Science & Technology Development, Vol 2, No 4 +5-1999

ỨNG DỤNG KỸ THUẬT KHAI MỎ DỮ LIỆU VÀO GIÁO DỤC-ĐÀO TẠO

Hoàng Kiếm - Đỗ Phúc
Trường Đại Học Khoa Học Tự Nhiên.
(Bài nhận ngày 14/04/1999).

TÓM TẮT : Trong bài báo này chúng tôi kết hợp thuật toán tự học trên mạng Kohonen, thuật toán di truyền, và ma trận thông tin để khám phá các luật từ các cụm khái niệm tiềm ẩn trong cơ sở dữ liệu lớn. Chúng tôi đã tiến hành thử nghiệm thuật toán đề nghị vào cơ sở dữ liệu chứa kết qủa học tập của hơn 1000 học sinh trung học ở Biên hoà, Đồng nai nhằm rút ra các tri thức yễm trợ công tác đào tạo học sinh giỏi.

1. GIỚI THIỆU :

Khám phá tri thức từ cơ sở dữ liệu (CSDL) là một hướng nghiên cứu mới của các nhà nghiên cứu CSDL. Mục đích của khám phá tri thức từ CSDL là phát hiện các tri thức tiềm ẩn bên trong CSDL. Theo thống kê của IBM với các công cụ SQL truyền thống, chúng ta chỉ có thể khai thác khoảng 80% thông tin từ CSDL. Khám phá tri thức từ CSDL tập trung việc khai thác lượng thông tin còn lại của CSDL. Các tri thức còn lại tuy nhỏ nhưng lại là những thông tin qúi báu yễm trợ tiến trình hoạch định kế hoạch, xây dụng các chiến lược giáo dục, kinh tế, xã hội [9].

Trong bài báo này chúng tôi tập trung vào bài toán phát hiện các luật từ các cụm khái niệm (conceptual clusters) tiềm ẩn trong CSDL, đây là bài toán hiện đang được nhiều nhà nghiên cứu quan tâm [4,8]. Bài toán phát hiện cụm khái niệm là bài toán gom n đối tượng thành m cụm sao cho các đối tượng trong cùng một cụm là giống nhau theo một ý nghĩa nào đó [1,3]. Có thể liệt kê một số cụm khái niệm tiêu biểu được phát hiện từ CSDL học sinh như sau:

- Cụm khái niệm 1: Học giỏi, hạnh kiểm tốt, có cha mẹ là giáo viên, có anh chị học đại học
- Cụm khái niệm 2: Thích học môn Toán, Ngoại ngữ, Tin học.

Từ các cụm khái niệm này, chúng ta có thể tạo ra các luật suy diễn gần đúng hay luật mờ [13] thể hiện mối quan hệ nhân qủa giữa gỉa thuyết và kết kuận của luật. Một số luật gần đúng tiêu biểu có dạng sau:

- Nếu học sinh có bố mẹ làm nghề giáo Thì học sinh học giỏi 3 năm liền.
- Nếu học sinh có tổng thu nhập gia đình trên 2 triệu /tháng Và Có anh chị học Đại học Thì học sinh đạt giải kỳ thi cấp quốc gia.

Mỗi luật sẽ có một hệ số tin cậy cho biết độ tin cậy của luật.

Có nhiều thuật toán truyền thống dựa trên tiếp cận thống kê để phát hiện cụm khái niệm nhưng những kỹ thuật này tỏ ra kém hiệu qủa khi số chiều tương đối lớn (trên trăm biến) và số lượng đối tượng nhiều (trên một ngàn đối tượng) [4,8,11,12].

Chúng tôi đã sử dụng thuật toán tự học trên mạng Kohonen và thuật toán di truyền để phát hiện cụm khái niệm [6] làm nền tảng cho việc xây dựng các luật.

Bài báo này được tổ chức như sau:Phần 1: giới thiệu. Phần 2: định nghĩa các khái niệm liên quan đến bài toán. Phần 3 : phát biểu và ý nghĩa của các bài toán. Phần 4: Ứng dụng thuật toán tự học trên mạng Kohonen để tìm các tập chỉ báo lớn. Phần 5: Sử dụng thuật toán di truyền và các tương ứng để phối kiểm cụm khái niệm. Phần 6: Ứng dụng thuật toán vào ma trận thông tin học sinh để phát hiện tri thức từ dữ liệu. Phần 7: kết luận, hướng phát triển

2. ĐỊNH NGHĨA CÁC KHÁI NIỆM CÓ LIÊN QUAN ĐẾN BÀI TOÁN

2.1. Định nghĩa 1: Ma trận thông tin B: Cho tập hợp $O = \{o_1,...,o_n\}$ gồm n đối tượng và tập hợp $I = \{i_1,...,i_m\}$ gồm m chỉ báo. Quan hệ hai ngôi R giữa O và I là tập con của tích Descartes OxI. Quan hệ hai ngôi R được mô tả bằng một ma trận B như sau:

$$B(b_{ij}), \text{với } i = 1,...,n \text{ và } j=1,...,m.$$

Phần tử b_{ij} có trị 1 nếu đối tương o_i có chỉ báo i_j và 0 nếu ngược lại.

Ma trận B được gọi là ma trận thông tin [5].

Ví dụ1: Một ma trận thông tin B lấy từ bảng dữ liệu điều tra học sinh sau đây:

Bảng 1: Ma trận thông tin

	i1	i2	i3	i4	i5	i6
o1	1	1	1			
o2	1	1	1			
o3	1	1	1	1	1	1
o4	1	1	1	1	1	1
o5			1	1	1	1
o6			1	1	1	1

Trong đó các chỉ báo có ý nghĩa như sau:

Bảng 2: Ý nghĩa các chỉ báo trong Ma trận thông tin ở bảng 1

Chỉ báo	Ý nghĩa
i1	Bố mẹ làm nghề giáo
i2	Học sinh giỏi trong 3 năm liền
i3	Có hạnh kiểm tốt
i4	Có đạt giải kỳ thi cấp quốc gia
i5	Có anh chị đang học đại học
i6	Có tổng thu nhập gia đình trên 2 triệu /tháng

Oi là học sinh thứ I trong ma trận thông tin. Mỗi học sinh sẽ tương ứng với một vector có 6 thành phần.

2.2. Định nghĩa 2: Các tương ứng

Gọi P(I) là tập các tập con của tập hợp I các chỉ báo. Gọi P(O) là tập các tập con của tập hợp O các đối tượng.

Ta định nghĩa hai tương ứng ρ và λ như sau:[15]

$\rho : P(I) \text{------} > P(O)$ và

$\lambda : P(O) \text{-----} > P(I)$

sao cho :

• Nếu $S \subseteq I$ thì $\rho(S) = \{ o \in O / \forall a \in S , (o,a) \in R \}$

• Nếu $X \subseteq O$ thì $\lambda(X) = \{ a \in I / \forall o \in X , (o,a) \in R \}$

Ví dụ2: Với ma trận thông tin trong bảng 1, ta có:

Nếu S ={i1, i2} thì X=ρ(S) ={o1, o2, o3, o4 }

Tập X này là tập các học sinh có "Bố mẹ làm nghề giáo và Học sinh giỏi trong 3 năm liền".

Nếu X={o4, o5, o6} thì S=λ(X)={i3, i4, i5, i6}

Tập X bao gồm các học sinh có các đặc điểm "Có hạnh kiểm tốt và Có đạt giải kỳ thi cấp quốc gia và Có anh chị học Đại học và Có tổng thu nhập gia đình trên 2 triệu /tháng ".

2.3. Định nghĩa 3: tập chỉ báo lớn(large item set)

Cho một ma trận thông tin B được xác định bởi tập đối tượng O và tập chỉ báo I.

Cho $S \subseteq I$ và MINSUP là một ngưỡng cho trước. S là một tập chỉ báo lớn của ma trận thông B với ngưỡng MINSUP nếu Card(ρ(S))/Card(O) >= MINSUP.

Ví dụ 3: Với ma trận thông tin trong bảng 1 và MINSUP = 4/6 , ta có các tập chỉ báo lớn sau đây {i1,i2,i3 } ; {i3, i4, i5,i6} và các tập hợp con của chúng.

2.4. Định nghĩa 4: Hệ số tin cậy của luật

Cho $S \subseteq I$, S là một tập chỉ báo lớn. $\forall L_i , L_j \subseteq S$ ta tạo luật kết hợp $L_i \text{ --> } L_j$.

Hệ số tin cậy của luật là tỉ số :

$Card(\rho(L_i),\rho(L_j)) / Card(\rho(L_i))$

2.5. Định nghĩa 5: cụm khái niệm

Cụm dữ liệu C là một cặp (X,S) với $X \subseteq O$ và $S \subseteq I$ thỏa tính chất sau :

i)Tính chất 1: $X \subseteq \rho(S)$ và $\lambda(X) = S$

ii)Tính chất 2: $\forall L_i , L_j \subseteq S$ và $card(L_i) = Card(L_j) = 1$ thì $\rho(L_i) \subseteq \rho(L_j)$

Ví dụ 4: Với ma trận thông tin trong bảng 1 và MINSUP = 2/6, chúng ta có các cụm khái niệm sau đây:

Cụm 1 : ({o1,o2} , {i1, i2})

Cụm 2 : ({o4,o5,o6 } , {i4, i5,i6})

3.PHÁT BIỂU VÀ Ý NGHĨA CỦA CÁC BÀI TOÁN

3.1. Bài toán tìm tất cả các tập chỉ báo lớn

a) Phát biểu bài toán: Cho một ma trận thông tin B và một ngưỡng MINSUP, hãy tìm tất cả các tập chỉ báo lớn của ma trận thông tin B.

b) Ý nghĩa của bài toán: Các tập chỉ báo lớn nhằm xác định tập các thuộc tính phổ biến trong nhiều đối tượng dữ liệu. Nói cách khác là các tập tính chung của một nhóm các đối tượng. Ngưỡng MINSUP cho phép ấn định mức độ phổ biến của tập chỉ báo lớn trong ma trận thông tin.

Ví dụ 5:

• Có khoảng 70% học sinh có "Bố mẹ làm nghề giáo và Học sinh giỏi trong 3 năm liền"

• Có khoảng 85% học sinh "Có hạnh kiểm tốt và Có đạt giải kỳ thi cấp quốc gia và có anh chị học Đại học và Có tổng thu nhập gia đình trên 2 triệu /tháng ".

Các hệ số 70% và 85% là hệ số tin cậy của luật.

3.2. Bài toán phát hiện cụm khái niệm

a) Phát biểu bài toán: Cho một ma trận thông tin B và một ngưỡng MINSUP, hãy tìm k cụm khái niệm $C_1,...,C_k$ với $Cj = (X_j,S_j)$ sao cho :

1. $\cap X_i = \{\}$ với i=1,...,k
2. $\cap S_i = \{\}$ với i=1,...,k
3. $Card(X_i) / Card(O) >= MINSUP$
4. Các $X_1,...,X_k$ phủ O càng nhiều càng tốt
5. Các C_i là các cụm khái niệm.

b) Ý nghĩa của bài toán: Các cụm khái niệm xác định các nhóm dữ liệu có chung một số thuộc tính mà các nhóm khác không có. Dựa trên các cụm khái niệm (X,S) , chúng ta có thể tạo các luật suy diễn có dạng :

$$L_i \text{ --- } > L_j \text{ sao cho } L_i \cup L_j = S.$$

Nghĩa là từ tập thuộc tính này chúng ta chắc chắn suy ra tập thuộc tính kia.

Ví dụ 6:

Từ cụm khái niệm :

({o1,o2} , {i1, i2}) ta có thể có luật suy diễn:

• Nếu học sinh có bố mẹ làm nghề giáo Thì học sinh học giỏi 3 năm liền.

Từ cụm khái niệm:

({o1,o2,o3} , {i4, i5,i6}) ta có thể có luật suy diễn:

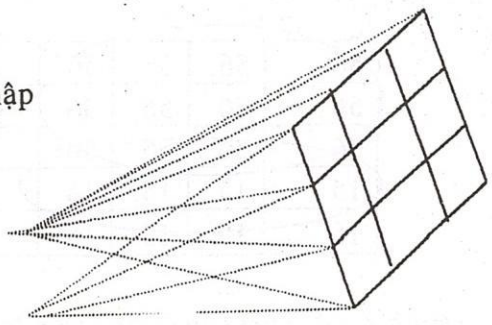
• Nếu học sinh có tổng thu nhập gia đình trên 2 triệu /tháng Và Có anh chị học Đại học Thì học sinh sẽ đoạt giải kỳ thi cấp quốc gia.

4. ỨNG DỤNG THUẬT TOÁN TỰ HỌC TRÊN MẠNG KOHONEN ĐỂ TÌM CÁC TẬP CHỈ BÁO LỚN

4.1. **Cấu trúc mạng Kohonen**: Mạng Kohonen do Teuvo Kohonen phát triển vào những 1979-1980. Có thể xem Mạng Kohonen là một phương tiện biến đổi các mẫu có nhiều chiều thành các tập các cụm có số chiều ít hơn [8].

Trong một mạng Kohonen tiểu biểu, các neuron xuất được sắp xếp trên một lưới hai chiều có kích thuốc MxN. Lưới này được gọi là lớp Kohonen. Tất c

ả các đầu vào đươc nối với các neu-on trên lưới.



Hình1: Cấu trúc của một mạng Kohonen tiêu biểu

4.2. **Phát hiện cụm khái niệm:** Mối liên kết giữa đầu vào và đầu ra được thể hiện qua một trọng số thể hiện mức độ nối kết. Tổng đầu vào mỗi neuron trong lớp Kohonen là tổng các trọng của các đầu vào. Mạng Kohonen sử dụng thuật toán học theo lối tự tổ chức. Các vector trọng ban đầu sẽ nhận các gía trị ngẫu nhiên. Tiến trình học sẽ phát hiện nhóm các nút trên lưới nằm gần nhau và có chung đáp ứng với dữ liệu vào. Thuật toán học của mạng Kohonen là thuật toán học theo lối tranh đua. Khi đưa một vector(mẩu học) vào mạng Kohonen, neuron có đáp ứng đầu ra cao nhất sẽ là neuron thắng cuộc. Có thể dùng khoảng cách Euclide để phát hiện neuron thắng cuộc đó là neuron có khoảng cách giữa vector nhập và vector trọng của nó nhỏ nhất. Các trọng ứng với các nút năm trong vùng lân cận của nút chứa neuron thắng cuộc sẽ được cập nhật.

Thuật toán học tự tổ chức có thể tóm tắt như sau:

1.Khởi tạo ngẩu nhiên các vector trọng

2. Chọn nút có khoảng cách dv nhỏ nhất từ vector nhập v(t) đến vector trọng ứng với nút đó. $dv(v,w_{icjc}) = \min \{ dv(v_i,w_{ij}) \}$

3.Cập nhật các vector trọngcủa các nút nằm trong vùng lân cận của nút (i_c,j_c):

$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)(v(t)-w_{ij}(t))$

Trong đó $i_c-N_c(t) <= i <= i_c+N_c(t)$ và

$j_c-N_c(t) <= j <= j_c+N_c(t)$

4. Cập nhật thời gian t = t+1, và xét vector nhập kế tiếp rồi nhảy đến bước (2)

Trong thuật toán trên, $\alpha(t)$ là độ lợi ($0 <= \alpha(t) <= 1$). $N_c(t)$ là bán kính của vùng lân cận. $N_c(t)$ và $\alpha(t)$ sẽ đơn điệu giảm theo thời gian.

4.3.Một ví dụ dùng thuật toán tự học để tìm tập chỉ báo lớn

Cho Ma trận thông tin B ở bảng 3:

Bảng 3: Ma trận thông tin thử nghiệm

	i1	i2	i3	i4	i5	i6
o1	1	1	1			
o2	1	1	1			
o3	1	1	1	1		
o4			1	1	1	1
o5				1	1	1
o6				1	1	1

Mỗi dòng của ma trận thông B sẽ là một vector nhập vào mạng Kohonen. Sau khi chạy thuật toán học theo lối tự tổ chức chúng ta có đầu ra của các neuron của lớp Kohonen:

56	56	56	56	7
56	56	56	56	7
56	56	56	56	7
15	15	15	15	0
15	15	15	15	0

Hình 2: các cụm của các neuron trên lớp Kohonen

Các cụm khái niệm được đặt trưng bằng các vector trọng sau đây:

Mã	i1	i2	i3	i4	i5	i6
56	0	0	0	1	1	1
15	1	1	1	1	0	0
7	1	1	1	0	0	0

Từ các vector trên, chúng ta rút ra được các tập chỉ báo lớn với ngưỡng MINSUP = 2/6 sau đây:

{i4, i5, i6} có Card(ρ({i4, i5, i6})) = 3

{i1, i2, i3} có Card(ρ({i1, i2, i3})) = 3

Tập {i1, i2, i3, i4} có Card(ρ({i1, i2, i3, i4 }))/Card(0) = 1/3 < MINSUP nên bị loại.

5. SỬ DỤNG THUẬT TOÁN DI TRUYỀN VÀ CÁC TƯƠNG ỨNG ĐỂ PHỐI KIỂM CỤM KHÁI NIỆM

5.1.Thuật toán di truyền

Thuật toán di truyền(genetic algorithm gọi tắt là GA) do John Holland đề xướng vào những năm 1970. Thuật toán di truyền dùng sự tiến hóa của tự nhiên qua nhiều bước tiến hoá để tìm kiếm lời giải tối ưu. Có thể tóm tắt thuật toán di truyền như sau{2,14]:

```
Procedure GA;
 Begin
    T = 0;
    Khởitạo P(t);
    LượngGía P(t);
    While  Not( KếtThúc) do
      Begin
        Chọn P(t) từ P(t-);
        K      p P(t) ;
        LượngGía P(t);
       End;
End;
```

Các thao tác cơ bản của GA bao gồm:

- Biểu diễn di truyền các lời giải tiềm năng của bài toán
- Tạo quần thể ban đầu
- Định nghĩa hàm đo mức độ phù hợp
- Định nghĩa các toán tử di truyền(đột biến và lai ghép)
- Định nghĩa các tham số do GA sử dụng.

5.2. Ứng dụng thuật toán di truyền để kiểm chứng và chọn lựa cụm khái niệm.

Các tập chỉ báo lớn được phát hiện bằng thuật toán học theo lối tự tổ chức không chắc sẽ trở thành các cụm khái niệm. Trong ví dụ ở mục 4.3, ta có {i1 , i2, i3} là một tập chỉ báo lớn, thế nhưng đây không phải là một cụm khái niệm vì :

ρ({i3} = { o1, o2,o3,o4} nhưng

ρ({i1, i2}) = {o1, o2,o3}

Do vậy ở bước này chúng ta sẽ xây dựng một thuật toán để giải bài toán phát hiện các cụm khái niệm [1]. Chúng ta dễ dàng kiểm chứng nhận xét sau: Nếu L là một tập chỉ báo lớn với ngưỡng MINSUP và S là một tập con của L thì S cũng là một tập chỉ báo lớn với ngưỡng MINSUP.

Xuất phát từ nhận định trên, chúng tôi sử dụng các tập chỉ báo được phát hiện từ thuật toán học trên mạng Kohonen làm ứng viên cho giai đoạn phát hiện cụm khái niệm.

Gọi L={L1,...,Lk} là tập các tập chỉ báo lớn, chúng ta sẽ sử dụng thuật toán di truyền để tìm tập C= {X1,...,Xk} sao cho Xi $\subseteq$ Li và thoả định nghĩa về cụm khái niệm. Các thao tác của thuật toán di truyền được định nghĩa như sau:

- Biểu diễn di truyền của nhiễm sắc thể là tập các BXi, mỗi BXi là một dãy bit biểu diễn tập hợp Xi tương ứng. Với hai tập chỉ báo lớn {i1, i2, i3} và {i4, i5, i6} ta có một nhiễm sắc thể sau đây:

i1	i2	i3	i4	i5	i6
1	1	1	1	1	1

Ứng với tập { {i1, i2, i3} , {i4, i5, i6} }

- Biểu diễn quần thể P là tập có nhiều nhiễm sác thể :

Ví dụ: Quần thể P với ba nhiễm sắc thể:P(t)= {111111, 100111,001100}

- Phép lai ghép:

Cho hai nhiễm sắc thể cha:

a1	a2	a3	a4	a5	a6

b1	b2	b3	b4	b5	b6

Trong đó các ai,bi $\in$ {0,1} và i=1,...,6.

Phép lai ghép cho phép hoán vị một bộ phận của hai nhiễm sác thể cha cho nhau để tạo ra hai nhiễm sắc thể con:

a1	a2	a3	b4	b5	b6

b1	b2	b3	a4	a5	a6

- Phép đột biến:

Cho một nhiễm sắc thể

a1	a2	a3	b4	b5	b6

Chọn ngẫu nhiên một vị trí từ 1 đến 6. Gọi ai là vị trí được, nếu ai đang là 1 thì sẽ đổi sang 0 và ngược lại.

- Hàm thích hợp của một nhiễm sắc thể được tính bằng tổng điểm của từng Xi có trong nhiễm sắc thể.

- Điểm của một Xi : Tạo các tập con có 2 phần tử của tập Xi. Gọi N là số tập con thu được. Gọi {a,b} là một trong N tập con thu được. Từ tập con này ta tạo ra hai luật a-- > b và b -- > a và tính điểm tin cậy của luật dựa trên định nghĩa 4. Điểm của tập Xi bằng tổng điểm của tất cả các luật chia cho 2xN .

- Điểm của nhiễm sắc thể bằng trung bình cộng điểm của các Xi có trong nhiễm sắc thể.

Đối với nhiễm sắc thể : 110110 ,ứng với các tập {i1, i2}, {i5, i6}
Tập {i1, i2} có hai luật là i1 -- > i2 và
i2 -- > i1 hai luật này đều có điểm là 1 nên điểm của tập này là 1.
Tập {i5, i6} có hai luật là i5 -- > i6 và
i6 -- > i5 hai luật này đều có điểm là 1 nên điểm của tập này là 1.
Điểm của nhiễm sắc thể này cũng là 1.
Đối với nhiễm sắc thể : 111110, ứng với các tập {i1, i2, i3}, {i5, i6}
Tập {i1, i2, i3} có các luật là i1 -- > i2
và i2--> i1 hai luật này đều có điểm là 1 nên điểm của tập này là 1.
Luật i1 -- > i3 có điểm là 1.
Luật i3 -- > i1 có điểm là ¾= 0.75.
Luật i2 -- > i3 có điểm là 1.
Luật i3 -- > i2 có điểm là ¾= 0.75.
Điểm của tập này (0.75+0.75+1)/3 = 0.83
Tập {i5, i6} có hai luật là i5 -- > i6
và i6 -- > i5 hai luật này đều có điểm là 1 nên điểm của tập này là 1.
Điểm của nhiễm sắc thể này sẽ là:
(1+0.89)/2 = 0.915

Ơ thế hệ t-1 , say khi đã dùng phép toán lai ghép và đột biện để tạo thế hệ kết tiếp, các nhiệm sắc thể sẽ được tính điểm dựa trên hàm phù hợp và được sắp xếp dựa trên giá trị của hàm phù hợp. Các nhiểm sắc thể có giá trị hàm phù hợp lớn sẽ được chọn cho thế hệ kế tiếp và tiến trình lặp tiếp diễn cho đến khi các nhiễm sắc thể đạt đến một ngưỡng định trước.

6. ỨNG DỤNG THUẬT TOÁN VÀO MA TRẬN THÔNG TIN HỌC SINH ĐỂ KHÁM PHÁ TRI THƯC TỪ DỮ LIỆU

Chúng tôi đã ứng dụng các thuật toán nêu trên vào ma trận thông tin có 1000 dòng, mỗi dòng là số liệu của một học sinh và 100 cột, mỗi cột là một chỉ báo.

Mạng Kohonen sữ dụng có kích thước 100x100. Một số chỉ báo mà chúng tôi đã thu nhận từ CSDL học sinh được nêu trong bảng 4:

Bảng 4: Tập một số chỉ báo của Ma trận thông tin.

STT	Chỉ báo
1	Là học sinh giỏi 3 năm liền
2	Có học lực giỏi
3	Có học lực khá
4	Có học lực trung bình
5	Có đoạt giải thưởng về Tin học
6	Có đoạt giải thưởng về Toán
7	Có đoạt giải thưởng về Ngoại ngữ
8	Có đoạt giải thưởng về Văn
9	Thích học môn toán
10	Thích học môn Lý
11	Thích học môn Hóa

12	Thích học môn Sinh
13	Thích học môn Ngoại ngữ
14	Nhà có điện thoại

Với MINSUP = 70% ta thu được một số tập chỉ báo lớn như đã nêu trong bảng 5.

Bảng 5: Một số tập chỉ báo lớn

STT	Tập chỉ báo lớn
1	Học sinh đoạt giải thưởng Có số giờ tự học > 6 giờ ngày
2	Học sinh loại giỏi Có Anh chị đang học Đại học Hoặc có bố mẹ lam nghể giáo
3	Học sinh thích Toán Học sinh thích Ngoại ngữ Học sinh thích Tin học
4	Học sinh học giỏi thích chọn nghể kỹ thuật, xây dựng, y

Chúng tôi đã sủ dựng các thông số sau đây cho thuật toán di truyền:

- Số lượng nhiễm sắc thể : 50
- Số lượng thế hệ 300
- Xác xuất lai ghép:0.1
- Xác suất đột biến:0.1

Chúng tôi phát hiện một số cụm khái niệm tiêu biểu được nêu trong bảng 6.

Bảng 6: Một số cụm khái niệm được phát hiện

STT	Cụm khái niệm
1	Học giỏi, hạnh kiểm tốt, có cha mẹ làm nghề giáo
2	Thích học môn toán, ngoại ngữ Tin học
3	Học sinh nông thôn có số anh chị > 4
4	Học sinh đoạt giải thưởng có số giờ tự học > 6 giờ ngày

7. KẾT LUẬN, HƯỚNG PHÁT TRIỂN

Việc kết hợp các thuật toán tự học trên Mạng Kohonen và thuật toán di truyền với ma trận thông tin bước đầu đã thu được các kết quả đáng khích lệ. Trong qúa trình triển khai, chúng tôi đã dùng các loại dữ liệu có kiểu rời rạc, chúng tôi đang tiếp tục triển khai các ma trận thông tin với các biến liên tục. Chúng tôi cũng đã sử dụng cách biểu diễn ma trận thông

	Học sinh thích Ngoại ngữ Học sinh thích Tin học
4	Học sinh học giỏi thích chọn nghề kỹ thuật, xây dựng, y

Chúng tôi đã sủ dựng các thông số sau đây cho thuật toán di truyền:

- Số lượng nhiễm sắc thể : 50
- Số lượng thế hệ 300
- Xác xuất lai ghép:0.1
- Xác suất đột biến:0.1

Chúng tôi phát hiện một số cụm khái niệm tiêu biểu được nêu trong bảng 6.

Bảng 6: Một số cụm khái niệm được phát hiện

STT	Cụm khái niệm
1	Học giỏi, hạnh kiểm tốt, có cha mẹ làm nghề giáo
2	Thích học môn toán, ngoại ngữ Tin học
3	Học sinh nông thôn có số anh chị > 4
4	Học sinh đoạt giải thưởng có số giờ tự học > 6 giờ ngày

7. KẾT LUẬN, HƯỚNG PHÁT TRIỂN

Việc kết hợp các thuật toán tự học trên Mạng Kohonen và thuật toán di truyền với ma trận thông tin bước đầu đã thu được các kết quả đáng khích lệ. Trong qúa trình triển khai, chúng tôi đã dùng các loại dữ liệu có kiểu rời rạc, chúng tôi đang tiếp tục triển khai các ma trận thông tin với các biến liên tục. Chúng tôi cũng đã sử dụng cách biểu diễn ma trận thông tin bằng ma trận BIT để có thể chứa nhiều thông tin trong bộ nhớ chính của Máy tính nhằm đẩy nhanh tiến trình khám phá tri thức từ các CSDL lớn .

USING DATA MINING IN EDUCATION AND TRAINING

Hoang Kiem - Do Phuc

ABSTRACT : In this paper, we combined the Kohonen self-learning algorithm, genetic algorithm and information matrix to discover the rules from conceptual clusters present implicitly in large databases. We applied the proposed algorithm to a database containing the result of learning of more than 1000 high school students in Bien hoa, Dong nai to discover the rules supporting the training of the outstanding students.

TÀI LIỆU THAM KHẢO.

[1] Bezdek, J. C. Cluster validity with fuzzy sets. In J. Cybernetics, Vol. 3, No.3, 1974, 58-73

[2] David E. Goldberg. Genetic Algorithm in Search, Optimization, and Machine learning, Addison-Wesley Publishing Company, Inc, 1989,147-214

[3] Hair Anderson, Multivariate Data Analysis, MacMillan Publishing Company, 1990, 340-435

[4] Eui-Hong Han Hypergraph Based Clustering in high dimensional dat sets,Data Engineering, IEEE , March 1998

[5] Hoang Kiem, Do Phuc, Using MDDM for mining association rules in a large database, proceedings of IT@EDU98 conference, VNU-HCMC, Vietnam, 1998, 6.6.1-6.6.8

[6] Hoang Kiem, Do Phuc A combined multi-dimensional data model, self-organizing algorithm and genetic algorithm for cluster discovery in data mining, the proceedings of the third conference on knowledge discovery, Bejing, China, 1999

[7] Jun Yan, Using Fuzzy logic, Prentice Hall, 1994,72-76

[8] L.P.J Veelenturf, Analysis and application of Artificial Neural Networks, Prentice hall, 1995,183-191

[9]Pieter Adrians, Dolf Zantige, Data Mining, Addision Wesley, Longman, 1996, 72-76

[10].G.D. Ramkumar Clustering data without distance functions, Data Engineering, IEEE , March 1998

[11] Raymond T Ng, Jiawei Han, Efficient and Effective clustering methods for spatial data mining, Proceedings of the 20th VLDB conference, Santiago, Chile, 1994

[12] O.Strout, Chemical Pattern Recognition, Research studies Press, England, 1986,

[13] Timothy J. Ross, Fuzzy logic with engineer application, Mac Graw Hill, 1995,

[14] Zbigniew Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag, 1992, 13-42

[15] Ho Tu Bao An automatic unsupervised learning on Galois lattice with the dynamic data, Proceedings of IOIT conference, Hanoi, Vietnam, 1996, 27-35