BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

HỒ MINH ĐÍCH

NGHIÊN CỨU GIẢI THUẬT DI TRUYỀN ỨNG DỤNG VÀO GIẢI MỘT SỐ BÀI TOÁN THỐNG KÊ

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2011

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS. Lê Văn Sơn**

Phản biện 1: **TS. Huỳnh Hữu Hưng**

Phản biện 2: **PGS.TS. Đoàn Văn Ban**

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 15 tháng 10 năm 2011

** Có thể tìm hiểu luận văn tại:*
- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng.

MỞ ĐẦU

1. Lý do chọn đề tài

Trong những năm gần đây, kỹ thuật lập trình tiến hóa là một trong những kỹ thuật lập trình rất phát triển trong lĩnh vực trí tuệ nhân tạo. Một công thức tương tự với công thức nổi tiếng của N.Wirth đưa ra trong lập trình cấu trúc được áp dụng cho kỹ thuật lập trình tiến hóa:

Cấu trúc dữ liệu + Giải thuật di truyền = chương trình tiến hóa

Thuật ngữ chương trình tiến hóa là một trong những khái niệm được dùng để chỉ các chương trình máy tính có sử dụng thuật toán tìm kiếm và tối ưu hóa dựa trên "nguyên lý tiến hóa tự nhiên". Ta gọi chung các thuật toán như vậy là thuật toán tiến hóa. Có một số thuật toán tiến hóa được công bố:

- Quy hoạch tiến hóa – EP, do D.B.Pogel đề xuất.
- Chiến lược tiến hóa, do T.Baeck, F.H.Hofmeister và H.P.Schwefel đề xuất.
- Thuật giải di truyền, do D.E.Golberg đề xuất, được L.Davis và Z.Michalevicz phát triển. Trong phạm vi luận văn chỉ nghiên cứu lập trình tiến hóa thông qua giải thuật di truyền và ứng dụng vào giải quyết hai lớp bài toán phân tích dữ liệu thống kê.

2. Đối tượng và phạm vi nghiên cứu

2.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài gồm:

- Giải thuật di truyền
- Phân lớp dữ liệu bằng các hàm phân biệt tuyến tính
- Phân tích hồi qui

2.2. Phạm vị nghiên cứu

Ứng dụng giải thuật di truyền để thiết kế giải thuật tìm giá trị Min (Max) của hàm nhiều biến làm công cụ để giải các bài toán thống kê đề ra trong luận văn. Cụ thể là hai bài toán:

- Bài toán phân tích dữ liệu hồi qui tuyến tính.
- Bài toán phân lớp dữ liệu bằng tập các hàm phân biệt tuyến tính.

3. Mục đích đề tài

Mục đích của đề tài là muốn tìm một cách tiếp cận mới bằng thuật giải di truyền để giải một số lớp bài toán thuộc lĩnh vực thống kê, đồng thời cũng muốn chứng minh tính năng vượt trội của giải thuật di truyền trong việc tìm lời giải cho nhiều dạng bài toán khác nhau.

4. Mục tiêu, ý nghĩa đề tài

Nghiên cứu và ứng dụng giải thuật di truyền vào hai lớp bài toán thuộc lĩnh vực thống kê là bài toán hồi quy tuyến tính và bài toán phân lớp dữ liệu dựa trên các hàm phân loại tuyến tính. Kết quả của bài toán mang lại vừa có tính năng của một hệ thống máy học, giúp dự báo, tính toán, phân lớp các dữ liệu không được học vừa có ý nghĩa đề xuất và đạt được kết quả khả quan về một phương pháp phân lớp dữ liệu cũng như việc thiết lập các mô hình toán học và phân tích tương quan cho các số liệu thực nghiệm dùng trong nghiên cứu khoa học.

Đối với thuật giải di truyền, ý tưởng xuyên suốt nhất của nó là mô phỏng quá trình tiến hóa tự nhiên để áp dụng tìm kiếm lời giải cho một bài toán trên máy tính.

Việc áp dụng giải thuật di truyền để giải quyết hai lớp bài toán nói trên là một phương pháp tiếp cận mới, tinh tế để giải quyết một số lớp bài toán trong lĩnh vực thống kê là những bài toán tốn rất nhiều công sức cho thao tác tính toán để tìm ra lời giải cho bài toán.

5. Cấu trúc luận văn

Nội dung chính của luận văn được trình bày trong 4 chương :

Chương 1. Cơ sở lý thuyết về giải thuật di truyền

Chương 2. Ứng dụng giải thuật di truyền tìm cực trị của hàm nhiều biến

Chương 3. Phân lớp dữ liệu bằng các hàm phân biệt tuyến tính

Chương 4. Bài toán hồi quy

CHƯƠNG 1. CƠ SỞ LÝ THUYẾT VỀ THUẬT GIẢI DI TRUYỀN

1.1. KHÁI NIỆM

Giải thuật di truyền(GA) là giải thuật tìm kiếm, chọn lựa các giải pháp tối ưu để giải quyết các bài toán thực tế khác nhau, dựa trên cơ chế chọn lọc của di truyền học: từ tập lời giải ban đầu, thông qua nhiều bước tiến hoá, hình thành

tập lời giải mới phù hợp hơn, và cuối cùng tìm ra lời giải tối ưu nhất. Giải thuật di truyền dựa trên quan điểm cho rằng quá trình tiến hoá của tự nhiên là quá trình hoàn hảo nhất, hợp lý nhất và tự nó đã mang tính tối ưu.

Ý tưởng chính của giải thuật di truyền là thay vì chỉ phát sinh một lời giải ban đầu chúng ta sẽ phát sinh một lúc nhiều lời giải cùng lúc. Sau đó, trong số lời giải được tạo ra, chọn ra những lời tốt nhất để làm cơ sở phát sinh ra nhóm các lời giải sau với nguyên tắc càng về sau càng tốt hơn. Quá trình cứ thế tiếp diễn cho đến khi tìm được lời giải tối ưu hoặc xấp xỉ tối ưu.

1.2. GIẢI THUẬT DI TRUYỀN

1.2.1 Định nghĩa :

GA được định nghĩa là một bộ 7: GA=($I, \Psi, \Omega, s, t, \mu, \lambda$) :

- I=B^t: Không gian tìm kiếm lời giải của bài toán.
- Ψ :I $\rightarrow$ R: Ký hiệu của hàm thích nghi (Eval function).
- Ω : Ký hiệu cho tập các phép toán di truyền.
- S: $I^{\mu+\lambda} \rightarrow I^{\mu}$ ký hiệu cho thao tác chọn; giữ lại μ cá thể.
- t: $I^{\varpi} \rightarrow \{True, false\}$ là tiêu chuẩn dừng.
- μ, λ : lần lượt là số cá thể trong thế hệ cha mẹ và thế hệ con cháu.

1.2.2. Những quá trình tiến hóa của giải thuật :

1.2.2.1. Quá trình lai ghép (Cross Over):

▪ Phép lai: Là quá trình hình thành nhiễm sắc thể mới trên cơ sở các nhiễm sắc thể cha mẹ bằng cách ghép một hay nhiều đoạn gen của hai (hay nhiều) nhiễm sắc thể cha-me với nhau, phép lai được thực hiện với xác suất p_c

1.2.2.2. Quá trình tái sinh (Preproduction) và lựa chọn (Selection):

▪ Tái sinh: Là quá trình trong đó các cá thể được sao chép dựa trên cơ sở độ thích nghi của nó.

▪ Phép lựa chọn: Là quá trình loại bỏ các cá thể xấu trong quần thể, chỉ giữ lại trong quần thể các cá thể tốt

1.2.2.3. Quá trình đột biến (Mutation):

Đột biến là hiện tượng cá thể con mang một số tính trạng không có trong mã di truyền của cha-mẹ.

1.2.3. Tổng quát về giải thuật di truyền :

```
Begin
    t:=0 ;
    Khởi tạo tập cá thể P(0) = {a_1(0), a_2(0), ..., a_μ(0)} ∈ I^μ;
    Đánh giá độ thích nghi cho các cá thể thuộc P(t) ;
        P(0) := {Ψ(a_1(0)), Ψ(a_2(0)), ..., Ψ(a_μ(0))};
    While (chưa thỏa điều kiện dừng) do
    Begin
    t:=t+1;
    Chọn cá thể  P'(t) từ P(t);
    Lai  Q(t) từ P(t-1);
    Đột biến R(t) từ P(t-1);
    P(t):=P(t-1) ∪ Q(t) ∪ R(t);
    Lượng Giá   P(t) := {Ψ(a_1(t)), Ψ(a_2(t)), ..., Ψ(a_n(t)), ..., Ψ(a_{n+λ}(t))};
    Sắp xếp P(t) theo thứ tự Ψ(a_1(t)) giảm dần;
    Loại bỏ λ cá thể cuối cùng ;
    End;
End;
```

Hình 1.1 Giải thuật di truyền tổng quát

1.2.4. Tính hội tụ trong giải thuật di truyền

Cho GA=($I, \Psi, \Omega, s, t, \mu, \lambda$) nếu các điều kiện sau thỏa:

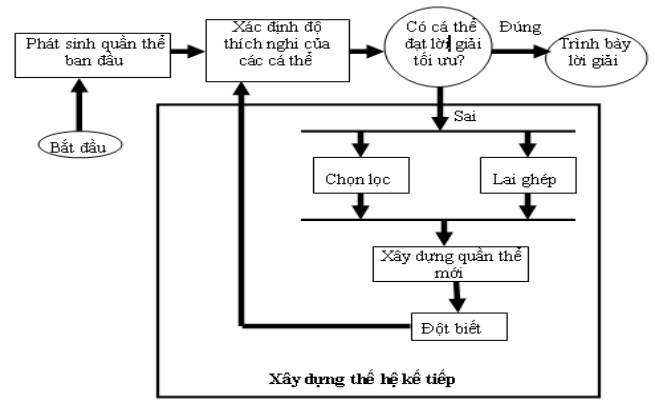
- I là không gian hữu hạn, đếm được;
- Lời giải tối ưu a* ∈ I

Thì giải thuật sẽ dừng và lời giải tìm được chính là lời giải tối ưu **a***

1.2.5. Nguyên lý hoạt động của của giải thuật :

- ***Bước 1***: Chọn một số tượng trưng cho toàn bộ các lời giải
- ***Bước 2***: Chỉ định cho mỗi lời giải một ký hiệu. Ký hiệu có thể là một dãy các bits 0, 1 hay dãy số thập phân
- ***Bước 3***: Tìm hàm số thích nghi và tính hệ số thích nghi
- ***Bước 4***: Tực hiện tái sinh và chọn.
- ***Bước 5***: Tính hệ số thích nghi cho các cá thể mới, iữ lại một số nhất định các cá thể tương đối tốt.

- ***Bước 6***: Nếu chưa tìm được lời giải tối ưu hay tương đối tốt nhất, quay lại bước 4 để tìm lời giải mới.
- ***Bước 7:*** Kế thúc giải thuật và báo cáo kết quả tìm được.



Hình 1.2 Sơ đồ tổng quát của giải thuật di truyền

1.2.6. Xây dựng mô hình giải thuật di truyền nâng cao :

Begin
t:=0;
Khởi tạo $P(0) = \{a_1(0), a_2(0), ..., a_\mu(0)\} \in I^\mu$;
Tính độ thích nghi cho các cá thể thuộc P(t)
$P(0) := \{\Psi(a_1(0)), \Psi(a_2(0)), ..., \Psi(a_\mu(0))\}$;
Khi *(chưa thỏa điều kiện dừng)* ***lặp lại***
t:=t+1;
Tái sinh *P'(t) từ P(t);*
Loại s cá thể có độ thích nghi kém nhất;
Chép $\frac{s}{2}$ *cá thể tốt nhất ở thế hệ (t-1) vào thế hệ t;*
Khởi tạo ngẫu nhiên $\frac{s}{2}$ *cá thể bổ sung cho thế hệ (t);*
Lai *Q(t) từ P(t-1);*
Đột biến *R(t) từ P(t-1);*
P(t):=P(t-1) $\cup$ *Q(t)* $\cup$ *R(t);*
Lượng giá $P(t) := \{\Psi(a_1(t)), \Psi(a_2(t)), ..., \Psi(a_n(t)), ..., \Psi(a_{n+s}(t))\}$;
Sắp xếp *P(t) theo thứ tự* $\Psi(a_i(t))$ ***giảm dần;***
Loại bỏ λ *cá thể cuối cùng ;*
Cuối lặp;
End;

Hình 1.3 Mô hình giải thuật di truyền nâng cao

1.3. SỰ KẾT HỢP GIỮA DI TRUYỀN VÀ LEO ĐỒI.

1.3.1 Khái niệm:

Sau khi tìm được lời giải tối ưu của bài toán thì vấn đề còn lại là phải chính xác hóa nghiệm tối ưu vừa tìm được, mà thuật toán leo đồi lại chỉ cho phép tìm được giải pháp tối ưu cục bộ.

1.3.2. Kết hợp di truyền và leo đồi

- ***Bước 1***: Chạy giải thuật di truyền cho đến khi cá thể thế hệ mới không tốt hơn nhiều so với thế hệ trước.
- ***Bước 2***: Gán n cá thể tốt nhất của giải thuật di truyền cho n điểm xuất phát của giải thuật leo đồi.
- ***Bước 3***: Chạy giải thuật leo đồi tìm được lời giải tối ưu

CHƯƠNG 2. ỨNG DỤNG GIẢI THUẬT DI TRUYỀN TÌM CỰC TRỊ CỦA HÀM NHIỀU BIẾN

2.1. ĐẶT VẤN ĐỀ

Hiện nay có rất nhiều phương pháp giải quyết bài toán tối ưu hàm số, nhưng các phương pháp chỉ dừng lại ở những lớp bài toán với những thông tin rõ ràng. Do đó, việc tìm ra một phương pháp mới để giải bài toán tối ưu hàm nhiều biến tổng quát là cần thiết.

Nhưng để giải quyết lớp hai bài toán trong luận văn này thì phải có một công cụ cần thiết phải thiết kế là bài toán tìm cực trị (giá trị Max hay Min) của một hàm số nhiều biến mà mỗi biến có thể nhận các giá trị số nằm trên một miền con hoặc toàn miền số thực (từ $-\infty$ đến $+\infty$).

2.2. BIỂU DIỄN BIẾN

Cho một hàm nhiều biến $y = f(x_1, x_2, ..., x_n)$ với $x_i \in D_i = [a_i, b_i] \subseteq R$.

Để biểu diễn x_i (i=1,…,n) sao cho có thể thực hiện các phép toán di truyền một cách hiệu quả, thì ta biểu diễn x_i bằng chuỗi bit nhị phân.

Giả sử x_i là một số thực có k chữ số thập phân sau dấu chấm. Thì giá trị của x_i là: $x_i = a_i + \text{decimal}(U)\dfrac{b_i - a_i}{2^{m_i} - 1}$

2.3. CÁC GIÁ TRỊ LỰA CHỌN TRONG GIẢI THUẬT DI TRUYỀN

2.3.1. Lựa chọn kích thước của quần thể

Để đảm bảo kích thước quần thể không quá lớn đồng thời cũng giúp tăng hiệu quả và tính chính xác của giải thuật khi hàm số có số biến lớn, thì ta nên chọn kích thước quần thể phụ thuộc vào số biến của hàm số: **μ = 100 +10 *NumVar** (NumVar là số biến của hàm số).

2.3.2. Lựa chọn số lần tiến hóa của giải thuật

Để đảm bảo tính chính xác của giải thuật ta chọn số lần tiến hóa NumGen = 100 + 10 * NumVar (NumVar là số biến của hàm số).

2.3.3. Lựa chọn xác suất lai ghép

Sự kết hợp các lời giải cha mẹ tạo sinh các cá thể mới trong giải thuật di truyền bằng toán tử lai ghép

2.3.4. Lựa chọn xác suất đột biến

Xác suất đột biến $P_M = \frac{1}{GenSize}$.

2.3.5. Lựa chọn khoảng giá trị của các biến

Xác định được khoảng giá trị của x thuộc khoảng [a,b] nào đó. Với lớp bài toán trong luận văn thì mỗi biến x_i sẽ thuộc $[-\infty,+\infty]$. Nhưng trong máy tính, mỗi kiểu dữ liệu được khai báo cho biến có giá trị khác nhau, giá trị ∞ có thể được quy ước bằng giá trị lớn nhất của kiểu dữ liệu đó.

2.4. HÀM ĐO ĐỘ THÍCH NGHI (EVAL FUNCTION)

2.4.1. Ánh xạ giá trị hàm mục tiêu f(x) sang giá trị thích nghi (Eval)

- Nếu bài toán tối ưu là tìm cực tiểu của một hàm đánh giá g(x) thì ta xây dựng như sau:

$$f(x) = \begin{cases} C_{Max} - g(x) & \text{khi } g(x) < C_{Max} \\ 0 & \text{Trong cac truong hop khac} \end{cases}$$

- Nếu bài toán tối ưu là tìm cực đại của một hàm đánh giá g(x) thì ta xây dựng như sau:

$$f(x) = \begin{cases} C_{Min} + g(x) & \text{khi } g(x) + C_{Min} > 0 \\ 0 & \text{Trong cac truong hop khac} \end{cases}$$

Trong đó C_{Max}, C_{Min} là một tham số đầu vào.

2.4.2. Điều chỉnh độ thích nghi

- Gọi G là độ tốt của cá thể, độ thích nghi của cá thể theo phương pháp điều chỉnh tuyến tính được xác định theo quy tắc sau:

F = a * G + b

- Giá trị độ thích nghi cuối cùng này lại nằm trong đoạn[0,1].

2.5 CÁC PHÉP TOÁN DI TRUYỀN

2.5.1. Khởi tạo quần thể ban đầu

```
Begin
        for i:=0 to PopSize-1 do
                for j:=0 to GenSize-1 do
                        QuanThe.CaThe[i][j]:=Flip(0.5);
End;
Flip(0.5) là hàm tạo các ngẫu nhiên 0 và 1 với xác suất 50%
```

Hình 2.3 Đoạn mã giả minh họa cho thao tác khởi tạo quần thể

2.5.2. Phép chọn cá thể (Selection)

Sử dụng phương pháp thông dụng là **quy tắc chọn theo bàn Roulete**. Quá trình này được thực hiện theo các bước:

- ***Bước 1***: Tính độ thích nghi cho từng cá thể trong quần thể.
- ***Bước 2***: Tính tổng độ thích nghi của tất cả các cá thể
- ***Bước 3***: Phát sinh một số ngẫu nhiên p nằm trong khoảng từ 0 đến tổng độ thích nghi của quần thể.
- ***Bước 4***: Trả về cá thể đầu tiên mà độ thích nghi của nó và độ thích nghi của các cá thể khác nhau trong quần thể trước đấy.

2.5.3. Phép lai ghép (CrossOver)

Phép lai được thực hiện trên hai cá thể cha mẹ được chọn với xác suất $p_c \in [0, 1]$ và được thực hiện theo nhiều cách khác nhau.

- Lai ghép đơn điểm
- Lai ghép đa điểm

CHƯƠNG 3. PHÂN LỚP DỮ LIỆU BẰNG CÁC HÀM PHÂN BIỆT TUYẾN TÍNH

3.1 PHÂN LỚP DỮ LIỆU

3.1.1 Khái niệm

Khi phân tích dữ liệu, thường dựa vào một số tiêu chuẩn hay các đại lượng khác nhau về bản chất. Ví dụ: Khi phân loại bệnh nhân mắc một chứng bệnh nào đó thì cần căn cứ vào một số tiêu chuẩn (huyết áp, các xét nghiệm, chụp X quang,…) của người đó và các bệnh đã có. Khi xác định được các dữ kiện liên quan đến những chỉ tiêu đã chọn, chúng ta có thể dự đoán được khả năng chẩn đoán bệnh của bệnh nhân đó đó với độ chính xác cao.

Thay vì nghiên cứu trên tập dữ liệu lớn thì sau khi phân lớp dữ liệu ta sẽ tiến hành phân tích trên các tập dữ liệu nhỏ hơn mà mỗi tập dữ liệu này có một số tính chất đặc thù tùy thuộc vào các chỉ tiêu lựa chọn để phân nhóm tập dữ liệu thu thập ban đầu.

3.1.2. Các bước để giải quyết bài toán phân lớp

Bước 1*:* Học (training).

Bước 2 : Kiểm tra và đánh giá.

3.1.3. Hàm phân lớp tuyến tính

Hàm phân lớp tuyến tính có ranh giới phân lớp là 1 siêu phẳng, vì vậy nó chỉ phân tách được 2 lớp.

Ví dụ: Xét hàm tuyến tính phân tách R^n thành 2 **nửa không gian** (*half-space*).

Để cho tiện, ta gán $w_1 = +1$, $w_2 = -1$, luật phân lớp khi sử dụng hàm phân lớp tuyến tính là:

$$f(x) = \theta((w, x) - b) \quad (3.1)$$

$$\theta(t) = \begin{cases} +1, & t \geq 0 \\ -1, & t < 0 \end{cases} \quad (3.2)$$

Trong đó, f(x) là **hàm phân lớp**, θ(t) là **hàm ngưỡng** (*threshold function*), (w, x) là *tích vô hướng* của w, x, w là **trọng số** (*weight*) trên các tọa độ/đặc trưng của x, b là **ngưỡng** (*threshold*).

3.2. HÀM PHÂN BIỆT TUYẾN TÍNH VÀ MẶT QUYẾT ĐỊNH

3.2.1. Định nghĩa:

Hàm phân biệt tuyến tính là một hàm số nhận một vector đầu vào x và gán nó cho một trong c lớp. Hàm phân biệt tuyến có dạng:

$$g(x) = W_0 + W_1X_1 + W_2X_2 + ... + W_kX_k = W_0 + \sum_{i=1}^{k} W_iX_i = W^tX + W_0 \quad (3.3)$$

Trong đó:

- $W = (W_1, W_2, ..., W_k)$ là **vectơ trọng số** và W_0 được gọi là **trọng số nền** hay **ngưỡng**.
- $X = (X_1, X_2, ... X_k)$ là các **biến độc lập**.

3.2.2. Trường hợp phân hai lớp

Nếu loại dữ liệu phân thành hai lớp thì phương trình (1) trở thành :

$$g(X) = W_0 + W_1X_1 \quad (3.4)$$

Dựa vào hàm phân biệt (2) sự phân chia dữ liệu thành hai lớp được thực hiện dựa trên quyết định sau: **Quyết định là thành phần dữ liệu thuộc vào W_1 nếu ta có g(X) > 0 và quyết định là W_2 nếu g(X) < 0.** Trường hợp g(X)= 0

$$W^tX_1 + W_0 = W^tX_2 + W_0 \text{ hay } W^t(X_1 - X_2) = 0 \quad (3.5)$$

Do đó khi g(X) > 0 thì X được gán đến W_1 (X nằm trên R_1), ngược lại thì X được gán đến W_2 (X nằm trên R_2). Khi X thuộc R_1 thì ta có thể nói **X thuộc phần dương của H** và khi X thuộc R_2 thì ta có thể nói **X thuộc phần âm của H**.

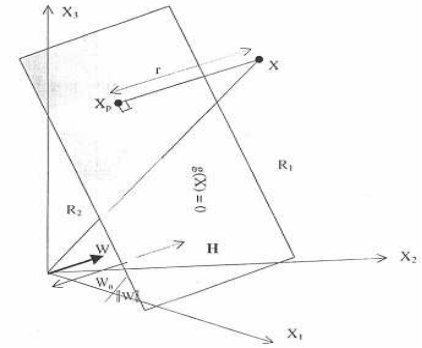
Hàm phân biệt tuyến tính g(X) chỉ ra khoảng cách đại số từ X đến siêu phẳng H. Vì vậy, có lẽ cách đơn giản nhất là biểu diễn X theo biểu thức sau:

$$X = X_p + r\frac{W}{\|W\|} \quad (3.6)$$

Trong đó:

- X_p là hình chiếu chuẩn của X trên H.

- r là khoảng cách đại số từ X đến siêu phẳng H.



Hình 3.2 Mặt quyết định tuyến tính H xác định bởi g(X) = WtX + W0, và chia không gian thành 2 nửa không gian R1(g(X)>0) và R2(g(X)<0)

Mà $g(X_p) = 0$ nên ta có:

$$g(X) = W^t X + W_0 = r\|W\| \qquad \text{hay} \qquad r = \frac{g(X)}{\|W\|} \tag{3.7}$$

3.2.3. Trường hợp phân nhiều lớp

Khi dữ liệu chia thành c lớp, ta sẽ chia không gian đặc tính thành tối đa $\frac{c^*(c-1)}{2}$ mặt quyết định, thì sẽ ta có $\frac{c^*(c-1)}{2}$ hàm phân biệt tuyến tính, mỗi hàm dùng cho một phần của sự phân lớp.

3.2.4. Tổng quát hóa các hàm phân biệt tuyến tính

Hàm phân biệt tuyến tính g(X) có thể viết dưới dạng :

$$g(X) = W_0 + \sum_{i=1}^{k} W_i X_i \tag{3.8}$$

Trong đó W_i là các thành phần của vectơ trọng số W

Việc tổng quát hóa các hàm phân loại có lợi ích là có thể viết g(X) dưới dạng thuần nhất dựa vào $a^t y$.

Trường hợp đặc biệt:

$$g(X) = W_0 + \sum_{i=1}^{k} W_i X_i = \sum_{i=0}^{k} W_i X_i \tag{3.9}$$

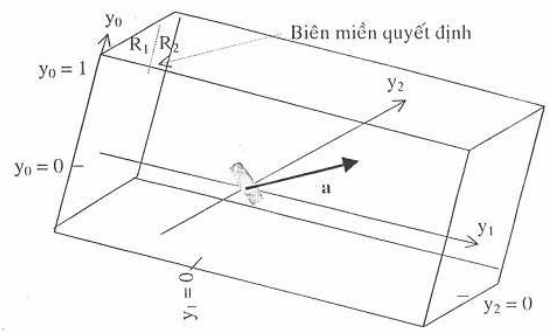
Nếu đặc $X_0 = 1$ thì ta có thể viết:

$$y = \begin{bmatrix} 1 \\ X_1 \\ \dots \\ X_k \end{bmatrix} = \begin{bmatrix} 1 \\ X \end{bmatrix} \quad (3.10)$$

y được gọi là vectơ gia tăng đặc tính.

Tương tự vectơ gia tăng trọng số có thể được viết dưới dạng sau:

$$a = \begin{bmatrix} W_0 \\ W_1 \\ \dots \\ W_k \end{bmatrix} = \begin{bmatrix} W_0 \\ W \end{bmatrix} \quad (3.11)$$



Hình 3.5 Minh họa chuyển đổi toạ độ từ không gian X-2 sang không gian y-3 chiều

3.3. TỔNG BÌNH PHƯƠNG SAI SỐ TỐI TIỂU

3.3.1 Trong trường hợp phân hai lớp (The Two-Category Case)

Giả sử có một tập hợp n mẫu $y_1, y_2, ..., y_n$

Trong đó một số mẫu được gán nhãn W_1 và một số được gán nhãn W_2. Để xác định vectơ trọng số a trong hàm phân biệt tuyến tính $g(X) = a^t y$. Mẫu y_i được phân lớp chính xác nếu $a^t y > 0$ và được gán nhãn là W_1 ngược lại thì y_i được gán nhãn là W_2.

Vậy, ta đã thay thế việc tìm giải pháp cho một tập hợp các bất phương trình tuyến tính bởi tìm giải pháp cho một tập hợp các phương trình tuyến tính.

$$\begin{pmatrix} y_{10} & y_{11} & \cdots & y_{1k} \\ y_{20} & y_{21} & \cdots & y_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ y_{n0} & y_{n1} & \cdots & y_{nk} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ \vdots \\ b_n \end{pmatrix} \quad \textbf{hay} \quad Ya = b \tag{3.12}$$

Ta có thể viết (12) dưới dạng: $a = Y^{-1}b$ (Nếu Y là ma trận khả nghịch) do đó, ta có thể tìm vectơ trọng số a sao cho sai số $\mathbf{Y^*a}$ và **b** là cực tiểu. Gọi vectơ e là: $e = Ya - b$ (3.13)

Thì ta cần phải tìm vectơ a sao cho:

$$J_s(a) = \left\|Ya\text{-}b\right\|^2 = (Ya-b)^t(Ya-b) = \sum (a^t y_i - b_i)^2 \tag{3.14}$$

Để tìm cực tiểu của tổng bình phương các sai số thì ta tìm bằng phương pháp đạo hàm:

$$\nabla J_s(a) = \sum_{i=1}^{n} 2(a^t y_i - b_i)y_i = 2Y^t(Ya-b) \tag{3.15}$$

Cho phương trình đạt giá trị 0 và giải ra ta được điều kiện:

$$Y^tYa = Y^tb \tag{3.16}$$

Vậy ta chỉ cần tìm nghiệm a thỏa mãn phương trình (3.16) là đủ. Giải ra ta được : $a = (Y^tY)^{-1}Y^tb = Y^*b$ (3.17)

$$Y^* = (Y^tY)^{-1}Y^t \tag{3.18}$$

3.3.2. Trong trường hợp phân nhiều lớp:

Ta có: $g_i(X) = W^t X + W_{i0}$ với i = 1, 2, …, c

Đặt y(X) là một vectơ k+1 chiều của các hàm X và khi đó,

$$g_i(X) = a_i^t y \qquad i=1, 2, \ldots, c \tag{3.19}$$

Khi đó, X được gán cho lớp W_i nếu $g_i(X) > g_j(X)$ với $\forall j \neq i$

Lúc này tồn tại một tập hợp các vectơ trọng số a_i (i = 1, 2, …,c) sao cho nếu mẫu

$$y_k \in Y_k \text{ thì } a_i^t y_k > a_j^t y_k \quad \forall j \neq i \tag{3.20}$$

Xem bài toán như là c bài toán con, mỗi bài toán con là mỗi bài toán phân loại 2 nhóm. Nghĩa là đối với bài toán con thứ i trọng số sẽ tìm vectơ trọng số a_i là kết quả của hệ phương trình:

$$\begin{cases} a_i^t y = 1 & \forall \mathbf{i} \in \mathbf{Y_t} \\ a_i^t y = -1 & \forall \mathbf{i} \notin \mathbf{Y_t} \end{cases} \quad (3.21)$$

• Ma trận Y trong trường hợp tổng quát sẽ là một ma trận cấp (nx(k+1)) của các mẫu được xét. Giả sử Y được phân hoạch có dạng:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_c \end{bmatrix} \quad (3.22)$$

• Tương tự gọi A là ma trận cấp ((k+1) x c) của các vectơ trọng số có dạng tổng quát là:

$$A = [a_1 \ a_2 \ \dots \ a_c] \quad (3.23)$$

• Ma trận B là ma trận cấp (n x c) có dạng

$$B = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_c \end{bmatrix} \quad (3.24)$$

Theo cách phát triển của ma trận bình phương lỗi $(YA - B)^t (YA - B)$ khi đó là kết quả của phương trình:

$$A = Y^* B \quad (3.25)$$

Bây giờ, việc tìm c hàm phân biệt tuyến tính thực hiện theo các bước sau:

• Bước 1: Tìm các vectơ trọng số a_i theo phương pháp MSE thõa hệ phương trình: $\begin{cases} a_i^t y = 1 & \forall \mathbf{i} \in \mathbf{Y_i} \\ a_i^t y = 0 & \forall \mathbf{i} \notin \mathbf{Y_i} \end{cases}$ (3.26)

• Bước 2: Sử dụng kết quả của bước 1, gán mẫu y_k cho nhóm W_i, nếu $a_i^t y_k > a_j^t y_k$ với $\forall i \neq j$

3.3.3. Qui trình thực hiện chương trình phân lớp dữ liệu

• Bước 1: Nhập dữ liệu gồm một tập mẫu ngẫu nhiên $(X_1^1, X_2^1, ..., X_k^1)$, $(X_1^2, X_2^2, ..., X_k^2)$, …, $(X_1^n, X_2^n, ..., X_k^n)$ thu được từ quan sát lưu trữ dưới dạng bảng dữ liệu.

• Bước 2: Tìm ước lượng của các hệ số của các vectơ trọng số a_i bằng thuật toán di truyền

• Bước 3: Vẽ đồ thị minh họa cho kết quả của sự phân lớp.

• Bước 4: Cho một bộ các giá trị ($X_1^*, X_2^*, ..., X_k^*$) xác định xem mẫu này sẽ thuộc vào lớp nào trong các phân nhóm.

CHƯƠNG 4. PHÂN TÍCH HỒI QUY

4.1. DẪN NHẬP

Hiện nay các vấn đề trong khoa học, kỹ thuật hay những lĩnh vực khác trong thực tế, có liên quan đến việc xác định mối liên hệ giữa một tập hợp các tiêu chuẩn hay các đại lượng (các biến) khác nhau về bản chất.

Chúng ta có thể làm rõ bản chất của hiện tượng hay sự việc cần nghiên cứu để tìm ra quy luật và dự đoán. Dạng đơn giản là, phương trình hồi quy:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_kX_k \quad (4.1)$$

4.2. ƯỚC LƯỢNG CÁC MÔ HÌNH TOÁN HỌC

4.2.1. Ước lượng các mô hình toán học

Các bước để ước lượng mô hình toán học bao gồm:

• Bước 1: Mô hình hóa đối tượng nghiên cứu để tiến hành thu thập số liệu thực nghiệm.

• Bước 2: Dự đoán mô hình toán học dựa trên cơ sở các số liệu đã thu thập được trong quá trình nghiên cứu.

• Bước 3: Xác định các hệ số của mô hình toán học

• Bước 4: Kiểm định sự phù hợp của mô hình toán đã dự đoán.

4.2.2. Mô hình hóa đối tượng nghiên cứu

Gọi X_1, X_2 ..., X_k là các nguyên nhân tác động gây nên hậu quả hay kết quả Y thì hàm $Y = f(X_1, X_2, ..., X_k) \rightarrow Y$.

4.2.3. Xây dựng mô hình toán học

4.2.3.1. Phương pháp "đồ thị thực nghiệm" và "tuyến tính hóa":

Mô hình toán học có thể dự đoán nhờ đồ thị thực nghiệm được phác họa từ các số liệu thu tập được.

4.2.3.2.Dự đoán mô hình toán học bằng phương pháp suy luận:

Chẳng hạn, mô hình gradient mật độ hay nồng độ có thể dự đoán được là **Y=aX + b**, với b là mật độ hay nồng độ ở trung tâm xuất phát điểm, X là khoảng cách từ trung tâm đó đến điểm đang xét. Ở đây, X có thể được thay thế bằng các đại diện của nó như lnX, 10^X, e^X, X^2,...

4.2.4. Tìm các hệ số của mô hình toán học

Hai phương pháp thường được sử dụng là :

- Phương pháp tối thiểu hóa tổng bình phương sai số

- Phương pháp Moment.

4.2.5. Kiểm định và đánh giá mức độ phù hợp của mô hình toán học

Mô hình toán học $Y = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$

hay $Y^* - \overline{Y} = b_1 (X_1 - \overline{X}) + b_2 (X_2 - \overline{X}) + ... + b_K (X_k - \overline{X})$

4.3. PHƯƠNG PHÁP TỐI TIỂU HÓA TỔNG BÌNH PHƯƠNG SAI SỐ (MINIMUM SUM SQUARED METHOD)

4.3.1. Phương pháp tối tiểu hóa tổng bình phương sai số

Phương pháp bình phương tối thiểu là phương pháp chuẩn để cụ thể hoá mô hình hồi quy tuyến tính và ước lượng các thông số chưa biết và tuân theo 4 giả thiết sau đây:

1. Các biến độc lập x_i không phải là các biến ngẫu nhiên.

2. Kỳ vọng toán của thành phần sai số (ε_i) bằng 0, tức là $E[\varepsilon_i]=0$

3. Có tính thuần nhất - phương sai của thành phần sai số cố định, tức là $var(\varepsilon_i) = \sigma_2$

4. Không có tự tương quan, tức là $cov(\varepsilon_i, \varepsilon_j) = 0, (i \neq j)$

Nếu f có dạng phi tuyến thì ta sẽ tiến hành tuyến tính hóa mô hình toán học trước khi tiến hành phân tích. Khi đó, phương trình hồi quy sẽ có dạng như phương trình (2):

$$Y = \varphi(X_1, X_2, ..., X_k) = f(X_1,X_2,...,X_k; b_0, b_1,...,b) \quad (4.2)$$

$$\text{Hay } Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... b_kX_k \quad (4.3)$$

Nếu giá trị Y hồi quy, hoàn toàn trùng khớp với giá trị Y thực nghiệm. Khi đó, ta có :

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... b_kX_k+e \quad (4.4)$$

$$Q_e = \sum_{i=1}^{n} (Y_i - Y_i^*)^2 \quad (4.5)$$

Để tìm các tham số b_0, b_1, ... b_k ta lấy đạo hàm riêng của Q_e theo các biến b_0, b_1, ... b_k, cho các giá trị đạo hàm này bằng 0, ta có hệ phương trình sau:

$$\begin{cases} \dfrac{\partial(Q_e)}{\partial(b_0)} = 0 \\ \dfrac{\partial(Q_e)}{\partial(b_1)} = 0 \\ \quad ... \qquad ... \\ \dfrac{\partial(Q_e)}{\partial(b_k)} = 0 \end{cases} \quad (4.6)$$

4.3.2. Tìm giá trị của các hệ số hồi quy bằng thuật giải di truyền

Để xác định các giá trị của các hệ số hồi quy b_0, b_1, ,,, b_k chúng ta sử dụng công cụ tìm giá trị tối thiểu của hàm nhiều biến bằng **thuật giải di truyền** đã được trình bày trong chương 2 **để tìm giá trị cực tiêu gần đúng của Q_e.** Từ đó xác định được các giá trị ước lượng của các tham số b_0, b_1, ..., b_k của phương trình hồi quy tuyến tuyến.

4.4. ƯỚC LƯỢNG HỒI QUY TUYẾN TÍNH

4.4.1. Ước lượng Hồi quy tuyến tính đơn

Cho hai đại lượng hồi quy tuyến tính ngẫu nhiên X và Y , khi đó mô hình hồi quy tuyến tính đơn tổng quát có dạng: $Y = b_0 + b_1X$

Trong đó b_0 và b_1 được xác định như sau:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i y_i - n\bar{x}\bar{y})}{\sum_{i=1}^{n}(x_i^2 - n\bar{x}^2)} \quad ; \quad b_0 = \bar{y} - b_1 \bar{x}$$

Việc phân tích hồi quy dựa trên mô hình toán học được thực hiện như sau:

• Bước 1: Tìm giá trị cực tiểu của một hàm nhiều biến số (hai biến) bằng thuật giải di truyền để xác định các hệ số hồi quy b_0, b_1 của mô hình toán học và giá trị gần đúng của Q_e.

• Bước 2: Kiểm định sự phù hợp theo các công thức:

▪ $$Q_x = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2 \qquad (4.7)$$

▪ $$Q_Y = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 \qquad (4.8)$$

▪ $$Q_e = \sum_{i=1}^{n}(Y_i - Y_i^*)^2 \qquad (4.9)$$

▪ $$Q_{Y^*} = \sum_{i=1}^{n}(Y_i^* - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i^*)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 \qquad (4.10)$$

▪ $$Q_Y = Q_{Y^*} + Q_e \qquad (4.11)$$

▪ $$F(1, n-2) = \frac{(n-2)Q_{Y^*}}{Q_e} \qquad (4.12)$$

▪ $$|r| = R = \sqrt{\frac{Q_R}{Q_Y}} = \sqrt{1 - \frac{Q_e}{Q_Y}} \qquad (4.13)$$

Bảng 4.1 Bảng kiểm định & đánh giá mức độ phù hợp của mô hình toán học

$Y = b_0 + b_1X$

Nguồn biến lượng	**Độ tự do**	**Tổng các bình phương**	**Biến lượng**
Hồi quy	1	Q_{Y^*}	$S^2 = \frac{Q_e}{(n-2)}$
Sai số ngẫu nhiên	n - 2	Q_e	

Tổng thực tế	n - 1	Q_Y	
$F(1,n-2)=\frac{(n-2)Q_{Y^*}}{Q_e}$			
$\|r\|=R=\sqrt{\frac{Q_R}{Q_Y}}=\sqrt{1-\frac{Q_e}{Q_Y}}$			

• Bước 3: Kiểm định giá trị của hệ số b_0 với giải thuyết tương đồng

$$b_0 - t_p(n-2)\sqrt{\frac{S^2\sum_{i=1}^{n}X_i^2}{nQ_x}} < b_0 < b_0 + t_p(n-2)\sqrt{\frac{S^2\sum_{i=1}^{n}X_i^2}{nQ_x}} \qquad (4.14)$$

• Bước 4: Xác định khoảng tin tưởng cho $Y_0 = b_0 + b_1X_0$

Và $Y_0 \in Y_0 \pm t_p(n-2)S_{Y^0}$ (4.15)

Với: $$S_{Y^0}=S\sqrt{\left[\frac{1}{n}+\left(\frac{X_0-\overline{X}}{Q_x}\right)^2\right]} \qquad (4.16)$$

4.4.2. Hồi quy tuyến tính bội :

Mô hình toán học tổng quát **hồi quy tuyến tính bội** có dạng:

$$Y = b_0 + b_1X_1 + b_2X_2, +...+ b_{k-1}X_{k-1}$$

Tiến hành phân tích hồi quy dựa trên mô hình toán học như sau:

• Bước 1: Tìm giá trị cực tiểu của một hàm nhiều biến số (số biến ≥3) bằng thuật giải di truyền để xác định các hệ số hồi quy b_0, b_1, b_2, ..., b_{k-1} của mô hình toán học và giá trị gần đúng của Q_e.

• Bước 2: Kiểm định sự phù hợp của mô hình toán học tìm được. Tương tự trường hợp **K = 2** mô hình toán học dạng tổng quát vẫn được tính theo công thức (9), hay là:

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(Y_i^*-\overline{Y})^2+\sum_{i=1}^{n}(Y_i-Y_i^*)^2$$

hay: $$Q_Y = Q_{Y^*} + Q_e \qquad (4.17)$$

▪ Tính các giá trị Q_Y, Q_{Y^*} theo các công thức (4.8) và (4.10).

▪ Tiến hành **kiểm định sự phù hợp**

$$F(K-1,n-1)=\frac{\frac{Q_{Y^*}}{K-1}}{\frac{Q_e}{n-K}} \quad (4.18)$$

▪ Để đánh giá mức độ phù hợp của mô hình toán học, chúng ta sử dụng hệ số tương quan đa phần R theo công thức (4.13), như sau:

$$R=\sqrt{\frac{Q_{Y^*}}{Q_Y}}=\sqrt{1-\frac{Q_e}{Q_Y}}$$

▪ Để kiểm định giá trị của hệ số tương quan đa phần R chúng ta sử dụng trắc nghiệm F với K-1 và n-K độ tự do và giả thuyết tương đồng H_0:b_1=0:

$$F(k-1,n-1)=\frac{\frac{R^2}{K-1}}{\frac{1-R^2}{n-K}} \quad (4.19)$$

Bảng 4.2 Bảng kiểm định và đánh giá mức độ phù hợp của mô hình toán học Y = $b_0 + b_1X_1 + b_2X_2, +...+ b_{k-1}X_{k-1}$

Nguồn biến lượng	Độ tự do	Tổng các bình phương	Biến lượng
Hồi quy	K - 1	$Y = b_0 + b_1X_1 + b_2X_2, +...+ b_{k-1}X_{k-1}$	$S^2=\frac{Q_e}{(n-K)}$
Sai số ngẫu nhiên	n - K	Q_e	
Tổng thực tế	n - 1	Q_Y	
$F(k-1,n-1)=\frac{\frac{R^2}{K-1}}{\frac{1-R^2}{n-K}}$			
$\lvert r\rvert=R=\sqrt{\frac{Q_{Y^*}}{Q_Y}}=\sqrt{1-\frac{Q_e}{Q_Y}}$			

Về ma trận hiệp phương sai: $K = S^2 A^{-1}$ (4.20)

Trong đó, A là ma trận:

$$\begin{bmatrix} Q_{x_1} & Q_{x_1x_2} & Q_{x_1x_3} & \dots & Q_{x_1x_{k-1}} \\ Q_{x_1x_2} & Q_{x_2} & Q_{x_2x_3} & \dots & Q_{x_2x_{k-1}} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ Q_{x_1x_{k-1}} & Q_{x_2x_{k-1}} & Q_{x_3x_{k-1}} & \dots & Q_{x_{k-1}} \end{bmatrix} \quad (4.21)$$

Trong đó:
$$Q_{x_i} = \sum_{j=1}^{n}(X_{ij} - \overline{X_i})^2 = \sum_{j=1}^{n} X_{ij}^2 - \frac{1}{n}\left(\sum_{j=1}^{n} X_{ij}\right)^2 \quad (4.22)$$

$$Q_{x_i x_j} = \sum_{i=1}^{n}(X_{ij} - \overline{X_i})(X_{ij} - \overline{X_j}) = \sum_{j=1}^{n} X_i X_j - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} X_j\right) \quad (4.23)$$

✓ Trong trường hợp K = 2 (mô hình toán học là $Y = b_1X + b_0$) ma trận A có dạng:

$$A = \begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix} \quad (4.24)$$

✓ Trong hai trường hợp hồi quy có dạng đường cong bậc hai (mô hình toán học là $Y = b_2X^2 + b_1X + b_0$) thì ma trận hiệp phương sai A có dạng:

$$A = \begin{bmatrix} n & \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 & \sum_{i=1}^{n} X_i^3 \\ \sum_{i=1}^{n} X_i^2 & \sum_{i=1}^{n} X_i^3 & \sum_{i=1}^{n} X_i^4 \end{bmatrix} \quad (4.25)$$

▪ Về phương sai của b_1:
$$S_{b_i}^2 = S^2 * C_{ii} \quad (4.26)$$

với C_{ii} là phần tử của ma trận A^{-1}

▪ Về phương sai của $Y_0 = b_0 + b_1X_{10} + b_2X_{20}, +... + b_{k-1}X_{k-10}$:

$$S_{Y_0}^2 = \frac{S^2}{n} + \left(X_1^0 - \overline{X_1}\right)^2 K_{11} + \left(X_2^0 - \overline{X_2}\right)^2 K_{22} + ... + 2\left(X_1^0 - \overline{X_1}\right)\left(X_2^0 - \overline{X_2}\right)C_{12} \quad (4.27)$$
$$+ ... + 2\left(X_2^0 - \overline{X_2}\right)\left(X_3^0 - \overline{X_3}\right)C_{13} + 2\left(X_2^0 - \overline{X_2}\right)\left(X_4^0 - \overline{X_4}\right)C_{14} + ...$$

Với K_{ij} là phần tử của ma trận hiệp phương sai K.

Trong trường hợp mô hình toán học là phù hợp với các số liệu thực nghiệm thu được thì ta tiến hành các bước sau:

• **Bước 3:** Kiểm định giá trị của các hệ số b_i bằng cách sử dụng khoảng tin tưởng của b_i với $P \le 0.05$ như sau:

$$b_i - t_p(n-K)S_{b_i} < b_t < b_t + t_p(n-K)S_{b_i} \quad (4.28)$$

• **Bước 4:** Xác định khoảng tin tưởng (dự đoán giá trị của Y_0 dựa vào tập giá trị X_{i0}). Cho $Y_0 = b_0 + b_1X_{10} + b_2X_{20}, +...+ b_{k-1}X_{k-10}$: Khoảng tin tưởng của Cho $Y_0 = b_0 + b_1X_{10} + b_2X_{20}, +...+ b_{k-1}X_{k-10}$ với mức sai lầm P được cho bởi:

$$Y_0 - t_p(n-2)S_{Y_{10}} < Y_0 < Y_0 + t_p(n-2)S_{Y_0}$$

hay $$Y_0 \in Y_0 \pm t_p(n-2)S_{Y_0} \quad (4.29)$$

4.4.3. Các bước thực hiện chương trình phân tích dự liệu hồi quy

• **Bước 1:** Nhập một tập mẫu ngẫu nhiên

$$(X_1^1, X_2^1, ..., X_{k-1}^1, Y_1), (X_1^2, X_2^2, ..., X_{k-1}^2, Y_2), ..., (X_1^n, X_2^n, ..., X_{k-1}^n, Y_n)$$

• **Bước 2:** Phác họa đồ thị của hàm số dựa theo các biến độc lập và phụ thuộc được chọn.

• **Bước 3:** Tìm ước lượng của các hệ số hồi quy b_j của phương trình: $Y = b_0 + b_1X_1 + b_2X_2, +...+ \beta_{k-1}X_{k-1}$ sao cho tổng giá trị của các sai số giữa các giá trị Y_i ,Y^* là nhỏ nhất.

• **Bước 4:** Kiểm định mức độ phù hợp của mô hình toán học

• **Bước 5:** Cho một bộ các giá trị $(X_1^1, X_2^1, ..., X_1^*, X_2^*, ..., X_k^*)$ của các biến độc lập X_i dự đoán giá trị Y^* của biến phụ thuộc Y.

• **Bước 6:** Tìm xem với giá trị nào của (X_1, X_2, ... X_k) thì Y đạt giá trị cực đại (Max) hay giá trị cực tiểu (Min).

• **Bước 7:** Vẽ đồ thị đường biểu diễn dữ liệu

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. KẾT LUẬN

Luận văn ứng dụng thuật giải di truyền để tìm cực trị của một hàm đa biến được trình bày trong chương 2. Kết quả này được sử dụng làm công cụ để giải quyết hai lớp bài toán thuộc lĩnh vực thống kê được đề cập ở hai chương tiếp theo của luận văn.

Mục tiêu của luận văn là giải hai lớp bài toán thuộc lĩnh vực thống kê đã nêu. Kết quả của các bài toán mang lại vừa có tính năng của một hệ thống máy học, giúp dự báo, tính toán, phân lớp các dữ liệu không được học vừa có ý nghĩa đề xuất và đạt được kết quả khả quan về một phương pháp phân lớp dữ liệu cũng như việc thiết lập các mô hình toán học.

Bài toán phân lớp dự liệu dựa trên tập các hàm phân biệt tuyến tính thực chất là tìm cách phân chia tập dữ liệu ban đầu (có kích thước lớn) thành những tập dữ liệu nhỏ hơn mà mỗi tập dữ liệu con này thỏa một số tính chất đặc thù nào đó, tạo điều kiện thuận lợi cho quá trình phân tích dữ liệu, nghiên cứu dữ liệu sau này nhẹ nhàng hơn, ít tốn công sức hơn nhưng vẫn có thể đạt được hiệu quả cao.

Bài toán phân tích hồi quy tuyến tính thực chất là tìm mối quan hệ mô tả sự phụ thuộc của các giá trị biến ngẫu nhiên độc lập vào giá trị của biến phụ thuộc xuất. Kiểm định độ tin cậy của mô hình tìm được, đồng thời cho phép ta dự báo các giá trị nằm ngoài tập thực nghiệm với độ chính xác cao mà không cần phải lưu trữ tập thực nghiệm nữa.

Việc áp dụng thuật giải di truyền để giải quyết hai lớp bài toán trên được trình bày một cách rõ ràng, cụ thể. Thể hiện một phương pháp tiếp cận mới, tinh tế để giải quyết một số lớp bài toán trong lĩnh vực thống kê là những bài toán tốn rất nhiều công sức cho thao tác tính toán để tìm lời giải cho bài toán. Cách tiếp cận bằng thuật toán di truyền chúng ta có thể giảm đi chi phí công sức cho việc tính toán rất nhiều mà vẫn đạt được kết quả tối ưu.

Các kết quả đạt được của luận văn đã góp phần xây dựng một phương pháp mới, một hướng tiếp cận mới để giải quyết một số lớp bài toán thống kê

ngoài các phương pháp toán học bằng giải tích truyền thống. Đồng thời cũng chứng minh được tiềm năng to lớn và tính ưu việt của thuật giải di truyền trong vấn đề tìm kiếm lời giải tối ưu cho nhiều dạng vấn đề khác nhau.

2. HƯỚNG PHÁT TRIỂN

Mặc dù đã đạt được một số kết quả nhất định nhưng vẫn chưa giải quyết rốt ráo các vấn đề liên quan đến hai lớp bài toán phân tích hồi quy và phân lớp dữ liệu như: Trong bài toán hồi quy tuyến tính chưa nghiên cứu vấn đề hồi quy phi tuyến để có thể giải quyết trọn vẹn bài toán hồi quy ở dạng tổng quát, còn bài toán phân lớp dữ liệu dựa trên các hàm phân biệt tuyến tính, chưa nghiên cứu đến các hàm phân biệt phi tuyến nên tính chính xác của kết quả chưa cao.

Trong tương lai, tôi mong muốn có được cơ hội tiếp tục tìm tòi, học hỏi thêm nhằm hoàn thiện đề tài cũng như có điều kiện nghiên cứu chuyên sâu hơn về thuật giải di truyền để giải quyết các bài toán có tính phức tạp cao như bài toán xếp lịch biểu.