MỘT TIÊU CHUẨN MỚI CHỌN NÚT XÂY DỰNG CÂY QUYẾT ĐỊNH

NGUYỄN THANH TÙNG

1. MỞ ĐẦU

Cho tập mẫu huấn luyện S gồm n đối tượng. Mỗi đối tượng x được mô tả bằng một véc tơ

$$x = \left(c_1(x), c_2(x), ..., c_p(x), d_{p+1}(x)\right),$$

trong đó $c_k(x)$ là giá trị của thuộc tính điều kiện c_k tại đối tượng x, $k = 1, 2, ..., p$; $d_{p+1}(x)$ là giá trị thuộc tính quyết định (nhãn 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 S.

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 [10].

Cây quyết định là một cấu trúc cây, biểu diễn một vấn đề quyết định. Mỗi nút trong (không phải nút lá) gắn với một thuộc tính điều kiện, mỗi nhánh từ nút trong gắn với một giá trị (hay một tập các giá trị) của thuộc tính điều kiện tương ứng, mỗi nút lá gắn với một giá trị thuộc tính quyết định (thuộc tính đích). 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. Quá trình xây dựng một cây quyết định cụ thể bắt đầu bằng một cây rỗng, toàn bộ tập mẫu huấn luyện và là như sau [8]:

1. Nếu tại nút hiện thời, tất cả các đối tượng huấn luyện đều thuộc vào một lớp nào đó thì cho nút này thành nút lá có tên là nhãn lớp chung của các đối tượng.
2. Trường hợp ngược lại, sử dụng một độ đo, chọn thuộc tính điều kiện phân chia tốt nhất tập mẫu huấn luyện có tại nút.
3. Tạo một lượng nút con của của nút hiện thời bằng số các giá trị khác nhau của thuộc tính được chọn. Gán cho mỗi nhánh từ nút cha đến nút con một giá trị của thuộc tính rồi phân chia các các đối tượng huấn luyện vào các nút con tương ứng.
4. Nút con t được gọi là thuần nhất, trở thành lá, nếu tất cả các đối tượng mẫu tại đó đều thuộc vào cùng một lớp. Lặp lại các bước 1-3 đối với mỗi nút chưa thuần nhất.

Trong bước 3, tiêu chuẩn sử dụng lựa chọn thuộc tính được hiểu là một số đo độ phù hợp, một số đo đánh giá độ thuần nhất, hay một quy tắc phân chia tập mẫu huấn luyện.

Vấn đề then chốt trong quá trình xây dựng cây quyết định là việc lựa chọn thuộc tính điều kiện kiểm tra tại mỗi nút (gọi tắt là chọn nút). Có nhiều phương pháp chọn nút dựa trên những tiêu chuẩn khác nhau đánh giá độ quan trọng của các thuộc tính. Hai tiêu chuẩn thường được sử dụng nhất là:

- Lượng thông tin thu thêm (Information Gain, thuật toán ID3 và C4.5 của Quinlan [8, 9, 12]).

- Độ phụ thuộc của thuộc tính quyết định vào thuộc tính điều kiện theo nghĩa lí thuyết tập thô của Pawlak [1, 2, 5].

Trong báo cáo này, dựa trên ý tưởng của lí thuyết tập thô, chúng tôi đưa ra một số đo mới đánh giá độ phụ thuộc của thuộc tính quyết định vào thuộc tính điều kiện. Số đo này được sử dụng làm tiêu chuẩn chọn nút trong quá trình phát triển cây. Kết quả tính toán thực nghiệm cho thấy thấy cây quyết định xây dựng được bằng cách sử dụng tiêu chuẩn mới này có kích thước nhỏ hơn kích thước của các cây sử dụng entropy hoặc độ phụ thuộc theo lí thuyết tập thô; độ phức tạp tính toán nhỏ hơn, các luật thu được gọn hơn, chính xác hơn.

2. MỘT SỐ KHÁI NIỆM CỦA LÍ THUYẾT TẬP THÔ

2.1. Hệ thống thông tin

Hệ thống thông tin là công cụ biểu diễn tri thức dưới dạng một bảng dữ liệu gồm p cột ứng với p thuộc tính và n hàng ứng với n đối tượng.

Định nghĩa 2.1.1. Hệ thống thông tin là một bộ tứ $S=(U,A,V,f)$ trong đó U là tập khác rỗng, hữu hạn các đối tượng; A là tập khác rỗng, hữu hạn các thuộc tính; $V=\prod_{a\in A} V_a$ với V_a là tập giá trị của thuộc tính $a\in A$; f là hàm thông tin, với mọi $a\in A$ và $x_i\in U$ hàm f cho giá trị $f(x_i,a)\in V_a$.

Dưới đây, giả sử tập các đối tượng U gồm n phần tử: $U=\{x_1,x_2,...,x_n\}$.

Xét hệ thống thông tin $S=(U,A,V,f)$. Mỗi tập con P của tập thuộc tính A xác định một quan hệ tương đương:

$$IND(P)=\left\{(x_i,x_j)\in U\times U \mid \forall a\in P, f(x_i,a)=f(x_j,a)\right\}.$$

Ký hiệu phân hoạch của U sinh bởi quan hệ $IND(P)$ là U/P và lớp tương đương chứa đối tượng x_i là $[x_i]_P$,

$$[x_i]_P=\left\{x_j \mid x_j\in U,(x_i,x_j)\in IND(P)\right\}.$$

Định nghĩa 2.1.2. Cho hệ thống thông tin $S=(U,A,V,f)$, P và Q là hai tập con của tập thuộc tính A. Ta nói:

1) $U/P=U/Q$ khi và chỉ khi $\forall x_i\in U,[x_i]_P=[x_i]_Q$;

2) $U/P\subseteq U/Q$ khi và chỉ khi $\forall x_i\in U,[x_i]_P\subseteq[x_i]_Q$;

3) $U/P\subset U/Q$ khi và chỉ khi $\forall x_i\in U,[x_i]_P\subseteq[x_i]_Q$ và tồn tại x_j sao cho $[x_j]_P\subset[x_j]_Q$.

Tính chất 2.1.1. ([6,7]) Xét hệ thống thông tin $S=(U,A,V,f)$ và $P,Q\subseteq A$. Nếu $P\subseteq Q$ thì $U/Q\subseteq U/P$.

Tính chất 2.1.2. ([6,7]) Xét hệ thống thông tin $S=(U,A,V,f)$ và $P,Q\subseteq A$. Với mọi $x_i\in U$ có:

$$[x_i]_{P\cup Q}=[x_i]_P \text{ I } [x_i]_Q \ .$$

Định nghĩa 2.1.3. Cho hệ thống thông tin $S=(U,A,V,f)$, $P\subseteq A$ và $X\subseteq U$. Khi đó các tập

$$\underline{P}X=\{x\in U \mid [x]_P\subseteq X\} \text{ và } \overline{P}X=\{x\in U \mid [x]_P \text{ I } X\neq\varnothing\}$$

lần lượt được gọi là P-xấp xỉ dưới và P-xấp xỉ trên của X trong S.

2.2. Bảng quyết định

Định nghĩa 2.2.1. Bảng quyết định là một dạng đặc biệt của hệ thống thông tin, trong đó tập các thuộc tính A bao gồm hai tập con rời nhau: tập các thuộc tính điều kiện C và tập các thuộc tính quyết định D. Như vậy, bảng quyết định là một hệ thống thông tin $DT=(U,C\cup D,V,f)$. trong đó $C\cap D=\varnothing$.

Không mất tính tổng quát có thể giả thiết D chỉ gồm một thuộc tính quyết định duy nhất d, (trường hợp có nhiều thuộc tính thì bằng một phép mã hoá luôn có thể quy về một thuộc tính). Như vậy, mỗi đối tượng x trong bảng quyết định được mô tả bằng một véc tơ

$$(c_1(x),c_2(x),...,c_p(x),d(x)).$$

Định nghĩa 2.2.2. Cho bảng quyết định $DT=(U,C\cup d,V,f)$. Ta gọi tập

$$POS_C(d)=\bigcup_{Y\in U/d}\underline{C}Y$$

là miền C-khẳng định của d.

Dễ thấy $POS_C(d)$ là tập các đối tượng được phân lớp đúng (như d) trong U nếu sử dụng tập các thuộc tính điều kiện C .

Định nghĩa 2.2.3. Xét bảng quyết định $DT=(U,C\cup d,V,f)$ và hai đối tượng $x,y\in U$. Ta nói x và y mâu thuẫn nhau trong DT nếu

$$C(x)=C(y) \text{ nhưng } d(x)\neq d(y).$$

Đối tượng x được gọi là nhất quán trong DT nếu không tồn tại một đối tượng y khác mâu thuẫn với x. DT được gọi là nhất quán nếu mọi đối tượng trong $x\in U$ đều là nhất quán.

Mệnh đề 2.1. ([6]) Xét bảng quyết định $DT=(U,C\cup d,V,f)$. Ta có

$$POS_C(d)=\{\{x\in U \mid x \text{ là đối tượng nhất quán}\}.$$

Hơn nữa, nếu DT là nhất quán thì $POS_C(d)=U$.

3. CÁC TIÊU CHUẨN CHỌN NÚT DỰA VÀO ENTROPY VÀ LÍ THUYẾT TẬP THÔ

3.1. Tiêu chuẩn dựa vào entropy

Xét bảng quyết định $DT = (U, C \cup d, V, f)$, số giá trị (nhãn lớp) có thể của d là k. Khi đó Entropy của tập các đối tượng trong DT được định nghĩa bởi:

$$entropy(DT) = -\sum_{i=1}^{k} p_i \log_2 p_i \quad (1)$$

trong đó p_i là tỉ lệ các đối tượng trong DT mang nhãn lớp i.

Lượng thông tin thu thêm (IG) là lượng entropy còn lại khi tập các đối tượng trong DT được phân hoạch theo một thuộc tính điều kiện c nào đó. IG xác định theo công thức sau:

$$IG(DT, c) = Entropy(DT) - \sum_{n \in values(c)} \frac{|DT_n|}{|DT|} Entropy(DT_n) \quad (2)$$

trong đó $values(c)$ là tập các giá trị của thuộc tính c, DT_n là tập các đối tượng trong DT có giá trị thuộc tính c bằng n. $IG(S, A)$ được J. R. Quinlan ([8]) sử dụng làm độ đo lựa chọn thuộc tính phân chia dữ liệu tại mỗi nút trong thuật toán xây dựng cây quyết định ID3. Thuộc tính được chọn là thuộc tính cho lượng thông tin thu thêm lớn nhất.

Nhược điểm của IG là, khi lựa chọn thuộc tính, nó thiên vị các đặc trưng có nhiều giá trị. Để khắc phục nhược điểm này, trong thuật toán cải tiến C4.5 của mình, J. R. Quinlan ([9]) đã sử dụng một độ đo mới, gọi là tỉ số thông tin thu thêm (Gain Ratio - GR). tỉ số thông tin thu thêm được từ lượng thông tin thu thêm bằng cách thêm vào IG một thành phần mới, đó là thông tin phân chia (Split Information). Thông tin phân chia của tập các đối tượng trong DT, khi được phân hoạch theo l giá trị của thuộc tính c, là đại lượng $Split(DT, c)$ xác định theo công thức sau:

$$Split(DT, c) = -\sum_{i=1}^{l} \frac{|DT_i|}{|DT|} \log_2 \frac{|DT_i|}{|DT|}$$

trong đó, DT_i, $i = 1, ..., l$ là các lớp đối tượng có giá trị thuộc tính c bằng i.

Với $Split(DT, c)$ xác định như trên, tỉ số thu thêm (GR - Gain Ratio) định nghĩa bởi công thức:

$$GR(DT, c) = \frac{IG(DT, c)}{Split(DT, c)}.$$

3.2. Tiêu chuẩn dựa vào độ phụ thuộc theo lí thuyết tập thô

Xét bảng quyết định $DT = (U, C \cup d, V, f)$ và tập con thuộc tính điều kiện $P \subseteq C$. Giả sử $U / d = \{Y_1, Y_2, ..., Y_m\}$, $U / P = \{X_1, X_2, ..., X_n\}$. Đặt

$$\gamma(d / P) = \frac{|POS_P(d)|}{|U|} = \frac{\left| \bigcup_{i=1}^{m} \underline{P} Y_i \right|}{|U|}.$$

$\gamma(d / P)$ được gọi là độ phụ thuộc của d vào P.

$\gamma(d / P)$ có các tính chất sau [1, 6]:

- $0 \le \gamma(d / P) \le 1$.
- Nếu $\gamma(d / P) = 1$ thì có phụ thuộc hàm $P \to d$
- Nếu $0 < \gamma(d / P) < 1$ thì d phụ thuộc một phần vào P
- Nếu $\gamma(d / P) = 0$ thì không có đối tượng nào của U có thể được phân lớp đúng (như d) dựa vào tập thuộc tính P.

Theo cách tiếp cận tập thô, $\gamma(d / c)$ được sử dụng làm tiêu chuẩn lựa chọn thuộc tính kiểm tra tại mỗi nút trong quá trình phát triển cây quyết định: Thuộc tính được chọn là thuộc tính c cho giá trị $\gamma(d / c)$ lớn nhất trong số các thuộc tính còn lại tại mỗi bước ([1,2,5]).

4. SỐ ĐO MỚI VỀ ĐỘ PHỤ THUỘC

Định nghĩa 4.1. Xét bảng quyết định $DT = (U, C \cup d, V, f)$ và tập con thuộc tính điều kiện $P \subseteq C$. Giả sử $U / d = \{Y_1, Y_2, ..., Y_m\}$, $U / P = \{X_1, X_2, ..., X_n\}$. Đặt

$$\beta(d / P) = 1 - \frac{m}{m-1} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|Y_i \cap X_j|}{|U|} \cdot \frac{|Y_i^C \cap X_j|}{|U|} .$$

Ta gọi $\beta(d / P)$ là độ phụ thuộc của thuộc tính quyết định d vào tập thuộc tính điều kiện P.

Bổ đề 4.1. Cho bảng quyết định $DT = (U, C \cup d, V, f)$ và tập con thuộc tính điều kiện $P \subseteq C$. Giả sử $U / d = \{Y_1, Y_2, ..., Y_m\}$, $U / P = \{X_1, X_2, ..., X_n\}$. Khi đó

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|Y_i \cap X_j|}{|U|} \cdot \frac{|Y_i^C \cap X_j|}{|U|} \ge 0$$

Dấu "=" xảy ra khi và chỉ khi $U / P \subseteq U / d$.

Chứng minh. Hiển nhiên $\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|Y_i \cap X_j|}{|U|} \cdot \frac{|Y_i^C \cap X_j|}{|U|} \ge 0$. Chỉ cần chứng minh dấu "=" xảy ra khi và chỉ khi $U / P \subseteq U / d$.

a) ($\Leftarrow$) Giả sử $U / P \subseteq U / d$, khi đó với mỗi $X_j \in U / P$ tồn tại $Y_k \in U / d$ sao cho $X_j \subseteq Y_k$. Suy ra $|Y_k^C \cap X_j| = 0$ và $|Y_i \cap X_j| = 0$ với mọi $i \ne k$. như vậy, trong mọi trường hợp ta đều có $|Y_i \cap X_j||Y_i^C \cap X_j| = 0$ với mọi $i = 1, 2, ..., m$ và $j = 1, 2, ..., n$. Do đó

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|Y_i \cap X_j|}{|U|} \cdot \frac{|Y_i^C \cap X_j|}{|U|} = 0 .$$

b) ($\Rightarrow$) Giả sử có $\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \cap X_j|}{|U|}\cdot\frac{|Y_i^C \cap X_j|}{|U|}=0$. Suy ra $|Y_i \cap X_j|\,|Y_i^C \cap X_j|=0$ với mọi $i=1,2,...,m$ và $j=1,2,...,n$.

Giả sử tồn tại X_k không phải là tập con của bất kỳ Y_i nào ($i=1,2,...,m$). Vì $\bigcup_{i=1}^{m} Y_i = U$, phải tồn tại l sao cho $Y_l \cap X_k \neq \varnothing$ và $Y_l^C \cap X_k \neq \varnothing$. Suy ra $|Y_l \cap X_k|\,|Y_l^C \cap X_k| \neq 0$. Điều này mâu thuẫn với $|Y_i \cap X_j|\,|Y_i^C \cap X_j|=0$ với mọi $i=1,2,...,m$ và $j=1,2,...,n$. Vậy với mọi $X_j \in U/P$ phải tồn tại $Y_i \in U/d$ thỏa mãn $X_j \subseteq Y_i$, tức $U/P \subseteq U/d$. ■

Bổ đề 4.2. Cho bảng quyết định $DT=(U, C\cup d, V, f)$ và tập con thuộc tính điều kiện $P\subseteq C$. Giả sử $U/d=\{Y_1, Y_2, ..., Y_m\}$ với $m>1$ và $U/P=\{X_1, X_2, ..., X_n\}$ với $n\geq 1$. Khi đó

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \cap X_j|}{|U|}\cdot\frac{|Y_i^C \cap X_j|}{|U|} \leq 1-\frac{1}{m.n}.$$

Dấu "=" xảy ra khi và chỉ khi $n=1$ và $\frac{|Y_1|}{|U|}=\frac{|Y_2|}{|U|}=...=\frac{|Y_m|}{|U|}=\frac{|U|}{m}$.

Chứng minh. Dễ thấy $|Y_i^C \cap X_j| = |X_j| - |Y_i \cap X_j|$. Do đó

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \cap X_j|}{|U|}\cdot\frac{|Y_i^C \cap X_j|}{|U|} = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \cap X_j|}{|U|}\cdot\left(\frac{|X_j|}{|U|}-\frac{|Y_i \cap X_j|}{|U|}\right)$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|X_j|}{|U|}\cdot\frac{|Y_i \cap X_j|}{|U|}-\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \cap X_j|^2}{|U|^2}=\sum_{j=1}^{n}\frac{|X_j|}{|U|}\sum_{i=1}^{m}\frac{|Y_i \cap X_j|}{|U|}-\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \cap X_j|^2}{|U|^2}$$

$$=\sum_{j=1}^{n}\frac{|X_j|^2}{|U|^2}-\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \cap X_j|^2}{|U|^2} \qquad (1)$$

Để ý rằng $U/P=\{X_1, X_2, ..., X_n\}$ là phân hoạch của U, nên

$$\frac{|X_j|}{|U|}>0 \text{ với mọi } j=1, 2, \ldots, n \text{ và } \sum_{j=1}^{n}\frac{|X_j|}{|U|}=1.$$

Suy ra $$\sum_{j=1}^{n}\frac{|X_j|^2}{|U|^2} \leq \left(\sum_{j=1}^{n}\frac{|X_j|}{|U|}\right)^2 = 1. \qquad (2)$$

Đẳng thức $\sum_{j=1}^{n}\frac{|X_j|^2}{|U|^2}=1$ xảy ra khi và chỉ khi $n=1$.

Lại có

$$-\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \cap X_j|^2}{|U|^2} \le -\frac{1}{m.n}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \cap X_j|}{|U|}\right)^2 = -\frac{1}{m.n} \quad (3)$$

Dấu "=" xảy ra khi và chỉ khi

$$\frac{|Y_1 \cap X_1|}{|U|} = ... = \frac{|Y_1 \cap X_n|}{|U|} = \frac{|Y_2 \cap X_1|}{|U|} = ... = \frac{|Y_2 \cap X_n|}{|U|} = ... = \frac{|Y_m \cap X_1|}{|U|} = ... = \frac{|Y_m \cap X_n|}{|U|}.$$

Từ (1), (2) và (3) suy ra

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \cap X_j|}{|U|}\cdot\frac{|Y_i^C \cap X_j|}{|U|} \le 1-\frac{1}{m.n}.$$

Dấu "=" xảy ra khi và chỉ khi $n = 1$ và $\frac{|Y_1|}{|U|} = \frac{|Y_2|}{|U|} = ... = \frac{|Y_m|}{|U|}$. ■

Từ Bổ đề 1. và 2. ta có kết quả sau.

Định lí 4.1. Cho bảng quyết định $DT = (U, C \cup d, V, f)$ và tập con P của tập thuộc tính điều kiện C. Giả sử $U/d = \{Y_1, Y_2, ..., Y_m\}$. Khi đó

a) $0 \le \beta(d/P) \le 1$.

b) $\beta(d/P) = 1$ khi và chỉ khi $U/P \subseteq U/d$, (tức là có phụ thuộc hàm $P \to d$).

c) $\beta(d/P) = 0$ khi và chỉ khi $n = 1$ và $\frac{|Y_1|}{|U|} = \frac{|Y_2|}{|U|} = ... = \frac{|Y_m|}{|U|}$.

Bổ đề 4.3. Xét bảng quyết định $DT = (U, C \cup d, V, f)$ và hai tập con khác rỗng các đối tượng $X, Y \subseteq U$. Giả sử $X = \bigcup_{j=1}^{k} X_j$, $X_p \cap X_q = \varnothing$ với mọi $p \ne q$. (tức $\{X_1, X_2, ..., X_k\}$ là một phân hoạch của X). Khi đó

$$\frac{|Y \cap X|}{|U|}\cdot\frac{|Y^C \cap X|}{|U|} \ge \sum_{j=1}^{k}\frac{|Y \cap X_j|}{|U|}\cdot\frac{|Y^C \cap X_j|}{|U|}.$$

Dấu "=" xảy ra khi $|Y \cap X_p|\cdot|Y^C \cap X_q| = 0$ với mọi $p \ne q$ và $p, q = 1, 2, ..., k$.

Chứng minh. Do $X_p \cap X_q = \varnothing$ với mọi $p \ne q$, ta có

$$\frac{|Y \cap X|}{|U|}\cdot\frac{|Y^C \cap X|}{|U|} = \frac{\left|Y \cap \left(\bigcup_{j=1}^{k} X_j\right)\right|}{|U|}\cdot\frac{\left|Y^C \cap \left(\bigcup_{p=1}^{k} X_p\right)\right|}{|U|}$$

$$= \frac{\left|\bigcup_{j=1}^{k}(Y \text{ I } X_j)\right|}{|U|} \cdot \frac{\left|\bigcup_{p=1}^{k}(Y^C \text{ I } X_p)\right|}{|U|} = \frac{\sum_{j=1}^{k}|Y \text{ I } X_j|}{|U|} \cdot \frac{\sum_{p=1}^{k}|Y^C \text{ I } X_p|}{|U|}$$

$$= \sum_{j=1}^{k}\sum_{p=1}^{k}\frac{|Y \text{ I } X_j|}{|U|} \times \frac{|Y^C \text{ I } X_p|}{|U|} \geq \sum_{j=1}^{k}\frac{|Y \text{ I } X_j|}{|U|} \times \frac{|Y^C \text{ I } X_j|}{|U|} .$$

Dấu "=" xảy ra khi $|Y \text{ I } X_p| \times |Y^C \text{ I } X_q| = 0$ với mọi $p \neq q$ và $p, q = 1, 2, ..., k$. ■

Định lí 4.2. Cho bảng quyết định $DT = (U, C \cup d, V, f)$ và hai tập con $P, Q \subseteq C$. Nếu $P \subset Q$. Khi đó

$$\beta(d / P) \leq \beta(d / Q).$$

Chứng minh. Do $P \subset Q$ nên $U / Q \subseteq U / P$; mỗi lớp của phân hoạch U / P sẽ là một hoặc hợp của một số lớp thuộc phân hoạch U / Q. Giả sử $U / d = \{Y_1, Y_2, ..., Y_m\}$, $U / P = \{X_1, X_2, ..., X_n\}$ và $U / Q = \{Z_1, Z_2, ..., Z_s\}$, trong đó

$$X_1 = \bigcup_{l=1}^{k_1} Z_l , \quad X_2 = \bigcup_{l=k_1+1}^{k_2} Z_l , \ \ldots , \quad X_n = \bigcup_{l=k_{n-1}+1}^{s} Z_l .$$

Theo bổ đề 4.3. với mỗi $i = 1, \ldots, m$ và $j = 1, \ldots, n$ ta có:

$$\frac{|Y_i \text{ I } X_j|}{|U|} \cdot \frac{|Y_i^C \text{ I } X_j|}{|U|} \geq \sum_{l=k_{j-1}+1}^{k_j} \frac{|Y_i \text{ I } Z_l|}{|U|} \cdot \frac{|Y_i^C \text{ I } Z_l|}{|U|} \quad (\text{đặt } k_0 = 0).$$

Suy ra

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \text{ I } X_j|}{|U|} \cdot \frac{|Y_i^C \text{ I } X_j|}{|U|} \geq \sum_{i=1}^{m}\sum_{l=1}^{t}\frac{|Y_i \text{ I } Z_l|}{|U|} \cdot \frac{|Y_i^C \text{ I } Z_l|}{|U|} ,$$

$$1 - \frac{m}{m-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_i \text{ I } X_j|}{|U|} \cdot \frac{|Y_i^C \text{ I } X_j|}{|U|} \leq 1 - \frac{m}{m-1}\sum_{i=1}^{m}\sum_{l=1}^{t}\frac{|Y_i \text{ I } Z_l|}{|U|} \cdot \frac{|Y_i^C \text{ I } Z_l|}{|U|} ,$$

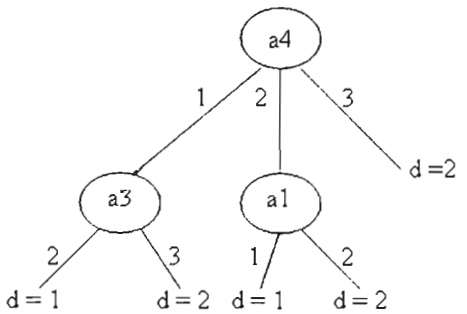
Tức là $\beta(d / P) \leq \beta(d / Q)$. ■

Các kết quả lí thuyết trên đây cho phép lấy $\beta(d / c)$ làm số đo đánh giá mức độ quan trọng của mỗi thuộc tính điều kiện đối với việc phân lớp các đối tượng. Từ đó có thể sử dụng $\beta(d / c)$ làm tiêu chuẩn lựa chọn thuộc tính kiểm tra tại mỗi nút trong quá trình phát triển cây quyết định: Thuộc tính được chọn là thuộc tính c sao cho $\beta(d / c)$ đạt giá trị lớn nhất trong số các thuộc tính còn lại tại mỗi bước.

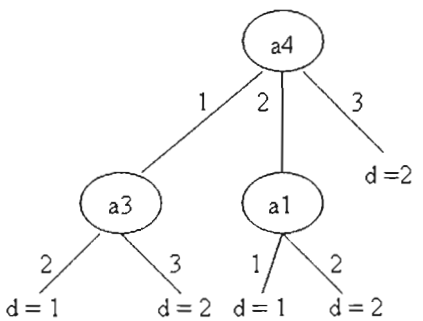
5. VÍ DỤ

Xét bảng quyết định DT sau đây.

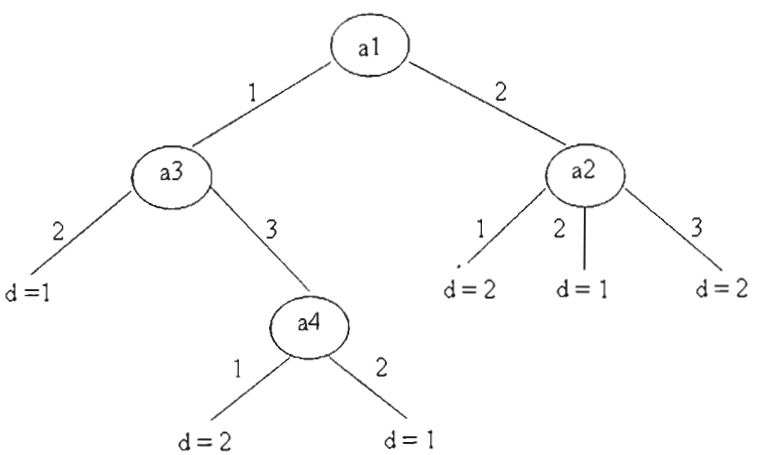
U	a1	a2	a3	a4	d	U	a1	a2	a3	a4	d
1	1	2	2	1	1	7	1	2	3	1	2
2	1	2	3	2	1	8	2	3	1	2	2
3	1	2	2	2	1	9	1	2	2	2	1
4	2	2	2	1	1	10	1	1	3	2	1
5	2	3	2	3	2	11	2	1	2	3	2
6	1	3	2	1	1	12	1	1	2	2	1



Hình 4.1. Cây quyết định sử dụng tiêu chuẩn $\beta(d/c)$: 8 nút, 5 luật.



Hình 4.2. Cây quyết định sử dụng tiêu chuẩn $\gamma(d/c)$ của lí thuyết tập thô: 8 nút, 5 luật



Hình 4.3. Cây quyết định sử dụng tiêu chuẩn $Gain(c,d)$ của lí thuyết thông tin: 10 nút, 6 luật

6. TÍNH TOÁN THỬ NGHIỆM VÀ ĐÁNH GIÁ

Để đánh giá độ hiệu quả của việc sử dụng $\beta(d/c)$ làm tiêu chuẩn chọn nút xây dựng cây quyết định, chúng tôi đã tiến hành tính toán thử nghiệm, so sánh kết quả thu được với các kết quả sử dụng tiêu chuẩn $\beta(d/c)$ và $\gamma(d/c)$. Các CSDL dùng để thử nghiệm là một số CSDL nhỏ lấy từ các tài liệu tham khảo và 3 CSDL lớn là Laborneg, Monk1, Monk2 lấy từ *UCI Repository of Machine Learning Databases* [4]. Chương trình nguồn C4.5 download từ [15]. Hai chương trình sử dụng số đo $\beta(d/c)$ và $\gamma(d/c)$ được xây dựng từ C4.5 bằng cách thay các lệnh tính *Gain(d,c)* bằng các lệnh tính $\beta(d/c)$ và $\gamma(d/c)$. Các tính toán được thực hiện trên máy PC Pentium 4, CPU 2.4Ghz, bộ nhớ 256MB.

Kết quả thử nghiệm cho thấy:

- Về thời gian tính toán, hai tiêu chuẩn $\beta(d/c)$ và *Gain(d,c)* là như nhau, tiêu chuẩn $\gamma(d/c)$ tiêu tốn nhiều hơn.
- Về kích thước, hầu hết các cây quyết định thu được sử dụng tiêu chuẩn $\beta(d/c)$ nhỏ hơn các cây sử dụng tiêu chuẩn *Gain(d,c)* , nhỏ hơn hoặc bằng các cây sử dụng tiêu chuẩn $\gamma(d/c)$.
- Do kích thước cây nhỏ hơn, các luật thu được từ cây ra sử dụng tiêu chuẩn $\beta(d/c)$ có số lượng và cấu trúc gọn hơn, chính xác hơn.

TÀI LIỆU THAM KHẢO

1. Jin-Mao Wei - Rough Set based Approach to Selection of Node, International Journal of Computational Cognition **1** (2) (2003) 25-40, http://www.YangSky.com/yangijcc.html
2. Longjun Huang, Minghe Huang, Bin Guo, Zhiming Zhang - A New Method for Constructing Decision Tree based on Rough Set Theory, Proceedings of the 2007 IEEE International Conference on Granular Computing, 2007, pp. 241-244.
3. Ming Li, Xiao-Feng Zhang - Knowledge Entropy in Rough Set Theory, Proceedings of Third International Conference on Machine Learning and Cybernetics, Shanghai, August 2004, 26-29.
4. Murphy P., Aha W. - UCI Repository of Machine Learning Databases. http://www.ics.ici.edu/~mlearn
5. Ning Yang, Tianrui Li, Jing Song - Construction of Decision Trees based Entropy and Rough Sets under Tolerance Relation.

   www.atlantis-press.com/php/download_paper.php?id=1485
6. Z. Pawlak - Rough sets – Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1991.
7. Z. Pawlak - Rough Set Theory and Its Application to Data Analysis, Cybernetics and Systems: An International Journal **29** (1998) 661-688.
8. R. Quinlan - Induction of Decision Trees, Machine Learning **1** (1) (1986)81-106.
9. J.R. Quinlan - C4.5: Programs for Machine Learning, The Morgan Kaufmann Series in Machine Learning Research **1** (2002) 1-23.

10. Safavian S. R., Landgrebe D. A - Survey of Decision Tree Classifier Methodology. IEEE Transactions on Systems, Man and Cybernetics **21** (3) (1991) 660-674.

Cobweb.ecn.purdue.edu/~landgreb/SMC91.pdf

11. Shannon C.E. - A mathematical theory of communication, Bell System and Technical Journal **27** (1948) 379-423, 623-656.

12. Yao Y.Y. - Information-Theoretic Measures for Knowledge Discovery and Data Mining. Studies in fuzziness and soft computing **119** (2003) 115-136.

13. [13] Yao, Y.Y., Wong, S.K.M. and Butz, C.J. On Information-Theoretic Measures of Attribute Importance, *Proceedings of PAKDD'99*, 133-137. 1999.

14. Ziarko W. - Variable Precision Rough set Model, Journal of Computer and System Science **46** (1993) 39-59.

15. Zhi-Hua Zhou - AI Softwares&Codes, 2004-02.

http://cs.nju.edu.cn/people/zhouzh/zhouzh.files/ai_resource/software.html

SUMMARY

A NEW NODE SELECTION MEASURE IN DECISION TREE GROWING

Classification is one of major tasks in Data Mining. It is to find the rules for assigning objects to one of several predefined categories based on training data set. Many classification techniques have been proposed in the literature, but decision tree is especially popular and efficient. The selection of an attribute used to split the data set at each decision tree node is fundamental to properly classify objects; a good selection will reduce the size of tree and improve the accuracy of classification rules. Different attribute selection measures were proposed in the literature, but two often used are entropy and dependecy measure from rough set theory. In this paper, based on rough set theory also, but we propose an another measure. Experimental computations shown that the decision tree, constructed by using our new measure, have smaller size in general than the trees induced by using entropy and dependency measure; the computation complexity is lower; the classification rules are shorter and more precise.

Địa chỉ:

Viện Công nghệ thông tin Viện KH và CN Việt Nam.

Nhận bài ngày 12 tháng 3 năm 2008