ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

HOÀNG HUYỀN TRANG

PHƯƠNG PHÁP PHÂN CỤM DỰA TRÊN TẬP THÔ VÀ GIẢI THUẬT DI TRUYỀN

Chuyên ngành: Hệ thống thông tin

Mã số: 60480104

TÓM TẮT LUẬN VĂN THẠC SĨ

Hà Nội - 2016

MỞ ĐẦU

Phân cụm dữ liệu là một trong những nghiên cứu quan trọng trong khai thác dữ liệu và được áp dụng cho đa lĩnh vực [7,8]. Mục tiêu chính trong phân cụm dữ liệu là để phân loại các đối tượng không có nhãn thành nhiều cụm mà các đối tượng thuộc cùng một cụm thì tương tự nhau và khác nhau đối với các cụm khác nhau. Phân cụm dữ liệu được chia làm hai loại là phân cụm cứng/rõ và phân cụm mềm [12,15].

Một kỹ thuật được sử dụng phổ biến trong phân cụm dữ liệu là thuật toán K-Means, thuộc phân cụm rõ, với sự hội tụ nhanh chóng và khả năng tìm kiếm địa phương mạnh mẽ. Trong quá trình phân cụm K-Means truyền thống, các đối tượng dữ liệu thu được trong cụm là nhất định. Tuy nhiên, trong thực tế giữa những đối tượng thường không có ranh giới rõ ràng. Để tăng hiệu quả và kết quả chính xác cho phân cụm việc sử dụng lý thuyết tập thô tiếp cận hỗ trợ phân cụm K-Meansđược đề xuất.

Mặc dù giải thuật K-Means thô có khả năng tìm kiếm địa phương mạnh mẽ nhưng lại dễ rơi vào cực trị địa phương. Một trong những biện pháp có thể khắc phục được hạn chế này là kết hợp với giải thuật di truyền là một thuật toán dựa trên nguyên tắc của sự tiến hóa sinh học, có lượng lớn số song song tiềm ẩn thực hiện không gian tìm kiếm lớn và cung cấp giải pháp tối ưu hóa toàn cầu giúp tránh được tối ưu địa phương.

Luận văn trình bày khảo cứu một cách hệ thống của bài báo [6] các kiến thức về phân cụm dữ liệu rõ, thô theo hướng K-Means và ứng dụng giải thuật di truyền để phân cụm dữ liệu thô. Trên cơ sở đó xây dựng chương trình thực nghiệm trên một số bộ dữ liệu, kết quả cho thấy ưu điểm của phương pháp mới. Cấu trúc của luận văn gồm 3 chương :

Chương I. Phân cụm dữ liệu và một số vấn đề liên quan.

Chương II. Phân cụm dựa trên tập thô và thuật toán di truyền.

Chương III. Cài đặt và phân tích thí nghiệm.

CHƯƠNG I. PHÂN CỤM DỮ LIỆU VÀ MỘT SỐ VẤN ĐỀ LIÊN QUAN

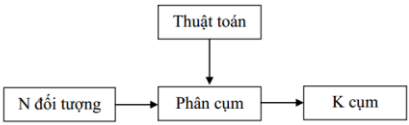
1.1. Giới thiệu về phân cụm dữ liệu

Khai phá dữ liệu tuộc quá trình khám phá tri thức. Về bản chất là giai đoạn duy nhất tìm ra được thông tin mới, tiềm ẩn có trong cơ sở dữ liệu chủ yếu phục vụ cho mô tả và dự đoán.

Phân cụm dữ liệu là một kỹ thuật trong khai phá dữ liệu với mục đích chính là khám phá cấu trúc của mẫu dữ liệu để thành lập các nhóm dữ liệu từ tập dữ liệu lớn, cho phép phân tích và nghiên cứu cho từng cụm dữ liệu nhằm khám phá và tìm kiếm các thông tin tiềm ẩn, hữu ích.

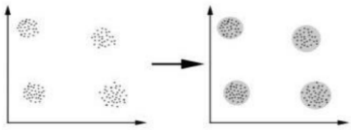
1.1.1. Khái niệm và mục đích của phân cụm dữ liệu

Bài toán phân cụm dữ liệu là một nhánh ứng dụng chính của lĩnh vực học không giám sát, mà dữ liệu mô tả trong bài toán là không được dán nhãn. Trong trường hợp này, thuật toán sẽ tìm cách phân cụm dữ liệu thành từng nhóm có đặc điểm tương tự nhau, nhưng đồng thời đặc tính giữa các nhóm đó lại phải càng khác biệt càng tốt. Số các cụm dữ liệu có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định theo thuật toán.



Hình 1.1. Quy trình phân cụm.

Độ tương tự được xác định dựa trên giá trị các thuộc tính mô tả đối tượng. Thông thường, phép đo khoảng cách thường được sử dụng để đánh giá độ tương tự hay phi tương tự. Vấn đề phân cụm có thể minh hoạ như hình 1,2:



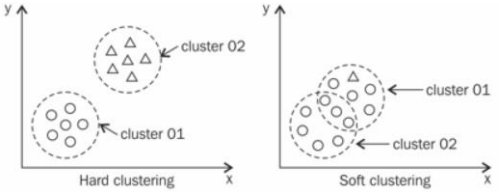
Hình 1.2. Mô phỏng sự phân cụm dữ liệu.

Ứng dụng của phân cụm dữ liệu: Được áp dụng trong rất nhiều lĩnh vực như: Kinh doanh; Sinh học; Thư viện; Bảo hiểm; www…

1.1.2. Phương pháp phân cụm dữ liệu

Phân cụm dữ liệu được chia làm hai loại là phân cụm dữ liệu cứng và phân cụm dữ liệu mềm:

- Phân cụm dữ liệu cứng (hay phân cụm rõ) là phương pháp gán mỗi đối tượng vào một và chỉ một cụm và xác định rõ ranh giới giữa các cụm. Một số thuật toán: Thuật toán K-Means, Thuật toán K-Medoids...
- Phân cụm dữ liệu mềm (hay phân cụm mờ) là phương pháp cho phép mỗi đối tượng có thể thuộc một hoặc nhiều cụm dữ liệu và có sự mơ hồ hoặc mờ ranh giới giữa các cụm: Thuật toán Fuzzy C-mean…



Hình 1.3. Mô tả phân cụm cứng/rõ và phân cụm mềm/mờ

Tùy theo đặc điểm về tính tương đồng của các đối tượng trong bài toán đang xét, có nhiều cách tiếp cận cho thuật toán phân cụm. Các kỹ thuật gồm:

- Phân cụm phân cấp (Hierarchical Data Clustering)
- Phân cụm phân hoạch (Partition Based Data Clustering)
- Phân cụm dựa trên mật độ (Density Based Data Clustering)
- Phân cụm dựa trên lưới (Grid Based Data Clustering)

1.1.3. Phân cụm với giải thuật K-Means

Thuật toán K-Means (MacQueen, 1967)[2] là một trong những thuật toán học không giám sát đơn giản nhất để giải quyết vấn đề phân cụm dữ liệu nổi tiếng, với số cụm được xác định trước là k cụm.

Thuộc nhóm phân cụm dữ liệu cứng/rõ, ý tưởng chính là để xác định k trọng tâm cho k cụm, một trọng tâm cho mỗi cụm. Những trọng tâm nên được đặt ở vị trí thích hợp nhất vì vị trí khác nhau gây ra kết quả khác nhau. Vì vậy, sự lựa chọn tốt hơn là đặt chúng càng nhiều càng tốt và cách xa nhau. Bước tiếp theo là với mỗi điểm thuộc tập dữ liệu cho trước và liên kết nó với trọng tâm gần nhất.

Giả sử thiết lập tập đối tượng $X=\{x_1,x_2,\ldots x_n\}$ và k trọng tâm cụm $C=\{C_1,C_2,\ldots C_k\}$; lấy $w_1,w_2,\ldots w_k$ của k cụm.

Công thức $C_j = \frac{1}{N_j}\sum_{x\in w_j} x$ với j=1, 2, …, j_k trong đó N_j là số lượng cụm thứ j. Xác định hàm mục tiêu như sau: $E = \sum_{i=1}^{k}\sum_{x\in w_i} d(x,c_i)$ trong đó c_i là tâm của cụm w_i tương ứng. Với $d(x,c_i)=\|x-c_i\|^2$ là khoảng cách Euclide từ điểm đối tượng của mỗi cụm đến các trung tâm cụm.

Thuật toán K-Means:

Quá trình phân cụm K-Means được biểu diễn bởi hình 1.4.

Đầu vào: k: số cụmX: tập dữ liệu chứa n đối tượng

Đầu ra: một tập hợp k các cụm

Bước 0. Xác định số lượng cụm k và điều kiện dừng

Bước 1. Khởi tạo ngẫu nhiên k trọng tâm cụm

Bước 2. Gom các đối tượng vào cụm mà nó gần tâm nhất

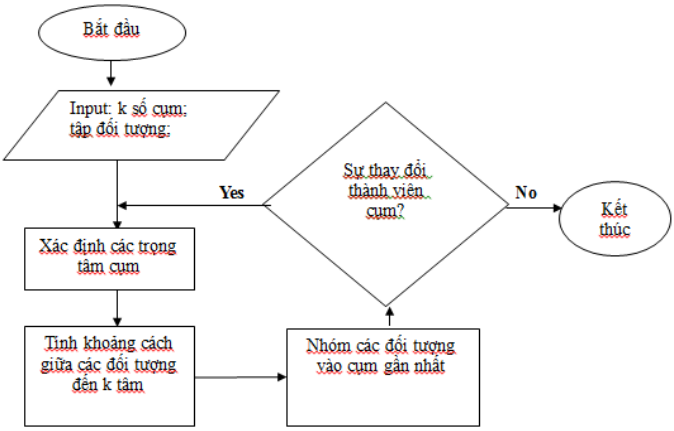
Bước 3. Tính lại các tâm theo đối tượng đã được phân hoạch ở bước 2

Lặp cho đến khi điều kiện dừng thỏa mãn.

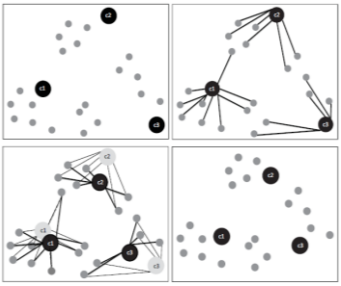
Điều kiện dừng thường chọn các điều kiện sau:

- Số lần lặp t=Tmax trong đó Tmax là số cho trước
- $|E^t - E^{t-1}|<\Delta$ trong đó Δ là hằng số bé cho trước
- Tới khi các cụm không đổi

Khi tập dữ liệu không quá lớn thì người ta dùng điều kiện dừng thứ 3.



Hình 1.4. Sơ đồ thuật toán phân cụm K-means



Hình 1.5. Mô phỏng quá trình phân cụm K-Means

Ví dụ quá trình phân cụm K-Means từ hình 1.5:

(1) Khởi tạo ngẫu nhiên trọng tâm cho mỗi cụm; ở đây là C1,C2,C3

(2) Gán mỗi đối tượng có khoảng cách gần nhất với trọng tâm vào một cụm;

(3) Xác định lại trọng tâm mới bằng cách tính toán giá trị trung bình của các đối tượng trong mỗi cụm;

(4) Lặp đến khi trọng tâm cụm không đổi.

1.2. Lý thuyết tập thô

Lý thuyết tập thô (Rough Set Theory - RST) được phát triển bởi Zdzislaw Pawlak, là phần mở rộng của lý thuyết tập hợp cho việc nghiên cứu hệ thống thông minh đặc trưng bởi các thông tin

không chính xác, không chắc chắn và không đầy đủ, chủ yếu dùng phân tích phân loại bảng dữ liệu.

Mục đích chính của việc phân tích tập thô là tổng hợp khái niệm xấp xỉ từ các dữ liệu thu được, trong đó có sự mơ hồ, thiếu giá trị hoặc dự phòng của các thuộc tính.

1.2.1. Hệ thông tin và quyết định

Một tập hợp dữ liệu được biểu diễn như một bảng trong đó mỗi hàng đại diện cho một trường hợp, một sự kiện, một mô hình hay đơn giản là một đối tượng. Mỗi cột đại diện cho một thuộc tính của từng đối tượng; các thuộc tính cũng có thể được cung cấp bởi một chuyên gia hay người sử dụng. Bảng này được gọi là hệ thông tin.

Hình thức hơn, hệ thông tin là một cặp I=(U,A) với U là tập hữu hạn đối tượng khác rỗng được gọi là miền và A là một tập hữu hạn các thuộc tính khác rỗng. Với mỗi $a \in A, a: U \to V_a$ tập V_a là tập giá trị của thuộc tính a.

Bảng 1.1. Hệ thống thông tin

	Color	Size
1	Green	Big
2	Green	Small
3	Yellow	Medium
4	Red	Medium
5	Yellow	Medium
6	Green	Big
7	Red	Small
8	Red	Small

Ví dụ [1] Bảng 1.1, Một hệ thông tin bao gồm 8 đối tượng:

U={u1,u2,u3,u4,u5,u6,u7,u8}

Tập thuộc tính A={Color, Size } và miền giá trị cho từng thuộc tính là V_{color} = {Green, Yellow, Red}, V_{Size} = {Small, Medium, Big }.

Bảng quyết định là dạng đặc biệt của hệ thông tin, kí hiệu S=(U,A) hay $S = (U, C \cup D)$ với tập thuộc tính A được phân thành hai tập rời nhau C và D, trong đó C là tập các thuộc tính điều kiện, D là tập các thuộc tính quyết định sao cho $C \cap D = \varnothing$

Định nghĩa 1.[6] Bảng quyết định $S = (U, C, D, V, f)$ Trong đó: - $U = \{x_1, x_2 \ldots, x_n\}$ là tập hữu hạn đối tượng khác rỗng, được gọi là miền;

- A Tập hữu hạn thuộc tính khác rỗng: $A = C \cup D$ và $C \cap D = \varnothing$ với C là một tập thuộc tính điều kiện, D là tập thuộc tính quyết định

- $V = \bigcup_{a \in C \cup D} V_a$ với V_a là tập giá trị của thuộc tính a

- $f : U \times C \cup D \to V$ là hàm thông tin sao cho $f(x,a) \in V_a$ với $\forall a \in C \cup D, x \in U$

Ví dụ [1] Bảng quyết định 1.2, bảng này có 8 đối tượng như trong bảng 1.1 nhưng có thêm thuộc tính quyết định (Shape). Trong bài toán phân lớp thì thuộc tính quyết định chính là lớp của đối tượng cần xếp lớp. Trong ví dụ này thuộc tính quyết định Shape có 3 giá trị là Circle, square và Triangle

Bảng 1.2. Bảng quyết định

	Color	Size	Shape[D]
1	Green	Big	Circle
2	Green	Small	Circle
3	Yellow	Medium	Square
4	Red	Medium	Square
5	Yellow	Medium	Triangle
6	Green	Big	Circle
7	Red	Small	Triangle
8	Red	Small	Triangle

1.2.2. Quan hệ bất khả phân biệt

Một hệ quyết định (hay bảng quyết định) đại diện cho kiến thức về đối tượng. Tồn tại hai khả năng dư thừa thông tin: Các đối tượng giống nhau, không phân biệt được lặp lại nhiều lần hoặc thậm chí một số các thuộc tính có thể là dư thừa.

Quan hệ hai ngôi $R \subseteq X \times X$ là tương đương khi:

- R có tính phản xạ : xRx , với mọi $x \in X$
- R có tính đối xứng: xRy → yRx, với mọi $x, y \in X$
- R có tính bắc cầu: xRy và yRz → xRz, với mọi $x, y, z \in X$

Một quan hệ tương đương R phân hoạch tập đối tượng thành các lớp tương đương. Lớp tương đương của một đối tượng $x \in X$ là tập tất cả các đối tượng $y \in X$ có quan hệ R với x: khí hiệu xRy

Với mỗi tập con thuộc tính $B \subseteq (C \cup D)$ tạo ra quan hệ tương đương kí hiệu :

$$IND(B) = \{(x, y) \in U \times U \mid \forall a \in B, f(x,a) = f(y,a)\}$$

IND(B) được gọi là B_quan hệ bất khả phân, nếu $(x, y) \in IND(B)$ thì x,y không thể phân biệt được với nhau qua tập thuộc tính B. Ngược lại là phân biệt được với B. Với mọi $x \in U$ lớp tương đương của x trong quan hệ IND(B) được ký hiệu bởi $[x]_B$. Quan hệ IND(B) phân hoạch tập đối tượng U thành các lớp tương đương được ký hiệu U/IND(B) hay U/B

$$U / IND(B) = U / B = \{[x]_B \mid x \in U\}$$

Trong hệ quyết định, nếu $x, y \in D, \forall a \in C, f(x,a) = f(y,a)$ tồn tại $f(x,D) \neq f(y,D)$ thì (x,y) là nhất quán, ngược lại là không nhất quán.

Ví dụ [1].Tập thuộc tính B= {Color, Size} trong Bảng 1.2 phân hoạch tập 8 đối tượng thành các lớp như sau (u1,u6),(u2),(u3,u5),(u4),(u7,u8)

Nhận xét: Ta thấy, các đối tượng u1và u6, u7 và u8 cùng một lớp tương đương nên chúng không thể phân biệt với nhau trên tập thuộc tính {Color, Size}.

1.2.3. Xấp xỉ tập hợp

Trong lý thuyết tập thô, các khái niệm không rõ ràng được thay bằng cặp khái niệm xấp xỉ: Xấp xỉ dưới bao gồm tất cả các đối tượng chắc chắn thuộc về khái niệm, xấp xỉ trên bao gồm tất cả các đối tượng có thể thuộc về khái niệm.

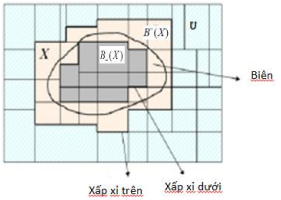
Định nghĩa 2.[6] Cho một bảng quyết định $S = (U, C, D, V, f)$ $\forall B \subseteq C, X \subseteq U$ với

$U / B = \{[x]_B \mid x \in U\} = \{B_1, B_2 \ldots, B_t\}$ thì :

$\underline{B}(X) = \{x \mid [x]_B \subseteq X\}$ là B_xấp xỉ dưới của X

$\overline{B}(X) = \{x \mid [x]_B \cap X \neq \varnothing\}$ là B_xấp xỉ trên của X

Tập $BN_B(X) = \bar{B}(X) - \underline{B}(X)$ được gọi là B_vùng biên của X chứa các đối tượng không thể xác định được có thuộc X hay không. Nếu $BN_B(X) = \varnothing$ (tức $\underline{B}(X) = \bar{B}(X)$) thì X được gọi là tập rõ (crisp), trái lại X được gọi là tập thô (rough)



Hình 1.6. Mô tả bộ xấp xỉ trên-dưới

Định nghĩa 3. Cho bảng quyết định $S = (U, C, D, V, f)$ $\forall B \subseteq C, X \subseteq U$ với $U/B = \{[x]_B \mid x \in U\} = \{B_1, B_2 \ldots, B_t\}$

Tập $POS_B(X) = \underline{B}(X)$ gọi là B_miền dương của X

$NEG_B(X) = U - \bar{B}(X)$ gọi là B_miền âm của X (hay gọi là miền ngoài chứa những đối tượng mà sử dụng thuộc tính trong B chắc chắn không thuộc tập X)

Bảng 1.3. Các triệu chứng cảm cúm

U	Đau đầu	Thân nhiệt	Cảm cúm
u1	Có	Bình thường	Không
u2	Có	Cao	Có
u3	Có	Rất cao	Có
u4	Không	Bình thường	Không
u5	Không	Cao	Không
u6	Không	Rất cao	Có
u7	Không	Cao	Có
u8	Không	Rất cao	Không

Ví dụ [1] Bảng 1.3, B={Đau đầu, Thân Nhiệt} ta có các lớp không phân biệt được là {u1},{u2}, {u3},{u4}, {u5,u7},{u6,u8}

Nếu đặt X={u|u(Cảm cúm)=Có}={u2,u3,u6,u7}, lúc đó ta có: $\underline{B}(X)$ = {u2,u3} và $\bar{B}(X)$= {u2,u3,u5,u7,u6,u8}.

Từ đó ta có B-miền biên của X là tập: $BN_B(X) = \bar{B}(X) - \underline{B}(X)$ ={ u5,u6,u7,u8}.

1.2.4. Thuộc tính thiết yếu và không thiết yếu

Định nghĩa 4.[6] Cho một bảng quyết định $S = (U, C, D, V, f)$, $\forall b \in B \subseteq C$

Nếu $POS_B(D) = POS_{B-\{b\}}(D)$ thì b là không thiết yếu (not necessary) trong B đối với D; Ngược lại, thì b là thiết yếu trong B đối với D

Bảng 1.4. Bảng quyết định về bệnh cúm

U	Đau đầu	Đau cơ	Thân nhiệt	Cảm cúm
u1	Có	Có	Bình thường	Không
u2	Có	Có	Cao	Có
u3	Có	Có	Rất cao	Có
u4	Không	Có	Bình thường	Không
u5	Không	Không	Cao	Không
u6	Không	Có	Rất cao	Có

Cho $\forall B \subseteq C$ nếu mọi phần tử trong B đối với D là thiết yếu, thì B đối với D là độc lập.

Bảng 1.4 có 2 tập rút gọn là {Đau đầu, Thân nhiệt} và {Đau cơ, Thân nhiệt}. Tập lõi {Thân nhiệt}.

1.3. Giải thuật di truyền

Giải thuật di truyền cung cấp cách tiếp cận dựa vào mô phỏng sự tiến hóa. Các giả thuyết thường được mô tả bằng các chuỗi bit, hoặc bằng các biểu thức ký hiệu, có khi ngay cả các chương trình máy tính. Giải thuật di truyền đã được ứng dụng thành công vào các lĩnh vực khác nhau và việc tối ưu hóa khác. Việc kết hợp giải thuật di truyền sẽ giúp quá trình phân cụm tối ưu hơn.

1.3.1. Thông tin

Giải thuật di truyền (Genetic Algorithm – GA), do John Holland (1975) và Goldberg (1989) đề xuất và phát triển. Ý tưởng của giải thuật di truyền là mô phỏng theo cơ chế của quá trình tiến hóa trong tự nhiên. Từ tập các lời giải ban đầu, thông qua nhiều bước tiến hóa để hình thành các tập mới với những lời giải tốt hơn, cuối cùng sẽ tìm được lời giải gần tối ưu nhất.

GA sử dụng các thuật ngữ lấy từ di truyền học:

- Lớp hay quần thể (Population) là một tập hợp các lời giải
- Một nhiễm sắc thể (NST) hay cá thể (Chromosome) biểu diễn cho mỗi lời giải
- NST được tạo thành từ các Gen

Quá trình tiến hóa được thực hiện trên một quần thể như là sự tìm kiếm các lời giải có thể của bài toán. Nó đòi hỏi sự cân bằng giữa hai mục tiêu chính là khai thác lời giải tốt nhất và tra xét toàn bộ không gian tìm kiếm. GA thực hiện tìm kiếm theo nhiều hướng: duy trì tập hợp các lời giải có thể và khuyến khích sự hình thành trao đổi thông tin giữa các hướng. Tập lời giải cần trải qua nhiều bước tiến hoá, một tập mới các cá thể được tạo ra tại mỗi thế hệ có chứa các cá thể thích nghi nhất trong thế hệ cũ. Đồng thời khai thác có hiệu quả thông tin trước đó để suy xét trên điểm tìm kiếm mới với mong muốn có được sự cải thiện qua từng thế hệ.

1.3.2. Các thành phần cơ bản trong giải thuật di truyền

1.3.2.1. Mã hóa nhiễm sắc thể

Trong giải thuật di truyền, cách mã hóa nhiễm sắc thể quyết định đến hiệu quả của giải thuật và ảnh hưởng đến việc lựa chọn các toán tử trong các bước lai ghép và đột biến. Với mỗi kiểu bài toán khác nhau có nhiều cách mã hóa nhiễm sắc thể. Cách mã hoá nhiễm sắc thể là một trong yếu tố quyết định trong xây dựng giải thuật di truyền gồm: mã hóa nhị phân; mã hóa hoán vị; mã hóa theo giá trị.

Mã hoá nhị phân

Phương pháp mã hoá nhị phân là phương pháp mã hoá NST đơn giản và phổ biến nhất. Trong đó, mỗi NST là một chuỗi nhị phân, mỗi bit trong nó có thể biểu diễn một đặc tính của lời giải.

Nhược điểm là tạo ra không gian mã hoá lớn hơn so với không gian giá trị của NST, hơn nữa có thể xảy ra trường hợp các toán tử lai ghép và đột biến tạo ra các cá thể không nằm trong không gian tìm kiếm và đòi hỏi phải có những phương pháp sửa chữa để làm cá thể tạo ra nằm trong không gian tìm kiếm.

Mã hoá hoán vị

Mã hoá hoán vị thường được sử dụng trong các bài toán liên quan đến thứ tự, trong đó mỗi NST là một chuỗi các số biểu diễn

một trình tự. Việc thao tác trên các NST chính là hoán vị các số trong chuỗi đó làm thay đổi trình tự của nó.

Mã hoá theo giá trị

Mã hoá trực tiếp theo giá trị dùng trong các bài toán sử dụng giá trị phức tạp như trong số thực. Trong đó, mỗi NST là một chuỗi các giá trị. Các giá trị có thể là bất cứ cái gì liên quan đến bài toán, từ số nguyên, số thực, kí tự cho đến các đối tượng phức tạp hơn. Trong cách mã hoá này ta thường phải phát triển các toán tử đột biến và lai ghép cho phù hợp với từng bài toán.

1.3.2.2. Quần thể ban đầu (The initial population)

Quần thể ban đầu[6] là một khía cạnh quan trọng của thuật toán di truyền, có các đặc tính ảnh hưởng quan trọng đến kết quả và hiệu quả. Để đạt được sự tối ưu toàn cầu, quần thể ban đầu trong không gian giải pháp nên được phân phối. Thuật toán di truyền chuẩn được khởi tạo ngẫu nhiên quần thể ban đầu, có thể dẫn đến phân phối không đồng đều trong không gian giải pháp, do đó ảnh hưởng đến hiệu suất của thuật toán, nếu kích cỡ quần thể quá nhỏ thì tính đa dạng của quần thể bị hạn chế; còn nếu quá lớn sẽ hao phí tài nguyên và làm chậm quá trình. Thường phụ thuộc vào kích thước chuỗi mã hóa.

1.3.2.3. Hàm thích nghi (The fitness function)

Trong các thuật toán di truyền, mức phạt là giải pháp tối ưu có thể đạt được bằng cách sử dụng các hàm thích nghi để đo độ tối ưu hóa của mỗi cá thể trong nhóm. Được định nghĩa là hàm đánh giá hay hàm mục tiêu thể hiện tính thích nghi của cá thể hay độ tốt của lời giải.

Trong bài, các hàm thích nghi được xây dựng [6] như sau:

$$F(v) = \beta.f(x) + p(x) = \beta \times \left(1 - \frac{card(x)}{card(C)} + \frac{card\left(POS_X(D)\right)}{card\left(POS_C(D)\right)}\right)$$

Trong đó :

$f(x) = \left(1 - \frac{card(x)}{card(C)}\right)$ chỉ ra các NST x không gồm tỷ lệ các thuộc tính điều kiện.

$\beta = \frac{1}{card\left(POS_C\left(D\right)\right)}$ đảm bảo tập siêu tính toán thuộc tính giảm vượt trội để giảm số lượng các thuộc tính điều kiện.

$p\left(x\right) = \frac{card\left(POS_x\left(D\right)\right)}{card\left(POS_C\left(D\right)\right)}$ chỉ ra khả năng phân biệt các thuộc tính x.

1.3.2.4. Chọn lọc

Quá trình chọn lọc và đấu tranh sinh tồn trong tự nhiên làm thay đổi các cá thể trong quần thể. Những cá thể có khả năng thích nghi tốt với điều kiện sống thì có sự đấu tranh lớn hơn, do đó có thể tồn tại và sinh sản. Ngược lại các thể sẽ dần mất đi. Dựa vào nguyên lý này, chọn lựa các cá thể trong GA chính là cách chọn các cá thể có độ thích nghi tốt để đưa vào thế hệ tiếp theo hoặc để cho lai ghép, với mục đích là sinh ra các cá thể mới tốt hơn. Mục tiêu cuối cùng của phép chọn lọc là chọn ra các cá thể tốt với khả năng cao hơn.

Cơ chế lựa chọn

Cơ chế lựa chọn áp dụng khi quần thể P(t+1) được tạo ra từ việc chọn các cá thể từ quần thể P(t) để thực hiện việc lai ghép và đột biến. Một số cơ chế hay áp dụng để lựa chọn các cá thể từ một quần thể được giới thiệu dưới đây

Cơ chế lựa chọn theo bánh xe Roulet được thực hiện bằng cách quay bánh xe Roulet n lần. Mỗi lần chọn một NST từ quần thể hiện hành vào quần thể mới.

Với cơ chế lựa chọn như thế này thì có một số NST sẽ được chọn nhiều lần. Điều này phù hợp với lý thuyết lược đồ: *Các NST tốt nhất thì có nhiều bản sao, NST trung bình thì không đổi , NST kém thì mất đi.*

Lựa chọn xếp hạng

Sắp xếp các NST trong quần thể theo độ thích nghi từ thấp đến cao.Đặt lại độ thích nghi cho quần thể đã sắp xếp theo kiểu.Theo phương pháp này việc một NST được chọn nhiều lần như trong lựa chọn theo kiểu bánh xe Roulet đã giảm đi. Nhưng

nó có thể dẫn đến sự hội tụ chậm và NST có độ thích nghi cao cũng không khác mấy so với các NST khác.

Lựa chọn theo cơ chế lấy mẫu ngẫu nhiên

Biểu diễn xác suất chọn các NST lên trên một đường thẳng.Đặt N điểm chọn lên đường thẳng. Các điểm chọn này cách nhau 1/N, điểm đầu tiên đặt ngẫu nhiên trong khoảng [0,1/N] Với một điểm chọn, NST gần với nó nhất về bên phải sẽ được chọn. Phương pháp này có đặc điểm là các điểm chọn được phân bố đều trên trục số.

Lựa chọn tranh đấu

Lấy một số NST trong quần thể, NST nào có độ thích nghi cao nhất được chọn. Lặp lại thao tác trên N lần.

1.3.2.5. Lai ghép

Sự kết hợp giữa tính trạng của cha mẹ để sinh ra thế hệ con là quá trình lai ghép trong tự nhiên. Trong giải thuật di truyền, lai ghép là sự tổ hợp lại các tính chất trong hai lời giải cha mẹ nào đó để sinh ra lời giải mới mà có đặc tính mong muốn tốt hơn thế hệ cha mẹ.

1.3.2.6. Đột biến

Đột biến được định nghĩa là sự biến đổi tại một hay một số gen của NST ban đầu để tạo ra một NST mới. Nó có thể tạo ra một cá thể mới tốt hoặc xấu hơn cá thể ban đầu.Tuy nhiên trong giải thuật di truyền, ta luôn muốn tạo ra những phép đột biến cho phép cải thiện lời giải qua từng thế hệ.

1.3.3. Quy trình thuật toán di truyền

Giải thuật di truyền tổng quát [5] được mô tả như sau:

Bước 1.[Bắt đầu] Tạo ngẫu nhiên quần thể P(t) có n NST
Bước 2.[Thích nghi] Đánh giá độ thích nghi f(x) cho mỗi NST x quần thể P(t).
Bước 3.[Quần thể mới] Tạo ra 1 quần thể mới bằng cách lặp lại các bước dưới đây đến khi quần thể mới hoàn thiện.
(a). [Lựa chọn] Lựa chọn 2 nhiễm sắc thể cha mẹ từ quần thể trên dựa vào độ thích nghi của chúng (độ thích nghi tốt hơn, cơ hội được lựa chọn lớn hơn).
(b). [Lai ghép] lai chéo các cha mẹ để tạo ra con mới.

Nếu lai chéo không được thực hiện, con cái chính là 1 bản y sao của cha mẹ.

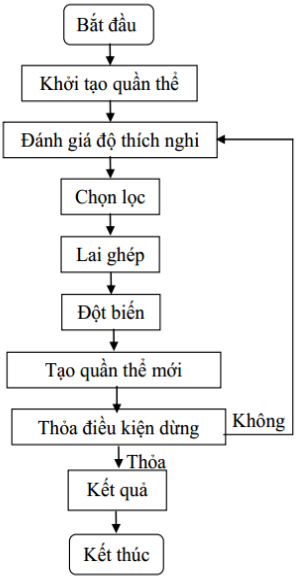
(c). [Đột biến] đột biến con mới tại mỗi quỹ tích (vị trí trong nhiễm sắc thể).

(d). [Chấp nhận] Đặt con mới vào trong quần thể mới

Bước 4. [Thay thế] Sử dụng quần thể mới đươc tạo ra cho bước lặp tiếp theo

Bước 5. [Kiểm thử] Nếu điều kiện cuối cùng được thỏa mãn, dừng lặp và trả về giải pháp tốt nhất trong quần thể đang xét.

Bước 6. [Lặp] Quay trở về bước lặp thứ 2



Hình 1.7. Sơ đồ giải thuật di truyền.

1.3.4. Các thông số cơ bản của giải thuật di truyền

- *Size(n) là kích thước(số lượng cá thể) của quần thể,*
- *Pc*: *xác suất lai ghép*
- *Pm*: *xác suất đột biến*

CHƯƠNG II. PHÂN CỤM DỮ LIỆU DỰA TRÊN TẬP THÔ VÀ GIẢI THUẬT DI TRUYỀN

2.1. Giới thiệu

Một chiến lược cải tiến mới kết hợp giữa thuật toán K-Means có khả năng tìm kiếm địa phương mạnh mẽ với thuật toán di truyền có tối ưu toàn cầu. Sử dụng lý thuyết tập thô để giải quyết sự thiếu chính xác và không đầy đủ tri thức. Thông qua các quy định phù hợp và áp dụng lợi thế của thuật toán, tính chính xác cụm được cải thiện. Các kết quả thực nghiệm cho thấy rằng thuật toán phân cụm được cải tiến tốt hơn cho việc phân cụm dữ liệu thông thường.

2.2. Phương pháp phân cụm tập thô

Khái niệm về phân cụm thô tương tự như lý thuyết tập thô - với bộ xấp xỉ dưới và trên - cho phép các đối tượng thuộc nhiều cụm trong tập hợp dữ liệu. Theo định nghĩa, xấp xỉ dưới của một cụm thô chứa các đối tượng mà nó chắc chắn thuộc về cụm đó, và các đối tượng thuộc về xấp xỉ trên có thể thuộc về nhiều hơn một cụm.

Đối với kỹ thuật phân cụm, lý thuyết tập thô được tiếp cận hỗ trợ phân cụm dựa vào hai hướng [3]:

- Cải tiến các thuật toán phân cụm cổ điển như K-Means, K-Medoid thành Rough K-Means, Rough K-Medoid... bằng cách kết hợp các khoảng cách hay độ tương đồng với các phép xấp xỉ.
- Hỗ trợ xác định số lượng phân cụm tối thiểu: Dựa trên số lượng phân cụm gợi ý ban đầu do người sử dụng cung cấp, các cụm sẽ được gom lại nếu các xấp xỉ trên các phân cụm giao nhau khác rỗng.

Trong phân cụm thô ta không xét tất cả các thuộc tính của tập thô. Tuy nhiên, bộ xấp xỉ trên và dưới được yêu cầu phải làm theo một số các thuộc tính tập thô cơ bản như sau: Cho tập đối tượng U với $X \subseteq U$, cặp $\left(\underline{A}(X), \overline{A}(X)\right)$ $v \in U$

- Một đối tượng v thuộc nhiều nhất một xấp xỉ dưới $\underline{A}(X)$.

- Một đối tượng v thuộc xấp xỉ dưới của một tập thì cũng thuộc xấp xỉ trên của nó ($v \in \underline{A}(X) \rightarrow v \in \overline{A}(X)$).

- Nếu một đối tượng v không thuộc bất kỳ xấp xỉ dưới $\underline{A}(X)$ thì nó thuộc hai hoặc nhiều hơn xấp xỉ trên

K-Means sử dụng lý thuyết tập thô (Rough K-Means)

Phương pháp phân cụm thô [13] phổ biến nhất được bắt nguồn từ phân cụm K-Means cổ điển. Tạo ngẫu nhiên k cụm từ n đối tượng. Giả định rằng các đối tượng được biểu diễn bằng vector m chiều.

Mục tiêu là để chỉ định các đối tượng n vào k cụm. Mỗi cụm cũng được đại diện bởi một vector m chiều, đó là trọng tâm hay vector cho cụm đó. Quá trình bắt đầu bằng cách chọn ngẫu nhiên k trọng tâm của k cụm. Các đối tượng được gán cho một trong những cụm k dựa trên giá trị tối thiểu của khoảng cách d(v, x) giữa các vector đối tượng $v = \{v_1, \ldots, v_j, ..., v_m\}$ và vector cụm $x = \{x_1, \ldots, x_j, ..., x_m\}$ với $1 \le j \le m$

Khoảng cách d (v, x) được cho như sau: $d(v, x) = \| v - x \|$ ở đây thường là chuẩn Euclide.

Sau khi phân tất cả các đối tượng vào các cụm khác nhau, các vector trọng tâm mới của các cụm được tính như sau: $x_j = \frac{\sum_{v \in x} v_j}{|x|}$ với |x| là size của cụm x. Quá trình dừng lại khi các trọng tâm của cụm ổn định, tức là các vector trọng tâm lặp lại trước đó trùng với trọng tâm cụm mới trong lần lặp hiện tại.

Kết hợp bộ thô vào phân cụm K-Means đòi hỏi việc bổ sung các khái niệm về xấp xỉ dưới và trên.

Đặc biệt: Tính toán các trọng tâm cần phù hợp và được quyết định liệu một đối tượng được gán cho một xấp xỉ thấp hơn hoặc trên của một cụm.

Các vấn đề sẽ giải chi tiết như sau:

(1) *Tính toán các trọng tâm*. Tính toán của các trọng tâm của cụm từ K-Means cổ điển cần phải được sửa đổi bao gồm có cả xấp xỉ dưới và xấp xỉ trên.

Nếu $[\underline{A}(x) \neq \varnothing \,\&\, \overline{A}(x) - \underline{A}(x) = \varnothing]$

Thì $\left[x_j = \frac{\sum_{v \in \underline{A}(x)} v_j}{|\underline{A}(x)|} \right]$

Ngược lại Nếu $[\underline{A}(x) = \varnothing \,\&\, \overline{A}(x) - \underline{A}(x) \neq \varnothing]$

thì $\left[x_j = \frac{\sum_{v \in (\overline{A}(x) - \underline{A}(x))} v_j}{|\overline{A}(x) - \underline{A}(x)|} \right]$

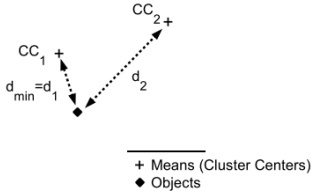
Ngược lại $\left[x_j = w_{lower} \frac{\sum_{v \in \underline{A}(x)} v_j}{|\underline{A}(x)|} + w_{upper} \frac{\sum_{v \in (\overline{A}(x) - \underline{A}(x))} v_j}{|\overline{A}(x) - \underline{A}(x)|} \right]$

với $1 \leq j \leq m$ và $w_{lower} + w_{uper} = 1$

Các thông số w_{lower} và w_{upper} tương ứng với độ tương đối của xấp xỉ dưới và xấp xỉ trên.

(2) *Quyết định xem một đối tượng được gán cho một xấp xỉ dưới hoặc trên của một cụm.*

Về cơ bản, một đối tượng sẽ được giao cho xấp xỉ dưới của một cụm khi khoảng cách giữa các đối tượng và các cụm trung tâm nhỏ hơn nhiều so với khoảng cách tới các trung tâm cụm còn lại khác (Hình 2.1).



Hình 2.1. Mô tả khoảng cách giữa đối tượng tới tâm cụm

Cụ thể, đối với mỗi vector đối tượng v, ***d(v, x_j)*** là khoảng cách giữa v và trọng tâm của cụm x_j.

Bước 1. Xác định trọng tâm gần nhất: $d_{\min} = d\left(v, x_i\right) = \min_{1 \le j \le k} d\left(v, x_j\right)$

Bước 2. Kiểm tra khoảng cách với trọng tâm cụm gần nhất và các trọng tâm khác $T = \left\{ \mathrm{t} : d\left(v, x_i\right) - d\left(v, x_j\right) \le \mathrm{Thresold}, \mathrm{i} \ne \mathrm{j} \right\}$

Nếu $T \ne \varnothing$ thì v thuộc xấp xỉ trên của 2 hoặc nhiều cụm

Nếu $T = \varnothing$ thì v thuộc xấp xỉ dưới chỉ của một cụm

Do đó, ta có được các nguyên tắc sau cho sự phân công của các đối tượng đến xấp xỉ:

Nếu $[T \ne \varnothing]$

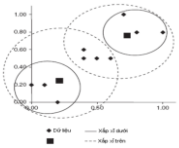
Thì $\left[v \in \overline{A}\left(x_i\right) \& v \in \overline{A}\left(x_j\right), \forall j \in T \right]$

Ngược lại $\left[v \in \overline{A}\left(x_i\right) \& v \in \underline{A}\left(x_i\right) \right]$

Ý tưởng thuật toán Rough K-Means có thể tóm tắt:

Bước 1. Khởi tạo: Chọn ngẫu nhiên *k* tâm các cụm xuất phát $x = \left\{ x_1, ..., x_k \right\}$

Bước 2. Gom cụm các đối tượng dựa vào xấp xỉ trên và xấp xỉ dưới (mỗi cụm gồm 2 tập: xấp xỉ dưới vàxấp xỉ trên - Hình 2.2)



Hình 2.2. Mô tả gom cụm vào bộ xấp xỉ trên - dưới

Xác định các thành viên của cụm như sau:

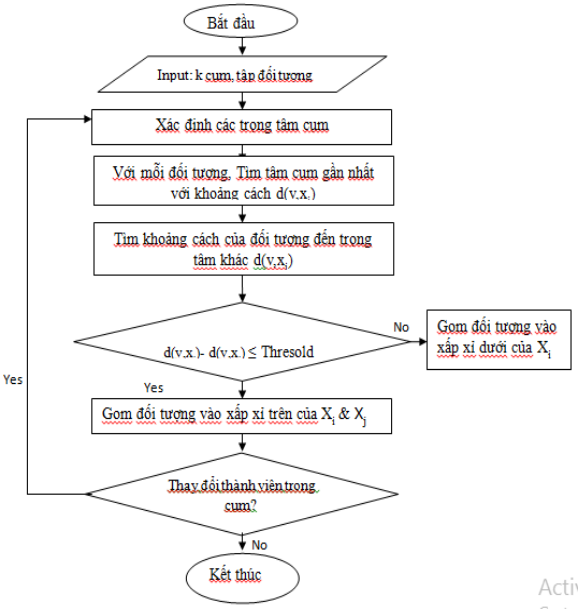
- Nếu $d\left(v, x_i\right) - d\left(v, x_j\right) \le \mathrm{Thresold}$ $1 \le$ i, j $\le$ k thì $v \in \overline{A}\left(x_i\right) \& v \in \overline{A}\left(x_j\right)$ và v sẽ không thuộc bất kỳ xấp xỉ dưới.
- Ngược lại, $v \in \overline{A}\left(x_i\right) \& v \in \underline{A}\left(x_i\right)$ như vậy là d (v, x_i) là tối thiểu, $1 \le \mathrm{i} \le \mathrm{k}$.

Bước 3. Cập nhật lại trọng tâm x_i bằng trọng tâm mới

$$x_j = \begin{cases} w_{lower} \times \dfrac{\sum_{v \in \underline{A}(x)} v_j}{|\underline{A}(x)|} + w_{upper} \times \dfrac{\sum_{v \in \overline{A}(x) - \underline{A}(x)} v_j}{|\overline{A}(x) - \underline{A}(x)|} & \text{Nếu } \overline{A}(x) - \underline{A}(x) \neq \varnothing \\ \dfrac{\sum_{v \in \underline{A}(x)} v_j}{|\underline{A}(x)|} & \text{Ngược lại} \end{cases}$$

Trong đó w_{lower}, w_{upper} là các trọng số thỏa $w_{lower} + w_{upper} = 1$.

- Nếu tiêu chuẩn hội tụ được đáp ứng, nghĩa là trung tâm cụm trùng với lần lặp trước thì dừng lại; Ngược lại đi đến bước 2.



Hình 2.3. Sơ đồ phân cụm K-Means thô

So sánh phân cụm thô và phân cụm K-Means

Voges [16] có sự so sánh phân cụm thô với phân cụm K-Means, nhận thấy hai kỹ thuật phân cụm này đều xác định số cụm cụ thể được sử dụng. Giải pháp phân cụm thô khác so với K-Means là khả năng nhóm đối tượng trong nhiều cụm khác nhau. Phân cụm thô cũng tạo nhiều cụm hơn phân cụm K-Means [16], với số lượng cụm cần thiết để mô tả dữ liệu phụ thuộc vào

khoảng cách đo. Nhiều cụm có nghĩa là một đối tượng có cơ hội cao trong hơn một cụm. Một giải pháp với quá ít các cụm không cung cấp một giải pháp hữu ích các phân vùng của dữ liệu. Mặt khác, quá nhiều cụm làm cho lời giải khó khăn. Ngoài ra, mức độ trùng lặp giữa các cụm được giảm thiểu để đảm bảo rằng mỗi cụm được cung cấp thông tin để hỗ trợ trong việc giải thích.

2.3. Phương pháp phân cụm dựa trên giải thuật di truyền

GA là một quá trình tìm kiếm dựa trên các nguyên tắc của sự tiến hóa thông qua chọn lọc tự nhiên. Các thành phần quan trọng bao gồm: gen, nhiễm sắc thể, quần thể, thế hệ, hàm thích nghi, lựa chọn, lai ghép và đột biến

K-Means sử dụng giải thuật di truyền

Thuật toán K-Means được sửa đổi để thích ứng với các nguyên tắc của GA.

Đầu vào: Số cụm k, kích thước của quần thể, tập dữ liệu D chứa n đối tượng, số thế hệ muốn tạo tMax

Đầu ra: Một tập hợp K cụm.

Bước 1: Khởi tạo Mỗi NST được tạo bằng cách chọn ngẫu nhiên k phần tử trong tập dữ liệu để làm k trọng tâm cụm

Bước 2: Lặp từ t =1 đến tMax

1, Đối với mỗi nhiễm sắc thể

- Đưa phần tử trong D vào cụm với trọng tâm cụm gần nhất
- Tính lại k tâm cụm là trung bình k cụm vừa tạo và thay thế vào NST đó
- Tính toán độ thích nghi cho NST hiện tại

2, Tạo thế hệ các NST mới sử dụng các phép toán lựa chọn, lai ghép và đột biến.

3, Sắp xếp các cá thể sau đột biến theo thứ hạng (Chọn ra cá thể có độ thích nghi tốt nhất)

Bước 3: In kết quả Tách ra k cụm đối với NST trong quần thể của thế hệ tạo ra sau cùng có độ thích nghi lớn nhất.

Điều kiện dừng: Lặp lại bước 2 đến khi thế hệ t = tMax.

So sánh Phân cụm K-Means và GA K-Means

Thuật toán phân cụm phổ biến thường được sử dụng là K-Means - là một kỹ thuật phân nhóm đơn giản và hiệu quả nhưng kết quả chưa chắc đã đạt giá trị tối ưu vì kết quả phụ thuộc vào việc lựa chọn trung tâm của các cụm ban đầu.

Giải thuật di truyền là một giải thuật tìm kiếm ngẫu nhiên dựa trên sự tiến hóa và di truyền học tự nhiên, đồng thời nó có một số lượng lớn các giá trị tiềm ẩn song song, vì vậy nó cung cấp các giải pháp tối ưu cho các đối tượng hoặc các hàm thích nghi. Bảng 2.1 sẽ đưa ra so sánh về hai giải thuật:

Bảng 2.1. So sánh về hai giải thuật K-Means, di truyền

K-Means	**K-Means sử dụng di truyền**
- Đầu vào: k, bộ dữ liệu; k trung tâm cụm được lựa chọn ngẫu nhiên	- Đầu vào: k, bộ dữ liệu; Quần thể P, p nhiễm sắc thể được chọn ngẫu nhiên; Tmax
- Phương pháp phân hoạch	- Phương pháp tiến hóa
- Mục tiêu: Giảm thiểu tổng bình phương khoảng cách	- Mục tiêu: Giảm thiểu tổng khoảng cách từ mỗi điểm dữ liệu đến trọng tâm của cụm
- Điều kiện dừng: Không có sự thay đổi trong trung tâm cụm mới	- Điều kiện dừng: Số lần lặp đạt giá trị tối đa
- Độ phức tạp: O(n*k*d*i) Trong đó: + n: là số điểm dữ liệu + k: số cụm + d: kích thước dữ liệu + i: số lần lặp	- Độ phức tạp: O(Tmax*p*n*k*d) Trong đó: + n: là số điểm dữ liệu + k: số cụm + d: kích thước dữ liệu + Tmax: số lần lặp + p: kích cỡ quần thể

2.4. Phương pháp phân cụm dựa trên tập thô và giải thuật di truyền

Phương pháp phân cụm hiệu quả dựa vào tập thô và thuật toán di truyền được cung cấp [6].

Đầu vào: n đối tượng dữ liệu, số cụm k

Đầu ra: Đầu ra là các trung tâm cụm tương ứng với các thành phần có giá trị hàm thích nghi lớn nhất.

Bước 1. **Khởi tạo** k số cụm, quần thể ngẫu nhiên P có p nhiễm sắc thể, chọn ra k tâm cụm, số thế hệ muốn lặp tMax. Mã hóa k cụm.

Bước 2. Phân cụm thô: Giải mã mỗi nhiễm sắc thể, gom các đối tượng tương ứng với mỗi k cụm phù hợp với nguyên tắc về khoảng cách, sau đó làm theo phân cụm K-Means thô để phân phối các đối tượng.

Bước 3. Tính toán các giá trị hàm thích nghi.

Bước 4. Lựa chọn, lai ghép và đột biến.

Bước 5. Đánh giá lại quần thể mới. Nếu số lần lặp bằng với giá trị tối đa được xác định, chuyển sang bước 6, nếu không, các thuật toán tiếp tục từ bước 2 đến bước 4.

Bước 6. Kết thúc thuật toán

Chiến lược mã hóa nhị phân: Nếu đối tượng trong tập dữ liệu thuộc biên hoặc miền âm trong các cụm, thì mã tương ứng của chuỗi nhiễm sắc thể là 1, ngược lại là 0.

Cơ chế để ngăn chặn cận huyết [6]

Để duy trì sự đa dạng của các quần thể khi lựa chọn các cá thể kết nối, ở đây sử dụng cơ chế để ngăn chặn sự cận huyết, hạn chế cá thể tương tự lại kết đôi. Cụ thể, chọn xác suất hai cá thể, nếu khoảng cách Hamming giữa chúng nhỏ hơn so với ngưỡng cho trước, thì lai gép chúng trong quần thể; nếu không, quay lại và tiếp tục chọn lần nữa.

Chiến lược Elitist [6] (Chọn lọc ưu tú)

Để bảo tồn các cá thể tốt nhất của giá trị hàm thích nghi trong cá thể, trong bài sử dụng chiến lược chọn lọc ưu tú, có nghĩa là sao chép cá thể có giá trị thích nghi cao nhất trong quần thể hiện tại sang quần thể mới, và các cá thể này không tham gia vào các hoạt động của lai ghép và đột biến.

CHƯƠNG III. CÀI ĐẶT VÀ PHÂN TÍCH THÍ NGHIỆM

3.1. Dữ liệu thử nghiệm

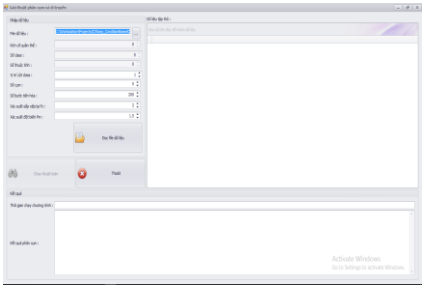
Để xác minh tính hợp lệ của các thuật toán phân cụm, chúng ta sử dụng bộ dữ liệu trong cơ sở dữ liệu UCI học máy để kiểm tra thuật toán này. Nguồn dữ liệu website:*ftp://ftp.ics.uci.edu/pub/machine-learning-databases*

Sử dụng bộ dữ liệu zoo để phân cụm, là bộ dữ liệu đơn giản có 17 thuộc tính (15 thuộc tính Boolean, 2 numerics) thuộc tính "type" là thuộc tính lớp. Số các trường hợp là: 101

Các thông số thí nghiệm như sau: số cụm k được thay đổi trong khi các tham số khác là cố định, kích thước quần thể ban đầu bằng số trường hợp trong bộ dữ liệu; các thuật toán chạy t= 100 lần liên tục; Pc=0.3, Pm=0.02.

3.2. Cài đặt thuật toán

Chương trình cài đặt thuật toán xây dựng đặc trưng dựa trên thuật toán K-Means kết hợp giải thuật di truyền để phân cụm tập dữ liệu thô được viết bằng ngôn ngữ C# trong môi trường .Net Framework, sử dụng bộ công cụ visual studio 2013 kết hợp DevExpress

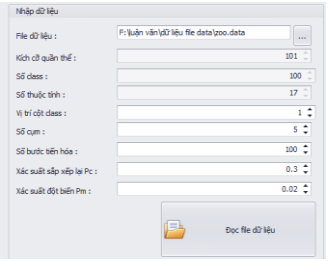


Hình 3.1. Giao diện chương trình

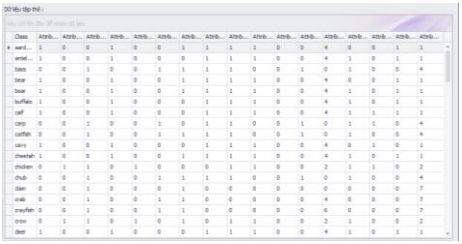
Chương trình gồm modul chính:

- Module 1: Khai báo các thuộc tính
- Module 2: Đọc file dữ liệu tập thô
- Module 3: Phân cụm một tập dữ liệu và đánh giá cụm

Chọn và tải dữ liệu bộ

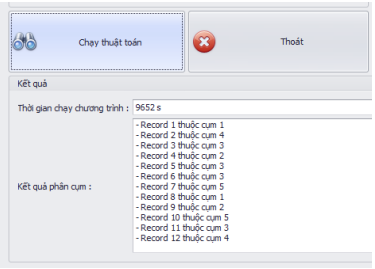


Hình 3.2. Giao diện nhập dữ liêu và các thuộc tính



Hình 3.3. Giao diện hiển thị file dữ liệu

Phân cụm tập dữ liệu và đánh giá cụm



Hình 3.4. Giao diện kết quả thuật toán

3.3. Kết quả thử nghiệm

Bảng 3.1. Kết quả thực nghiệm với phân cụm K-Means thông thường

K-Means						
Lượt test	1	2	3	4	5	Giá trị trung bình
Cụm số	K=3					
1	41,6%	41,6%	41,6%	41,6%	41,6%	**41,6%**
2	27,7%	27,7%	27,7%	27,7%	27,7%	**27,7%**
3	30,7%	30,7%	30,7%	30,7%	30,7%	**30,7%**
Thời gian chạy	0.01	0.01	0.02	0	0	**0.008**
	K=5					
1	40,6%	40,6%	40,6%	40,6%	40,6%	**40,6%**
2	17,8%	17,8%	17,8%	17,8%	17,8%	**17,8%**
3	20,8%	20,8%	20,8%	20,8%	20,8%	**20,8%**
4	15,8%	15,8%	15,8%	15,8%	15,8%	**15,8%**
5	5%	5%	5%	5%	5%	**5%**
Thời gian chạy	0.02	0	0.02	0	0.01	**0.01**
	K=7					
1	5,9%	5,9%	5,9%	5,9%	5,9%	**5,9%**
2	18,8%	18,8%	18,8%	18,8%	18,8%	**18,8%**
3	10,9%	10,9%	10,9%	10,9%	10,9%	**10,9%**
4	12,9%	12,9%	12,9%	12,9%	12,9%	**12,9%**
5	7,9%	7,9%	7,9%	7,9%	7,9%	**7,9%**
6	34,7%	34,7%	34,7%	34,7%	34,7%	**34,7%**
7	8,9%	8,9%	8,9%	8,9%	8,9%	**8,9%**
Thời gian chạy	0.02	0	0.02	0	0.02	**0.012**

Bảng 3.2. Kết quả thực nghiệm với phân cụm dựa trên tập thô và giải thuật di truyền

GA rough K-Means						
Lượt test	1	2	3	4	5	Giá trị trung bình
Cụm số	K=3					
1	39,6%	40,6%	30,7%	37,6%	44,6%	**38,6%**
2	29,7%	28,7%	36,6%	35,7%	28,7%	**31,9%**
3	30,7%	30,7%	32,7%	26,7%	26,7%	**29.5%**
Thời gian chạy	9440	10437	9016	9407	9054	**9470.8**
	K=5					
1	19,8%	21,8%	23,8%	23,8%	24,8%	**22,8%**
2	20,8%	20,8%	19,8%	18,8%	16,8%	**19,4%**
3	13,8%	13,8%	13,8%	16,8%	23,8%	**16,4%**
4	18,8%	24,8%	18,8%	18,8%	17,8%	**19,8%**
5	26,8%	18,8%	23,8%	21,8%	16,8%	**21.6%**
Thời gian chạy	9661	9650	9513	9667	9545	**9607.2**
	K=7					
1	17,8%	10,9%	7,8%	11,8%	18,8%	**13,4%**
2	14,9%	18,8%	14,9%	11,8%	14,9%	**15.1%**
3	12,9%	14,9%	13,9%	13,9%	8,9%	**12,9%**
4	9,8%	14,9%	17,8%	13,9%	12,9%	**13,9%**
5	12,9%	15,8%	14,9%	19,8%	17,8%	**16.2%**
6	13,9%	12,9%	13,9%	13,9%	14,9%	**13,9%**
7	17,8%	11,8%	16,8%	14,9%	11,8%	**14.6%**
Thời gian chạy	9944	9944	9865	10027	10120	**9980**

Từ bảng 3.1 và 3.2 cho thấy sự so sánh của giải thuật K-Means thông thường với GA thô K-Means. Kết quả bao gồm giá trị tỉ lệ gom các đối tượng vào các cụm và giá trị trung trung bình thời gian từ bộ thử nghiệm. Có thể thấy GA thô K-Means cải thiện kết quả của K-Means qua từng lần thí nhiệm với số cụm xác định trước. Thời gian tính toán của phân cụm dựa trên tập thô và giải thuật di truyền có chậm hơn nhưng việc chọn lọc các đối tượng vào các cụm là đa dạng, đồng đều hơn cho mỗi lần chạy.

Kết quả thực nghiệm đối với thuật toán mới kết hợp tập thô và thuật toán di truyền, đã làm cho độ chính xác phân cụm ưu việt hơn của phân cụm K-Means thông thường. Thuật toán đã đưa ra giải pháp tối ưu toàn cầu và có được kết quả phân cụm tốt hơn.

KẾT LUẬN

Luận văn trình bày khảo cứu một cách có hệ thống của bài báo [6] các kiến thức cơ bản về lý thuyết phân cụm dữ liệu, thuật toán phân cụm K-Means; các khái niệm về lý thuyết tập thô và giải thuật di truyền. Tìm hiểu giải thuật chung cho phân cụm rõ, thô theo hướng thuật toán K-Means và ứng dụng giải thuật di truyền trong phân cụm thô. Tiến hành cài đặt thử nghiệm với bộ dữ liệu trên UCI.

Luận văn đã tìm hiểu chiến lược cải tiến mới là phân cụm dựa trên lý thuyết tập thô và thuật toán di truyền để cải thiện chất lượng phân cụm.

Trên cơ sở các kết quả đạt được, hướng nghiên cứu tiếp như sau:

- Tiếp tục nghiên cứu một số giải thuật phân cụm dựa trên tập thô và giải thuật di truyền.
- Xây dựng tiếp chương trình chạy thử nghiệm các giải thuật phân cụm, cải thiện thuật toán để có chất lượng phân cụm tốt nhất.
- Tìm kiếm các cách thức ứng dụng giải thuật vào thực tiễn.

Do thời gian và hiểu biết về lĩnh vực còn nhiều hạn chế nên luận văn không tránh khỏi những khiếm khuyết.

Tôi xin tiếp thu những góp ý của quý thầy cô, các đọc giả, khắc phục những hạn chế, tiếp tục phát triển đề tài theo hướng đã chọn ứng dụng hữu ích trong công việc và cuộc sống.

TÀI LIỆU THAM KHẢO

I. TÀI LIỆU TIẾNG VIỆT

[1] Nguyễn Văn Chức, *"Ứng dụng lý thuyết tập thô trong khai phá dữ liệu",* tại bis.net.vn năm 2013.

[2] Hoàng Xuân Huấn (2012), "*Giáo trình Nhận dạng mẫu"*, Trường Đại học công nghệ – Đại Học Quốc Gia Hà Nội.

[3] Nguyễn Đức Thuần, "*Lý thuyết tập thô trong khai phá dữ liệu*", trong Tập san tin học Quản lý, tập 02, số 2, 2012, 25-32.

[4] Vũ thị Anh Trâm, "*Sử dụng phương pháp xây dựng đặc trưng dựa trên di truyền để toám tắt dữ liệu",* luận văn ths năm 2012, ĐH Công nghệ- ĐHQGHN.

II. TÀI LIỆU TIẾNG ANH

[5] Bashar Al-Shboul, and Sung-Hyon Myaeng,"*Initializing K-Means using Genetic Algorithms*", in World Academy of Science, Engineering and Technology 54 2009

[6] Jianyong Chen and Changsheng Zhang "*Efficient Clustering Method Based on Rough Set and Genetic Algorithm*" in College of Physics and Electronic Information Engineering, Wenzhou University, Wenzhou, 325035, China; Procedia Engineering 15 (2011) 1498 – 1503.

[7] Jiawei Han, Micheline Kamber. Data Mining: *Concepts and Techniques*[M]. US Kaufmann Publishers, Inc, 2001: p.223-262.

[8] Grabmeier J, Rudolph A. Techniques of cluster algorithms in data mining[J]. *Data Mining and Knowledge Discovery*, 2005,6(4):303-360.

[9] Guoyin Wang, Yiyu Yao, Hong Yu. "*A Survey on Rough Set Theory and Applications*[J]", Chinese Journal of Computers,2009. 32(7):1229-1246.

[10] Kevin E. Voges , and Nigel K. Ll. Pope, "*Rough Clustering Using an Evolutionary Algorithm".*

[11] Parvesh Kumar and Siri Krishan Wasan, *"Comparative Study of K-Means , Pam and Rough K-Means Algorithms Using Cancer Datasets"*, in 2009 International Symposium on

Computing, Communication, and Control (ISCCC 2009) Proc.of CSIT vol.1 (2011) © (2011) IACSIT Press, Singapore.

[12] Pawan Lingras, "*Interval Set Clustering of Web Users with Rough K-Means [J]*". Journal of Intelligent Information System,2004, 23: 15-16.

[13] Pawan Lingras and Georg Peter, "*Applying Rough Set Concepts to Clustering*".

[14] Pawlak Z. "*Rough set theory and its application to data analysis[J]".* Cybernetics and Systems, 1998, 9: 661-668.

[15] Ting Lin, Haixiang Guo, Kejun Zhu, Siwei Gao. "*An Improved Genetic K-Means Algorithm for Optimal Clustering[J*]".Mathematic in Practice and Theory, 2007, 37(8):104-111.

[16] Voges, K. E., N. K. Ll. Pope, and M. R. Brown, "*Cluster Analysis of Marketing Data Examining On-line Shopping Orientation: A Comparison of K-Means and Rough Clustering Approaches*", in Abbass, H. A., R. A. Sarker, and C. S. Newton (eds.), Heuristics and Optimization for Knowledge Discovery, Idea Group Publishing, Hershey, PA, 2002, pp. 207-224.