CẢI TIẾN THUẬT TOÁN CÂY QUYẾT ĐỊNH C4.5 CHO VẤN ĐỀ PHÂN NHÓM TRẺ TỰ KỶ

A MODIFIED DECISION TREE ALGORITHM C4.5 FOR AUTISM CHILDREN CLASSIFICATION PROBLEM

Nguyễn Văn Hiệu

Trường Đại học Bách khoa, Đại học Đà Nẵng; nvhieuqt@dut.udn.vn

Tóm tắt - Bài báo đề xuất hướng tiếp cận cải tiến các kỹ thuật phân nhóm để từ đó có thể vận dụng xây dựng hệ thống hỗ trợ trong dự đoán bệnh tự kỷ ở trẻ em. Trên cơ sở kiến thức cơ bản về rối loạn phổ tự kỷ ở trẻ em, nhóm tác giả sử dụng thuật toán di truyền để tối ưu kết quả của cây quyết định C4.5 và từ đó đưa ra quy trình chẩn đoán rối loạn phổ tự kỷ. Ngoài ra, bài báo đã biến đổi các triệu chứng bệnh thành các thuộc tính của dữ liệu vào và biến đổi các kết luận bệnh thành thuộc tính của dữ liệu ra, sau đó tiến hành cài đặt ứng dụng. Nghiên cứu này cũng góp phần phát triển phương pháp luận phục vụ trong việc chẩn đoán phổ tự kỷ ở trẻ em, giúp các bậc cha mẹ, thầy cô giáo, y bác sĩ có thể phát hiện bệnh sớm nhằm nâng cao hiệu quả trong điều trị bệnh.

Từ khóa - tự kỷ; thuật toán di truyền; cây quyết định; phân nhóm; chẩn đoán tự kỷ.

Abstract - This research is aimed at improving the techniques used to manipulate the group to build support systems in predicting autism in children. From basic knowledge of autism in children, the author uses a genetic algorithm to optimize the results of C4.5 decision tree and launches autism diagnosis process. In addition, the author changes the symptoms of the disease into attributes of the input data and the conclusion of disease into attributes of the output data and then proceed to install the application. This research also contributes to developing methodologies to serve in predicting autism in children, thereby helps parents, teachers and doctors detect the disease early in order to improve efficiency in the treatment.

Key words - autism; genetic algorithm; decision tree; classification; autism diagnosis.

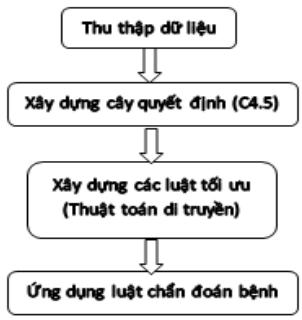
1. Giới thiệu

Hiện nay, ở Việt Nam chưa có số liệu chính thức về tỷ lệ trẻ em mắc bệnh tự kỷ. Tuy nhiên, theo số liệu thống kê mới nhất từ Khoa Phục hồi Chức năng thuộc Bệnh viện Nhi Trung ương thì số trẻ em mắc bệnh tự kỷ không ngừng tăng lên: Năm 2008 có 963 trẻ điều trị tự kỷ, năm 2010 có 1.752 trẻ, năm 2012 có 2.200 trẻ, năm 2014 có 2.640 trẻ điều trị tự kỷ. Và bệnh tự kỷ dường như đang là nỗi lo lắng lớn nhất của gia đình Việt. Tuy nhiên, các bậc làm cha, làm mẹ còn thiếu kiến thức về căn bệnh tự kỷ [5]. Hơn thế nữa, ở Việt Nam có rất ít hệ thống chẩn đoán trẻ tự kỷ, gần đây có phần mềm A365 để sàng lọc chậm phát triển và can thiệp sớm cho trẻ tự kỷ tại nhà do nhóm của Tiến sĩ Vũ Song Hà phát triển. Phần mềm này sử dụng 2 bộ công cụ để sàng lọc cho trẻ đó là ASQ và MCHAT. Công cụ này đánh giá sàng lọc dựa vào tổng điểm so với bảng điểm sàng lọc đã được xác lập vì vậy sẽ phát hiện tối đa dương tính các trường hợp có nguy cơ bị tự kỷ nên hiệu quả dự đoán chưa cao. Ngoài ra, hai công cụ này gồm rất nhiều câu hỏi trong 5 lĩnh vực, yêu cầu người dùng phải trả lời hết, vì vậy phải thử nghiệm với trẻ tất cả các nội dung có trong bộ câu hỏi để có cơ sở trả lời, điều này mất nhiều thời gian và có thể bỏ sót những trường hợp mà trẻ có biểu hiện đặc biệt khác.

Phân lớp có vai trò và ý nghĩa hết sức quan trọng trong việc dự đoán những nhãn phân lớp cho các bộ dữ liệu mới. Thuật toán phân lớp giữ vai trò quyết định tới sự thành công của mô hình phân lớp. Một trong những thuật toán phân lớp bằng cây quyết định khá dễ hiểu, thích hợp được với các dữ liệu liên tục, dữ liệu bị thiếu, bị nhiễu, đó là thuật toán C4.5. Tuy nhiên, bản chất của thuật toán C4.5 là thuật toán tham lam, nên mặc dầu thuật toán chạy nhanh nhưng lại không tạo được các luật tối ưu. Chính vì vậy, bài báo đề xuất mô hình mới trên cơ sở thuật toán di truyền để tối ưu kết quả của thuật toán C4.5. Sau đó sử dụng kết quả nhận được để chẩn đoán bệnh tự kỷ.

2. Đề xuất ứng dụng vào chẩn đoán bệnh tự kỷ

Hệ thống chẩn đoán bệnh tự kỷ bằng kỹ thuật phân nhóm được thực hiện theo quy trình như sau:



***Hình 1**. Quy trình chẩn đoán bệnh tự kỷ*

2.1. Thu thập dữ liệu

Để có nguồn dữ liệu đáng tin cậy phục vụ việc tập huấn luyện, nhóm tác giả đã sử dụng cẩm nang DSM-V để trích rút ra các triệu chứng, đồng thời tham khảo thêm các tài liệu y học chuyên ngành, các bài báo chuyên môn cũng như ý kiến của một số bác sỹ, các nhà tâm lý có kinh nghiệm trong việc điều trị tự kỷ.

2.2. Xây dựng cây quyết định

Cây quyết định được xây dựng theo thuật toán C4.5, sau đó sử dụng thuật toán di truyền để tối ưu kết quả của cây quyết định. Dữ liệu vào/ra của thuật toán được mô tả ở Bảng 1.

2.3. Xây dựng luật

Từ cây quyết định của thuật toán C4.5 có thể xây dựng được các luật dưới dạng IF-THEN. Mỗi luật là một đường đi từ nút gốc đến nút lá. Luật có dạng:

IF R_1 and R_2 and … and R_n **THEN** Kết quả = G

Trong đó:

- $R_1,\ldots, R_n$: là các biểu thức logic mà vế trái là thuộc tính và vế phải là giá trị của thuộc tính.
- G là kết quả cần dự đoán: P_1, P_2, P_3, P_4.

Sau khi có các luật cụ thể, ta tiến hành tối ưu tập luật theo thuật toán di truyền. Dựa độ chính xác, độ hỗ trợ, tính đơn giản, và độ đo Gain được trình bày trong phần III, tiến hành xây dựng hàm thích nghi. Giá trị thích nghi càng lớn, luật càng tốt. Sử dụng phép toán lai ghép và đột biến của thuật toán di truyền để điều chỉnh hàm thích nghi. Nên giá trị hàm thích nghi sẽ tiến đến giá trị lớn nhất và luật sẽ được tốt nhất.

2.4. Ứng dụng luật trong chẩn đoán bệnh

Hệ thống sử dụng các luật để ứng dụng vào chẩn đoán bệnh tự kỷ. Người dùng cung cấp cho hệ thống các thông tin liên quan đến bệnh nhân: thông tin về quan hệ xã hội, khả năng bắt chước, đáp ứng tình cảm, các động tác cơ thể, sử dụng đồ vật,… Hệ thống sẽ suy diễn dựa vào tập luật được tạo ra và cho ra kết quả về nguy cơ mắc bệnh ứng với các thông tin đã được cung cấp. Các kết quả chẩn đoán bệnh có thể là một trong các trường hợp: P1, P2, P3, P4 tương ứng với kết quả trẻ không bị tự kỷ, trẻ bị tự kỷ ở mức độ nhẹ, trẻ bị tự kỷ ở mức độ trung bình và trẻ bị tự kỷ.

Bảng 1. *Thuộc tính thể hiện triệu chứng của bệnh tự kỷ*

Thuộc tính	Giải thích	Kiểu	Giá trị mã hóa
A	Quan hệ xã hội	Rời rạc	A1, A2, A3, A4
B	Khả năng bắt chước	Rời rạc	B1, B2, B3, B4
C	Khả năng đáp ứng tình cảm	Rời rạc	C1, C2, C3, C4
D	Các động tác cơ thể	Rời rạc	D1, D2, D3, D4
E	Khả năng sử dụng đồ vật	Rời rạc	E1, E2, E3, E4
F	Khả năng thích nghi với sự thay đổi	Rời rạc	F1, F2, F3, F4
G	Khả năng phản ứng thị giác	Rời rạc	G1, G2, G3, G4
H	Phả năng ứng thính giác	Rời rạc	H1, H2, H3, H4
I	Khả năng phản ứng qua vị, khứu, xúc giác	Rời rạc	I1, I2, I3, I4
J	Sợ hãi hoặc hồi hộp	Rời rạc	J1, J2, J3, J4
K	Giao tiếp bằng lời	Rời rạc	K1, K2, K3, K4
L	Giao tiếp không lời	Rời rạc	L1, L2, L3, L4
M	Mức độ hoạt động	Rời rạc	M1, M2, M3, M4
N	Đáp ứng trí tuệ	Rời rạc	N1, N2, N3, N4
P	Mức độ tự kỷ	Rời rạc	P1, P2, P3, P4

3. Mô hình cải tiến thuật toán C4.5

Mục tiêu của mô hình là sử dụng thuật toán di truyền để tối ưu kết quả của cây quyết định C4.5.

3.1. Mã hóa cho các luật

Thuật toán di truyền sử dụng mã nhị phân để biểu thị một cá thể. Phương pháp này sử dụng một chuỗi được tạo bởi kí tự {0, 1} để biểu thị một cá thể. Mỗi mã hóa đáp ứng với một thuộc tính điều kiện và giá trị thuộc tính sẽ xác định độ dài mã hóa. Minh họa, một thuộc tính có K loại giá trị, do đó mã hóa cá thể sẽ phân phối K bit cho nó và mỗi bit tương ứng với các giá trị có thể. Khi giá trị là 0, có nghĩa là các cá thể sẽ không lấy giá trị thuộc tính. Khi giá trị là 1, các cá thể sẽ lấy giá trị thuộc tính. Sự biến đổi của phương pháp này là đơn giản và mỗi nhiễm sắc thể có chiều dài cố định. Tuy vậy, cây quyết định có một đặc tính là nút không chỉ là thuộc tính rời rạc mà còn là thuộc tính số. Mã hóa nhị phân không phải là cách hữu dụng.

Trong phương pháp, mỗi nhiễm sắc thể thể hiện một luật phân lớp. Một số nhiễm sắc thể sẽ trở thành giải pháp của vấn đề. Những luật cuối cùng sẽ được sắp xếp theo chất lượng của luật. Khi các luật được dùng để nhận dạng mẫu mới, luật tốt nhất sẽ được xem xét đầu tiên. Nếu luật tốt nhất không thể nhận ra mẫu thì chúng ta có thể lựa chọn luật tiếp theo. Nếu các luật trong tập luật không thể nhận ra các mẫu khác thì các mẫu sẽ được phân lớp thành lớp mặc định. Nhiễm sắc thể sẽ cạnh tranh với nhau theo độ ưu tiên trong quần thể.

Giả sử rằng dữ liệu gồm n thuộc tính nên mỗi nhiễm sắc thể sẽ được chia thành n gen, gen thứ i tương ứng với thuộc tính thứ i. Mỗi cá thể đại diện cho một quy tắc phân lớp và mỗi gen đại diện phía bên trái hay bên phải của luật phân lớp. Các gen bên trái của mỗi luật phân lớp gọi là những gen đặc trưng, còn phía bên phải gọi là gen lớp. Trong quá trình tiến hóa gen, các gen đặc trưng sẽ tham gia vào sự phát triển, nhưng gen lớp thì không. Mỗi nhiễm sắc thể có chiều dài cố định và có một số gen. Bên trong của mỗi gen bao gồm bốn phần: {Trọng số, phép toán, giá trị, độ đo Gain}.

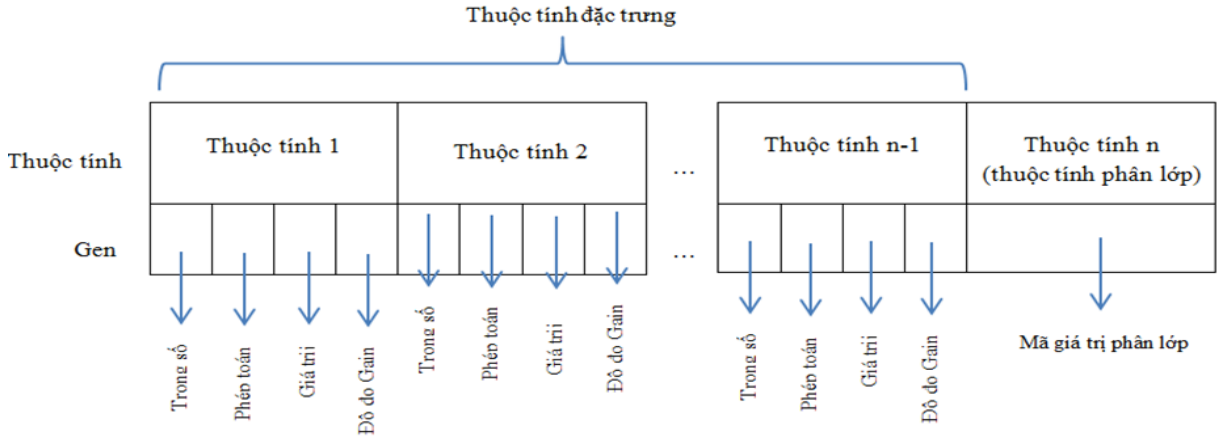
Trọng số: Trọng số là một biến logic; nó đại diện cho tình trạng gen tương ứng với các thuộc tính xuất hiện; nếu trọng số là 1, các thuộc tính tương ứng với các gen sẽ xuất hiện trong luật. Ngược lại, trọng số là 0, có nghĩa là các thuộc tính tương ứng với các gen sẽ không xuất hiện trong các luật.

Phép toán: Nó biểu thị những phép toán kết hợp gen. Đối với các thuộc tính rời rạc, các giá trị là "=" hoặc "#"; đối với các thuộc tính liên tục, giá trị là "≥" hoặc "<".

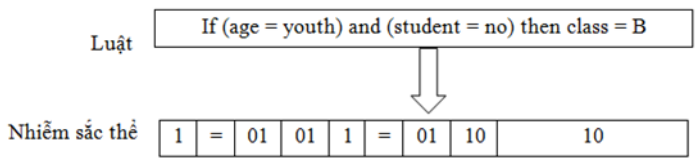
Giá trị: Giá trị biểu thị giá trị của các thuộc tính. Đối với các thuộc tính rời rạc, giá trị tương đương với các giá trị thực tế trong miền giá trị. Giá trị của các thuộc tính liên tục bằng với giá trị thực.

Độ đo Gain: Biểu thị tốc độ tăng của thông tin thuộc tính.

Trước khi bắt đầu thuật toán di truyền, tính toán và lưu lại tất cả các Information Gain của thuộc tính của cá thể. Cấu trúc của nhiễm sắc thể thể hiện trong Hình 2. Trong phương pháp, mặc dù độ dài của nhiễm sắc thể là cố định, nhưng độ dài của luật có thể thay đổi, và nó sẽ tạo ra luật đơn giản hơn.



Hình 2. *Cấu trúc nhiễm sắc thể*



Hình 3. *Ví dụ mã hóa luật*

3.2. Hàm thích nghi cho luật

Trong thuật toán di truyền, hàm thích nghi là một độ đo đánh giá tính tốt hay xấu của một cá thể. Trong phương pháp này, ta có thể chia mẫu thành 4 lớp:

(1) T_T: Thể hiện số luật dự đoán tập mẫu là đúng và thực tế là đúng;

(2) T_F: Thể hiện số luật dự đoán tập mẫu là đúng và thực tế là sai;

(3) F_T: Thể hiện số luật dự đoán tập mẫu là sai và thực tế là đúng;

(4) F_F: Thể hiện số luật dự đoán tập mẫu là sai và thực tế là sai;

Công thức tính độ chính xác (ký hiệu ac) được tính:

$$\text{ac} = \frac{\text{T\_T} + \text{F\_F}}{\text{T} + \text{F}} \qquad (1)$$

Trong đó, T là số lượng mẫu đúng và F là số lượng mẫu sai. Độ chính xác có thể là mức độ chính xác của các luật hoạt động trên dữ liệu đào tạo. Giá trị càng cao, mẫu phân lớp càng chính xác.

Công thức tính độ hỗ trợ (kí hiệu su) được tính:

$$\text{su} = \frac{\text{T\_T} + \text{F\_T}}{\text{T} + \text{F}} \qquad (2)$$

Giá trị càng lớn, tỉ lệ luật trong không gian dữ liệu càng lớn, nghĩa là luật có ý nghĩa tốt hơn.

Công thức tính độ đơn giản (kí hiệu si) được tính:

$$\text{si} = \frac{\text{N} - \text{n}}{\text{N}} \qquad (3)$$

Khi đó N là số thuộc tính trong dữ liệu khởi tạo và n là số thuộc tính trong luật. Độ đơn giản của cá thể càng cao thì luật càng đơn giản và dễ hiểu.

Chúng ta có thể lấy thông tin độ đo Gain để xây dựng hàm thích nghi theo công thức sau:

$$F = w_1\, si + w_2\, su + w_3\, ac + w_4\, Gain \qquad (4)$$

Với $w_i \in [0,1], i = 1, \dots, 4,\ \sum_1^4 w_i = 1$

3.3. Phép toán lai ghép và đột biến cho luật

Trong phương pháp này, ta nên lựa chọn một tập mẫu R trong tập dữ liệu đào tạo mà có thuộc tính phân lớp là C_i ngẫu nhiên và sau đó mã hóa R thành xâu mã hóa cá thể dựa trên quy tắc mã hóa. Với cách này, những cá thể mới hiệu quả hơn cá thể trước. Nó sẽ giảm không gian tìm kiếm của thuật toán tham lam và nâng cao tốc độ của thuật toán.

Lai ghép hai điểm được sử dụng trong nhiễm sắc thể; đầu tiên tạo ra các số thực ngẫu nhiên S_c trong đoạn [0,1]. Nếu S_c nhỏ hơn xác suất lai ghép P_c thì lựa chọn ngẫu nhiên cá thể a_i và a_j để lai ghép.

Tạo ra một số thực ngẫu nhiên S_m trong đoạn [0,1]. Nếu S_m nhỏ hơn xác suất đột biến của P_m, ta sẽ đột biến trên cá thể. Đối với gen trong phương pháp này có 4 phần, vì vậy ta phải xem xét cấu trúc gen đầy đủ. Vì vậy nó sẽ bao gồm 3 phép đột biến (độ đo Gain của gen sẽ không thay đổi trong các phép toán).

Trọng số đột biến: Nếu trọng số của gen ban đầu là 1,

sau đó biến nó thành 0; nếu trọng số của gen ban đầu là 0, sau đó đột biến nó là 1. Nếu trọng số đột biến từ 1 đến 0, các thuộc tính của gen tương ứng sẽ không xuất hiện trong luật. Ví dụ, như thể hiện trong Hình 4, thông qua đột biến trọng số, có thể nhận được luật sau:

IF (student = no) then Class = B

Đột biến phép toán: Đối với các thuộc tính rời rạc, nếu phép toán của gen ban đầu là "=", sau đó biến nó thành "#"; nếu phép toán của gen ban đầu là "#", sau đó biến nó thành "=". Đối với thuộc tính liên tục, nếu phép toán của gen ban đầu là "≥", sau đó đột biến nó thành "<"; nếu các phép toán ban đầu là "<", sau đó biến đổi nó thành "≥". Ví dụ, như thể hiện trong Hình 5, thông qua các đột biến phép toán, chúng ta có thể nhận được luật sau:

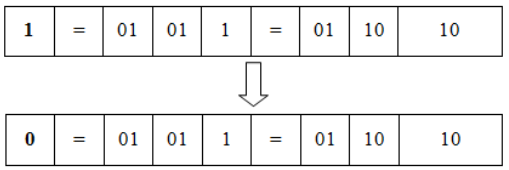
IF (age ≠ youth) and (student = no), then Class = B

Điều đó có nghĩa là:

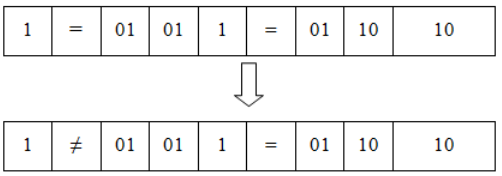
IF (age = middle_age or age = senior) and (student = no), then Class = B

Đột biến giá trị: Đối với các thuộc tính rời rạc, chọn một giá trị trong các thuộc tính để thay thế các giá trị trong bản gốc, ngẫu nhiên; đối với các thuộc tính liên tục, tạo ngẫu nhiên một số thập phân và sau đó cộng hoặc trừ các số thập phân vào giá trị ban đầu. Ví dụ, như thể hiện trong Hình 6, thông qua đột biến giá trị, chúng ta có thể nhận được các luật sau:

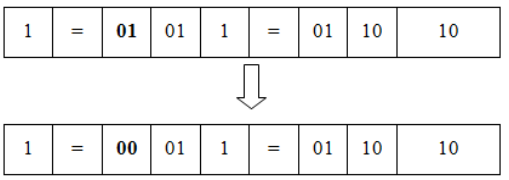
IF (age = middle_age) and (student = no), then Class = B



Hình 4. *Ví dụ cho đột biến trọng số*



Hình 5. *Ví dụ cho đột biến phép toán*



Hình 6. *Ví dụ cho đột biến giá trị*

Mỗi đột biến có thể là phép toán đột biến hoặc bất kì sự phối hợp của phép toán đột biến nào.

3.4. Thuật toán di truyền

Dữ liệu vào: Tập dữ liệu khởi tạo R, các tham số cho thuật toán di truyền.

Dữ liệu ra: Luật phân lớp tối ưu.

Mô tả thuật toán:

Bước 1: Khởi tạo quần thể một tập mẫu R với S bản ghi là lựa chọn ngẫu nhiên từ tập dữ liệu đào tạo có thuộc tính phân lớp với giá trị là Ci. Biến thích nghi trung bình (avg) của quần thể khởi tạo được gán bằng 0.

Bước 2: Tiền xử lý hoạt động được tạo ra dựa trên mẫu R bao gồm làm sạch dữ liệu, rời rạc hóa thuộc tính liên tục, tính toán độ đo Gain của mỗi thuộc tính đặc trưng và mã hóa các bản ghi dữ liệu. Cuối cùng chúng ta có quần thể ban đầu được mã hóa P(r).

Bước 3: Tính toán độ thích nghi của mỗi cá thể trong quần thể và sau đó độ thích nghi trung bình được tìm ra.

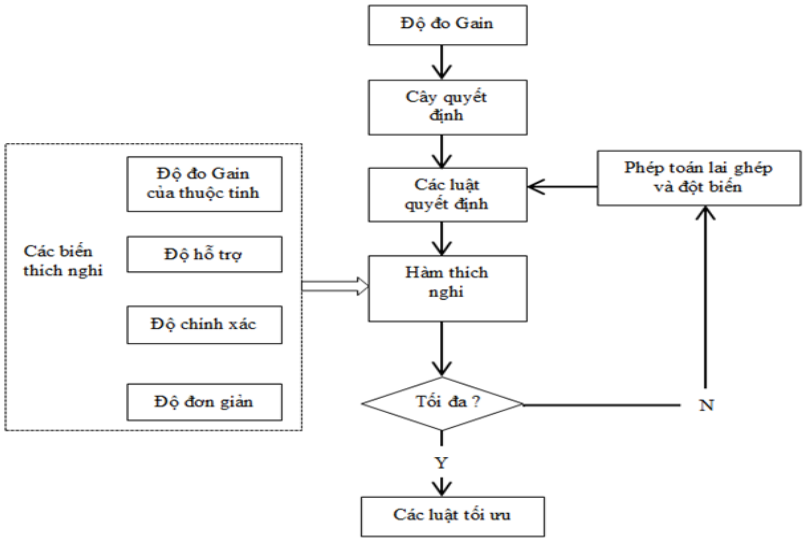
Bước 4: Nếu số thế hệ vượt quá 1 ngưỡng cho trước hoặc $\text{avg}_i - \text{avg}_{i-1} > \varepsilon$ thì lặp lại bước 5, 6 và bước 7. Nếu không thì chuyển qua bước 8.

Bước 5: Tính toán độ thích nghi trung bình của thế hệ được lựa chọn này, lai ghép và đột biến được tiến hành trên quần thể này. Vì vậy quần thể con được tạo ra.

Bước 6: Thay thế các cá thể có độ thích nghi thấp trong tập bố mẹ bởi cá thể có độ thích nghi cao trong quần thể con. Vì thế, một thế hệ mới được hình thành.

Bước 7: Tính toán độ thích nghi của mỗi cá thể trong thế hệ mới và tính độ thích nghi trung bình.

Bước 8: Những cá thể có độ thích nghi thấp hơn ngưỡng thích nghi thấp nhất được loại ra. Quần thể tối ưu là tập luật tối ưu.



Hình 7. *Quy trình cải tiến*

Đánh giá thuật toán:

Thuật toán cải tiến dựa trên thuật toán di truyền đã sử dụng khả năng tối ưu của thuật toán di truyền. Cụ thể:

- Thuật toán đề xuất này đã được cải thiện so với thuật toán cây quyết định bình thường về độ chính xác. Như kết quả ở Bảng 3 cho thấy, thuật toán cải tiến dựa trên thuật toán di truyền có độ chính xác cao hơn thuật toán C4.5.

- Các luật tối ưu dễ hiểu hơn so với các thuật toán khác.

4. Kết quả thực nghiệm

Hiện nay, tiếp cận với internet hằng ngày trở thành một thói quen của rất nhiều người. Vì vậy, để các bậc phụ huynh có thể theo dõi và chẩn đoán sớm bệnh tự kỷ ở con em mình một cách nhanh chóng và thuận tiện nhất, tác giả đã phát triển ứng dụng theo hướng trở thành một website.

Bảng 2. *Giá trị tham số thử nghiệm*

Tham số	**Giá trị**
Xác suất lai ghép	0,85
Xác suất đột biến	0,09
Số lần lặp tối đa	100
Các hệ số	$w_1 = 0{,}1, w_2 = 0{,}1;$ $w_3 = 0{,}1, w_4 = 0{,}7$

Để chứng minh hiệu quả của phương pháp đề xuất, tác giả đã tiến hành cài đặt thuật toán C4.5 và sử dụng thuật toán di truyền để tối ưu kết quả. Thuật toán di truyền sử dụng các tham số khởi tạo Bảng 2. Các luật đã tối ưu sẽ được áp dụng vào phân nhóm các mức độ của trẻ tự kỷ. Với tập dữ liệu mẫu có kích thước trên 500 bản ghi gồm 14 thuộc tính đặc trưng và 1 thuộc tính phân lớp, thuật toán C4.5 sinh ra 37 luật, thuật toán cải tiến dựa trên thuật toán di truyền sinh ra 30 luật.

Sau khi tiến hành thực hiện đánh giá độ chính xác của các mô hình phân lớp bằng phương pháp k-fold cross validation, với k bằng 10 ta có kết quả so sánh thể hiện ở Bảng 3.

Thông qua việc so sánh, chúng ta có thể kết luận rằng, thuật toán đề xuất dựa trên thuật toán di truyền đã được cải thiện so với thuật toán cây quyết định C4.5 về độ chính xác.

Bảng 3. *Độ chính xác thuật toán C4.5 và thuật toán di truyền*

Thuật toán	Độ chính xác
C4.5	96%
Di truyền	98%

Về cơ bản, chương trình đã đạt được mục tiêu đề ra là xây dựng thành công công cụ hỗ trợ chẩn đoán bệnh tự kỷ ở trẻ em, là giải pháp có khả năng ứng dụng cao trong thực tế. Tuy vậy, bộ dữ liệu mẫu chủ yếu được tác giả sưu tầm vào tổng hợp qua internet. Để tăng độ chính xác của hệ thống chẩn đoán, cần bổ sung nhiều hơn nữa các bộ dữ liệu mẫu được lấy từ các bác sỹ, chuyên gia tâm lý trong lĩnh vực nghiên cứu về tự kỷ ở trẻ em.

5. Kết luận

Bài báo này đề xuất một cây quyết định mới dựa trên thuật toán di truyền; nó sử dụng khả năng tối ưu của thuật toán di truyền. Trước tiên, xây dựng các quy trình ứng dụng vào chẩn đoán bệnh tự kỷ, và sau đó làm một thử nghiệm với thuật toán C4.5 và thuật toán di truyền; thông qua việc so sánh, chúng ta có thể kết luận rằng, thuật toán di truyền đã được cải thiện so với thuật toán cây quyết định bình thường về độ chính xác. Cuối cùng, áp dụng thuật toán này để phân loại trẻ em vào các nhóm: trẻ không bị tự kỷ, trẻ bị tự kỷ ở mức độ nhẹ, trẻ bị tự kỷ ở mức độ trung bình, trẻ bị tự kỷ. 30 luật với độ chính xác cao hơn đã được tạo ra với thuật toán di truyền. Tuy nhiên, cần bổ sung thêm dữ liệu tập huấn để mô hình phân nhóm có độ tin cậy cao hơn và hoạt động hiệu quả hơn, liên kết với các bác sĩ, nhà tâm lý học về bệnh tự kỷ để tăng thêm tính chuẩn xác trong quá trình chẩn đoán bệnh, đồng thời tìm hiểu nhu cầu thực tế, để từ đó cải tiến chương trình, cài đặt lại bài toán theo các thuật toán đã nghiên cứu để làm việc tốt hơn với các cơ sở dữ liệu lớn.

TÀI LIỆU THAM KHẢO

[1] Nguyễn Văn Siêm, *Tâm bệnh học trẻ em và thanh thiếu niên,* Nhà xuất bản Đại học Quốc gia Hà Nội, năm 2007.

[2] Trung tâm Nghiên cứu Giáo dục và Chăm sóc Trẻ em, *Những điều cần biết về hội chứng tự kỷ*, Nhà xuất bản Đại học Sư Phạm, năm 2011.

[3] Võ Nguyễn Tinh Vân, *Để hiểu chứng tự kỷ,* Nhóm tương trợ phụ huynh Việt Nam có con khuyết tật và chậm phát triển tại New South Wales, Nxb Bamboo, Australia, năm 2012.

[4] Nguyễn Thị Thùy Linh, *Nghiên cứu các thuật toán phân lớp dữ liệu dựa trên cây quyết định,* Khóa luận Đại học, Đại học Công nghệ, Đại học Quốc gia Hà Nội, năm 2005.

[5] Đỗ Thị Thu Hà, *Nghiên cứu và ứng dụng kỹ thuật cây quyết định xây dựng hệ thống dự đoán bệnh tự kỷ ở trẻ em*, Luận văn thạc sỹ, Đại học Đà Nẵng, năm 2015.

[6] Nguyễn Văn Hiệu, Đỗ Thị Thu Hà, "Hệ thống chẩn đoán bệnh tự kỷ sử dụng cây quyết định", *Tạp chí Khoa học Công nghệ - Đại học Đà Nẵng*, Số: 11(96), trang: 96-101, năm 2015.

[7] Lior Rokach; Oded Maimon, "Data mining with decision trees: theory and applications", *World Scientific Pub Co Inc. ISBN 978-9812771711*, năm 2008.

[8] Ihsan A. Kareem*, Mehdi G. Duaimi, "Modified Decision Tree Classification Algorithm for Large Data Sets", Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq, năm 2014.

(BBT nhận bài: 10/05/2017, hoàn tất thủ tục phản biện: 15/05/2017)