

HỆ THỐNG CHẨN ĐOÁN BỆNH TỰ KỶ SỬ DỤNG CÂY QUYẾT ĐỊNH

A SYSTEM FOR DIANOSING AUTISM BASED ON THE DECISION TREE

Nguyễn Văn Hiệu¹, Đỗ Thị Thu Hà²

¹*Trường Đại học Bách khoa, Đại học Đà Nẵng; nvhieutq@dut.udn.vn*

²*Trường Cao đẳng Sư phạm Gia Lai*

Tóm tắt - Những năm gần đây, số lượng trẻ em mắc chứng tự kỷ ở Việt Nam không ngừng tăng lên và dần trở thành nỗi lo lắng chung. Bệnh có thể xảy ra ở bất kì trẻ nào, ảnh hưởng đến sự phát triển não bộ ở trẻ và gây ra hậu quả nghiêm trọng, nếu không phát hiện và điều trị kịp thời. Thế nhưng, hiểu biết của người dân Việt Nam về bệnh tự kỷ còn rất hạn chế. Hơn nữa, khoảng 65% dân số Việt Nam là ở nông thôn, nơi mà cơ sở vật chất y tế còn hạn chế. Điều này khiến cho việc kiểm tra sức khỏe định kỳ cho trẻ và phát hiện bệnh càng khó khăn. Để giải quyết những vấn đề trên, cần xây dựng một công cụ để sử dụng và giúp phát hiện sớm nguy cơ mắc bệnh tự kỷ ở trẻ. Trong bài báo, chúng tôi sẽ đi sâu vào việc nghiên cứu cây quyết định trong việc khai phá dữ liệu về bệnh tự kỷ và ứng dụng cây quyết định trong việc xây dựng nên hệ thống cung cấp chức năng chẩn đoán nguy cơ mắc bệnh tự kỷ ở trẻ.

Từ khóa - chẩn đoán; bệnh tự kỷ; cây quyết định; ID3; mô hình chẩn đoán.

Abstract - In recent years, the number of autistic children in Viet Nam has been increasingly on the rise, which has become a public concern. Autism can appear in any child, affecting the growth of the brain and leave serious complications if it is not quickly detected and promptly treated. However, what the Vietnamese people know about autism is still very limited. In addition, about 65% of the Vietnamese population live in rural areas, where medical facilities are few, which makes it even more difficult to conduct periodic health examination for children and detect autism. To solve the above problems, it is essential to create a user-friendly tool that enables early detection of children's autism. In this paper, we are to conduct an in-depth investigation into the study of the decision tree in examining autism data and apply this decision tree in building up a system which provides functionality for the diagnosis of autism in children.

Key words - diagnose; autism; decision tree; iterative dichotomiser 3; diagnostic model.

1. Đặt vấn đề

Tự kỷ là một tình trạng khiếm khuyết phức tạp về các khả năng phát triển ở não bộ, tình trạng này có thể xảy ra ở bất kỳ một đứa trẻ nào, không lệ thuộc vào dân tộc, xã hội hay trình độ của cha mẹ, vì vậy sẽ rất khó phát hiện trẻ bị tự kỷ, nếu chỉ quan sát một số biểu hiện bên ngoài [2], [3]. Trong thực tế, nhiều người hay nhầm lẫn biểu hiện của trẻ mắc bệnh tự kỷ với những biểu hiện bình thường ở trẻ mới lớn, điều này gây ảnh hưởng lớn đến việc chữa trị cho trẻ. Vì vậy, cần có một công cụ áp dụng hỗ trợ người dùng trong việc tìm hiểu về bệnh và dễ dàng tự kiểm tra nguy cơ mắc bệnh tự kỷ của trẻ đơn giản và thường xuyên. Trong y học, để chẩn đoán bệnh tự kỷ cần trải qua 2 giai đoạn. Thứ nhất, ở giai đoạn khám đoán lâm sàng, bác sĩ sẽ dựa vào những thông tin thu thập được bằng cách hỏi thăm người nhà, quan sát trẻ như ánh mắt, cách giao tiếp, hành động... để đưa ra đánh giá sơ bộ về tình trạng của trẻ. Ở giai đoạn khám cận lâm sàng, trẻ sẽ được thực hiện các khám xét khác như điện não đồ, chụp CT, X-quang... để đưa ra kết quả chính xác nhất.

Với những hỗ trợ từ y học và sự phát triển của công nghệ thông tin, chúng tôi xây dựng nên hệ thống cho phép người dùng xem thông tin và tự chẩn đoán bệnh tự kỷ ở trẻ trên môi trường web bằng cách nhập thông tin về các biểu hiện của trẻ, để từ đó hệ thống đưa ra kết quả về nguy cơ mắc bệnh tự kỷ. Với những chức năng trên, hệ thống đóng vai trò như một công cụ hữu ích thay thế một phần cho giai đoạn khám lâm sàng cho bệnh tự kỷ, giúp người dùng dễ dàng kiểm tra, chẩn đoán bệnh tự kỷ ở trẻ mọi lúc mọi nơi, mà không cần phải thường xuyên đến bệnh viện hoặc trung tâm y tế.

Bài báo tập trung vào việc ứng dụng kỹ thuật cây quyết định trong khai phá dữ liệu, để xây dựng hệ thống chẩn đoán bệnh tự kỷ ở trẻ.

2. Lý thuyết về cây quyết định

Trong lĩnh vực khai phá dữ liệu, cây quyết định (Decision Tree – DT) là một mô hình dự đoán thuộc lớp các bài toán phân lớp, dùng để xác định lớp của các đối tượng cần dự đoán [4]. Bản chất cây quyết định dựa vào dãy các luật IF ... THEN để dự đoán lớp của đối tượng. Mỗi nút trong (*internal node*) của DT tương ứng với một biến, đường nối giữa một nút trong với nút con của nó thể hiện một giá trị cụ thể biến đó. Mỗi nút lá (*leaf*) đại diện cho giá trị dự đoán. Cây quyết định học để dự đoán giá trị của các biến phân loại bằng cách dựa vào tập dữ liệu huấn luyện (*training data*) để chọn ra nút gốc (*root node*) để phân tách cây bằng cách tính độ lợi thông tin (*Information Gain - IG*), quá trình phân tách cây được thực hiện một cách đệ quy cho đến khi không thể tiếp tục thực hiện việc phân tách cây được nữa [1], [4].

Cây quyết định được chia làm 2 loại:

Cây hồi quy dùng để dự đoán giá trị của biến phân loại có kiểu dữ liệu giá trị như dự đoán doanh thu, lợi nhuận, giá thành sản phẩm... Thuật toán phổ biến dùng để xây dựng cây hồi quy là CART.

Cây phân lớp dùng để dự đoán giá trị của biến phân loại có kiểu dữ liệu phi giá trị như dự đoán khả năng mua hàng, khả năng bị bệnh, kết quả học tập của sinh viên (xuất sắc, giỏi, khá, trung bình, yếu)... Thuật toán phổ biến dùng để xây dựng cây phân lớp là ID3, J48, C4.5, C5.0.

Hệ thống chẩn đoán bệnh tự kỷ sử dụng thuật toán ID3 để xây dựng cây quyết định và trong thuật toán này có sử dụng độ đo của nhà toán học Claude Shannon (*Information Gain - IG*) để xác định điểm chia.

3. Xây dựng hệ thống chẩn đoán bệnh tự kỷ

3.1. Mô tả chức năng hệ thống

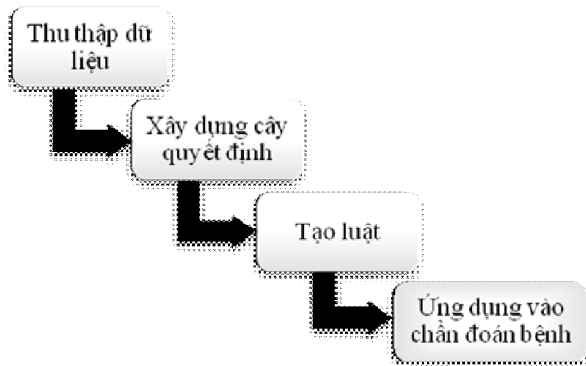
Hệ thống được xây dựng theo cấu trúc hệ thống website quản lý và website người dùng.

Với hệ thống quản lý này, hệ thống cho phép người quản lý có thể quản lý bộ dữ liệu huấn luyện bằng cách thêm, xóa, chỉnh sửa dữ liệu. Dựa vào bộ dữ liệu huấn luyện đó, hệ thống sẽ tự động tạo các bộ luật và lưu trữ lại để sử dụng trong việc chẩn đoán bệnh. Bên cạnh đó, hệ thống quản lý cho phép người quản lý có thể thêm vào những luật lấy từ bác sĩ, chuyên gia nghiên cứu về bệnh tự kỷ để tăng thêm độ chính xác trong quá trình chẩn đoán.

Với website người dùng, hệ thống cho phép người dùng đăng nhập vào sử dụng chức năng chẩn đoán bệnh một cách dễ dàng. Ngoài ra, hệ thống còn cung cấp những thông tin cần thiết về bệnh tự kỷ.

3.2. Kịch bản triển khai hệ thống chẩn đoán

Hệ thống chẩn đoán bệnh tự kỷ được xây dựng để đáp ứng các bước sau:



Hình 1. Kịch bản triển khai hệ thống

Bước 1: Thu thập dữ liệu (tạo bộ dữ liệu training)

Muốn xây dựng hệ thống chẩn đoán có độ chính xác cao thì cần có một bộ dữ liệu training đủ lớn. Để đảm bảo điều này, hệ thống cung cấp chức năng cho phép mở rộng bộ dữ liệu bằng việc người quản lý có thể thêm dữ liệu training thường xuyên và dễ dàng (dữ liệu càng lớn thì độ chính xác khi chẩn đoán càng cao).

Cấu trúc của bảng dữ liệu training:

Tên thuộc tính	Miền giá trị
A: Quan hệ xã hội	A1: Không biểu hiện khó khăn hoặc bất thường trong quan hệ với mọi người. A2: Một chút bất thường trong quan hệ với mọi người. A3: Bất thường ở mức độ trung bình trong quan hệ với mọi người. A4: Bất thường ở mức độ nghiêm trọng trong quan hệ với mọi người.
B: Khả năng bắt chước	B1: Bắt chước giống như trẻ bình thường khác. B2: Bất thường ở mức độ nhẹ về kỹ năng bắt chước. B3: Bất thường mức độ trung bình về kỹ năng bắt chước. B4: Bất thường ở mức độ nghiêm trọng về kỹ năng bắt chước.
C: Đáp ứng tình cảm	C1: Đáp ứng tình cảm phù hợp với tuổi và phù hợp với hoàn cảnh. C2: Bất thường ở mức độ nhẹ về đáp ứng tình cảm. C3: Bất thường ở mức độ trung bình về đáp

	ứng tình cảm. C4: Bất thường ở mức độ nặng trong đáp ứng tình cảm.
D: Các động tác cơ thể	D1: Các động tác của cơ thể phù hợp với tuổi. D2: Bất thường ở mức độ nhẹ các động tác cơ thể. D3: Bất thường (ở mức độ trung bình) các động tác cơ thể. D4: Bất thường ở mức độ nghiêm trọng các động tác cơ thể.
E: Sử dụng đồ vật	E1: Sử dụng đồ vật phù hợp, thích thú với đồ chơi và các đồ vật khác. E2: Thiếu thích hợp nhỏ trong việc sử dụng đồ vật, thiếu thích thú với đồ chơi và các đồ vật khác. E3: Thiếu thích hợp trung bình trong việc sử dụng đồ chơi và các đồ vật khác, bất thường trong yêu thích đồ chơi và các đồ vật khác. E4: Thiếu thích hợp một cách nghiêm trọng trong việc sử dụng đồ chơi và các đồ vật khác, bất thường nghiêm trọng trong việc thích thú đồ vật.
F: Thích nghi với sự thay đổi	F1: Đáp ứng lại sự thay đổi phù hợp với tuổi. F2: Bất thường nhỏ trong việc thích nghi với sự thay đổi. F3: Bất thường (mức độ trung bình) trong sự thích nghi với những thay đổi. F4: Bất thường trầm trọng trong việc thích nghi với sự thay đổi.
G: Phản ứng thị giác	G1: Phản ứng thị giác bình thường và phù hợp với tuổi. G2: Bất thường nhỏ về thị giác, thị thoảng phải nhắc nhở trẻ chú ý đến mục tiêu. G3: Bất thường mức trung bình về thị giác nhìn. G4: Bất thường nghiêm trọng về thị giác nhìn.
H: Phản ứng thính giác	H1: Phản ứng thính giác phù hợp với tuổi. H2: Bất thường nhỏ về hoạt động của thính giác. H3: Bất thường ở mức độ trung bình về khả năng nghe. H4: Bất thường nghiêm trọng về hoạt động của thính giác.
I: Phản ứng qua vị, khứu, xúc giác và khả năng sử dụng các giác quan này	I1: Hoạt động bình thường của các hành vi: ngửi, nếm, sờ mó, đụng chạm. I2: Bất thường nhẹ trong hoạt động của các hành vi: ngửi, nếm, sờ mó, đụng chạm. I3: Bất thường ở mức độ trung bình trong hoạt động của các hành vi: ngửi, nếm, sờ mó, đụng chạm. I4: Bất thường nghiêm trọng trong hoạt động của các hành vi: ngửi, nếm, sờ mó, đụng chạm.
J: Sợ hãi hoặc hồi hộp	J1: Sợ hãi và hồi hộp một cách bình thường. J2: Bất thường nhẹ về sợ hãi và hồi hộp. J3: Bất thường ở mức độ trung bình về sợ hãi và hồi hộp. J4: Bất thường nghiêm trọng về sợ hãi và hồi hộp.

K: Giao tiếp bằng lời	K1: Giao tiếp bằng lời phù hợp với tuổi và tình huống. K2: Bất thường nhẹ về giao tiếp bằng lời, trẻ chậm nói. K3: Bất thường mức độ trung bình về giao tiếp bằng lời. K4: Bất thường nghiêm trọng về giao tiếp bằng lời.
L: Giao tiếp không lời	L1: Giao tiếp không lời bình thường, phù hợp với tuổi và tình huống. L2: Bất thường nhẹ về giao tiếp không lời. L3: Bất thường mức độ trung bình. L4: Bất thường nghiêm trọng về giao tiếp không lời.
M: Mức độ hoạt động	M1: Mức độ hoạt động bình thường, phù hợp với tuổi và tình huống. M2: Bất thường nhẹ về mức độ hoạt động. M3: Bất thường trung bình về mức độ hoạt động. M4: Bất thường nghiêm trọng về mức độ hoạt động.
N: Đáp ứng trí tuệ	N1: Đáp ứng trí tuệ bình thường, phù hợp với tuổi. N2: Bất thường nhẹ về trí tuệ. N3: Bất thường trung bình về trí tuệ. N4: Bất thường nghiêm trọng về trí tuệ.
P: kết quả	P1: Trẻ không bị tự kỷ. P2: Trẻ có khả năng mắc tự kỷ nhưng ở mức độ nhẹ. P3: Trẻ có khả năng mắc tự kỷ ở mức độ trung bình P4: Trẻ bị tự kỷ

Bước 2: Xây dựng cây quyết định

Hệ thống tích hợp thuật toán ID3 [5] cho phép xây dựng cây quyết định trực tiếp trên môi trường web mà không cần sử dụng công cụ thứ 3. Hơn nữa, việc lưu trữ trực tiếp bộ dữ liệu mẫu trong hệ thống giúp giảm thời gian và chi phí thực hiện xây dựng cây quyết định.

Thuật toán ID3:

<i>Thuật toán tạo cây:</i> <i>Dữ liệu vào:</i> Bảng quyết định DT = (U, C $\cup$ {d}) <i>Dữ liệu ra:</i> Mô hình cây quyết định. <pre> Function Create_tree (U, C, {d}) Begin If tất cả các mẫu thuộc cùng nhãn lớp d_i Then Return một nút lá được gán nhãn d_i Else IF C = null Then Return nút lá có nhãn d_j là lớp phổ biến nhất trong DT Else Begin bestAttribute:= getBestAttribute (U, C); // Chọn thuộc tính tốt nhất để chia C:= C - {bestAttribute}; End End End End </pre>
--

<pre> //xóa bestAttribute khỏi tập thuộc tính với mỗi v in bestAttribute Begin U_v:= [U]_v //U_v là phân hoạch của U ChildNode:= Create_tree(U_v, C, {d}); End End End </pre>
<i>Hàm getBestAttribute():</i> <i>Dữ liệu vào:</i> Bảng quyết định DT = (U, C $\cup$ {d}) <i>Dữ liệu ra:</i> Thuộc tính điều kiện tốt nhất <pre> Function getBestAttribute (U, C); Begin maxIG:= 0; Với mỗi c in C Begin temp:= IG(U, c); // Tính lượng thông tin thu thêm IG(U,c) If (temp > maxIG) Then Begin maxIG:= temp; kq:= c; End End End Return result; //Hàm trả về thuộc tính có lượng thông tin thu thêm IG là lớn nhất End </pre>

Bước 3: Rút tri thức từ cây quyết định

Từ kết quả đạt được khi xây dựng cây quyết định, các tri thức sẽ được phát hiện và lưu trữ thành một bộ các luật theo dạng IF...THEN như sau:

IF (A = "A1" AND B = "B2" AND... AND N = "N1")THEN P = "P1"

Trong đó:

A = "A1": là biểu thức logic về giá trị tương ứng của các thuộc tính biểu hiện.

P: Kết quả của biến cần chẩn đoán.

Ví dụ:

IF (A = "A3" AND D = "D4" AND K = "K3") THEN P = "P3"

(Nếu trẻ có các biểu hiện: bất thường ở mức độ trung bình trong quan hệ với mọi người, bất thường ở mức độ nghiêm trọng các động tác cơ thể và bất thường mức độ trung bình về giao tiếp bằng lời thì trẻ có khả năng mắc bệnh tự kỷ ức mức độ trung bình).

Bước 4: Ứng dụng tri thức vào chẩn đoán bệnh

Với tri thức phân tích được việc khai phá bộ dữ liệu mẫu, hệ thống cung cấp website tương tác với người dùng cho phép người dùng cung cấp thông tin liên quan đến trẻ tương ứng với dữ liệu đầu vào của hệ thống bằng cách chọn những biểu hiện, từ đó hệ thống đưa ra kết quả chẩn đoán khả năng mắc bệnh tự kỷ.

Kết quả thử nghiệm chẩn đoán:

b1: Chọn thông tin các biểu hiện của trẻ (Hình 3).

b2: Chọn nút chẩn đoán.

Kết quả: Trẻ không có dấu hiệu của bệnh tự kỷ (Hình 4).

