KẾT LUẬN VÀ KIẾN NGHỊ

Những vấn đề luận án đã giải quyết

1. Đề xuất và chứng minh công thức biểu diễn mối quan hệ giữa độ đo hỗ trợ với độ đo chính xác và độ đo phủ của các luật quyết định.

2. Đề xuất thuật toán theo tiếp cận gia tăng phát hiện các luật quyết định mới khi các giá trị thuộc tính trong bảng dữ liệu thay đổi. Ưu điểm của thuật toán là chỉ cần cập nhật lại ma trận độ hỗ trợ, dựa trên đó tính ma trận độ chính xác và ma trận độ phủ, rồi sinh luật.

3. Đưa ra và chứng minh các định lý và hệ quả làm cơ sở cho tính đúng đắn của thuật toán theo tiếp cận gia tăng phát hiện các luật quyết định mới khi làm thô, làm mịn các giá trị thuộc tính điều kiện và khi làm thô, làm mịn các giá trị thuộc tính quyết định. Đã đưa ra các mệnh đề đánh giá độ phức tạp của thuật toán.

4. Đưa ra các mệnh đề đánh giá độ phức tạp của các thuật toán (tính gia tăng ma trận độ chính xác và ma trận độ phủ khi bổ sung, loại bỏ đối tượng) theo mô hình của Liu.

5. Đề xuất thuật toán theo tiếp cận gia tăng dựa trên cập nhật ma trận độ hỗ trợ nhằm phát hiện các luật quyết định mới khi bổ sung, loại bỏ đối tượng ra khỏi bảng dữ liệu. Chứng minh tính đúng đắn của thuật toán được đề xuất trên cơ sở chỉ ra sự cập nhật gia tăng ma trận độ hỗ trợ tương ứng với ma trận gia tăng. Đã đưa ra các mệnh đề đánh giá độ phức tạp của thuật toán. Nhờ đó, chứng tỏ thuật toán được đề xuất tốt hơn thuật toán của Liu.

Những vấn đề cần tiếp tục nghiên cứu:

Xây dựng thuật toán để phát hiện các luật quyết định trên bảng dữ liệu có tập các thuộc tính thay đổi hoặc khi bảng dữ liệu có thuộc tính đa trị hoặc bảng dữ liệu không đầy đủ./.

MỞ ĐẦU

Tính cấp thiết của đề tài

Khai phá luật quyết định trên bảng dữ liệu động nhằm nghiên cứu vấn đề trích rút luật quyết định có ý nghĩa trong cơ sở dữ liệu thay đổi theo thời gian về các giá trị thuộc tính, về số các thuộc tính và về số các đối tượng. Tiếp cận gia tăng theo tiếp cận tập thô để giải quyết bài toán khai phá các luật quyết định trên bảng dữ liệu động nhằm giảm chi phí về thời gian và bộ nhớ đòi hỏi sự quan tâm của nhà nghiên cứu.

Trong luận án này đề nghị một cách tiếp cận gia tăng để "Khai phá luật quyết định trên bảng dữ liệu động" trên cơ sở sử dụng độ chính xác và độ phủ của luật làm hai nhân tố đánh giá chất lượng mô tả của các tri thức (luật) quan tâm được trích rút.

Đối tượng nghiên cứu

Đối tượng nghiên cứu chính của luận án là bảng dữ liệu có tập các đối tượng thay đổi và tập các giá trị thuộc tính thay đổi. Mục đích là xây dựng các thuật toán học các tri thức (luật) quan tâm trên bảng dữ liệu động như vậy.

Nội dung, phương pháp nghiên cứu, bố cục của luận án.

Nội dung: Hai nội dung nghiên cứu chính của luận án là (1) xây dựng thuật toán khai phá các luật quyết định từ bảng dữ liệu khi làm thô, làm mịn các giá trị thuộc tính; (2) cải tiến thuật toán khai phá các luật quyết định khi bổ sung, loại bỏ các đối tượng ra khỏi bảng dữ liệu. Cả hai nội dung này đều được phân tích và xem xét dựa trên các công cụ của lý thuyết tập thô mà nền tảng là quan hệ "không thể phân biệt".

Phương pháp nghiên cứu: Tiếp cận gia tăng theo tiếp cận tập thô để giải quyết bài toán khai phá luật quyết định trên bảng dữ liệu động.

Bố cục của luận án

Luận án gồm phần mở đầu, 03 chương nội dung và phần kết luận, danh mục các bài báo đã được công bố và tài liệu tham khảo.

Chương 1: Trình bày tổng quan về khai phá dữ liệu, khai phá luật quyết định trên bảng dữ liệu động, một số khái niệm cơ bản về lý thuyết tập thô, luật quyết định và các độ đo của chúng.

Chương 2: Nghiên cứu một số tính chất của các lớp tương đương; xây dựng thuật toán khai phá các luật quyết định có ý nghĩa khi các giá trị thuộc tính điều kiện hoặc giá trị thuộc tính quyết định được làm thô hoặc làm mịn. Đánh giá độ phức tạp thuật toán được đề nghị.

Chương 3: Trình bày mô hình và thuật toán của Liu để khai phá các luật quyết định có ý nghĩa khi thực hiện việc bổ sung, loại bỏ các đối tượng. Đề xuất thuật toán cải tiến thuật toán của Liu. Đưa ra các mệnh đề đánh giá độ phức tạp của các thuật toán.

Chương 1
TỔNG QUAN

1.1. Khai phá dữ liệu

Khám phá tri thức trong cơ sở dữ liệu (KDD) là một quá trình tìm kiếm trong cơ sở dữ liệu các mẫu đúng đắn, mới, có ích tiềm tàng và có thể hiểu được đối với người sử dụng. KDD là một quá trình gồm nhiều pha, mỗi pha có vai trò và tầm quan trọng riêng. Khai phá dữ liệu (DM) là một pha quan trọng trong toàn bộ tiến trình khám phá tri thức, sử dụng các thuật toán đặc biệt để chiết xuất các mẫu từ dữ liệu.

giả thiết bài toán như nhau; Cùng sử dụng cách tiếp cận gia tăng theo tiếp cận tập thô; Cùng xét trên bảng quyết định đầy đủ với các giá trị thuộc tính đã được rời rạc hóa; Cùng chọn cả độ chính xác và độ phủ để mô tả các tri thức quan tâm. Cho cùng một kết quả khi chạy trên cùng một bảng dữ liệu.

✓ *Khác nhau*

	Phương pháp của Liu	**Phương pháp đề nghị trong luận án**
Phương pháp thực hiện	Lưu và cập nhật đối với cả ma trận độ chính xác và ma trận độ phủ	Chỉ lưu và cập nhật đối với ma trận độ hỗ trợ.
Tính ma trận Acc, Cov tại thời điểm t+1	Trong mỗi lần cập nhật, phải cập nhật tất cả các phần tử của dòng/cột tương ứng với lớp điều kiện, lớp QĐ bị thay đổi của cả 2 ma trận này.	Trong mỗi lần cập nhật, cập nhật trực tiếp cho phần tử của ma trận Sup tương ứng với các lớp bị thay đổi. Việc tính 2 ma trận Acc, Cov chỉ một lần.
Độ phức tạp	- Thời gian:$O(\|U\|^3)$ - Không gian: $O(2\|U\|^2)$	- Thời gian: $O(\|U\|^2)$ - Không gian: $O(\|U\|^2)$

3.5 Kết luận chương 3

Trong chương này, trình bày mô hình và thuật toán của Liu tính gia tăng ma trận độ chính xác và ma trận độ phủ phát hiện các luật quyết định khi bổ sung, loại bỏ đối tượng. Đề xuất thuật toán cải tiến thuật toán của Liu, chứng minh tính đúng đắn của thuật toán trên cơ sở chỉ ra sự cập nhật gia tăng ma trận độ hỗ trợ tương ứng với ma trận gia tăng. Đưa ra các mệnh đề đánh giá độ phức tạp của các thuật toán, nhờ đó chứng tỏ tính hiệu quả của thuật toán cải tiến so với thuật toán của Liu.

Định lý 3.1

Thuật toán tính gia tăng ma trận độ hỗ trợ để phát hiện các luật quyết định khi bổ sung, loại bỏ đối tượng ra khỏi bảng dữ liệu có cùng kết quả với thuật toán tính gia tăng ma trận độ chính xác và ma trận độ phủ khi chạy trên cùng tập dữ liệu.

3.3.3 Độ phức tạp thuật toán

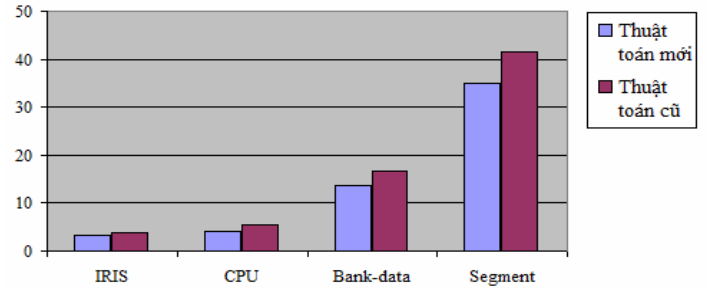
Độ phức tạp thời gian của thuật toán tính gia tăng ma trận độ hỗ trợ để trích rút các luật quyết định có ý nghĩa khi bổ sung, loại bỏ đối tượng là $O(|U|^2)$ và độ phức tạp không gian của nó là $O(|U|^2)$.

3.3.4 Thực nghiệm

Chọn 4 cơ sở dữ liệu từ kho dữ liệu học máy UCI (bảng 3.3) để làm thực nghiệm. Kết quả thực nghiệm được chỉ ra trong hình 3.4.

Bảng 3.3: Các thông tin cơ bản của bốn cơ sở dữ liệu thực nghiệm

Tên tệp dữ liệu	IRIS	CPU	Bank-data	Segment
Số các đối tượng	*150*	*209*	*600*	*1500*
Số thuộc tính điều kiện	*4*	*6*	*10*	*19*
Số thuộc tính quyết định	*1*	*1*	*1*	*1*



Hình 3.4: Thời gian (giây) chạy trung bình của hai thuật toán

3.4 So sánh hai phương pháp phát hiện luật quyết định

✓ *Giống nhau*

Cả hai phương pháp phát hiện luật quyết định cùng sử dụng mô hình bổ sung, loại bỏ đối tượng ra khỏi bảng dữ liệu với yêu cầu và

1.2 Khai phá luật quyết định

Khai phá các luật quyết định là quá trình xác định những luật quyết định trên bảng quyết định cho trước, phục vụ cho việc phân lớp của các đối tượng mới. Khai phá luật quyết định đã được nhiều chuyên gia trong và ngoài nước quan tâm trên cả hai phương diện lý thuyết và ứng dụng. Các nghiên cứu này tập trung chủ yếu xét trên các bảng dữ liệu tĩnh.

Trong thực tế, dữ liệu thường xuyên thay đổi theo thời gian. Đã có một số nghiên cứu về các khía cạnh khác nhau để khai phá các luật quyết định trên bảng dữ liệu động, tập trung chủ yếu vào ba trường hợp: (1) Tập các giá trị thuộc tính thay đổi; (2) Tập các đối tượng thay đổi; (3) Tập các thuộc tính thay đổi.

Trường hợp (1), năm 2010 Chen đã đề nghị một phương pháp học gia tăng để cập nhật các xấp xỉ dưới và xấp xỉ trên của một khái niệm (một lớp quyết định) khi làm thô, làm mịn các giá trị thuộc tính điều kiện. Tuy nhiên, trong cách tiếp cận này, thuật toán được đưa ra chưa đề cập đến trường hợp khi các giá trị thuộc tính quyết định thay đổi, hơn nữa cũng chưa xem xét đến vấn đề làm thế nào để sinh các luật quyết định khi xét đồng thời với nhiều lớp quyết định. Mặt khác, khi xét với mỗi lớp quyết định, thuật toán phải thực hiện lại việc phân lớp các đối tượng khi các giá trị thuộc tính điều kiện thay đổi, chưa tận dụng được các tính chất của các lớp tương đương khi các giá trị thuộc tính thay đổi.

Trường hợp (2), Shan và Ziarko đã đề nghị một phương pháp học gia tăng dựa trên ma trận phân biệt để tìm tất cả các luật quyết định chắc chắn. Một trong những hạn chế chính trong thuật toán của Shan và Ziarko đó là chưa xem xét đến việc trích rút các luật trong bảng quyết định không nhất quán. Để khắc phục hạn chế trên, Bian

đã đề nghị thuật toán cải tiến bằng cách sử dụng ma trận quyết định mở rộng. Tuy nhiên, trong cả hai cách tiếp cận trên đều không đưa ra được các luật quyết định không chắc chắn (đây cũng là các luật có ý nghĩa trong bảng quyết định). Tong và An đã đề xuất thuật toán mới dựa trên ∂ - ma trận quyết định để học gia tăng các luật quyết định, các tác giả đã đưa ra bảy trường hợp có thể xẩy ra khi một đối tượng mới được bổ sung, nhưng chưa đề cập đến vấn đề đối tượng bị loại bỏ. Năm 2009, Liu đề xuất mô hình và thuật toán để phát hiện các luật quyết định khi bổ sung và loại bỏ đối tượng ra khỏi bảng dữ liệu dựa trên việc tính toán gia tăng ma trận độ chính xác và ma trận độ phủ làm cơ sở để sinh các luật quyết định. Nghiên cứu của Liu tiêu tốn nhiều thời gian tính và không gian bộ nhớ do phải cập nhật và lưu trữ đối với cả ma trận độ chính xác và ma trận độ phủ.

Trường hợp (3), Chan đã sử dụng khái niệm phân cấp động được cung cấp bởi người sử dụng để cập nhật gia tăng các xấp xỉ của một khái niệm; Li. đã trình bày một phương pháp để cập nhật các xấp xỉ của khái niệm trong hệ thông tin không đầy đủ dựa trên các quan hệ đặc trưng.

Trong nước, năm 2008 Trọng N.H. đã đề xuất thuật toán khai phá các luật kết hợp khi bảng dữ liệu được gia tăng theo chiều dọc dựa trên việc phân hoạch dữ liệu thành nhiều phần nhỏ tương ứng với các mục dữ liệu và lưu chúng ở bộ nhớ ngoài, mỗi lần xử lý chỉ đưa một số tập phân hoạch vào bộ nhớ trong, hoặc khi bảng dữ liệu gia tăng theo chiều ngang dựa vào cấu trúc của cây quyết định. Tuy nhiên, trong nghiên cứu này cũng chưa đề cập đến vấn đề loại bỏ đối tượng và trường hợp bảng dữ liệu có tập các giá trị thuộc tính thay đổi.

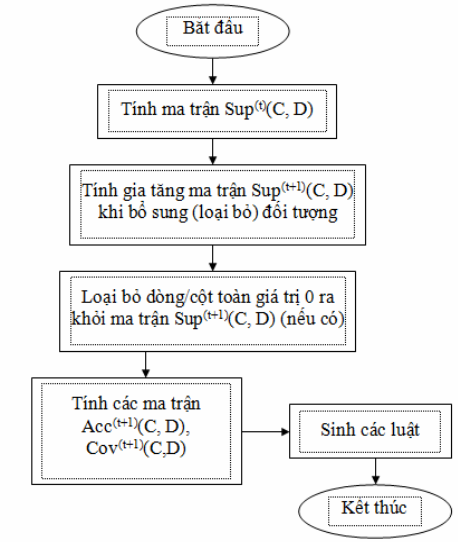
Trong khuôn khổ luận án, tập trung nghiên cứu, xây dựng thuật toán khai phá các luật quyết định trên bảng dữ liệu động theo hướng

Ra: Ma trận Sup tại thời điểm t+ 1

Phương pháp:

- Tìm lớp điều kiện và lớp quyết định mà x thuộc vào
- Cập nhật phần tử của ma trận Sup tương ứng

Kết thúc.



Hình 3.3: Các bước cơ bản của thuật toán tính gia tăng ma trận Sup

Thuật toán 3.6 Tính toán gia tăng ma trận độ hỗ trợ khi xóa đối tượng

Vào: - Tập lớp điều kiện C_i; Tập lớp quyết định D_j;

- Tập DM gồm M đối tượng bị loại bỏ;
- Ma trận Sup tại thời điểm t

Ra: Ma trận Sup tại thời điểm t+ 1

Phương pháp:

Tương tự như thuật toán 3.5

Kết thúc.

3.2.5 Độ phức tạp thuật toán

Độ phức tạp thuật toán của Liu tính toán gia tăng ma trận độ chính xác và ma trận độ phủ là $O(|U|^3)$ và độ phức tạp không gian của nó là $O(2|U|^2)$.

3.3 Tính toán gia tăng ma trận độ hỗ trợ

3.3.1 Cơ sở lý thuyết

Căn cứ mô hình của Liu và yêu cầu bài toán được đặt ra ở trên, thấy rằng khi bổ sung (loại bỏ) đối tượng thực chất là bổ sung (loại bỏ) đối với ma trận độ hỗ trợ. Khi đó ta có: $Sup^{(t+1)}(C_i, D_j) = Sup^{(t)}(C_i, D_j) + N_{ij} - M_{ij}$ với i = 1,…m+p; j=1,…,n+q, trong đó $M_{ij} = 0$ và $Sup^{(t)}(C_i, D_j) = 0$ $\forall$ i=m+1,…,m+q, j=n+1,…,n+q (vì ta chỉ xóa các đối tượng có chỉ số i từ 1 đến m và chỉ số j từ 1 đến n).

Như vậy, thay vì việc phải cập nhật các phần tử của dòng/cột tương ứng trong cả 2 ma trận độ chính xác và ma trận độ phủ, ta chỉ cần tìm lớp tương đương bị thay đổi và cập nhật trực tiếp cho ma trận độ hỗ trợ tương ứng. Việc tính ma trận độ chính xác và ma trận độ phủ làm cơ sở cho việc sinh các luật quyết định có ý nghĩa được suy ra từ ma trận độ hỗ trợ sau khi đã được cập nhật.

3.3.2 Thuật toán

Hình 3.3 biểu thị các bước cơ bản của thuật toán, trong đó sử dụng thuật toán 2.1 để tính ma trận Sup tại thời điểm t; thuật toán 2.6 để tính ma trận Acc, Cov và thuật toán 2.7 để trích rút luật quyết định. Các thuật toán để thực hiện các bước còn lại được trình bày dưới đây.

Thuật toán 3.5 Tính toán gia tăng ma trận độ hỗ trợ khi bổ sung đối tượng

Vào: - Tập lớp điều kiện C_i; Tập lớp quyết định D_j;

- Tập AN gồm N đối tượng được bổ sung;

- Ma trận Sup tại thời điểm t

tiếp cận gia tăng đối với hai trường hợp thay đổi của bảng dữ liệu đó là: Bảng dữ liệu có các giá trị thuộc tính thay đổi và bảng dữ liệu có tập các đối tượng thay đổi.

1.3 Lý thuyết tập thô

1.3.1 Hệ thông tin

Định nghĩa 1.1

Hệ thông tin là một bộ bốn IS = (U, A, V, f), trong đó U là tập hữu hạn, khác rỗng các đối tượng gọi là tập vũ trụ, A là tập hữu hạn khác rỗng các thuộc tính, $V = \bigcup_{a \in A} V_a$ là tập các giá trị thuộc tính, trong đó V_a là tập giá trị của thuộc tính a, $f: U \times A \rightarrow V$ là hàm thông tin sao cho $\forall x \in U$, $\forall a \in A$ ta có $f(x, a) \in V_a$. Ta gọi f(x, a) là giá trị của đối tượng x trên thuộc tính a, tập $X \neq \phi$, $X \subseteq U$ gọi là một khái niệm trong IS.

1.3.2 Quan hệ bất khả phân biệt

Giả sử IS = (U, A, V, f) là một hệ thông tin. Với mỗi tập thuộc tính $P \subseteq A$ xác định một quan hệ tương đương, ký hiệu là IND(P), gọi là quan hệ bất khả phân biệt, được định nghĩa là $IND(P) = \{(x, y) \in U \times U: \forall a \in P, f(x, a) = f(y, a)\}$. Quan hệ IND(P) chia U thành họ các lớp tương đương, tạo thành một phân hoạch của U, ký hiệu là U/P.

Với mỗi đối tượng $x \in U$, lớp tương đương chứa x theo quan hệ IND(P), được ký hiệu là $[x]_P$ được định nghĩa là $[x]_P = \{y \in U: (x, y) \in IND(P)\}$. Điều này có nghĩa rằng, hai đối tượng thuộc cùng một lớp tương đương khi và chỉ khi chúng có giá trị giống nhau trên các thuộc tính trong P. Do đó, để xác định các lớp tương đương, ta có thể sắp xếp các đối tượng trong U theo một thứ tự tùy ý (thông thường sắp xếp theo thứ tự từ điển).

Định nghĩa 1.2

Cho hệ thông tin IS = (U, A, V, f), , P, $Q \subseteq A$ là tập các thuộc tính, $U/P = \{P_1, ...,P_m\}$, $U/Q = \{Q_1, ...,Q_n\}$ là các phân hoạch được

sinh bởi P, Q, ta nói Q thô hơn (coarser) P hoặc P mịn hơn (refiner) Q khi và chỉ khi $\forall P_i \in U/P, \exists Q_j \in U/Q$ (i = 1,...,m; j= 1,...,n) sao cho $P_i \subseteq Q_j$.

1.3.3 Xấp xỉ tập hợp

Định nghĩa 1.3

Cho hệ thông tin IS = (U, A, V, f), $P \subseteq A$ là tập các thuộc tính, $X \subseteq U$ là tập các đối tượng, khi đó các tập $\underline{P}X = \{x \in U: [x]_P \subseteq X\}$ và $\overline{P}X = \{x \in U: [x]_P \cap X \neq \phi\}$ tương ứng được gọi là P-xấp xỉ dưới và P-xấp xỉ trên của X trong IS. Vùng $BN_P(X) = \overline{P}X - \underline{P}X$ được gọi là P – vùng biên của X. Nếu $BN_P(X) = \phi$ thì X gọi là tập rõ (crisp), trái lại X gọi là tập thô.

1.3.4 Bảng quyết định

Một trường hợp đặc biệt của hệ thông tin gọi là bảng quyết định nếu tập thuộc tính A được phân thành hai tập rời nhau C và D, trong đó C là tập các thuộc tính điều kiện, D là tập các thuộc tính quyết định sao cho $C \cap D = \phi$, $C \cup D = A$. Bảng quyết định được ký hiệu là: DS = (U, C ∪ D, V, f) hoặc DS = (U, C ∪ D).

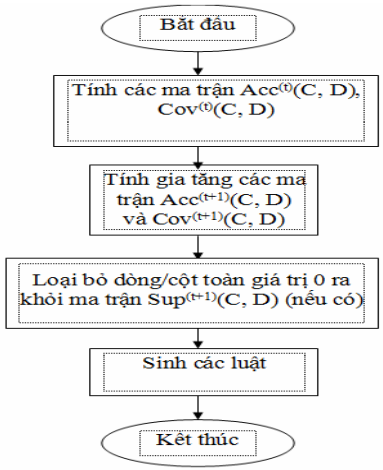
Giả sử $U/C = \{C_1, C_2,\ldots,C_m\}$ và $U/D = \{D_1, D_2,\ldots,D_n\}$ tương ứng là các phân hoạch được sinh bởi tập các thuộc tính điều kiện C và tập các thuộc tính quyết định D, $\forall i = 1,\ldots,m$; $\forall j = 1,\ldots,n$, C_i, D_j tương ứng được gọi là các lớp tương đương điều kiện và các lớp tương đương quyết định.

1.3.5 Luật quyết định

Định nghĩa 1.4

Cho bảng quyết định DS = (U, C ∪ D), $U/C = \{C_1, \ldots, C_m\}$; $U/D = \{D_1, D_2,\ldots,D_n\}$ tương ứng là các phân hoạch được sinh bởi C, D. Một luật quyết định được biểu diễn dưới dạng $C_i \rightarrow D_j$ ở đây $C_i \in U/C$, $D_j \in U/D$ (i=1,…,m; j=1,…,n).

Định nghĩa 1.5



Hình 3.2: Các bước cơ bản thuật toán tính gia tăng ma trận độ chính xác và ma trận độ phủ

Các bước còn lại được trình bày bởi các thuật toán dưới đây:

Thuật toán 3.2: Tính toán gia tăng ma trận độ chính xác và ma trận độ phủ tại thời điểm t+1 khi bổ sung đối tượng x.

Vào: - Các lớp tương đương C_i; các lớp tương đương D_j

- Tập AN chứa N đối tượng được bổ sung

Ra: $Acc^{(t+1)}(C, D)$ và $Cov^{(t+1)}(C ,D)$.

Phương pháp:

Thực hiện trường hợp 1 mục 3.3.2

Kết thúc.

Thuật toán 3.3: Tính toán gia tăng ma trận độ chính xác và ma trận độ phủ tại thời điểm t+1 khi xóa đối tượng x'.

Vào: - Các lớp tương đương C_i ;Các lớp tương đương D_j

- Tập DM chứa M đối tượng bị loại bỏ;

Ra: $Acc^{(t+1)}(C, D)$ và $Cov^{(t+1)}(C ,D)$.

Phương pháp:

Thực hiện trường hợp 2 mục 3.3.2

Kết thúc.

- Trường hợp 1.1: Sinh lớp điều kiện mới và lớp quyết định mới. Khi đó, ta có $x \notin C_i$ (i=1,…,m) và $x \notin D_j$ (j=1,…,n), tức việc bổ sung x sẽ sinh một lớp điều kiện mới C'_{m+1} và một lớp quyết định mới D'_{n+1}.

-Trường hợp 1.2: Chỉ sinh lớp điều kiện mới. Khi đó, ta có $x \notin C_i$ (i=1,…,m) và $\exists j^* \in \{1, 2, \ldots, n\}$: $x \in D_{j^*}$ tức việc bổ sung x sẽ sinh lớp điều kiện mới C'_{m+1} và bổ sung lực lượng cho D_{j^*}.

- Trường hợp 1.3: Chỉ sinh lớp quyết định mới. Khi đó $\exists i^* \in \{1, 2,\ldots, m\}$ sao cho $x \in C_{i^*}$ và $x \notin D_j$ (j=1,…,n), tức là việc bổ sung x sẽ sinh một luật không chắc chắn mới và x hình thành một lớp quyết định mới D'_{n+1}.

- Trường hợp 1.4: Không sinh lớp điều kiện mới cũng không sinh lớp quyết định mới. Khi đó $\exists i^* \in \{1, 2, \ldots, m\}$ sao cho $x \in C_{i^*}$ và $\exists j^* \in \{1, 2, \ldots, n\}$ sao cho $x \in D_{j^*}$. Do đó, việc bổ sung x làm gia tăng độ hỗ trợ của luật $C_{i^*} \rightarrow D_{j^*}$.

Trường hợp 2: Loại bỏ đối tượng x' ra khỏi hệ thống:

$\exists i^* \in \{1, 2, \ldots, m\}$ sao cho $x' \in C_{i^*}$, $\exists j^* \in \{1, 2, \ldots, n\}$ sao cho $x' \in D_{j^*}$. Do đó, sự thay đổi này làm ảnh hưởng đến dòng i* và cột j* của các ma trận độ chính xác và ma trận độ phủ.

3.2.4 Thuật toán

Các bước cơ bản của thuật toán (hình 3.2), trong đó sử dụng thuật toán 2.7 để sinh luật quyết định có ý nghĩa.

Thuật toán 3.1: Tính ma trận độ chính xác và ma trận độ phủ tại thời điểm t

Vào: - Các lớp tương đương C_i; các lớp tương đương D_j

Ra: Ma trận độ chính xác $Acc^{(t)}(C,D)$ và ma trận độ phủ $Cov^{(t)}(C, D)$;

Phương pháp:

Áp dụng định nghĩa 1.5

Kết thúc.

Cho bảng quyết định DS = (U, C ∪ D). Giả sử $C_i \in U/C$; $D_j \in U/D$ (i = 1,…,m; j = 1,…, n). Độ hỗ trợ, độ chính xác và độ phủ của luật quyết định $C_i \rightarrow D_j$ tương ứng được định nghĩa như sau:

- ✓ Độ hỗ trợ: $Sup(C_i, D_j) = |C_i \cap D_j|$
- ✓ Độ chính xác: $Acc(C_i, D_j) = \frac{|C_i \cap D_j|}{|C_i|}$
- ✓ Độ phủ: $Cov(C_i, D_j) = \frac{|C_i \cap D_j|}{|D_j|}$

trong đó, |.| biểu thị lực lượng của tập hợp.

Khi xem xét các bảng dữ liệu lớn, để đơn giản chúng tôi biểu diễn các độ đo này dưới dạng các ma trận độ đo như sau:

- ✓ Ma trận độ hỗ trợ: $Sup(C,D) = (Sup(C_i, D_j))_{m \times n}$
- ✓ Ma trận chính xác: $Acc(C,D) = (Acc(C_i, D_j))_{m \times n}$
- ✓ Ma trận độ phủ: $Cov(C,D) = (Cov(C_i, D_j))_{m \times n}$

Định nghĩa 1.6

Nếu $Acc(C_i, D_j) = 1$ thì $C_i \rightarrow D_j$ gọi là luật quyết định chắc chắn, nếu $0 < Acc(C_i, D_j) < 1$ thì $C_i \rightarrow D_j$ gọi là luật quyết định không chắc chắn (i=1,…,m; j=1,…,n).

Mệnh đề 1.1

$\forall C_i \in U/C$; $\forall D_j \in U/D$ (i=1,…,m; j = 1,…,n), ta có

$$Acc(C_i,D_j) = \frac{Sup(C_i, D_j)}{\sum_{q=1}^{n} Sup(C_i, D_q)}$$

$$Cov(C_i,D_j) = \frac{Sup(C_i, D_j)}{\sum_{p=1}^{m} Sup(C_p, D_j)}$$

Định nghĩa 1.7

Giả sử $C_i \in U/C$; $D_j \in U/D$ (i=1,…,m; j = 1,…,n), nếu $Acc(C_i,D_j) \geq \alpha$ và $Cov(C_i,D_j) \geq \gamma$ thì ta gọi luật $C_i \rightarrow D_j$ là luật

quyết định có ý nghĩa, trong đó α, γ là hai ngưỡng cho trước, với α, $\gamma \in (0,1)$.

1.4 So sánh kỹ thuật phân lớp dựa trên luật kết hợp và phân lớp dựa trên tập thô

Có thể so sánh Kỹ thuật phân lớp dựa trên luật kết hợp (ký hiệu là A_C) và kỹ thuật phân lớp dựa trên tập thô (ký hiệu là R_C) ở hai khía cạnh đó là: độ chính xác phân lớp và số lượng các luật được sinh ra. Các kết quả thử nghiệm cho thấy, trong hầu hết các tập dữ liệu, độ chính xác phân lớp của A_C xấp xỉ với R_C, cá biệt đối với một vài tập dữ liệu thì độ chính xác phân lớp của A_C cao hơn R_C. Về số lượng các luật được sinh ra, trong hầu hết các trường hợp A_C sinh nhiều luật hơn R_C.

1.5 Kết luận chương 1

Chương một trình bày tổng quan về khai phá dữ liệu, khai phá các luật quyết định và một số vấn đề cơ bản của lý thuyết tập thô, luật quyết định, đưa ra công thức biểu diễn mối quan hệ giữa các độ đo của các luật quyết định. Đây là những vấn đề cơ bản để nắm bắt và trình bày các kết quả trong các chương sau của luận án.

Chương 2
KHAI PHÁ LUẬT QUYẾT ĐỊNH TRÊN BẢNG DỮ LIỆU CÓ CÁC GIÁ TRỊ THUỘC TÍNH THAY ĐỔI

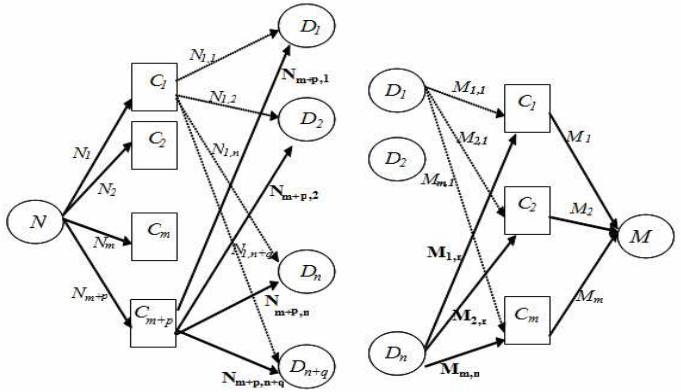
2.1 Giới thiệu

Sự thay đổi các giá trị thuộc tính nói chung được chia thành hai loại: hoặc là một vài giá trị thuộc tính được kết hợp với nhau thành một giá trị mới (làm thô); hoặc là một vài giá trị thuộc tính được tách thành hai giá trị mới (làm mịn). Như vậy, khi làm thô, làm mịn các giá trị thuộc tính thì các phân hoạch được sinh bởi các thuộc tính

bảng quyết định mới. Ký hiệu, $U'/C = \{C'_1,\ldots,C'_m,\ldots,C'_{m+p}\}$, $U'/D = \{D'_1,\ldots,D'_n,\ldots,D'_{n+q}\}$ tương ứng là tập các lớp tương đương điều kiện mới và tập các lớp tương đương quyết định mới; $Sup^{(t+1)}(C,D)$, $Acc^{(t+1)}(C,D)$ và $Cov^{(t+1)}(C, D)$ tương ứng là các ma trận độ hỗ trợ, ma trận độ chính xác và ma trận độ phủ của các luật sau khi tập các đối tượng thay đổi.

3.2.2 Mô hình

Giả sử trong N đối tượng được bổ sung, có N_i đối tượng được bổ sung cho lớp C_i (i = 1,…,m+p); trong N_i đối tượng bổ sung cho lớp C_i có N_{ij} đối tượng được bổ sung cho lớp D_j (j=1,2,..,n+q). Tương tự, trong M đối tượng bị loại bỏ, có M_i đối tượng bị loại khỏi lớp C_i (i=1,…,m); trong M_i đối tượng bị loại khỏi lớp C_i có M_{ij} đối tượng bị loại khỏi lớp D_j (j=1,2,…,n) (hình 3.1).



Hình 3.1: Tiến trình bổ sung/loại bỏ các đối tượng

3.2.3 Tính toán gia tăng ma trận độ chính xác và ma trận độ phủ

Khi bổ sung, loại bỏ đối tượng ra khỏi bảng dữ liệu, xẩy ra 4 trường hợp, trong mỗi trường hợp ta xét sự thay đổi của các dòng/cột của các ma trận Acc, Cov và cập nhật nó. Sau lần cập nhật cuối cùng sẽ thu được các ma trận này tại thời điểm t+1.

Trường hợp 1: Bổ sung đối tượng x, xẩy ra 4 trường hợp

trị thuộc tính thay đổi. Đồng thời, cũng đưa ra các mệnh đề đánh giá độ phức tạp của các thuật toán được đề nghị.

Chương 3
KHAI PHÁ LUẬT QUYẾT ĐỊNH TRÊN BẢNG DỮ LIỆU CÓ TẬP ĐỐI TƯỢNG THAY ĐỔI

3.1 Giới thiệu

Năm 2009, Liu, D. đã đề xuất một mô hình và thuật toán để phát hiện các luật quyết định khi tập đối tượng thay đổi dựa trên việc tổ chức lưu trữ và cập nhật đối với cả hai ma trận độ chính xác và độ phủ làm cơ sở cho việc sinh luật, vì vậy tiêu tốn thời gian tính và không gian bộ nhớ cần thiết. Trong chương này, đề xuất thuật toán cải tiến thuật toán của Liu nhằm giảm chi phí về thời gian và bộ nhớ.

3.2 Mô hình của Liu tính toán gia tăng ma trận độ chính xác và ma trận độ phủ

3.2.1 Yêu cầu và giả thiết bài toán

Cho bảng quyết định DS = (U, C $\cup$ D). Giả sử thêm vào DS N đối tượng và xóa đi M đối tượng. Yêu cầu đặt ra là: Rút ra các luật quyết định thỏa mãn đồng thời cả ngưỡng độ chính xác và ngưỡng độ phủ cho trước sau khi tập đối tượng thay đổi.

Giả sử tiến trình cập nhật tri thức diễn ra từ thời điểm t đến thời điểm t+1. Tại thời điểm t, tập các lớp tương đương điều kiện và tập các lớp tương đương quyết định tương ứng được ký hiệu là U/C = $\{C_1,\ldots,C_m\}$ và U/D = $\{D_1,\ldots,D_n\}$; $Sup^{(t)}(C, D)$, $Acc^{(t)}(C, D)$ và $Cov^{(t)}(C, D)$ tương ứng là các ma trận độ hỗ trợ, ma trận độ chính xác và ma trận độ phủ của tất cả các luật. Tại thời điểm t+1, giả sử AN là tập N đối tượng được bổ sung, N đối tượng được bổ sung hình thành thêm p lớp tương đương điều kiện mới và q lớp tương đương quyết định mới; DM là tập M đối tượng bị loại bỏ; DS' = (U', C $\cup$ D) là

cũng trở nên thô hay mịn hơn. Khi đó các luật quyết định đã thu được trước đó có thể bị thay đổi, không còn giá trị tại thời điểm mới.

Để thu được các luật quyết định có ý nghĩa tại thời điểm mới, trong chương này luận án đề xuất thuật toán trích rút các luật quyết định có ý nghĩa khi làm thô, làm mịn các giá trị thuộc tính điều kiện và khi làm thô, làm mịn các giá trị thuộc tính quyết định.

2.2 Khái niệm làm thô, làm mịn các giá trị thuộc tính

Định nghĩa 2.1

Cho hệ thông tin IS = (U, A, V, f), a $\in$ P $\subseteq$ A, V_a là tập giá trị của thuộc tính a. Giả sử $f(x_p, a) = w$, $f(x_q, a) = y$ tương ứng là giá trị của đối tượng x_p, x_q trên thuộc tính a ($p \neq q$). Nếu tại thời điểm nào đó ta có $f(x_p, a) = f(x_q, a) = z$ ($z \notin V_a$) thì ta gọi hai giá trị w, y của thuộc tính a được làm thô thành giá trị mới z.

Định nghĩa 2.2

Cho hệ thông tin IS = (U, A, V, f), a $\in$ P $\subseteq$ A. Giả sử $Z = \{x_s \in U \mid f(x_s, a) = z\}$ là tập các đối tượng có giá trị z trên thuộc tính a. Nếu tại thời điểm nào đó, Z được phân hoạch thành hai tập hợp con W, Y sao cho Z = W $\cup$ Y, W $\cap$ Y = ϕ, trong đó $W = \{x_p \in U \mid f(x_p, a) = w, w \notin V_a\}$ và $Y = \{x_q \in U \mid f(x_q, a) = y, y \notin V_a\}$ thì ta gọi giá trị z của thuộc tính a là được làm mịn thành hai giá trị mới là w và y.

2.3 Tiến trình cập nhật tri thức khi làm thô, làm mịn các giá trị thuộc tính

2.3.1 Yêu cầu và giả thiết bài toán

Cho bảng quyết định DS = (U, C $\cup$ D, V, f), V_a, V_d tương ứng là tập các giá trị của thuộc tính điều kiện a và thuộc tính quyết định d. Yêu cầu đặt ra là: Trích rút các luật quyết định sau khi làm thô, làm mịn các giá trị của thuộc tính điều kiện hoặc thuộc tính quyết định,

các luật quyết định được rút ra thỏa mãn đồng thời cả ngưỡng độ chính xác và ngưỡng độ phủ cho trước.

Giả sử tập thuộc tính quyết định D chỉ gồm một thuộc tính d, tiến trình học các luật quyết định khi các giá trị thuộc tính thay đổi diễn ra từ thời điểm t đến thời điểm t+1; U/C = $\{C_1,\ldots,C_m\}$, U/D = $\{D_1, \ldots,D_n\}$ tương ứng là các phân hoạch được sinh bởi C, D ($0<m, n\leq|U|$);

Tại thời điểm t, ký hiệu $f_t(x, a)$, $f_t(C_i, a)$ và $f_t(x, d)$, $f_t(D_j, d)$ tương ứng là giá trị của x, giá trị của lớp tương đương điều kiện C_i trên thuộc tính a và là giá trị của x, giá trị của lớp tương đương quyết định D_j trên thuộc tính d.

Tương tự, tại thời điểm t+1, ta cũng ký hiệu lần lượt các giá trị này là $f_{t+1}(x, a)$, $f_{t+1}(C_i, a)$ và $f_{t+1}(x, d)$, $f_{t+1}(D_j, d)$.

2.3.2 Cơ sở toán học

2.3.2.1 Làm thô các giá trị thuộc tính điều kiện

Định lý 2.1:

Giả sử sau thời điểm t, hai giá trị w, y của thuộc tính $a \in C$ được làm thô thành giá trị mới z, $z \notin V_a$. Tại thời thời điểm t+1, tồn tại hai lớp tương đương điều kiện C_p, C_q nào đó được làm thô thành lớp tương đương điều kiện mới C_s, khi và chỉ khi $\forall a_j \neq a$, $f_t(C_p, a_j) = f_t(C_q, a_j)$.

Hệ quả 2.1:

Nếu sau thời điểm t, hai lớp tương đương điều kiện C_p, C_q nào đó được làm thô thành lớp điều kiện mới C_s thì tại thời điểm t+1, ta có: (i) $C_p \cup C_q = C_s$; (ii) $\forall D_j \in U/D$, $Sup(C_p, D_j) + Sup(C_q, D_j) = Sup(C_s, D_j)$ ở đây j=1,...,n.

2.3.2.2 Làm mịn các giá trị thuộc tính điều kiện

Định lý 2.2:

Một số nhận xét về thuật toán đề xuất và thuật toán của Chen

	Thuật toán đề xuất trong luận án	Thuật toán của Chen
Kết quả đầu ra	Các luật quyết định có ý nghĩa	Các xấp xỉ dưới, xấp xỉ trên của một khái niệm (một lớp quyết định)
Phạm vi đề cập của thuật toán	- Xem xét cả 4 trường hợp: làm thô, làm mịn các giá trị của một thuộc tính điều kiện và làm thô, làm mịn các giá trị thuộc tính quyết định. - Thao tác với tất cả n lớp tương đương quyết định.	- Chỉ xem xét đến trường hợp làm mịn các giá trị của một số thuộc tính điều kiện. - Thao tác chỉ với một lớp tương đương quyết định.
Phương pháp thực hiện	Không thực hiện lại việc phân lớp khi các giá trị thuộc tính thay đổi	Thực hiện lại việc phân lớp khi các giá trị thuộc tính thay đổi.
Độ phức tạp	$O(\|U\|^2)$	$O(\|U\|^2)$

Nếu cả 2 thuật toán khi cùng giải quyết vấn đề làm mịn các giá trị của |C| thuộc tính điều kiện với n lớp quyết định, đồng thời cùng bỏ qua bước trích rút luật và coi độ phức tạp của bước này trong cả hai thuật toán là tương tự nhau. Khi đó, thuật toán của Chen mở rộng với n lớp quyết định có độ phức tạp $O(|U|^3)$, thuật toán được đề xuất trong luận án khi mở rộng với |C| thuộc tính điều kiện có độ phức tạp $O(|C||U|^2)$.

2.4 Kết luận chương 2

Trong chương này trình bày các khái niệm làm thô, làm mịn các giá trị thuộc tính. Đưa ra và chứng minh một số tính chất làm cơ sở đề xuất thuật toán phát hiện các luật quyết định có ý nghĩa khi các giá

Thuật toán 2.5: Tính ma trận độ hỗ trợ tại thời điểm t+1 khi làm mịn các giá trị thuộc tính quyết định.

Vào: - Ma trận Sup tại thời điểm t;

- Tập D_w là tập các đối tượng có giá trị trên thuộc tính d là z được làm mịn thành giá trị w;

- Tập D_y là tập các đối tượng có giá trị trên thuộc tính d là z được làm mịn thành giá trị y;

Ra: Ma trận Sup tại thời điểm t+1 sau khi làm mịn d;

Phương pháp:

- Tìm lớp D_{z*} nào đó được tách thành 2 lớp D_y, D_w mới

- Tính ma trận Sup tại thời điểm t+1

Kết thúc.

Thuật toán 2.6: Tính ma trận độ chính xác và ma trận độ phủ tại thời điểm t+1

Vào: Ma trận độ hỗ trợ tại thời điểm t+1

Ra: Ma trận độ chính xác và ma trận độ phủ tại thời điểm t+1.

Phương pháp:

Áp dụng mệnh đề 1.1

Kết thúc.

Thuật toán 2.7: Trích rút luật quyết định có ý nghĩa

Vào: - Ma trận Acc, Cov tại thời điểm t+1; các ngưỡng α, γ

Ra: Các luật quyết định có ý nghĩa

Phương pháp:

Áp dụng định nghĩa 1.7

Kết thúc.

2.3.4 Độ phức tạp thuật toán

Độ phức tạp thời gian của thuật toán trích rút các luật quyết định có nghĩa khi làm thô, làm mịn các giá trị thuộc tính là $O(|U|^2)$ và độ phức tạp không gian của nó là $O(|U|^2)$.

Giả sử sau thời điểm t, giá trị z của thuộc tính $a \in C$ được làm mịn thành hai giá trị mới w và y ($w, y \notin V_a$). Tại thời điểm t+1, tồn tại một lớp tương đương điều kiện C_s nào đó được làm mịn thành hai lớp tương đương điều kiện mới C_p, C_q, khi và chỉ khi: (i) $f_t(C_s, a) = z$; (ii) $C_s \cap W \neq \phi$ với $W = \{x \in C_s : f_{t+1}(x, a) = w\}$; (iii) $C_s \cap Y \neq \phi$ với $Y = \{x \in C_s : f_{t+1}(x, a) = y\}$.

Hệ quả 2.2:

Nếu sau thời điểm t, lớp tương đương điều kiện C_s nào đó được làm mịn thành hai lớp tương đương điều kiện mới C_p, C_q. Tại thời điểm t+1 ta có: (i) $C_s = C_p \cup C_q$; (ii) $\forall D_j \in U/D$, $Sup(C_s, D_j) = Sup(C_p, D_j) + Sup(C_q, D_j)$ ở đây j=1,...,n

2.3.2.3 Làm thô các giá trị thuộc tính quyết định

Giả sử sau thời điểm t, hai giá trị w, y của thuộc tính quyết định d được làm thô thành giá trị mới z ($z \notin V_d$). Tại thời điểm t+1, tồn tại hai lớp tương đương quyết định D_w, D_y nào đó được làm thô thành một lớp tương đương quyết định mới D_z, có nghĩa là $D_w \cup D_y = D_z$. với $D_w = \{x \in U : f_t(x, d) = w, w \notin V_d\}$, $D_y = \{x \in U: f_t(x, d) = y, y \notin V_d\}$

Hệ quả 2.3 :

$\forall C_i \in U/C$ ta có: $Sup(C_i, D_w) + Sup(C_i, D_y) = Sup(C_i, D_z)$ ở đây i = 1,...,m.

2.3.2.4 Làm mịn các giá trị thuộc tính quyết định

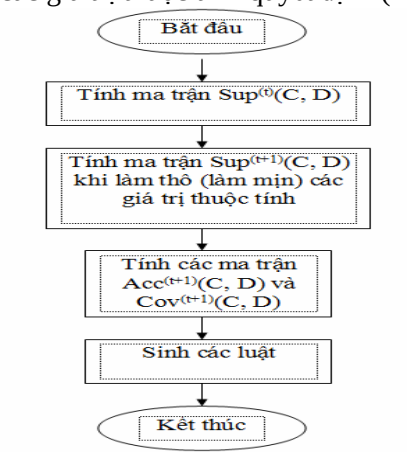
Giả sử sau thời điểm t, giá trị z của thuộc tính quyết định d được làm mịn thành hai giá trị mới w và y ($w, y \notin V_d$). Tại thời điểm t+1, tồn tại một lớp tương đương quyết định D_z nào đó được làm mịn thành hai lớp tương đương quyết định mới D_w và D_y.

Hệ quả 2.4:

$\forall C_i \in U/C$, ta có $Sup(C_i, D_z) = Sup(C_i, D_w) + Sup(C_i, D_y)$ ở đây i = 1,...,m.

2.3.3 Thuật toán

Các bước cơ bản của thuật toán trích rút các luật quyết định có ý nghĩa khi làm thô, làm mịn các giá trị thuộc tính điều kiện hoặc khi làm thô, làm mịn các giá trị thuộc tính quyết định (hình 2.1).



Hình 2.1: Các bước cơ bản của thuật toán trích rút luật quyết định khi làm thô/mịn các giá trị thuộc tính.

Các thuật toán để thực hiện các bước này được trình bày dưới đây.

Thuật toán 2.1 Tính ma trận độ hỗ trợ tại một thời điểm t nào đó

Vào:- Các lớp tương đương điều kiện C_i

- Các lớp tương đương quyết định D_j

Ra: Ma trận độ hỗ trợ (Sup) tại thời điểm t

Phương pháp :

Áp dụng định nghĩa 1.5

Kết thúc.

Thuật toán 2.2: Tính ma trận độ hỗ trợ tại thời điểm t+1 khi làm thô các giá trị thuộc tính điều kiện

Vào:- Ma trận độ hỗ trợ Sup tại thời điểm t

- Thuộc tính điều kiện a* được làm thô

- Các giá trị w, y của a* được làm thô thành z

Ra: Ma trận độ hỗ trợ Sup tại thời điểm t+1 sau khi làm thô thuộc tính a*;

Phương pháp:

- Tìm tất cả các cặp lớp tương đương điều kiện C_p, C_q nào đó được hợp thành lớp tương đương điều kiện C_s mới

- Tính ma trận Sup tại thời điểm t+1

Kết thúc.

Thuật toán 2.3 Tính ma trận độ hỗ trợ tại thời điểm t+1 khi làm mịn các giá trị thuộc tính điều kiện

Vào: - Ma trận Sup tại thời điểm t

- Thuộc tính điều kiện a* được làm mịn

- Tập W các đối tượng mà có giá trị z trên thuộc tính a* được làm mịn thành w

- Tập Y các đối tượng có giá trị z trên thuộc tính a* được làm mịn thành y

Ra: Ma trận Sup tại thời điểm t+1 sau khi làm mịn thuộc tính a*;

Phương pháp:

- Tìm lớp điều kiện C_s nào đó được tách thành 2 lớp mới C_p, C_q

- Tính Sup tại thời điểm t+1

Kết thúc.

Thuật toán 2.4 Tính ma trận độ hỗ trợ tại thời điểm t+1 khi làm thô các giá trị thuộc tính quyết định

Vào: - Ma trận Sup tại thời điểm t ;

- Giá trị w, y của thuộc tính quyết định d được làm thô thành z

Ra: Ma trận Sup tại thời điểm t+1 sau khi làm thô thuộc tính d;

Phương pháp:

- Tìm 2 lớp D_{w*}, D_{y*} nào đó được kết hợp thành lớp mới D_z

- Tính ma trận Sup tại thời điểm t+1

Kết thúc.