phát hiện ra những thông tin, tri thức hữu ích cho người sử dụng một cách tự động, nhanh chóng và chính xác. Một trong những phương pháp quan trọng của kỹ thuật khai phá dữ liệu mà đề tài đi tìm hiểu để làm cơ sở dữ liệu cho một số thuật toán rút gọn trên bảng quyết định nhất quán. Trong khoảng thời gian không dài đề tài đã tổng kết những kiến thức cơ bản nhất để phục vụ cho việc nghiên cứu một số thuật toán liên quan đến tập rút gọn trên bảng quyết định nhất quán. Có thể nói đề tài là một tài liệu tham khảo khá khá đầy đủ, rõ ràng về các kiến thức cơ bản trong khi nghiên cứu một số thuật toán liên quan đến tập rút gọn trên bảng quyết định nhất quán. Thông qua đó đã cài đặt thuật toán **"Tìm tập tất cả các thuộc tính rút gọn trên bảng quyết định nhất quán"** và chạy thử chương trình trên máy PC nhiều lượt với các bộ dữ liệu khác nhau.

2. Hướng nghiên cứu tiếp theo

Trên cơ sở những nghiên cứu đã được trình bày trong luận văn, tiếp tục nghiên cứu sâu hơn một số thuật toán liên quan tới tập rút gọn trên bảng quyết định nhất quán. Nhằm loại bỏ các thuộc tính dư thừa không cần thiết mà vẫn bảo toàn thông tin bài toán. Thông qua việc loại bỏ các thuộc tính dư thừa, các bài toán khai phá dữ liệu trở nên đơn giản hơn, phù hợp với giai đoạn hiện nay... Trong quá trình học tập, tìm hiểu và nghiên cứu cùng với khoảng thời gian làm luận văn, tôi đã cố gắng tập trung tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên do thời gian và điều kiện nghiên cứu có hạn nên không tránh khỏi những thiếu sót, rất mong nhận được sự nhận xét và những đóng góp ý kiến của quý thầy cô giáo và những ai quan tâm để luận văn được hoàn thiện hơn.