ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

HOÀNG TIẾN HIẾU

RÚT GỌN THUỘC TÍNH VÀ TRÍCH LỌC LUẬT TRÊN BẢNG QUYẾT ĐỊNH KHÔNG ĐẦY ĐỦ DỰA TRÊN MÔ HÌNH TẬP THÔ DUNG SAI

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC
TS. Nguyễn Long Giang

Thái Nguyên – 2013

MỤC LỤC

Danh mục các thuật ngữ

Thuật ngữ tiếng Việt	**Thuật ngữ tiếng Anh**
Tập thô	*Rough Set*
Hệ thông tin	*Information System*
Hệ thông tin đầy đủ	*Complete Information System*
Hệ thông tin không đầy đủ	*Incomplete Information System*
Hệ thông tin không nhất quán	*Inconsistent Information System*
Bảng quyết định	*Decision Table*
Bảng quyết định đầy đủ	*Complete Decision Table*
Bảng quyết định không đầy đủ	*Incomplete Decision Table*
Bảng quyết định không nhất quán	*Inconsistent Decision Table*
Quan hệ không phân biệt được	*Indiscernibility Relation*
Quan hệ dung sai	*Tolerance Relation*
Xấp xỉ dưới	*Lower Approximation*
Xấp xỉ trên	*Upper Approximation*
Rút gọn thuộc tính	*Attribute Reduction*
Tập rút gọn	*Reduct*
Tập lõi	*Core*
Luật quyết định	*Decision Rule*
Khoảng cách	*Distance*

Danh sách bảng

MỞ ĐẦU

Lý thuyết tập thô - do Zdzislaw Pawlak [16] đề xuất vào những năm đầu thập niên tám mươi của thế kỷ hai mươi - được xem là công cụ hữu hiệu để giải quyết các bài toán phân lớp, phát hiện luật…chứa dữ liệu không đầy đủ, không chắc chắn. Từ khi xuất hiện, lý thuyết tập thô đã được sử dụng hiệu quả trong các bước của quá trình khai phá dữ liệu và khám phá tri thức, bao gồm tiền xử lý số liệu, khai phá dữ liệu và đánh giá kết quả thu được. Rút gọn thuộc tính và trích lọc luật quyết định (luật phân lớp) là hai ứng dụng chính của lý thuyết tập thô trong khai phá dữ liệu. Rút gọn thuộc tính thuộc giai đoạn tiền xử lý dữ liệu còn trích lọc luật thuộc giai đoạn khai phá dữ liệu. Mục tiêu của rút gọn thuộc tính là loại bỏ các thuộc tính dư thừa nhằm tím tập con nhỏ nhất của tập thuộc tính điều kiện (tập rút gọn) mà bảo toàn thông tin phân lớp của bảng quyết định. Dựa trên tập rút gọn thu được, việc sinh luật và phân lớp đạt hiệu quả cao nhất.

Trong các bài toán thực tế, các bảng quyết định thường thiếu giá trị trên miền giá trị thuộc tính, gọi là các bảng quyết định không đầy đủ. Trên bảng quyết định không đầy đủ, Kryszkiewicz [10] đã mở rộng quan hệ tương đương trong lý thuyết tập thô truyền thống thành quan hệ dung sai và đề xuất mô hình tập thô dung sai nhằm trích lọc luật trực tiếp không qua bước xử lý giá trị thiếu. Dựa trên mô hình tập thô dung sai, một số công trình công bố trong mấy năm gần đây đã đề xuất một số độ đo không chắc chắn nhằm giải quyết bài toán rút gọn thuộc tính và trích lọc luật, đáng chú ý là các công trình [6, 7, 8, 11, 13, 12, 23]. Tuy nhiên, các kết quả nghiên cứu về các phương pháp rút gọn thuộc tính và trích lọc luật trên các bảng quyết định không đầy đủ còn nhiều hạn chế.

Luận văn đặt ra hai mục tiêu chính:

1) Tổng hợp các công bố mới nhất về các phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ theo tiếp cận mô hình tập thô dung sai, bao gồm: phân nhóm các phương pháp và nghiên cứu mối liên hệ giữ các nhóm; nghiên cứu các độ đo đánh giá hiệu năng tập luật quyết định. Trên cơ sở đó, luận văn đề

xuất các độ đo cải tiến đánh giá hiệu năng tập luật quyết định và nghiên cứu sự thay đổi giá trị các độ đo này trên các tập rút gọn nhằm đánh giá các phương pháp rút gọn thuộc tính về mặt định lượng.

2) Tổng hợp các công bố về các phương pháp rút gọn thuộc tính sử dụng khoảng cách. Từ đó, xây dựng phương pháp rút gọn thuộc tính sử dụng khoảng cách Hamming (một trong những khoảng cách đơn giản và hiệu quả) và thử nghiệm phương pháp trên các bộ số liệu mẫu từ kho dữ liệu UCI.

Đối tượng nghiên cứu của luận văn là các *bảng quyết định không đầy đủ* với kích thước trung bình và kích thước lớn.

Phạm vi nghiên cứu của luận văn *tập trung vào bài toán rút gọn thuộc tính ở* bước tiền xử lý số liệu trong quá trình khai phá dữ liệu.

Phương pháp nghiên cứu của luận văn là nghiên cứu lý thuyết và nghiên cứu thực nghiệm. Về nghiên cứu lý thuyết: các mệnh đề được chứng minh chặt chẽ dựa vào các kiến thức cơ bản và các kết quả nghiên cứu đã công bố. Về nghiên cứu thực nghiệm: luận văn thực hiện cài đặt các thuật toán, chạy thử nghiệm thuật toán với các bộ số liệu lấy từ kho dữ liệu UCI, so sánh và đánh giá kết quả thực nghiệm so với kết quả nghiên cứu lý thuyết, từ đó kết luận tính đúng đắn của kết quả nghiên cứu.

Bố cục của luận văn gồm phần mở đầu và hai chương nội dung, phần kết luận và danh mục các tài liệu tham khảo.

Chương 1 trình bày các khái niệm cơ bản về mô hình tập thô truyền thống, phương pháp rút gọn thuộc tính trong mô hình tập thô truyền thống và mô hình tập thô mở rộng dựa trên quan hệ dung sai, phương pháp rút gọn thuộc tính trong mô hình tập thô dung sai. Tổng kết các công bố mới nhất về các phương pháp rút gọn thuộc tính, mối liên hệ, phân nhóm các phương pháp

Chương 2 đề xuất phương pháp rút gọn thuộc tính sử dụng khoảng cách Hamming. Trích lọc luật quyết định từ tập rút gọn theo phương pháp rút gọn thuộc tính sử dụng khoảng cách Hamming.

Chương 3 cài đặt, thử nghiệm, đánh giá phương pháp trên các bộ số liệu mẫu từ kho dữ liệu UCI.

Cuối cùng, phần kết luận nêu những đóng góp của luận văn, hướng phát triển tiếp theo.

Chương 1. RÚT GỌN THUỘC TÍNH THEO TIẾP CẬN LÝ THUYẾT TẬP THÔ

Mô hình tập thô truyền thống do Pawlak đề xuất [16] là công cụ hiệu quả để giải quyết bài toán phân lớp trên các hệ thông tin đầy đủ dựa trên quan hệ tương đương. Tuy nhiên trong thực tế, các hệ thông tin thường thiếu giá trị trên miền giá trị của thuộc tính, gọi là các hệ thông tin không đầy đủ. Trong hệ thông tin không đầy đủ, Kryszkiewicz [10] được xem là người đầu tiên mở rộng quan hệ tương đương thành quan hệ dung sai và xây dựng mô hình tập thô mở rộng dựa trên quan hệ dung sai, gọi là mô hình tập thô dung sai. Trong chương này, tôi trình bày các khái niệm cơ bản về mô hình tập thô truyền thống và mô hình tập thô dung sai.

1.1. Rút gọn thuộc tính theo tiếp cận mô hình tập thô truyền thống

1.1.1 Hệ thông tin đầy đủ và mô hình tập thô truyền thống

1) Hệ thông tin đầy đủ

Hệ thông tin là công cụ biểu diễn tri thức dưới dạng một bảng dữ liệu gồm p cột ứng với p thuộc tính và n hàng ứng với n đối tượng. Một cách hình thức, hệ thông tin được định nghĩa như sau.

Định nghĩa 1.1. Hệ thông tin là một bộ tứ $IS = U, A, V, f$ trong đó U là tập hữu hạn, khác rỗng các đối tượng; A là tập hữu hạn, khác rỗng các thuộc tính; $V = \bigcup_{a \in A} V_a$ với V_a là tập giá trị của thuộc tính $a \in A$; $f : U \times A \to V_a$ là hàm thông tin, $\forall a \in A, u \in U$ $f(u, a) \in V_a$.

Với mọi $u \in U$, $a \in A$, ta ký hiệu giá trị thuộc tính a tại đối tượng u là $a(u)$ thay vì $f(u, a)$. Nếu $B = \{b_1, b_2, ..., b_k\} \subseteq A$ là một tập con các thuộc tính thì ta ký hiệu bộ các giá trị $b_i(u)$ bởi $B(u)$. Như vậy, nếu *u* và *v* là hai đối tượng, thì ta viết $B(u) = B(v)$ nếu $b_i(u) = b_i(v)$ với mọi $i = 1, ..., k$.

Cho hệ thông tin $IS = U, A, V, f$, nếu tồn tại $u \in U$ và $a \in A$ sao cho $a(u)$ *thiếu giá trị* (missing value) thì *IS* được gọi là *hệ thông tin không đầy đủ*, trái lại *IS*

được gọi là *hệ thông tin đầy đủ*. Trong luận văn này, *hệ thông tin đầy đủ* được gọi tắt là *hệ thông tin*.

Xét hệ thông tin $IS = U, A, V, f$. Mỗi tập con các thuộc tính $P \subseteq A$ xác định một quan hệ hai ngôi trên U, ký hiệu là $IND\ P$, xác định bởi

$$IND\ P = \{u, v \in U \times U \mid \forall a \in P,\ a\ u = a\ v\}.$$

$IND\ P$ là quan hệ *P-không phân biệt được*. Dễ thấy rằng $IND\ P$ là một quan hệ tương đương trên U. Nếu $u, v \in IND\ P$ thì hai đối tượng u và v không phân biệt được bởi các thuộc tính trong P. Quan hệ tương đương $IND\ P$ xác định một phân hoạch trên U, ký hiệu là $U / IND\ P$ hay U / P. Ký hiệu lớp tương đương trong phân hoạch U / P chứa đối tượng u là u_P, khi đó $u_P = \{v \in U \mid u, v \in IND\ P\}$.

2) Mô hình tập thô truyền thống

Cho hệ thông tin $IS = U, A, V, f$ và tập đối tượng $X \subseteq U$. Với một tập thuộc tính $B \subseteq A$ cho trước, chúng ta có các lớp tương đương của phân hoạch U / B, thế thì một tập đối tượng X có thể biểu diễn thông qua các lớp tương đương này như thế nào?

Trong lý thuyết tập thô, để biểu diễn X thông qua các lớp tương đương của U / B (còn gọi là biểu diễn X bằng tri thức có sẵn B), người ta xấp xỉ X bởi hợp của một số hữu hạn các lớp tương đương của U / B. Có hai cách xấp xỉ tập đối tượng X thông qua tập thuộc tính B , được gọi là *B-xấp xỉ dưới* và *B-xấp xỉ trên* của X, ký hiệu là lượt là $\underline{B}X$ và $\overline{B}X$, được xác định như sau:

$$\underline{B}X = \{u \in U \mid u_B \subseteq X\}\ ,\ \overline{B}X = \{u \in U \mid u_B \cap X \neq \varnothing\}.$$

Tập $\underline{B}X$ bao gồm tất cả các phần tử của U chắc chắn thuộc vào X, còn tập $\overline{B}X$ bao gồm các phần tử của U có thể thuộc vào X dựa trên tập thuộc tính B. Từ hai tập xấp xỉ nêu trên, ta định nghĩa các tập

$$BN_B\ X = \overline{B}X - \underline{B}X : \textit{B-miền biên của } X,\quad U - \overline{B}X : \textit{B-miền ngoài của } X.$$

B-miền biên của *X* là tập chứa các đối tượng có thể *thuộc* hoặc *không thuộc X*, còn *B-miền ngoài* của *X* chứa các đối tượng *chắc chắn không thuộc X*. Sử dụng các lớp của phân hoạch *U/B*, các xấp xỉ dưới và trên của *X* có thể viết lại

$$\underline{B}X = \bigcup Y \in U/B \,|\, Y \subseteq X \quad , \quad \overline{B}X = \bigcup Y \in U/B \,|\, Y \cap X \neq \varnothing \ .$$

Trong trường hợp $BN_B\ X = \varnothing$ thì *X* được gọi là *tập chính xác (exact set)*, ngược lại *X* được gọi là *tập thô (rough set)*.

Với $B, D \subseteq A$, ta gọi *B*-miền dương của D là tập được xác định như sau

$$POS_B(D) = \bigcup_{X \in U/D} \underline{B}X$$

Rõ ràng $POS_B(D)$ là tập tất cả các đối tượng u sao cho với mọi $v \in U$ mà $u\ B = v\ B$ ta đều có $u\ D = v\ D$. Nói cách khác, $POS_B(D) = \{u \in U \,|\, u_B \subseteq u_D\}$.

Ví dụ 1.1. Xét hệ thông tin biểu diễn các triệu chứng cúm của bệnh nhân cho ở Bảng 1.1.

Bảng 1.1. *Bảng thông tin về bệnh cúm*

U	**Đau đầu**	**Thân nhiệt**	**Cảm cúm**
u_1	Có	Bình thường	Không
u_2	Có	Cao	Có
u_3	Có	Rất cao	Có
u_4	Không	Bình thường	Không
u_5	Không	Cao	Không
u_6	Không	Rất cao	Có
u_7	Không	Cao	Có
u_8	Không	Rất cao	Không

Ta có: $U/\{Đau\ đầu\} = \{u_1, u_2, u_3\}\ ,\ \{u_4, u_5, u_6, u_7, u_8\}$

$U/\{Thân\ nhiệt\} = \{u_1, u_4\}\ ,\ \{u_2, u_5, u_7\}\ ,\ \{u_3, u_6, u_8\}$

$U/\{Cảm\ cúm\} = \{u_1, u_4, u_5, u_8\}\ ,\ \{u_2, u_3, u_6, u_7\}$

$U/\{Đau\ đầu, Cảm\ cúm\} = \{u_1\}\ ,\ \{u_2, u_3\}\ ,\ \{u_4, u_5, u_8\}\ ,\ \{u_6, u_7\}$

Như vậy, các bệnh nhân u_2, u_3 không phân biệt được về đau đầu và cảm cúm, nhưng phân biệt được về thân nhiệt.

Các lớp không phân biệt được bởi *B* = {Đau đầu, Thân nhiệt} là: