HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Vũ Thị Mai

NGHIÊN CỨU ỨNG DỤNG LÝ THUYẾT TẬP THÔ TRONG TRÍCH CHỌN DỮ LIỆU

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01
TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2012

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS. TS. Nguyễn Hoàng Phương

Phản biện 1: ……

Phản biện 2: ……

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

Ngày nay, phát hiện tri thức (Knowledge Discovery) và khai phá dữ liệu (Data mining) là lĩnh vực nghiên cứu đang phát triển mạnh mẽ. Khai phá dữ liệu được sử dụng với những cái tên như là *sự thăm dò và phân tích bằng cách tự động hoặc bán tự động* của một số lượng lớn dữ liệu theo một thứ tự để tìm kiếm được những mẫu có ích hoặc các luật.

Mặc khác, trong môi trường cạnh tranh khốc liệt như hiện nay, người ta ngày càng cần có nhiều thông tin với tốc độ nhanh để trợ giúp việc ra quyết định và ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Với những lý do như vậy dẫn tới sự phát triển một khuynh hướng kỹ thuật mới đó là kỹ thuật phát hiện tri thức và khai phá dữ liệu (*Knowledge Discovery and Data ming – KDD*)

Lý thuyết tập thô được nhà logic học Balan Zdzislak Pawlak giới thiệu vào đầu những năm 80 [20] được xem như là một cách tiếp cận mới để phát hiện tri thức. Nó cung cấp một công cụ để phân tích, trích chọn dữ liệu từ các dữ liệu không chính xác để phát hiện ra mối quan hệ giữa các đối tượng và những tiềm ẩn trong dữ liệu. Nó cho ta một cách nhìn đặc biệt về mô tả, phân tích và thao tác dữ liệu cũng như một cách tiếp cận đối với tính không chắc chắn và không chính xác của dữ liệu.

Mục đích của lý thuyết tập thô là sự phân loại của dữ liệu ở dạng bảng biểu gọi là hệ thông tin. Mỗi hàng biểu diễn một đối tượng (object), mỗi cột biểu diễn một thuộc tính. Nó cung cấp một hệ thống trợ giúp phân loại tập dữ liệu, rút trích các thông tin hữu ích từ tập dữ liệu…Với việc áp dụng lý thuyết tập thô vào việc trích chọn dữ liệu giúp làm giảm đi mức độ đồ sộ của hệ thống dữ liệu, giúp chúng ta có thể nhận biết trước loại dữ liệu được xử lý.

Ở Việt Nam lý thuyết tập thô được chú ý trong một vài năm gần đây. Có nhiều đề tài nghiên cứu cho kết quả khả quan và đã được đưa vào ứng dụng như xử lý ảnh trong y tế, khai phá dữ liệu y tế, nhận dạng, trí tuệ nhân tạo,…

Cho nên tôi chọn đề tài: "***Nghiên cứu ứng dụng lý thuyết tập thô trong trích chọn dữ liệu***" là một kế thừa, phát triển, đóng góp vào những nghiên cứu về lý thuyết tập thô.

CHƯƠNG 1: CÁC PHƯƠNG PHÁP DÙNG TRONG TRÍCH CHỌN DỮ LIỆU

1.1. Tổng quan về khai phá dữ liệu và phát hiện tri thức

1.1.1. Khái niệm về phát hiện tri thức và khai phá dữ liệu

Phát hiện tri thức là lĩnh vực nghiên cứu và ứng dụng tập trung vào dữ liệu, thông tin và tri thức.

Phát hiện tri thức (Knowledge discovery) trong cơ sở dữ liệu là quá trình phát hiện các mẫu hay các mô hình đúng đắn, mới lạ, có lợi ích tiền tàng và có thể hiểu được trong dữ liệu [11].

Khai phá dữ liệu (Data mining) là một bước quan trọng của quá trình phát hiện tri thức bao gồm các giải thuật khai phá dữ liệu để tìm ra các mẫu hay các mô hình trong dữ liệu dưới khả năng có thể chấp nhận được của máy tính điện tử [11].

1.1.2. Quá trình phát hiện tri thức

Các bước của quá trình phát hiện tri thức mô tả hình 1.1

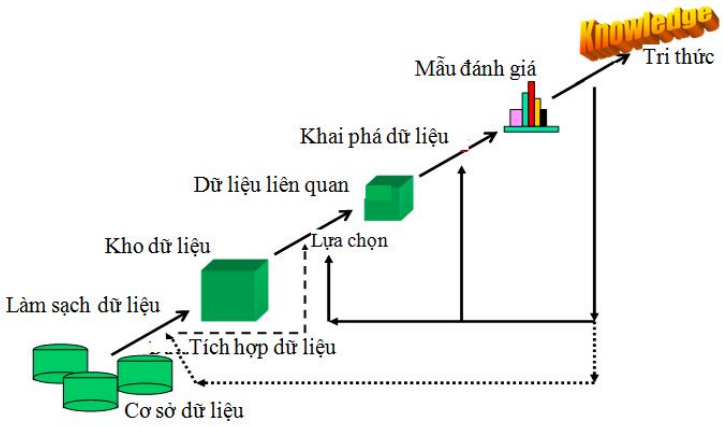
Bước đầu tiên là khảo sát miền ứng dụng và xác định, phát biểu vấn đề.

Bước thứ hai là thu thập và tiền xử lý dữ liệu.

Bước thứ ba là sử dụng các phương pháp khai phá dữ liệu để trích rút ra các dạng và các mô hình ẩn trong dữ liệu.

Bước thứ tư là giải thích tri thức được phát hiện, sau đó lấy trung bình các kết quả để đánh giá hiệu năng các luật.

Bước cuối cùng là đưa tri thức được phát hiện sử dụng trong thực tế.



Hình 1.1. Quá trình phát hiện tri thức

1.1.3. Các nhiệm vụ của phát hiện tri thức và khai phá dữ liệu

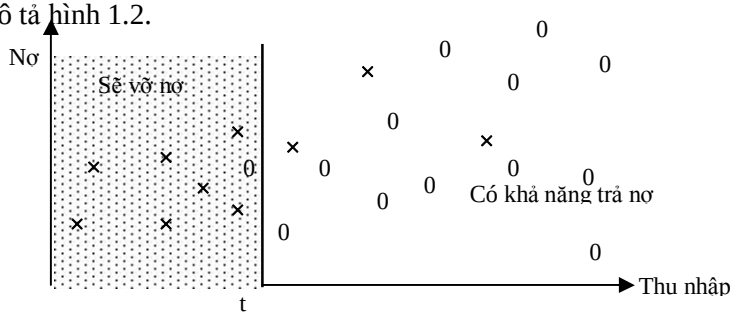
- Phát triển sự hiểu biết của miền ứng dụng
- Tạo dữ liệu mục tiêu (dữ liệu đầu ra)
- Làm sạch dữ liệu tiền xử lý
- Rút gọn dữ liệu và dự báo
- Chọn nhiệm vụ khai phá dữ liệu
- Chọn phương pháp khai phá dữ liệu
- Khai phá dữ liệu để trích xuất các mẫu/mô hình
- Giải thích và đánh giá các mẫu/mô hình

1.1.4. Các thách thức của phát hiện tri thức

- Các cơ sở dữ liệu lớn.
- Dữ liệu nhiều chiều.
- Hiện tượng quá phù hợp (over – fitting).
- Đánh giá ý nghĩa thống kê.
- Dữ liệu động.
- Dữ liệu thiếu và nhiễu.
- Các quan hệ phức tạp giữa các trường.
- Khả năng biểu đạt của mẫu.
- Sự tương tác với người dùng và tri thức có sẵn.
- Tích hợp với các hệ thống khác.

1.2. Các phương pháp trích chọn dữ liệu

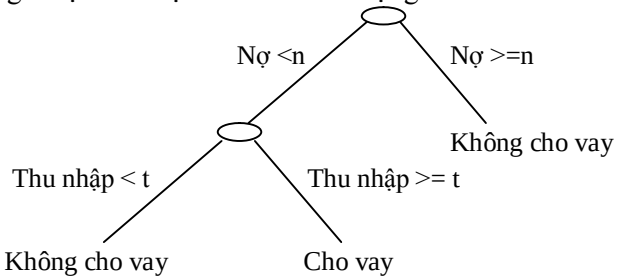
Để minh họa cho quá trình trích chọn dữ liệu tôi xin trình bày ví dụ sau: Một tập dữ liệu hai chiều gồm 23 điểm mẫu. Mỗi điểm biểu thị cho một khách hàng, trục hoành biểu thị thu nhập, trục tung biểu thị tổng dư nợ. Dữ liệu được chia thành hai lớp: dấu x biểu thị cho khách hàng bị vỡ nợ, dấu 0 biểu thị cho khách hàng có khả năng trả nợ. "Nếu thu nhập < t đồng thì khách hàng vay sẽ bị vỡ nợ" như mô tả hình 1.2.



Hình 1.2. Tập dữ liệu hai chiều

1.2.1. Cây quyết định

Cây quyết định mô tả tri thức dạng đơn giản nhằm phân loại các đối tượng dữ liệu thành một số lớp nhất định. Các nút của cây được gán nhãn là tên các thuộc tính, các cạnh được gán các giá trị có thể của các thuộc tính, các lá mô tả các lớp khác nhau. Các đối tượng được phân lớp theo các đường đi trên cây, qua các cạnh tương ứng với các giá trị của thuộc tính của đối tượng tới lá.



Hình 1.3. Cây quyết định

Hình 1.3 mô tả một mẫu đầu ra có thể của quá trình khai phá dữ liệu dùng phương pháp cây quyết định với tập dữ liệu khách hàng xin vay vốn.

1.2.2. Phân cụm (Clustering)

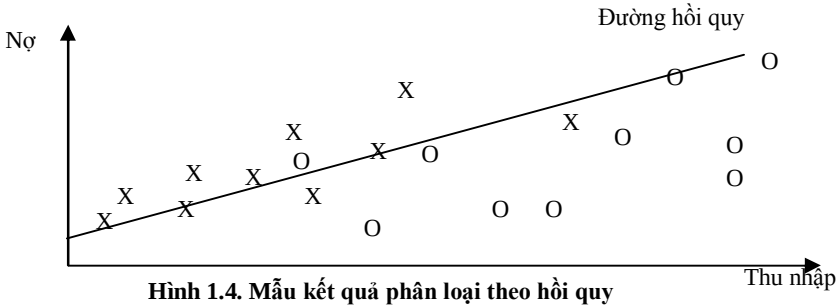
Phân cụm hay nhóm là việc tìm ra các nhóm trong dữ liệu. Các phương pháp phân cụm có thể phân thành hai loại:

- Phân cụm có thứ bậc: Mỗi điểm trong dữ liệu được xem như một cụm riêng biệt được kết hợp một cách liên tiếp dựa vào các quan hệ của nó với các dạng khác.
- Các phương pháp tối ưu hóa dựa trên hàm đối tượng: các phương pháp này sử dụng một chỉ số hiệu năng để giúp cho việc phát triển các phân chia tốt của các điểm dữ liệu.

1.2.3. Hồi quy (Regression)

Hồi quy là việc học một hàm ánh xạ từ một mẫu dữ liệu thành một biến dự đoán có giá trị thực.

Hình 1.4 mô tả mẫu kết quả dự đoán tổng dư nợ của khách hàng với phương pháp khai phá dữ liệu là hồi quy. Đường hồi quy tuyến tính cho thấy rằng những khách hàng có thu nhập càng cao thì tổng dư nợ càng lớn. Mẫu kết quả này không phù hợp với quy luật.

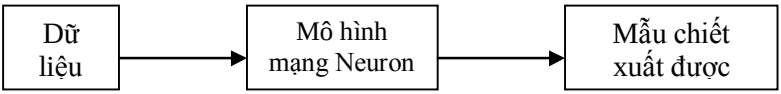


Hình 1.4. Mẫu kết quả phân loại theo hồi quy

1.2.4. Mạng nơron (neural networks)

Mạng nơron là tiếp cận tính toán mới liên quan đến việc phát triển các cấu trúc toán học với khả năng học. Phương pháp là kết quả của việc nghiên cứu mô hình học của hệ thống thần kinh con người.

Một trong số những ưu điểm phải kể đến của mạng nơron là khả năng tạo ra các mô hình dự đoán có độ chính xác cao, có thể áp dụng được cho rất nhiều loại bài toán khác nhau, đáp ứng được nhiệm vụ đặt ra của khai phá dữ liệu như phân loại, phân nhóm, mô hình hóa, dự báo các sự kiện phụ thuộc vào thời gian, v.v…



Hình 1.5. Sơ đồ quá trình khai phá dữ liệu bằng mạng nơron

1.2.5. Lý thuyết tập thô

Tập thô có quan điểm hoàn toàn khác với quan điểm truyền thống về 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 thuộc tập hợp. Rõ ràng có thể tồn tại một số đối tượng giống nhau ở một số thông tin nào đó, và ta nói rằng chúng có quan hệ không thể phân biệt được. Đây chính là quan hệ mấu chốt và chính là điểm xuất phát của lý thuyết tập thô; biên giới của tập thô là không rõ ràng, chúng ta phải xấp xỉ nó bằng các tập hợp khác nhau, nhằm mục đích cuối cùng là trả lời được rằng một đối tượng nào đó thuộc tập hợp hay không. Lý thuyết tập thô với các tiếp cận như vậy đã được ứng dụng rất rộng rãi. Ở chương sau sẽ trình bày ở hơn về lý thuyết tập thô.

CHƯƠNG 2: LÝ THUYẾT TẬP THÔ ỨNG DỤNG TRONG KHAI PHÁ DỮ LIỆU

Lý thuyết tập thô rất hiệu quả trong khai phá dữ liệu, tìm kiếm thông tin, hỗ trợ quyết định, máy học, các hệ cơ sở tri thức.

Lý thuyết tập thô phát huy tác dụng đối với tính không chắc chắn và không chính xác của dữ liệu. Trong lý thuyết tập thô, mỗi khái niệm không chính xác được thay thế bởi một cặp khái niệm chính xác được gọi là xấp xỉ dưới (lower approximation) và xấp xỉ trên (upper approximation). Xấp xỉ dưới gồm tất cả các đối tượng chắc chắn có thể thuộc về khái niệm và xấp xỉ trên bao gồm tất cả đối tượng có thể thuộc về khái niệm. Hiệu của xấp xỉ trên và dưới tạo thành một khoảng ranh giới (boundary region) của khái niệm không rõ ràng.

Lý thuyết tập thô (Pawlak, 1980) [20] và lý thuyết tập mờ (Zadeh, 1965) [15] là những lý thuyết độc lập, nhưng có mối quan hệ khăng khít với nhau và bổ sung cho nhau trong việc biểu diễn và xử lý thông tin không chính xác, không đầy đủ. Trong lý thuyết tập mờ, tính không chính xác được biểu hiện bởi một hàm thuộc, trong khi cách tiếp cận tập thô lại dựa trên tính không phân biệt được và các xấp xỉ.

2.1. Các hệ thống thông tin

2.1.1. Hệ thông tin

Hệ thông tin (information system) là tập hợp dữ liệu được biểu diễn theo dạng bảng, trong đó mỗi dòng là một đối tượng, mỗi cột biểu diễn một thuộc tính.

Xét hệ thông tin S là một bộ bốn S=<U, Q, V,f>

Trong đó:

U={x1,x2,x3,…,xn} là tập hữu hạn đối tượng

Q: Tập hữu hạn thuộc tính, Q=C∪D. C tập các thuộc tính điều kiện, Q thuộc tính quyết định.

$V = \sum_{q \in Q} V_q$ và V_q là vùng xác định của thuộc tính q

f: U x Q $\rightarrow$ V là hàm tổng thể sao cho $f(x,q) \in V_q$ với mọi $q \in Q$ và $x \in U$. f được gọi là hàm thông tin

Ví dụ 2.1: Cho hệ thông tin T1

Bảng 2.1. Bảng thông tin T1

Bệnh nhân	Đau đầu	Đau cơ	Sốt	Cúm
P1	Có	Không	Cao	Có
P2	Không	Có	Cao	Có
P3	Có	Có	Rất cao	Có
P4	Không	Có	Bình thường	Không
P5	Có	Không	Cao	Không
P6	Không	Có	Rất cao	Có

Tập đối tượng U={P1, P2, P3, P4, P5, P6}

Tập thuộc tính Q={Đau đầu, đau cơ, sốt, cúm}

Tập giá trị thuộc tính: $V_{đau\ đầu} = V_{đau\ cơ} = V_{cúm}$ ={có, không};

$V_{sốt}$ ={bình thường, cao, rất cao}

Hàm thông tin f: f(P1, đau đầu) = có; f(P1, đau cơ) = không;

f(P2,đau đầu)=Không; f(P2, sốt) = Cao,…

2.1.2. Hệ quyết định

Hệ thông tin S=<U, C∪D, V,f> được gọi là quyết định nếu và chỉ nếu $C \rightarrow D$; ngược lại, nó là không quyết.

Trong bảng thông tin T1 có thể xem là một hệ quyết định vì có thuộc tính quyết định là cúm. Ta có thể rút ra luật như sau:

"*Nếu đau đầu = có và đau cơ = không và sốt = cao thì cúm = có*"

Trong quá trình tạo tập luật sau này chúng ta thường chú trọng đến việc rút gọn vế trái của luật.

2.2. Tính bất khả phân

2.2.1. Quan hệ tương đương

Quan hệ R trên tập X gọi là quan hệ tương đương nếu thỏa mãn 3 tính chất: Tính phản xạ, tính đối xứng, tính bắc cầu.

2.2.2. Lớp tương đương

Với mỗi phần tử $x \in X$, ta định nghĩa lớp tương đương chứa x, ký hiệu [x], là tập hợp tất cả những phần tử thuộc X và có quan hệ R với x:

$[x]=\{y \in X: yRx\}$

2.2.3. Quan hệ bất khả phân

Giả sử: S = <U, Q, V, f> là một hệ (bảng) thông tin

$P \subseteq Q$, $X \subseteq U$ và $x, y \in U$ (x, y là hai đối tượng trong tập vũ trụ U)

Quan hệ không thể phân biệt theo P (Indiscernibility relation), ký hiệu IND(P) được định nghĩa như sau:

IND(P) = {(x, y) ∈ U x U: f(x,q) = f(y,q) ∀ q∈P}

Quan hệ không thể phân biệt là một quan hệ tương đương và chia tập đối tượng U thành một họ các lớp tương đương. Họ này được gọi là sự phân loại (classification) và ký hiệu U|IND(P) hay U|P. Các đối tượng trong cùng một lớp tương đương là bất khả phân biệt đối với P. Với ∀x∈U, lớp tương đương (equivalence class) của x trong quan hệ IND(P) được biểu diễn là I_p.

Ví dụ 2.2:

Hệ thông tin T1 của bảng 2.1 ở ví dụ 2.1 có một số quan hệ không thể phân biệt như sau:

IND{(Sốt)} = {(P1,P2), (P1,P5), (P2,P5), (P3,P6)}

U|IND({Sốt}) = {{P1, P2, P5}, {P3, P6}, {P4}}

Với P = {Đau đầu, sốt}

IND(P) = {(P1, P5)}

U|IND(P) = {{P1, P5}, {P2}, {P3}, {P4}, {P6}}

2.3. Xấp xỉ tập hợp

2.3.1. Không gian xấp xỉ

Cho hệ thông tin S = <U, Q, V, f> và P ⊆ Q

Một cặp có thứ tự PS = (U, IND(P)) được gọi là một không gian xấp xỉ (approximation space)

Mô tả của tập P-cơ bản X∈U|P được định nghĩa:

$Des_p(X)$ = {(q,v): f(x,q) = v, ∀x∈X, ∀q ∈ P}

2.3.2. Tập xấp xỉ

Cho hệ thông tin S = <U, Q, V, f>. P⊆Q và X ⊆ U.

P – xấp xỉ dưới (P lower approximation) của X trong PS, ký hiệu $\underline{P}(X)$: $\underline{P}(X)$ = {x∈U; I_p (x) ⊆ X}

Những phần tử của $\underline{P}(X)$ là và chỉ là những đối tượng x∈U thuộc vào lớp tương đương sinh ra từ quan hệ không thể phân biệt được I_p chỉ nằm trong X.

P – xấp xỉ trên (P upper approximation) của X trong PS, ký hiệu $\overline{P}(X)$: $\overline{P}(X) = \bigcup_{x \in X} I_p(x)$

Những phần tử $\overline{P}(X)$ là và chỉ là những đối tượng $x \in U$ thuộc vào lớp tương đương sinh ra từ quan hệ không thể phân biệt được, chứa ít nhất một phần tử $x \in X$.

P-biên (P – boundary) của X trong S hay vùng không chắc chắn (Doubtful region) được ký hiệu là $Bn_p(X)$ và tính như sau:

$Bn_p(X) = \overline{P}(X) - \underline{P}(X)$

$Bn_p(X)$ là tập các phần tử mà sử dụng tập thuộc tính P ta không thể xác định chúng có thuộc vào X hay không.

2.3.3. Tập thô

Định nghĩa: Tập hợp X được gọi là tập thô nếu $Bn_p(X)$ là khác rỗng

Ví dụ 2.3. Với bảng thông tin T1 (bảng 2.1)

Thuộc tính cúm = có. X = {P1, P2, P3, P6}

Với P = {Đau đầu, sốt}

U|IND(P) = {{P1, P5}, {P2}, {P3}, {P4}, {P6}}

$\underline{P}(X)$ = {P2, P3, P6}

$\overline{P}(X)$ = {P1, P2, P3, P5, P6}

Bnp(X) = {P1, P5}→ Tập thô

2.3.4. Các tính chất trên tập xấp xỉ

Cho hệ thông tin S = <U, Q, V, f>. $P \subseteq Q$ và $X \subseteq U$.

1. $\underline{P}(X) \subseteq X \subseteq \overline{P}(X)$
2. $\underline{P}(\phi) = \overline{P}(\phi) = \phi, \underline{P(U)} = U$
3. $\overline{P}(X \cup Y) = \overline{P}(X) \cup \overline{P}(Y)$
4. $\underline{P}(X \cap Y) = \underline{P}(X) \cap \underline{P}(Y)$

…

2.3.5. Các loại tập thô

- Tập thô xác định: Tập thô X được gọi là tập thô xác định nếu và chỉ nếu $\underline{P}(X) \neq \varnothing$ và $\overline{P}(X) \neq U$.

- Tập thô không xác định trong: Tập thô X được gọi là tập thô không xác định trong nếu và chỉ nếu $\underline{P}(X) = \varnothing$ và $\overline{P}(X) \neq U$.

-Tập thô không xác định ngoài: Tập thô X được gọi là tập thô không xác định ngoài nếu và chỉ nếu $\underline{P}(X) \neq \varnothing$ và $\overline{P}(X) = U$.

- Tập thô không xác định: Tập thô X được gọi là tập thô không xác định nếu và chỉ nếu $\underline{P}(X) = \varnothing$ và $\overline{P}(X) = U$.

2.3.6. Hệ số xấp xỉ

Hệ số chính xác (acuracy coefficient) là hệ số để đánh giá độ chính xác của xấp xỉ (acuracy approximation). Tập thô có thể đặc trưng hóa dưới hình thức số bằng hệ số phản ánh độ chính xác của xấp xỉ ký hiệu $\alpha_p(X)$: $\alpha_p(X) = \dfrac{|\underline{P}(X)|}{|\overline{P}(X)|}$ $(0 \leq \alpha_p(X) \leq 1)$

Trong đó |X| biểu diễn lực lượng (số phần tử) của tập $X \neq \varnothing$

Nếu $\alpha_p(X) = 1$ thì X là tập rõ đối tượng với quan hệ P

Nếu $\alpha_p(X) < 1$ thì X là tập thô đối với P

2.4. Hàm thuộc thô

Cho $P \subseteq Q$ và $X \subseteq U$, sử dụng khái niệm lớp tương đương, ta có định nghĩa của hàm thuộc thô (rough membership function) – Độ chắc chắn như sau:

$$\mu_x^p(x) = \frac{|X \cap I_p(x)|}{|I_p(x)|}, \ \mu_x^p(x) \in [0, 1].$$

Hàm thuộc thô có một số tính chất:

1. $\mu_x^p(x) = 1$ nếu và chỉ nếu $x \in \underline{P}(X)$
2. $\mu_x^p(x) = 0$ nếu và chỉ nếu $x \in \overline{P}(X)$
3. $0 < \mu_x^p(x) < 1$ nếu và chỉ nếu $x \in Bn_p(X)$

…

2.5. Tập thuộc tính thu gọn - Reduct

2.5.1 Rút gọn các thuộc tính – Reduct

Chỉ giữ lại những thuộc tính không làm ảnh hưởng đến quan hệ bất khả phân và do đó không ảnh hưởng đến tập xấp xỉ. Những tập thuộc tính như vậy gọi là tập thuộc tính thu gọn Reduct.

Cho hệ thông tin $S = <U, Q, V, f>$. $P \subseteq Q$ và $X \subseteq U$.

Tập con P' của P là rút gọn của P (kí hiệu Red(P)) nếu P' là không phụ thuộc và $I_P=I_{P'}$ hoặc U|IND(P) = U|IND(P')

Có thể có nhiều hơn một Y rút gọn của P trong bảng thông tin. Tập chứa tất cả các thuộc tính không thể bỏ được trong P gọi là Y_lõi (Y_Core).

$$CoreY(P)=\bigcap RedY(P)$$

Ví dụ 2.4: Với bảng 2.1 (bảng thông tin T1) trong ví dụ 2.1 ta có thể tìm được các tập lõi là tập rút gọn như sau:

Red_Y={{đau đầu, sốt},{đau cơ, sốt}}; $Core_Y$={Sốt}

2.5.2. Ma trận khả phân (ma trận phân biệt)

Cho hệ thông tin S=<U,Q> với n đối tượng U={x_1, x_2, …, x_n}, ma trận phân biệt (discernibility matrix) của S, ký hiệu M(S) là một ma trận đối xứng n x n với các giá trị c_{ij} được định nghĩa như sau:

(c_{ij}) = {p∈Q: p(x_i) ≠ p(x_j)} đối với i,j = 1, 2, …, n

Lõi có thể định nghĩa là hợp tất cả các tập một phần tử trong ma trận phân biệt được:

CORE(Q) = {p∈Q: c_{ij}={p} với i, j nào đó}

Cho Q'⊆Q có thể dễ dàng thấy rằng Q' là rút gọn của Q, nếu Q' là tập con cực tiểu của Q (đối với phép bao hàm) sao cho: Q' ∩ c ≠ ∅ với mọi phần tử khác rỗng c trong M(S)

Ví dụ 2.6: Cho hệ thông tin S = (U, {a, b, c, d}) như bảng 2.3 từ đó xây dựng ma trận phân biệt, tìm các tập rút gọn và lõi.

Bảng 2.3. Bảng thông tin T2

U	a	b	c	d
x_1	0	1	2	0
x_2	1	2	0	2
x_3	1	0	1	0
x_4	2	1	0	1
x_5	1	1	0	2

Ma trận phân biệt được là đối xứng, do vậy ta chỉ cần xác định các phần tử nằm dưới đường chéo chính của ma trận. Ma trận phân biệt được với bảng 2.3 là như sau:

Bảng 2.4. Ma trận phân biệt biến đổi từ bảng 2.3

	x_1	x_2	x_3	x_4	x_5
x_1					
x_2	a, b, c, d				
x_3	a, b, c	b, c, d			
x_4	a, c, d	a, b, d	a, b, c, d		
x_5	a, c, d	b	b, c, d	a,d	

Từ bảng 2.4 và theo định nghĩa trên ta xác định được lõi chỉ chứa thuộc tính b (vì Core(Q) = {b}, b ∈ Q và c_{52} = {b}) và có 2 tập thuộc tính rút gọn {a, b} hoặc {b, d} trong hệ thông tin.

2.5.3. Hàm khả phân (hàm phân biệt)

Tất cả các rút gọn của một hệ thông tin có thể tìm được thông qua hàm khả phân. Với hệ thông tin S = (U, Q) có ma trận phân biệt M(S) = c_{ij} với (c_{ij}) = {p∈Q: p(x_i) ≠ p(x_j)} và i,j = 1, 2, …, n. Hàm phân biệt fs là một hàm Boolean của m biến Boolean a^*_1, a^*_2, …,a^*_m (ứng với các thuộc tính a_1, a_2, …, a_m) được xây dựng dưới dạng chuẩn tắc tuyển như sau:

fs(a^*_1, a^*_2, …,a^*_m) = ∧ { ∨ c_{ij} | 1 ≤ j ≤ i ≤ n, c_{ij} ≠ ∅}

Trong đó: c^*_{ij} = {a* | a ∈ c_{ij}}

Tập các đơn thức của fs xác định tập rút gọn của S.

Ví dụ 2.7: Theo ví dụ 2.6, ta đã xây dựng được ma trận phân biệt, từ đó ta xác định được hàm phân biệt như sau

fs(a,b,c,d)=(a∨b∨c∨d)∧(a∨b∨c)∧(b∨c∨d)∧(a∨c∨d)∧(a∨b∨d) ∧(a∨b∨c∨d)∧(a∨c∨d)∧b∧(b∨c∨d) ∧(a∨d)

Rút gọn hàm ta được:

fs(a,b,c,d)= b∧(a∨d) = (a ∧ b) ∨(b ∧ d)

Hai tập thuộc tính rút gọn {a,b}; {b,d}

2.5.4. Hàm k-khả phân

Định nghĩa: Hàm k-khả phân là hàm số bool được tạo ra từ việc chỉ xét các mối kết hợp trên một cột k trong ma trận khả phân (thay vì tất cả các cột trong ma trận)

2.5.5. k-Reduct

Định nghĩa: Từ hàm k-khả phân ta tìm ra được các Recduct của hệ thông tin S. Mỗi k-Reduct là tập thuộc tính tối tiểu để nhận ra

được lớp tương đương U|IND(P_k) từ các đối tượng khác trong không gian thông tin.

2.5.6. Không gian quyết định

Định nghĩa: Cho hệ quyết định S = (U, Q ∪ {d}). Với d là thuộc tính quyết định. Số lượng phần tử của tập d(U) = {v | d(x) = v, x ∈ U} được gọi là không gian quyết định của thuộc tính quyết định d.

- Ký hiệu là r(d).

Gọi V_d là miền giá trị của d. V_d xác định như sau:

$V_d = \{ v_d^1, v_d^2, ..., v_d^{r(d)} \}$

2.5.7. Lớp quyết định

Từ thuộc tính quyết định d ta có thể phân chia không gian thông tin như sau: $CLASS_Q(d) = \{ X_S^1, X_S^2, ..., X_S^{r(d)} \}$

Với $X_S^k = \{x \in U \mid d(x) = v_d^k\}$, k= 1, 2, …, r(d).

Định nghĩa:

- $CLASS_Q(d)$ gọi là sự phân loại các đối tượng trong hệ quyết định S dựa trên thuộc tính quyết định d.

- Tập X_S^i gọi là lớp quyết định thứ i của hệ quyết định S.

- $X_Q(u)$: lớp quyết định {x ∈ U | d(x) = d(u)} của mọi u ∈ U.

2.5.8. Reduct quan hệ quyết định

Cho hệ quyết định nhất quán S = (U, Q ∪ {d}). Ma trận khả phân tương ứng $M(S) = (c_{ij})$. Có ma trận quyết định khả phân tương ứng:

$M^d(S) = (c_{ij}^d)$ với $c_{ij}^d = \varnothing$ nếu $d(x_i) = d(x_j)$, i,j = 1, 2, …, n.

Các reduct có được từ hàm quyết định khả phân fs_M^d của ma trận quyết định khả phân $M^d(S)$ gọi là reduct quan hệ quyết định của S.

2.5.9. Thuật toán thu gọn không gian thuộc tính điều kiện

Input: Hàm khả phân $fs=fs_1 \vee fs_2 \vee \ldots \vee fs_n$

Output: Các tập thuộc tính thu gọn của hệ thông tin S

1. Với mỗi phần hội, áp dụng luật hút để loại bỏ những phần hội là tập cha của nó.

2. Thay tất cả các thuộc tính tương đương mạnh bởi các thuộc tính đại diện.

3. Với mỗi phần hội fs_i, áp dụng luật mở rộng nếu được để tách thành hai hàm khả phân $fs_i = fs_{i1} \vee fs_{i2}$.

4. Quay lại 1 cho đến khi không thể thực hiện được (3), ta được các fs_i ở dạng đơn giản

5. Thay thế các thuộc tính đại diện bởi các thuộc tính ban đầu.

6. Phân rã fs_i theo luật phân phối ta được $Red(fs_i)$

7. Các phần giao nhỏ nhất của các $Red(fs_i)$ là các tập thuộc tính thu gọn của hệ thông tin S.

2.6. Sự phụ thuộc của các thuộc tính

Cho D và C là các tập thuộc tính con của Q. Ta nói rằng D phụ thuộc hoàn toàn vào C nếu và chỉ nếu $I(C) \subseteq I(D)$.

Ta nói D phụ thuộc C ở mức k ($0 \le k \le 1$; k được gọi là mức độ phụ thuộc), ký hiệu là $C \Rightarrow_k D$, nếu:

$$k = \gamma(C,D) = \frac{|POS_C D|}{|U|}, \text{ trong đó } POS_c D = \bigcup_{x \in U/D} \underline{C}(X)$$

- Nếu k=1 ta nói rằng D phụ thuộc hoàn toàn vào C
- Nếu k<1, ta nói rằng D phụ thuộc một phần (theo mức độ k) vào C bằng tập thuộc tính C

2.7. Độ quan trọng của các thuộc tính và khái niệm Reduct – xấp xỉ

2.7.1. Độ quan trọng của thuộc tính

Định nghĩa: Cho hệ quyết định S = (U, C ∪ D), D là thuộc thuộc tính quyết định. Độ quan trọng của một thuộc tính a trong hệ quyết định S có thể được ước lượng bằng cách đánh giá mức độ ảnh hưởng của việc loại bỏ thuộc tính a thuộc tập C trong vùng khẳng định của S được tính bằng công thức sau:

$$\sigma_{(C,D)}(a) = \frac{(\gamma(C,D) - \gamma(C - \{a\}, D))}{\gamma(C,D)} = 1 - \frac{\gamma(C - \{a\}, D)}{\gamma(C,D)}$$

2.7.2. Reduct-xấp xỉ

Định nghĩa: Mọi tập con B của C được gọi là Reduct-Xấp xỉ của C với độ sai lệch:

$\varepsilon_{(C,\ D)}(B) = \dfrac{(\gamma(C,D)-\gamma(B,D))}{\gamma(C,D)} = 1-\dfrac{\gamma(B,D)}{\gamma(C,D)}$, mô tả độ chính xác của các thuộc tính B xấp xỉ tập các thuộc tính điều kiện C.

2.8. Phương pháp rút trích đặc trưng

2.8.1. Lượng tử hóa giá trị thuộc tính (Khái niệm các tập nhát cắt)

Cho hệ quyết định S = (U, Q∪{d}). Gọi $V_q = [v_q, w_q)$ là một khoảng các giá trị thực của thuộc tính q ∈ Q. Đối với mọi q trong Q ta tìm các phần P_q có dạng $v_1<v2<\ldots<v_k$ trong V_q (gọi là những nhát cắt – cuts). Việc tìm ra tập các nhát cắt $\{P_q\}_{q\in Q}$ thỏa một số điều kiện cơ bản là một phần trong quy trình lượng tử hóa.

Nhát cắt là một cặp (q, c), với q ∈ Q và c ∈ V_q. Ý tưởng cắt như sau: giá trị của c được định tại điểm giữa của khoảng giá trị của các thuộc tính ở trên.

Từ các nhát cắt ở trên, ta tạo ra được các thuộc tính điều kiện mới với miền giá trị nhị phân tương ứng sao cho, với một nhát cắt (a,c), giá trị của thuộc tính mới:

- 0 nếu q(x) < c
- 1 nếu ngược lại.

Vì thế, nhát cắt nhận ra được những đối tượng có vị trí nằm ở hai mặt của đường thẳng a=c.

2.8.2. Gom nhóm giá trị biểu tượng của thuộc tính

Cho hệ quyết định S = (U, Q ∪ {d}), phân hoạch P_q của miền giá trị của thuộc tính q ∈ Q là hàm số được định nghĩa như sau:

$P_q: V_q \rightarrow \{1, 2,.., m_q\}$, $m_q \le |V_q|$

Thứ hạng của P_{qi} được tính bởi công thức: $rank(P_i) = | P_{qi}(V_{qi})|$

Tập các phân hoạch $\{P_q\}_{q\in B}$ là nhất quán với B nếu và chỉ nếu với mọi (u,u')∈U, (u,u') ∉ IND(B/{d}) thì ∃q ∈ B, P_q(u, u') ∉ IND(B/{d}). Nghĩa là, nếu hai đối tượng (u,u')∈U là khả phân dựa trên tập thuộc tính điều kiện B thì (u,u')∈U cũng sẽ khả phân khi dựa trên các phân hoạch $\{P_q\}_{q\in B}$.

Phân hoạch giá trị biểu tượng: Cho bảng quyết định S = (U, Q∪{d}), và tập các thuộc tính điều kiện B ∈ Q. Việc tập các thuộc tính tối thiểu B tương đương với tìm các tập phân hoạch nhất quán.

Để phân biệt giữa các cặp đối tượng trong không gian thông tin, ta sử dụng biến bool $q_v^{v'}$, với:

v = q(x); v' = q(y); $q_v^{v'}(x,y) = 1$ nếu $v \neq v'$

2.9. Các luật quyết định (decision rules)

Giả sử U|IND(C) là một họ tất cả các tập C cơ bản được gọi là các lớp điều kiện (condition class), kí hiệu X_i (i=1, 2, …, k). Giả sử thêm rằng U|IND(D) là họ các tập cơ bản được gọi là lớp quyết định (decision class), ký hiệu Y_j (j = 1, 2,…,n)

$Des_C(X_i) \Rightarrow Des_D(Y_j)$ được gọi là luật quyết định (C,D).

Những luật là các phát biểu logic "Nếu … thì …" liên kết mô tả các lớp điều kiện với các lớp quyết định. Tập các luật quyết định cho mỗi lớp quyết định Y_j (j=1, 2,…, n) được biểu thị bởi $\{r_{ij}\}$.

$$\{r_{ij}\} = \{Des_C(X_i) \Rightarrow Des_D(Y_j): X_i \rfloor \quad Y_j = \varnothing, i = 1, \ldots, k\}$$

Luật $\{r_{ij}\}$là có tính quyết định nếu và chỉ nếu $X_i \subseteq Y_j$ ngược lại $\{r_{ij}\}$ là không có tính quyết định.

Các luật chắc chắn sẽ được sinh ra từ các đối tượng nằm trong tập xấp xỉ dưới với các thuộc tính đã được rút gọn để sinh ra luật, hoặc có thể tạo ra tập luật tối thiểu bao hàm từ tập lõi.

Các luật không chắc chắn sẽ được sinh ra từ các đối tượng nằm trong vùng biên của với thuộc tính đã được rút gọn để sinh ra luật

Để đánh giá độ chính xác của tập luật có thể dùng hệ số chính xác của xấp xỉ.

Ví dụ 2.8. Từ hệ thông tin T1 của bảng 2.1 ta có thể tính một số xấp xỉ như sau:

- Thuộc tính quyết định D = {Cúm} có V_d = {Có, Không}

Phân loại của U theo giá trị của D là:

$D^* = \{Y_1 = \{P1, P2, P3, P6\}, Y_2 = \{P_4, P_5\}\}$

$Des_D(Y_1) = (\{Cúm\} = Có)$; $Des_D(Y_2) = (\{Cúm\} = Không)$

- Tập các thuộc tính rút gọn A1 = {đau đầu, sốt}; A2 ={đau cơ, sốt}

- Lớp tương đương của các tập rút gọn A1, A2

U|IND(A1)= U|IND(A2) ={X_1 ={P1, P5}, X_2 = {P2}, X_3 ={P3}, X_4={P4}, X_5 ={P6}}

Thiết kế các luật cho lớp Y_1(Có). Vì

$X_1 \cap Y_1 = \{P1\}$ $\qquad$ $X_2 \cap Y_1 = \{P2\}$

$X_3 \cap Y_1 = \{P3\}$ $X_5 \cap Y_1 = \{P6\}$

$X_4 \cap Y_1 = \varnothing$

*** Định nghĩa các luật quyết định cho lớp Y_1 (Cúm = Có) là:**

- Thuộc tính rút gọn A1={đau đầu, sốt}

$\tau_{A111} \Rightarrow Des_D(Y_1)$ $\tau_{A131} \Rightarrow Des_D(Y_1)$

$\tau_{A121} \Rightarrow Des_D(Y_1)$ $\tau_{A151} \Rightarrow Des_D(Y_1)$

- Thuộc tính rút gọn A2={đau cơ, sốt}

$\tau_{A211} \Rightarrow Des_D(Y_1)$ $\tau_{A231} \Rightarrow Des_D(Y_1)$

$\tau_{A221} \Rightarrow Des_D(Y_1)$ $\tau_{A251} \Rightarrow Des_D(Y_1)$

*** Các luật được viết lại cho Y1 (Cúm = Có):**

τ_{A111}: IF (Đau đầu = Có) and (Sốt = Cao) THEN (Cúm = Có). Độ chắc chắn $\mu = 0.5$

τ_{A121}: IF (Đau đầu = Không) and (Sốt = Cao) THEN (Cúm = Có). Độ chắc chắn $\mu = 1$

τ_{A131}: IF (Đau đầu = Có) and (Sốt = Rất Cao) THEN (Cúm = Có). Độ chắc chắn $\mu = 1$

τ_{A151}: (Đau đầu = Không, Sốt = Rất Cao) THEN (Cúm =Có). Độc chắc chắn $\mu = 1$

τ_{A211}: (Đau cơ = Không, Sốt = Cao) THEN (Cúm = Có) $\mu = 0.5$

τ_{A221}: (Đau cơ = Có, Sốt = Cao) THEN (Cúm = Có) $\mu = 1$

τ_{A231}: (Đau cơ = Có, Sốt = Rất Cao) THEN (Cúm = Có) $\mu = 1$

τ_{A251}: (Đau cơ = Có, Sốt = Rất Cao) THEN (Cúm = Có) $\mu = 1$

2.10. Ứng dụng lý thuyết tập thô trong y tế

2.10.1. Ứng dụng lý thuyết tập thô trong phân đoạn ảnh y tế

Phân đoạn ảnh là một bước cơ bản để có thể thực hiện việc phân tích các ảnh thu được. Phân đoạn hình ảnh y tế là một nhiệm vụ quan trọng, phần lớn các nghiên cứu trong phân đoạn ảnh y tế thường gắn liền với việc sử dụng các hình ảnh chụp MRI. MRI (Magnetic Resonance Imaging) là một kỹ thuật chuẩn đoán y khoa tạo ra hình ảnh giải phẫu của cơ thể nhờ sử dụng từ trường và sóng radio.

Lý thuyết tập thô được đề xuất với Pawlak là một công cụ toán học để phân tích sự không rõ ràng và không chắc chắn trong việc

quyết định. Nó phân tích và tìm ra được mối quan hệ của dữ liệu, không gian xấp xỉ với các tập xấp xỉ trên và xấp xỉ dưới.

Phân đoạn hình ảnh y tế được thực hiện thông qua các phương thức khác nhau như: chuẩn đoán hình ảnh (MRI), tính toán cắt lớp (Computed tomography), siêu âm (ultrasound),…

2.10.2. Ứng dụng lý thuyết tập thô trong khai phá dữ liệu y tế

Ứng dụng của tập thô trong lĩnh vực này bao gồm các luật tạo ra từ cơ sở dữ liệu bằng cách sử dụng tập thô trước khi sử dụng những quy tắc trong một hệ chuyên gia.

2.10.3. Ứng dụng lý thuyết tập thô trong hỗ trợ ra quyết định y tế

Quá trình chẩn đoán y tế có thể được hiểu là một quá trình ra quyết định, trong đó các bác sỹ sẽ đưa ra các chuẩn đoán cho một bệnh nhân mới mà các dữ liệu lâm sàng về bệnh nhân này chưa có trong dữ liệu lâm sàng. Quá trình này có thể được máy tính đưa ra thông qua các thủ tục chẩn đoán một cách hợp lý, kịp thời, nhanh chóng và độ chính xác cao.

Trong thực tế từ hai đến ba thập kỷ gần đây hệ thống hỗ trợ chẩn đoán ra quyết định y tế đang trở thành một công cụ hỗ trợ rất tốt cho các bác sỹ và nó đã trở thành một phần của kỹ thuật công nghệ trong y tế.

2.11. Kết luận

Trong chương này tôi đã trình bày một số khái niệm về lý thuyết tập thô như quan hệ tương đương, các tập xấp xỉ trên và xấp xỉ dưới, các cách tìm các tập rút gọn, tập lõi bằng cách tính toán quy nạp dựa trên các xấp xỉ và cách tìm các tập rút gọn, tập lõi bằng cách sử dụng ma trận bất khả phân biệt, sử dụng các khái niệm nhát cát.

Bên cạnh đó cũng giới thiệu một cách tổng quan về các ứng dụng của lý thuyết tập thô trong xử lý thông tin y tế. Một số ứng dụng hiệu quả của tập thô đã chứng minh được tiềm năng của phương pháp này và sẽ được tiếp tục nghiên cứu cải tiến và mở rộng hơn.

CHƯƠNG 3: ÁP DỤNG LÝ THUYẾT TẬP THÔ TRONG TẠO SINH LUẬT CHẨN ĐOÁN Y TẾ

3.1. Tại sao phải tạo luật trong y học?

Cúm là một bệnh truyền nhiễm do virus, có khả năng lây lan cao qua đường hô hấp. Cúm lây truyền mạnh, có thể thành dịch, biểu hiện bởi sốt, viêm đường hô hấp trên, các biến chứng về phế quản và phổi, nặng hay nhẹ tùy theo từng vụ dịch và tùy theo cơ địa của mỗi bệnh nhân.

Cúm nếu không được chẩn đoán sớm và điều trị, bệnh diễn biến kéo dài sẽ đi đến nhiều hậu quả nghiêm trọng, hay gặp nhất là viêm phế quản, tiêu chảy, viêm phổi, viêm tai giữa, viêm não, viêm ngang tủy, và kết quả cuối cùng là dẫn đến tử vong. Vì vậy việc chẩn đoán sớm bệnh cúm đóng vai trò vô cùng quan trọng trong việc cải thiện sức khỏe cho bệnh nhân cúm và ngăn chặn những biến chứng chết người của bệnh.

Lý thuyết tập thô là một công cụ tương đối mới và đã bắt đầu cho thấy tầm quan trọng của nó trong việc hỗ trợ chẩn đoán nhiều bệnh. Trước mối nguy hiểm và sự phát triển không ngừng của bệnh cúm mà đặc biệt là cúm do virus ở Việt Nam cũng như trên toàn thế giới. Trong luận văn này sẽ mang tới một công cụ hỗ trợ trong việc rút gọn các thuộc tính (triệu chứng) của cơ sở dữ liệu chẩn đoán bệnh cúm thông qua việc tạo luật chẩn đoán mới cho các hệ chuyên gia hỗ trợ chẩn đoán bệnh trong y tế.

3.2. Mô tả dữ liệu bài toán

Trước hết chúng ta hãy đi tìm hiểu về quy trình chẩn đoán. Hiện nay khi một bệnh nhân đến khám tại một bệnh viện, bác sỹ sẽ tiến hành chuẩn đoán các bước sau:

Giai đoạn 1: Khám lâm sàng

Ủ bệnh: 2-3 ngày (có thể đến 5 ngày).

Khởi bệnh: đột ngột, với sốt, nhức đầu, đau lưng, mệt mỏi.

Nếu hết giai đoạn này, bác sỹ không có ghi ngờ gì về bệnh cúm, thì bác sỹ sẽ đưa ra câu trả lời phủ định bệnh cúm có thể gợi ý khả năng bệnh nhân mắc một bệnh khác. Bệnh nhân sẽ được khuyên là nên quay lại nếu bệnh nặng hơn mà không rõ căn nguyên

Ngược lại, nếu tối cuối giai đoạn lâm sàng bệnh nhân bị ghi ngờ là đã mắc bệnh thì giai đoạn chuẩn đoán thứ hai sẽ được tiến hành để có kết luận chắc chắn.

Giai đoạn 2: Khám cận lâm sàng
- Chụp X quang
- Xét nghiệm máu
-…

Hầu hết các triệu chứng cận lâm sàng đều có ảnh hưởng rất mạnh đến khả năng mắc bệnh của bệnh nhân. Vì vậy bệnh trạng được khẳng định hoặc loại trừ một cách chắc chắn trong giai đoạn này. Sau đó, bác sĩ sẽ có kết luận và đưa ra một phương pháp điều trị.

Cơ sở dữ liệu y tế về các bệnh nhân cúm được cung cấp bởi bác sĩ Nguyễn Thị Năm bệnh viện Đa khoa tỉnh Hưng Yên. Cơ sở dữ liệu ban đầu có 50 bệnh nhân (*đính kèm trong phần phụ lục*). Mỗi bệnh nhân gồm 12 thuộc tính và thuộc tính quyết định (Cúm = {Có, Không}). Thông tin về các thuộc tính như sau:

1. Daudau (Đau đầu): Có, không
2. Dauco (Đau cơ): Có, không
3. Sot (Sốt): Bình thường, cao, rất cao
4. Onlanh (Ớn lạnh): Có, không
5. Chongmat (Chóng mặt): Có, không
6. Metmoi (Mệt mỏi): Có, không
7. Ho (Ho): Có, không
8. Dauhong (Đau họng): Có, không
9. Chaynuocmui (Chảy nước mũi): Có, không
10. Nghetmui (Nghẹt mũi): Có, không
11. Non (Nôn): Có, không
12. Tieuchay (Tiêu chảy): Có, không
13. Cum (Cúm): Có, không

3.3. Mục đích của bài toán

Từ cơ sở dữ liệu lớn với nhiều thuộc tính (12 thuộc tính điều kiện) cho một bệnh nhân. Mỗi một bệnh nhân lại có những giá trị khác nhau trong cùng một thuộc tính (Có, không,..). Ta có luật

Nếu đau đầu = không và đau cơ = có và sốt = cao và ớn lạnh = có và chóng mặt = không và mệt mỏi = có và đau họng = có và chảy nước mũi = không và nghẹt mũi = không và nôn = không và tiêu chảy = không thì cúm = có.

…

Vấn đề đặt ra: Tìm luật rút gọn cho các thuộc tính điều kiện từ đó đưa ra các luật quyết định để dùng vào cơ sở tri thức của các hệ chuyên gia nhằm mục đích chẩn đoán bệnh. Số thuộc tính rút gọn trong các tập rút gọn phải nhỏ hơn số thuộc tính ban đầu (nhỏ hơn 12 thuộc tính) và có giá trị như nhau trong việc đưa ra các luật quyết định. Luật mới tạo ra có số thuộc tính nhỏ hơn 12 mà không ảnh hưởng tới việc ra quyết định.

Ví dụ luật mới thu được: Nếu đau đầu = không và đau cơ = có và sốt = cao và ớn lạnh = có thì Cúm = có.

…

Để tìm ra được các tập rút gọn để đưa ra được các luật quyết định, trong luận văn tôi áp dụng lý thuyết tập thô để tìm như sau: Bước đầu tôi phải tìm các lớp con tương, sau đó tìm các tập xấp xỉ trên, dưới. Sau đó tìm các lớp con tương đương của các tập thuộc tính con, so sánh nếu các lớp con tương đương của thuộc tính con nhỏ nhất bằng lớp con tương đương của tất cả 12 thuộc tính thì đó chính là tập rút gọn.

3.4. Cài đặt chương trình

Chương trình được viết bằng ngôn ngữ C# trong bộ Visual Studio 2010 của Microsoft và cơ sở dữ liệu được xây dựng trên hệ quản trị cơ sở dữ liệu SQL server 2008 trên hệ điều hành window7.

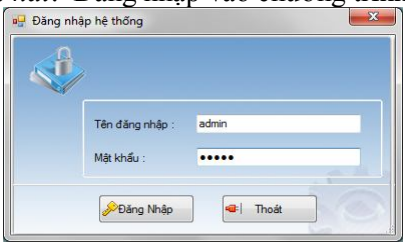
3.5. Kết quả thử nghiệm

Chương trình được thiết kế với năm cửa sổ màn hình, mỗi cửa sổ thực hiện các chức năng riêng biệt.

- *Màn hình đầu tiên*: là phần giới thiệu và trang trí của chương trình

Hình 3.1. Màn hình giao diện chương trình

- *Màn hình thứ hai*: Đăng nhập vào chương trình



Hình 3.2. Màn hình đăng nhập hệ thống

- *Màn hình thứ ba*: Thay đổi mật khẩu của người dùng.

- *Màn hình thứ tư:* Giúp người dùng thao tác với cơ sở dữ liệu được load lên từ hệ quản trị cơ sở dữ liệu SQL server với các chức năng thêm, sửa, xóa các bệnh án. Ở cửa sổ này người dùng tìm được lớp tương đương, tập xấp xỉ trên, tập xấp xỉ dưới, tập biên, hệ số xấp xỉ theo thuộc tính quyết định cúm = {Có, không}

Để thêm bệnh nhân người dùng đưa con trỏ xuống bản ghi trắng cuối cùng và thực hiện thao tác thêm, sau đó chọn cập nhật. Sửa dữ liệu ta sửa trực tiếp vào dữ liệu của bản ghi cần sửa, chọn cập nhật. Xóa bản ghi (bệnh nhân) ta bấm chuột phải vào bản ghi (bệnh nhân) cần xóa và chọn Delete.

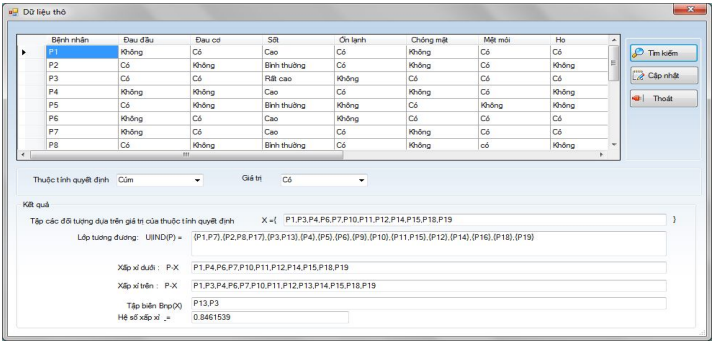
Khi tìm lớp tương đương, tập xấp xỉ, tập biên, hệ số xấp xỉ. Đầu tiên người dùng phải chọn giá trị cho thuộc tính quyết định cúm (có, không), sau đó chọn tìm kiếm. Chương trình thực hiện tìm kiếm tập đối tượng (bệnh nhân) có giá trị của thuộc tính quyết định cúm (có, không). Tiếp đến là tìm tập các lớp con tương đương (*tập các đối tượng có cùng các thuộc tính điều kiện giống nhau)*

Tập xấp xỉ dưới là tập các phần tử chỉ là những đối tượng (bệnh nhân) thuộc vào lớp tương đương sinh ra từ quan hệ không thể phân biệt được.

Tập xấp xỉ trên là tập các phần tử chỉ là những đối tượng (bệnh nhân) thuộc vào lớp tương đương sinh ra từ quan hệ không thể phân biệt được, chứa ít nhất một phần tử thuộc lớp tương đương.

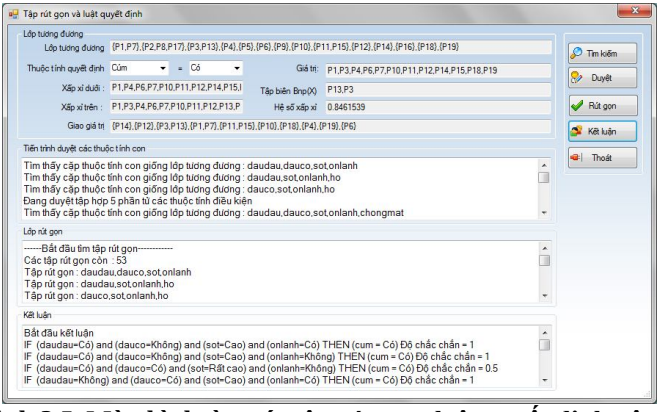
Tập biên Bnp(X) là tập các phần tử không thể xác định có thuộc X hay không (vùng không chắc chắn). Tập biên Bnp(X)=Tập xấp xỉ trên – tập xấp xỉ dưới.

Hệ số xấp xỉ = số phần tử của tập xấp xỉ dưới chia cho số phần tử của tập xấp xỉ trên.



Hình 3.3. Màn hình xem, soạn thảo các bệnh án, tìm lớp tương đương, các tập xấp xỉ, tập biên, hệ số xấp xỉ.

- *Màn hình thứ năm:* Cho phép người sử dụng tìm các lớp tương đương của các thuộc tính con, tìm các tập thuộc tính rút gọn, tìm các luật quyết định và tập lõi.



Hình 3.5. Màn hình tìm các tập rút gọn, luật quyết định, tập lõi

Để tìm được tập rút gọn, chương trình duyệt tìm tất cả các tập con của thuộc tính điều kiện mà có các lớp tương đương bằng với các lớp đương của tất cả các thuộc tính điều kiện. Các tập rút gọn là các tập con nhỏ nhất của các tập vừa tìm được.

Trước khi đưa ra kết luận chương trình tìm giao (phần chung) của từng lớp con với các đối tượng của thuộc tính quyết định. Nếu

giao khác rỗng → đó là một luật, nếu giao bằng rỗng → đó không phải là luật.

Độ chắc chắn μ = số phần tử chung của từng lớp con với các đối tượng của thuộc tính quyết định/ Số phần tử của từng lớp con đó.

3.6. Nhận xét kết quả chương trình

Sau khi kết thúc chương trình tạo luật tự động từ cơ sở dữ liệu bệnh cúm, dựa trên thuật toán phân tích quyết định của lý thuyết tập thô, chương trình thu được các kết quả như sau:

Từ 12 thuộc tính điều kiện ta thu được những tập rút gọn có số thuộc tính điều kiện nhỏ hơn số thuộc tính điều kiện ban đầu. Hệ số xấp xỉ bằng 0.8461539 thực hiện với 50 đối tượng (bệnh nhân):

Red_1 = {Daudau, Dauco, Sot, Onlanh}

Red_2 = {Daudau, Sot, Onlanh, Ho}

Red_3 = {Dauco, Sot, Onlanh, Ho}

…

Và theo đánh giá của bác sỹ thì các tập luật tạo ra tốt cho chẩn đoán bệnh cúm. Rõ ràng việc rút gọn các thuộc tính đã đem lại một lợi ích không nhỏ cho quá trình khai phá dữ liệu khi đã giảm thiểu được sự dư thừa mà hiệu quả mang lại hầu như không thay đổi.

Kết thúc chương trình sẽ thu được các tri thức mới được phát hiện trong cơ sở dữ liệu biểu diễn dưới dạng các luật quyết định có độ chắc chắn kèm theo. Các luật tạo ra sau khi được sự tham khảo của bác sỹ Nguyễn Thị Năm tại bệnh đa khoa tỉnh Hưng Yên bước đầu đã cho kết quả tương đối tốt. Cấu trúc của mỗi luật có dạng như sau:

"*IF (Daudau=Có) AND (Dauco=Không) AND (Sot=Cao) AND (Onlanh=Có) THEN (Cum=Có) Độ chắc chắn = 1.*

IF (daudau=Có) and (dauco=Có) and (sot=Rất cao) and (onlanh=Không) and (dauhong=Không) and (chaynuocmui=Có) THEN (cum = Có) Độ chắc chắn = 0.5

IF (daudau=Không) and (dauco=Có) and (sot=Cao) and (onlanh=Có) and (dauhong=Có) and (chaynuocmui=Không) THEN (cum = Có) Độ chắc chắn = 0.6666667

* Ý nghĩa: Từ cơ sở dữ liệu tri thức lưu trữ, qua lý thuyết tập thô trích chọn ra một số luật nhất định với các thuộc tính rút gọn với độ chắc chắn khác nhau. Đóng góp tri thức của hệ chuyên gia, phát hiện tri thức, luật rút gọn làm cho hệ chuyên gia thông minh hơn

KẾT LUẬN

Luận văn nghiên cứu về lý thuyết tập thô và ứng dụng của nó trong lĩnh vực y tế với mục đích xây dựng chương trình tạo luật cho hệ chuyên gia hỗ trợ trong chuẩn đoán bệnh cúm trong y tế. Những kết quả mà luận văn đã đạt được:

Lý thuyết:

- Đã nghiên cứu các phương pháp khai phá dữ liệu và phát hiện tri thức từ cơ sở dữ liệu.
- Đã nghiên cứu một cách có hệ thống các khái niệm cơ bản của lý thuyết tập thô.
- Nghiên cứu các ứng dụng của lý thuyết tập thô trong lĩnh vực y tế như phân đoạn ảnh, hỗ trợ chẩn đoán y tế.
- Nghiên cứu phương pháp, mô hình chẩn đoán trong y học, thu thập các dữ liệu tri thức về bệnh cúm để dùng trong việc trích chọn, tạo sinh luật y tế.
- Phương pháp trên giúp để xây dựng các luật cho hệ chuyên gia dựa trên cơ sở dữ liệu (tri thức) được lưu trữ. Các luật này kết hợp với tri thức (luật) kinh nghiệm của chuyên gia tạo ra hệ chuyên gia thông minh hơn trong y học.

Ứng dụng:

Trên cơ sở nghiên cứu lý thuyết, đã xây dựng một chương trình tạo luật cho hệ chuyên gia, tạo ra được một tập luật hỗ trợ chẩn đoán bệnh cúm.

Chương trình xây dựng đã tìm kiếm được các quan hệ tương đương, các xấp xỉ, các tập rút gọn và lõi, từ đó đưa ra các luật quyết định dựa vào lý thuyết tập thô.

Hướng nghiên cứu tiếp theo:

- Nâng cao hiệu quả để chương trình chạy nhanh hơn, trích chọn được dữ liệu kể cả dữ liệu đầu vào lớn
- Nghiên cứu chương trình ứng dụng phát triển thành Hệ chuyên gia chẩn đoán bệnh.
- So sánh kết quả với các phương pháp khác, mở rộng hướng nghiên cứu sang lý thuyết tập thô mở rộng và sử dụng thuật tán MD – Heuristics để tìm tập rút gọn.
- Nghiên cứu tạo ra chương trình trích chọn dữ liệu là văn bản,…