ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

PHÙNG THỊ NGA

PHÂN LỚP MIỀN XÁC ĐỊNH THUỘC TÍNH TRONG BÀI TOÁN KHAI PHÁ DỮ LIỆU MỜ

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2015

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

PHÙNG THỊ NGA

PHÂN LỚP MIỀN XÁC ĐỊNH THUỘC TÍNH TRONG BÀI TOÁN KHAI PHÁ DỮ LIỆU MỜ

Chuyên ngành: Khoa học máy tính

Mã số: 60.48.01.01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: TS. TRẦN THÁI SƠN

THÁI NGUYÊN - 2015

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn trực tiếp của **Ts. Trần Thái Sơn**.

Mọi trích dẫn sử dụng trong báo cáo này đều được ghi rõ nguồn tài liệu tham khảo theo đúng qui định.

Mọi sao chép không hợp lệ, vi phạm quy chế đào tạo, hay gian trá, tôi xin chịu hoàn toàn trách nhiệm.

Thái Nguyên, ngày ... tháng ... năm 2014

Tác giả

Phùng Thị Nga

LỜI CẢM ƠN

Luận văn được viết dưới sự hướng dẫn tận tình và nghiêm khắc của TS. Trần Thái Sơn. Lời đầu tiên, tác giả xin bày tỏ lòng kính trọng và biết ơn sâu sắc tới thầy.

Xin chân thành gửi lời cảm ơn tới thầy về những đóng góp quý báu trong quá trình nghiên cứu cũng như trong thời gian hoàn thành luận văn. Tác giả xin chân thành gửi lời cảm ơn đến Phòng Đào tạo sau đại học đã tạo điều kiện thuận lợi trong quá trình học tập, nghiên cứu và hoàn thành luận văn, đảm bảo tiến độ.

Cuối cùng, tác giả xin chân thành cảm ơn các thành viên trong gia đình, những người luôn dành cho tác giả những tình cảm nồng ấm và sẻ chia những lúc khó khăn trong cuộc sống, luôn động viên giúp đỡ tác giả trong quá trình nghiên cứu.

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

Các ký hiệu

$A\ X$	Đại số gia tử tuyến tính
$\underline{A\ X}$	Đại số gia tử tuyến tính đầy đủ
$A\ X^2$	Đại số 2 gia tử
$\mu(h)$, $fm(x)$	Độ đo tính mờ gia tử h và của hạng từ xυ Giá trị định lượng theo điểm của giá trị ngôn ngữ
$\mu A(v)$	Hàm định lượng của giá trị ngôn ngữ A (đo độ thuộc của v)
$sm(x,y)$	Hàm xác định mức độ gần nhau của hai hạng từ x và y
$\Im$	Khoảng tính mờ của giá trị ngôn ngữ
$\boldsymbol{X}k$	Tập các hạng từ có độ dài đúng k
$\boldsymbol{X}(k)$	Tập các hạng từ có độ dài không quá k
$\boldsymbol{I}k$	Hệ khoảng tính mờ mức k của các giá trị ngôn ngữ
$\boldsymbol{I}(k)$	Hệ khoảng tính mờ từ mức 1 đến mức k của các giá trị ngôn ngữ
Tg	Khoảng tương tự bậc g của giá trị ngôn ngữ
$S(k)$	Hệ khoảng tương tự ở mức k của các giá trị ngôn ngữ

Các chữ viết tắt

CSDL	Cơ sở dữ liệu
ĐSGT	Đại số gia tử
ĐS2GT	Đại số 2 gia tử
ĐLNN	Định lượng ngữ nghĩa
RB	Rule-Base
FB	Fuzzy Base
HAFRG	Hedge Algebras based Fuzzy Rules Generation

MOGA	Thuật giải di truyền đa đối tượng
NST	Nhiễm sắc thể

DANH MỤC CÁC HÌNH

DANH MỤC BẢNG BIỂU

MỞ ĐẦU

1. Lý do chọn đề tài

Trong lĩnh vực khai phá dữ liệu, một khó khăn thường gặp là hệ thống phải xử lý khối lượng thông tin rất lớn, đòi hỏi phải có những thuật toán hữu hiệu để khai thác các tri thức ngầm chứa trong khối thông tin to lớn đó.

Một trong những bài toán cơ bản đặt ra trong lĩnh vực nghiên cứu này là cho trước một Cơ sở dữ liệu (thường là CSDL số, tức các giá trị của CSDL là các số thực), từ đó, bằng các phương pháp xử lý nhất định, rút ra một hệ tri thức phản ánh các quy luật chứa trong CSDL số này. Các quy luật này có thể biểu diễn dưới dạng hệ luật IF X is A and Y is B THEN Z is C, trong đó X, Y, Z là các biến mờ (thường là các biến ngôn ngữ), A, B, C là các giá trị biến ngôn ngữ (thường là các tập mờ). Thí dụ luật IF *đường* là *xa* và *tốc độ di chuyển* là *trung bình* THEN *thời gian đến đích* sẽ là *lâu*. Để có thể sinh ra những luật như vậy, đầu tiên ta phải chuyển hóa miền giá trị của các thuộc tính "khoảng cách", "tốc độ", "thời gian" thành các miền mờ, hay nói cách khác là phân chia các miền giá trị đó thành các miền mờ cho các bước xử lý tiếp theo. Chẳng hạn, có thể chia miền giá trị thuộc tính độ dài (có các giá trị min, max tương ứng chẳng hạn là 0km, 200km) thành các miền mờ "gần" (0km- 50km), "trung bình" (51km-100km), "xa" (100km-200km). Trong lý thuyết tập mờ, mỗi miền mờ như vậy được coi là một tập mờ và ứng với một hàm thuộc (MF- membership function) nhằm xác định độ "thuộc" của giá trị biến vào tập mờ đã cho. Khi đó, một giá trị của một thuộc tính CSDL sẽ ứng với một tập các giá trị của các hàm thuộc ứng với với các tập mờ của thuộc tính đó. Và ta sẽ xây dựng hệ luật mờ dựa trên việc xử lý tập giá trị độ thuộc này thay vì xử lý bản thân giá trị ban đầu của CSDL. Việc xây dựng các MF phân chia miền xác định thuộc tính là bước đầu tiên nhưng rất quan trọng trong quy trình xây dựng hệ luật mờ vì chỉ có trên cơ sở phân chia hợp lý các miền xác định thuộc tính ta mới có thể có các tập mờ ngôn ngữ phản ánh

tương đối chính xác ngữ nghĩa định tính của nhãn ngôn ngữ dùng trong hệ luật được xây dựng tiếp theo.Phương pháp tiếp cận theo lý thuyết tập mờ cho ta một cách xử lý dữ liệu khá mềm dẻo, nhanh chóng so với các phương pháp xử lý số cổ điển. Tuy vậy, vẫn còn nhiều vấn đề đặt ra như việc phân chia các miền mờ thế nào cho hợp lý, làm sao xây dựng được các hàm thuộc nhanh chóng, phù hợp và cách xử lý các hàm thuộc này thế nào để giữ được ngữ nghĩa gắn với chúng... Đại số gia tử (ĐSGT) ra đời dựa trên một cấu trúc thứ tự tốt trong tập các giá trị ngôn ngữ của biến ngôn ngữ có thể khắc phục phần nào những điểm yếu đó. Luận văn đặt mục tiêu sử dụng cách tiếp cận ĐSGT trong việc xác định các MF tối ưu phân chia miền mờ cho các thuộc tính của CSDL, để có thể xây dựng được các hệ luật mờ tốt trong các bước tiếp theo nhằm giải quyết các bài toán quan tâm trong lĩnh vực khai phá dữ liệu hay điều khiển mờ.

Được sự đồng ý của trường Đại học Công nghệ thông tin và Truyền thông với sự hướng dẫn của Thầy giáo em xin mạnh dạn nhận đề tài: ***"Phân lớp miền xác định thuộc tính trong bài toán khai phá dữ liệu mờ"*** làm đề tài luận văn của mình.

2. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của luận văn là cơ sở dữ liệu đầu vào dùng để khai phá dữ liệu. Lý thuyết tập mờ và đại số gia tử cũng được nghiên cứu như là công cụ để giải bài toán đặt ra.

3. Hướng nghiên cứu của đề tài

Luận văn nghiên cứu các phương pháp giải bài toán phân lớp miền xác định thuộc tính của các tác giả trong nước cũng như trên thế giới, ưu, khuyết điểm của các phương pháp đã có và nghiên cứu cách giải bài toán theo cách tiếp cận của Đại số gia tử, sử dụng giá trị định lượng ngữ nghĩa của các giá trị biến ngôn ngữ, phân chia miền thuộc tính tiến hành khai phá dữ liệu

4. Phương pháp nghiên cứu

Tìm hiểu các lý thuyết về tập mờ, các dạng tập mờ, tìm hiểu cách biểu diễn tập giá trị chân lý ngôn ngữ cho tập mờ. Tìm hiểu mối quan hệ giữa các dạng biểu diễn tập mờ với hàm định lượng ngữ nghĩa của đại số gia tử, tìm hiểu cách thức chuyển đổi giá trị chân lý ngôn ngữ thành một giá trị số.

Phân tích, đối sánh, liệt kê, nghiên cứu tài liệu, tổng hợp các kết quả của các nhà nghiên cứu liên quan đến lĩnh vực nghiên cứu.

5. Ý nghĩa khoa học

Bài toán phân chia miền xác định thuộc tính nói chung đóng vai trò quan trọng trong quá trình khai phá dữ liệu và do đó nó có ý nghĩa ứng dụng rộng lớn, đặc biệt loại bài toán liên quan đến thông tin mờ vì con người thường quyết định thông qua thông tin mờ ngôn ngữ. Cho đến nay các phương pháp giải bài toán này chủ yếu dựa trên các tập mờ.

Giải bài toán phân chia miền xác định thuộc tính theo cách tiếp cận Đại số gia tử cho ta một phương pháp tương đối đơn giản nhưng khá hữu hiệu trong các cách mà Đại số gia tử nói riêng và lý thuyết tập mờ nói chung có thể sử dụng.