XÂY DỰNG CHÙM CÁC HÀM MẬT ĐỘ XÁC SUẤT TỪ DỮ LIỆU RỜI RẠC

Võ Văn Tài[1] và Nguyễn Trang Thảo[1]

ABSTRACT

This article presents some conceptions, theoretical results and algorithms for building clusters of the probability density functions. With programs written by Matlab, we solve the computing problem of clustering probability density functions. This technique can illustrate the real discrete data about the extra-practicing and studying marks of the students from CONS (College of Naturel Science), Can Tho University.

Keywords: Cluster, cluster width, hierarchical method, non–hierarchical method

Title: Bulding clusters of probability densitiy functions from discrete data

TÓM TẮT

Bài báo trình bày một số khái niệm, kết quả lý thuyết và thuật toán để xây dựng chùm các hàm mật độ xác suất. Với các chương trình được viết bằng Matlab, chúng tôi giải bài toán với máy tính để xây dựng chùm các hàm mật độ xác suất. Kỹ thuật này có thể minh giải các dữ liệu rời rạc thực tế về điểm rèn luyện và điểm học tập của sinh viên Khoa Khoa học Tự Nhiên, Trường Đại học Cần Thơ.

Từ khóa: Chùm, độ rộng chùm, phương pháp thứ bậc, phương pháp không thứ bậc

1 GIỚI THIỆU

Khi làm việc với tập dữ liệu lớn, đến từ nhiều nguồn khác nhau, người ta có nhu cầu phân chia chúng thành những nhóm với những phần tử "gần" nhau theo một dấu hiệu được chọn lựa, từ đó bài toán phân tích chùm ra đời. *Phân tích chùm là việc nhóm các phần tử trong tập hợp đã cho thành các chùm sao cho các phần tử trong cùng chùm tương tự nhau theo những dấu hiệu được chọn lựa. Khi chùm được xây dựng, những phần tử trong cùng một chùm sẽ có sự tương tự nhiều hơn so với những phần tử của chùm khác*. Có rất nhiều ứng dụng cụ thể trong những lĩnh vực khác nhau của bài toán phân tích chùm: y học, sinh học, kinh tế, kỹ thuật, xã hội,…và trong bất kỳ lĩnh vực nào nơi việc nhóm những phần tử lại với nhau được đòi hỏi. Một số tác giả như Sibson (1973), Defays (1977), Rohlf (1982),…đã đưa ra những thuật toán cụ thể cho những dữ liệu rời rạc. Fukunaga (1990), Webb (2002) đã tổng kết những phương pháp liên quan đến phân tích chùm. Nhưng vấn đề phân tích chùm cũng chỉ xét cho dữ liệu rời rạc với tiêu chuẩn đánh giá "gần" và "xa" bởi khoảng cách truyền thống mà không dựa vào sự phân bố của dữ liệu. Do đó, trong một số trường hợp nó tạo ra sự nghịch lý: phần tử đúng lý phải được xếp vào chùm này nhưng lại được xếp vào chùm kia. Năm 2010 nhóm tác giả Võ Văn Tài, Phạm Gia Thụ đã đưa ra khái niệm *độ rộng chùm* làm tiêu chuẩn phân tíchchùm các hàm mật độ xác suất. Độ rộng chùm được định nghĩa qua tích phân hàm cực đại của các hàm mật độ xác suất, vì vậy khi đánh giá sự tương tự của các phần tử, yếu tố phương sai đã được xem xét. Điều này thể hiện sự hợp lý hơn trong phân tích chùm. Tuy nhiên, trong việc giải quyết bài toán chùm các hàm mật độ xác suất, vấn đề ước lượng hàm mật độ xác suất từ số liệu rời rạc và việc tính độ

rộng chùm vẫn còn gặp nhiều khó khăn. Trong bài viết này chúng tôi có bổ sung kết quả lý thuyết liên quan đến độ rộng chùm và vấn đề tính toán qua các chương trình được viết trên phần mềm Matlab. Một ví dụ với số liệu thực về điểm rèn luyện và điểm học tập của sinh viên Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ được đưa ra để kiểm chứng các thuật toán, các chương trình đã viết và cũng để minh họa cho các ứng dụng của bài toán phân tich chùm.

2 SỰ TƯƠNG TỰ VÀ ĐỘ RỘNG CHÙM CÁC HÀM MẬT ĐỘ XÁC SUẤT

2.1 Sự tương tự của các hàm mật độ xác suất

Tiêu chuẩn đánh giá sự tương tự của hai phần tử rời rạc là khoảng cách truyền thống. Người ta cũng có nhiều định nghĩa khác nhau về khoảng cách của hai chùm rời rạc, tuy nhiên việc chọn khoảng cách nào là tối ưu để đánh giá sự tương tự của các phần tử rời rạc là câu hỏi đã được nhiều nhà toán học quan tâm, nhưng hiện còn bỏ ngõ. Trong trường hợp 2 hàm mật độ xác suất, sự tương tự của chúng thông thường cũng được đánh giá qua khái niệm khoảng cách như: Khoảng cách Chernoff, khoảng cách Bhattacharyya, khoảng cách Divergence,…Khi có nhiều hơn hai hàm mật độ xác suất, nghiên cứu về tính tương tự của nó chưa được các nhà toán học quan tâm nhiều. Có hai khái niệm cổ điển được đưa ra ở trường hợp này. Đó là khái niệm *độ đo tách rời* của Glick (1973) và *affinity* của Matusita (1967) cũng như của Toussaint (1972).

***Định nghĩa* 1**: *Một hàm đối xứng* s *được gọi là độ đo* $k\ (k \geq 2)$ *điểm tách rời cho tập* S *trong không gian véc tơ với chuẩn* $\|.\|$ *nếu với mọi phần tử* $a_1, a_2, ..., a_k \in S$ *nó thỏa mãn điều kiện*

$$\max_{i<j}\|a_i - a_j\| \leq s(a_1, a_2, ..., a_k) \leq \sum_{i<j}\sum\|a_i - a_j\| \quad (1)$$

Từ (1) có nhiều định nghĩa cụ thể về hàm s đã được chỉ ra.

***Định nghĩa* 2**: *Cho k hàm mật độ xác suất* $f_1, f_2, ..., f_k$, $(k \geq 2)$*, ta có các định nghĩa affinity như sau:*

i) Affinity của Matusita: $D_M(f_1, f_2, ..., f_k) = \int_{R^n} (f_1.f_2...f_k)^{1/k}\, d\boldsymbol{x}$ (2)

ii) Affinity của Toussaint: $D_T(f_1, f_2, ..., f_k)^{\alpha} = \int_{R^n} \prod_{j=1}^{k} [f_j(x)]^{\alpha_j}\, d\boldsymbol{x}$ (3)

Trong đó $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k)$, $\alpha_j \in (0,1)$, $\sum_{j=1}^{k} \alpha_j = 1$.

Trong trường hợp đặc biệt $\alpha_1 = \alpha_2 = \cdots = \alpha_k = \frac{1}{k}$ thì affinity của Toussaint trở thành affinity của Matusita, và khi $k = 2$ thì nó trở thành affinity của Hellinger.

2.2 Độ rộng chùm

a) Định nghĩa

***Định nghĩa* 3**: *Cho* k *hàm mật độ xác suất trên* $\boldsymbol{R}^n$: $\{f_1, f_2, ..., f_k\}$, $k \geq 2$, *độ rộng của chùm* $\{f_1, f_2, ..., f_k\}$ *được định nghĩa như sau*:

$$w(f_1, f_2, ..., f_k) = \int_{R^n} f_{\max}(\boldsymbol{x})d\boldsymbol{x} - 1 \quad (4)$$

***Định nghĩa* 4**: *Cho* $g, (g_1, g_2, ..., g_n), (f_1, f_2, ..., f_m)$ *là các hàm mật độ xác suất, chúng ta định nghĩa độ rộng của chùm* $\{g, (f_1, f_2, ..., f_m)\}$ *là* $w[g \cup \{f_1, f_2, ..., f_m\}]$ *và độ rộng của chùm* $\{(f_1, f_2, ..., f_m), (g_1, g_2, ..., g_n)\}$ *là* $w[\{f_1, f_2, ...f_m\} \cup \{g_1, g_2, ...g_n\}]$.

b) Định lý về độ rộng chùm

Cho $f_1, f_2, ..., f_k, f_{k+1}$ là hàm mật độ xác suất của $k+1$ tổng thể. Chúng ta có các kết quả sau về độ rộng của chùm:

a) $w(f_1, f_2, ..., f_{k+1}) - w(f_1, f_2, ..., f_k) = 1 - \int_{R^n} \min\{h_1(\boldsymbol{x}), f_{k+1}(\boldsymbol{x})\}d\boldsymbol{x}$ (5)

Trong đó $h_1(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_k(\boldsymbol{x})\}$, $k \geq 3$.

b) $w(f_1, f_2, ..., f_k) = w(f_1, f_2, ..., f_n) + w(f_{n+1}, f_{n+2}, ..., f_k) + 1 - A$ (6)

Trong đó n, $k \geq 3$, $n < k$ và $A = \int_{R^n} \min\{k_1(\boldsymbol{x}), k_2(\boldsymbol{x})\}d\boldsymbol{x}$ với

$k_1(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_n(\boldsymbol{x})\}$, $k_2(\boldsymbol{x}) = \max\{f_{n+1}(\boldsymbol{x}), f_{n+2}(\boldsymbol{x}), ..., f_k(\boldsymbol{x})\}$.

c) $\max_{i<j}\{w(f_i, f_j)\} - \frac{1}{2} \leq w(f_1, f_2, ..., f_k) \leq \frac{1}{k}\sum_i \sum_{<j} w(f_i, f_j)$ (7)

Chứng minh (5), (6) và (7) khá dài, chúng tôi xin phép không trình bày ở đây.

Nhận xét: i) Kết quả (6) được hiểu như sau:

$w(f_1, f_2, ..., f_i) + w(f_{i+1}, ..., f_k)$ là tổng độ rộng của hai chùm trước khi ghép.

$1 - A$ là khoảng cách ngoài hay khoảng cách giữa hai chùm.

Độ rộng chùm phản ánh sự gần nhau của những phần tử trong chùm, trong khi khoảng cách ngoài phản ánh sự xa nhau giữa hai chùm. Bởi vì $w(f_1, f_2, ..., f_i, f_{i+1}, ...f_k)$ là hằng số, nên độ rộng chùm và khoảng cách ngoài biến thiên theo hướng trái ngược nhau. Khi ghép hai chùm thành một chùm, chúng ta cực tiểu tổng độ rộng vì vậy cũng có nghĩa cực đại khoảng cách giữa hai chùm.

ii) Độ rộng chùm có mối quan hệ với các khái niệm được trình bày bởi (1), (2) và (3).

3 PHƯƠNG PHÁP XÂY DỰNG CHÙM

3.1 Phương pháp thứ bậc

a) Bài toán (*Bài toán* 1)

Có *n* phần tử với biến quan sát đã biết. Chúng ta chia những phần tử này thành những chùm với số lượng giảm dần theo từng bước. Tại mỗi bước ta ghép 2 chùm thành 1 chùm mới có độ rộng chùm nhỏ nhất so với việc ghép 2 chùm khác. Ở bước cuối cùng, tất cả các phần tử sẽ được kết hợp thành một chùm. Kết quả thực hiện sẽ được sử dụng để thành lập một cây phân loại.

b) Thuật toán (*Thuật toán* 1)

Bước 1: Bắt đầu với *n* chùm, mỗi chùm chứa một đối tượng. Tính từng đôi độ rộng chùm của hai phần tử. Thành lập ma trận đối xứng *D* của các độ rộng chùm $w(f_i, f_j)$ với *j, i* =1…*n, j* ≠ *i*.

Bước 2: Trong ma trận *D,* tìm độ rộng chùm nhỏ nhất của hai chùm khác nhau, tức hai chùm có sự tương tự nhiều nhất.

Bước 3: Gọi $w(u,v)$ là khoảng cách giữa hai chùm *U* và *V* có sự tương tự nhau nhất. Hợp nhất chùm *U* và *V* thành chùm mới là (*UV*). Tính toán lại ma trận độ rộng chùm mới theo hai bước:

i) Xóa dòng và cột chứa chùm *U* và *V,*
ii) Thêm dòng và cột đại diện cho chùm (*UV*), tìm độ rộng chùm giữa chùm (*UV*) với các chùm còn lại.

Bước 4: Lặp lại bước 2 và bước 3 (lặp lại $n - 1$ lần) cho đến khi các đối tượng được nhóm lại thành một chùm duy nhất.

3.2 Phương pháp không thứ bậc

a) Bài toán (*Bài toán* 2)

Có *n* phần tử với biến quan sát đã biết cần chia những phần tử này thành *k* chùm với *k* cho trước, sao cho một phần tử trong chùm có độ rộng đến chùm nó đang thuộc nhỏ hơn độ rộng đến các chùm khác.

b) Thuật toán (*Thuật toán* 2)

Bước 1: Chia *n* phần tử thành *k* chùm một cách ngẫu nhiên (số lượng phần tử trong mỗi chùm là tùy ý).

Bước 2: Tính độ rộng chùm từ mỗi phần tử đến tất cả các chùm. Nếu độ rộng chùm từ một phần tử đến chùm nó đang thuộc là nhỏ nhất thì ta giữ phần tử đó trong chùm ban đầu. Nếu tồn tại một chùm khác mà độ rộng chùm từ phần tử đang xét đến chùm đo là nhỏ nhất thì ta gán phần tử đang xét vào chùm này, bỏ phần tử trong chùm nó đang thuộc. Nếu phần tử được di chuyển đến chùm khác thì cần phải tính lại giá trị trọng tâm của hai chùm mới có sự thay đổi.

Bước 3: Quay lại bước 2 và dừng lại khi ta có *k* chùm, sao cho một phần tử bất kỳ trong chùm có khoảng cách đến chùm nó đang thuộc nhỏ hơn khoảng cách đến các chùm khác.

4 VẤN ĐỀ TÍNH TOÁN VÀ VÍ DỤ ÁP DỤNG

4.1 Vấn đề ước lượng hàm mật độ xác suất từ dữ liệu rời rạc

Trong thực tế, hầu như mọi dữ liệu có nhu cầu phân tích chùm là dữ liệu rời rạc, do đó để phân tích chùm các hàm mật độ xác suất có ý nghĩa thật sự, việc đầu tiên phải làm là ước lượng hàm mật độ xác suất từ dữ liệu rời rạc. Có nhiều phương pháp tham số cũng như phi tham số để ước lượng hàm mật độ xác suất. Trong bài viết này, chúng tôi sử dụng phương pháp hàm hạt nhân, một phương pháp có nhiều ưu điểm nhất hiện nay.

Gọi $\{x_1, x_2, ..., x_N\}$ là các dữ liệu rời rạc *n* chiều cần ước hàm mật độ xác suất. Hàm mật độ xác suất cần ước lượng theo phương pháp hạt nhân có dạng

$$\hat{f}(x) = \frac{1}{N} \frac{1}{h_1 h_2 ... h_N} \sum_{i=1}^{N} \prod_{j=1}^{n} K_j \left(\frac{x - x_{ij}}{h_j} \right) \quad (8)$$

Trong đó h_j là tham số trơn cho biến thứ *j* và K_j là hàm hạt nhân của biến thứ *j*.

Có rất nhiều bàn luận về việc chọn tham số trơn, nhưng cũng không có việc chọn nào là tối ưu. Khi chọn tham số trơn nhỏ thì hàm mật độ ước lượng sẽ không được trơn, nhưng khi tham số trơn lớn sẽ làm giảm tính chính xác của ước lượng. Tham số trơn đóng vai trò quan trọng trong ước lượng. Trong bài viết này chúng tôi chọn tham số trơn theo Scott (1992):

$$h_j = \left(\frac{4}{N(n+2)} \right)^{\frac{1}{n+4}} \sigma_j \qquad (9)$$

với σ_j là độ lệch chuẩn mẫu của biến thứ *j*.

Theo Webb có nhiều hàm hạt nhân khác nhau được đề nghị như dạng tam giác, hình chữ nhật, song lượng, Epanechnikov,... Ở đây chúng tôi chọn hàm hạt nhân dạng chuẩn:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right) \qquad (10)$$

Sử dụng phần mềm Matlab, chúng tôi đã viết các chương trình ước lượng hàm mật độ xác suất như sau:

a) Chương trình 1: Ước lượng hàm mật độ xác suất một chiều

```
function fa=uocluong(dla);
    syms x;
    fa=sym('x');
    sa=sym('x');
    ha=1.06*std(dla)*(length(dla)^-0.2);
    sa=0;
    for i=1:length(dla)
    sa=sa+1/(2*pi)^.5*exp(-(((x-dla(1,i))/ha)^2/2));
    end;
     sa;
    fa=(1/ha/length(dla)*sa);
```

Khi cần ước lượng hàm mật độ xác suất của một tổng thể nào đó ta chỉ cần sử dụng lệnh:

```
syms x
uocluong([dữ liệu cần ước lượng])
```

b) Chương trình 2: Ước lượng hàm mật độ xác suất hai chiều.

```
function f=uocluong2(dl1,dl2) %dl1, dl2 lần lượt là hai chiều của dữ liệu
    syms x1 x2
    s=sym('s(x1,x2)');
    f=sym('f(x1,x2)');
    h1=std(dl1)/(length(dl1))^(1/6);
    h2=std(dl2)/(length(dl2))^(1/6);
    s=0;
    for i=1:length(dl1)
    s=s+(1/(2*pi)^.5*exp(-(((x1-dl1(1,i))/h1)^2/2)))*(1/(2*pi)^.5*exp(-(((x2-
    dl2(1,i))/h2)^2/2)));
```

```
end;
f=1/(length(dl1)*h1*h2)*s;
```

Khi cần ước lượng hàm mật độ xác suất của một tổng thể nào đó ta chỉ cần sử dụng lệnh:

```
syms x1 x2 ;
uocluong2([chiều thứ nhất],[chiều thứ hai])
```

4.2 Tính độ rộng chùm

Khi có được các hàm mật độ xác suất, để thực hiện bài toán phân tích chùm vấn đề quan trọng là phải tính được độ rộng chùm. Giải quyết vấn đề này là một việc không dễ dàng, bởi vì chúng ta phải xác định hàm cực đại của các hàm mật độ xác suất và phải tính được tích phân trên R^n của hàm cực đại này. Chương trình tìm biểu thức giải tích cụ thể cho hàm cực đại của các hàm mật độ xác suất một chiều để từ đó tính độ rộng chùm đã được viết, tuy nhiên trường hợp nhiều chiều vẫn chưa được giải quyết. Trong bài viết này, chúng tôi tính độ rộng chùm dựa trên việc tính gần đúng tích phân hàm cực đại bằng phương pháp Moncte- Carlo.

Sử dụng cách tính gần đúng hàm cực đại của các hàm mật độ xác suất bằng phương pháp Moncte-Carlo, chúng tôi đã viết các chương trình tính độ rộng chùm các hàm mật độ cho trường hợp một chiều, cũng như nhiều chiều. Sau đây là một chương trình minh họa cho trường hợp hai chiều với ba tổng thể:

Chương trình 3: Tính độ rộng chùm

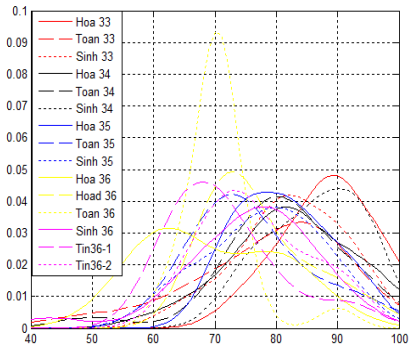
```
syms i x1 x2 y1 y2 gtmax;
fx=sym('f(x1,x2)');
fy=sym('f(y1,y2)');
fz=sym('f(z1,z2)');
fx=uocluong2([ dữ liệu mẫu 1]);
fy=uocluong2([ dữ liệu mẫu 2]);
fz=uocluong2([ dữ liệu mẫu 3]);
f=[fx fy fz];
a1=[chiều thứ nhất của dãy điểm mô phỏng ngẫu nhiên];
a2=[chiều thứ hai của dãy điểm mô phỏng ngẫu nhiên];
for i=1:length(a1)
        gtmax(1,i)=max(subs(subs(f,x1,a1(1,i)),x2,a2(1,i)));
end
gtmax;
double(gtmax)
gttp=sum(gtmax)/length(a1)*(max(a1)-min(a1))*(max(a2)-min(a2));
        drchum = gttp-1;
double(drchum)
```

4.3 Ví dụ áp dụng

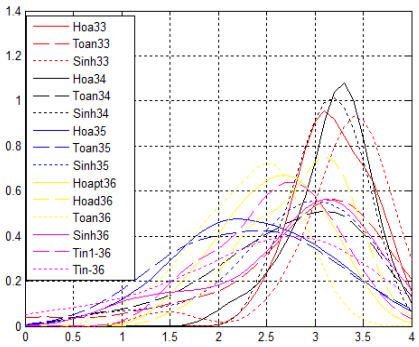
Trong phần này, chúng tôi sẽ tiến hành phân tích chùm điểm học tập, chùm điểm rèn luyện cũng như chùm tổng hợp điểm học tập và điểm rèn luyện ở học kỳ I năm học 2010 -2011 của sinh viên 15 lớp thuộc Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ. Mục đích của việc nghiên cứu này là xem xét mức độ tương đồng về

hai điểm số của sinh viên trong Khoa để có những nhận xét về tình hình học tập và rèn luyện của sinh viên cũng như mức độ đánh giá các ngành học của Thầy Cô trong Khoa. Khoa Khoa học Tự nhiên hiện có 15 lớp: Toán ứng dụng (K33, K34, K35, K36), Hóa học (K33, K34, K35, K36), Hóa dược (K36), Sinh học (K33, K34, K35, K36), Tin học 1 và 2 (K36). Sau khi có số liệu được cung cấp bởi Phòng Công tác sinh viên, chúng tôi chọn ngẫu nhiên mỗi lớp 20 sinh viên, lấy điểm rèn luyện và điểm học tập đưa vào tập dữ liệu (số liệu cụ thể được cho trong phần phụ lục) và chuẩn hóa điểm học tập về thang điểm 100 như điểm rèn luyện.

Gọi $f_1, f_2, f_3, ..., f_{15}$ lần lượt là các hàm mật độ xác suất được ước lượng từ điểm trung bình học tập đã được tổng hợp ở phần trên của sinh viên các lớp Hóa K33, Toán K33, Sinh K33,…, Tin học 2-K36. Sử dụng chương trình 1 và chương trình 2, ta ước lượng được 15 hàm mật độ xác suất cho điểm học tập và rèn luyện mà chúng được minh họa bởi các đồ thị (vẽ trong Matlab) như hình 1 và hình 2 sau:



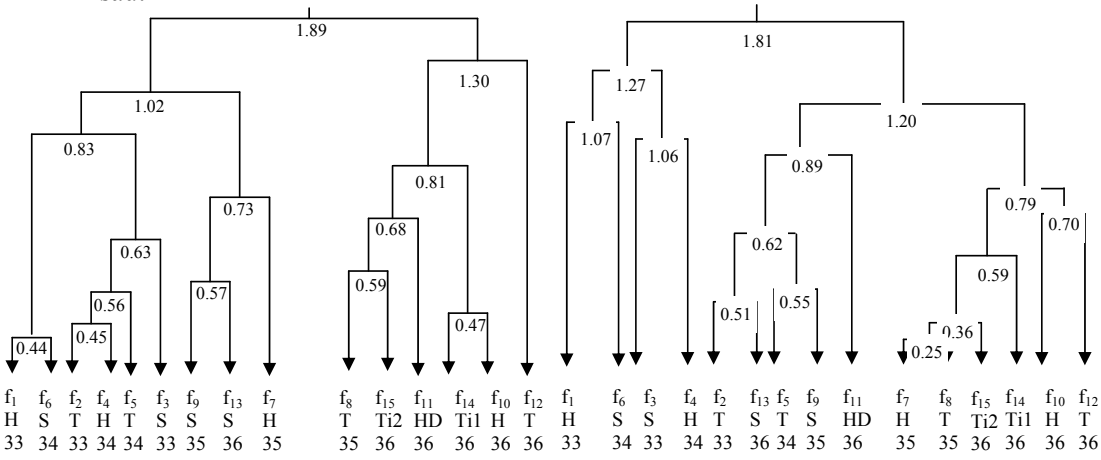
Hình 1: Đồ thị 15 hàm mật độ xác suất điểm rèn luyện



Hình 2: Đồ thị 15 hàm mật độ xác suất điểm học tập

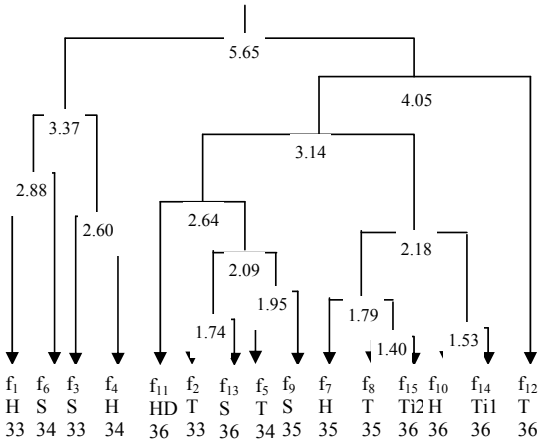
a) Kết quả của phương pháp thứ bậc

Qua 14 bước tính toán cho mỗi trường hợp, các cây phân loại được thành lập như sau:

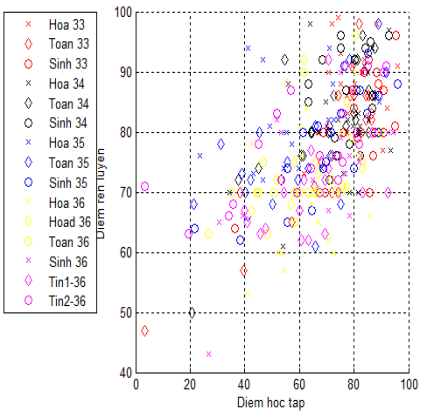


Hình 3: Cây phân loại điểm rèn luyện

Hình 4: Cây phân loại điểm học tập



Hình 5: Cây phân loại điểm học tập và rèn luyện



Hình 6: Tập các điểm học tập và rèn luyện trên mặt phẳng tọa độ

b) Kết quả của phương pháp không thứ bậc

Chia kết quả học tập cũng như điểm rèn luyện của các lớp thành 4 chùm một cách tùy ý: Chùm 1={*Hóa 33, Toán 33, Sinh 33*}, chùm 2 = {*Hóa 34, Toán 34, Sinh 34*}, n chùm 3 = {*Hóa 35, Toán 35, Sinh 35},* chùm 4 = {*Hóa 36, Hóa Dược 36, Toán 36, Sinh 36, Tin1 36, Tin2 36*}.

i) Qua 13 vòng lặp, ta có kết quả 4 chùm về điểm học tập như sau:

Chùm 1 = {*Hóa 33, Sinh 33*}, chùm 2 = {*Hóa 34, Sinh 34*}, chùm 3 = {*Hóa 35, Toán 35, Hóa 36, Toán 36, Tin1 36, Tin2 36*}, chùm 4 = {*Toán 33, Toán 34, Sinh 35, Hóa Dược 36, Sinh 36*}.

ii) Qua 10 vòng lặp, ta có kết quả 4 Chùm điểm rèn luyện như sau:

Chùm 1 ={*Hóa 33, Toán 33, Sinh 34, Hóa 36*}, chùm 2 ={*Sinh 33, Hóa 34, Toán 34, Hóa 35, Sinh 35, Sinh 36*}, chùm 3 ={*Toán 35, Hóa dược 36, Tin1 36, Tin2 36*}, chùm 4 ={*Toán 36*}.

iii) Qua 12 vòng lặp ta có kết quả 4 chùm kết hợp điểm học tập và rèn luyện như sau:

Chùm 1 = {*Hóa 33, Sinh 33*}, chùm 2 = {*Hóa 34, Sinh 34*}, chùm 3 ={*Toán 33, Toán 34, Hóa 35, Sinh 35, Hóa Dược 36, Sinh 36, Tin2-36* }, chùm 4 = {*Toán 35, Hóa 36, Toán 36, Tin1-36* }.

Nhận xét:

i) Hình 4 cho thấy kết quả điểm rèn luyện của sinh viên các lớp thuộc Khoa Khoa học Tự nhiên chia làm hai chùm rõ rệt. Chùm A gồm các lớp khóa 33, 34 và 35 (chỉ có Sinh 36 là ngoại lệ), chùm B gồm các lớp khóa 36 (chỉ có Toán 35 là ngoại lệ). Trong chùm A lại chia thành hai chùm nhỏ: một chùm gồm các lớp khóa 33 và 34 và chùm còn lại gồm các lớp khóa 35. Kết quả phân tích của phương pháp không thứ bậc cũng gần giống với phương pháp thứ bậc, trong đó, hợp của chùm 1 và chùm 2 của phương pháp không thứ bậc gần giống với chùm A của phương pháp thứ bậc (có thêm lớp Hóa 36), và hợp của chùm 3 và chùm 4 thì gần giống chùm B của phương pháp thứ bậc (không có lớp Hóa 36). Một trường hợp đáng quan tâm khác là điểm rèn luyện của lớp Toán 36 được tách hẳn thành một

chùm riêng biệt, điều này cho thấy điểm rèn luyện của lớp Toán 36 có sự khác biệt rất lớn so với các lớp khác, nhìn vào đồ thị các hàm mật độ xác suất hình 1, chúng ta thấy điểm rèn luyện của lớp Toán 36 đạt trung bình vào khoảng 70 và có độ phân tán không cao, nghĩa là hầu như đều xấp xỉ 70 điểm, trong khi các lớp khác có số điểm trung bình khác và có độ phân tán cao hơn điểm rèn luyện của lớp Toán 36. Như vậy, điểm rèn luyện của các lớp Khoa Khoa học Tự nhiên phụ thuộc vào khóa học một cách rõ rệt, cụ thể các lớp khóa 36 có sự khác biệt với các lớp khóa cũ, trong các lớp khóa cũ thì khóa 35 lại khác với khóa 33 và 34.

ii) Kết quả phân tích điểm học tập của phương pháp thứ bậc lại chỉ ra điểm trung bình của các lớp Hóa, Sinh các năm cuối (Hóa 33, Sinh 33, Hóa 34, Sinh 34) có sự khác biệt đối với các lớp còn lại. Kết quả chia chùm của phương pháp không thứ bậc cũng chỉ ra điều đó, cụ thể hợp của chùm 1 và chùm 2 của phương pháp không thứ bậc chính là chùm các lớp Hóa, Sinh những năm cuối, chùm 3 gồm một số lớp năm một và năm hai, những lớp còn lại được phân vào chùm 4. Kết quả trên cho thấy điểm học tập một phần phụ thuộc vào ngành học, song song đó, một phần cũng phụ thuộc vào khóa học. Kết quả phân tích đồng thời hai biến điểm học tập và rèn luyện lại là sự tổng hợp các kết quả đã có ở trên, các lớp Hóa, Sinh các khóa 33 và 34 được nhóm thành một chùm, trong chùm còn lại thì lớp Toán 36 lại có sự khác biệt so với các lớp khác.

5 KẾT LUẬN

Trong thời đại thông tin, nhiều đối tượng của nhiều lĩnh vực ngày càng phải tiếp nhận và xử lý nhiều loại dữ liệu đa dạng, vì vậy nhu cầu thực hiện bài toán phân tích chùm ngày càng được ứng dụng trong nhiều lĩnh vực khác nhau. Tiêu chuẩn độ rộng chùm đã cải thiện việc đánh giá mức độ gần và xa của các phần tử trong xây dựng chùm, vì vậy kết quả phân tích chùm được hợp lý hơn. Với các chương trình đã viết cho việc tính độ rộng chùm, việc xây dựng chùm các hàm mật độ xác suất đã có một bước tiến bộ quan trọng. Chương trình ước lượng hàm mật độ xác suất từ dữ liệu rời rạc giúp ta có thể giải quyết được yêu cầu phân tích chùm từ số liệu quan sát thực tế.

TÀI LIỆU THAM KHẢO

Defays, D. (1977), "An efficient algorithm for a complete link method", *Computer Journal*, 20(4), pp.354–366.

Fukunaga, K., (1990), *Introduction to statistical pattern recognition*, 2nd Ed., Academic Press, New York.

Glick, N., (1973), "Separation and probability of correct classification among two or more distributions", *Annals Inst. Stat Math.* 25, pp.373–382.

Martinez, W.L. and Martinez, A.R., (2008), *Computational statistics handbook with Matlab*, Chapman & Hall/CRC, Boca Raton.

Matusita, K. (1967), "On the notion of affinity of several distributions and some of its applications", *Ann. Inst. Statist. Math.* 19, pp.181–192.

Pham–Gia, T. Turkkan, N. and Tai, Vovan., (2008), "The maximum function in statistical discrimination analysis",*Commun. in Stat–Simulation computation* 37(2), pp. 320 – 336.

Rohlf. F.J., (1982), "Single – link clustering algorithms", i*n P.R. Krishnaiah and L.N. Kanal*, *eds, Handbook of Statistics*, North Holland, Amsterdam, vol.2, pp. 267–284.

Scott, David W. (1992), *Mutivariate density estimation: theory, practice and visualization*, John Wiley & Son, New York.

Sibson, R., "Slink: an optimally efficient algorithm for the single – link cluster method", *Computer Journal* 16(1), pp. 20–34.

Tai, VoVan., Pham – Gia,T., (2010), "Clustering probability distributions", *Journal of applied statistics,* 37(11), pp. 1891-1910.

Toussaint G.T., (1972), "Some inequalities between distance measures for feature evaluation", *I.E.E.E Trans. Comput.* 21, pp.409-429.

Webb, A., (2002), *Statistical pattern recognition*, 2nd Ed., John Wiley & Sons, New York.

PHỤ LỤC

Bảng kết quả chọn mẫu điểm học tập và điểm rèn luyện của 20 sinh viên được chọn ngẫu nhiên từ các lớp của Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ

Hóa 33		Toán 33		Sinh 33		Hóa 34		Toán 34		Sinh 34		Hóa 35		Toán 35		Sinh 35		Hóa 36		Hóa Dược 36		Toán 36		Sinh 36		Tin1-36		Tin2-36	
X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
3.84	91	3.50	86	3.82	96	3.26	80	3.08	81	3.21	92	0.94	76	3.14	92	0.87	64	3.00	80	3.33	74	2.53	70	3.07	73	3.33	70	2.97	72
3.21	88	3.46	82	1.46	64	3.05	73	1.80	74	2.53	88	2.84	94	2.85	76	3.00	83	2.83	63	2.27	68	1.07	63	2.60	72	3.70	70	0.13	71
2.97	99	3.45	83	3.30	80	3.29	80	0.83	50	3.47	84	2.24	88	1.68	72	3.39	83	1.63	53	2.93	70	2.07	72	3.17	87	3.07	91	3.40	91
3.09	86	2.73	75	3.50	80	3.11	82	3.23	82	3.37	94	3.53	86	3.00	68	3.41	93	2.40	81	2.33	75	2.33	65	2.13	80	1.63	65	1.43	68
3.11	83	2.97	86	3.64	80	3.13	70	2.13	72	3.00	94	1.38	70	2.38	72	2.22	65	2.17	57	3.10	71	2.67	70	2.80	81	2.83	74	2.37	70
3.53	98	2.62	71	2.75	80	3.41	90	3.20	83	3.00	96	2.73	74	1.25	78	3.56	80	2.50	84	3.10	71	2.40	70	3.13	76	2.77	72	0.77	63
2.67	80	0.13	47	3.20	84	2.68	80	3.42	87	3.63	90	2.21	80	1.56	73	2.45	80	1.70	78	3.27	76	1.80	70	2.07	82	1.83	63	1.37	66
3.32	90	3.17	91	3.37	94	3.26	83	2.83	80	2.94	76	2.39	74	2.53	74	2.79	75	3.20	81	2.93	86	1.87	75	1.23	65	2.57	77	2.57	74
3.53	98	3.47	86	3.53	90	2.16	61	2.58	80	3.24	87	1.87	92	1.61	70	3.84	88	3.43	66	2.63	75	2.70	70	3.70	80	2.17	70	2.13	72
3.58	87	3.28	98	3.48	76	2.97	82	2.84	81	2.92	79	2.17	74	3.56	98	3.56	86	2.90	60	3.03	71	3.00	70	3.30	76	2.97	70	3.67	91
2.88	98	1.53	70	3.43	70	3.71	97	2.45	76	3.47	86	1.97	81	0.85	68	2.75	74	2.13	60	2.47	92	2.37	70	3.00	79	3.47	80	1.80	78
3.69	84	3.38	82	3.42	92	3.74	77	3.50	86	3.44	95	2.89	88	2.63	81	2.58	67	3.00	80	3.60	80	1.93	63	1.57	66	2.53	62	3.03	75
3.00	93	1.58	57	3.17	87	2.56	98	3.50	94	3.29	76	2.29	78	1.65	66	3.06	91	2.13	60	3.17	79	1.40	70	3.00	90	1.60	67	2.70	76
2.47	73	3.15	86	2.77	70	3.42	82	3.42	80	3.08	87	1.86	72	1.82	80	2.86	80	3.23	90	3.23	96	1.77	70	3.77	80	1.90	64	3.63	90
3.35	92	2.88	70	3.63	87	3.35	91	2.18	92	2.84	72	1.65	94	2.91	86	3.29	87	2.67	70	3.33	75	2.70	70	2.83	67	2.40	67	2.07	83
3.17	91	3.20	86	3.56	88	3.39	93	3.25	92	3.14	78	2.83	82	2.00	68	1.53	62	2.83	62	2.33	70	2.03	70	3.27	80	2.43	62	2.97	76
3.10	80	3.42	86	3.80	81	3.29	81	2.43	76	3.71	96	1.72	73	2.33	75	3.03	75	2.10	60	1.70	67	2.63	70	3.33	90	3.23	80	3.03	83
3.62	77	3.38	86	3.25	79	3.26	70	2.57	80	2.74	92	3.18	81	3.67	90	3.20	74	2.57	68	2.23	88	2.87	72	3.13	66	2.83	92	2.27	87
3.00	88	2.29	65	3.35	90	3.47	94	3.33	96	3.24	80	3.57	85	2.22	75	2.22	74	2.43	76	3.10	85	2.57	74	3.27	75	2.47	72	3.27	78
2.83	90	2.88	76	3.04	80	2.79	85	1.50	72	2.53	85	2.91	82	2.63	61	2.67	81	2.67	65	2.67	73	2.47	90	1.07	43	2.77	63	2.70	70

X: điểm học tập.
Y: điểm rèn luyện.