Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Huế, ngày 07-08/6/2019
DOI: 10.15625/vap.2019.00013

DATA MINING IN HEALTHCARE SYSTEM ON PATIENTS CLINICAL SYMPTOMS DATASET

Trần Đình Toàn[1], Huỳnh Thị Châu Lan[1], Trần Văn Thọ[1], Lê Minh Hưng[2], Trần Văn Lăng[3]

[1]Trường Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh
[2]Trường Đại học Công nghệ Thông tin - Đại học Quốc gia Thành phố Hồ Chí Minh
[3]Viện Hàn lâm Khoa học và Công nghệ Việt Nam

toantd@cntp.edu.vn, lanhtc@hufi.edu.vn, thotv@hufi.edu.vn, hungml@uit.edu.vn, langtv@vast.ac.vn

TÓM TẮT: *Bài báo này khảo sát một số các kỹ thuật khai phá dữ liệu (KPDL) thường được sử dụng trong hệ thống chăm sóc sức khỏe thông minh (E-Health Care Systems). Tại Việt Nam, việc ứng dụng KPDL trong lĩnh vực y khoa còn nhiều hạn chế đặc biệt là việc tiếp cận và thu thập dữ liệu bệnh nhân. Trong khảo sát này, chúng tôi tiến hành hai thực nghiệm áp dụng một số phương pháp học có giám sát để phân loại bệnh trên ba bộ dữ liệu cận lâm sàng là bệnh tim, bệnh thận mãn tính và ung thư vú, trong đó Thực nghiệm 1 chúng tôi tiến hành kiểm tra đặc tính của dữ liệu có phân bố tuyến tính hay phi tuyến, Thực nghiệm 2 tiến hành phân loại bệnh và so sánh các kết quả đạt được. Từ đó đánh giá tính hiệu quả việc của các kỹ thuật trên từng bộ dữ liệu khác nhau.*

Từ khóa*: Khai phá dữ liệu, chăm sóc sức khỏe cộng đồng, học có giám sát, máy vector hỗ trợ.*

I. GIỚI THIỆU

Thời gần đây, do nhu cầu ngày càng cao về các phương pháp phân tích dữ liệu y khoa để phát hiện thông tin có giá trị nên kỹ thuật khai phá dữ liệu đang được sử dụng phổ biến. Các kỹ thuật khai phá dữ liệu được phát triển trong những năm gần đây bao gồm khái quát hóa, đặc tính hóa, phân loại, phân cụm, kết hợp, so khớp mẫu, trực quan hóa dữ liệu và lựa chọn đặc trưng. Trong lĩnh vực y khoa, áp dụng kỹ thuật khai phá dữ liệu đem đến một số lợi ích cho các nhà chuyên môn và bệnh nhân như phát hiện sớm các bệnh, đưa ra giải pháp y tế cho bệnh nhân với chi phí thấp hơn, phát hiện nguyên nhân bệnh và xác định các phương pháp điều trị y tế, và đồng thời phát hiện gian lận trong chăm sóc sức khỏe. Nó cũng giúp các nhà quản lý và nhà nghiên cứu chăm sóc sức khỏe thực hiện các chính sách chăm sóc sức khỏe hiệu quả, phát triển hồ sơ y tế của các cá nhân, xây dựng hệ thống khuyến cáo thuốc,... Các kỹ thuật khai phá dữ liệu cũng được sử dụng để phân tích các yếu tố khác nhau đối với các bệnh như loại thực phẩm, môi trường làm việc khác nhau, trình độ học vấn, điều kiện sống, nguồn nước sạch, dịch vụ chăm sóc sức khỏe, văn hóa, môi trường. Các phương pháp khai phá dữ liệu gần như thuộc một trong hai phương pháp phân tích chính của máy học (Machine Learning): học có giám sát (Supervised learning) và không giám sát (Unsupervised learning). Mục đích chính của phương pháp học có giám sát là phát triển một mô hình để dự đoán tình huống dựa trên nhãn lớp, trong khi phương pháp học không giám sát phân lớp các loại bệnh dựa trên dữ liệu không có nhãn lớp mà dựa trên những tính chất của dữ liệu; mục đích là mô hình hóa phân phối trong dữ liệu để tìm hiểu thêm về dữ liệu.

Bài báo này khảo sát các kết quả đã công bố trong thời gian gần đây về việc sử dụng các kỹ thuật khai phá dữ liệu trong hệ thống hỗ trợ chăm sóc sức khỏe, các bài viết được thu thập dựa trên việc ứng dụng các kỹ thuật khai phá dữ liệu trong chăm sóc sức khỏe, không có phân loại cụ thể. Deepika và cộng sự đã phân loại bệnh nhân đau tim bằng cách đề xuất sử dụng quy tắc kết hợp [1]. K. Srinivas và cộng sự đã dự đoán các cơn đau tim (heart attack) dựa trên thuật toán Naïve Bayes, K-NN, cây quyết định, trong đó cây quyết định đạt được hiệu suất tốt nhất [2]. Tương tự, để dự đoán các bệnh đột quỵ một số thuật toán phân loại bao gồm Naïve Bayes, cây quyết định và mạng neuron được sử dụng, kết quả thử nghiệm cho thấy mạng neuron hoạt động tốt hơn nhiều so với hai thuật toán còn lại [3]. Jabbar và cộng sự đã dự đoán đau tim khi đề xuất khai phá quy tắc kết hợp dựa trên số thứ tự và phân cụm, trong đó các mẫu được trích xuất từ cơ sở dữ liệu với tính toán trọng số [4]. Shouman và cộng sự đã dự đoán bệnh tim khi kết hợp phân cụm K-means với phương pháp cây quyết định trên tập hợp con gồm 13 thuộc tính đầu vào, nghiên cứu này cho thấy rằng việc kết hợp phân cụm K-means và cây quyết định có thể chẩn đoán bệnh nhân bị bệnh tim đạt được độ chính xác cao hơn các phương pháp truyền thống khác [5]. Olatubosun và cộng sự đã chẩn đoán bệnh mạch máu não với đề xuất sử dụng mạng neuron nhân tạo với cơ chế lan truyền ngược [6]. M. Anbarasi và cộng sự đề xuất dự đoán bệnh tim với lựa chọn đặc trưng bằng thuật toán di truyền [7]. Singh và cộng sự đã đề xuất sử dụng phương pháp lựa chọn đặc trưng di truyền kết hợp với phương pháp Naïve Bayes để dự đoán bệnh tim [8]. Takci đã tìm phương pháp học máy tốt nhất và phương pháp lựa chọn đặc trưng để dự đoán các cơn đau tim, trong đó SVM với nhân tuyến tính kết hợp với đặc trưng Relief–Based đạt được hiệu suất tốt nhất [9]. Hưng M.L. cùng cộng sự đã đề xuất một phương pháp lựa chọn đặc trưng và kỹ thuật khai phá dữ liệu để dự đoán các nhóm bệnh tim khác nhau [10]. Patel cùng cộng sự đề xuất giảm số lượng thuộc tính đầu vào bằng cách sử dụng các kỹ thuật phân loại cây trong khai phá dữ liệu bao gồm Nave Bayes, cây quyết định và phân loại theo cụm, trong đó cây quyết định đạt được hiệu suất tốt nhất [11]. Tương tự, Suganya và cộng sự đề xuất một phương pháp lựa chọn đặc trưng mới cho dự đoán bệnh tim trên 13 thuộc tính được chọn với tổng số 303 trường hợp của số liệu bệnh nhân [12]. Mirmozaflari cùng cộng sự áp dụng phương pháp phân cụm trong công cụ WEKA trên một tập dữ liệu bệnh nhân với 8 thuộc tính và tổng số 209 trường hợp chẩn đoán bệnh tim [13]. Uma cùng cộng sự đã áp dụng một số thuật toán phân loại và phương pháp chọn lọc đặc trưng trên tập con của tập dữ liệu với 18 thuộc tính và tổng số 689 trường hợp, kết quả đã chứng minh rằng SVM đạt được hiệu suất tốt nhất trong số các

bộ phân loại và hầu hết các phương pháp chọn đặc trưng được chấp nhận đạt được độ chính xác gần như nhau [14]. Trong khi đó để chẩn đoán bệnh nhân mắc bệnh tim Ziasabounchi và Askerzade đã đề xuất sử dụng phương pháp phân cụm dựa trên PCA [15].

Một cách cụ thể, bài báo có những đóng góp chính như sau:

- Cung cấp một cách nhìn tổng quát về các phương pháp phổ biến được ứng dụng trong lĩnh vực chăm sóc sức khỏe thời gian gần đây. Đồng thời cũng chỉ ra các phương pháp phân loại khác nhau có những ưu điểm và nhược điểm, và tùy theo đặc tính của dữ liệu, cần cân nhắc việc sử dụng các bộ phân loại thích hợp để đạt được hiệu suất cao nhất.
- Để phân biệt được dữ liệu có phân bố tuyến tính hay phi tuyến trên không gian đa chiều bài báo đã chỉ ra cách thức phân tích đặc tính của dữ liệu.
- Bài báo chỉ ra sử dụng bộ phân lớp phù hợp sẽ cho hiệu suất hệ thống tốt nhất.

Phần còn lại của bài báo được tổ chức như sau: trong phần 2 chúng tôi trình bày giản lược một số kỹ thuật đại diện cho học có giám sát để phân nhóm các bệnh; phần 3 trình bày về kết quả thực nghiệm và bàn luận, phần 4 kết luận.

II. PHƯƠNG PHÁP NGHIÊN CỨU

Trong khảo sát này, chúng tôi trình bày một số kỹ thuật học có giám sát (SVM, J48, Naïve Bayes) thường được sử dụng trong khai phá các dữ liệu y khoa cận lâm sàng [16] [21] [22] để tiến hành phân lớp dữ liệu:

2.1. Phương pháp máy vector hỗ trợ SVM (Support Vector Machines)

SVM là một thuật toán học máy có giám sát được sử dụng rất phổ biến ngày nay trong các bài toán phân lớp (classification) hay hồi quy (Regression) [10]. SVM được đề xuất bởi Vladimir N. Vapnik và các đồng nghiệp của ông vào năm 1963 tại Nga và sau đó trở nên phổ biến trong những năm 90 nhờ ứng dụng giải quyết các bài toán phi tuyến tính (nonlinear) bằng phương pháp Kernel Trick.

Ý tưởng chính của SVM là xây dựng một siêu phẳng để phân tách và tối đa hóa lề của các lớp dương (có nguy cơ) và âm (không có nguy cơ) với lề lớn nhất. Giả sử $\{(x_i, y_i)\}_{i=1}^{N}$ là tập mẫu huấn luyện mà chứa các thuộc tính phân biệt nhất, (x_i, y_i) là đặc trưng đầu vào thứ i và đầu ra tương ứng của nó. Ranh giới quyết định sự phân tách được thực hiện thông qua phương trình:

$$w^T x_i + b \geq 0 \text{ với } y_i = +1 \text{ (lớp dương)} \quad (1)$$

$$w^T x_i + b < 0 \text{ với } y_i = -1 \text{ (lớp âm)} \quad (2)$$

Trong đó, w là một vector trọng số điều chỉnh, x là vector đầu vào và b là một hằng số (bias). Vấn đề tối ưu hóa của SVM có thể được định nghĩa như sau:

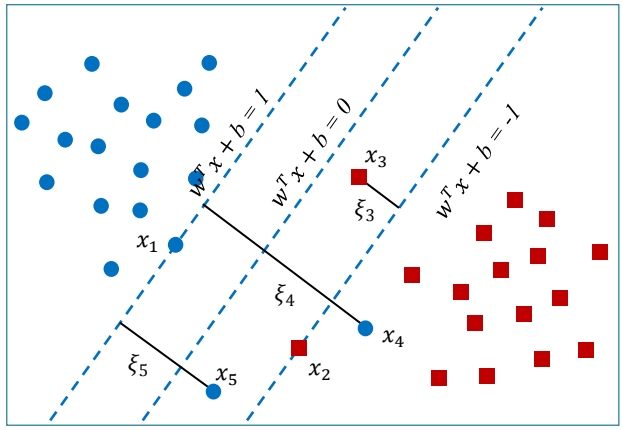
$$(\mathrm{w}, \mathrm{b}) = \arg\min_{w,b} \frac{1}{2}||w||_2^2 \; s.t \; y_i(w^T.x_i + b) \geq 1, \forall i = 1, 2, \dots, N. \quad (3)$$

SVM thường làm việc với các đặc trưng tách biệt tuyến tính. Tuy nhiên, trong một số trường hợp khi có nhiễu, đặc trưng thuộc về một lớp mà rất gần với lớp khác. Với trường hợp này, SVM sẽ tạo ra một siêu phẳng có lề rất nhỏ, rất nhạy với nhiễu. Nếu thuật toán loại bỏ được nhiễu thì SVM có thể tạo ra một siêu phẳng với biên độ tốt hơn để phân tách tốt nhất hai lớp. Một số trường hợp khác là khi hai lớp có thể phân tách tuyến tính gần nhau, trong đó tồn tại một số lượng nhỏ các trường hợp xuất hiện không đáng tin cậy, thuật toán tối ưu hóa lề SVM là không khả thi. Tương tự, nếu thuật toán bỏ qua các trường hợp đó, SVM cũng tạo ra một lề tốt hơn mà hầu hết có thể tách hai lớp. Kỹ thuật này được gọi là SVM với lề mềm (Soft Margin). Việc hình thành bài toán tối ưu hóa SVM có thể được viết lại như sau:

$$(\mathrm{w}, \mathrm{b}, \xi) = \arg\min_{w,b,\xi} \frac{1}{2}\left|\left|w\right|\right|_2^2 + C\sum_{i=1}^{N} \xi_i \; s.t \; 1 - \xi_i - y_i(w^T.x_i + b) \geq 1,$$

$$\forall i = 1, 2, \dots, N, \xi_i \geq 0, C > 0 \quad (4)$$

Trong đó C là hằng số được sử dụng để tránh vượt quá giới hạn, $\xi = [\xi_1, \xi_2, \dots, \xi_N]$ là tập hợp các biến slack. Như được hiển thị trong Hình 1, đối với các biến nằm trên lề an toàn, thì $\xi_i = 0$ (ví dụ: x_1, x_2). Đối với các biến không nằm trong lề an toàn, nhưng vẫn ở phía bên phải của lớp, thì $0 < \xi_i < 1$ (ví dụ x_3). Đối với các biến số nằm ở bên trái của lớp thì $\xi_i > 1$ (ví dụ: x_4, x_5).



Hình 1. SVM với kernel mềm cho các trường hợp khác nhau của biến slack [10]

2.2. Cây quyết định (J48)

Phân loại là quá trình xây dựng mô hình các lớp từ một tập hợp các mẫu tin có chứa nhãn lớp. Thuật toán cây quyết định là tìm ra cách làm việc của vector thuộc tính cho một số trường hợp. Trên cơ sở huấn luyện, các lớp mới sẽ được tìm thấy [17]. Thuật toán này tạo ra các quy tắc để dự đoán biến mục tiêu với sự trợ giúp của thuật toán phân loại cây quyết định [18].

J48 là một thuật toán C4.5 được cài đặt theo ngôn ngữ Java mã nguồn mở. Việc phân loại được thực hiện đệ quy cho đến khi đạt được mọi nút lá, đó là phân loại dữ liệu tốt nhất có thể. Thuật toán J48 tạo ra các quy tắc để có thể nhận dạng dữ liệu. Mục tiêu là dần dần khái quát hóa cây quyết định cho đến khi nó đạt được trạng thái cân bằng linh hoạt và chính xác [19].

Các đặc trưng của thuật toán:

(i) Cả hai thuộc tính rời rạc và liên tục đều được xử lý. Giá trị ngưỡng được quyết định bởi C4.5 để xử lý các thuộc tính liên tục thành các thuộc tính có giá trị rời rạc.

(ii) Thuật toán này cũng xử lý các giá trị còn thiếu trong dữ liệu huấn luyện.

(iii) Sau khi cây được xây dựng đầy đủ, thuật toán thực hiện việc cắt tỉa cây để loại bỏ các nhánh không cần thiết.

2.3. Naïve Bayes

Việc phân lớp dựa trên lý thuyết Bayes được gọi là phân lớp Bayes. Định lý Bayes cung cấp cơ sở cho phân loại Naïve Bayes và Mạng Belief Bayes (BBN). Vấn đề chính với phân lớp Naïve Bayes là nó giả định rằng tất cả các thuộc tính độc lập với nhau trong khi các thuộc tính thuộc lĩnh vực y tế như triệu chứng bệnh và trạng thái sức khỏe có mối tương quan với nhau. Mặc dù giả định các thuộc tính độc lập, phân lớp Naïve Bayes đã cho thấy hiệu quả về độ chính xác vì vậy nếu trong lĩnh vực y tế, các thuộc tính độc lập với nhau thì chúng ta có thể sử dụng phương pháp này. Định lý Bayes tập trung vào phân phối xác suất trước, sau và rời rạc của các mục dữ liệu (data items). Mạng Belief Bayes được sử dụng cho bệnh nhân bị ung thư phổi. Và nó đã được sử dụng rộng rãi bởi nhiều nhà nghiên cứu trong lĩnh vực chăm sóc sức khỏe.

Liu và cộng sự phát triển hệ thống hỗ trợ quyết định bằng BBN để phân tích rủi ro liên quan đến sức khỏe. Curiac và cộng sự, phân tích dữ liệu bệnh nhân tâm thần bằng BBN trong việc đưa ra quyết định quan trọng liên quan đến sức khỏe bệnh nhân và thực hiện thí nghiệm trên dữ liệu thực có được từ Bệnh viện thành phố Lugoj [19].

III. KẾT QUẢ THỰC NGHIỆM VÀ BÀN LUẬN

A. Dữ liệu

1. Dữ liệu bệnh tim (Heart disease)

Cơ sở dữ liệu bệnh tim được sử dụng trong bài khảo sát là tập dữ liệu công khai từ UCI Machine Learning Repository [20]. Bao gồm 4 bộ dữ liệu được thu thập từ 4 bệnh viện khác nhau:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias P_sterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Chúng tôi chọn 3 tập dữ liệu với tổng số 699 mẫu từ tập dữ liệu Cleveland (282 mẫu), Hungarian (294 mẫu) và Switzerland (123 mẫu). Các mẫu trong tập dữ liệu gốc được gán nhãn vào 5 lớp khác nhau: lớp không có nguy cơ và lớp có nguy cơ theo 4 cấp độ. Tổng số mẫu thuộc lớp không nguy cơ là 353 và lớp có nguy cơ là 346. Mỗi mẫu gồm 76 thuộc tính bao gồm các thuộc tính chẩn đoán khác nhau và thông tin y tế được thu thập từ mỗi bệnh nhân. Tuy nhiên, do có khá nhiều thuộc tính không có dữ liệu hoặc thiếu dữ liệu khá nhiều nên được loại bỏ. Vì thế thực nghiệm đã được thực hiện trên tập dữ liệu gồm 58 thuộc tính được mô tả trong Bảng 1. Không giống như hầu hết các nghiên cứu gần đây chỉ thực nghiệm trên tập chỉ có 14 thuộc tính hoặc 6 thuộc tính từ cơ sở dữ liệu này, khảo sát này chúng tôi khám phá đầy đủ hầu hết các thông tin được cung cấp trong bộ dữ liệu gốc.

Bảng 1. Mô tả 58 thuộc tính của bộ dữ liệu bệnh tim

Stt	Thuộc tính	Stt	Thuộc tính
1	Age	30	tpeakbpd: peak exercise blood pressure
2	Sex	31	trestbpd: resting blood pressure
3	painloc: chest pain location	32	exang: exercise induced angina (1 = yes; 0 = no)
4	painexer (1 = provoked by exertion; 0 = otherwise)	33	xhypo: (1 = yes; 0 = no)
5	relrest (1 = relieved after rest; 0 = otherwise)	34	oldpeak = st depression induced by exercise relative to rest
6	cp: chest pain type	35	slope: the slope of the peak exercise st segment
7	trestbps: resting blood pressure	36	rldv5: height at rest
8	Htn	37	rldv5e: height at peak exercise
9	chol: serum cholesterol in mg/dl	38	ca: number of major vessels (0-3) colored by fluoroscopy
10	cigs (cigarettes per day)	39	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
11	years (number of years as a smoker)	40	thalsev
12	fbs: (fasting blood sugar > 120 mg/dl)	41	thalpul
13	famhist: family history of coronary artery disease	42	cmo: month of cardiac
14	restecg: resting electrocardiographic results	43	cday: day of cardiac
15	20 ekgmo (month of exercise ECG reading)	44	cyr: year of cardiac
16	ekgday(day of exercise ECG reading)	45	Lmt
17	ekgyr (year of exercise ECG reading)	46	ladprox
18	dig (digitalis used furring exercise ECG)	47	laddist
19	24 prop (Beta blocker used during exercise ECG)	48	diag
20	nitr (nitrates used during exercise ECG)	49	cxmain
21	pro (calcium channel blocker used during exercise ECG)	50	ramus
22	diuretic (diuretic used during exercise ECG)	51	oml
23	proto: exercise protocol	52	om2
24	thaldur: duration of exercise test in minutes	53	Reaprox
25	thaltime: time when ST measure depression was noted	54	Rcadist
26	met: mets achieved	55	lvx3
27	thalach: maximum heart rate achieved	56	lvx4
28	thalrest: resting heart rate	57	Lvf
29	tpeakbps: peak exercise blood pressure	58	Cathef

2. Dữ liệu bệnh ung thư vú (Breast Cancer)

Bộ dữ liệu Breast Cancer thu được từ University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Tập dữ liệu này có tất cả 286 mẫu thuộc vào 2 lớp khác nhau: lớp thứ nhất (no-recurrence-events) có 201 mẫu và lớp thứ 2 (recurrence-events) có 85 mẫu. Mỗi mẫu được mô tả bởi 9 thuộc tính như Bảng 2, trong đó một số thuộc tính là tuyến tính và một số là định danh.

Bảng 2. Mô tả 9 thuộc tính của bộ dữ liệu bệnh ung thư vú

Stt	Thuộc tính	Stt	Thuộc tính	Stt	Thuộc tính
1	age	4	inv-nodes	7	Breast
2	menopause	5	node-caps	8	breast-quad
3	tumor-size	6	deg-malig	9	Irradiat

3. Dữ liệu bệnh thận mãn tính (Chronic Kidney disease)

Bộ dữ liệu Chronic Kidney disease thu được từ Dr. P. Soundarapandian. M. D., D. M (Senior Consultant Nephrologist), Apollo Hospitals, Managiri, Madurai Main Road, Karaikudi, Tamilnadu, India. Tập dữ liệu này có tất cả 400 mẫu thuộc vào 2 lớp khác nhau: lớp thứ nhất (no-recurrence-events) có 250 mẫu và lớp thứ 2 (recurrence-events) có 150 mẫu. Mỗi mẫu được mô tả bởi 25 thuộc tính như trong Bảng 3.

Bảng 3. Mô tả 25 thuộc tính của bộ dữ liệu thận mãn tính

Stt	Thuộc tính	Stt	Thuộc tính
1	age - age	14	pot – potassium
2	bp - blood pressure	15	hemo - hemoglobin
3	sg - specific gravity	16	pcv - packed cell volume
4	al - albumin	17	wc - while blood cell count
5	su - sugar	18	rc - red blood cell count
6	rbc - red blood cells	19	htn - hypertension
7	pc - pus cell	20	dm - diabetes mellitus
8	pcc - pus cell clumps	21	cad - coronary artery disease
9	ba - bacteria	22	appet - appetite
10	bgr - blood glucose random	23	pe - pedal edema
11	bu - blood urea	24	ane - anemia
12	sc - serum creatinine	25	class - class
13	sod - sodium		

B. Kết quả thực nghiệm

1. Chuẩn hóa dữ liệu

Trong nghiên cứu này, 70% dữ liệu được dùng để huấn luyện và 30% dữ liệu còn lại dùng để đánh giá. Trước khi áp dụng các kỹ thuật, phương pháp z-score được sử dụng để chuẩn hóa dữ liệu. Phương pháp z-score dựa trên giá trị trung bình và độ lệch chuẩn. Giả sử ta có giá trị cũ *v* tương ứng với giá trị trung bình ($\bar{A}$) và độ lệch chuẩn σ_A. Một cách cụ thể, giá trị mới của *v* (ký hiệu v') được tính theo công thức:

$$v' = \frac{v - \bar{A}}{\sigma_A} \tag{5}$$

2. Đánh giá thực nghiệm

Dựa vào độ chính xác (Accurracy), độ nhạy (Sensitivity) và độ đặc hiệu (Specificity) để đánh giá hiệu suất phân loại của hệ thống chẩn đoán bệnh của các thực nghiệm. Diện tích dưới đường cong (Area under the curve) AUC trên đồ thị theo các kỹ thuật thử nghiệm thu được (ROC) cũng cung cấp thông tin cho phân loại nhị phân. Độ chính xác, độ nhạy và độ đặc hiệu cụ thể được xác định như sau:

$$\text{Độ chính xác} = \frac{TP+TN}{TP+FP+FN+TN} \tag{6}$$

$$\text{Độ nhạy} = \frac{TP}{TP+FN} \tag{7}$$

$$\text{Độ đặc } hiệu = \frac{TN}{TN+FP} \tag{8}$$

Trong đó: TP, FP, TN và FN lần lượt là True Positive, False Posotive, True Negative và False Negative. Trong khảo sát này, chúng tôi xét hiệu suất tốt nhất về độ chính xác và AUC.

Trong bài báo này, các thực nghiệm được thực hiện trên ngôn ngữ MATLAB.

3. Mô tả thực nghiệm

Trong bài khảo sát này tiến hành hai thực nghiệm trên 3 bộ dữ liệu khác nhau. Trong đó, 70% dữ liệu được dùng để huấn luyện và 30% dữ liệu còn lại được dùng để kiểm thử. Trước khi tiến hành thực nghiệm phân lớp dữ liệu với các kỹ thuật nêu trên, chúng tôi đã thực hiện thực nghiệm 1 nhằm mục đích kiểm tra dữ liệu phân bố tuyến tính hay phi tuyến. Điều này có thể giúp ích cho việc lựa chọn phương pháp phân loại hiệu quả.

Thực nghiệm 1: Kiểm tra đặc tính của dữ liệu

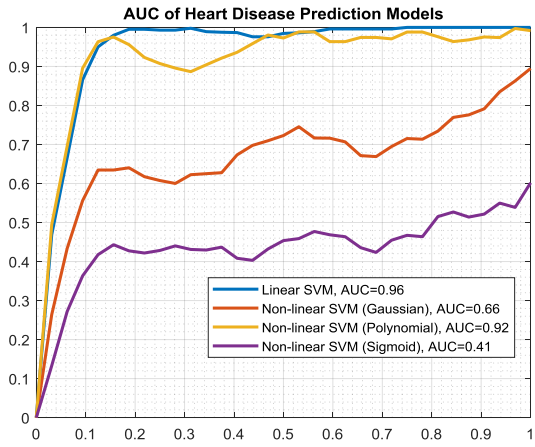
Trong thực nghiệm này chúng tôi đề xuất sử dụng kỹ thuật SVM để kiểm tra dữ liệu phân bố tuyến tính hay phi tuyến. Cụ thể thực nghiệm sử dụng kỹ thuật Linear SVM và Non-linear SVM với 3 nhân (kernel) khác nhau (Gaussian, Polynomial và Sigmoid).

a. Bệnh tim

Bảng 4. So sánh đặc tính bệnh tim sử dụng kỹ thuật SVM tuyến tính và phi tuyến

Methods	Accuracy (%)	AUC	Sensitivity	Specificity
Linear SVM	89,93	0,96	0,87	0,93
Non-linear SVM (Gaussian)	49,64	0,66	0,00	1,00
Non-linear SVM (Polynomial)	83,45	0,92	0,85	0,81
Non-linear SVM (Sigmoid)	49,64	0,41	0,00	1,00

Dựa vào bảng kết quả thực nghiệm cho thấy kỹ thuật linear SVM đạt kết quả cao nhất với độ chính xác 89,93% và AUC 0,96. Trong khi đó, kỹ thuật Non-linear SVM với hàm nhân Gaussian và Sigmoid lại cho kết quả khá thấp như mô tả trong Bảng 4 và Hình 2. Nguyên nhân có thể là do nhiễu hoặc dữ liệu quá khớp. Với kết quả thực nghiệm thu được thì dữ liệu bệnh tim có phân bố tuyến tính.



Hình 2. Kết quả AUC của thực nghiệm 1 với bệnh tim

b. Bệnh thận mãn tính

Bảng 5. So sánh đặc tính bệnh thận mãn tính sử dụng kỹ thuật SVM tuyến tính và phi tuyến

Methods	Accuracy (%)	AUC	Sensitivity	Specificity
Linear SVM	72,50	0,90	0,96	0,59
Non-linear SVM (Gaussian)	100	1,00	1,00	1,00
Non-linear SVM (Polynomial)	98,75	1,00	1,00	0,96

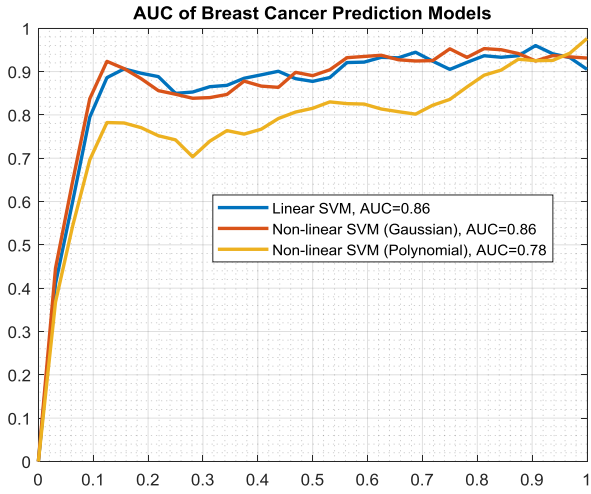
Với thực nghiệm tương tự trên bộ dữ liệu bệnh thận mãn tính thì kỹ thuật Non-linear SVM cho kết quả rất cao với cả 2 nhân (Gaussian và Polynomial) và đạt độ chính xác 100% với AUC 1 như trong Bảng 5. Trong khi Linear SVM chỉ đạt độ chính xác 72,50% và AUC 0,90. Điều này cho thấy, dữ liệu bệnh thận mãn tính có phân bố phi tuyến.

c. Bệnh ung thư vú

Bảng 6. So sánh đặc tính bệnh ung thư vú sử dụng kỹ thuật SVM tuyến tính và phi tuyến

Methods	Accuracy (%)	AUC	Sensitivity	Specificity
Linear SVM	82,61	0,86	0,75	0,90
Non-linear SVM (Gaussian)	86,96	0,86	0,77	1,00
Non-linear SVM (Polynomial)	78,26	0,78	0,72	0,83

Kết quả thực nghiệm tương tự tiếp theo trên bộ dữ liệu bệnh ung thư vú cho thấy kỹ thuật Non-linear SVM với nhân Gaussian và Linear SVM đều cho kết quả khá cao với độ chính xác 86,96% và có cùng AUC 0,86. Nhưng Non-linear SVM với nhân Polynomial có độ chính xác cũng khá cao như thể hiện trong Bảng 6. Điều này cho thấy, dữ liệu bệnh ung thư vú có phân bố phi tuyến, nhưng cũng có thể là tuyến tính.



Hình 3. Kết quả AUC của thực nghiệm 1 với bệnh ung thư vú

Thực nghiệm 2: So sánh các kỹ thuật phân loại khác nhau

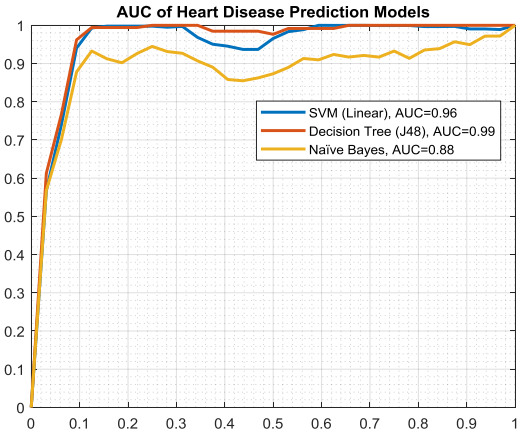
Chúng tôi thử nghiệm 5 phương pháp phân loại (SVM, J48, Naïve Bayes) trên 3 bộ dữ liệu (bệnh tim, thận mãn tính và ung thư vú). Kết quả chi tiết như sau:

a. Bệnh tim

Bảng 7. So sánh 5 kỹ thuật phân loại trên bệnh tim

Methods	Accuracy (%)	AUC	Sensitivity	Specificity
SVM (Linear)	89,93	0,96	0,87	0,93
Decision Tree (J48)	99,28	0,99	0,98	1,00
Naïve Bayes	74,82	0,88	0,89	0,68

Kết quả thực nghiệm cho thấy kỹ thuật cây quyết định (J48) cho kết quả tốt nhất với độ chính xác 99,28% và AUC 0,99; xếp sau là kỹ thuật SVM với độ chính xác 89,93% và AUC 0,96; kỹ thuật Naïve Bayes cho kết quả không cao như mô tả trong Bảng 7 và Hình 4.



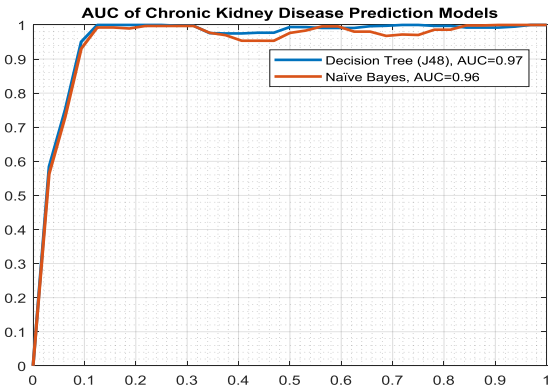
Hình 4. Kết quả AUC của thực nghiệm 2 với bệnh tim

b. Bệnh thận mãn tính

Bảng 8. So sánh 5 kỹ thuật phân loại trên bệnh thận mãn tính

Methods	Accuracy (%)	AUC	Sensitivity	Specificity
SVM (Gaussian)	100	1,00	1,00	1,00
Decision Tree (J48)	93,75	0,97	1,00	0,86
Naïve Bayes	90,00	0,96	0,90	0,90

Kết quả thực nghiệm cho thấy kỹ thuật SVM với nhân Gaussian cho kết quả tốt nhất với độ chính xác 100% và AUC 1; xếp sau là kỹ thuật cây quyết định (J48) cho kết quả với độ chính xác 93,75% và AUC 0,97; kỹ thuật Naïve Bayes cho kết quả với độ chính xác 90% và AUC 0,96; như mô tả trong Bảng 8 và Hình 5.



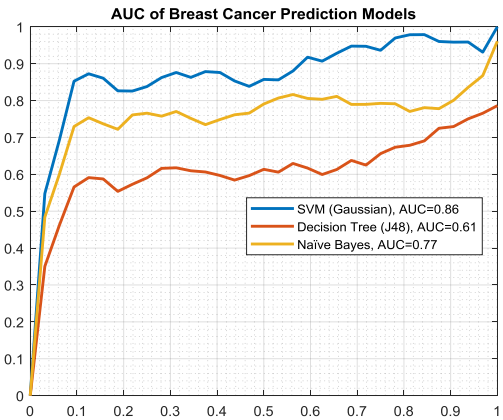
Hình 5. Kết quả AUC của thực nghiệm 2 với bệnh thận mãn tính

c. Bệnh ung thư vú

Bảng 9. So sánh 5 kỹ thuật phân loại trên bệnh ung thư vú

Methods	Accuracy (%)	AUC	Sensitivity	Specificity
SVM (Gaussian)	86,96	0,86	0,77	1,00
Decision Tree (J48)	69,57	0,61	0,67	0,71
Naïve Bayes	69,57	0,77	0,80	0,67

Kết quả thực nghiệm cho thấy kỹ thuật SVM với nhân Gaussian cho kết quả tốt nhất với độ chính xác 86,96% và AUC 0,86; Các kỹ thuật còn lại cho kết quả khá thấp như mô tả trong Bảng 9 và Hình 6.



Hình 6. Kết quả AUC của thực nghiệm 2 với bệnh ung thư vú

IV. KẾT LUẬN

Trong khảo sát này, chúng tôi đã tiến hành khảo sát 2 thực nghiệm trên 3 bộ dữ liệu của 3 bệnh là bệnh tim, bệnh thận mãn tính và bệnh ung thư vú. Thực nghiệm 1 sử dụng phương pháp SVM với các nhân khác nhau để bước đầu phân loại và đánh giá dữ liệu là tuyến tính hay phi tuyến. Thực nghiệm 2 dùng để so sánh các kỹ thuật phân loại khác nhau trên 3 bộ dữ liệu trên là SVM (Linear), Decision Tree (J48) và Naïve Bayes. Kết quả thực nghiệm cho thấy rằng: (1) Với bệnh tim thì Decision Tree (J48) cho kết quả tốt nhất; (2) Với bệnh thận mãn tính thì SVM với nhân Gaussian cho kết quả tốt nhất và; (3) Với bệnh ung thư vú thì SVM với nhân Gaussian cũng cho kết quả tốt nhất. Tuy nhiên tùy vào đặc tính khác nhau của dữ liệu mà các phương pháp khác nhau có thể cho kết quả khác nhau. Nhìn chung SVM với các nhân khác nhau đều cho kết quả tốt và đáng tin cậy với độ chính xác trung bình là khoảng 92,30% và trung bình của AUC là khoảng 0,94. Qua thực nghiệm cho thấy, việc xác định dữ liệu phân bố tuyến tính hay phi tuyến để dùng phương pháp phân loại dữ liệu phù hợp sẽ tăng hiệu quả và độ tin cậy của hệ thống.

Chúng tôi đã khảo sát và áp dụng thành công một số kỹ thuật khai phá dữ liệu áp dụng trong hỗ trợ chăm sóc khỏe cộng đồng trên dữ liệu cận lâm sàng. Kết quả thực nghiệm cho thấy SVM với các nhân khác nhau đều cho ra kết quả chính xác hơn các phương pháp còn lại. Bộ dữ liệu sử dụng trong thực nghiệm của bài báo này là bộ dữ liệu công cộng dành cho nghiên cứu về y khoa, nhưng xét về mặt đặc tính thì các thông tin của bệnh nhân đều khá tương đồng giữa các chủng tộc người khác nhau trên thế giới, do đó kết quả của bài báo này có thể áp dụng tốt trong hỗ trợ các bác sĩ chẩn đoán bệnh tại Việt Nam. Công việc tương lai của chúng tôi là tiếp tục khảo sát thêm những kỹ thuật phân loại dữ liệu khác, từ đó tìm ra thêm những kỹ thuật phân loại dữ liệu tốt nhất trên nhiều tập dữ liệu chăm sóc sức khỏe khác nhau. Ngoài ra, chúng tôi sẽ tiến hành thu thập thêm các dữ liệu tại Việt Nam để có thể hỗ trợ tốt nhất cho quy trình khám chữa bệnh ở Việt Nam.

V. LỜI CẢM ƠN

Bài viết được hoàn thành dưới sự hỗ trợ của đề tài VAST01.03/19-20 của Viện Hàn lâm Khoa học và Công nghệ Việt Nam

VI. TÀI LIỆU THAM KHẢO

[1] N. Deepika, K. Chandrashekar, "Association rule for classification of heart attack patients", International Journal of Advanced Engineering Science and Technologies, vol. 11, no. 2, pp. 253-57, 2011.

[2] K. Srinivas, B. Kavitha Rani, Dr. A. Govrdhan, "Application of data mining techniques in healthcare and prediction of heart attacks", International Journal on Computer Science and Engineering, vol.2, no.2, pp. 250-255, 2011.

[3] A. Sudha, P. Gayathiri, N. Jaisankar, "Effective analysis and predictive model of stroke disease using classification methods", International Journal of Computer Applications, vol. 43, no. 14, pp. 0975-8887, 2012.

[4] A. Jabbar, Priti Chandra, and B.L. Deekshatulu, "Cluster based association rule mining for heart attack prediction", Journal of Theoretical and Applied Information Technology, vol. 32, no. 2, pp. 196-201, 2011.

[5] Shouman, Mai, Tim Turner, and Rob Stocker, "Integrating decision tree and K-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients", Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[6] Olatubosun, Bola Titilayo, "Cerebrovascular accident attack classification using multilayer feed forward artificial neural network with back propagation error", Journal of Computer Science, vol. 8, no. 1, pp.18-25, 2012.

[7] M. Anbarasi, E. Anupriya, and N.CH.S.N. Iyenga, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm", International Journal of Engineering Science and Technology, vol. 2, no. 10, pp. 5370-5376, 2010.

[8] N. Singh, P. Ferozepur, S. Jindal, "Heart disease prediction using classification and feature selection techniques", International Journal of Advance Research, Ideas and Innovations in Technology, vol. 4, no. 2, 2018.

[9] H. Takci, "Improvement of heart attack prediction by the feature selection methods", Turkish Journal of Electrical Engineering & Computer Sciences, vol. 26, no. 1, pp. 1-10, 2018.

[10] Le Minh Hung, Tran Dinh Toan, Tran Van Lang, "Automatic heart disease prediction using feature Selection and data mining technique", Journal of Computer Science and Cybernetics, ISSN: 1813-9663, 2018.

[11] Patel, Yadav, Shukla. "Predict the diagnosis of heart disease patients using classification mining techniques", IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), 4(2), 61-64, 2013.

[12] Suganya, Rajaram, Abdullah, Rajendran, "A Novel Feature Selection method for predicting heart diseases with Data Mining Techniques", Asian Journal of Information Technology, 15(8), 1314-21, 2016.

[13] Mirmozaffari, Alinezhad, Gilanpour, "Heart Disease Prediction with Data Mining Clustering Algorithms", Int'l Journal of Computing, Communications & Instrumentation Engineering (IJCCIE), 4(1), 2017.

[14] Uma, Hanumathappa, "Heart Disease Prediction Using Classification Techniques with Feature Selection Method", Adarsh Journal of Information Technology, 5(2), 22-29, 2016.

[15] Ziasabounchi, Negar; Askerzade, Iman, "A Comparative Study of Heart Disease Prediction Based on Principal Component Analysis and Clustering Methods", Turkish Journal of Mathematics and Computer Science (TJMCS), 16.17: 18, 2014.

[16] T. Balasubramanian, R. Umarani, "An Analysis on the Impact of Fluoride in Human Health (Dental) using Clustering Data Mining Technique", Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, (2012) March 21-23.

[17] V. Rogeith, S. Magesh, "A Survey On Health Care Data Using Data Mining Techniques" International Journal of Pure and Applied Mathematics, Volume 117 No. 16 2017, 665-672

[18] Nadali, Kakhky, Nosratabadi, "Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system", Electronics Computer Technology (ICECT), 2011 3rd International Conference on, vol.6, no., pp.161,165, 8- 10 April 2011

[19] Korting, Thales Sehn, "C4.5 algorithm and Multivariate Decision Trees". Image Processing Division, National Institute for Space Research--INPE.

[20] http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names, The contents of the heart-disease directory.

[21] Lashari, Saima Anwar, Ibrahim, Senan, Taujuddin, "Application of Data Mining Techniques for Medical Data Classification: A Review", MATEC Web of Conferences. Vol. 150. EDP Sciences, 2018.

[22] Jothi, Neesha, and Wahidah Husain, "Data mining in healthcare–a review", Procedia Computer Science 72 (2015): 306-313.

DATA MINING IN HEALTHCARE SYSTEM ON PATIENTS CLINICAL SYMPTOMS DATASET

Tran Dinh Toan, Huynh Thi Chau Lan, Tran Van Tho, Hoang Tung, Le Minh Hung, Tran Van Lang

ABSTRACT: *This paper survey some of the data mining techniques commonly used in the intelligent healthcare system. In Vietnam, the application of data mining techniques in the medical field has many limitations, especially the access and collection of patient data. In this survey, we conducted two experiments using a number of supervised learning methods to classify the disease on three subclinical datasets: heart disease, chronic kidney disease, and breast cancer, in which Experiment 1, we conduct the test of the characteristics of data with linear or nonlinear distribution, Experiment 2 conducted disease classification and comparison of achieved results. Thereby assessing the effectiveness of the techniques on each different dataset.*