TẠP CHÍ KHOA HỌC, Đại học Huế, Số 53, 2009

PHỤ THUỘC DỮ LIỆU VÀ TÁC ĐỘNG CỦA NÓ ĐỐI VỚI BÀI TOÁN PHÂN LỚP CỦA KHAI PHÁ DỮ LIỆU

Lê Văn Tường Lân
Trường Đại học Khoa học, Đại học Huế

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. Trên cây quyết định, chúng ta dễ dàng tìm ra các luật, những luật này cung cấp thông tin để ra quyết định giải quyết một vấn đề nào đó. Xây dựng một cây quyết định phục vụ khai phá dữ liệu hiệu quả phụ thuộc 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 việc chọn tốt bộ dữ liệu mẫu còn gặp nhiều khó khăn.

Trong bài báo này, chúng tôi tập trung phân tích sự phụ thuộc tự nhiên và sự phụ thuộc theo tương quan hàm số của dữ liệu, nhằm loại bỏ những tính toán dư thừa trong thuật toán học quy nạp và các sự phụ thuộc dữ liệu ở mẫu huấn luyện, tạo dựng cây quyết định có khả năng dự đoán cao, nhằm hỗ trợ ra quyết định trong các bài toán phân tích dữ liệu.

***Từ khoá**: Khai phá dữ liệu, phát hiện tri thức, cây quyết định, mẫu huấn luyện, phụ thuộc hàm, phụ thuộc hàm xấp xỉ, phân lớp dữ liệu.*

I. Đặt vấn đề

Một trong những đích khai phá dữ liệu trong thực tế nhằm đạt đến là mô tả các mẫu dữ liệu, mỗi một sự mô tả là thể hiện những tri thức được khai phá. Sự phân lớp là quá trình nhằm đến một trong những mục đích ấy. Cây quyết định là một trong những giải pháp trực quan và hữu hiệu để mô tả quá trình phân lớp dữ liệu. Do cây quyết định rất hữu dụng nên đã có nhiều nghiên cứu để xây dựng nó mà nổi bật là các thuật toán học quy nạp như CATD, ID3, C45,...[3, 4, 5, 7, 9, 10].

Xây dựng cây quyết định có khả năng dự đoán cao, là một trong những mục tiêu quan trọng của khai phá dữ liệu. Để xây dựng được một cây quyết định có hiệu quả thì ngoài các thuật toán học quy nạp tốt, việc chọn mẫu huấn luyện đóng một vai trò đáng kể. Khi chọn mẫu huấn luyện, sự phụ thuộc tự nhiên giữa các thuộc tính dữ liệu trong mẫu cần phải được đề cập và ứng dụng để loại trừ nó, nhằm nâng cao hiệu quả cho cây được xây dựng [3, 5, 8, 9]. Hơn nữa, có nhiều trường hợp trong thực tế, các nhóm thuộc tính mặc dầu giữa chúng không có sự phụ thuộc theo định nghĩa của phụ thuộc hàm thông thường mà lại phụ thuộc theo kiểu tương quan hàm số nào đó, ta gọi là phụ thuộc hàm xấp xỉ. Các nhóm thuộc tính này làm phức tạp việc xác định 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. Ở đây, chúng ta sẽ xét đến các phụ thuộc dữ liệu loại này nhằm xây dựng cây quyết định có khả năng dự đoán cao.

II. Xây dựng cây quyết định

2.1. Xây dựng cây quyết định

Cho mẫu huấn luyện như ở bảng 1 với thuộc tính quyết định là thuộc tính "MuaÔtô". Chúng ta hãy dự đoán khả năng mua ô tô cho một khách hàng nào đó.

Bảng 1. *Bảng dữ liệu điều tra khách mua ô tô*

Họ và tên	Thành phần GĐ	Công việc	Phụ cấp công việc	Khu vực	Phụ cấp khu vực	Thu nhập	Mua ôtô
Phù Trọng Hưng	Khá	Bác sỹ	80	Thị xã	20	4500	Không
Dương Quang Khai	Trung bình	Bác sỹ	82	Thị xã	20	4000	Không
Trần Trọng Minh Khang	Khá	Giám đốc	110	Thị xã	20	5200	Có
Nguyễn Ngọc Duy Khuê	Khá	Bán hàng	50	T.Phố loại 2	20	2300	Có
Lê Trung Kiên	Khá	Bán hàng	51	T.Phố loại 1	30	5000	Có
Thái Xuân Lãm	Trung bình	Bán hàng	49	T.Phố loại 1	30	6000	Không
Trần Thị Kim Liễu	Trung bình	Giám đốc	110	T.Phố loại 1	30	6500	Có
Đỗ Khánh Long	Khá	Bác sỹ	80	T.Phố loại 2	20	2350	Không
Trần Công Mẫn	Khá	Bác sỹ	81	T.Phố loại 1	30	6000	Có
Võ Quang Mẫn	Khá	Bán hàng	49	T.Phố loại 2	20	5000	Có
Nguyễn Văn Nam	Trung bình	Bác sỹ	83	T.Phố loại 2	20	6000	Có
Trần Thị Hạnh Nguyên	Trung bình	Giám đốc	112	T.Phố loại 2	20	4000	Có
Cao Thọ Ninh	Khá	Giám đốc	108	Thị xã	20	5500	Có
Nguyễn Bảo Phong	Trung bình	Bán hàng	50	T.Phố loại 2	20	5000	Không

Để xây dựng cây quyết định, tại mỗi nút của cây thì các thuật toán đều tính lượng thông tin nhận được trên các thuộc tính và chọn thuộc tính tương ứng có lượng thông tin tối đa làm nút phân tách trên cây - tức là các thuộc tính chia tập mẫu thành các lớp mà mỗi lớp có một phân loại duy nhất hay ít nhất thuộc tính phải có triển vọng đạt được điều này, nhằm để đạt được cây có ít nút nhưng có khả năng dự đoán cao. Như thế, thuộc tính X được chọn phải có có lượng thông tin đạt được tối đa đối với mẫu M trên

thuộc tính quyết định Y, tức là X được chọn phải đạt: Gain(X,Y,M) = max{gain(X_i,Y,M), i = 1,…,n} [5, 8, 10].

Do đối với các thuộc tính riêng biệt X ta phải tính lượng thông tin nhận được cho X tại mỗi giá trị xi nhằm xác định vị trí tốt nhất x* cho việc phân lớp. Giá trị x* được chọn phải có có lượng thông tin đạt được tối đa đối với mẫu M trên thuộc tính quyết định Y, tức là x* được chọn phải đạt: Gain(x*|X,Y,M) = max{gain(x_i|X,Y,M), i = 1,…,n} [8, 10]. . Tại bước lặp đầu tiên ta có:

Lượng thông tin của cây đối với Y trên M là S(Y|M1) = 0,940

Gain(CôngViệc,Y,M1) = 0,246

Gain(ThànhPhầnGĐ,Y,M1) = 0,048

Gain(SốNgườiGĐ,Y,M1) = 0,029

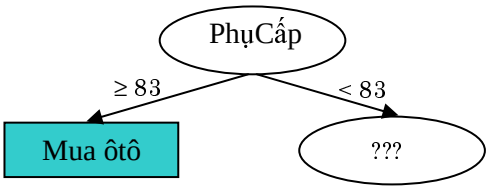
Bảng 2. *Lợi ích của thuộc tính Thu Nhập*

x_i	E(ThuNhậpGĐ)	Gain(ThuNhậpGĐ)
6500	0,8926	0,0477
6000	0,9253	0,0150
5500	0,8950	0,0453
5200	0,8500	0,0903
5000	0,8380	0,1022
4500	0,9152	0,0251
4000	0,9300	0,0103
2350	0,8926	0,0477

Bảng 3. *Lợi ích của thuộc tính Phụ Cấp*

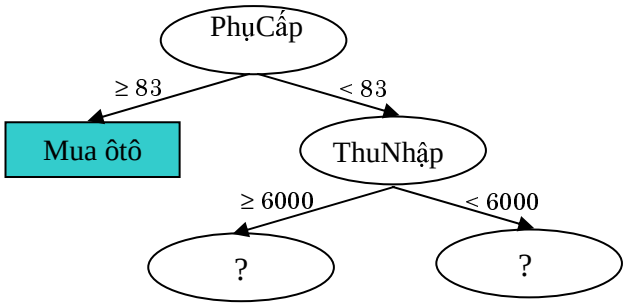
x_i	E(PhụCấp)	Gain(PhụCấp)
112	0,8926	0,0477
110	0,7810	0,1593
108	0,7143	0,2260
83	0,6371	0,3032
82	0,8500	0,0903
81	0,7885	0,1518
80	0,9371	0,0032
51	0,9152	0,0251
50	0,9300	0,0103

Tương tự cho các thuộc tính còn lại, ta tìm được hàm Gain(x_i|PhụCấp,Y,M1) tại giá trị x^* = 83 là lớn nhất nên ta chọn để làm điểm phân tách cây tại bước này. Cây quyết định thu được cho ở hình 2.



Hình 2. *Cây quyết định tại bước1 trên thuộc tính PhụCấp*

Tương tự, cây sau bước lặp thứ 2 được cho ở hình 3.



Hình 3. *Cây quyết định sau bước lặp thứ 2 trên thuộc tính Thu Nhập*

2.2. Ảnh hưởng của phụ thuộc hàm khi xây dựng cây quyết định

Cho mẫu huấn luyện M gồm có m thuộc tính, n bộ. Mỗi thuộc tính X ϵ M có các giá trị là $\{x_1, x_2,...,x_n\}$. Thuộc tính quyết định trong mẫu được đánh dấu là Y còn các thuộc tính còn lại gọi là thuộc tính dự đoán. Với thuộc tính $X = \{x_1, x_2,...,x_n\}$, ta ký hiệu |X| là số các giá trị khác nhau của của tập $\{x_1, x_2,...,x_n\}$ gọi là lực lượng của X; số lần xuất hiện giá trị x_i trong X ký hiệu là $|x_i|$. Giá trị của bộ r trên thuộc tính X được ký hiệu là r|X.

Định nghĩa 1. Với 2 thuộc tính bất kỳ $X_i, X_j \in$ M, ta nói rằng X_i xác định hàm đối với X_j (hay X_j phụ thuộc hàm đối với X_i) nếu với mọi bộ bất kỳ $r_1, r_2 \in$ M mà ta có $r_1|X_i = r_2|X_i$ thì cũng có $r_1|X_j = r_2|X_j$. Ký hiệu $X_i => X_j$.

Mệnh đề 1. Trên mẫu M với thuộc tính quyết định Y, nếu có phụ thuộc hàm $X_1 -> X_2$ và nếu đã chọn X_1 làm nút phân tách trên cây thì mọi nút con của nó sẽ không nhận X_2 làm nút phân tách.

Thật vậy, giả sử $|X_1| = k$, khi chọn X_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 $X = x_i$, i = 1,.,.k. Do $X_1 \rightarrow X_2$ nên tại nhánh bất kỳ thì trên mẫu huấn luyện tương ứng M', X_2 cũng sẽ có cùng 1 giá trị. Như thế $Gain(X_2,Y,M') = 0$ là nhỏ nhất nên X_2 không thể chọn để làm nút phân tách cây.

Mệnh đề 2. Trên mẫu M với thuộc tính quyết định Y, nếu có phụ thuộc hàm $X_1 \rightarrow X_2$ thì lượng thông tin nhận được trên X_1 không nhỏ hơn lượng thông tin nhận được trên X_2.

Thật vậy, giả sử thuộc tính quyết định Y có k giá trị. Do $X_1 \rightarrow X_2$ nên $|X_1| \geq |X_2|$. Theo [8, 10] thì lượng thông tin nhận được trên thuộc tính X là Gain(X,Y,M) được xác định theo công thức (C1).

$$Gain(X,Y,M) = S(Y|M) - \sum_{\forall x_i \in \{X\}} E(X, x_i, Y, M) \qquad (C1)$$

Nếu $|X_1| = |X_2|$ thì trên X_1 hay X2 đều có k phân hoạch như nhau nên $Gain(X_1,Y,M) = Gain(X_2,Y,M)$.

Ngược lại nếu $|X_1| > |X_2|$ tức tồn tại $x_{1i}, x_{1j} \ni X_1$, $x_{1i} \neq x_{1j}$ mà trên tương ứng trên

X_2 thì $x_{2i} = x_{2j}$. Lúc này 2 phân hoạch trên X_1 được gộp thành 1 phân hoạch trên X_2 nên entropy tương ứng trên X_2 lớn hơn. Vậy $Gain(X_1,Y,M) > Gain(X_2,Y,M)$.

Mệnh đề 3. Nếu thuộc tính X là khoá của mẫu M thì loại X ra khỏi M để thu được cây quyết định có khả năng dự đoán tốt hơn.

Thật vậy, giả sử $X = \{x_1, x_2,\ldots..,x_n\}$. Do X là khoá nên ta có $x_i \neq x_j$, $\forall$ i ≠ j. Như thế, mẫu M được phân ra làm n phân hoạch, mà mỗi phân hoạch chỉ có 1 bộ nên hàm $E(X,x_i,Y,M) = 0$, $\forall\ x_i \in X$. Hàm xác định thông tin nhận được trên thuộc tính X $Gain(X,Y,M) = S(Y \mid M) - \sum_{i=1}^{n} \frac{1}{n} E(X, x_i, Y, M)$ = S(Y|M)đạt giá trị cực đại, vì thế chọn X làm điểm phân tách cây. Tại đây, cây được phân chia làm n nút, mỗi cạnh tương ứng được gán nhãn x_i. 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 X ra khỏi M để thu được cây quyết định có khả năng dự đoán tốt hơn.

Hệ quả 1. Nếu có phụ thuộc hàm $X_1 \rightarrow X_2$ mà X_1 không phải là thuộc tính khóa của mẫu M thì thuộc tính X_2 không được chọn làm nút phân tách cây.

Hệ quả này được suy ra từ 3 mệnh đề trên.

III. Phụ thuộc hàm xấp xỉ và ảnh hưởng của nó đến bài toán phân lớp dữ liệu

Như đã nói ở mục 2, sự phụ thuộc hàm giữa các thuộc tính đã được tính đến để làm giảm các chi phí tính toán trong quá trình xây dựng cây. Tuy nhiên, trong một số trường hợp, mặc dầu ta không có được sự phụ thuộc hàm như đã xét nhưng dữ liệu giữa các thuộc tính cũng không thật sự là độc lập với nhau. Ví dụ, nếu nghề nghiệp là ‘bác sỹ’ thì lương ở trong khoảng {1000$ - 1100$}, nếu là ‘Giáo viên’ thì lương lại ở trong khoảng {500 $ - 550 $},… Như vậy, ta phải giải quyết các trường hợp này như thế nào?

Như các nghiên cứu đã đề cập [3, 4, 5], để có thể dự đoán chúng ta xây dựng cây quyết định nhằm phân lớp khả năng mua ô tô của khách hàng. Mẫu huấn luyện trong trường hợp này được chọn là M1 = (ThànhPhầnGĐ, SốNgườiGĐ, CôngViệc, PhụCấp, Lương, ThuNhập, MuaÔtô), trong đó MuaÔtô là thuộc tính quyết định còn lại là các thuộc tính dự đoán và các thuộc tính PhụCấp, Lương, ThuNhập là các thuộc tính có giá trị riêng biệt.

Việc tính $Gain(x^*|X,Y,M)$ của thuộc tính X tại mỗi bước lặp của mỗi nút có độ phức tạp tính toán là $O(n^2)$ nên việc phân lớp tại các thuộc tính PhụCấp, Lương, ThuNhập mất rất nhiều thời gian.

Cây quyết định thu được không cô đọng, xuất hiện nhánh quá ngắn và nhánh quá dài nên không phản ánh ý nghĩa thực tiễn của mô hình [4, 5]. Khảo sát một số thuộc tính trong mẫu, chẳng hạn thuộc tính CôngViệc và PhụCấp, ta thấy mặc dầu giữa chúng không có sự phụ thuộc hàm như đã đề cập ở mục 2 nhưng giá trị của chúng không thật sự là độc lập với nhau. Ví dụ, nếu giá trị của thuộc tính CôngViệc là ‘Bác sỹ’ thì giá trị

của thuộc tính PhụCấp nằm trong miền giá trị {80,81,82,83},… Như thế, vấn đề được giải quyết như thế nào?

Giải quyết vấn đề này, ta thấy ngay có thể thay mẫu M bởi mẫu M' bằng cách chia khoảng giá trị thuộc tính rồi thay các giá trị trong khoảng bằng giá trị trung bình của nó. Cách làm này đơn giản, tuy nhiên, nó làm thay đổi dữ liệu thực tế và có nhiều sai số. Cần để ý rằng, trong thực tế thì rất nhiều trường hợp mặc dù giá trị của chúng là khác nhau nhưng chênh lệch trong một khoảng xác định và sự thay đổi giá trị của dữ liệu có khoảng cách là không đáng kể. Ở đây, chúng ta sẽ tập trung xem xét các trường hợp này.

Định nghĩa 2. Thuộc tính X được gọi là tính được nếu:

i. Các phần tử của X có thể so sánh với nhau theo một quan hệ thứ tự toàn phần nào đó nghĩa là giữa 2 phần tử bất kỳ luôn tìm được giá trị lớn hơn và nhỏ hơn.

ii. $\forall xi, xj \in X$ ta có thể tính được độ lệch giá trị giữa chúng và giá trị này là |xi - xj|.

Định nghĩa 3. Cho thuộc tính $X_i \in M$ là tính được và 2 bộ bất kỳ $r_1, r_2 \in M$. Khoảng cách giá trị giữa 2 bộ r_1, r_2 trên X_i là một giá trị, ký hiệu là $d(r_1|X_i, r_2|X_i)$, được xác định như sau:

$$d(r_1|X_i, r_2|X_i) = \frac{|(r_1|X_i - r_2|X_i)|}{Max(|(r_1|X_i)|, |(r_2|X_i)|)}$$

Khi $Max(|(r_1|X_i)|, |(r_2|X_i)|) = 0$ thì ta quy ước $d(r_1|X_i, r_2|X_i) = 0$. Như thế ta luôn có: $d(r_1|X_i, r_2|X_i) \geq 0$.

Định nghĩa 4. Với 2 thuộc tính bất kỳ $X_i, X_j \in M$ và độ xấp xỉ khoảng cách giá trị chấp nhận cho trước ε, gọi là xấp xỉ giá trị ε, $0 \leq \varepsilon < 1$. Ta nói rằng X_i xác định hàm xấp xỉ ε đối với X_j (hay X_j phụ thuộc hàm xấp xỉ ε đối với X_i) nếu với mọi bộ bất kỳ $r_1, r_2 \in M$ mà ta có $d(r_1|X_i, r_2|X_i) \leq \varepsilon$ thì cũng có $d(r_1|X_j, r_2|X_j) \leq \varepsilon$, ký hiệu $X_i =>_\varepsilon X_j$.

Mệnh đề 4. Với 2 thuộc tính bất kỳ X_i, X_j và 2 độ xấp xỉ $\varepsilon_1, \varepsilon_2$ thoả $0 \leq \varepsilon_1 \leq \varepsilon_2 < 1$. Nếu $X_i =>_{\varepsilon 1} X_j$ thì $X_i =>_{\varepsilon 2} X_j$.

Thật vậy, do $\varepsilon_1 \leq \varepsilon_2$ nên đặt $e = \varepsilon_2 - \varepsilon_1 \geq 0$.

Vì $X_i =>_{\varepsilon 1} X_j$ nên $\forall r_1, r_2 \in X$ ta có $d(r_1|X_i, r_2|X_i) \leq \varepsilon \Rightarrow d(r_1|X_j, r_2|X_j) \leq \varepsilon$ mà $e \geq 0$ nên suy ra $d(r_1|X_i, r_2|X_i) \leq \varepsilon + e \Rightarrow d(r_1|X_j, r_2|X_j) \leq \varepsilon + e$ tức là $X_i =>_{\varepsilon 2} X_j$.

Chọn mẫu M như đã cho trên bảng 1, với độ xấp xỉ $\varepsilon = 0{,}03$ ta có phụ thuộc hàm Lương $\Rightarrow_\varepsilon$ ThuNhập. Với độ xấp xỉ $\varepsilon = 0{,}04$ ta có phụ thuộc hàm xấp xỉ CôngViệc $\Rightarrow_\varepsilon$ PhụCấp, Lương $\Rightarrow_\varepsilon$ ThuNhập.

Định lý 1. Một phụ thuộc hàm đúng trên một quan hệ R thì nó cũng là một phụ thuộc hàm xấp xỉ ε trên quan hệ R, với mọi độ xấp xỉ ε thoả $0 \leq \varepsilon < 1$.

Thật vậy, tính đúng của định lý này được suy ra từ mệnh đề trên với ε1=0 và ε2=ε.

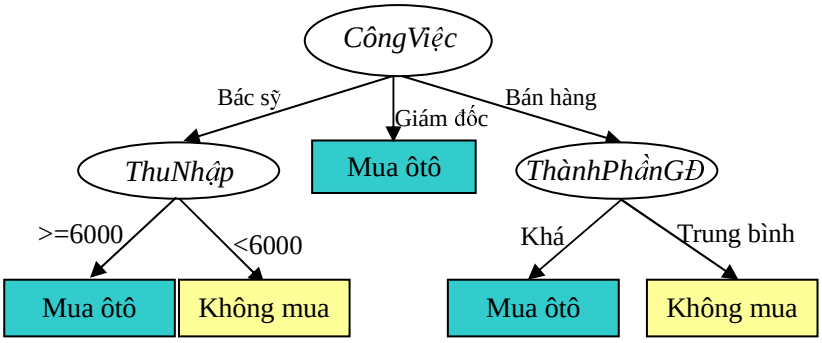
Từ định lý 1 và hệ quả 1, ta suy ra được hệ quả 2 như sau:

Hệ quả 2. Trong mẫu huấn luyện M với độ xấp xỉ giá trị ε. Nếu có phụ thuộc hàm $X_i =>_\varepsilon X_j$ thì:

Nếu X_i không phải là thuộc tính riêng biệt thì thuộc tính X_j trong mẫu M không được chọn làm nút phân tách cây.

Nếu X_i là thuộc tính riêng biệt thì thuộc tính có lực lượng lớn hơn không được chọn làm nút phân tách cây.

Như vậy, cho dữ liệu huấn luyện như bảng 1 với độ xấp xỉ giá trị ε=0,04 thì ta có các phụ thuộc hàm xấp xỉ CôngViệc $\Rightarrow_\varepsilon$ PhụCấp, Lương $\Rightarrow_\varepsilon$ ThuNhập. Theo hệ quả trên thì mẫu phải chọn M2 = (ThànhPhầnGĐ, SốNgườiGĐ, CôngViệc, ThuNhập, MuaÔtô), cây quyết định sau khi học như sau hình 4.



Hình 4. *Cây quyết định của mẫu huấn luyện M2*

IV. So sánh và đánh giá

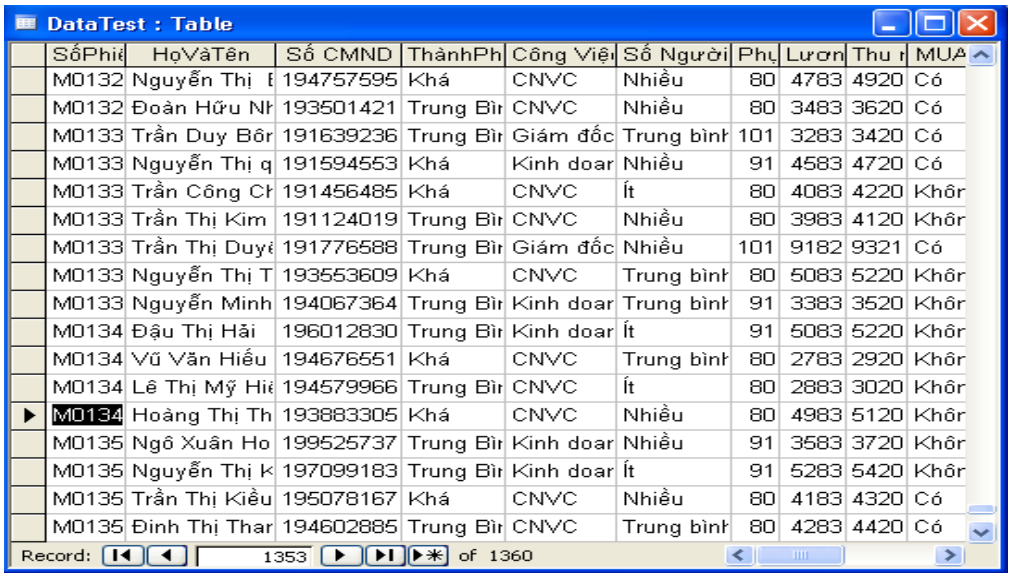
Chúng tôi đã cho huấn luyện trên mẫu gồm 8.492 bản ghi, sau đó kiểm thử trên tập gồm 1.360 bản ghi và tiến hành so sánh thì thu được kết quả như ở bảng 4 và bảng 5.

Bảng 4. *Mẫu huấn luyện và kiểm tra*

DataTraining : Table

SốPhiế	HọVàTên	Số CMND	ThànhPh	Công Việ	Số Người	Phụ	Lươn	Thu n	MUA
M0208	Trần Bình	198875584	Khá	Kinh doar	Trung bình	91	3183	3320	Có
M0301	Lê Bá Linh	191568422	Khá	CNVC	Trung bình	80	3283	3420	Khôn
M0108	Lê Văn Bình	191098234	Trung Bì	CNVC	Nhiều	80	3783	3920	Khôn
M0408	Nguyễn Hoa	196001278	Trung Bì	Giám đốc	Trung bình	101	3783	3920	Có
M0104	Nguyễn Văn .	191003117	Khá	CNVC	Nhiều	80	4283	4420	Khôn
M0508	Trần Quế Chi	193567450	Trung Bì	Kinh doar	Trung bình	91	4783	4920	Khôn
M0204	Trần Thị Hươ	198234309	Khá	Kinh doar	Ít	91	4783	4920	Có
M0403	Lý Thị Hoa	198235457	Khá	Kinh doar	Trung bình	91	4783	4920	Có
M0204	Nguyễn Thị H	196986568	Khá	Giám đốc	Nhiều	101	4983	5120	Có
M0504	Lê Xuân Hoa	196224003	Khá	Giám đốc	Nhiều	101	5283	5420	Có
M0403	Bạch Ân	197543457	Khá	CNVC	Ít	80	5783	5920	Có
M0404	Vũ Quang Bì	196679345	Trung Bì	CNVC	Trung bình	80	5783	5920	Có
M0302	Vũ Tuấn Hoa	198692720	Trung Bì	Giám đốc	Ít	101	6283	6420	Có
M0000	Ngô Thị Nhật	192095078	Khá	Kinh doar	Ít	91	3083	3220	Có

Record: 3 of 8491



Bảng 5. *Bảng so sánh kết quả*

	C45		C45-Theo xấp xỉ ε=0,005	
Số lượng sai	256	18,82%	148	10,88%
Số lỗi	0	0,00%	0	0,00%
Số đúng	1104	81,18%	1212	89,12%
Thời gian chạy	2s		2s	

Như vậy, với việc nhận ra các phụ thuộc hàm theo giá trị xấp xỉ ε=0,005: *CôngViệc* $\Rightarrow_{\varepsilon}$ *PhụCấp*, *Lương* $\Rightarrow_{\varepsilon}$ *ThuNhập*, số lượng lỗi trong quá trình dự đoán đã giảm 108 trường hợp trên 1353 mẫu dự đoán (tương đương 7.94%).

V. Kết luận

Sự phụ thuộc dữ liệu giữa các thuộc tính có ảnh hưởng lớn đến việc trích chọn mẫu huấn luyện nhằm xây dựng cây quyết định có hiệu quả. Việc nhận ra sự phụ thuộc dữ liệu góp phần làm cải thiện hiệu quả trong bài toán phân lớp. Với sự phụ thuộc tự nhiên của dữ liệu thì ta dễ dàng nhận ra và xử lý, tuy nhiên, trong các bài toán thực tế thì còn có các phụ thuộc xấp xỉ do bản chất của dữ liệu nghiệp vụ. Việc nhận định được giá trị xấp xỉ của dữ liệu trong khi huấn luyện đã làm tăng thêm đáng kể độ chính xác.

TÀI LIỆU THAM KHẢO

1. B.Liu, W. Hsu, Y. Ma. *Integrating classification and association mining*, Proc. Int. Cnf. Knowledge Discovery and Data Mining (KDD'98), New York, (1998), 80-86.

2. Đoàn Văn Ban. *Phương pháp thiết kế và khai thác kho dữ liệu*, Đề tài nghiên cứu cấp TT KHTN & CNQG, Hà Nội, 1997.

3. Đoàn Văn Ban, Lê Mạnh Thạnh, Lê Văn Tường Lân. *Một phương pháp để xây dựng cây quyết định có hiệu quả trong khai phá dữ liệ*u, Kỷ yếu hội thảo khoa học quốc gia về Công nghệ phần mềm & Công nghệ tri thức, (2006), 38-48.

4. 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 liệ*u, Tạp chí Tin học và Điều khiển học, T23, S4, 2007.

5. Lê Thanh Huệ, Lê Văn Tường Lân, Đặng Đình Đường. *Một cách tiếp cận mới trong khai phá dữ liệu*, Tạp chí Khoa học Kỹ thuật Mỏ Địa chất Chuyên đề Công nghệ Thông tin, S20, 2007.

6. Đỗ Văn Thành, Phạm Thọ Hoàn. *Một cách tiếp cận nghiên cứu phát hiện tri thức trong các cơ sở dữ liệu trợ giúp quyết định*, Tuyển tập hệ mờ mạng nơron và ứng dụng, Nhà xuất bản Khoa học và Kỹ thuật, 2001

7. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, S., and Uthurusamy, R. *Advances in Knowledge Discovery and Data Mining*, M.I.T. Press, 1996.

8. Ho Tu Bao. *Introduction to knowledge discovery and data mining*, Institute of Information Technology National Center for Natural Science and Technology, 2000. http://www.jaist.ac.jp/~bao

9. J. Gehrke and W. Loh. *Advances in Decision Tree Construction*, KDD, 2001.

10. Quinlan,J.R.: *Simplifying decision trees*, International Journal of Man-Machine Studies, 27, 221-234, 1987. http://www.mlrg.cecs.ucf.edu/MLRG_documents/c4.5.pdf

11. Yka Huhtala, Juha Kahkkainen, Pasi Porkka, Hannu Toivonen. *An efficient algorithm for discovering function and appoximate dependencies*, Proc.14th Int, Conf. on Data Engineering (ICDE'98), IEEE. Computer Society Press (1998), 392 - 402.

12. Vũ Đức Thi. *Cơ sở dữ liệu - kiến thức và thực hành*, XNB thống kê, Hà Nội, 1997.

13. Zhang, J. and Honavar. *Learning Decision Tree Classifiers from Attribute-Value Taxonomies and Partially Specified Data*, Proceedings of the International Conference on Machine Learning. Washington DC, 2003.

THE EFFECTS OF DEPENDENCY DATA IN DATA MINING'S CLASSTIFICATION

Le Van Tuong Lan
College of Sciences, Hue University

SUMMARY

Decision tree is one of the effective and visual solutions to describe the characteristics of mined data. From the decision tree, we can easily find the rules which provide information on solving a certain issue. 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.

In this article, we have analysed natural dependency data and approximate dependency data… to build an effective decision tree of high predictability for supporting decision making in data analysis problems.

***Keyword**: Data mining, knowledge discovery, decision tree, training set, functional dependency, approximate functional dependency, classification*