Phân cụm dữ liệu sử dụng giải thuật di truyền

Hoàng Hà Đức*, Lưu Thị Thanh Hà**

**ThS. Trường Cao Đẳng Yên Bái*

Received: 10/01/2024; Accepted: 18/01/2024; Published: 22/01/2024

Abstract: *The main purpose of algorithms is to find high-quality solutions, and using artificial intelligence techniques is especially necessary when solving problems with large search spaces.*
Genetic Algorithm is one of the techniques for finding optimal solutions that has met the requirements of many problems and applications. Currently, genetic algorithms are widely applied in complex fields.

Keywords**:** *Data clustering, genetic algorithms*

1. Đặt vấn đề

Tìm kiếm lời giải tối ưu cho các bài toán là vấn đề được các nhà khoa học máy tính đặc biệt rất quan tâm. Mục đích chính của các thuật toán là tìm kiếm thuật giải chất lượng cao và sử dụng kỹ thuật trí tuệ nhân tạo đặc biệt rất cần thiết khi giải quyết các bài toán có không gian tìm kiếm lớn.

Giải thuật di truyền (Genetic Algorithm GA) là một trong những kỹ thuật tìm kiếm lời giải tối ưu đã đáp ứng được yêu cầu của nhiều bài toán và ứng dụng. Hiện nay, thuật toán di truyền được ứng dụng rất rộng rãi trong các lĩnh vực phức tạp. Thuật toán di truyền chứng tỏ được hiệu quả của nó trong các vấn đề khó có thể giải quyết bằng các phương pháp thông thường hay các phương pháp cổ điển, nhất là trong các bài toán cần có sự lượng giá, đánh giá sự tối ưu của kết quả thu được.

2. Nội dung nghiên cứu

2.1.Thuật toán phân cụm dữ liệu dựa trên giải thuật di truyền

2.1.1 Giải thuật di truyền.

Thuật giải di truyền (GA) là kỹ thuật chung giúp giải quyết vấn đề bài toán bằng cách mô phỏng sự tiến hóa của con người hay của sinh vật nói chung (dựa trên thuyết tiến hóa muôn loài của Darwin) trong điều kiện qui định sẵn của môi trường. GA là một thuật giải, nghĩa là mục tiêu của GA không nhằm đưa ra lời giải chính xác tối ưu mà là đưa ra lời giải tương đối tối ưu.

Method

1. Khởi tạo một quần thể ban đầu với n cá thể.

2. Lặp m bước, mỗi bước phát sinh một quần thể mới theo quy trình sau.

2.1. Lai ghép:

- Chọn ngẫu nhiên một cặp hai cá thể cha mẹ B và M theo xác xuất Pl

- Sinh hai cá thể mới C1 và C2 từ B và M.

- Thay thế C1 và C2 cho B và M.

2.2. Đột biến:

- Chọn ngẫu nhiên một cá thể X theo xác xuất Pd

- Đột biến cá thể X.

2.3. Lặp nhận:

- Tính lại độ thích nghi của các cá thể.

- Chọn các cá thể có độ thích nghi tốt đưa vào quá trình mới.

3. Lấy nghiệm.

End.

Biểu diễn Gen bằng chuỗi nhị phân.

Quy tắc biểu diễn gen qua chuỗi nhị phân: Chọn chuỗi nhị phân ngắn nhất nhưng đủ thể hiện được tất cả kiểu gen.Để biểu diễn chuỗi nhị phân, ta thường dùng các cách sau: Mảng byte, mảng bit biểu diễn bằng mảng byte, mảng bit biểu diễn bằng mảng INTEGER.Mảng byte và mảng bit bây giờ ít sử dụng. Đối với máy tính ngày nay, người ta thường dùng mảng integer để tối ưu truy xuất.

Biểu diễn gen bằng chuỗi số thực.

Đối với những vấn đề bài toán có nhiều tham số, việc biểu diễn gen bằng chuỗi số nhị phân đôi lúc sẽ làm cho kiểu gen của cá thể trở nên quá phức tạp. Dẫn đến việc thi hành các thao tác trên gen trở nên kém hiệu quả. Khi đó, người ta sẽ chọn biểu diễn kiểu gen dưới dạng một chuỗi số thực.

Biểu diễn gen bằng cấu trúc cây.

Một loại cây thường được sử dụng trong thuật giải di truyền là dạng cây hai nhánh (ở đây chúng tôi dùng chữ hai nhánh để phân biệt với loại cây nhị phân – thường dùng trong sắp xếp và tìm kiếm).

Nguyên lý về xác định tính thích nghi.

"Tính tốt của một cá thể (lời giải) trong một quần thể chỉ là một cơ sở để xác định tính thích nghi của cá thể (lời giải) đó". Nguyên lý này ban đầu có vẻ hơi bất ngờ một khi chúng ta đã hiểu những ý tưởng chung của thuật giải di truyền. Thật đơn giản, người leo lên ngọn đồi cao nhất trong thế hệ hiện tại vẫn có khả năng bị"kẹt" trong các thế hệ sau cũng như một lời giải chưa tốt ở thế hệ hiện tại vẫn còn khả năng tiềm tàng dẫn đến lời giải tối ưu.

Độ thích nghi tiêu chuẩn.

Hàm mục tiêu là hàm dùng để đánh giá độ tốt của một lời giải hoặc cá thể. Hàm mục tiêu nhận vào một tham số là gen của một cá thể và trả ra một số thực. Tùy theo giá trị của số thực này mà ta biết độ tốt của cá thể đó (chẳng hạn với bài toán tìm cực đại thì giá trị trả ra càng lớn thì cá thể càng tốt, và ngược lại, với bài toán tìm cực tiểu thì giá trị trả ra càng nhỏ thì cá thể càng tốt)

Mã hóa (encoding).

Mã hóa bằng số nhị phân (Binary Encoding)

Mã hóa bằng số nhị phân là phương pháp chính. Bởi vì là phương pháp đầu tiên GA dung để mã hóa và nó đơn giản.

Mã hóa vị trí (Permutation Encoding)

Những vấn đề dựa trên thứ tự có thể dùng mã hóa vị trí, ví dụ như bài toán người du lịch hoặc thao tác thứ tự vấn đề.

Mã hóa theo giá trị (Value Encoding)

Mã hóa theo giá trị có thể dùng trong nhiều vấn đề, ở một vài giá trị phức tạp (ví dụ: giá trị thực). Dùng mã hóa nhị phân để giải quyết vấn đề này rất khó.

Cây mã hóa (Tree Encoding)

Cây mã hóa dùng trong chương trình tiến hóa hoặc biểu thức. cho lập trình tiến hóa. Trong cây mã hóa mỗi nhiễm sắc thể là một cây, ví dụ hàm và lệnh trong ngôn ngữ lập trình.

Các phương pháp chọn (Selection).

Chọn lọc cá thể thông qua kết quả, hay mục đích của vấn đề dựa trên mức độ thích nghi của cá thể. Vì vậy, đánh giá độ thích nghi của cá thể để tìm ra cá thể tốt nhất. Thông thường, đặt mỗi vấn đề nhỏ tương ứng với một giá trị điểm thích nghi, kết quả đánh giá gồm tổng các số điểm đó. Cá thể tốt nhất sẽ có điểm thấp nhất hoặc lớn nhất.

Chọn lọc Roulette (Roulette Wheel Selection).

Các cá thể được chọn theo độ thích nghi của chúng. Nhiễm sắc thể tốt hơn có cơ hội cao hơn để tham dự vào thế hệ tiếp theo.

Chọn lọc xếp hạng (Rank Selection).

Phương pháp này sẽ sắp hạng cá thể dựa trên độ thích nghi của chúng. Cá thể xấu nhất sẽ có giá trị 1, kế tiếp là 2… Và cá thể tốt nhất có độ thích nghi N (N là số các nhiễm sắc thể trong quần thể).

Chọn lọc cạnh tranh (Tournament Selection).

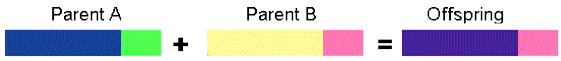
- Chọn lọc cạnh tranh 2 (2- Tournament Selection)
- Chọn lọc cạnh tranh 3 (3- Tournament Selection)

Các phương pháp lai tạo (crossover) và đột biến (mutation).

Lai ghép và đột biến là hai phép cơ bản được thực hiện trong giải thuật di truyền trên nhiều vấn đề. Kiểu và thực thi của phép thực hiện trên mã hóa và ngoài ra trên vấn đề. Có nhiều phương pháp lai ghép và đột biến. Ở đây chúng ta chỉ miêu tả một số thường dùng.

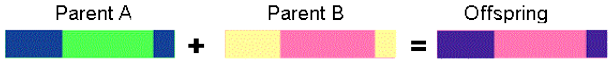
Lai ghép (Crossover)

Lai ghép ở một vị trí (Single point crossover) – Từ hai nhiễm sắc thể cha mẹ ban đầu ta cắt ở một vị trí sau đó ghép lại với nhau thành nhiễm sắc thể con.



11001011+11011111 = 11001111

Lai ghép ở hai vị trí (Two point crossover)– Từ hai nhiễm sắc thể cha mẹ ban đầu ta cắt ở hai vị trí sau đó ghép chúng với nhau thành nhiễm sắc thể con.

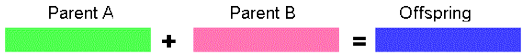


11001011 + 11011111 = 11011111

Lai ghép đồng dạng(Uniform crossover) – Những bit được copy ngẫu nhiên từ nhiễm sắc thể cha thứ nhất sang nhiễm sắc thể cha thứ hai và ngược lại.

11001011 + 11011101 = 11011111

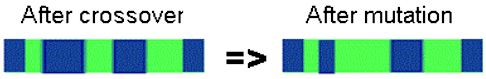
Lai ghép số học (Arithmetic crossover) – Một vài phép tính số học được thực hiện khi lai ghép để tạo ra nhiễm sắc thể con. (AND,OR,NOT…)



11001011 + 11011111 = 11001001 (AND)

Đột biến (Mutation)

Chèn bit (Bit inversion) – chọn một số bit sau đó chèn vào nhiễm sắc thể cha, tạo ra nhiễm sắc thể mới.



11001001 => 10001001

Thuật Toán K-Means

Thuật toán K-Means thực hiện qua các bước chính sau:

1. Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.

2. Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean)

3. Nhóm các đối tượng vào nhóm gần nhất

4. Xác định lại tâm mới cho các nhóm

5. Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng

Thuật toán Kmean sử dụng giải thuật di truyền

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

Output: Một tập hợp K cụm

Begin

Bước 1: Khởi tạo

Mỗi nhiễm sắc thể được tạo bằng cách chọn ngẫu nhiên k phần tử trong tập dữ liệu để làm k trọng tâm cụm.

Bước 2: For t = 1 to tMax

1. Đối với mỗi nhiễm sắc thể

a. Đưa phần tử trong D vào cụm với trọng tâm cụm gần nhất

b. Tính toán lại k trọng tâm cụm là trung bình k cụm vừa tạo và thay thế vào nhiễm sắc thể đó.

c. Tính toán độ thich snghi cho nhiễm sắc thể hiện tại.

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

Bước 3: In kết quả

Tách ra k cụm đối với nhiễm sắc thể trong quần thể của thế hệ tạo ra sau cùng có độ thích nghi lớn nhất.

Điều kiện dừng:

Lặp lại các bước 2 cho đến khi thế hệ t=Max

End.

So sánh giữa K-means và K-means sử dụng giải thuật di truyền:

Bảng dưới đây sẽ đưa ra so sánh về hai giải thuật trên:

K-means	*K-means sử dụng giải thuật di truyền*
- Đầu vào: k, bộ dữ liệu, k cụm trung tâm được lựa chọn ngẫu nhiên.	- Đầu vào: k, bộ dữ liệu, p, P nhiễm sắc thể được chọn ngẫu nhiên, Tmax.
- 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
- Cuối cùng nhóm có thể hội tụ về giá trị tối ưu cục bộ.	- Giải thuật di truyền là nền tảng trên phương pháp tiếp cận toàn cục với giá trị tiềm ẩn song song.
- 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: (Tmax*p*n*k*d) Trong đó: + n: số điểm dữ liệu + k: số cụm + d: kích thước dữ liệu + Tmax: số lần lặp + P: kích cỡ dân số

So sánh K-means và K-means có sử dụng giải thuật di truyền

2.2.2. Thực nghiệm phân cụm dữ liệu về sinh viên

a) Mô tả bài toán.

Hiện nay công tác kiểm tra đánh giá điểm rèn luyện của sinh viên trường rất được quan tâm và là một trong các nội dung quan trọng. Công tác đánh giá trình điểm rèn luyện được thể hiện thông qua kết quả khảo sát, đánh giá, thống kê tỷ lệ phân loại học sinh qua quá trình học tập, rèn luyện từ đó trường có thể tự đánh giá trình độ học vấn, rèn luyện đạo đức của sinh viên trong trường.

b) Xây dựng chương trình.

Các chức năng của chương trình

Bài báo đã sử dụng được viết trên ngôn ngữ lập trình C# xây dựng chương trình sử dụng giải thuật di truyền để phân cụm dữ liệu sinh viên trường:

- Đọc số liệu phân cụm
- Xây dựng cấu trúc dữ liệu
- Chọn số cụm đánh giá
- Chọn các môn đánh giá điểm rèn luyện
- Hiển thị kết quả.
- Phân tích kết quả để đưa ra các nhận xét, đánh giá

c) Giao diện chương trình

Từ việc khảo sát, thống kê tập hợp dữ liệu điểm rèn luyện của sinh viên đã xây dựng được chương trình tương đối hoàn chỉnh để giải quyết được bài toán khảo sát, đánh giá, thống kê đảm bảo những yêu cầu đã đề ra ban đầu.

d) Kết quả thực nghiệm.

Dữ liệu đầu vào dựa trên 200 học sinh, sinh viên của trường để phân cụm đánh giá về ý thức học tập.

3. Kết luận

Bài báo trình bày khái niệm cơ sở lý thuyết của khai phá dữ liệu và phân cụm dữ liệu; giới thiệu giải thuật chung cho giải thuật phân cụm sử dụng giải thuật di truyền; thực hiện cài đặt thử nghiệm giải thuật phân cụm Kmeans sử dụng giải thuật di truyền.

Trên cơ sở các kết quả đã đạt được, có thể tiếp tục nghiên cứu một số vấn đề như sau: Xây dựng tiếp các chương trình thử nghiệm các thuật toán phân cụm và các giải thuật phân cụm có sử dụng giải thuật di truyền; tìm thêm các ứng dụng giải thuật vào thực tiễn.

Tài liệu tham khảo

1.Shoa-Yei Yeong, Al-Salihy (2009). "Combination of neural network based clustering and genetic algorithm for multi-objective 802.11n planning".

3. Guojun Gan, Chaoqun Ma, Jianhong Wu (2007). "Data ClusteringTheory, Algorithms, and Applications". ASA-SIAM Series on Statistics and AppliedProbability, SIAM, Philadelphia, ASA, Alexandria, VA.