Tạp chí Tin học và Điều khiển học, T. 18, S. 4 (2002), 309–316

CÁCH TIẾP CẬN TẬP THÔ TRONG VIỆC PHÁT HIỆN TRI THỨC TRONG CƠ SỞ DỮ LIỆU

NGUYỄN ĐĂNG KHOA

Abstract. In this paper we present the state and perspective of rough sets in knowledge database discovery (KDD). We concentrate also on some searching methods for global elementary block, related with data extraction which follows the description of each decision class. Besides of some algorithms for template generation, we introduce an improved method for short reducts generation.

Tóm tắt. Báo cáo đề cập đến hiện trạng và triển vọng của tập thô trong việc phát hiện tri thức trong cơ sở dữ liệu phục vụ trợ giúp quyết định. Theo cách tiếp cận tập thô, chúng tôi tập trung vào một số bài toán tìm kiếm những khối sơ cấp có liên quan đến trích chọn dữ liệu để từ đó tìm ra các mô tả của mỗi lớp quyết định. Bên cạnh một số thuật toán tạo sinh mẫu, chúng tôi cũng giới thiệu một giải thuật cải tiến cho việc tìm kiếm rút gọn.

1. MỞ ĐẦU

Triết lý tập thô là một cách tiếp cận toán học mới cho tính không chính xác, tính mơ hồ và tính không chắc chắn. Quan hệ không phân biệt được là cơ sở toán học của lý thuyết tập thô. Trong lý thuyết tập thô, mỗi khái niệm không chính xác được thay thế bởi một cặp các khái niệm chính xác được gọi là xấp xỉ dưới và xấp xỉ trên ([9]).

Lý thuyết tập thô ngày càng được nghiên cứu mạnh mẽ và có nhiều ứng dụng trong các lĩnh vực học máy, thu nhận tri thức, phân tích quyết định, phát hiện tri thức từ các CSDL, các hệ chuyên gia, lập luận quy nạp và nhận dạng mẫu. Nó cũng giữ một vai trò quan trọng đặc biệt đối với các hệ trợ giúp quyết định ([10]).

2. CÁC KHÁI NIỆM CƠ BẢN CỦA LÝ THUYẾT TẬP THÔ

2.1. Các xấp xỉ và hàm thành viên thô

Giả sử có cơ sở tri thức $K = (U, \mathbf{R})$, trong đó $U \neq \varnothing$ là một tập hữu hạn được gọi là vũ trụ và $\mathbf{R}$ là một họ các quan hệ tương đương trên U. Với mỗi tập con $X \subseteq U, R \in \mathbf{R}$, xấp xỉ R-dưới và xấp xỉ R-trên của X được định nghĩa theo thứ tự như sau:

$$\underline{R}(X) = \{x \in U : R(x) \subseteq X\},$$

$$\overline{R}(X) = \{x \in U : R(x) \cap X \neq \varnothing\},$$

Trong đó $R(x)$ ký hiệu tập tất cả các đối tượng không phân biệt được với x, nghĩa là lớp tương đương chứa x được xác định bởi quan hệ tương đương R.

$$POS_R(X) = \underline{R}(X), \quad R - \text{miền dương (miền khẳng định) của } X$$

$$NEG_R(X) = U - \overline{R}(X), \quad R - \text{miền âm (miền phủ định) của } X$$

$$BN_R(X) = \overline{R}(X) - \underline{R}(X), \quad R - \text{miền đường biên của } X$$

Tập thô X được đặc trưng bởi hệ số $\alpha_R(X)$ (gọi là độ chính xác của xấp xỉ) được xác định như sau:

$$\alpha_R(X) = \frac{card\ \underline{R}(X)}{card\ \overline{R}(X)} = \frac{|\underline{R}(X)|}{|\overline{R}(X)|}$$

trong đó $X \neq \varnothing$ và $|X|$ ký hiệu lực lượng của tập X.

Hiển nhiên $0 \leq \alpha_R(X) \leq 1$ với mỗi $R \in \mathbf{R}$ và $X \subseteq U$. Nếu $\alpha_R(X) = 1$ thì miền đường biên của X là rỗng và tập X là chính xác đối với R; ngược lại nếu $\alpha_R(X) < 1$ thì tập X có miền đường biên nào đó và là thô (còn gọi là mơ hồ) đối với R.

Hàm thành viên thô có thể được định nghĩa bằng cách dùng quan hệ tương đương (quan hệ không phân biệt được) là:

$$\mu_X^R(x) = \frac{|X \cap R(x)|}{|R(x)|},$$

hiển nhiên, $0 \leq \mu_X^R(x) \leq 1$.

Hàm thành viên thô còn dùng để định nghĩa các xấp xỉ và miền biên của một tập:

$$\underline{R}(X) = \{x \in U : \mu_X^R(x) = 1\}; \quad \overline{R}(X) = \{x \in U : \mu_X^R(x) > 0\};$$

$$BN_R(X) = \{x \in U : 0 < \mu_X^R(x) < 1\}.$$

Các định nghĩa trên cho thấy ở đây tồn tại một liên kết chặt giữa tính mơ hồ và tính không chắc chắn trong lý thuyết tập thô. Tính mơ hồ liên quan tới các tập, trong khi tính không chắc chắn liên quan tới các phần tử của tập. Do vậy các xấp xỉ là cần thiết khi nói về các khái niệm mơ hồ, trong khi hàm thành viên thô lại cần đến khi xem xét dữ liệu không chắc chắn.

2.2. Rút gọn tri thức và các phụ thuộc

Các đối tượng có thể được phân lớp theo nhiều mẫu, cho nên ta có thể có một họ các quan hệ không phân biệt được $\mathbf{R} = \{R_1, R_2, ..., R_n\}$ trên vũ trụ U.

Giao của các quan hệ tương đương $\{R_1, R_2, ..., R_n\}$ ký hiệu: $\bigcap \mathbf{R} = \bigcap_{i=1}^{n} R_i$ cũng là một quan hệ tương đương. Trong trường hợp này, các tập sơ cấp là các lớp tương đương của quan hệ tương đương $\bigcap \mathbf{R}$. Vì các tập sơ cấp xác định duy nhất tri thức của chúng ta về vũ trụ, nên một số mẫu phân lớp có thể được loại bỏ mà vẫn không làm thay đổi họ các tập sơ cấp, nói cách khác là bảo toàn tính không phân biệt được. Tập cực tiểu **R'** của **R** sao cho $\bigcap \mathbf{R} = \bigcap \mathbf{R'}$ sẽ được gọi là một rút gọn của **R**. Tất nhiên **R** có thể có nhiều rút gọn và có nhiều phương pháp tìm các rút gọn.

Vấn đề quan trọng khác là mối quan hệ giữa các mẫu phân lớp khác nhau, nghĩa là giữa các quan hệ tương đương trong họ **R**. Ta nói rằng quan hệ không phân biệt được R phụ thuộc vào quan hệ không phân biệt được R' ký hiệu $R' \to R$, nếu $R' \subseteq R$, nghĩa là mỗi lớp của R' được chứa trong một lớp tương đương nào đó của R. Nói cách khác, điều đó nghĩa là nếu $R' \to R$, thì mỗi tập sơ cấp được sinh bởi R' được chứa trong một tập sơ cấp nào đó được sinh bởi R, nghĩa là mọi hạt nhân của tri thức liên kết với R' là một phần của hạt nhân nào đó của tri thức liên kết với R. Do vậy, sự phụ thuộc giải thích mối quan hệ giữa nhiều mẫu phân lớp ([5, 10]).

2.3. Hệ thông tin và bảng quyết định

Một *hệ thông tin* là một cặp **A**$= (U, A)$ trong đó U là tập không rỗng được gọi là vũ trụ và A là tập không rỗng các thuộc tính. Mọi $a \in A$ là một ánh xạ, $a : U \to V_a$, trong đó V_a là *tập trị* của a. Một *bảng quyết định* là một hệ thông tin dạng $\mathbf{A} = (U, A \cup \{d\})$, trong đó $d \notin A$ là một thuộc tính riêng được gọi là *quyết định*. Quyết định d xác định phân hoạch $\{X_1 ... X_{r(d)}\}$ của vũ trụ U, trong đó $X_i = \{x \in U : d(x) = i\}$ với $i \in V_d$ và $r(d) = |V_d|$. Tập X_i được gọi là *lớp quyết định thứ i* của **A**.

+ Với $B \subseteq A$, *quan hệ B -không phân biệt được* được định nghĩa bởi $IND(B) = \{(x, x') \in U \times U,$ với mọi $a \in B,\ a(x) = a(x')\}$.
+ *Tập các rút gọn* được ký hiệu bởi $RED(\mathbf{A})$ và chứa tất cả các tập con cực tiểu (đối với phép lấy bao hàm) $B \subseteq A$ sao cho $IND(A) = IND(B)$.

Dùng các phép toán xấp xỉ trên và xấp xỉ dưới trong lý thuyết tập thô người ta có thể giải quyết một loạt các bài toán cơ bản như mô tả các đối tượng theo các trị thuộc tính, các phụ thuộc giữa các trị thuộc tính, rút gọn các thuộc tính và tạo sinh các luật quyết định... Ví dụ sau với dữ liệu về 6 công chức minh hoạ những điều nói ở trên (Bảng 1).

Mỗi hàng của bảng có thể được xem như thông tin về một công chức cụ thể. Ví dụ công chức CC1 được đặc trưng trong bảng bởi tập trị thuộc tính sau: (ĐH, Không), (CM, Có), (QL, Cao), (LĐ, Có) tạo nên thông tin về công chức CC1. Trong bảng:

+ CC2, CC3, và CC5 là không phân biệt được đối với thuộc tính ĐH;
+ CC3 và CC6 là không phân biệt được đối với các thuộc tính CM và LĐ;
+ CC2 và CC5 là không phân biệt được đối với các thuộc tính ĐH, CM và QL.

Công chức	Tốt nghiệp đại học (ĐH)	Năng lực chuyên môn (CM)	Trình độ quản lý (QL)	Xét bổ nhiệm lãnh đạo (LĐ)
CC1	Không (K)	Có (C)	Cao	Có
CC2	Có	Không	Cao	Có
CC3	Có	Có	Rất cao	Có
CC4	Không	Có	Bình thường	Không
CC5	Có	Không	Cao	Không
CC6	Không	Có	Rất cao	Có

Vì thế ví dụ thuộc tính ĐH sinh ra hai tập sơ cấp {CC2, CC3, CC5} và {CC1, CC4, CC6}, trong khi các thuộc tính ĐH và CM tạo thành các tập sơ cấp {CC1, CC4, CC6}, {CC2, CC5} và {CC3}. Tương tự người ta có thể định nghĩa tập sơ cấp được sinh bởi tập bất kỳ của các thuộc tính. *Hệ số đo độ chính xác* của khái niệm "xét bổ nhiệm lãnh đạo lãnh đạo-LĐ" là:

$$\alpha_R(\text{Xét bổ nhiệm lãnh đạo}) = \frac{|\{CC1, CC3, CC6\}|}{|\{CC1, CC2, CC3, CC5, CC6\}|} = 3/5$$

và với khái niệm "không xét bổ nhiệm lãnh đạo" ta có

$$\alpha_R(\text{Không xét bổ nhiệm lãnh đạo}) = \frac{|\{CC4\}|}{|\{CC2, CC4, CC5\}|} = 1/3$$

Ta cũng có thể tính trị thành viên cho mỗi công chức theo khái niệm "xét bổ nhiệm lãnh đạo" hoặc "không xét bổ nhiệm lãnh đạo". Trị hàm thành viên của công chức CC1, CC2 được tính như sau:

$$\mu^R_{LD}(CC1) = \frac{|\{CC1, CC2, CC3, CC6\} \cap \{CC1\}|}{|\{CC1\}|}$$
$$\mu^R_{LD}(CC2) = \frac{|\{CC1, CC2, CC3, CC6\} \cap \{CC1, CC5\}|}{|\{CC2, CC5\}|} = 1/2$$

Tương tự ta có $\mu^R_{LD}(CC3) = 1$; $\mu^R_{LD}(CC4) = 0$; $\mu^R_{LD}(CC5) = 1/2$; $\mu^R_{LD}(CC6) = 1$.

Việc giải thích các hệ số của độ chính xác cũng như các độ thuộc dành cho các độc giả quan tâm.Vấn đề loại bỏ các thuộc tính dư thừa được gọi là rút gọn toàn bộ tập các thuộc tính. Người ta có thể tính rằng đối với ví dụ chỉ ra ở bảng 1, ta có hai rút gọn: {ĐH, QL} và {CM, QL}. Điều đó nghĩa là hoặc thuộc tính ĐH hoặc CM có thể được loại khỏi bảng mà không thay đổi các tập sơ cấp của nó.

3. NHỮNG VẤN ĐỀ KHAI PHÁ DỮ LIỆU TRONG TẬP THÔ

3.1. Một số giải pháp cho bài toán khai phá dữ liệu

+ *Dữ liệu cực lớn*: Rút gọn tập dữ liệu theo hàng ngang bằng cách trộn các bộ giống nhau sau khi thay thế một trị thuộc tính bằng trị mức cao hơn trong một hệ thống thứ bậc tổng quát của các thuộc tính phạm trù hoặc theo hàng dọc bằng cách áp dụng một số phương pháp lựa chọn đặc điểm. Một trong các chiến lược được dùng là *cách tiếp cận hướng thuộc tính* cho việc học khái niệm quy nạp ([1]).

+ *Dữ liệu nhiễu* (Noisy Data): Phương pháp phát hiện tri thức có thể ít bị ảnh hưởng bởi nhiễu trong tập dữ liệu khi việc nghiên cứu dựa trên sự biến thiên của các cây quyết định quy nạp, phụ thuộc vào nơi và số lượng nhiễu xuất hiện.

+ *Các trị Null* (Null Values): Trị null không chỉ có nghĩa là một trị chưa biết mà còn có thể là trị *không thể áp dụng được*. Một số nghiên cứu gần đây về việc xử lý với các trị null có thể tìm thấy trong [3].

+ *Dữ liệu không đầy đủ hoặc dữ liệu dư thừa*: Khi dữ liệu không đầy đủ, mô hình phát hiện tri thức có thể có khả năng cung cấp các quyết định gần đúng với mức tin cậy nào đó. Có nhiều giải pháp gần tối ưu hoặc tối ưu, với độ phức tạp thời gian hợp lý, nhằm loại bỏ các thuộc tính dư thừa (hoặc không quan trọng) khỏi tập các thuộc tính đã cho bằng cách sử dụng các trọng số của các thuộc tính riêng biệt hoặc kết hợp một số thuộc tính.

+ *Dữ liệu động*: Hầu hết các CSDL trực tuyến là *động*, ta khó có thể chụp ảnh nhanh loại dữ liệu này. Các hệ thống CSDL đang hoạt động luôn sẵn sàng cung cấp các điều kiện dễ dàng kích trigger (hoặc các luật $if - then$) có thể được dùng nhiều để thực hiện các phương pháp phát hiện dữ liệu ngày càng gia tăng.

3.2. Các truy vấn khai phá cơ sở dữ liệu

+ *Truy vấn phụ thuộc dữ liệu*: Các phụ thuộc dữ liệu trong DBMS được dùng để rút gọn số lượng các thuộc tính đã cho trong một tập dữ liệu hoặc xây dựng một đồ thị phụ thuộc dữ liệu. Truy vấn phụ thuộc dữ liệu thông qua việc xác định các kết hợp trọng số của các thuộc tính chắc chắn, được gọi là kiểm định giả thiết (*hypothesistesting*).

+ *Truy vấn phân lớp*: Loại truy vấn này bao gồm việc sinh một hàm phân lớp (còn được hiểu là sinh một phân lớp, khái niệm học hoặc phân biệt mô tả các lớp), nó phân hoạch một tập các bộ cho trước thành các lớp con có ý nghĩa đối với các nhãn được định nghĩa bởi người dùng hoặc các trị của một số thuộc tính quyết định.

+ *Truy vấn phân cụm*: Ta gọi phân hoạch không giám sát được của các bộ trong một bảng quan hệ là một truy vấn phân cụm. Có nhiều thuật toán phân cụm trải từ các phương pháp truyền thống của nhận dạng mẫu đến các kỹ thuật phân cụm trong học máy. Các tham số được định nghĩa bởi người dùng như số cực đại các bộ trong một phân cụm hoặc số các phân cụm có thể ảnh hưởng tới kết quả của một truy vấn phân cụm. Một phân lớp có thể được thiết kế thành một tập các mẫu nhỏ hơn được gán nhãn và sau đó cho phép nó hoạt động mà không có sự giám sát trên một tập các bộ lớn không được gán nhãn. Có thể ứng dụng các kỹ thuật phân cụm tương hỗ nhằm liên kết sức mạnh của các máy tính với tri thức của con người ([2]).

4. HIỆN TRẠNG CỦA TẬP THÔ TRONG KHAI PHÁ DỮ LIỆU

Lý thuyết tập thô dựa trên hoặc các không gian xấp xỉ đại số hoặc các không gian xấp xỉ xác suất, được dùng để lập luận về dữ liệu chứa thông tin không chắc chắn cho nhiệm vụ phát hiện dữ liệu đặc biệt. Dưới đây, chúng tôi nhìn lại hiện trạng của tập thô đối với vấn đề khai phá dữ liệu.

Lý thuyết tập thô dựa trên giả thuyết vũ trụ được bàn luận (hoặc tập các đối tượng) là hữu hạn, nó xem xét một chụp ảnh nhanh của một CSDL, nó không thể là một thừa nhận đúng, nếu tri thức nền thực chất là *động*. Một liệu pháp hợp lý cho vấn đề này là thiết kế một phương pháp tăng dần và tách tóm tắt và kết quả dùng cho bước sau. Trong [2], trình bày một phương pháp phân lớp

bên trên để phát triển một phân lớp thô tăng dần.

Trong không gian đại số,lý thuyết tập thô đã đưa ra các khái niệm sử dụng tập khái niệm trên và dưới. Tính không chắc chắn trong một tập dữ liệu là do dữ liệu bị *nhiễu* hoặc *không đầy đủ*, cách tiếp cận đối với loại dữ liệu này ít được mong đợi. Tuy nhiên có một số phương pháp xấp xỉ thô dựa trên các định nghĩa khác nhau về các miền dương (và miền biên) ([4]). Một tập sơ cấp được ánh xạ tới miền dương của một khái niệm chưa biết nếu độ thuộc của nó lớn hơn trị ngưỡng được định nghĩa bởi người dùng. Một tiếp cận khác có thể làm thay đổi miền của bài toán từ không gian đại số tới không gian xác xuất, nếu người ta có thể gán các phép đo xác xuất ưu tiên đối với các tập định nghĩa được ([6, 7]).

Dữ liệu dư thừa có thể được loại bỏ bằng cách bỏ bớt một số thuộc tính không quan trọng trong hệ thông tin sao cho tập thuộc tính của hệ được rút gọn là độc lập và không có thuộc tính nào có thể bị loại bỏ khỏi hệ thống mà không làm mất thông tin ([5]). Việc tìm kiếm toàn diện trên không gian các thuộc tính là hàm mũ với số các thuộc tính. Khi kiểm lại *các truy vấn khai phá dữ liệu* đối với phương pháp luận tập thô, ta thấy rằng phân tích phụ thuộc thuộc tính và phân lớp các chủ đề được quan tâm nghiên cứu hơn. *Kiểm định giả thiết và sự kết hợp giữa các trị của một thuộc tính* có thể được giải quyết một cách dễ dàng bởi phương pháp luận tập thô ([8]). Một bài viết có tính lý thuyết trong [4] mở rộng các khái niệm của xấp xỉ và sự bằng nhau thô cho việc phân tích khái niệm hình thức. Trong việc truy vấn khai phá CSDL bởi phương pháp luận tập thô, phân nhóm thô đối mặt với một vấn đề khi một đối tượng mới (đến từ bên ngoài tập dữ liệu) được đưa vào và việc mô tả của đối tượng là không tìm thấy trong lớp tương ứng. Nói cách khác, vấn đề là tìm ra các bao đóng của đối tượng cho trước để biết được khái niệm. Giải pháp thông thường cho vấn đề này là ánh xạ các trị không-định lượng thành một thang số (scale) và dùng một hàm khoảng cách cho việc đánh giá.

5. CÁC HƯỚNG TƯƠNG LAI

Như trình bày ở trên, một số khía cạnh của bản chất dữ liệu (nghĩa là dữ liệu không đầy đủ, dư thừa và không chắc chắn) đã được nghiên cứu trong phương pháp luận tập thô, nhưng chúng cần được trắc nghiệm trong các CSDL lớn. Đã có một số báo cáo về việc sử dụng phương pháp luận tập thô dựa trên các công cụ phát hiện tri thức trên các dữ liệu không trực tuyến như: KDD-R, một hộp công cụ mở có tính thí nghiệm; LERS, một hệ học máy từ các ví dụ và một Data/Logic/R, một sản phẩm thương mại cho việc khai phá CSDL và trợ giúp quyết định. Dưới đây, chúng tôi trình bày những hướng nghiên cứu tương lai, đó là các ứng dụng khai phá dữ liệu.

+ Các xấp xỉ thô gia tăng: Cần dự phòng thuật toán quyết định xử lý liên tục trong mô hình tập thô và tri thức nền là động. Các lược đồ phân lớp thô tiến hóa có thể được phát triển, nếu bảng quyết định được điều tiết với trường gia tăng phức hợp chứa tần suất các hàng.

+ Sự gần nhau của hai luật: Xác định luật gần nhất,trong trường hợp mô tả một đối tượng cho trước không sánh hợp với các đối tượng của khái niệm đã biết. Các *truy vấn phân cụm* là một chủ đề rất quan trọng cần được nghiên cứu trong cộng đồng tập thô.

+ Các trị Null: Trị Null của thuộc tính tổng quát hơn là trị chưa biết của thuộc tính đó và suy luận dẫn về các trị null còn là vấn đề ngỏ trong việc nghiên cứu khai phá dữ liệu. Các *trị thuộc tính chưa biết* đã được Grzlymala-Buse nghiên cứu và được thực hiện trong hệ học máy LERS.

+ Truy vấn đặc trưng hóa: Phân tích phụ thuộc dữ liệu được ứng dụng để đặc trưng các khái niệm, song nó còn thiếu một cấu trúc rõ ràng như hệ thống thứ bậc của các khái niệm bền (persistent) để khám phá các phụ thuộc. Chủ đề này đã được nghiên cứu và sử dụng cho mô hình khái niệm. Người ta đã nghiên cứu việc mở rộng phương pháp luận tập thô tới lĩnh vực phân tích khái niệm hình thức.

+ Các khía cạnh tính toán của phương pháp luận tập thô: Các ứng dụng khai thác dữ liệu đòi hỏi những kỹ thuật hiệu quả để tăng độ chính xác của thuật toán quyết định. Trong tương lai rất cần đến việc nghiên cứu toàn diện về chủ đề này.

Khai phá CSDL là một bài toán thực tế, nó dẫn dắt nghiên cứu lý thuyết hướng tới việc hiểu lập luận về dữ liệu *lớn* và *đang tồn tại*.

6. MỘT SỐ PHƯƠNG PHÁP TẠO SINH MẪU TIÊU BIỂU

Có nhiều phương pháp hiệu quả để tổng hợp các luật quyết định hoặc tìm ra các quy luật đối với các bảng dữ liệu tương đối nhỏ. Dưới đây, chúng tôi trình bày một phương pháp mới được dùng cho các bảng dữ liệu lớn, dựa trên sự phân tách các bảng dữ liệu lớn thành các bảng dữ liệu nhỏ hơn sao cho sự xấp xỉ của hàm quyết định toàn cục có thể thu được từ hàm quyết định cục bộ, liên quan đến các bảng nhỏ đó. Tập các bảng nhỏ hơn này có thể được xem như một tập các phần tử sinh (generators). Dùng các phần tử sinh là "các mẫu" được mô tả bởi các biểu thức *thuộc tính = trị* để thực hiện việc xấp xỉ đủ dùng cho hàm quyết định toàn cục. Chúng ta cũng cố gắng tìm kiếm các mẫu có chất lượng cao chẳng hạn được đặc trưng bởi số các đối tượng trợ giúp một mẫu nhân với số cặp *thuộc tính = trị* mô tả mẫu và mong đợi tìm ra các quy tắc (luật) mạnh cho các miền con có nhiều đặc điểm chung được tách từ một vũ trụ. Ở đây trình bày một giải thuật cải tiến cho việc tìm kiếm rút gọn ngắn trên dữ liệu được rút gọn và một số thuật toán tạo sinh mẫu.

Trong hệ thông tin $\mathbf{A} = (U, A)$ như đã nêu, *một mẫu* đối với $A = \{a_1, a_2, ..., a_n\}$ là một xâu $v_1...v_n$ trong đó $v_i \in V_{a_i} \cup \{*\}$ với mọi i và '*' là một ký hiệu "không cần quan tâm". Vị trí thứ i của một mẫu tương ứng với thuộc tính a_i. Nếu '*' xuất hiện ở vị trí thứ i thì nghĩa là mẫu "bỏ qua" thuộc tính a_i. Một đối tượng x là sánh được *một mẫu* cho trước nếu và chỉ nếu $a_i(x)$ là bằng phần tử thứ i của mẫu nếu phần tử này khác '*' với mọi i. Chúng ta ký hiệu $n_A(a, v)$ là số đối tượng trong $\mathbf{A}$ mà trên đó thuộc tính a có trị v. Tìm kiếm mẫu có thể được thực hiện theo nhiều cách như sau:

6.1. Tìm kiếm các mẫu với độ dài Max

Mục đích của phương pháp Max là tìm kiếm các mẫu với độ dài lớn và độ khớp không nhỏ hơn một cận dưới s nào đó. Ta tăng dần mẫu T bởi các mô tả $T \wedge (a = v_a)$ đến khi độ khớp của mẫu được sinh nhỏ hơn s. Khi chọn ngẫu nhiên một mô tả trong tổng số các mô tả theo một xác xuất nào đó $P(a = v_a) = (n_A(a, v_a))/\sum n_A(a, v_{a_i})$ ta thu được nhiều hơn một mẫu nhờ giải thuật sau:

1) $T = \varnothing$;
2) while $l(T) < m$ and $fitness_A(T) > s$ do
3) begin
   3.1) for mọi thuộc tính $a \notin T$ sắp xếp các đối tượng trong U theo các trị của a để xác định trị v_a thường hay xuất hiện trong bảng nhất, nghĩa là $n_A(a, v_a) = \max\{n_A(a, v) : v \in V_a\}$ trong đó V_a là tập trị của a;
   3.2) Chọn một cách ngẫu nhiên mô tả $a = v_a$ với xác xuất $P(a = v_a) = (n_A(a, v_a))/\sum n_A(a, v_{a_i})$
   3.3) U = tập các đối tượng trong U sánh hợp với mẫu $a = v_a$; $A = A - \{a\}$;
   3.4) $T = T \cup \{a = v_a\}$;
4) end

6.2. Tìm các mẫu bằng việc sử dụng đối tượng và các trọng số thuộc tính

Tất cả các đối tượng trong bảng quyết định được gán các trọng số thích hợp, mô tả tiềm năng của các đối tượng thuộc về một mẫu "tốt" theo một nghĩa nào đó.

+ Các trọng số phản ánh sự giống nhau của các đối tượng: Cho $\mathbf{A}= (U, A)$ và $x \in U$. Với mọi $y \in U$, ta tính $g_{x,y} = \{a : a(x) = a(y)\}$ nghĩa là số các thuộc tính có cùng trị trên x và y. Số này ký hiệu "tính gần" của y đối với x.
  Sau đó, với mọi thuộc tính $a \in A$, ta tính $w_a(x) = \sum_{y:a(x)=a(y)} g_{x,y}$ và cuối cùng tính trọng số $w(x) = \sum_{a \in A} w_a(x)$. Ta có $w(x) = \sum_y g^2_{x,y}$.

+ Các trọng số của các đối tượng được dẫn xuất từ tần suất trị thuộc tính: Cho $\mathbf{A}= (U, A)$ và $x \in U$. Với mọi $a \in A$ gọi $w_a(x) = n_A(a, a(x))$ và $w(x) = \sum_{a \in A} w_a(x)$.
  Các trọng số này cho phép việc phân cụm rất thỏa đáng các đối tượng vào các mẫu trong khi các trị "ngây thơ" hơn của trọng số làm giảm chất lượng các kết quả.

Ý tưởng sử dụng *trọng số các thuộc tính* cho việc tạo sinh mẫu rất giống phương pháp "các trọng số đối tượng". Cho **A**= (U, A), $m = |U|$, $n = |A|$. Ta có thể sắp thứ tự các trị thuộc tính của a theo giá trị $n_A(a, v)$ với bất kỳ $a \in A$. Ta ký hiệu v_i^a là trị thứ i của thuộc tính a trong trật tự này. Trị v_i^a là trị xuất hiện thường xuyên nhất của a trong A. Ta chọn ngẫu nhiên thứ tự giữa các trị v và u nếu $n_A(a, v) = n_A(a, u)$. Ta ký hiệu $w_A(a)$ là

$$\frac{m}{\sum_{i=1}^{|v_a|} i.n_A(a, v_i^a)}$$

Dễ dàng thấy $w_A(a) \in (0, 1]$. Với mọi trị u của thuộc tính a ta có thể định nghĩa trọng số của u là $w_A^a(u) = n_A(a, u)/m$. Ta có $w_A^a(u) \in (0, 1]$ và $\sum_{v \in V_a} w_A^a(v) = 1$ với mọi $a \in A$. Các giải thuật tạo sinh các mẫu tốt sử dụng các trọng số thuộc tính và các trọng số đối tượng được mô tả chi tiết trong [4].

6.3. Tìm các mẫu bằng cách sử dụng các giải thuật di truyền

Tìm kiếm các mẫu lớn theo phương pháp di truyền được mô tả chi tiết trong [4]. Các mẫu được biểu diễn bởi các xâu nhị phân có độ dài N (chỉ ra các thuộc tính nào được cố định) và mọi đối tượng sánh hợp với chúng (được gọi là một đối tượng cơ sở). Với một s thì khi cho trước một đối tượng cơ sở x_b, ta tìm một mẫu tốt nhất sánh hợp với x_b. Để làm được điều đó, ta sử dụng một *giải thuật tham lam* (greedy). Mỗi một mẫu cực đại cục bộ có thể được tìm thấy bằng cách sử dụng giải thuật này - kết quả phụ thuộc vào thứ tự các thuộc tính. Mục đích của ta là tìm ra thứ tự riêng của các thuộc tính và ta dùng các giải thuật di truyền để làm công việc này. Nhiễm sắc thể của ta sẽ là một hóa n vị τ có độ dài N. Trị của hàm khớp bằng kích thước của mẫu tốt nhất được tìm thấy.

6.4. Các mẫu được tổng quát hóa

Ý tưởng các mẫu có thể được mở rộng thành mẫu tổng quát hóa, nghĩa là các mẫu có dạng:

$$GT = (a_{i_1} = v_{i_1} \vee ... \vee a_{i_1} = v_{i_n}) \wedge ... \wedge (a_{j_k} = v_{j_1} \vee ... \vee a_{j_k} = v_{j_m}).$$

Sự khác nhau chính là ở chỗ thay vì một trị ta lại có nhiều vị trí đa trị của mẫu GT. Ta nói rằng đối tượng x thỏa mô tả được tổng quát hóa $a = v_1 \vee ... \vee a = v_m$ nếu trị của a trên x thuộc về tập $\{v_1, ..., v_m\}$. Một đối tượng x thỏa mẫu tổng quát GT nếu nó thỏa tất cả các mô tả của GT. Một sự mở rộng khác của ý tưởng này có thể được thực hiện bởi các mẫu với các mô tả không-rời rạc, nghĩa là:

$$a \in [v_{i_1}, v_{i_2}] \vee ... \vee a \in [v_{m_1}, v_{m_2}]$$

Các phương pháp tìm các mẫu trình bày ở trên có thể dễ dàng được chấp nhận để tìm kiếm các mẫu tổng quát hóa trong các bảng dữ liệu lớn. Trong trường hợp các mẫu toàn cục, người ta có thể biến đổi hàm khớp nếu $|V_a| > 1$ thì $s(a) = (|V_a| - k)/(V_a - 1)$ với mọi $a \in A$, trong đó k bằng độ dài của mô tả tổng quát của a; ngược lại thì $s(a) = 1$. Số $s(a)$ ký hiệu mức độ thuộc tính a không phải là trị ('*') "không cần quan tâm" tức là $s(a) = 0$, nghĩa là trên a ta có * và $s(a) = 1$ nghĩa là mẫu chỉ có một trị trên a. Do vậy trong giải thuật, ta có thể làm cực đại số các đối tượng thỏa mãn được nhân lên bởi $\sum_{a \in A} s(a)$.

6.5. Các mẫu quyết định

Ta có thể dùng các mẫu được sinh bởi các giải thuật tìm kiếm ở trên cho tính chính quy trong CSDL. Vậy tìm các luật ràng buộc các thuộc tính và các quyết định như thế nào. Giả sử cho một bảng quyết định **A**. Ta quan tâm đến việc mô tả lớp quyết định thứ i bởi một tập luật quyết định nghĩa là bởi giải thuật quyết định cho lớp này. Muốn vậy, ta tạo ra tập các mẫu phủ lớp quyết định, nghĩa là phần lớn các đối tượng trong lớp, sánh hợp một trong các mẫu, trong khi đối với các lớp khác, càng ít đối tượng sánh hợp với chúng thì càng tốt. Giải thuật di truyền được mô tả ở trên có

thể được thích nghi để tìm kiếm loại mẫu mới này: ta có thể dễ dàng thay đổi công thức cho việc khớp mẫu. Chất lượng của giải thuật quyết định có được bằng cách áp dụng phương pháp này phụ thuộc vào hai yếu tố: nó xấp xỉ tốt các lớp quyết định như thế nào và mô tả của nó dài bao nhiêu- Ta hướng tới sản xuất các luật càng đơn giản càng tốt. Các kết quả được trình bày trong [7] hình như thiên về phương pháp này. Sử dụng giải thuật mô tả trong [8], ta có thể phủ toàn bộ lớp quyết định với các mẫu. Ta xuất phát với một đối tượng ngẫu nhiên trong lớp làm đối tượng cơ bản, sau đó mẫu thu được được lưu giữ trong bộ nhớ. Đối tượng cơ bản tiếp theo được chọn một cách ngẫu nhiên trong lớp, nhưng ta không tính tới các đối tượng được phủ bởi các mẫu đã có sẵn rồi.

6.6. Quan hệ dung sai và trích chọn mẫu

Trong các mục trước ta đã gợi ý tìm kiếm các hình mẫu (pattern) theo dạng các mẫu. Loại khác của hình mẫu có thể được định nghĩa bởi các quan hệ dung sai, nghĩa là các quan hệ nhị phân phản xạ và đối xứng trong miền đối tượng. Ta xem xét các quan hệ dung sai được xây dựng từ các hàm tương tự được xác định trên các trị thuộc tính. Ta gọi quan hệ dung sai là tối ưu nếu nó bao gồm một tập cực đại các cặp đối tượng có cùng quyết định. Các chiến lược tìm kiếm một dung sai (một dung sai con tối ưu) trong các lớp quan hệ dung sai khác nhau được trình bày trong [5]. Có các hình mẫu, ta có thể phân tách bảng bằng cách nhóm lớp các đối tượng tương tự (theo nghĩa của quan hệ dung sai). Các quan hệ dung sai được trích chọn cũng có thể được dùng cho việc xấp xỉ các lớp quyết định (bởi việc xây dựng các cụm) và dùng phân loại các đối tượng mới.

7. KẾT LUẬN

Trên đây chúng tôi đã giới thiệu một số hướng trợ giúp quyết định theo cách tiếp cận tập thô và một số giải thuật hiệu quả cho việc tạo sinh mẫu. Xây dựng lược đồ tổng quát cho việc mô tả xấp xỉ các lớp quyết định và các phương pháp đã trình bày ở trên để tạo sinh mẫu nếu được kết hợp với các phương pháp phân tách các bảng dữ liệu lớn thành các bảng nhỏ hơn tỏ ra có nhiều hứa hẹn trong việc phát hiện tri thức và cần được tiếp tục nghiên cứu. Chúng ta tin rằng khai phá dữ liệu là một lĩnh vực ứng dụng, trong đó các nghiên cứu lý luận của lý thuyết tập thô có thể được khẳng định, giúp chúng ta hiểu rõ những mặt mạnh và những mặt yếu của lý thuyết này trong thực hành.

TÀI LIỆU THAM KHẢO

[1] J.W. Grzymala-Busse *On the unknow attribute values in learning from examples* 1991.
[2] J. S. Deogun, V.V.Raghavan and H.Sever. Rough set based classification methods and extended decision tables. *In Proc. of The Int. Workshop on Rough Sets and Soft* (1998).
[3] J. D. Ullman, *Principles of Database and Knowledge- Base Systems*, Vol. 1 Computer Science Press, Rockville, Maryland, 1988.
[4] Komorowski J., Pawlak Z, Polkowski L., Skowron A *A rough set Perspective on Data and Knowledge,* 1999.
[5] Nguyễn Đăng Khoa, Rút gọn tập thô, Tuyển tập Báo cáo tại Hội nghị Khoa học nhân dịp kỷ niệm 45 năm ngày thành lập trường ĐH Bách Khoa Hà Nội, 2001.
[6] Pawlak Z., *Rough sets:Theoretical aspects of reasoning about data*, Kluwer, Dortrecht 1991.
[7] Pawlak Z., *Rough sets: Present State and Further Prospect* 1995.
[8] Pawlak Z., *Rough sets approach to knowledge-based decision support* Artificial Intelligence, IOS Press, Amsterdam 1995 (220–238).
[9] S. H. Nguyen, L. Polkowski, A. Skowron, P. Synak, J.Wro' blewski *Searching for approximate description of decision classes,* 1997 (4–5).
[10] S. H. Nguyen, Skowron A., Synak P., Wróblewski *Discovery in Databases: Rough Set Approach, 1999.*

Nhận bài ngày 20 - 5 - 2002

Học viện Hành chính Quốc gia