Tạp chí Tin học và Điều khiển học, T.23, S.4 (2007), 317–326

MỘT CÁCH CHỌN MẪU HUẤN LUYỆN VÀ THUẬT TOÁN HỌC ĐỂ XÂY DỰNG CÂY QUYẾT ĐỊNH TRONG KHAI PHÁ DỮ LIỆU

ĐOÀN VĂN BAN[1], LÊ MẠNH THẠNH[2], LÊ VĂN TƯỜNG LÂN[2]

[1] *Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Việt Nam*
[2] *Đại học Huế*

Abstract. Data mining for the purpose of discovering useful implicit information from data warehouse, i.e. knowledge discovery to serve supporting decision making in our activities, has become more and more important. Therefore, there exists a lot of methods and techniques focusing on the studies and applications for data mining and knowledge discovery. Decision tree is known to be one of the effective solutions to describe the characteristics of mined data. 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 way of choosing sample training set from business database. Based on this, we will apply learning algorithm to build an effective decision tree of high predictability for supporting decision making in data analysis problems. The obtained results show that proposed this method is more efficient.

Tóm tắt. Khai phá dữ liệu để phát hiện các thông tin bổ ích tiềm ẩn từ các kho dữ liệu, tức là phát hiện các tri thức nhằm phục vụ cho việc hỗ trợ ra quyết định trong các hoạt động của chúng ta ngày càng trở nên quan trọng. Do vậy, đã có nhiều phương pháp, kỹ thuật tập trung nghiên cứu và triển khai ứng dụng để phục vụ cho công việc khai phá dữ liệu và phát hiện tri thức. Cây quyết định là một trong những giải pháp hữu hiệu để mô tả các đặc tính dữ liệu đã được khai phá. Việc 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 chọn tốt bộ dữ liệu mẫu còn gặp nhiều khó khăn. Nếu chúng ta chọn bộ mẫu không đặc trưng thì cây quyết định được sinh ra sẽ không có khả năng dự đoán cao. Trong bài viết này, chúng tôi phân tích và đã chỉ ra một cách chọn tập mẫu huấn luyện tốt từ cơ sở dữ liệu nghiệp vụ, từ đó đưa vào thuật toán học để 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. Kết quả đã được kiểm tra trên thực nghiệm và đã chứng tỏ tính hiệu quả của thuật toán.

1. GIỚI THIỆU

Sự phân lớp là một 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 thuộc về một lớp nào đó [1, 2, 4, 5, 9]. Khi đã tìm được các đặc tính mô tả mẫu dữ liệu khai phá thì cây quyết định là một mô hình trực quan và hữu hiệu để mô tả. Trên cây quyết

định, chúng ta dễ dàng duyệt cây để tìm ra các luật. Các luật này cho chúng ta thông tin để giải quyết một vấn đề nào đó tức là cho chúng ta tri thức về lĩnh vực cần nghiên cứu. 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 CLS, ID3, C45,...[7, 9, 11, 12, 13, 15] với độ phức tạp thuật toán là $O(m \times n \times \log n)$, trong đó m là số thuộc tính, n là số thể hiện của tập huấn luyện.

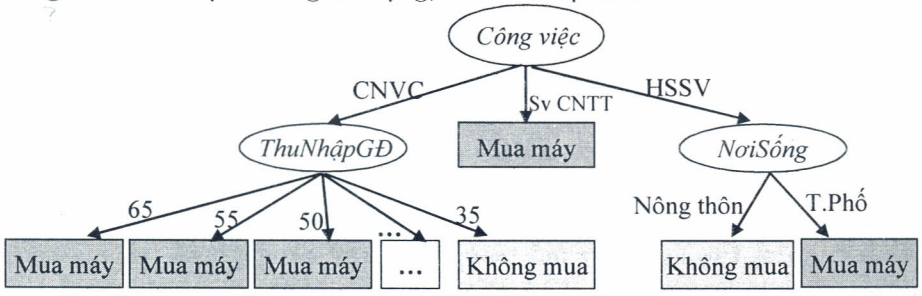
Việc xây dựng một cây quyết định có hiệu quả phụ thuộc vào việc chọn tập mẫu huấn luyện. Trong thực tế, dữ liệu nghiệp vụ rất đa dạng vì chúng được lưu trữ để phục vụ nhiều công việc khác nhau, nhiều thuộc tính cung cấp các thông tin có khả năng dự đoán sự việc nhưng cũng có nhiều thuộc tính không có 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 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

SốPhiếu ĐT	HọVàTên	SốCMND	Nơi Sống	Công Việc	SốNgười GĐ	ThuNhập GĐ	Mua Máy
M01045	Nguyễn Văn An	193567450	T.Phố	CNVC	Nhiều	45	Không
M01087	Lê Văn Bình	191568422	NôngThôn	CNVC	Nhiều	40	Không
M02043	Nguyễn Thị Hoa	196986568	T.Phố	SvCNTT	Nhiều	52	Có
M02081	Trần Bình	191003117	T.Phố	HSSV	Trung bình	34	Có
M02046	Trần Thị Hương	196001278	T.Phố	HSSV	Ít	50	Có
M03087	Nguyễn Thị Lài	198235457	NôngThôn	HSSV	Ít	60	Không
M03025	Vũ Tuấn Hoa	198875584	NôngThôn	SvCNTT	Ít	65	Có
M03017	Lê Bá Linh	191098234	T.Phố	CNVC	Trung bình	35	Không
M04036	Bạch Ân	196224003	T.Phố	CNVC	Ít	60	Có
M04037	Lý Thị Hoa	196678578	T.Phố	HSSV	Trung bình	50	Có
M04042	Vũ Quang Bình	197543457	NôngThôn	CNVC	Trung bình	60	Có
M04083	Nguyễn Hoa	192267457	NôngThôn	SvCNTT	Trung bình	40	Có
M05041	Lê Xuân Hoa	198234309	T.Phố	SvCNTT	Nhiều	55	Có
M05080	Trần Quế Chung	196679345	NôngThôn	HSSV	Trung bình	50	Không
...	...	...	...	...	...	...	...

Giả sử ta chọn tập M_1 = (NơiSống, CôngViệc, SốNgườiGĐ, ThuNhậpGĐ, Mua máy) gồm các bản ghi trên Bảng 1 để làm mẫu huấn luyện cho việc xây dựng cây. Lúc này cây quyết định thu được ở Hình 1 có sự phân chia tại nút ThuNhậpGĐ rất lớn. Trên cây ở Hình 1, lượng thông tin thu được không cô đọng, rất khó dự đoán.



Hình 1. Cây quyết định của mẫu huấn luyện $M1$

Chẳng hạn ta cần dự đoán trường hợp sau có mua máy hay không?

HọVàTên = "Nguyễn Văn B", CôngViệc = "CNVC",
ThuNhậpGĐ = 49, NơiSống = "Nông thôn" (1)

Như vậy, với tập huấn luyện M ta phải khảo sát bản chất của các thuộc tính trước khi thực hiện huấn luyện cây.

2. TÁCH CÂY Ở THUỘC TÍNH RIÊNG BIỆT

Cho mẫu huấn luyện M gồm có m thuộc tính, n bộ. Mỗi thuộc tính bất kỳ $X \in M$ có các giá trị là $\{x_1, x_2, ..., x_n\}$ và ta viết X = $\{x_1, x_2, ..., x_n\}$. 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 $|x_i|$.

Thuộc tính quyết định trong mẫu được đánh dấu là Y và các thuộc tính còn lại gọi là thuộc tính dự đoán. Như thế, trên cây quyết định Hình 1 với mẫu $M1$ thì thuộc tính MuaMáy là thuộc tính quyết định Y, các thuộc tính còn lại là thuộc tính dự đoán.

Để xây dựng cây quyết định, 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 [7, 9, 10, 13]. Tuy nhiên như mẫu huấn luyện $M1$ đã xét, khi dựa vào lượng thông tin nhận được trên thuộc tính ThuNhậpGĐ, cây quyết định thu được ở Hình 1 có sự phân chia tại nút này lớn, nên khả năng dự đoán là không tốt.

Khảo sát tại nút ThuNhậpGĐ trên cây ở Hình 1, ta thấy cây có nhiều nhánh ứng với mỗi một giá trị của thuộc tính quyết định Y. Chiều rộng của cây tại nút này lớn hơn chiều sâu của cây và các giá trị trên từng nhánh là riêng biệt với nhau mặc dù có chung giá trị dự đoán, điều này dẫn đến khó có sự trùng khớp khi thực hiện việc dự đoán trên cây, ví dụ như ở (1).

Định nghĩa 1. 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. Thuộc tính $X \in M$ được gọi là thuộc tính có giá trị riêng biệt (gọi tắt là thuộc tính riêng biệt) nếu như $|X| > (m-2) \times |Y|$. Tập các thuộc tính có giá trị riêng biệt trong M ký hiệu là M^*. Như thế trên mẫu huấn luyện $M1$, ThuNhậpGĐ là thuộc tính riêng biệt.

Định lý 1. *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.*

Chứng minh. Thật vậy, mẫu M có m thuộc tính nên có $m-1$ thuộc tính dự đoán và chiều sâu tối đa của cây là $m-2$ ([9, 13]). Với thuộc tính X là riêng biệt và nó được chọn làm điểm phân tách cây thì theo các thuật toán xây dựng cây [7, 9, 13], tại nút này có ít nhất $((m-2) \times |Y| + 1)$ nhánh nên đây là cây dàn trải. Trên cây ở Hình 1, nút ThuNhậpGĐ là phân chia trên thuộc tính riêng biệt của mẫu $M1$ nên đây là cây dàn trải.

Giả sử $X \in M$ là thuộc tính riêng biệt, được chọn làm điểm phân tách cây, chẳng hạn thuộc tính ThuNhậpGĐ trong mẫu $M1$. Khi đó chúng ta không thể phân tách cây trên tập

giá trị này mà phải phân ngưỡng các giá trị của thuộc tính [3, 8, 9, 14]. Ngưỡng phân tách trên X thường được chọn là giá trị gần nhất sao cho nhỏ hơn hoặc bằng giá trị trung bình của toàn bộ dữ liệu đối với tập các giá trị này [6, 9, 13, 14]. Cách chọn ngưỡng này tương đối hiệu quả khi các giá trị của X phân bố đều trên miền giá trị, còn nếu chúng tập trung vào một số miền con của miền giá trị thì cách này không thật sự hiệu quả do ta chọn phải giá trị mà xác suất xuất hiện không đủ lớn.

Cho mẫu M với thuộc tính quyết định Y. Để ý rằng, chúng ta có thể chọn một giá trị bất kỳ $x_i \in X$ để làm điểm phân tách thì X sẽ có 2 phân hoạch là: $X' = \{x_j \text{ mà } x_j \leq x_i\}$ và $X'' = \{x_j \text{ mà } x_j > x_i\}$, mẫu M lúc này tương ứng sẽ được chia thành 2 mẫu là M' và M''. Vấn đề là phải chọn x_i như thế nào?

Ta nhận thấy là khi chọn thuộc tính để phân tách, thuộc tính $X \in M$ được chọn là thuộc tính có lượng thông tin nhận được $Gain(X, Y, M)$ đạt giá trị lớn nhất [3, 6, 9, 13, 14], nên ở đây ta cũng tính lượng thông tin nhận được cho X tại mỗi giá trị x_i. 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:

$$Gain(x^*|X, Y, M) = \max\{gain(x_i|X, Y, M),\ i = 1, ..., n\},$$

trong đó $gain(x_i|X, Y, M) = S(Y|M) - E(X, x_i, Y, M)$.

$S(Y|M) = \sum_{j=1}^{n} p(y_j) \log(p(y))$ là lượng thông tin (Entropy) của cây đối với thuộc tính quyết định Y trên mẫu huấn luyện M.

$p(y_j) = \dfrac{y_j}{Y}$ là xác suất của y_j trên Y.

$E(X, x_i, Y, M) = \dfrac{|X'|}{|X|} S(Y|M') + \dfrac{|X''|}{|X|} S(Y|M'')$ là kỳ vọng cần thiết để hoàn chỉnh cây khi lấy theo phân hoạch tại giá trị x_i của thuộc tính X làm gốc, mẫu M lúc này được chia ra 2 phân hoạch M' và M''.

Như thế, ở mẫu huấn luyện $M1$ với thuộc tính quyết định Y là MuaMáy ta có:

$S(Y|M1) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0,9403,$

$Gain(\text{CôngViệc}, Y, M1) = S(Y|M1) - 5/14.Entropy(SCNVC) - 4/14.Entropy(SSvCNT)$
$- 5/14.Entropy(SHSSV) = 0,2467,$

$Gain(\text{NơiSống}, Y, M1) = 0,0481,$

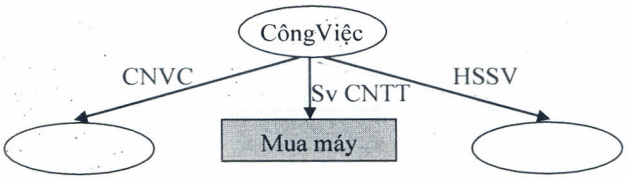
$Gain(\text{SốNgườiGĐ}, Y, M1) = 0,0292.$

Trong Bảng 2, thuộc tính riêng biệt ThuNhậpGĐ có hàm lợi ích cho từng giá trị x_i là $Gain(x_i|\text{ThuNhậpGĐ}, Y, M1)$.

Bảng 2. Lợi ích của thuộc tính ThuNhậpGĐ

x_i	E(ThuNhậpGĐ)	$Gain$(ThuNhậpGĐ)
65	0,8926	0,0477
60	0,9253	0,0150
55	0,8950	0,0453
52	0,8500	0,0903
50	0,8380	0,1022
45	0,9152	0,0251
40	0,9300	0,0103
35	0,8926	0,0477
35	0,8926	0,0477

Ta chọn thuộc tính CôngViệc vì $Gain$(CôngViệc,$Y, M1$) là lớn nhất và cây quyết định tại bước này như Hình 2.



Hình 2. Cây quyết định tại nhánh CôngViệc

Lặp lại đối với các nhánh của cây ở Hình 2, với nhánh thứ nhất tương ứng mẫu $M1$ mới theo nhánh này, ta có:

$$S(Y|M1) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0,9710.$$

$$Gain(\text{NơiSống}, Y, M1) = 0,0200.$$

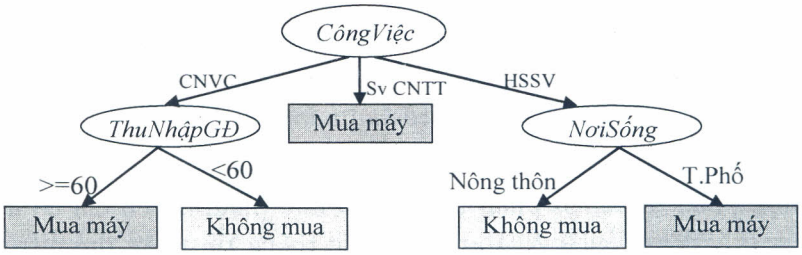
$$Gain(\text{SốNgườiGĐ}, Y, M1) = 0,5710.$$

$Gain(x_i|\text{ThuNhậpGĐ}, Y, M1)$ được tính cho từng giá trị ở Bảng 3.

Bảng 3. Lợi ích của thuộc tính ThuNhậpGĐ của nhánh CNVC

x_i	E(ThuNhậpGĐ)	Gain(ThuNhậpGĐ)
60	0,0000	**0,9710**
45	0,5510	0,4200
40	0,8000	0,1710
35	0,9710	0,0000

Do hàm $Gain(x_i|$ ThuNhậpGĐ$, Y, M1)$ tại giá trị $x^* = 60$ là lớn nhất, ta chọn để làm điểm phân tách cây tại bước này. Thực hiện trên tất cả các nhánh của cây ta thu được cây quyết định như Hình 3.



Hình 3. Cây quyết định của mẫu huấn luyện $M1$

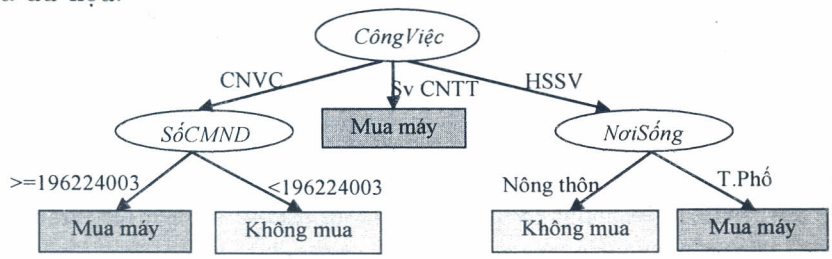
Trên cây này ta có thể dự đoán trường hợp đã nêu ở (1)

HọVàTên = "Nguyễn Văn B", CôngViệc = "CNVC", ThuNhậpGĐ = 49, NơiSống = "Nông thôn" là "Không mua".

Tuy nhiên, nếu ta chọn mẫu $M2$ = (SốCNND, NơiSống, CôngViệc, SốNgườiGĐ, MuaMáy) gồm các bản ghi trên Bảng 1 để làm mẫu huấn luyện để xây dựng cây. Theo cách xây dựng cây ở trên, thuộc tính SốCMND là thuộc tính riêng biệt nên tính ngưỡng tương tự, cây quyết định thu được như ở Hình 4.

Trên cây quyết định thu được, việc phân tách lớp theo thuộc tính SốCMND sẽ cho quyết định: nếu SốCMND từ "196224033" trở lên thì "Mua máy" rõ ràng là một quyết định không

chính xác. Như thế, đây là cây không có khả năng dự đoán do không phản ánh được bản chất thực tế của dữ liệu.



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

Vậy, trên cơ sở dữ liệu nghiệp vụ được lưu trữ, ta phải chọn tập huấn luyện M như thế nào để xây dựng cây quyết định có khả năng dự đoán cao.

3. CHỌN TẬP MẪU HUẤN LUYỆN CHO VIỆC HỌC ĐỂ XÂY DỰNG CÂY

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

Định nghĩa 3. Thuộc tính $X = \{x_1, x_2, ..., x_n\} \in M$ mà giữa các phần tử x_i, x_j với $i \neq j$ không sánh được thì ta gọi X là thuộc tính ghi nhớ trong tập mẫu. Tập các thuộc tính này trong M ký hiệu là M^G.

Mệnh đề 1. *Nếu $X \in M$ là thuộc tính ghi nhớ thì ta loại X ra khỏi mẫu M mà không làm ảnh hưởng đến cây quyết định thu được sau huấn luyện.*

Hiển nhiên, bởi ta không thể so sánh giữa các phần tử x_i với x_j của X để tính hàm $Gain(X, Y, M)$, do đó không tồn tại lợi ích thông tin của mỗi bộ trên X. Vì thế X không bao giờ được chọn là nút để phân tách cây nên ta loại X ra khỏi mẫu M mà không làm thay đổi cây quyết định thu được sau huấn luyện. Như vậy, khi xét mẫu để học được rút từ dữ liệu thực tế, các thuộc tính không định kiểu sẽ bị loại.

Mệnh đề 2. *Nếu thuộc tính X là khóa của mẫu M thì loại X ra khỏi M để cây quyết định thu được không phải là cây dàn trải không có khả năng dự đoán.*

Thật vậy, giả sử $X = \{x_1, x_2, , x_n\}$. Do X là khóa nên $\forall i \neq j$, ta có $x_i \neq x_j$. Như thế, mẫu M được phân ra làm n phân hoạch, mà mỗi phân hoạch chỉ có một bộ nên hàm $E(X, x_i, Y, M) = 0$, với mọi $x_i \in X$. Hàm xác định thông tin nhận được trên thuộc tính X 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, đây là một cây dàn trải theo chiều ngang tại nút X. Hơn nữa, do tính duy nhất của khóa 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. Vậy phải loại X ra khỏi M để thu được cây quyết định có khả năng dự đoán tốt hơn.

Ở Bảng 1, thuộc tính PhiếuĐT và SốCMND là khóa nên nó không thể có mặt trong bất kỳ mẫu huấn luyện nào để xây dựng cây.

Định nghĩa 4. Nếu $X = \{x_1, x_2, ..., x_n\}$ là thuộc tính riêng biệt mà ta không thể phân nhóm cho các giá trị x_i của X theo các phép tính toán thông thường thì ta gọi X là thuộc tính tự do. Tập các thuộc tính này trong M ký hiệu là M^T.

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

Thật vậy, do $X \in M^T$ nên không thể phân cụm để tạo cây theo thuộc tính riêng biệt như ở Mục 2. Mặt khác, do X là riêng biệt nên $|X| > (m-2) \times |Y|$. Như thế đây là nút dàn trải trên cây, theo Định lý 1. Vậy phải loại X ra khỏi M.

Theo Bảng 1, thuộc tính HọVàTên của khách hàng là tự do nên ta không thể đưa vào mẫu để xây dựng cây.

4. THUẬT TOÁN XÂY DỰNG CÂY QUYẾT ĐỊNH TỪ CƠ SỞ DỮ LIỆU NGHIỆP VỤ

4.1. Thuật toán chọn mẫu huấn luyện để xây dựng cây quyết định từ dữ liệu nghiệp vụ

Với dữ liệu nghiệp vụ D đã được lưu trữ có k thuộc tính, thuật toán chọn mẫu huấn luyện M để xây dựng cây quyết định như sau:

Procedure ChonMauHuanLuyen (D, Y, M)
Input: Tập dữ liệu nghiệp vụ D, gồm n bộ, k thuộc tính dự đoán $D_1, ..., D_k$ và thuộc tính quyết định Y
Output: Tập mẫu huấn luyện M trên thuộc tính quyết định Y
Begin
    $M := \{\}$;
    $M^* := \{\}$;
    For $i := 1$ to k do
        Begin
            Kiểm tra tính chất D_i;
            If D_i là riêng biệt then $M^* := M^* \cup D_i$;
            Else If $D_i \notin$ {khoá, tự do, ghi nhớ} then $M := M \cup D_i$;
        End;
    $M = M \cup M^*$;
End;

Theo Mục 3, thuật toán tìm đúng mẫu huấn luyện trên D. Thuật toán có độ phức tạp $O(k)$ đối với "Kiểm tra tính chất D_i;" trên k thuộc tính nên với mẫu có n bộ thì độ phức tạp thuật toán là $O(kn)$. Đến đây, ta có thể cải tiến thuật toán học quy nạp để xây dựng cây quyết định từ dữ liệu nghiệp vụ như sau.

4.2. Cải tiến thuật toán học quy nạp để xây dựng cây quyết định từ dữ liệu nghiệp vụ

Procedure TaoCay(D, Y, T)
Input: Tập mẫu D có n bộ, k thuộc tính dự đoán và thuộc tính quyết định Y
Output: Cây quyết định trên thuộc tính quyết định Y
Begin
    ChonMauHuanLuyen(D, Y, M);

Khởi tạo nút T ứng với mẫu M; {Cây ban đầu với nút gốc là T};
XayDungCay(M, Y, T);
End;

Với thuật toán XayDungCay được xây dựng như sau:
Procedure XayDungCay(M, Y, T)
Begin
If $(y_i = y_j \forall y_i, y_j \in Y)$ then Gán nhãn nút $T := y_i$;
Else
Begin
$S(Y|M) = \sum_{j=1}^{n} -p(y_1) \times \log(p(y_j))$;
IF $X \in M^*$ then
For each $x_i \in X$ do
Begin
$E(X, x_i, Y, M) = \frac{|X'|}{|X|} S(Y|M') + \frac{|X''|}{|X|} S(Y|M'')$;
$Gain(x_i|X, Y, M) = S(Y|M) - E(X, x_i, Y, M)$;
End
Else $Gain(X, Y, M) = S(Y|M) - E(X, Y, M)$;
Chọn X^* có $Gain()$ đạt giá trị cực đại;
Gán nhãn nút $T := X^*$;
If $X^* \notin M^*$ then
Begin
For each $x_i \in X^*$ do
Begin
$M_i := \{m_i \in M : x_j = x_i; x_j, x_i \in X^*\}$
Tạo nút T_i là con của T;
Gán nhãn cho cung từ T đến T_i là x_i;
End;
For each M_i do
Begin
Loại X^* ra khỏi M_i;
Call XayDungCay(M_i, Y, T_i);
End;
End
Else {Chuyển thành phân tách nhị phân theo Mục 3}
Begin
Chọn $x^* \in X^*$ có $Gain(x_i|X)$ đạt cực đại;
$M1 := \{m_i \in M; x_i \geq x^*, x_j \in X^*\}$;
$M2 := \{m_i \in M; x_i < x^*, x_j \in X^*\}$;
Tạo 2 nút T_1 và T_2 là con của T;
Gán nhãn cho cung T đến T_1 là $X^* \geq x^*$;
Gán nhãn cho cung T đến T_2 là $X^* < x^*$;
For $i := 1$ to 2 do

```
        Begin
          Loại X* ra khỏi M_i;
          Call XayDungCay(M_i, Y, T_i);
        End;
      End;
    End;
```

Với tập dữ liệu nghiệp vụ D gồm có n bộ, k thuộc tính dự đoán và thuộc tính quyết định Y, thuật toán xây dựng cây quyết định TaoCay ở trên gồm có các bước chính:

1) *Chọn mẫu huấn luyện M trên D*

Trong các thuật toán học để xây dựng cây quyết định, chẳng hạn [13], việc chọn các thuộc tính cho mẫu huấn luyện tuỳ thuộc vào người chọn nên kết quả của cây phụ thuộc vào sự chọn lựa này, như đã chỉ ra ở các mẫu $M1$ và $M2$. Trong thuật toán TaoCay ở trên, với chi phí thời gian là $O(kn)$ để đánh giá và chọn mẫu huấn luyện, thuật toán đã loại bỏ các thuộc tính không có khả năng dự đoán nên đã tránh được sự lựa chọn chúng một cách ngẫu nhiên từ người sử dụng. Hơn thế nữa, thay vì phải nêu rõ bản chất dữ liệu của từng thuộc tính ở mẫu, mà điều này đòi hỏi trình độ chuyên gia của người xác định, việc duyệt mẫu trong thuật toán TaoCay cũng đã phân loại thuộc tính để việc xây dựng cây đạt hiệu quả.

Như đã phân tích ở Mục 2 và 3, trên dữ liệu DIEUTRA đã cho ở Hình 3 thì thuật toán TaoCay chọn mẫu huấn luyện là $M1$, mà không chọn mẫu $M2$ để học nên tránh được cây không có khả năng dự đoán như ở Hình 4.

2) *Xây dựng cây trên mẫu M*

Thuật toán TaoCay cũng xây dựng cây quyết định dựa trên việc tính hàm $Gain()$ cho các thuộc tính ở mẫu huấn luyện như các thuật toán học khác, chẳng hạn ở [13]. Tuy nhiên, ở các thuộc tính thuộc phân hoạch M^*, do việc tính hàm giá trị $Gain()$ cần thêm chi phí $O(n)$ nên với m thuộc tính riêng biệt, chi phí cho việc xây dựng cây theo thuật toán TaoCay là $O((k-m) \times n \times \log(n) + m \times n^2 \log(n))$. Tuy vậy, việc xác định các thuộc tính $X^* \in M^*$ để chuyển từ điểm phân tách đa phân mà nó sẽ tạo nút dàn trải trên cây thành phân tách nhị phân là thực sự cần thiết, vì nó làm tăng khả năng dự đoán cho cây như đã phân tích ở Mục 3, tránh được cây như ở Hình 1.

Như thế, thuật toán TaoCay ở trên cần thêm thời gian cho việc huấn luyện cây. Tuy nhiên trên thực tế thì chi phí gia tăng này không quá lớn vì chúng ta chỉ học một lần nhưng sẽ dùng cây này để dự đoán cho nhiều lần, mà việc xây dựng cây quyết định hiệu quả nhằm có thể dự đoán chính xác các trường hợp xảy ra là thật sự cần thiết như đã được phân tích ở Mục 2 và 3.

Chúng tôi đã cho huấn luyện trên mẫu gồm 9 trường với 8414 bản ghi, sau đó cho kiểm thử trên tập có 1410 bản ghi. Kết quả được cho ở Bảng 4.

Bảng 4. Bảng so sánh kết quả thực nghiệm

	C45		**SQL analysis**		**Thuật toán đã nêu**	
Số lượng sai	253	17.94%	182	12.91%	186	13.19%
Số lỗi	0	0.00%	6	0.43%	0	0.00%
Số đúng	1157	82.06%	1222	86.67%	1224	86.81%
Thời gian chạy	2s		2s		3s	

Như thế, với việc phân tích bản chất của dữ liệu trước khi chọn mẫu huấn luyện, thuật toán trên bước đầu đã cho kết quả khá tốt.

5. KẾT LUẬN

Phân lớp là một quá trình quan trọng trong khai phá dữ liệu và cây quyết định là một trong những giải pháp hữu hiệu để mô tả các đặc tính dữ liệu đã được khai phá. Tuy vậy, việc xây dựng cây quyết định lại phụ thuộc rất lớn vào tập mẫu huấn luyện. Trong thực tế, việc chọn tốt tập mẫu để xây dựng cây quyết định từ dữ liệu nghiệp vụ còn nhiều hạn chế do tính phức tạp của dữ liệu được lưu trữ.

Bài báo đã phân tích và chỉ ra một cách chọn tốt mẫu huấn luyện từ dữ liệu nghiệp vụ, từ đó đưa vào thuật toán cải tiến học thống kê để xây dựng cây quyết định nhằm phân lớp dữ liệu có hiệu quả, làm tăng khả năng dự đoán trên thực tế. Kết quả kiểm tra trên cùng một bộ mẫu đã cho thấy khi áp dụng thuật toán này thì kết quả đã có sự cải tiến đáng kể so với khi không áp dụng.

TÀI LIỆU THAM KHẢO

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

[2] 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.

[3] Nguyễn Đình Hiền, *Giáo trình xác suất thống kê*, NXB Đại Học Sư Phạm, 2003.

[4] M. Berry, G. Linoff, *Data Mining Techniques for Marketing, Sales and Customer Support*, John Wiley & Sons, Inc., 1997.

[5] J. Bischoff, T. Alexander, *Data Warehouse: Practical Advice from the Experts*, Prentice Hall, 2002.

[6] P. Dorian, *Data Preparation for Data Mining*, Morgan Kaufmann, 1999.

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

[8] S.J. Hong, Use of contextual information for feature ranking and discretization, *IEEE Transactions on Knowledge and Data Eng.* **9** (5) (September/Octorber 1997) 718–730.

[9] Ho Tu Bao, Introduction to knowledge discovery and data mining, 2000. http://www.jaist.ac.jp/~bao

[10] J. Gehrke and W. Loh, *Advances in Decision Tree Construction*, KDD, 2001.

[11] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[12] T.D. Nguyen, and T.B. Ho, An interactive-graphic system for decision tree Induction, *Journal of the Japanese Society for Aritifical Intelligence* **14** (1) (1999) 131–138.

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

[14] C. Westphal, and T. Blaxton, *Data Mining Solutions: Methods and Tools for Real-World Problems*, Wiley, 1998.

[15] J. Zhang, 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.

Nhận bài ngày 24 - 5 - 2007
Nhận lại sau sửa ngày 30 - 8 -2007