ĐẠI HỌC HUẾ
KHOA KỸ THUẬT VÀ CÔNG NGHỆ

BÁO CÁO ĐỒ ÁN
NĂM HỌC 2022-2023

Giảng viên hướng dẫn: Quách Hải Thọ
Lớp: KHDL & TTNT K2
Môn: Học máy 2
Sinh viên thực hiện:

1	Hoàng Nữ Thu Phương
2	Ngô Ngọc Hà
3	Lê Minh Quân
4	Nguyễn Thị Hòa

Số phách
(Do hội đồng chấm thi ghi)

Thừa Thiên Huế, ngày 26 tháng 05 năm 2023

ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ

PHIẾU ĐÁNH GIÁ ĐỒ ÁN

Học kỳ II, năm học 2022 - 2023

Cán bộ chấm thi 1	Cán bộ chấm thi 2
Nhận xét: .... **Điểm đánh giá của CBCT1:** Bằng số: ..... Bằng chữ:	**Nhận xét:** .... **Điểm đánh giá của CBCT2:** Bằng số: ..... Bằng chữ: .....

Điểm kết luận: .....

Bằng số:.....

Bằng chữ: .....

Thừa Thiên Huế, ngày tháng năm 2023

Cán bộ chấm thi 1	Cán bộ chấm thi 2
(Ký và ghi rõ họ và tên)	(Ký và ghi rõ họ và tên)

LỜI CẢM ƠN

Đầu tiên, nhóm chúng em xin gửi lời cảm ơn chân thành đến Khoa Kỹ thuật và công nghệ - Đại học Huế đã đưa vào chương trình giảng dạy môn học Học máy. Đặc biệt, nhóm nghiên cứu xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn – thầy Quách Hải Thọ đã đồng hành cùng nhóm trong suốt quá trình học tập và nghiên cứu về chủ đề **"Phân lớp dữ liệu bằng cây quyết định mờ dựa trên đại số gia tử".** Nhóm nghiên cứu đã nhận được sự quan tâm, hướng dẫn tận tình và chỉ tiết của thầy để nhóm có đủ kiến thức và vận dụng chúng vào bài đồ án này .

Trong quá trình thực hiện bài nghiên cứu, chắc chắn khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong thầy xem xét và góp ý để đồ án của nhóm được hoàn thiện hơn. Nhóm nghiên cứu kính chúc thầy nhiều sức khỏe, thành công và hạnh phúc.

Nhóm nghiên cứu xin chân thành cảm ơn!

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Viết tắt	Viết đầy đủ
ĐSGT	Đại số gia tử
CART	Classification and Regression Trees
HA	Hedge Algebra – Đại số gia tử
GĐ1	Giai đoạn 1
GĐ2	Giai đoạn 2

DANH MỤC CÁC KÝ HIỆU

Ký hiệu	Diễn giải ý nghĩa
A_i	Thuộc tính A_i
D	Tập mẫu huấn luyện
D_{A_i}	Tập các giá trị kinh điển của A_i
f	Ánh xạ
$f_h(S)$	Hàm đánh giá tính hiệu quả của cây
$f_n(S)$	Hàm đánh giá tính đơn giản của cây
I_k	Tập tất cả các khoảng mờ mức k của các giá trị ngôn ngữ
LD_{A_i}	Tập các giá trị ngôn ngữ của A_i
$O(\log n)$	Độ phức tạp *logarit* của thuật toán
$\mu_A(v)$	Hàm định lượng của giá trị ngôn ngữ A (đo độ thuộc của v)
S	Cây quyết định
$sim(x, y)$	Mức độ gần nhau của x và y
v	Giá trị định lượng theo điểm của giá trị ngôn ngữ
$\underline{X}$	Đại số gia tử
Y	Thuộc tính phân lớp

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ

MỞ ĐẦU

1. Lý do chọn đề tài

Trong cuộc sống con người, ngôn ngữ được hình thành một cách tự nhiên để đáp ứng nhu cầu trao đổi thông tin của xã hội. Hơn thế, ngôn ngữ là công cụ để con người mô tả các sự vật, hiện tượng trong thế giới thực và dựa trên đó để tư duy, lập luận đưa ra những nhận định, phán quyết nhằm phục vụ cho cuộc sống xã hội của chúng ta. Trong thực tế, các khái niệm mờ luôn tồn tại, ví dụ như *trẻ, rất trẻ, hơi già, quá già, hơi xinh*... nên với việc quan niệm các đối tượng đượcsử dụng phải luôn rõ ràng ở trong logic cổ điển sẽ không đủ miêu tả các vấn đề của thế giới thực.

Năm 1965, L. A. Zadeh đã đề xuất hình thức hóa toán học của khái niệm mờ, từ đó lý thuyết tập mờ được hình thành và ngày càng thu hút nhiều nhà nghiên cứu. Bằng các phương pháp tiếp cận khác nhau, nhiều nhà nghiên cứu như Dubois, Prade [10], Mariana [17], Ishibuchi [13], Herrera [6], Yakun Hu [19],… đã đưa ra những kết quả cả về lý thuyết và ứng dụng cho nhiều lĩnh vực như: điều khiển mờ, cơ sở dữ liệu mờ, khai phá dữ liệu mờ. Ý tưởng nổi bật của Zadeh là từ những khái niệm trừu tượng về ngữ nghĩa của thông tin mờ, không chắc chắn như *trẻ-già, nhanh-chậm, cao-thấp*,… và đã tìm ra cách biểu diễn chúng bằng một khái niệm toán học, được gọi là tập mờ. Tuy nhiên, việc mô hình hóa quá trình tư duy lập luận của con người là một vấn đề khó luôn thách thức các nhà nghiên cứu bởi đặc trưng giàu thông tin của ngôn ngữ và cơ chế suy luận không những dựa trên tri thức mà còn là kinh nghiệm, trực quan cảm nhận theo ngữ cảnh của con người. Cấu trúc thứ tự cảm sinh trên các khái niệm mờ biểu thị bằng các giá trị ngôn ngữ không được thể hiện trên các tập mờ vì hàm thuộc của chúng lại không sánh được với nhau. Hơn thế nữa, việc thiết lập các tập mờ của các giá trị ngôn ngữ một cách cố định dựa theo chủ quan của người thiết lập, trong khi một giá trị ngôn ngữ sẽ mang ngữ nghĩa tương đối khác nhau trong các bài toán khác nhau.

Thêm vào đó, với sự bùng nổ dữ liệu của thời đại thông tin như hiện nay, lượng dữ liệu được tạo ra hàng ngày là rất lớn. Khối lượng thông tin dữ liệu

khổng lồ này vượt khỏi giới hạn khả năng ghi nhớ và xử lý của con người. Nhu cầu cần thiết là nghĩ đến các quá trình tự động tìm kiếm các thông tin hữu ích, các quan hệ ràng buộc dữ liệu trong các kho dữ liệu lớn để phát hiện các tri thức, các quy luật hay khuynh hướng dữ liệu hỗ trợ con người phán đoán, nhận xét, ra quyết định. Nhằm đáp ứng các nhu cầu đó, nhiều nhà khoa học đã đề xuất, nghiên cứu và phát triển các phương pháp mới trong khai phá dữ liệu. Các bài toán được biết đến trong lĩnh vực này như phân lớp và nhận dạng mẫu, hồi quy và dự báo, phân cụm, khai phá luật kết hợp,... với rất nhiều kết quả đã được công bố.

Trong việc phân lớp dữ liệu bằng cây quyết định, quá trình xây dựng tại mỗi nút của cây, các thuật toán đều tính lượng thông tin và chọn thuộc tính tương ứng có lượng thông tin tối đa làm nút phân tách trên cây. Các thuộc tính này sẽ chia tập mẫu thành các lớp mà mỗi lớp có một phân loại duy nhất hay ít nhất phải có triển vọng đạt được điều này, nhằm để đạt được cây có ít nút nhưng có khả năng dự đoán cao. Tuy vậy, các cách tiếp cận cho việc huấn luyện cây quyết định hiện nay vẫn còn nhiều vấn đề cần giải quyết:

- Manish Mehta, Jorma Rissanen, Rakesh Agrawal [14], [15], Narasimha Prasad, Mannava Munirathnam Naidu [18], Zhihao Wang, Junfang Wang, Yonghua Huo, Hongze Qiu [23], Haitang Zhang và các cộng sự [12] dựa vào việc tính hệ số *Gini* và tỷ lệ hệ số *Gini* của các thuộc tính để lựa chọn điểm phân chia. Theo hướng tiếp cận này, chúng ta không cần đánh giá mỗi thuộc tính mà chỉ cần tìm điểm chia tách tốt nhất cho mỗi thuộc tính đó. Tuy nhiên, tại mỗi thời điểm chúng ta phải tính một số lượng lớn hệ số *Gini* cho các giá trị rời rạc nên chi phí về độ phức tạp tính toán cao và cây kết quả mất cân xứng vì phát triển nhanh theo chiều sâu, số nút trên cây lớn.

- B. Chandra [7], Chida A. [8], Daveedu Raju Adidela, Jaya Suma. G, Lavanya Devi. G [9], Hesham A. Hefny, Ahmed S. Ghiduk [11], Hou Yuan-long, Chen Ji-lin, Xing Zong-yi [12], Marcos E. Cintra, Maria C. Monard [16], Zeinalkhani M., Eftekhari M. [22] và các cộng sự đã thông qua lý thuyết tập mờ để tính lợi ích thông tin của các thuộc tính mờ cho quá trình phân lớp. Hướng tiếp cận này đã giải quyết được các giá trị mờ trong tập huấn luyện thông qua việc xác định các hàm thuộc, từ đó các bộ giá trị này có thể tham gia vào quá trình huấn luyện. Cách làm này đã giải quyết được hạn chế là bỏ qua các giá trị dữ liệu mờ của cách tiếp phân lớp rõ. Tuy vậy, hiện vẫn còn gặp phải những hạn

chế xuất phát từ bản thân nội tại của lý thuyết tập mờ: hàm thuộc của chúng không so sánh được với nhau, xuất hiện sai số lớn tại quá trình xấp xỉ, phụ thuộc vào sự chủ quan, giá trị ngôn ngữ còn thiếu một cơ sở đại số làm nền tảng.

Thêm vào đó, tất cả các thuật toán học phân lớp bằng cây quyết định hiện có đều phụ thuộc lớn vào việc chọn tập mẫu của người huấn luyện. Khi chúng ta chọn tập mẫu không đặc trưng thì cây quyết định được sinh ra sẽ không có khả năng dự đoán. Mà trong thế giới thực, việc lưu trữ dữ liệu tại các kho dữ liệu nghiệp vụ nhằm nhiều mục đích khác nhau. Nhiều thông tin phục vụ tốt cho việc dự đoán nhưng nhiều thông tin khác chỉ có ý nghĩa lưu trữ thông thường, phục vụ cho việc diễn giải thông tin. Các nhóm thuộc tính này làm phức tạp mẫu nên tăng chi phí cho quá trình huấn luyện, quan trọng hơn là chúng gây nhiễu nên cây được xây dựng không có hiệu quả cao. Vì vậy, làm sao để phân lớp dữ liệu bằng cây quyết định đạt hiệu quả là vấn đề mà các nhà khoa học hiện nay vẫn đang quan tâm, nghiên cứu.

Xuất phát từ việc tìm hiểu, nghiên cứu các đặc điểm và các thách thức về các vấn đề của phân lớp dữ liệu bằng cây quyết định, đồ án đã chọn đề tài là: "*Phân lớp dữ liệu bằng cây quyết định mờ dựa trên đại số gia tử*".

2. Đối tượng và phạm vi nghiên cứu

Phân lớp dữ liệu là vấn đề lớn và quan trọng của khai phá dữ liệu. Cây quyết định là giải pháp hữu hiệu của bài toán phân lớp, nó bao gồm từ mô hình cho quá trình học đến các thuật toán huấn luyện cụ thể để xây dựng cây. Đồ án tập trung nghiên cứu mô hình linh hoạt cho quá trình huấn luyện cây từ tập mẫu huấn luyện, nghiên cứu phương pháp xử lý giá trị ngôn ngữ và xây dựng các thuật toán học phân lớp dữ liệu bằng cây quyết định mờ đạt nhằm đạt hiệu quả trong dự đoán và đơn giản đối với người dùng.

3. Phương pháp nghiên cứu

Đồ án tập trung vào các phương pháp chính:

- *Phương pháp nghiên cứu tài liệu, tổng hợp và hệ thống hóa:* tìm kiếm, thu thập tài liệu về các công trình nghiên cứu đã được công bố ở các bài báo đăng ở các hội thảo và tạp chí lớn; nghiên cứu các phương pháp xây dựng cây quyết định đã có, nhằm phân tích những thuận lợi và khó khăn trong quá trình học phân lớp dữ liệu bằng cây quyết định. Đề xuất các thuật toán học phân lớp

bằng cây quyết định mờ theo hướng tăng độ chính xác cho quá trình sử dụng cây kết quả để dự đoán nhằm thỏa mãn mục tiêu cụ thể của người dùng.

- *Phương pháp thực nghiệm khoa học:* sử dụng các bộ dữ liệu chuẩn không chứa giá trị mờ Northwind và các bộ dữ liệu có chứa giá trị mờ Mushroom và Adult cho quá trình thử nghiệm, đánh giá. Thực hiện việc thử nghiệm, đánh giá các thuật toán đã đề xuất trong các công trình trước đây với các thuật toán được đề xuất trong đồ án nhằm minh chứng cho tính hiệu quả về độ chính xác trong quá trình dự đoán.

4. Mục tiêu và nội dung của đồ án

Sau khi nghiên cứu và phân tích các vấn đề về phân lớp dữ liệu bằng cây quyết định của các nghiên cứu trong và ngoài nước, đồ án đưa ra mục tiêu nghiên cứu chính như sau:

- Xây dựng mô hình học phân lớp dữ liệu bằng cây quyết định mờ và phương pháp trích chọn đặc trưng để chọn tập mẫu huấn luyện cho quá trình học phân lớp. Đề xuất phương pháp xử lý giá trị ngôn ngữ của các thuộc tính chưa thuần nhất dựa vào ĐSGT.

- Đề xuất các thuật toán học bằng cây quyết định mờ cho bài toán phân lớp nhằm đạt hiệu quả trong dự đoán và đơn giản đối với người dùng.

Để đáp ứng cho các mục tiêu nghiên cứu trên, đồ án tập trung nghiên cứu các nội dung chính sau:

- Nghiên cứu các thuật toán học cây truyền thống CART, ID3, C45, C50, SLIQ, SPRINT trên mỗi tập mẫu huấn luyện để tìm phương pháp học đạt hiệu quả dự đoán cao.

- Nghiên cứu xây dựng phương pháp trích chọn đặc trưng để chọn tập mẫu huấn luyện cho việc học cây quyết định từ các kho dữ liệu nghiệp vụ.

- Nghiên cứu xây dựng một mô hình học phân lớp dữ liệu bằng cây quyết định linh hoạt từ tập mẫu huấn luyện.

- Nghiên cứu để đề xuất phương pháp xử lý giá trị ngôn ngữ của các thuộc tính chưa thuần nhất trên tập mẫu huấn luyện dựa vào bản chất của ĐSGT.

- Nghiên cứu để đề xuất các thuật toán học phân lớp bằng cây quyết định mờ nhằm đạt hiệu quả trong dự đoán và đơn giản đối với người dùng. Phân tích và đánh giá kết quả của các thuật toán học đã đề xuất với các thuật toán khác

trên các bộ mẫu chuẩn không chứa giá trị mờ Northwind và các bộ dữ liệu có chứa giá trị mờ Mushroom, Adult để đối sánh.

5. Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học

- Xây dựng mô hình học phân lớp dữ liệu bằng cây quyết định mờ từ tập mẫu huấn luyện. Đề xuất phương pháp trích chọn đặc trưng để chọn tập mẫu huấn luyện cho việc học phân lớp bằng cây quyết định từ các kho dữ liệu nghiệp vụ, nhằm hạn chế sự phụ thuộc ý kiến của chuyên gia trong quá trình chọn tập mẫu huấn luyện.

- Đề xuất phương pháp xử lý giá trị ngôn ngữ của các thuộc tính chưa thuần nhất trên tập mẫu huấn luyện dựa vào bản chất của ĐSGT.

Ý nghĩa thực tiễn

- Góp phần chứng tỏ khả năng ứng dụng phong phú của ĐSGT trong biểu diễn và xử lý thông tin mờ, không chắc chắn.

- Góp phần vào việc giải quyết vấn đề định lượng cho các giá trị ngôn ngữ mà không phụ thuộc cố định vào miền trị *Min-Max* của các giá trị kinh điển của thuộc tính mờ trong tập mẫu.

- Dựa trên các khái niệm về khoảng mờ và khoảng mờ lớn nhất, đồ án đã đề xuất các thuật toán cho quá trình học cây, nhằm tăng khả năng dự đoán cho bài toán phân lớp dữ liệu bằng cây quyết định. Làm phong phú thêm các phương pháp học cho bài toán phân lớp nói chung và phân lớp bằng cây quyết định nói riêng.

6. Bố cục của đồ án

Ngoài phần mở đầu, kết luận và tài liệu tham khảo, đồ án được chia làm3 chương nội dung:

Chương 1: Cơ sở lý thuyết về đại số gia tử và tổng quan phân lớp dữ liệu bằng cây quyết định. Chương này tập trung nghiên cứu, phân tích và đánh giá các vấn đề liên quan mật thiết đến đồ án như: khái niệm mờ, tập mờ và khái niệm biến ngôn ngữ, phương pháp lập luận xấp xỉ trực tiếp trên ngôn ngữ, khái niệm và tính chất về ĐSGT. Đồ án cũng trình bày các vấn đề cơ bản của bài toán phân lớp dữ liệu bằng cây quyết định, các hạn chế trên cây quyết định truyền thống và sự cần thiết của bài toán phân lớp bằng cây quyết định mờ. Ở

đây, đồ án đã phát biểu hình thức bài toán phân lớp dữ liệu bằng cây quyết định và cũng tập trung nghiên cứu, phân tích và đánh giá các công trình nghiên cứu đã công bố gần đây, chỉ ra các vấn đề còn tồn tại để xác định mục tiêu và nội dung cần giải quyết.

Chương 2: Phân lớp dữ liệu bằng cây quyết định mờ theo phương pháp đối sánh điểm mờ dựa trên đại số gia tử. Chương này của tập trung phân tích sự ảnh hưởng của tập mẫu huấn luyện đối với hiệu quả cây kết quả thu được, trình bày một phương pháp nhằm trích chọn được tập mẫu huấn luyện đặc trưng phục vụ cho quá trình huấn luyện; phân tích, đưa ra các khái niệm về tập mẫu không thuần nhất, giá trị ngoại lai và xây dựng thuật toán để có thể thuần nhất cho các thuộc tính có chứa các giá trị này. Đề xuất các thuật toán MixC4.5 và FMixC4.5 phục vụ quá trình học cây quyết định trên tập mẫu không thuần nhất; thử nghiệm trên các cơ sở dữ liệu không chứa dữ liệu mờ Northwind và có chứa thông tin mờ Mushroom để đối sánh về khả năng dự đoán của cây kết quả sau khi huấn luyện.

Chương 1.

CƠ SỞ LÝ THUYẾT VỀ ĐẠI SỐ GIA TỬ VÀ TỔNG QUAN PHÂN LỚP DỮ LIỆU BẰNG CÂY QUYẾT ĐỊNH

Với mục tiêu nhằm giải quyết các vấn đề của bài toán phân lớp dữ liệu bằng cây quyết định mờ, Chương 1 của đồ án trình bày một số vấn đề liên quan đến bài toán phân lớp dữ liệu bằng cây quyết định, cây quyết định mờ và các kiến thức cơ bản của đại số gia tử dùng để nghiên cứu trong quá trình học phân lớp dữ liệu bằng cây quyết định. Nội dung của chương này bao gồm: tập mờ, đại số gia tử và các phương pháp học phân lớp dữ liệu bằng cây quyết định.

1.1. Lý thuyết tập mờ

1.1.1. Tập mờ và thông tin không chắc chắn

Thực tế đã chứng minh khái niệm mờ luôn tồn tại, hiện hữu trong các bài toán ứng dụng, trong cách suy luận của con người, ví dụ như *trẻ, rất trẻ, hơi già, quá già*, hơi xinh... Vì thế, với việc quan niệm các đối tượng được sử dụng phải luôn rõ ràng ở trong logic cổ điển sẽ không không đủ tốt cho việc miêu tả các vấn đề củabài toán thế giới thực. Như vậy, rất cần một tiếp cận nghiên cứu mới so với logic cổ điển. Ý tưởng nổi bật của khái niệm tập mờ của Zadeh là từ những khái niệm trừu tượng về ngữ nghĩa của thông tin mờ, không chắc chắn như *trẻ-già, nhanh- chậm, cao-thấp, xấu-đẹp*,… ông đã tìm cách biểu diễn chúng bằng một khái niệm toán học, được gọi là tập mờ, như là một sự khái quát trực tiếp của kháiniệm tập hợp kinh điển.

Cho một tập vũ trụ V khác rỗng. Một tập mờ A trên tập vũtrụ V được đặc trưng bởi hàm thuộc:

$$\mu_A(x): V \rightarrow [0,1]$$

với $\mu_A(x)$ là độ thuộc của phần tử x trong tập mờ A.

$$A = \frac{\mu_{A(x_1)}}{x_1} + \frac{\mu_{A(x_2)}}{x_2} + \cdots + \frac{\mu_{A(x_n)}}{x_n}$$

Một tập mờ hữu hạn được ký hiệu bởi:

Một tập mờ vô hạn được ký hiệu bởi:

$$A = \int \mu_A(x)/x$$

Giá trị hàm $\mu_A(x)$ càng gần tới 1 thì mức độ thuộc của x trong A càng cao. Tập mờ là sự mở rộng của khái niệm tập hợp kinh điển. Thật vậy, khi A là một tập hợp kinh điển, hàm thuộc của nó, $\mu_A(x)$, chỉ nhận 2 giá trị 1 hoặc 0, tương ứng với x có nằm trong A hay không.

1.1.2. Biến ngôn ngữ

Khái niệm biến ngôn ngữ đã được L. A. Zadeh giới thiệu, là một công cụ quan trọng để phát triển phương pháp lập luận xấp xỉ dựa trên logic mờ [20], [21]. Ông đã viết: "*Khi thiếu hụt tính chính xác bề ngoài của những vấn đề phức tạp cố hữu, một cách tự nhiên là tìm cách sử dụng các biến gọi là biến ngôn ngữ; đó là các biến mà các giá trị của chúng không phải là các số mà là các từ hoặc các câu trong một ngôn ngữ tự nhiên hoặc nhân tạo. Động lực cho việc sử dụng các từ, các câu hơn các số là đặc trưng ngôn ngữ của các từ, các câu thường là ít xác định hơn của số*".

Trong các nghiên cứu của L. A. Zadeh về biến ngôn ngữ và lập luận xấp xỉ, ông luôn nhấn mạnh hai đặc trưng quan trọng sau đây của biến ngôn ngữ:

1. *Tính phổ quát:* miền giá trị của hầu hết các biến ngôn ngữ có cùng cấu trúc cơ sở theo nghĩa các giá trị ngôn ngữ tương ứng là giống nhau ngoại trừphần tử sinh nguyên thủy.

2. *Tính độc lập ngữ cảnh của gia tử và liên từ:* ngữ nghĩa của các gia tử và liên từ hoàn toàn độc lập với với ngữ cảnh, khác với giá trị nguyên thủy của các biến ngôn ngữ lại phụ thuộc vào ngữ cảnh. Do đó khi tìm kiếm mô hình cho các gia tử và liên từ chúng ta không phải quan tâm đến giá trị nguyên thủy của biến ngôn ngữ đang xét.

Vấn đề sử dụng tập mờ để biểu diễn các giá trị ngôn ngữ và dùng các phép toán trên tập mờ để biểu thị các gia tử ngôn ngữ đã cho phép thực hiện các thao tác dữ liệu mờ, một phần nào đã đáp ứng được nhu cầu thực tế của con người. Tuy nhiên, theo cách sử dụng tập mờ cho thấy vẫn có nhiều hạn chế do việc xây dựng các hàm thuộc và xấp xỉ các giá trị ngôn ngữ bởi các tập mờ còn mang tính chủ quan, phụ thuộc nhiều vào ý kiến chuyên gia cho nên dễ mất mát thông tin. Mặc khác, bản thân các giá trị ngôn ngữ có một cấu trúc thứ tự nhưng ánh xạ gán nghĩa sang tập mờ, không bảo toàn cấu trúc đó nữa. Do đó, vấn đề đặt ra là cần có một cấu trúc toán học mô phỏng chính xác hơn cấu trúc ngữ nghĩa của một khái niệm mờ.

1.2. Đại số gia tử

1.2.1. Khái niệm đại số gia tử

Xét một ví dụ có miền ngôn ngữ của biến chân lý *TRUTH* gồm các từ sau: *Dom*(*TRUTH*) = {*đúng, sai, rất đúng, rất sai, ít nhiều đúng, ít nhiều sai, khả năng đúng, khả năng sai, xấp xỉ đúng, xấp xỉ sai, ít đúng, ít sai, rất khả năng đúng, rất khả năng sai*,...}, trong đó *đúng, sai* là các từ nguyên thuỷ, các từ nhấn *rất, ít nhiều, khả năng, xấp xỉ, ít,…* được gọi là các gia tử. Khi đó, miền ngôn ngữ $T = Dom(TRUTH)$ có thể biểu thị như một đại số $\underline{X} = (X, G, H, \leq)$, trong đó G là tập các từ nguyên thuỷ {*thấp, cao*} được xem là các phần tử sinh. $H = H^+ \cup H^-$ là tập các gia tử dương, âm và được xem như là các phép toán một ngôi, quan hệ $\leq$ trên các từ là quan hệ thứ tự được "*cảm sinh*" từ ngữ nghĩa tự nhiên. Tập X được sinh ra từ G bởi các phép tính trong H.

Như vậy, mỗi phần tử của X sẽ có dạng biểu diễn $x = h_n h_{n-1}.....h_1 c$, $c \in G$. Tập tất cả các phần tử được sinh ra từ một phần tử x được ký hiệu là $H(x)$. Nếu G có đúng hai từ nguyên thuỷ mờ, thì một được gọi là phần tử sinh dương ký hiệu là c^+, một gọi là phần tử sinh âm ký hiệu là c^- và ta có $c^- < c^+$. Trong ví dụ trên *đúng* là phần tử sinh dương còn *sai* là phần tử sinh âm.

Như vậy, một cách tổng quát, cho ĐSGT $\underline{X} = (X, G, H, \leq)$, với $G = \{0, c^-, W, c^+, 1\}$, trong đó c^+ và c^- tương ứng là phần tử sinh dương và âm, X là tập nền. $H = H^+ \cup H^-$ với giả thiết $H^+ = \{h_1, h_2, .., h_p\}$, $H^- = \{h_{-q}, ..., h_{-1}\}$, $h_1 < h_2 < ... < h_p$ và $h_{-q} < ... < h_{-1}$ là dãy các gia tử.

Trong ĐSGT tuyến tính, chúng ta bổ sung thêm vào hai phép tính Σ và Φ với ngữ nghĩa là cận trên đúng và cận dưới đúng của tập $H(x)$, tức là $\Sigma x = supremum\ H(x)$ và $\Phi x = infimum\ H(x)$, khi đó ĐSGT tuyến tính được gọi là ĐSGT tuyến tính đầy đủ và được ký hiệu $\underline{X} = (X, G, H, \Sigma, \Phi, \leq)$.

1.2.2. Khoảng mờ và các mối tương quan của khoảng mờ

Khoảng mờ *I(x)* của một phần tử *x* là một đoạn con của[0, 1], được xác định bằng cách quy nạp theo độ dài của *x* như sau:

1. Với độ dài *x* bằng 1($l(x) = 1$), tức là $x \in \{ c^+, c^-\}$, $I_{fm}(c^-)$ và $I_{fm}(c^+)$ là các khoảng con và tạo thành một phân hoạch của [0, 1], thỏa $I_{fm}(c^-) \leq I_{fm}(c^+)$. Tức là $\forall u \in I_{fm}(c^-)$ và $\forall u \in I_{fm}(c^+)$: $u \leq v$. Điều này hoàn toàn phù hợp với thứ tự ngữ nghĩa của c^- và c^+.

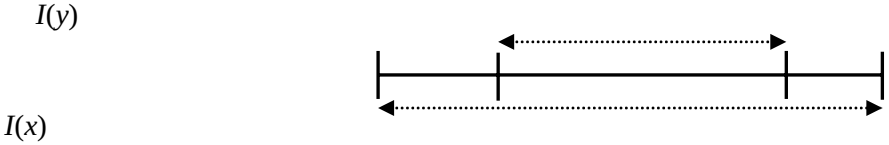
Ký hiệu độ dài của $I_{fm}(x)$ là $|I_{fm}(x)|$. Ta có $|I_{fm}(c^-)| = I_{fm}(c^-)$ và $|I_{fm}(c^+)| = I_{fm}(c^+)$.

2. Giả sử $\forall x \in X$ độ dài bằng *k* ($l(x) = k$) có khoảng mờ là $I_{fm}(x)$ và $|I_{fm}(x)| = fm(x)$. Các khoảng mờ của $y = h_i x$, $\forall i \in [-q, -q+1,\ldots,-1, 1, 2,..., p]$, lúc này $l(y) = k + 1$, là tập $\{I_{fm}(h_i x)\}$ thoả mãn một phân hoạch của $I_{fm}(x)$, $|I_{fm}(h_i x)| = I_{fm}(h_i x)$ và có thứ tự tuyến tính tương ứng với thứ tự của tập $\{h_{-q}x, h_{-q+1}x, \ldots, h_p x\}$.

Khi $l(x) = k$, ta ký hiệu $I(x)$ thay cho $I_{fm}(x)$, $X_k = \{\forall x \in X: l(x) = k\}$ là tập các phần tử trong *X* có độ dài đúng bằng *k*, $I_k = \{I_k(x) : x \in X_k\}$ là tập tất cả các khoảng mờ mức *k*.

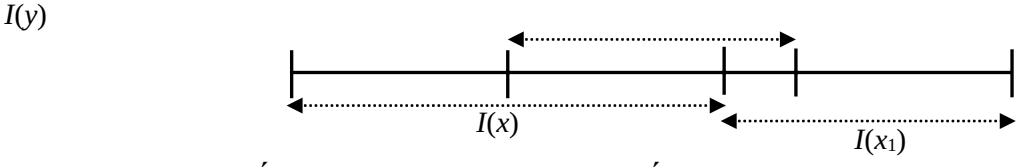
Cho một ĐSGT $\underline{X} = (X, G, H, \leq)$, với $x, y \in X$ ta có:

1. Nếu $I_L(x) \leq I_L(y)$ và $I_R(x) \geq I_L(y)$ thì ta nói giữa *y* và *x* có mối tương quan $I(y) \subseteq I(x)$, ngược lại ta nói $I(y) \not\subset I(x)$.



Hình 1.1. Mối tương quan $I(y) \subseteq I(x)$

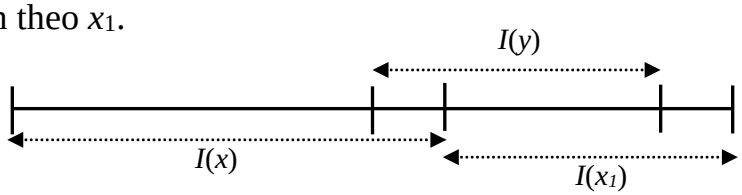
2. Khi $I(y) \not\subset I(x)$, với $x_1 \in X$ và giả sử $x < x_1$, nếu $|I(y) \cap I(x)| \geq |I(y)|/£$ với £ là số đoạn $I(x_i) \subseteq [0, 1]$ sao cho $I(y) \cap I(x_i) \neq \varnothing$ thì ta nói ta nói *y* có mối tương quan được đối sánh theo *x*.



Hình 1.2. Mối tương quan của *y* được đối sánh theo *x*, khi $I(y) \not\subset I(x)$

Ngược lại, nếu $|I(y) \cap I(x_1)| \geq |I(y)|/£$ thì ta nói ta nói *y* có mối tương quan

được đối sánh theo x_1.



Hình 1.3. Mối tương quan của y được đối sánh theo x_1, khi $I(y) \not\subset I(x)$

1.3. Phân lớp dữ liệu bằng cây quyết định

1.3.1. Bài toán phân lớp trong khai phá dữ liệu

Mục đích của khai phá dữ liệu nhằm phát hiện các tri thức mà mỗi tri thức được khai phá đó sẽ được mô tả bằng các mẫu dữ liệu. Sự phân lớp là quá trình quan trọng trong khai phá dữ liệu, nó chính là việc đi tìm những đặc tính của đối tượng, nhằm mô tả một cách rõ ràng phạm trù mà các đối tượng đó thuộc về một lớp nào đó. Quá trình phân lớp gồm có 02 tiến trình:

1. *Xây dựng mô hình:* với tập các lớp đã được định nghĩa trước, mỗi bộ mẫu phải được quyết định để thừa nhận vào một nhãn lớp. Tập các bộ dùng cho việc xây dựng mô hình gọi là tập dữ liệu huấn luyện, tập huấn luyện có thể được lấy ngẫu nhiên từ các cơ sở dữ liệu nghiệp vụ được lưu trữ.

2. *Sử dụng mô hình:* ước lượng độ chính xác của mô hình. Dùng một tập dữ liệu kiểm tra có nhãn lớp được xác định hoàn toàn độc lập với tập dữ liệu huấn luyện để đánh giá độ chính xác của mô hình. Khi độ chính xác của mô hình được chấp nhận, ta sẽ dùng mô hình để phân lớp các bộ hoặc các đối tượng trong tương lai mà nhãn lớp của nó chưa được xác định từ tập dữ liệu chưa biết.

Vậy, bài toán phân lớp có thể được phát biểu tổng quát như sau:

Cho $U = \{A_1, A_2,\ldots, A_m\}$ là tập có m thuộc tính, $Y = \{y_1, ..., y_n\}$ là tập các nhãn của các lớp; với $D = A_1 \times ... \times A_m$ là tích Đề-các của các miền của m thuộc tính tương ứng, có n số lớp và N là số mẫu dữ liệu. Mỗi dữ liệu $d_i \in D$ thuộc một lớp $y_i \in Y$ tương ứng tạo thành từng cặp $(d_i , y_i) \in (D, Y)$.

1.3.2. Cây quyết định

Một cây quyết định là một mô hình logic được biểu diễn như một cây, cho biết giá trị của một biến mục tiêu có thể được dự đoán bằng cách dùng các giá trị của một tập các biến dự đoán. Trên mô hình cây quyết định, mỗi một nút trong tương ứng với một biến dự đoán, đường nối giữa nó với nút con của nó thể hiện

một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Nó có thể hiểu như là một cách biểu diễn các quy tắc để đưa về kết quả là một giá trị cụ thể hay thuộc một lớp nào đó.

Giải bài toán phân lớp dựa trên mô hình cây quyết định chính là xây dựng một cây quyết định, ký hiệu *S*, để phân lớp. S đóng vai trò như một ánh xạ từ tập dữ liệu vào tập nhãn:

$$S : D \rightarrow Y$$

Cây quyết định biểu diễn cho tri thức về bài toán, nó không chỉ phản ánh đúng với tập dữ liệu mẫu huấn luyện mà còn phải có khả năng dự đoán và cung cấp giúp cho người dùng phán đoán, ra quyết định đối với đối tượng trong tương lai mà nhãn lớp của nó chưa được xác định từ tập dữ liệu chưa biết. Quá trình học cây quyết định gồm có 3 giai đoạn:

1. *Tạo cây.* Sử dụng các thuật toán phân lớp để phân chia tập dữ liệu huấn luyện một cách đệ quy cho đến khi mọi nút lá đều thuần khiết, tức là nút mà tại đó tập mẫu tương ứng có cùng một giá trị trên thuộc tính quyết định Y. Sự lựa chọn các thuộc tính trong quá trình xây dựng cây được dựa trên việc đánh giá lượng lợi ích thông tin tại mỗi thuộc tính đang xét.

2. *Cắt tỉa cây.* Sau khi tạo cây, cắt tỉa cây quyết định là việc làm rất cần thiết để khắc phục những khiếm khuyết của cây. Cắt tỉa cây là cố gắng loại bỏ những nhánh không phù hợp hay những nhánh gây ra lỗi.

3. *Kiểm định cây kết quả.* Để bảo đảm độ chính xác của cây trước khi đưa vào ứng dụng trong thực tế, ta cần phải đánh giá độ chính xác của cây từ đó đưa ra tiêu chí đánh giá độ tin cậy theo tỷ lệ phần trăm được dự đoán chính xác.

Việc tạo cây là giai đoạn quan trọng nhất, nó chính là quá trình tạo ra mô hình logic cho cây. Để xây dựng cây quyết định, tại mỗi nút trong cần xác định một thuộc tính thích hợp để kiểm tra, phân chia dữ liệu thành các tập con.

1.3.3. Lợi ích thông tin và tỷ lệ lợi ích thông tin

a. Entropy

Một bít là một chữ số nhị phân nên ta sử dụng một bít để đại diện cho đối tượng thì ta chỉ phân biệt được hai đối tượng, với n bít sẽ phân biệt được 2^n đối tượng khác nhau. Theo đó chúng ta có thể phân biệt n đối tượng bằng $log_2(n)$ bít.

Một bộ mã P thiết kế để phân biệt các phần tử của tập $\{x\}$, để nhận diện được $\{x\}$, chúng ta cần $-log_2P(x)$ bít. Nếu muốn xác định một phân phối thì ít nhất ta cần phải dùng số bít kỳ vọng để nhận diện một phần tử là:

$$\sum_x P(x)\ log\ P(x)$$

b. Lợi ích thông tin

Lợi ích thông tin được tính theo *Entropy*, nó đại diện cho giá trị thông tin của thuộc tính được chọn trong tập mẫu. Với thuộc tính quyết định Y của tập D chưa thuần nhất, được phân phối trong n lớp và giả sử tỉ lệ của các lớp của Y trong D là p_1, p_2,... p_n . Khi đó, *Entropy* của Y trong D là:

$$E(Y,D) = \sum_{i=1}^{n} -p_i \log_2 p_i$$

Giả sử thuộc tính $A_i \in D$ có m giá trị được chọn làm thuộc tính phân lớp và giả thiết A_i sẽ chia tập huấn luyện D thành m tập con D_1, D_2,… D_n,. Lúc này, *Entropy* mà ta nhận được khi phân lớp trên thuộc tính A_i là:

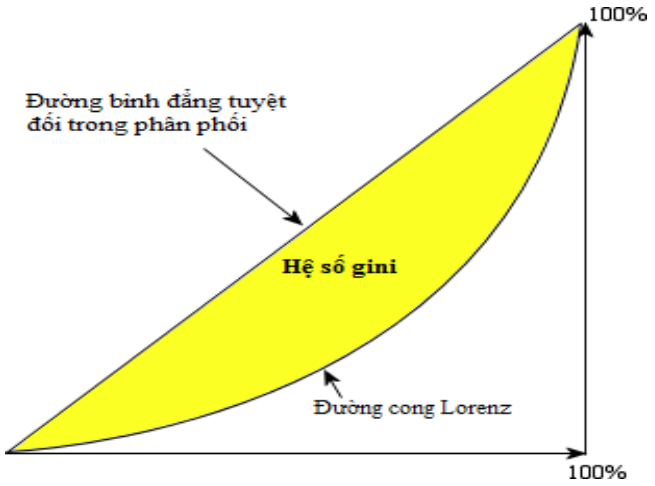
$$E(A_i,D) = \sum_{j=1}^{m} \frac{|D_j|}{|D|} Entropy\ (D_j)$$

Lợi ích thông tin của thuộc tính A_i trong D được tính:

$$Gain(A_i, D) = E(Y, D) - E(A_i, D)$$

c. Hệ số *Gini* và tỷ lệ hệ số *Gini*

Hệ số *Gini* là tỷ lệ phần trăm giữa diện tích của vùng nằm giữa đường bình đẳng tuyệt đối và đường cong Lorenz với diện tích của vùng nằm giữa đường bình đẳng tuyệt đối và đường bất bình đẳng tuyệt đối. Hệ số *Gini* được đưa ra dựa vào hàm phân bố xác xuất, nó dựa trên việc tính bình phương các xác suất thành viên cho mỗi thể loại đích trong nút.



Hình 1.4. Minh họa hình học về chỉ số Gini

Giả sử tập *D* được chia làm *n* lớp khác nhau, tần suất xuất hiện của lớp *i* trong *D* là p_i, chỉ số *Gini* của tập *D* được ký hiệu là *Gini*(*D*), được cho bởi công thức:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{n} p_i^2$$

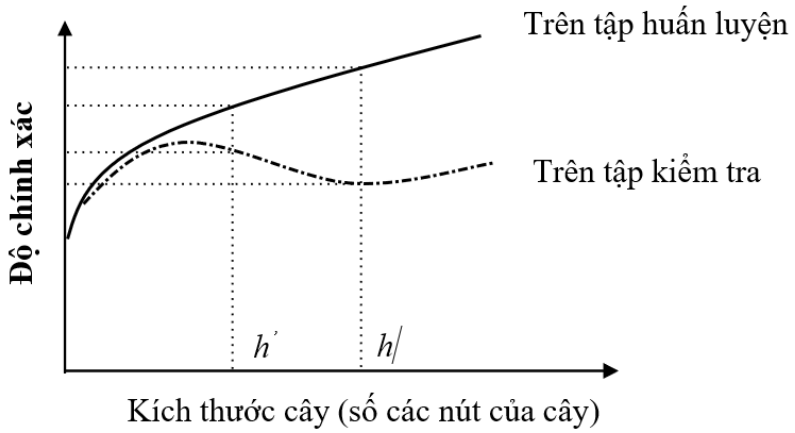
Nếu tập *D* được tách thành 2 tập con D_1, D_2 thì hệ số *Gini* của tập D khi được chia tách được gọi là tỷ lệ hệ số Gini (*GiniSplitIndex)* ký hiệu là $Gini(D)_{Split}$ được xác định như công thức:

$$Gini(D)_{Split} = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

1.3.4. Vấn đề quá khớp trong mô hình cây quyết định

Trong quá trình học cây quyết định, mỗi nhánh của cây vừa đủ sâu để phân lớp hoàn hảo các mẫu huấn luyện, điều này chính là chiến lược phù hợp. Song trong thực tế nó có thể dẫn đến nhiều khó khăn khi có độ nhiễu của dữ liệu huấn luyện hoặc số mẫu huấn luyện là quá nhỏ để đem lại một mô hình quá lý tưởng. Vì vậy, đôi lúc các mẫu dữ liệu cho ta một khái niệm trong quá trình học nhưng điều này chưa hẳn là có thể dự đoán tốt đối với các mẫu chưa gặp. Hơn thế nữa, khi số lượng các mẫu của tập huấn luyện tăng lên thì cũng không bảo đảm được rằng chương trình học sẽ hội tụ đến khả năng đúng khi dự đoán, ta gọi là "*quá khớp*" trong quá trình huấn luyện. Trong thực tế, khó có câu trả lời cho câu hỏi: "cần bao nhiêu mẫu để nhận ra một khái niệm đúng". Như vậy, "*quá khớp*" là một vấn đề khó khăn đáng kể trên thực tế đối với việc học phân lớp bằng cây quyết định .

Hình 1.5. Vấn đề "*quá khớp*" trong cây quyết định



1.4. Phân lớp dữ liệu bằng cây quyết định mờ

1.4.1. Các hạn chế của phân lớp dữ liệu bằng cây quyết định rõ

♦ **Hướng tiếp cận dựa vào việc tính lợi ích thông tin của thuộc tính:** Dựa vào khái niệm Entropy thông tin để tính lợi ích thông tin và tỷ lệ lợi ích thông tin của các thuộc tính tại thời điểm phân chia của tập mẫu huấn luyện, từ đó lựa chọn thuộc tính tương ứng có lợi ích thông tin lớn nhất làm điểm phân chia. Sau khi chọn được thuộc tính để phân lớp, nếu thuộc tính là kiểu rời rạc thì phân lớp theo giá trị phân biệt của chúng, nếu thuộc tính là liên tục thì ta phải tìm ngưỡng của phép tách để chia thành 2 tập con theo ngưỡng đó. Việc tìm ngưỡng cho phép tách cũng dựa theo tỷ lệ lợi ích thông tin của các ngưỡng trong tập huấn luyện tại nút đó. Với *m* là số thuộc tính, *n* là số thể hiện của tập huấn luyện thì độ phức tạp của các thuật toán là $O(m \times n \times \log n)$.

Tuy hướng tiếp cận này cho chúng ta các thuật toán có độ phức tạp thấp nhưng việc phân chia *k-phân* trên các thuộc tính rời rạc làm cho số nút của cây tại một cấp tăng lên nhanh, làm tăng chiều rộng của cây, dẫn đến việc cây dàn trải theo chiều ngang nên dễ xảy ra tình trạng quá khớp, khó để có thể dự đoán. Hơn nữa, cách chia này có khả năng dẫn đến lỗi - khi dữ liệu không thể đoán nhận được lớp - điều này dẫn đến việc dự đoán sẽ cho kết quả không chính xác.

♦ **Hướng tiếp cận dựa vào việc tính hệ số Gini của thuộc tính:** Dựa vào việc tính hệ số Gini và tỷ lệ hệ số Gini của các thuộc tính để lựa chọn điểm phân chia cho tập huấn luyện tại mỗi thời điểm. Theo cách tiếp cận này, chúng ta không cần đánh giá mỗi thuộc tính mà chỉ cần tìm điểm tách tốt nhất cho mỗi thuộc tính đó. Thêm vào đó, với việc sử dụng kỹ thuật tiền xử lý sắp xếp trước trên mỗi một thuộc tính, nên hướng tiếp cận này đã giải quyết được vấn đề thiếu bộ nhớ khitập huấn luyện lớn.

Tuy nhiên, vì tại thời điểm phân chia với thuộc tính rời rạc, hoặc luôn lựa chọn cách phân chia theo *nhị phân tập hợp* của SLIQ nên cây kết quả mất cân xứng vì phát triển nhanh theo chiều sâu. Thêm vào đó, tại mỗi thời điểm chúng ta phải tính một số lượng lớn hệ số Gini cho các giá trị rời rạc nên chi phí về độ phức tạp tính toán cao.

Thêm vào đó, việc học phân lớp bằng cây quyết định theo các hướng tiếp cận đòi hỏi tập mẫu huấn luyện phải thuần nhất và chỉ chứa các dữ liệu kinh điển. Tuy nhiên, do bản chất luôn tồn tại các khái niệm mờ trong thế giới thực

nên điều kiện này không đảm bảo trong các cơ sở dữ liệu hiên đại. Vì vậy, việc nghiên cứu bài toán phân lớp dữ liệu bằng cây quyết định mờ là vấn đề tất yếu.

1.4.2. Bài toán phân lớp dữ liệu bằng cây quyết định mờ

Như đã trình bày, cho $U = \{A_1, A_2,\ldots, A_m\}$ là tập có m thuộc tính, $Y = \{y_1, ..., y_n\}$ là tập các nhãn của các lớp; với $D = A_1 \times ... \times A_m$ là tích Đề-các của các miền của m thuộc tính tương ứng, có n số lớp và N là số mẫu dữ liệu. Mỗi dữ liệu $d_i \in D$ thuộc một lớp $y_i \in Y$ tương ứng tạo thành từng cặp $(d_i , y_i) \in (D, Y)$. Ta có bài toán phân lớp dữ liệu bằng cây quyết định là một ánh xạ từ tập dữ liệu vào tập nhãn:

$$S : D \rightarrow Y$$

Như vậy, mô hình cây quyết định S phải đạt các mục tiêu như hiệu quả phân lớp cao, tức là sai số phân lớp cho các dữ liệu dự đoán ít nhất có thể và cây có ít nút để thuận tiện cho việc biểu diễn và duyệt cây. Mục tiêu về hiệu quả phân lớp nhằm đáp ứng tính đúng đắn của của mô hình đối với tập dữ liệu mẫu được cho của bài toán, còn mục tiêu sau với mong muốn mô hình cây quyết định nhận được phải đơn giản đối với người dùng.

Ta ký hiệu $\underline{S}$ là tập tất cả các cây có thể được tạo ra từ tập huấn luyện S trên thuộc tính quyết định Y. Gọi $f_h(S) : \underline{S} \rightarrow \mathbb{R}$ là hàm đánh giá khả năng dự đoán của cây quyết định S và $f_n(S) : \underline{S} \rightarrow \mathbb{N}$ là hàm thể hiện số nút của cây kết quả nhằm đánh giá tính đơn giản của cây đối với người dùng. Lúc này, mục tiêu của bài toán phân lớp dữ liệu bằng cây quyết định mờ:

$$S : D \rightarrow Y$$

nhằm đạt được:

$$f_h(S) \rightarrow max \text{ và } f_n(S) \rightarrow min$$

Hai mục tiêu trên khó có thể đạt được đồng thời. Khi số nút của cây giảm đồng nghĩa với lượng tri thức về bài toán giảm nên nguy cơ phân lớp sai sẽ tăng lên, nhưng khi có quá nhiều nút cũng có thể gây ra sự quá khớp thông tin trong quá trình phân lớp.

Bên cạnh đó, sự phân chia tại mỗi nút ảnh hưởng đến tính phổ quát hay cá thể tại nút đó. Nếu sự phân chia tại một nút là nhỏ sẽ làm tăng tính phổ quát và ngược lại nếu sự phân chia lớn sẽ làm tăng tính cá thể của nút đó. Tính phổ quát của nút trên cây sẽ làm tăng khả năng dự đoán nhưng nguy cơ gây sai số lớn, trong khi tính cá thể giảm khả năng dự đoán nhưng lại tăng tính đúng đắn nhưng

nó cũng là nguyên nhân của tình trạng quá khớp trên cây.

Các phương pháp giải quyết bài toán mô hình cây quyết định đều phải thỏa hiệp giữa các mục tiêu này để đạt được kết quả cuối cùng.

1.5. Kết luận chương 1

Với mục tiêu nghiên cứu bài toán phân lớp dữ liệu bằng cây quyết định mờ dựa trên ĐSGT, chương này tập trung nghiên cứu, phân tích và đánh giá các vấn đề liên quan mật thiết đến đồ án.

Đầu tiên đồ án đã trình bày về khái niệm mờ, vấn đề mô hình hóa toán học cho khái niệm mờ chính là các tập mờ và khái niệm biến ngôn ngữ. Tiếp theo là phương pháp lập luận xấp xỉ trực tiếp trên ngôn ngữ, ở phần này những khái niệm và tính chất về ĐSGT lần lượt được nêu ra, đây là những kiến thức cơ sở cần thiết cho việc nghiên cứu các chương tiếp theo của đồ án. Ở đây, đồ án đã phát biểu hình thức bài toán phân lớp dữ liệu bằng cây quyết định và cũng tập trung nghiên cứu, phân tích và đánh giá các công trình nghiên cứu đã công bố gần đây, chỉ ra các vấn đề còn tồn tại để định hướng cho mục tiêu và nội dung cần giải quyết cho đồ án.

Chương 2.

PHÂN LỚP DỮ LIỆU BẰNG CÂY QUYẾT ĐỊNH MỜ THEO PHƯƠNG PHÁP ĐỐI SÁNH ĐIỂM MỜ DỰA TRÊN ĐẠI SỐ GIA TỬ

2.1. Giới thiệu

Trong chương này, trên cơ sở phân tích mối tương quan giữa các thuật toán học cây quyết định nền tảng và phân tích sự ảnh hưởng tập mẫu huấn luyện đối với cây kết quả thu được, đồ án trình bày một cách có hệ thống phương pháp lựa chọn tập mẫu huấn luyện và đề xuất thuật toán phục vụ việc học cây quyết định linh hoạt. Đồng thời, đồ án cũng đưa ra mô hình học khi tập mẫu huấn luyện có chứa giá trị mờ, định nghĩa các giá trị ngôn ngữ ngoại lai và đề xuất thuật toán nhằm thuần nhất miền trị cho các thuộc tính theo tiếp cận đại số gia tử. Cuối cùng sẽ trình bày thuật toán FMixC4.5 phục vụ cho việc học cây quyết định trên tập huấn luyện mờ.

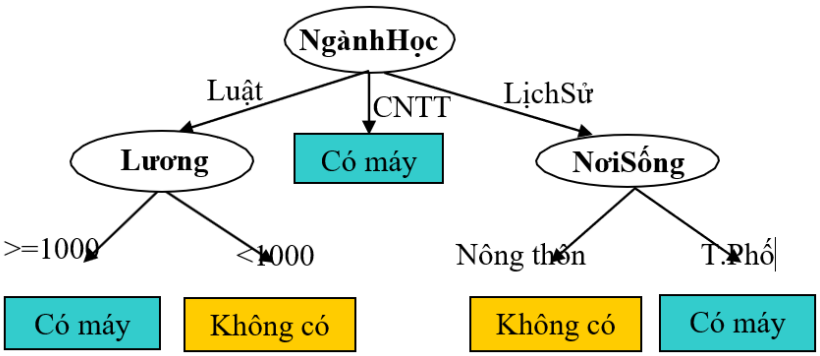
2.2. Phương pháp chọn tập mẫu huấn luyện đặc trưng cho bài toán phân lớp dữ liệu bằng cây quyết định

Ví dụ 2.1. Với dữ liệu DIEUTRA được khảo sát về tình hình có sử dụng máy tính xách tay của nhân viên như Bảng 2.1, cần chọn tập mẫu huấn luyện để huấn luyện cây quyết định cho bài toán dự đoán.

Bảng 2.1. Bảng dữ liệu DIEUTRA

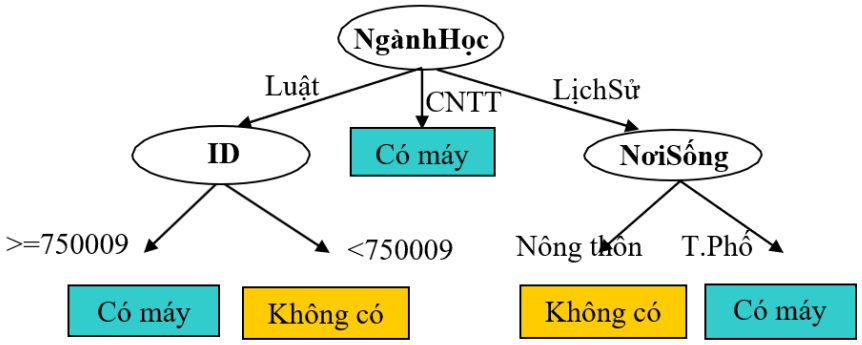
ID	PhiếuĐT	HọVàTên	NơiSống	NgànhHọc	KinhTếGĐ	Lương	PhụCấp	MáyTính
750001	M01045	Nguyễn Văn An	T.Phố	Luật	Chưa tốt	450	45	Không
750002	M01087	Lê Văn Bình	NôngThôn	Luật	Chưa tốt	400	40	Không
750003	M02043	Nguyễn Thị Hoa	T.Phố	CNTT	Chưa tốt	520	52	Có
750004	M02081	Trần Bình	T.Phố	LịchSử	Trung bình	340	34	Có
750005	M02046	Trần Thị Hương	T.Phố	LịchSử	Khá	500	50	Có
750006	M03087	Nguyễn Thị Lài	NôngThôn	LịchSử	Khá	1000	100	Không
750007	M03025	Vũ Tuấn Hoa	NôngThôn	CNTT	Khá	2000	200	Có

Giả sử chọn ***M1 = (NơiSống, NgànhHọc, KinhTếGĐ, Lương)*** làm tập mẫu huấn luyện, cây quyết định thu được áp dụng thuật toán C4.5 cho kết quả:



Hình 2.1. Cây quyết định được tạo từ tập mẫu huấn luyện M1

Tuy nhiên, nếu chọn ***M2 = (ID, NơiSống, NgànhHọc, KinhTếGĐ)*** làm tập mẫu huấn luyện, cũng áp dụng thuật toán C4.5 ta lại có cây kết quả thu được:



Hình 2.2. Cây quyết định được tạo từ tập mẫu huấn luyện M2

Như vậy, so sánh giữa cây Hình 2.1 và Hình 2.2, chúng ta dễ dàng nhận thấy cây ở Hình 2.2. là một cây không có khả năng dự đoán theo nhánh ***ID***, do không phản ánh được bản chất thực tế của dữ liệu cần học.

2.3. Phân lớp dữ liệu bằng cây quyết định dựa trên ngưỡng miền trị thuộc tính

2.3.1. Cơ sở của việc xác định ngưỡng cho quá trình học phân lớp

Như chúng ta đã xét, các thuật toán học quy nạp cây quyết định đều dựa vào việc chọn thuộc tính có lượng thông tin tốt nhất để phân tách cây và sự phân chia tại mỗi nút phụ thuộc vào kiểu của thuộc tính là liên tục hay rời rạc. Tất cả các thuật toán đều cố định cách phân chia cho mọi thuộc tính rời rạc của tập huấn luyện theo *nhị phân* hoặc *k-phân*.

- Đối với cách chia *k-phân*, một điều dễ thấy là nếu thuộc tính *A* có lực

lượng lớn sẽ làm cho số nút của cây tại một cấp tăng lên nhanh. Điều này làm tăng chiều rộng của cây nên cây sẽ dàn trải theo chiều ngang. Hơn nữa, cách chia này có khả năng dẫn đến lỗi khi dữ liệu không thể đoán nhận được lớp. Mặc dù vậy chia *k-phân* theo thuộc tính rời rạc có ưu điểm là độ phức tạp thấp, bởi vì sau khi phân thì thuộc tính đó không cần phải sử dụng lại nữa.

- Cách chia *nhị phân theo giá trị tại điểm phân chia* không làm tăng chiều rộng của cây, bởi cho dù *k* có lớn bao nhiêu cũng chỉ chia theo 2 nút, một nút là giá trị được chọn và một nút là tập còn lại. Tuy nhiên, điều này lại làm tăng nhanh chiều sâu của cây. Cách chia *nhị phân theo tập hợp tại điểm phân chia* luôn tách thuộc tính rời rạc làm 2 tập con nên và chi phí tính toán rất lớn và khó khăn trong việc duyệt cây kết quả cho quá trình dự đoán.

Từ những nhận định trên, ta nhận thấy cần phải xây dựng một thuật toán học với cách chia hỗn hợp *nhị phân, k-phân* theo từng thuộc tính nhằm có được cây với chiều rộng và chiều sâu hợp lý cho quá tình huấn luyện.

2.3.2. Thuật toán MixC4.5 dựa trên ngưỡng miền trị thuộc tính

Với tập mẫu huấn luyện *D* với *m* thuộc tính A_1, A_2, ... A_m có lực lượng tương ứng của mỗi thuộc tính là $|A_1|$, $|A_2|$, … $|A_m|$. Ta gọi *k* là ngưỡng giới hạn sự phân chia tại mỗi thuộc tính theo *nhị phân,* tức là nếu lực lượng của thuộc tính nhỏ hơn một giá trị được lựa chọn *k* cho trước thì sẽ phân theo *k-phân*, ngược lại phân theo *nhị phân*.

Với tập mẫu dữ liệu nghiệp vụ huấn luyện *D* có *m* thuộc tính, ta có thuật toán MixC4.5 xây dựng cây quyết định *S*. Với *m* là số thuộc tính, *n* là số thể hiện của tập huấn luyện. Tuần tự, chúng ta mất $O(m \times n)$ vì phải duyệt qua toàn bộ mẫu để xác định ngưỡng *k* cho *m* thuộc tính, là ngưỡng xác định sẽ chia tách theo nhị phân hay *k-phân* và tuần tự. Sau đó chúng ta mất chi phí $O(m^2 \times n)$ để lựa chọn các thuộc tính đặc trưng cho tập mẫu huấn luyện nhằm tránh tình trạng quá khớp trên cây.

Trong quá trình huấn luyện, với thuộc tính liên tục MixC4.5 hoàn toàn trùng khớp với C4.5 và SPRINT. Đối với thuộc tính rời rạc, MixC4.5 được thiết kế dựa trên sự tổng hợp của C4.5 và SPRINT, khi lực lượng của thuộc tính đang xét chưa vượt ngưỡng k, do chúng ta sử dụng k-phân theo C4.5 nên độ phức tạp lúc này là O(m × n × log n). Ngược lại, khi vượt quá ngưỡng k, chúng ta phân chia nhị phân theo giá trị theo SPRINT nên độ phức tạp lúc này là O(m × n 2 ×

log n). Vậy độ phức tạp của thuật toán MixC4.5 là O(m × n 2 × log n). Tính đúng và tính dừng của thuật toán được rút ra từ các thuật toán C4.5 và SPRINT do MixC4.5 được kết hợp từ hai thuật toán này.

2.3.3. Cài đặt thử nghiệm và đánh giá thuật toán MixC4.5

Tập mẫu huấn luyện gồm 1500 bảng ghi và 500 bộ giá trị kiểm tra được lấy từ 2155 bộ dữ liệu từ các bảng Customers, Details, OrderDetails, Products của cơ sở dữ liệu Northwind.

Thu được kết quả:

Thuật toán	Thời gian (s)	Tổng số nút	Độ chính xác (%)
C4.5	11.2	389	69.2
SLIQ	220.2	89	76.4
SPRINT	89.2	122	79.8
MixC4.5	**73.3**	**130**	**78.2**

Bảng 2.2. Bảng so sánh kết quả huấn luyện của thuật toán MixC4.5 với 1000 mẫu trên cơ sở dữ liệu Northwind

Thuật toán	Thời gian (s)	Tổng số nút	Độ chính xác (%)
C4.5	20.4	552	76.4
SLIQ	523.3	162	82.4
SPRINT	184.0	171	83.2
MixC4.5	**186.6**	**172**	**86.6**

Bảng 2.3. Bảng so sánh kết quả huấn luyện của thuật toán MixC4.5 với 1500 mẫu trên cơ sở dữ liệu Northwind

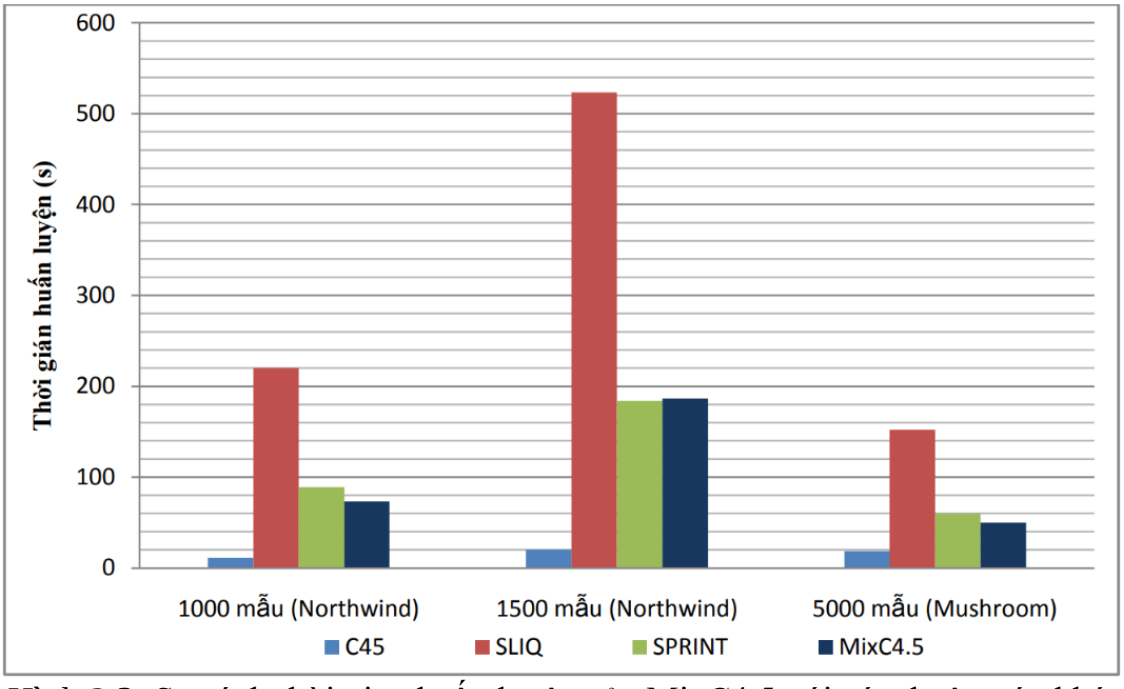
Với tập mẫu huấn luyện gồm 5000 bảng ghi và các bộ dữ liệu kiểm tra gồm 500 và 1000 mẫu được lấy từ 8000 bộ dữ liệu của cơ sở dữ liệu Mushroom. Kết quả thực nghiệm của quá trình huấn luyện và kiểm tra như sau:

Thuật toán	Thời gian huấn luyện (s)	Độ chính xác (%) 500 mẫu kiểm tra	Độ chính xác (%) 1000 mẫu kiểm tra
C4.5	18.9	54.8	51.2
SLIQ	152.3	51.8	52.2
SPRINT	60.1	54.2	54.6
MixC4.5	**50.2**	**54.8**	**54.6**

Bảng 2.4. Bảng so sánh kết quả của thuật toán MixC4.5 với 5000 mẫu huấn luyện trên cơ sở dữ liệu có chứa thuộc tính mờ Mushroom

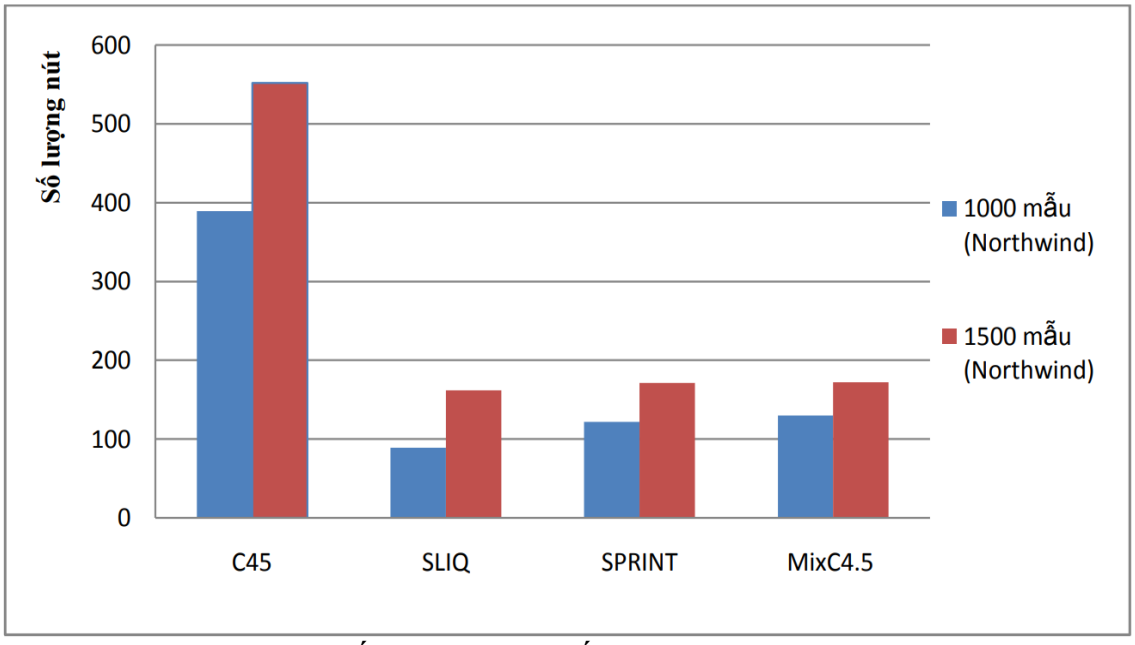
Nhận xét:

♦ **Thời gian huấn luyện:** Thuật toán C4.5 luôn thực hiện k-phân tại các thuộc tính rời rạc và loại bỏ các thuộc tính của tập huấn luyện ở mỗi bước phân chia, nên C4.5 luôn đạt tốc độ thực hiện nhanh nhất. Thời gian xử lý của SLIQ là lớn nhất do phải thực hiện các phép tính Gini trên mỗi giá trị rời rạc của thuộc tính rời rạc. Do cách phân chia của MixC4.5 trộn lẫn giữa C4.5 và SPRINT và C4.5 nhanh hơn SPRINT nên thời gian huấn luyện của MixC4.5 khá tương đồng tốt với SPRINT, thể hiện ở Hình 2.3

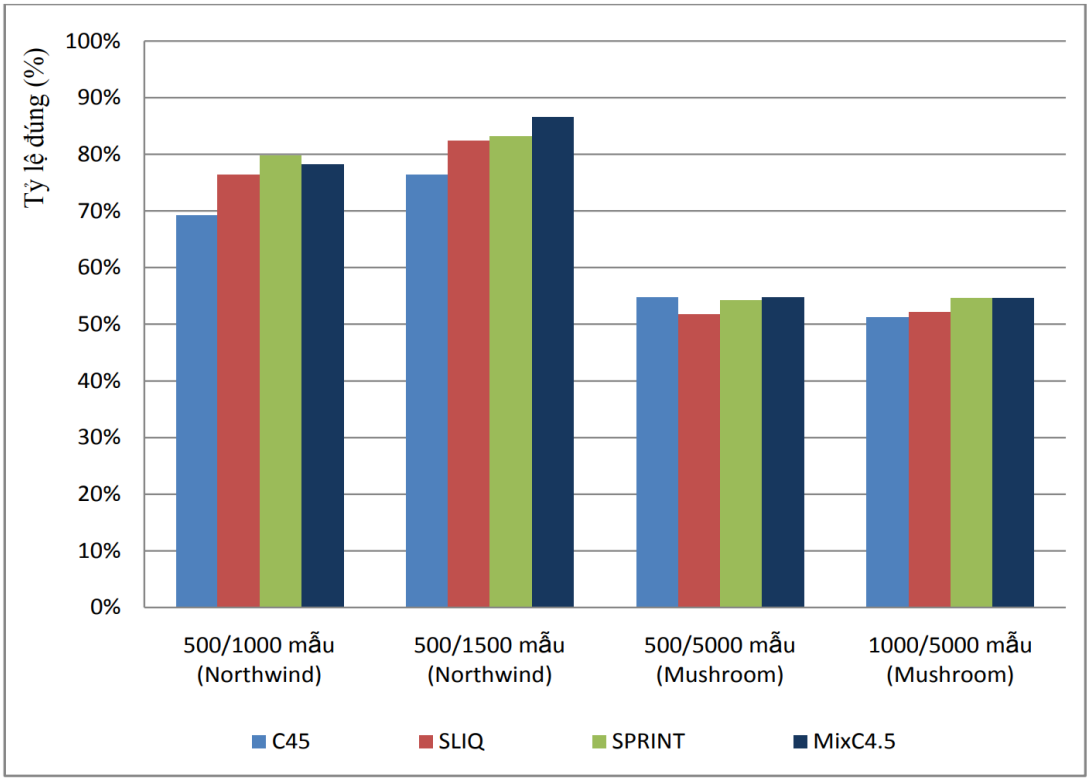


Hình 2.3. So sánh thời gian huấn luyện của MixC4.5 với các thuật toán khác

♦ **Kích thước cây kết quả:** SLIQ do thực hiện cách chia nhị phân theo tập hợp nên số nút của nó luôn nhỏ nhất và C4.5 luôn phân chia k-phân nên số nút luôn lớn nhất. Thuật toán MixC4.5 tương đồng kém với SPRINT do số lượng nút của thuật toán SPRINT ít hơn C4.5, Hình 2.4.



Hình 2.4. So sánh số nút trên cây kết quả của MixC4.5 với các thuật toán khác.



Hình 2.5. So sánh tỷ lệ đúng trên kết quả của MixC4.5 với các thuật toán khác.

♦ **Hiệu quả dự đoán:** Thuật toán MixC4.5 cải tiến từ sự kết hợp các thuật toán C4.5 và SPRINT nên cho cây kết quả có khả năng dự đoán khả quan hơn các thuật toán kinh điển. Trong tất cả các trường hợp dự đoán trên cả dữ liệu rõ và dữ liệu có chứa giá trị mờ, MixC4.5 luôn cho kết quả dự đoán phù hợp hơn C4.5, SLIQ và SPRINT do chúng ta đã hạn chế được tình huống "quá khớp" trên cây kết quả.

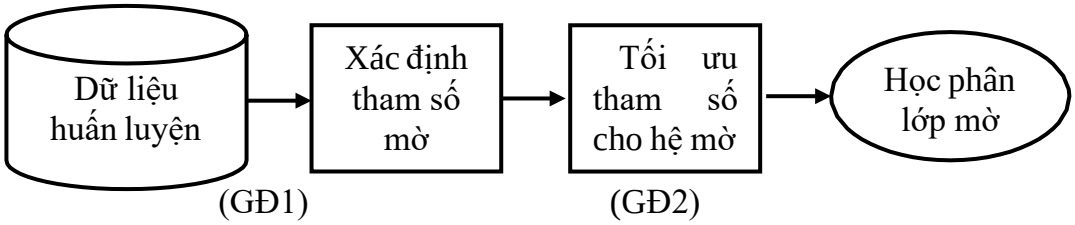
Tuy nhiên, đối sánh giữa các tập huấn luyện không có thuộc tính mờ (Northwind) và các tập huấn luyện có chứa thuộc tính mờ (Mushroom) thì khả năng dự đoán của MixC4.5 còn có sự chênh lệch lớn, khả năng dự đoán khi dữ liệu có chứa giá trị mờ chưa cao. Trong tất cả các trường hợp, các thuật toán kinh điển và thuật toán đề xuất MixC4.5 đều cho kết quả dự đoán đúng có tỷ lệ nhỏ hơn 60%, Hình 2.5. Điều này hoàn toàn hợp lý vì trong quá trình học các thuật toán đang xét không thể xử lý nên chọn giải pháp bỏ qua các giá trị mờ, vì thế kết quả dự đoán có sai số lớn.

2.4. Phân lớp dữ liệu bằng cây quyết định mờ dựa trên đối sánh điểm mờ

2.4.1. Xây dựng mô hình học phân lớp dữ liệu bằng cây quyết định mờ

Như chúng ta đã biết, bài toán phân lớp mờ đã và đang được nhiều tác giả nghiên cứu và ứng dụng, các phương pháp được biết đến như lập luận xấp xỉ mờ ,hệ nơ-ron mờ, luật kết hợp mờ, cây quyết định mờ….Các phương pháp này sử dụng các phép toán truyền thống trên tập mờ để lập luận cho kết quả đầu ra. Mô hình thể hiện cho quá trình phân lớp mờ này bao gồm 2 giai đoạn, thể hiện như ở Hình 2.6.

- *Giai đoạn 1:* xác định hệ mờ bao gồm việc lựa chọn các biến vào, các tham số mờ, phân hoạch các khoảng mờ của các biến vào, lập luận đầu ra cho hệ mờ.



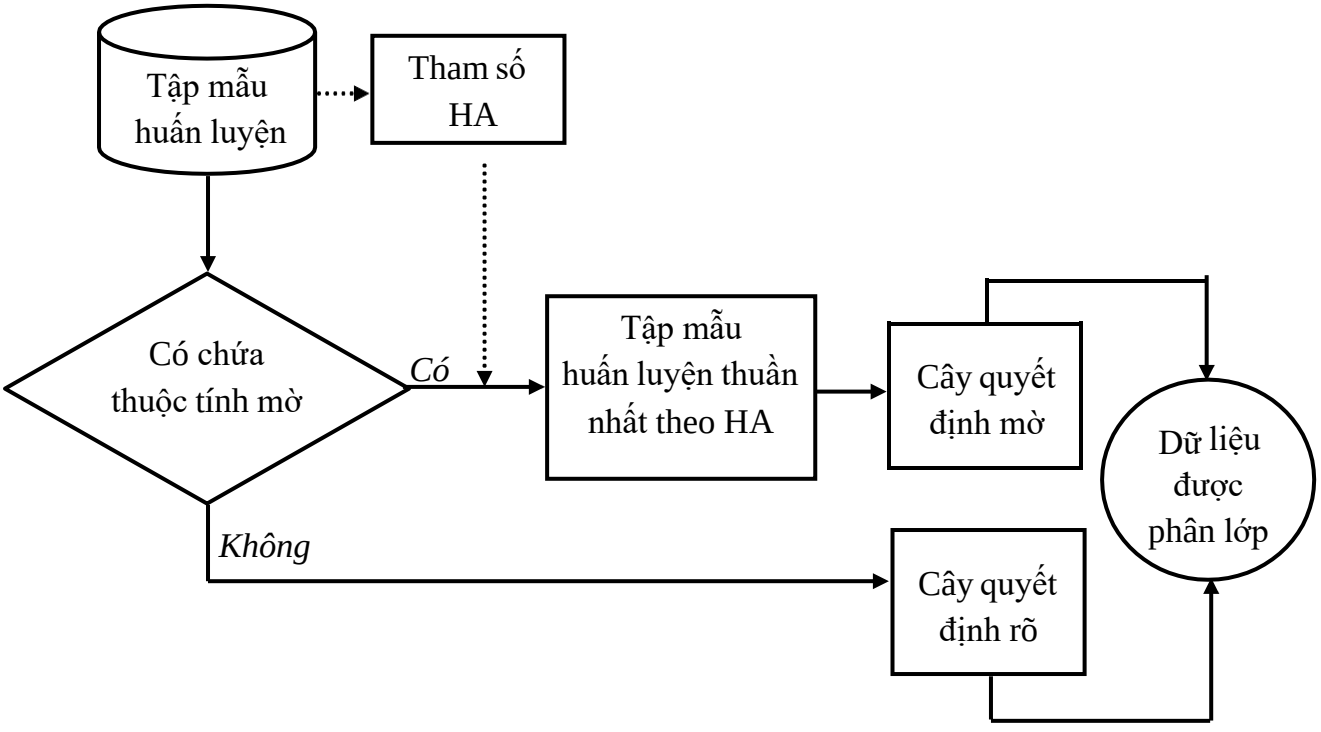
Hình 2.6. Mô hình cho quá trình học phân lớp mờ

- *Giai đoạn 2:* tối ưu các tham số của hệ mờ nhằm nâng cao hiệu quả của việc học phân lớp.

Do nhu cầu phản ánh thế giới thực nên các kho dữ liệu nghiệp vụ rất đa dạng. Vì vậy, tập mẫu huấn luyện được lấy từ các kho dữ liệu sẽ có các thuộc

tính chứa cả giá trị rõ và giá trị mờ - thường được biểu diễn bằng các giá trị ngôn ngữ và về bản chất, đây không phải là thuộc tính rời rạc. Các phương pháp học cây quyết định truyền thống mặc dầu có độ phức tạp thấp nhưng lại xem là các thuộc tính mờ như là thuộc tính rời rạc và tiến hành *k-phân* theo giá trị tại điểm này. Do vậy, cây kết quả nhận được sẽ dàn trải theo chiều ngang nên sẽ dẫn đến tình trạng "quá khớp", vì vậy mô hình nhận đượcsau quá trình học không thật sự hiệu quả.

Phân lớp dữ liệu mờ nói chung và phân lớp dữ liệu mờ bằng cây quyết định nói riêng sử dụng lý thuyết tập mờ luôn gặp phải những hạn chế xuất phát từ bản thân nội tại của lý thuyết tập mờ đó là cấu trúc thứ tự cảm sinh trên các khái niệm mờ biểu thị bằng các giá trị ngôn ngữ không được thể hiện trên các tập mờ. Thêm vào đó, việc áp dung các phương pháp học truyền thống để học cây quyết định mờ từ tập huấn luyện mờ chưa thể hiện rõ tính mờ trên cây kết quả. Do vậy, khai thác từ những đặc tính về tính có cấu trúc thứ tự của các phần tử là các giá trị ngôn ngữ và các phép toán mờ của đại số gia tử, đã gợi ý chúng ta cần xây dựng một mô hình linh hoạt và phù hợp hơn cho quá trình học phân lớp dữ liệu bằng cây quyết định mờ. Mô hình cho quá trình học được đồ án đề xuất như Hình 2.7.



Hình 2.7. Mô hình đề nghị cho việc học phân lớp dữ liệu bằng cây quyết định

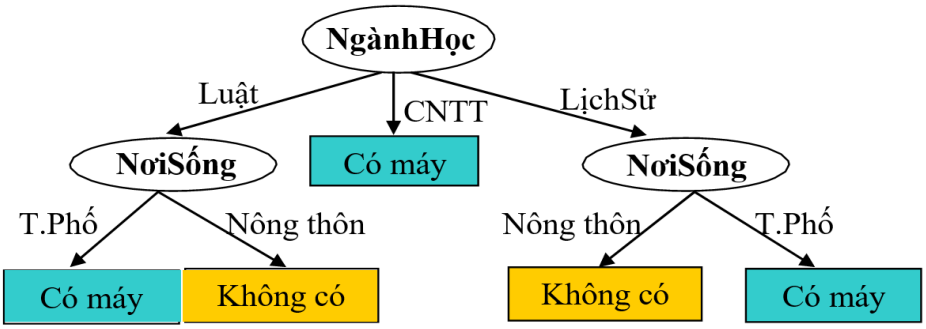
- *Giai đoạn 1:* giai đoạn này trước hết sẽ thiết lập các tham số HA và tối

ưu cho các tham số này bằng cách lựa chọn từ chính dữ liệu mẫu đang có, phân hoạch các khoảng mờ cho miền trị của các thuộc tính, chuyển tập mẫu huấn luyện nghiệp vụ về tập mẫu chứa các giá trị ngôn ngữ sử dụng đại số gia tử (hoặc các đoạn con của [0, 1]). Giai đoạn này cũng bao hàm việc xử lý tập mẫu huấn luyện nhằm loại bỏ các thuộc tính không hữu ích hay xử lý các giá trị ngoại lai nếu có.

- *Giai đoạn 2:* Áp dụng các phương pháp học để xây dựng cây quyết định rõ hay mờ (cây quyết định có nhãn là các giá trị ngôn ngữ hay các khoảng con của [0, 1]). Trường hợp xây dựng cây quyết định mờ với nhãn ngôn ngữ (hoặc đoạn con của [0, 1]), ta có thể sử dụng các giải thuật rõ ở đã biết và phân hoạch đa phân theo giá trị ngôn ngữ tại điểm phân chia mờ này hoặc phân lớp nhị phân dựa trên thứ tự của các giá trị ngôn ngữ trong đại số gia tử.

2.4.2. Vấn đề với tập mẫu huấn luyện không thuần nhất

Trong thế giới thực, dữ liệu nghiệp vụ rất đa dạng vì chúng được lưu trữ để phục vụ nhiều công việc khác nhau, nhiều thuộc tính đã được thuần nhất miền giá trị trước khi lưu trữ, nhưng cũng tồn tại nhiều thuộc tính mà miền trị của nó chứa cả dữ liệu rõ và dữ liệu mờ. Khi các thuộc tính chưa thuần nhất này xuất hiện trong tập mẫu huấn luyện, các thuật toán xây dựng cây truyền thống không thể tiến hành. Do đó, chúng ta cần phải tiền xử lý dữ liệu để có được tập mẫu huấn luyện thuần nhất. Vấn đề đặt ra là ta phải xử lý như thế nào để có được kết quả là khả quan. Với dữ liệu ở Bảng 2.7, nếu chúng ta lựa chọn giải pháp bỏ các mẫu "*lỗi*" tức là các mẫu chứa giá trị ngôn ngữ, ta sẽ thu được cây kết quả "*lệch*" như ở Hình 2.8.



Hình 2.8. Cây quyết định kết quả "sai lệch" khi tập mẫu huấn luyện bị loại bỏ giá trị ngôn ngữ

Một cách tự nhiên là chúng ta sử dụng tính chất có thứ tự của các giá trị

ngôn ngữ trong đại số gia tử để thuần nhất dữ liệu, tức là chuyển toàn bộ các dữ liệu kinh điển về các giá trị ngôn ngữ hoặc ngược lại.

2.5. Kết luận Chương 2

Với mục tiêu khắc phục các hạn chế của các thuật toán học cây quyết định truyền thống như C4.5, SPLIQ, SPRINT trên các tập huấn luyện có chứa thuộc tính mờ, chương này của đồ án tập trung:

1. Phân tích mối tương quan giữa các thuật toán học cây quyết định nền tảng và phân tích sự ảnh hưởng của tập mẫu huấn luyện đối với hiệu quả cây kết quả thu được, trình bày một phương pháp nhằm trích chọn được tập mẫu huấn luyện đặc trưng phục vụ cho quá trình huấn luyện và đề xuất thuật toán MixC4.5 phục vụ quá trình học.

2. Nhận thấy việc xây dựng mô hình cây quyết định hoặc cây quyết định mờ phụ thuộc vào mục đích và tập mẫu huấn luyện, đồ án đã đưa ra mô hình học nhằm đáp ứng yêu cầu cho yêu cầu này. Đồng thời đồ án cũng phân tích, đưa ra các khái niệm về tập mẫu không thuần nhất, giá trị ngoại lai và xây dựng thuật toán để có thể thuần nhất cho các thuộc tính có chứa các giá trị này.

3. Trên cơ sở phân tích, chỉ ra cách thuần nhất cho các thuộc tính không thuần nhất của tập mẫu và khái niệm cùng cách thức xử lý giá trị ngoại lai, chương này của đồ án cũng đã xây dựng thuật toán FMixC4.5 nhằm phục vụ cho quá trình học xây dựng cây quyết định trên tập huấn luyện này. Các kết quả thực nghiệm được đối sánh đã cho thấy khả năng dự đoán của MixC4.5, FMixC4.5 tốt hơn các thuật toán truyền thống khác.

KẾT LUẬN

Đồ án tập trung nghiên cứu, phân tích và đánh giá các ưu nhược điểm của các kết quả đã được nghiên cứu cho việc học phân lớp bằng cây quyết định. Kết quả chính của đồ án là nghiên cứu, đề xuất mô hình và các phương pháp cho việc học cây quyết định nhằm thu được cây kết quả đạt hiệu quả cao cho quá trình phân lớp và đơn giản, dễ hiểu đối với người dùng. Nội dung chính của đồ án đã đạt được các kết quả cụ thể như sau:

1. Đề xuất mô hình linh hoạt cho quá trình học cây quyết định từ tập mẫu huấn luyện thực tế và phương pháp nhằm trích chọn được tập mẫu huấn luyện đặc trưng phục vụ cho quá trình huấn luyện. Phân tích, đưa ra các khái niệm về tập mẫu không thuần nhất, giá trị ngoại lai và xây dựng thuật toán để có thể thuần nhất cho các thuộc tính có chứa các giá trị này.

2. Đề xuất thuật toán xây dựng cây MixC4.5 trên cơ sở tổng hợp các ưu và nhược điểm của các thuật toán truyền thống CART, C4.5, SLIQ, SPRINT. Với việc chỉ ra các hạn chế của thuật toán FDT và FID3 cho việc học cây quyết định mờ, đồ án đề xuất thuật toán FMixC4.5 phục vụ quá trình học cây quyết định trên tập mẫu không thuần nhất. Cả hai thuật toán MixC4.5 và FMixC4.5 đều được đánh giá thực nghiệm trên các cơ sở dữ liệu Northwind và Mushroom và kết quả có khả quan dự đoán tốt hơn các thuật toán truyền thống C4.5, SLIQ, SPRINT.

3. Đề xuất phương pháp đối sánh dựa trên khoảng mờ và xây dựng thuật toán học phân lớp dựa trên khoảng mờ HAC4.5. Xây dựng phương pháp nhằm có thể định lượng cho các giá trị của thuộc tính không thuần nhất, chưa xác định *Min - Max* của tập huấn luyện.

4. Mặc dầu vậy, trong việc lựa chọn tham số để xây dựng đại số gia tử nhằm định lượng giá trị ngôn ngữ trên tập mẫu huấn luyện, đồ án đang sử dụng kiến thức của chuyên gia để xác định các tham số mà chưa có nghiên cứu nhằm đưa ra một phương pháp hoàn chỉnh cho việc lựa chọn này.

TÀI LIỆU THAM KHẢO

TIẾNG VIỆT

[1]. Nguyễn Công Hào: *Cơ sở dữ liệu mờ với thao tác dữ liệu dựa trên đại số gia tử*, Đồ án Tiến sĩ Toán học, Viện Công nghệ Thông tin, 2008.

[2]. Nguyễn Cát Hồ, *Cơ sở dữ liệu mờ với ngữ nghĩa đại số gia tử*, Bài giảng trường Thu - Hệ mờ và ứng dụng, Viện Toán học Việt Nam, 2008.

[3]. Lê Anh Phương, *Một tiếp cận xây dựng miền giá trị chân lý ngôn ngữ trong các hệ logic*, Đồ án Tiến sĩ Toán học, Viện Công nghệ Thông tinvà Truyền Thông – Đại học Bách Khoa Hà Nội, 2013.

[4]. Lê Xuân Việt, *Định lượng ngữ nghĩa các giá trị của biến ngôn ngữ dựa trên đại số gia tử và ứng dụng*, Đồ án Tiến sĩ Toán học, Viện Công nghệ Thông tin, 2008.

[5]. Lê Xuân Vinh, *Về một cơ sở đại số và logíc cho lập luận xấp xỉ và ứng dụng*, Đồ án Tiến sĩ Toán học, Viện Công nghệ Thông tin - Viện Khoa học và Công nghệ Việt Nam, 2006.

TIẾNG ANH

[6]. Alberto Fernández, María Calderón, Francisco Herrera, *Enhancing Fuzzy Rule Based Systems in Multi-Classication Using Pairwise Coupling with Preference Relations*, University of Navarra, Spain, 2009.

[7]. B. Chandra, *Fuzzy SLIQ Decision Tree Algorithm*, IEEE, 2008.

[8]. Chida A., *Enhanced Encoding with Improved Fuzzy Decision Tree Testing Using CASP Templates*, Computational Intelligence Magazine, IEEE, 2012.

[9]. Daveedu Raju Adidela, Jaya Suma. G, Lavanya D. G., *Construction of Fuzzy Decision Tree using Expectation Maximization Algorithm*, International Journal of Computer Science and Management Research , Vol 1 Issue 3 October 2012.

[10]. Dubois D., Prade H., *Fuzzy Sets in Approximate Reasoning and Information Systems*, Kluwer Academic Publishers, USA, 1999.

[11]. Hesham A. Hefny, Ahmed S. Ghiduk, Ashraf Abdel Wahab, *Effective Method for Extracting Rules from Fuzzy Decision Trees based on Ambiguity and Classifiability*, Universal Journal of Computer Science and Engineering Technology, Cairo University, Egypt., pp. 55-63, 2010

[12]. Hongze Qiu, Haitang Zhang, *Fuzzy SLIQ Decision Tree Based on Classification Sensitivity*, Modern Education and Computer Science (MECS), pp. 18-25, 2011.

[13]. Ishibuchi H., Nojima Y., Kuwajima I., *Parallel distributed genetic fuzzy rule selection*, SpringerLink, vol. 13, no. 5, 2009.

[14]. Manish Mehta, Jorma Rissanen, Rakesh Agrawal, *SLIQ: A Fast Scalable Classifier for Data Mining*, IBM Almaden Reseach Center, 1996.

[15]. Manish Mehta, Jorma Rissanen, Rakesh Agrawal, *SPRINT: A Fast Scalable Classifier for Data Mining*, IBM Almaden Reseach Center, 1998.

[16]. Marcos E. Cintra, Maria C. Monard, Heloisa A. Camargo, *A Fuzzy Decision Tree Algorithm Based on C4.5*, Mathware & Soft Computing Magazine. Vol. 20, Num. 1, pp. 56-62, 2013.

[17]. Mariana V. Ribeiro, Luiz Manoel S. Cunha, Heloisa A. Camargo, Luiz Henrique A. Rodrigues, *Applying a Fuzzy Decision Tree Approach to Soil Classification*, Springer International Publishing Switzerland, pp. 87–96, 2014.

[18]. Narasimha Prasad, Mannava Munirathnam Naidu, *CC-SLIQ: Performance Enhancement with 2k Split Points in SLIQ Decision Tree Algorithm*, International Journal of Computer Science, 2014.

[19]. Yakun Hu, Dapeng Wu, Antonio Nucci, *Fuzzy-Clustering-Based Decision Tree Approach for Large Population Speaker Identification*, IEEE, pp. 1-13, 2010.

[20]. Zadeh L. A., *Fuzzy sets*, Information and Control 8, pp.338-358, 1965.

[21]. Zadeh L. A., *Fuzzy sets and fuzzy information granulation theory*, Beijing Normal University Press, China, 2000.

[22]. Zeinalkhani M., Eftekhari M., *Comparing Different Stopping Criteria For Fuzzy Decision Tree Induction Through IDFID3*, Iranian Journal Of

Fuzzy Systems Vol. 11, No. 1, pp. 27-48, 2014.

[23]. Zhihao Wang, Junfang Wang, Yonghua Huo, Yanjun Tuo, Yang Yang, *A Searching Method of Candidate Segmentation Point in SPRINT Classification*, Journal of Electrical and Computer Engineering, Hindawi Publishing Corporation, 2016.