MỘT CÁCH TIẾP CẬN CHỌN TẬP MẪU HUẤN LUYỆN CÂY QUYẾT ĐỊNH DỰA TRÊN ĐẠI SỐ GIA TỬ

Lê Văn Tường Lân[1], Nguyễn Mậu Hân[1], Nguyễn Công Hào[2]
[1] Khoa Công nghệ Thông tin, Đại học Khoa học, Đại học Huế, Huế, Việt Nam.
[2]Trung tâm Công nghệ Thông tin, Đại học Huế, Huế, Việt Nam.
{lvtlan, nmhan2005}@yahoo.com, nchao@hueuni.edu.vn

Tóm tắt. Cây quyết định là một trong những giải pháp trực quan và hữu hiệu để mô tả quá trình phân lớp dữ liệu. Xây dựng một cây quyết định phục vụ khai phá dữ liệu hiệu quả phụ thuộc lớn vào việc chọn tập mẫu huấn luyện. Trong thực tế, dữ liệu nghiệp vụ được lưu trữ rất đa dạng và phức tạp cho nên chọn tốt tập dữ liệu mẫu huấn luyện gặp rất nhiều khó khăn. Trong bài viết này, chúng tôi nêu lên một cách để chọn tập mẫu huấn luyện tốt từ cơ sở dữ liệu nghiệp vụ dựa trên đại số gia tử, nhằm xây dựng được cây quyết định mờ có khả năng dự đoán cao cho các bài toán phân lớp dữ liệu.

Từ khoá: Đại số gia tử, khai phá dữ liệu, phát hiện tri thức, cây quyết định, cây quyết định mờ, mẫu huấn luyện.

AN APPROACH FOR CHOOSING A TRAINING SET TO BUILD A DECISION TREE BASED ON HEDGE ALGEBRA

Abstract. A decision tree is one of the intuitive and effective solutions to describe the process of data classification. Building an effective decision tree depends on the selection of training set.In practice, business data have been stored in multiform and of complexity, which consequently leads to the difficulty in selecting a good sample training set. If an untypical sample of training set is chosen, it will lead to low practicability in the corresponding decision tree. In this article, we have analysed and presented one effective approach for choosing sample training set from business database base on hedge algebra to build an fuzzy effective decision tree of high predictability for supporting decision making in data analysis problems.

1 ĐẶT VẤN ĐỀ

Cho một tập huấn luyện, tất cả các mẫu của tập đều có chung một cấu trúc, gồm những cặp <*thuộc tính, giá trị*>, một trong những thuộc tính này đại diện cho lớp và ta gọi là thuộc tính dự đoán hay thuộc tính phân lớp. Bài toán phân lớp là bài toán tìm quy tắc xếp các đối tượng vào một trong các lớp đã cho dựa trên tập mẫu huấn luyện. Có nhiều phương pháp tiếp cận bài toán phân lớp: Hàm phân biệt tuyến tính Fisher, Naïve Bayes, Logistic, Mạng nơ-ron, Cây quyết định, … trong đó phương pháp cây quyết định là phương pháp phổ biến do tính trực quan, dễ hiểu và hiệu quả của nó [2, 14, 17].

Cây quyết định được xây dựng dựa trên một tập dữ liệu huấn luyện bao gồm các đối tượng mẫu. Mỗi đối tượng được mô tả bởi một tập giá trị các thuộc tính và nhãn lớp. 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. Trên mẫu huấn luyện M, về cơ bản, các thuật toán phân lớp phải thực hiện 2 bước sau:

Bước 1: Chọn thuộc tính U_J có các giá trị $u_1, u_2,\ldots, u_n$;

Bước 2: Với thuộc tính U_J được chọn, ta tạo một nút của cây và sau đó chia các mẫu ứng với nút này thành các tập tương ứng $M_1, M_2, \ldots, M_{n,}$; Sau đó lại tiếp tục [17].

Đây là bước phân chia với kết quả nhận được từ bước 1, điều này có nghĩa là chất lượng của cây kết quả phụ thuộc phần lớn vào cách chọn thuộc tính và cách phân chia các mẫu tại mỗi nút. Chính vì điều này, các thuật toán đều phải tính lượng thông tin nhận được trên các thuộc tính và chọn thuộc tính tương ứng có lượng thông tin tốt nhất để làm nút phân tách trên cây, nhằm để đạt được cây có ít nút nhưng có khả năng dự đoán cao [2][14][17].

Việc xây dựng một cây quyết định có hiệu quả phụ thuộc vào việc chọn tập mẫu huấn luyện. Trong thực tế, dữ liệu nghiệp vụ rất đa dạng vì chúng được lưu trữ để phục vụ nhiều công việc khác nhau, nhiều thuộc tính cung cấp các thông tin có khả năng dự đoán sự việc nhưng cũng có nhiều thuộc tính không có khả năng phản ánh thông tin dự đoán mà chỉ có ý nghĩa lưu trữ, thống kê bình thường. Điều này gây khó khăn cho chúng ta khi chọn tốt tập mẫu huấn luyện để xây dựng cây.

Cho bảng dữ liệu DIEUTRA lưu trữ về tình hình mua máy tính xách tay của khách hàng tại một công ty như bảng 1, cần chọn mẫu huấn luyện để xây dựng cây quyết định cho việc dự đoán khách hàng mua máy hay không.

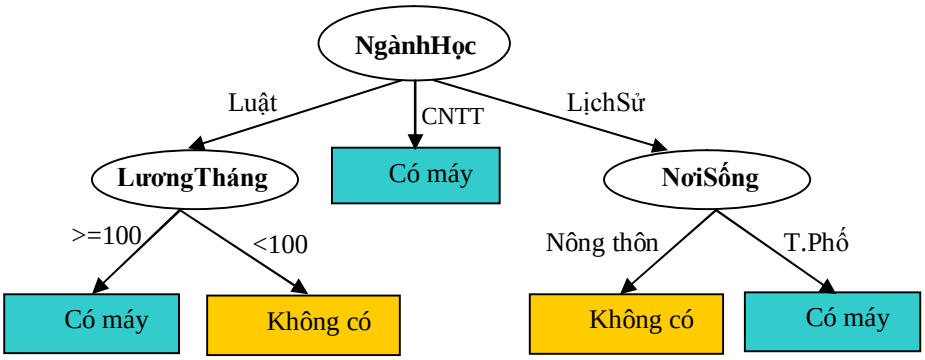
Bảng 1: Bảng dữ liệu DIEUTRA

PhiếuĐT	HọVàTên	SốCMND	NơiSống	NgànhHọc	KinhTếGĐ	LươngTháng	MáyTính
M01045	Nguyễn Văn An	193567450	T.Phố	Luật	Chưa tốt	45	Không
M01087	Lê Văn Bình	191568422	NôngThôn	Luật	Chưa tốt	40	Không
M02043	Nguyễn Thị Hoa	196986568	T.Phố	CNTT	Chưa tốt	52	Có
M02081	Trần Bình	191003117	T.Phố	LịchSử	Trung bình	34	Có
M02046	Trần Thị Hương	196001278	T.Phố	LịchSử	Khá	50	Có
M03087	Nguyễn Thị Lài	198235457	NôngThôn	LịchSử	Khá	100	Không
M03025	Vũ Tuấn Hoa	198875584	NôngThôn	CNTT	Khá	200	Có
M03017	Lê Bá Linh	191098234	T.Phố	Luật	Trung bình	35	Không
M04036	Bạch Ân	196224003	T.Phố	Luật	Khá	100	Có
M04037	Lý Thị Hoa	196678578	T.Phố	LịchSử	Trung bình	50	Có
M04042	Vũ Quang Bình	197543457	NôngThôn	Luật	Trung bình	100	Có
M04083	Nguyễn Hoa	192267457	NôngThôn	CNTT	Trung bình	40	Có
M05041	Lê Xuân Hoa	198234309	T.Phố	CNTT	Chưa tốt	55	Có
M05080	Trần Quế Chung	196679345	NôngThôn	LịchSử	Trung bình	50	Không

Ta thấy, việc xây dựng cây sẽ phụ thuộc vào nhiều yếu tố: Phụ thuộc vào việc chọn tập huấn luyện, Phụ thuộc vào tính nhất quán của dữ liệu huấn luyện.

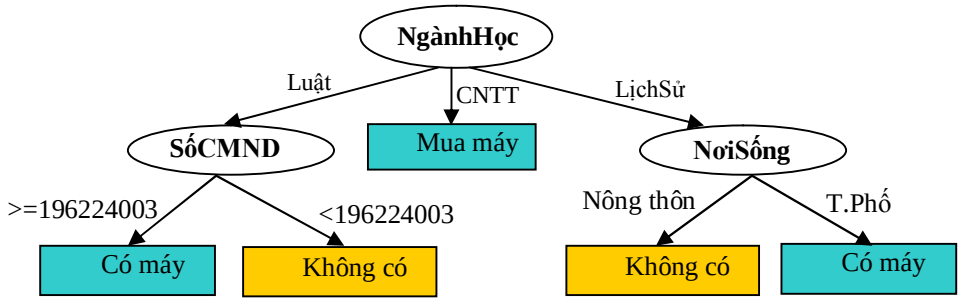
1.1 Phụ thuộc vào việc chọn tập huấn luyện

Giả sử ta chọn tập ***M1 = (*NơiSống, NgànhHọc, *KinhTếGĐ*, LươngTháng)*** làm tập mẫu huấn luyện cho việc xây dựng cây. Lúc này cây quyết định thu được ở hình 1[2, 17].



Hình 1: Cây quyết định được tạo từ tập mẫu huấn luyện M1

Tuy nhiên, nếu ta chọn tập ***M2 = (SốCNND, NơiSống,* NgànhHọc, *KinhTếGĐ)*** làm tập mẫu huấn luyện để xây dựng cây, ta sẽ được cây như hình 2 và ta dễ dàng thấy đây là một cây không hiệu quả do không phản ánh được bản chất thực tế của dữ liệu.



Hình 2: Cây quyết định không có hiệu quả được tạo từ tập huấn luyện M2

1.2 Phụ thuộc vào tính nhất quán của dữ liệu huấn luyện

Bảng 2: Tập mẫu có thuộc tính với dữ liệu không nhất quán (**LươngTháng**)

PhiếuĐT	HọVàTên	SốCMND	NơiSống	NgànhHọc	KinhTếGĐ	LươngTháng	MáyTính
M01045	Nguyễn Văn An	193567450	T.Phố	Luật	Chưa tốt	45	Không
M01087	Lê Văn Bình	191568422	NôngThôn	Luật	Chưa tốt	Thấp	Không
M02043	Nguyễn Thị Hoa	196986568	T.Phố	CNTT	Chưa tốt	52	Có
M02081	Trần Bình	191003117	T.Phố	LịchSử	Trung bình	34	Có
M02046	Trần Thị Hương	196001278	T.Phố	LịchSử	Khá	Cao	Có
M03087	Nguyễn Thị Lài	198235457	NôngThôn	LịchSử	Khá	Cao	Không
M03025	Vũ Tuấn Hoa	198875584	NôngThôn	CNTT	Khá	Rất cao	Có
M03017	Lê Bá Linh	191098234	T.Phố	Luật	Trung bình	35	Không
M04036	Bạch Ân	196224003	T.Phố	Luật	Khá	100	Có
M04037	Lý Thị Hoa	196678578	T.Phố	LịchSử	Trung bình	50	Có
M04042	Vũ Quang Bình	197543457	NôngThôn	Luật	Trung bình	Rất cao	Có
M04083	Nguyễn Hoa	192267457	NôngThôn	CNTT	Trung bình	Ít thấp	Có
M05041	Lê Xuân Hoa	198234309	T.Phố	CNTT	Chưa tốt	55	Có
M05080	Trấn Quế Chung	196679345	NôngThôn	LịchSử	Trung bình	50	Không

Trong trường hợp này, các thuật toán học có thể lựa chọn cách thức bỏ qua các bộ dữ liệu "lỗi" hay nhờ ý kiến chuyên gia để xác định các giá trị "lỗi". Tuy vậy việc này sẽ làm mất dữ liệu hay phụ thuộc lớn vào trình độ của chuyên gia, ví dụ nhiệt độ rất cao là 40 nhưng tuổi thọ thì lại rất thấp [5, 6, 11, 22] nên kết quả thu được chưa cao.

2 RÚT GỌN MẪU HUẤN LUYỆN DỰA VÀO SỰ PHÂN TÍCH BẢN CHẤT THUỘC TÍNH CỦA TẬP MẪU

Cho mẫu huấn luyện M với thuộc tính quyết định Y gồm có n bộ, m thuộc tính.

Với thuộc tính $U = \{u_1, u_2,....,u_n\}$, ta ký hiệu $|U|$ là số các giá trị khác nhau của của tập $\{u_1, u_2,....,u_n\}$ gọi là lực lượng của U; số lần xuất hiện giá trị u_i trong U ký hiệu là $|u_i|$.

Định nghĩa 1. [2] Một cây quyết định được gọi là cây dàn trải nếu số nhánh phân chia tại một nút bất kỳ lớn hơn tích của |Y| với chiều sâu của cây.

Định nghĩa 2. [2] Thuộc tính $U \in M$ được gọi là thuộc tính có giá trị riêng biệt (gọi tắt là thuộc tính riêng biệt) nếu như $|U| > (m-2)*|Y|$. Tập các thuộc tính có giá trị riêng biệt trong M ký hiệu là M*.

Định lý 1. [2] Quá trình xây dựng cây nếu có một nút bất kỳ được tạo dựa trên thuộc tính riêng biệt thì kết quả thu được là một cây dàn trải.

Thật vậy, mẫu M có m thuộc tính nên có m-1 thuộc tính dự đoán và chiều sâu tối đa của cây là m-2 [2, 17]. Với thuộc tính U là riêng biệt và nó được chọn làm điểm phân tách cây thì theo các thuật toán xây dựng cây [2], tại nút này có ít nhất((m-2)*|Y|+1) nhánh nên đây là cây dàn trải.

Định nghĩa 3. [2] Thuộc tính $U = \{u_1, u_2,.....,u_n\} \in M$ mà giữa các phần tử u_i, u_j với $i \neq j$ không sánh được thì ta gọi U là thuộc tính ghi nhớ trong tập mẫu. Tập các thuộc tính này trong M ký hiệu là M^G.

Mệnh đề 1. Nếu $U \in M$ là thuộc tính ghi nhớ thì ta loại U ra khỏi mẫu M mà không làm thay đổi cây quyết định thu được.

Hiển nhiên, bởi ta không thể so sánh giữa các phần tử u_i với u_j của U để tính hàm Gain(U,Y,M) nên không tồn tại lợi ích thông tin của mỗi bộ trên U. Vì thế U không thể xuất hiện trên cây kết quả nên ta loại U ra khỏi M mà không làm thay đổi cây quyết định thu được.

Mệnh đề 2. Nếu thuộc tính U là khoá của mẫu M thì loại U ra khỏi M để thu được cây quyết định có khả năng dự đoán tốt hơn.

Thật vậy, giả sử $U = \{u_1, u_2,.....,u_n\}$. Do U là khoá nên ta có $u_i \neq u_j$, $\forall\, i \neq j$. Như thế, mẫu M được phân ra làm n phân hoạch, mà mỗi phân hoạch chỉ có 1 bộ nên hàm $E(U,u_i,Y,M) = 0$, $\forall\, u_i \in U$. Hàm xác định thông tin nhận được trên thuộc tính U là $\text{Gain}(U, Y, M) = S(Y \mid M) - \sum_{i=1}^{n} \frac{1}{n} E(U, u_i, Y, M) = S(Y|M)$ đạt giá trị cực đại, vì thế chọn U làm điểm phân tách cây. Tại đây, cây được phân chia làm n nút, mỗi cạnh tương ứng được gán

nhãn u_i, đây là một cây dàn trải theo chiều ngang tại nút U. Tuy vậy, do tính duy nhất của khoá nên không có giá trị trùng khớp khi so sánh tại nút này trong quá trình dự đoán. Do vậy cây không có khả năng dự đoán nên phải loại U ra khỏi M.

Định nghĩa 4. [2] Nếu U = {u_1, u_2,…..,u_n} là thuộc tính riêng biệt mà ta không thể phân nhóm cho các giá trị u_i của U theo các phép tính toán thông thường thì ta gọi U là thuộc tính tự do. Tập các thuộc tính này trong M ký hiệu là M^T.

Mệnh đề 3. Nếu U là thuộc tính tự do của mẫu huấn luyện M thì ta loại U ra khỏi M để cây quyết định thu được không phải là cây dàn trải.

Thật vậy, do U є M^T nên không thể phân cụm để tạo cây theo thuộc tính riêng biệt như đã nói. Mặt khác, do U là riêng biệt nên |U| > (m-2)*|Y|. Như thế đây là nút dàn trải trên cây, theo định lý 1. Vậy phải loại U ra khỏi M.

Mệnh đề 4. Trên mẫu M với thuộc tính quyết định Y, nếu có phụ thuộc hàm U_1->U_2 và nếu đã chọn U_1 làm nút phân tách trên cây thì mọi nút con của nó sẽ không nhận U_2 làm nút phân tách.

Thật vậy, giả sử $|U_1|$ = k, khi chọn U_1 làm nút phân tách trên cây thì tại nút này ta có k nhánh. Không mất tính tổng quát, các nhánh của cây lần lượt được gán các giá trị U=u_i, i=1,.,.k. Do U_1->U_2 nên tại nhánh bất kỳ thì trên mẫu huấn luyện tương ứng M', lúc này trên thuộc tính U_2 sẽ có cùng 1 giá trị. Như thế Gain(U_2,Y,M')=0 là nhỏ nhất nên U_2 không thể chọn để làm nút phân tách cây.

Mệnh đề 5. Trên mẫu M với thuộc tính quyết định Y, nếu có phụ thuộc hàm U_1->U_2 thì lượng thông tin nhận được trên U_1 không nhỏ hơn lượng thông tin nhận được trên U_2.

Thật vậy, giả sử thuộc tính quyết định Y có k giá trị. Do U_1->U_2 nên $|U_1| \geq |U_2|$. Theo [7,10] thì lượng thông tin nhận được trên thuộc tính U là Gain(U,Y,M) được xác định theo công thức :

$$\text{Gain}(U,Y,M) = S(Y|M) - \sum_{\forall u_i \in \{U\}} E(U, x_i, Y, M)$$

Nếu $|U_1| = |U_2|$ thì trên U_1 hay U_2 đều có k phân hoạch như nhau nên Gain(U_1,Y,M)=Gain(U_2,Y,M) .

Ngược lại nếu $|U_1| > |U_2|$ tức tồn tại u_{1i}, u_{1j} ɔ U_1, $u_{1i} \neq u_{1j}$ mà trên tương ứng trên U_2 thì $u_{2i} = u_{2j}$. Lúc này 2 phân hoạch trên U_1 được gộp thành 1 phân hoạch trên U_2 nên entropy tương ứng trên U_2 lớn hơn. Vậy Gain(U_1,Y,M)>Gain(U_2,Y,M).

Hệ quả. Nếu có phụ thuộc hàm U_1->U_2 mà U_1 không phải là thuộc tính khóa của mẫu M thì thuộc tính U_2 không được chọn làm nút phân tách cây.

Như vậy, với dữ liệu đã cho ở bảng 1 hay bảng 2, các thuộc tính sau sẽ bị loại bỏ nếu nó có xuất hiện trong mẫu huấn luyện:

- Thuộc tính *ThuNhậpGĐ* là thuộc tính riêng biệt.
- Thuộc tính *SốCMND, PhiếuĐT* là khoá.
- Thuộc tính *HọVàTên* của khách hàng là tự do.

Đánh giá:

Sau khi kết thúc quá trình xử lý, tập huấn luyện đã được giới hạn chỉ còn các thuộc tính hữu dụng. Như tập mẫu đã cho ở bảng 2 thì kết quả sau khi xử lý như ở bảng 3. Do đó, người dùng không thể sinh ra cây không có hiệu quả như đã đề cập ở hình 2.

Bảng 3. Tập mẫu sau khi đã thực hiện quá trình loại bỏ các thuộc tính không khả dụng.

NơiSống	NgànhHọc	KinhTếGĐ	LươngTháng	MáyTính
T.Phố	Luật	Chưa tốt	45	Không
NôngThôn	Luật	Chưa tốt	Thấp	Không
T.Phố	CNTT	Chưa tốt	52	Có
T.Phố	LịchSử	Trung bình	34	Có
T.Phố	LịchSử	Khá	Cao	Có
NôngThôn	LịchSử	Khá	Cao	Không
NôngThôn	CNTT	Khá	Rất cao	Có
T.Phố	Luật	Trung bình	35	Không
T.Phố	Luật	Khá	100	Có
T.Phố	LịchSử	Trung bình	50	Có
NôngThôn	Luật	Trung bình	Rất cao	Có
NôngThôn	CNTT	Trung bình	Ít thấp	Có
T.Phố	CNTT	Chưa tốt	55	Có
NôngThôn	LịchSử	Trung bình	50	Không

Tuy vậy, với tập mẫu có chứa các thuộc tính không thuần nhất, như ở bảng 3, ta khó để xây dựng được cây nếu không loại bỏ các giá trị không thuần nhất này, như thuộc tính *LươngTháng*. Như đã đề cập, việc loại bỏ theo ý kiến của chuyên gia phụ thuộc nhiều vào trình độ của chuyên gia. Bước tiếp ở đây, ta sẽ sử dụng đại số gia tử để giải quyết vấn đề này.

3 SỬ DỤNG ĐẠI SỐ GIA TỬ ĐỂ THUẦN NHẤT GIÁ TRỊ CHO CÁC THUỘC TÍNH CỦA MẪU HUẤN LUYỆN

Xét miền trị của biến ngôn ngữ của thuộc tính chưa thuần nhất trong tập mẫu, ở đây là *LươngTháng, ta có* Dom(Lương Tháng) = {*cao, thấp, rất cao, rất thấp, ít cao, ít thấp, cao hơn, thấp hơn, khả năng cao, khả năng thấp, ...}* trong đó *cao, thấp* là các từ nguyên thuỷ, các từ nhấn *rất, hơn, ít, khả năng* gọi là các gia tử.

Khi đó miền ngôn ngữ T = Dom(*LươngTháng*) có thể biểu thị như một đại số $\underline{X}$=(**X**, G, H, ≤), trong đó G là tập các từ nguyên thuỷ {thấp, cao} được xem là các phần tử sinh. H là tập các gia tử được xem như là các phép toán một ngôi, quan hệ ≤ 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 $x = h_n h_{n-1}.....h_{1x}$, x∈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 $c^- < c^+$. Ở đây, *cao* là dương còn *thấp* là âm.

Cho đại số gia tử $\underline{X}$ = (**X**, G, H, ≤), với G = {c^+,c^-}, trong đó c^+ và c^- tương ứng là phần tử sinh dương và âm, **X** là tập nền. H =$H^+ \cup H^-$ với H^- = {h_1,h_2,.., h_p} và H^+ = {h_{p+1}, ..., h_{p+q}}, $h_1 > h_2 > ... > h_p$ và $h_{p+1} < ... < h_{p+q}$.

Định nghĩa 5. [4-6] Hàm *fm*: $\boldsymbol{X} \to [0,1]$ được gọi là *độ đo tính mờ* trên ***X*** nếu thoả mãn các điều kiện sau:

(1) $fm(c^-) = W > 0$ và $fm(c^+) = 1 - W > 0$

(2) Với $c \in \{c^-, c^+\}$ thì $\sum_{i=1}^{p+q} fm(h_i c) = fm(c)$.

(3) Với $\forall x,y \in \boldsymbol{X}$, $\forall h \in H$, $\frac{fm(hx)}{fm(x)} = \frac{fm(hy)}{fm(y)} = \frac{fm(hc)}{fm(c)}$, với $c \in \{c^-, c^+\}$,

nghĩa là tỉ số này không phụ thuộc vào *x* và *y*, được kí hiệu là $\mu(h)$ gọi là độ đo tính mờ của gia tử *h*.

Mệnh đề 6. [4-6]

(1) fm(hx) = μ(h)fm(x), với ∀x∈X

(2) $\sum_{i=1}^{p+q} fm(h_i c) = fm(c)$, trong đó c ∈ {c–, c^+}

(3) $\sum_{i=1}^{p+q} fm(h_i x) = fm(x)$, với ∀x∈X

(4) $\sum_{i=1}^{p} \mu(h_i) = \alpha$ và $\sum_{i=p+1}^{p+q} \mu(h_i) = \beta$, với α, β > 0 và α + β = 1.

Định nghĩa 6. [4-6] Giả sử cho trước độ đo tính mờ của các gia tử $\mu(h)$, và các giá trị độ đo tính mờ của các phần tử sinh $fm(c^-)$, $fm(c^+)$ và w là phần tử trung hòa. Hàm định lượng ngữ nghĩa (*quantitatively semantic function*) ν của *X* được xây dựng như sau với $x = h_{im}...h_{i2}h_{i1}c$:

(1) $\nu(c^-) = W - \alpha.fm(c^-)$ và $\nu(c^+) = W + \alpha.fm(c^+)$

(2) $\nu(h_j x) = \nu(x) + Sign(h_j x) \times \left[\sum_{i=j}^{p} fm(h_i x) - \frac{1}{2}(1 - Sign(h_j x) Sign(h_1 h_j x)(\beta - \alpha)) fm(h_j x) \right]$ với 1≤ j ≤ p,

và $\nu(h_j x) = \nu(x) + Sign(h_j x) \times \left[\sum_{i=p+1}^{j} fm(h_i x) - \frac{1}{2}(1 - Sign(h_j x) Sign(h_1 h_j x)(\beta - \alpha)) fm(h_j x) \right]$ với j > p

Định nghĩa 7[4-6]. Cho $Dom(A_i) = D_{Ai} \cup LD_{Ai}$, ν là hàm định lượng ngữ nghĩa của $Dom(A_i)$. Hàm *IC*: $Dom(A_i) \to [0,1]$ được xác định như sau:

Nếu $LD_{Ai} = \emptyset$ và $D_{Ai} \neq \emptyset$ thì $\forall \omega \in Dom(A_i)$ ta có $IC(\omega) = 1 - \frac{\psi_{\max} - \omega}{\psi_{\max} - \psi_{\min}}$ với $Dom(A_i) = [\psi_{\min}, \psi_{\max}]$ là miền trị kinh điển của A_i.

Nếu $D_{Ai} \neq \varnothing$, $LD_{Ai} \neq \varnothing$ thì $\forall \omega \in Dom(A_i)$ ta có $IC(\omega) = \{\omega_* \nu(\psi_{maxLV})\}/\psi_{max}$, với $LD_{Ai} = [\psi_{minLV}, \psi_{maxLV}]$ là miền trị ngôn ngữ của A_i.

Định nghĩa 8. Cho đại số gia tử $\underline{X}=(X, G, H, \leq)$, ν là hàm định lượng ngữ nghĩa của X.
Φ_k: $[0,1] \rightarrow X$ gọi là hàm ngược của hàm ν theo mức k được xác định:
$\forall a \in [0,1]$, $\Phi_k(a) = x^k$ khi và chỉ khi $a \in I(x^k)$, với $x^k \in X^k$.

Như thế, với bất kỳ một thuộc tính không thuần nhất U, ta sẽ chuyển về giá trị ngôn ngữ để rồi có thể chuyển về giá trị số thuần nhất.

Như vậy, trong mẫu đã cho ở bảng 3, ta sẽ xây dựng 1 ĐSGT để tính cho thuộc tính không thuần nhất *LươngTháng* như sau:

$\underline{X}_{LươngTháng} = (X_{LươngTháng}, G_{LươngTháng}, H_{LươngTháng}, \leq)$, với $G_{LươngTháng}$ = {*cao, thấp*}, $H^+_{LươngTháng}$ = {hơn, *rất*}, $H^-_{LươngTháng}$ = {*khả năng, ít*} với quan hệ ngữ nghĩa: *rất > hơn* và *ít >khả năng*.

$W_{LươngTháng}$ = 0.6, *fm*(*thấp*) = 0.4, *fm*(*cao*) = 0.6, *fm*(*rất*) = 0.35, *fm*(*hơn*) = 0.25, *fm*(*khả năng*) = 0.20, *fm*(*ít*) = 0.20.

Lúc này ta có : *fm*(*rất thấp*) = 0.35x0.4=0.14, *fm*(*hơn thấp*) =025x0.4=0.10, *fm*(*ít thấp*) = 0.2x0.4=0.08, *fm*(*khả năng thấp*) =0.2x0.4= 0.08.

Vì *rất thấp* < *hơn thấp* < *thấp* < *khả năng thấp* < *ít thấp* nên : *I*(*rất thấp*) = [0,0.14], *I*(*hơn thấp*) = [0.14,0.24], *I*(*khả năng thấp*) = [0.24,0.32], *I*(*ít thấp*) = [0.32,0.4].

Ta lại có: *fm*(*rất cao*) = 0.35x0.6=0.21, *fm*(*hơn cao*) =025x0.6=0.15, *fm*(*ít cao*) = 0.2x0.6=0.12, *fm*(*khả năng cao*) =0.2x0.6= 0.12.

Vì *ít cao* < *khả năng cao* < *cao* < *hơn cao* < *rất cao* nên : *I*(*ít cao*) = [0.4,0.52], *I*(*khả năng cao*) = [0.52,0.64], *I*(*hơn cao*) = [0.64,0.79], *I*(*rất cao*) = [0.79,1].

Trong tập mẫu ở bảng 3, với thuộc tính không thuần nhất *LươngTháng*, ta có
$U_{LươngTháng}$ = {45, Thấp, 52, 34,Cao, Cao, Rất cao, 35,100, 50, Rất cao, Ít thấp, 55, 50},
chọn ψ_1= 100$\in X_{LươngTháng}$ khi đó $\forall \omega \in Num(LươngTháng)$, $IC(\omega)$ = {0.45, 0.24, 0.52, 0.34, 0.64, 0.64, 0.79, 0.35, 1, 0.50, 0.79, 0.4, 0.55, 0.50}.

Do đó $\Phi_2(0.45)$ = *ít cao* vì $0.45 \in I$(*ít cao*), tương tự cho các giá trị còn lại, ta có thuộc tính *LươngTháng* theo ngữ nghĩa sẽ như sau: *{Ít cao, Thấp, Khả năng cao, Ít thấp, Cao, Cao, Rất cao, Ít thấp, Rất cao, Khả năng cao, Rất cao, Ít thấp, Khả năng cao, Khả năng cao*}. Lúc này, thuộc tính *LươngTháng* sẽ được làm thuần nhất theo giá trị là: {45, 24, 52, 34, 64, 64, 79, 35,100, 50, 79, 40, 55, 50}.

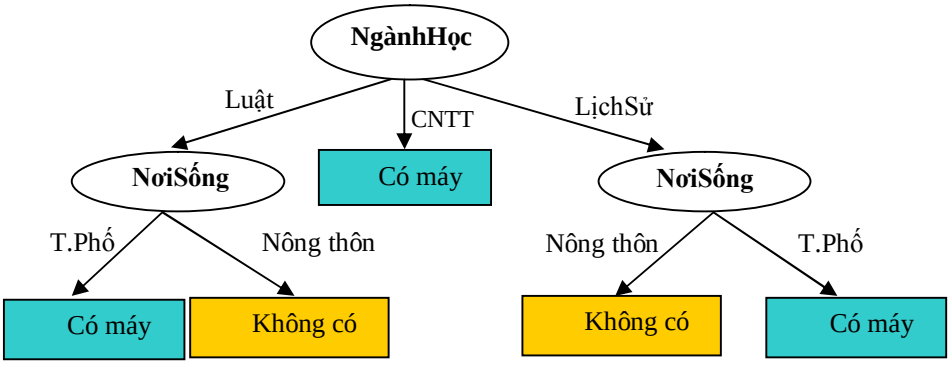
Đánh giá:

a. Khi ta chọn tập mẫu đã cho ở bảng 3 làm tập huấn luyện, ta không thể tạo cây do tính không thuần nhất của thuộc tính *LươngTháng*. Nếu ta loại bỏ các mẫu "lỗi" ta sẽ được tập huấn luyện như ở bảng 4. Tiến hành xây dựng cây trên tập huấn luyện ở bảng 4, ta thu được cây kết quả ở hình 5.

Bảng 4. Tập mẫu sau khi đã thực hiện quá trình loại bỏ các mẫu "lỗi".

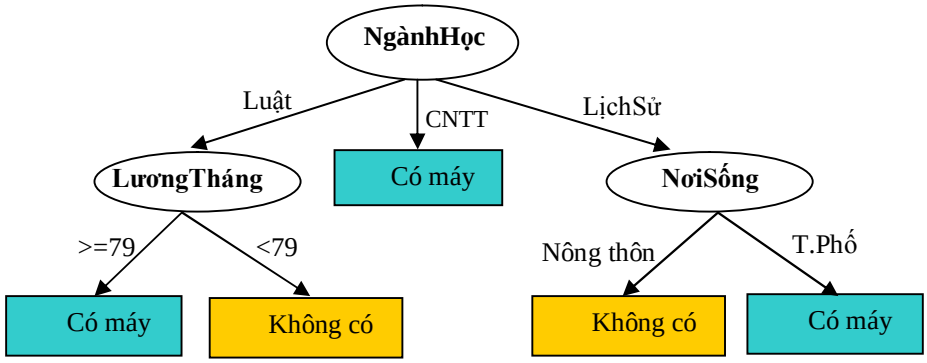
NơiSống	NgànhHọc	KinhTếGĐ	LươngTháng	MáyTính
T.Phố	Luật	Chưa tốt	45	Không
T.Phố	CNTT	Chưa tốt	52	Có
T.Phố	LịchSử	Trung bình	34	Có
T.Phố	Luật	Trung bình	35	Không
T.Phố	Luật	Khá	100	Có
T.Phố	LịchSử	Trung bình	50	Có
T.Phố	CNTT	Chưa tốt	55	Có
NôngThôn	LịchSử	Trung bình	50	Không

Như vậy, quá trình loại các mẫu "lỗi" đã làm cho thuộc tính *LươngTháng* trong mẫu 4 không đủ hữu hiệu để xất hiện trên cây kết quả. Đối sánh với cây được tạo ra khi tập mẫu huấn luyện thuần nhất ở bảng 1 là cây ở hình 1 ta nhận thấy cây thu được ở hình 5 không phản ánh hết thực tế. Điều này là hoàn toàn phù hợp do việc xử lý bằng cách này đã làm mất dữ liệu một cách nghiêm trọng.



Hình 5. Cây quyết định không phản ánh đúng thực tế ứng với tập huấn luyện ở bảng 4.

b. Sau khi ta làm thuần nhất giá trị cho thuộc tính *LươngTháng* theo đại số gia tử $\underline{X}_{LươngTháng}$ đã xây dựng ở trên, ta có tập giá trị của thuộc tính *LươngTháng* mới là: {45, 24, 52, 34, 64, 64, 79, 35,100, 50, 79, 40, 55, 50}. Tiến hành xây dựng cây, thu được cây kết quả ở hình 6. Kết quả này là hữu hiệu, nó cho thấy sự tương đồng với cây ở hình 1 khi tập mẫu là hoàn toàn thuần nhất và với ý kiến chuyên gia tốt nhất.



Hình 6. Cây quyết định được tạo sau khi làm thuần nhất giá trị cho thuộc tính *LươngTháng* ở bảng 3 dựa theo ĐSGT.

4 KẾT LUẬN

Bài báo đã đánh giá tính phức tạp của dữ liệu huấn luyện được chọn từ dữ liệu nghiệp vụ, phân tích và chỉ ra các thông tin bổ ích để có thể chọn được tập mẫu huấn luyện tốt. Trên cơ sở của đại số gia tử, bài báo cũng đã xem xét một cách trọn vẹn việc làm thuần nhất giá trị cho các thuộc tính chưa thuần nhất trong mẫu theo giá trị ngôn ngữ hay theo giá trị thực để từ đó ta có thể xây dựng được cây quyết định phù hợp. Chúng tôi sẽ tiếp tục nghiên cứu phương pháp tạo cây mờ theo giá trị ngôn ngữ và đối sánh trực tiếp với cây được tạo theo giá trị truyền thống trong những báo cáo tiếp theo.

Tài liệu tham khảo

Tài liệu tiếng Việt

1. Dương Thăng Long: Phương pháp xây dựng hệ mờ dạng luật với ngữ nghĩa dựa trên đại số gia tử và ứng dụng trong bài toán phân lớp, Luận án Tiến sĩ Toán học, Viện Công nghệ Thông tin (2010).
2. Đoàn Văn Ban, Lê Mạnh Thạnh, Lê Văn Tường Lân: Một cách chọn mẫu huấn luyện và thuật toán học để xây dựng cây quyết định trong khai phá dữ liệu, Tạp chí Tin học và Điều khiển học, T23, S4 (2007).
3. Nguyễn Cát Hồ: Lý thuyết tập mờ và Công nghệ tính toán mềm, Tuyển tập các bài giảng về Trường thu hệ mờ và ứng dụng (2006).
4. 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).
5. Nguyễn Công Hào, Nguyễn Cát Hồ: Một cách tiếp cận xấp xỉ dữ liệu trong cơ sở dữ liệu mờ, Tạp chí Tin học và Điều khiển học (2006).

6. Nguyễn Công Hào: Cơ sở dữ liệu mờ với thao tác dữ liệu dựa trên đại số gia tử, Luận án Tiến sĩ Toán học, Viện Công nghệ Thông tin (2008)
7. Lê Văn Tường Lân: Phụ thuộc dữ liệu và tác động của nó đối với bài toán phân lớp của khai phá dữ liệu, Tạp chí khoa học Đại học Huế, Tập:19, Số:53 (2009).

Tài liệu tiếng Anh

8. A.K. Bikas, E. M. Voumvoulakis and N. D. Hatziargyriou: Neuro-Fuzzy Decision Trees for Dynamic Security Control of Power Systems, Department of Electrical and Computer Engineering, NTUA,Athens, Greece (2008)
9. Chida, A: Enhanced Encoding with Improved Fuzzy Decision Tree Testing Using CASP Templates, Computational Intelligence Magazine, IEEE (2012).
10. Chang, Robin L. P. Pavlidis: Theodosios, Fuzzy Decision Tree Algorithms, Man and Cybernetics, IEEE (2007).
11. Dorian, P.: Data Preparation for Data Mining, Morgan Kaufmann (1999).
12. Daveedu Raju Adidela, Jaya Suma. G, Lavanya Devi. G: Construction of Fuzzy Decision Tree using Expectation Maximization Algorithm, International Journal of Computer Science and Management Research (2012).
13. Fernandez A., Calderon M., Barrenechea E.: Enhancing Fuzzy Rule Based Systems in Multi-Classication Using Pairwise Coupling with Preference Relations, EUROFUSE Workshop Preference Modelling and Decision Analysis, Public University of Navarra, Pamplona, Spain (2009).
14. FA. Chao Li, Juan sun, Xi-Zhao Wang: Analysis on the fuzzy filter in fuzzy decision trees, Proceedings of the Second International Conference on Machine Learxung and Cybernetics (2003).
15. Kavita Sachdeva, Madasu Hanmandlu, Amioy Kumar: Real Life Applications of Fuzzy Decision Tree, International Journal of Computer Applications (2012).
16. 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. (2010).
17. Ho Tu Bao: Introduction to knowledge discovery and data mining, Institute of Information Technology National Center for Natural Science (2000).
18. Ho N. C. and Nam H. V.: An algebraic approach to linguistic hedges in Zadeh's fuzzy logic, Fuzzy Sets and Systems, vol.129, pp.229-254 (2002).
19. Moustakidis, S. Mallinis, G. ; Koutsias, N. ; Theocharis, J.B. ; Petridis, V. : SVM-Based Fuzzy Decision Trees for Classification of High Spatial Resolution Remote Sensing Images, Geoscience and Remote Sensing, IEEE (2012).
20. Oleksandr Dorokhov, Vladimir Chernov: Application of the fuzzy decision trees for the tasks of alternative choices, Transport and Telecommunication Institute, Lomonosova, Latvia , Vol.12, No 2 (2011).