

Chuyên san

CÁC CÔNG TRÌNH NGHIÊN CỨU, PHÁT TRIỂN VÀ ỨNG DỤNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

ISSN 1859 - 3526

ẤN PHẨM KHOA HỌC CỦA TẠP CHÍ THÔNG TIN VÀ TRUYỀN THÔNG - BỘ THÔNG TIN VÀ TRUYỀN THÔNG

CÔNG NGHỆ TTTT

Tập 2021
Số 2
Tháng 12

SỐ ĐẶC BIỆT VNICT 2021

Journal of Research,
Development and
Application on
Information and
Communication
Technology

CÔNG NGHỆ TTTT

Tổng biên tập

Nguyễn Văn Bá

Thường trực Hội đồng biên tập

Chủ tịch

GS. TS. Lê Hoài Bắc,
*Trường Đại học Khoa học Tự nhiên,
Đại học Quốc gia TP. Hồ Chí Minh*

Phó Chủ tịch

GS. TS. Trần Xuân Nam,
Trường Đại học Kỹ thuật Lê Quý Đôn

PGS. TS. Nguyễn Khanh Văn,
Trường Đại học Bách khoa Hà Nội

Thành viên

GS. TS. Nguyễn Thanh Thủy,
Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

PGS. TSKH. Nguyễn Xuân Huy,
Viện Công nghệ thông tin, Viện Hàn Lâm KHCN Việt Nam

GS. TS. Từ Minh Phương,
Học viện Công nghệ Bưu chính Viễn thông

GS. TS. Tzung-Pei Hong,
National University of Kaohsiung, Đài Loan

GS. TS. Nguyễn Xuân Huân,
Middlesex University, Anh

TS. Giuseppe D'Aniello,
University of Salerno, Italia

Thư ký

TS. Nguyễn Quang Hưng

Địa chỉ liên hệ:

115 Trần Duy Hưng, Hà Nội
Tel: (84)-24-37737136, Fax: (84)-24-37737130,
Website <http://ictmag.vn>
Email: chuyensanbcvt@mic.gov.vn
Chi nhánh TP. Hồ Chí Minh
27 Nguyễn Bình Khiêm, Quận 1, TP. Hồ Chí Minh
Tel/Fax: (84)-28-38992898

Liên hệ quảng cáo phát hành:

Phạm Thị Lâm
Email: ptlam@mic.gov.vn, Tel: 84-948503999

Technical Editor-in-Chief

Le Hoai Bac
*University of Science,
Vietnam National University Hochiminh City*

Associate Technical Editor-in-Chief

Tran Xuan Nam
Le Quy Don University of Technology, Vietnam
Nguyen Khanh Van
Hanoi University of Science and Technology

Editor

Nguyen Thanh Thuy,
*University of Engineering and Technology,
Vietnam National University Hanoi*

Nguyen Xuan Huy
*Institute of Information Technology,
Vietnam Academy of Science and Technology*

Tu Minh Phuong
*Posts and Telecommunications Institute of Technology,
Vietnam*

Tzung-Pei Hong
National University of Kaohsiung, Taiwan

Nguyen Xuan Huan
Middlesex University, London, England

Giuseppe D'Aniello
University of Salerno, Italia

Secretary

Nguyen Quang Hung
*Journal of Information and Communication,
Vietnam Ministry of Information and communications*

Giấy phép xuất bản số 365/GP-BTTTT, ngày 19/12/2014;
sửa đổi số 233/GP-BTTTT ngày 23/5/2017.

In tại Công ty CP In Hà Nội.

In xong và nộp lưu chiểu tháng 12/2021.

Giá: 40.000 đồng

MỤC LỤC

Chẩn đoán bệnh trong Y học cổ truyền: Hướng tiếp cận dựa trên Đồ thị tri thức mờ dạng cặp	
..... <i>Cù Kim Long, Trần Mạnh Tuấn, Lê Hoàng Sơn</i>	
..... <i>Phạm Minh Chuẩn, Lương Thị Hồng Lan, Phạm Văn Hải, Nguyễn Thọ Thông</i>	59
Xây dựng hệ thống nhận dạng hành động sử dụng cảm biến gia tốc và các phương pháp học máy trên vi điều khiển hiệu năng thấp	
..... <i>Đào Tô Hiệu, Ngô Văn Công, Nguyễn Quang Huy</i>	
..... <i>Trần Đức Nghĩa, Trần Đức Tân</i>	69
A Reliable Numerical Scheme for a Computer Virus Propagation Model	
..... <i>Phạm Hoài Thu, Hoàng Mạnh Tuấn</i>	77
Nâng cao hiệu quả tra cứu ảnh nhãn hiệu sử dụng cây quyết định và phản hồi liên quan	
..... <i>Nguyễn Việt Anh, Vũ Văn Hiệu, Nguyễn Thị Trang, Lê Ngọc Minh</i>	84
Quá trình tiến hóa sinh kháng thuốc của vi-rút HIV với cây đột biến di truyền theo mô hình Markov	
..... <i>Nguyễn Văn Thế, Tạ Văn Nhân, Nguyễn Thị Kim Duyên</i>	
..... <i>Trịnh Mai Phương, Nguyễn Thị Hồng Minh</i>	94
Ngân sách tối thiểu phát hiện nguồn thông tin sai lệch trên mạng xã hội trực tuyến, đảm bảo đạt ít nhất một ngưỡng cho trước	
..... <i>Phạm Văn Dũng, Nguyễn Thị Tuyết Trinh</i>	
..... <i>Vũ Chí Quang, Hà Thị Hồng Vân, Nguyễn Việt Anh</i>	104
Searching 3D Motion Patterns of Vietnamese Traditional Dances	
..... <i>Le Thi Thanh Van, Vu Ngoc Quang</i>	
..... <i>Phạm Thanh Huyền, Ma Thị Châu, Le Thanh Hà</i>	112

Research, Development and Application on **Information and Communication Technology**

<https://www.ictmag.vn>

Volume 2021, No.2, December
VNICT 2021 Special Issue

TABLE OF CONTENTS

Disease Diagnosis in the Traditional Medicine: A Novel Approach based on FKG-Pairs	
..... <i>Cu Kim Long, Tran Manh Tuan, Le Hoang Son</i>	
..... <i>Pham Minh Chuan, Luong Thi Hong Lan, Pham Van Hai, Nguyen Tho Thong</i>	59
Building Human Activity Recognition System using Accelerometers and Machine Learning Methods on Low-Performance Microcontrollers	
..... <i>Dao To Hieu, Ngo Van Cong, Nguyen Quang Huy</i>	
..... <i>Tran Duc Nghia, Tran Duc Tan</i>	69
A Reliable Numerical Scheme for a Computer Virus Propagation Model	
..... <i>Pham Hoai Thu, Hoang Manh Tuan</i>	77
Improving Image Trademark Retrieval Effectiveness using Decision Trees and Relevance Feedback	
..... <i>Nguyen Viet Anh, Vu Van Hieu, Nguyen Thi Trang, Le Ngoc Minh</i>	84
Evolution of Drug Resistance of HIV Virus with Mutagenetic Tree according to Markov Model ..	
..... <i>Nguyen Van The, Ta Van Nhan, Nguyen Thi Kim Duyen</i>	
..... <i>Trinh Mai Phuong, Nguyen Thi Hong Minh</i>	94
Minimum Budget for Misinformation Source Detection in Online Social Networks with Guaranteed to Reach at Least a Given Threshold	
..... <i>Pham Van Dung, Nguyen Thi Tuyet Trinh</i>	
..... <i>Vu Chi Quang, Ha Thi Hong Van, Nguyen Viet Anh</i>	104
Searching 3D Motion Patterns of Vietnamese Traditional Dances	
..... <i>Le Thi Thanh Van, Vu Ngoc Quang</i>	
..... <i>Pham Thanh Huyen, Ma Thi Chau, Le Thanh Ha</i>	112

Chẩn đoán bệnh trong y học cổ truyền: Hướng tiếp cận dựa trên đồ thị tri thức mờ dạng cặp

Cù Kim Long^{1,5}, Trần Mạnh Tuấn², Lê Hoàng Sơn³, Phạm Minh Chuẩn³, Lương Thị Hồng Lan², Phạm Văn Hải¹,
Nguyễn Thọ Thông²

¹ Trường Công nghệ thông tin và Truyền thông, Trường Đại học Bách khoa Hà Nội

² Khoa Công nghệ thông tin, Trường Đại học Thủy lợi

³ Viện Công nghệ thông tin, Đại học Quốc gia Hà Nội

⁴ Khoa Công nghệ thông tin, Trường Đại học Sư phạm Kỹ thuật Hưng Yên

⁵ Trung tâm Công nghệ thông tin, Bộ Khoa học và Công nghệ

Tác giả liên hệ: Trần Mạnh Tuấn, tmtuan@tlu.edu.vn

Ngày nhận bài: 30/07/2021, ngày sửa chữa: 28/10/2021, ngày duyệt đăng: 15/11/2021

Định danh DOI: 10.32913/mic-ict-research-vn.v2021.n2.1003

Tóm tắt: Chẩn đoán bệnh trong y học cổ truyền thường được cốt lại trong bốn cách là “Vọng - Văn - Vấn - Thiết” (hay còn được gọi là “Tứ chẩn”). Trong những năm gần đây, đội ngũ lương y, bác sĩ đã sử dụng kết hợp giữa phác đồ điều trị trong y học cổ truyền với kết quả khám, xét nghiệm trong y học hiện đại nhằm nâng cao chất lượng chẩn đoán bệnh. Điều này đã thúc đẩy mạnh mẽ công tác nghiên cứu y học cổ truyền dân tộc và áp dụng công nghệ thông tin vào trong quá trình khám và chẩn đoán bệnh nhằm mục đích hỗ trợ đội ngũ lương y, bác sĩ tại bệnh viện y học cổ truyền ở các địa phương. Gần đây, hướng nghiên cứu suy luận dựa trên đồ thị tri thức mờ với ưu điểm cho phép thực hiện suy luận trong những trường hợp thiếu tri thức trong kho dữ liệu đã nhận được nhiều sự quan tâm và chú ý của các nhà nghiên cứu trong và ngoài nước. Tuy nhiên, các nghiên cứu chủ yếu tập trung vào khai phá tri thức trong mạng xã hội hay các bài toán ứng dụng trong các lĩnh vực khác mà chưa tập trung vào bài toán y tế. Trong nghiên cứu này, một hướng tiếp cận mới dựa trên mô hình đồ thị tri thức mờ dạng cặp được đề xuất ứng dụng cho bài toán chẩn đoán bệnh trong y học cổ truyền. Mô hình đề xuất được áp dụng đối với bài toán chẩn đoán bệnh cho các sản phụ và đã chứng minh được hiệu quả trong việc suy luận xấp xỉ, chẩn đoán bệnh trong y học cổ truyền.

Từ khóa: Đồ thị tri thức mờ, đồ thị tri thức mờ dạng cặp, suy luận xấp xỉ, chẩn đoán bệnh, y học cổ truyền.

Title: Disease Diagnosis in the Traditional Medicine: A Novel Approach based on FKG-Pairs.

Abstract: Disease diagnosis in traditional medicine is often summarized in four ways: “Vong - Van - Van - Thiet” (or so-called “Four diagnostics”). In recent years, the team of physicians and doctors have used a combination of the treatment regimens in traditional medicine with the results of examination and tests in modern medicine to improve the quality of disease diagnosis. This has strongly promoted the research of traditional medicine and the application of information technology in the process of medical examination and diagnosis in order to support the team of physicians and doctors at the local traditional medical hospitals. Recently, the direction of inferential research based on fuzzy knowledge graphs with the advantage of allowing inferences in cases of lack of knowledge in data warehouses has received much attention from researchers domestically and internationally. However, the researches mainly focus on knowledge mining in social networks or applied problems in many sectors but without the medical sector. In this study, a new approach based on the FKG-Pairs model is proposed to apply to the problem of disease diagnosis in traditional medicine. The proposed model is also applied to validate the pregnant women diagnosis and has proven the efficiency in approximate reasoning to diagnose diseases in traditional medicine.

Keywords: Knowledge graph, FKG-Pairs, approximate reasoning, disease diagnosis, traditional medicine.

I. GIỚI THIỆU

Chẩn đoán bệnh trong y học là quá trình xác định bệnh hoặc căn nguyên tạo ra các triệu chứng của người bệnh. Chẩn đoán bệnh trong y học cổ truyền là tổng hợp của các kiến thức, kỹ năng và thực hành dựa trên lý thuyết, niềm tin và kinh nghiệm bản địa của các nền văn hóa khác nhau, dù có thể giải thích hay không, được sử dụng để chẩn đoán bệnh cả về thể chất [1]. Phương pháp chẩn đoán bệnh trong y học cổ truyền được cốt lại trong bốn cách là “Vọng - Văn - Vấn - Thiết” hay còn được gọi là “Tứ chẩn” [2].

Một hướng nghiên cứu suy luận xấp xỉ dựa trên đồ thị tri thức và đồ thị tri thức mờ với ưu điểm cho phép thực hiện suy luận trong những trường hợp thiếu tri thức trong kho dữ liệu đã nhận được nhiều sự quan tâm và chú ý của các nhà nghiên cứu trong và ngoài nước. Đồ thị tri thức (*Knowledge Graph*) [3] là một kỹ thuật được sử dụng để giải quyết các bài toán phân loại, hỗ trợ ra quyết định rất hiệu quả. Tuy nhiên, đồ thị tri thức đã gặp khó khăn trong các quá trình biểu diễn đồ thị và suy luận xấp xỉ trên các tập dữ liệu đầu vào có chứa các thông tin nhập nhằng, mơ hồ, không rõ ràng. Gần đây, đồ thị tri thức mờ (*Fuzzy Knowledge Graph - FKG*) lần đầu được giới thiệu vào năm 2020 và được tích hợp trong mô hình M-CFIS-FKG [4]. Nó đã được đưa ra để giải quyết những hạn chế của đồ thị tri thức bằng cách sử dụng các nhãn ngôn ngữ cho các thuộc tính trong tập dữ liệu huấn luyện. Đầu tiên, một đồ thị tri thức mờ được xây dựng sau giai đoạn chuẩn bị dữ liệu (hay còn gọi là biểu diễn đồ thị). Sau đó, quá trình suy luận xấp xỉ được sử dụng để tìm nhãn của các bản ghi mới trong tập dữ liệu kiểm thử. Đồ thị tri thức mờ suy luận dựa trên luật IF-THEN, trong đó tác động của nhãn ngôn ngữ có khả năng tạo ra nhãn đầu ra tương ứng. Chẳng hạn, trong bài toán chẩn đoán tiền sản giật [5], việc bác sĩ đưa ra quyết định chẩn đoán là “*Bình thường*”, “*Tiền sản giật*”, hay “*Tiền sản giật nặng*” phụ thuộc vào nhãn ngôn ngữ “*Cao*”, “*Bình thường*” hay “*Thấp*” của thuộc tính “*Huyết áp*”, và nhãn ngôn ngữ “*Cao*” hay “*Bình thường*” của thuộc tính “*LDH*” (*Axit Lactic Dehydrogenase*), với độ chính xác khoảng 90%. Trong trường hợp này, chúng ta có một luật mờ được biểu diễn như sau:

IF “*Huyết áp*” là “*Cao*” AND “*LDH*” là “*Cao*” then “*Tiền sản giật nặng*”, với độ tin cậy là 90%.

Bằng cách tích lũy từng sự kiện đơn lẻ (hoặc cặp đơn trong đồ thị tri thức), kết quả sẽ xác định được nhãn của bản ghi mới. Tuy nhiên, mô hình M-CFIS-FKG [4] chỉ sử dụng các cặp đơn (*Single Pairs*) trong quá trình suy luận. Điều này dẫn đến những khó khăn trong việc áp dụng đồ thị tri thức mờ để giải quyết các bài toán phân loại, trợ giúp quyết định trong các lĩnh vực như y học, nông nghiệp, môi trường v.v.. Ngoài ra, các nghiên cứu gần đây [3, 6–11] liên quan đến đồ thị tri thức và đồ thị tri thức mờ cũng chủ yếu

tập trung vào khai phá tri thức trong mạng xã hội hay các bài toán ứng dụng trong các lĩnh vực khác mà chưa tập trung vào các bài toán trong lĩnh vực y tế. Trong nghiên cứu này, một hướng tiếp cận mới dựa trên mô hình đồ thị tri thức mờ dạng cặp (*FKG-Pairs*) được đề xuất để ứng dụng cho bài toán chẩn đoán bệnh trong y học cổ truyền. Đây là một sự kết hợp giữa y học cổ truyền và y học hiện đại. Cụ thể, nhóm tác giả đã kết hợp sử dụng các phác đồ điều trị trong y học cổ truyền cùng với các tập dữ liệu được thu thập từ kết quả đo, xét nghiệm trong y học hiện đại. Với hướng tiếp cận mới, các nhãn đầu ra được xác định bằng cách kết hợp nhiều nhãn ngôn ngữ, được gọi là các cặp thuộc tính (cặp đôi, cặp ba, ...). Những đóng góp chính của bài báo này về mặt khoa học bao gồm:

- Đề xuất mô hình mới cho bài toán chẩn đoán bệnh trong y học cổ truyền ở Việt Nam dựa trên đồ thị tri thức mờ dạng cặp (*FKG-Pairs*).
- Đề xuất cải tiến các công thức tính bộ trọng số trong mô hình M-CFIS-FKG từ cặp đơn thành dạng cặp k (k pairs). Điều này không chỉ có ý nghĩa quan trọng về mặt học thuật mà còn có ý nghĩa về mặt ứng dụng giải quyết bài toán hỗ trợ ra quyết định trong nhiều lĩnh vực khác nhau thông qua việc sử dụng đồ thị tri thức mờ dạng cặp.
- Hệ thống đề xuất cũng góp phần cải thiện độ chính xác đối với bài toán chẩn đoán bệnh tiểu đường thai phụ.

Phần còn lại của bài báo được cấu trúc như sau: Phần II trình bày các khái niệm cơ bản được sử dụng để nghiên cứu và phát triển mô hình đề xuất đối với bài toán chẩn đoán bệnh trong y học cổ truyền. Xây dựng đồ thị tri thức mờ dạng cặp cho bài toán chẩn đoán bệnh trong y học cổ truyền ở Việt Nam được đề xuất trong Phần III. Phần IV đưa ra kết quả thực nghiệm để kiểm chứng mô hình đề xuất. Phần cuối cùng đưa ra kết luận và một số hướng nghiên cứu mới.

II. KHÁI NIỆM CƠ BẢN

Trong mục này, bài báo trình bày các kiến thức cơ bản được sử dụng để nghiên cứu và phát triển mô hình đề xuất đối với bài toán chẩn đoán bệnh trong y học cổ truyền.

1. Hệ suy diễn mờ

Để áp dụng một hệ suy diễn mờ (*Fuzzy Inference System*) [9, 12–14], một số điều kiện sau cần được thỏa mãn: *Thứ nhất*, một thang đánh giá rõ ràng về các tham số của mỗi hệ suy diễn mờ phải được thống nhất. *Thứ hai*, các thuật ngữ về mức độ của các tham số này phải được đưa ra. *Cuối cùng*, mỗi thuật ngữ để xác định độ thuộc của các giá trị cũng phải được xác định. Sau khi thỏa mãn các điều kiện

này, quá trình áp dụng hệ suy diễn mờ được mô tả ngắn gọn theo một số bước sau:

- **Bước 1.** Áp dụng phương pháp mờ (*Fuzzification*) để chuyển đổi các giá trị rõ thành các giá trị mờ dựa trên các hàm thuộc (*Membership Functions*).
- **Bước 2.** Sử dụng các toán tử mờ (*AND, OR, NOT*) để đưa ra các luật biểu diễn các mệnh đề IF-THEN.
- **Bước 3.** Áp dụng công cụ suy diễn (*Inference Engine*) để tìm kết quả đầu ra.
- **Bước 4.** Kết hợp tất cả các tập mờ đại diện cho mọi đầu ra của luật để tổng hợp vào trong một tập mờ duy nhất.
- **Bước 5.** Áp dụng phương pháp giải mờ (*Defuzzification*) để chuyển đổi các giá trị mờ thành các giá trị rõ.

2. Đồ thị tri thức mờ

Như đã đề cập ở Phần I, thuật ngữ đồ thị tri thức mờ (FKG) lần đầu được giới thiệu vào năm 2020 và được tích hợp trong mô hình M-CFIS-FKG [4]. Như đã đề cập ở Phần I, thuật ngữ đồ thị tri thức mờ (FKG) lần đầu được giới thiệu vào năm 2020 và được tích hợp trong mô hình M-CFIS-FKG. Về mặt hình thức, nó cũng khá giống với một đồ thị tri thức [3, 6, 8, 11]. Các đỉnh đại diện cho các nhân của thuộc tính/nhân đầu ra của luật, và các cạnh tương ứng với cung nối giữa các đỉnh.

Trong [4], thuật toán FISA được áp dụng để suy luận và tìm ra nhân của các bản ghi mới dựa vào kết quả tính toán trọng số của các cạnh $\sum A_{ij}^t$ và $\sum B_{il}^t$ bằng cách sử dụng các công thức (1) và (2) tương ứng dưới đây.

$$A_{ij}^t = \frac{|X_i \text{ có liên quan với } X_j \text{ trong luật thứ } t|}{|R|} \quad (1)$$

trong đó $t = \overline{1, k}$, $1 \leq i \leq j \leq m$ và

$$B_{il}^t = \left(\sum A_{ij}^t \right) \times \frac{|X_i \text{ quan hệ với nhân } l \text{ trong luật thứ } t|}{|R|} \quad (2)$$

trong đó $t = \overline{1, k}$, $1 \leq i \leq j \leq m$, $l = \overline{1, C}$

3. Suy luận xấp xỉ

Suy luận xấp xỉ (Approximate Reasoning) đã được giới thiệu bởi Zadeh [15] trong bối cảnh việc thực hiện suy diễn dựa trên thông tin mơ hồ, không đầy đủ. Đây là một lý thuyết về sự không chắc chắn dựa trên logic mờ và liên quan đến việc định lượng và lập luận bằng cách sử dụng ngôn ngữ tự nhiên mà ở đó các từ có nghĩa nhập nhằng, không rõ ràng. Nó dựa trên sự tổng quát hóa của Modus

Bảng I
HỆ CƠ SỞ LUẬT MỜ PHỨC

S_j P_i	S_1	S_2	...	S_{m-1}	S_m	Kết luận
P_1	Cao	Cao	...	Rất cao	Cao	1
P_2	B.thường	B.thường	...	B.thường	Thấp	2
...	...	...	...	...	...	...
P_{n-1}	B.thường	B.thường	...	B.thường	B.thường	2
P_n	Thấp	B.thường	...	Thấp	Thấp	3
Triệu chứng của bệnh nhân	Cao B.thường Thấp	Cao B.thường	...	Rất cao Cao B.thường Thấp	Cao B.thường Thấp	1,2, ... C

Ponens cổ điển [16, 17]. Trong [18–20], để xây dựng các sự kiện và cơ sở luật cần có vai trò của chuyên gia nhằm đánh giá sự liên quan giữa các dữ liệu chuyên ngành. Tri thức của chuyên gia sẽ được biểu diễn thành các mệnh đề dưới dạng:

IF A THEN B hoặc IF A_1, A_2 THEN B_1, B_2
trong đó, A, A_1, A_2 là các mệnh đề điều kiện và B, B_1, B_2 là các kết luận.

III. XÂY DỰNG ĐỒ THỊ TRI THỨC MỜ DẠNG CẬP CHO BÀI TOÁN Y HỌC CỔ TRUYỀN Ở VIỆT NAM

1. Bài toán chẩn đoán bệnh trong y học cổ truyền

Đầu vào: Giả sử chúng ta có một cơ sở luật mờ của bài toán chẩn đoán bệnh trong y học cổ truyền sau khi áp dụng quá trình tiền xử lý đối với các phác đồ điều trị trong y học cổ truyền và tập dữ liệu kết quả đo, xét nghiệm được thu thập bởi các thầy thuốc. Cơ sở luật mờ của bài toán chẩn đoán bệnh trong y học cổ truyền thu được bao gồm n luật $P_1, P_2, \dots, P_n$ đại diện cho số lượng bệnh nhân, m thuộc tính $S_1, S_2, \dots, S_m$ đại diện cho các triệu chứng; và C nhân đầu ra 1, 2, 3, ..., C đại diện cho kết luận chẩn đoán bệnh của thầy thuốc với người bệnh được mô tả như trong Bảng I.

Ngoài ra, có thêm một bệnh nhân mới được biểu diễn như sau:

IF S_1 là “Thấp” và S_2 là “Thấp” và S_3 là “Cao” và S_4 là “Rất cao” và S_5 là “Thấp” và ... và S_{m-1} là “Cao” và S_m là “Thấp” THEN kết luận = ?

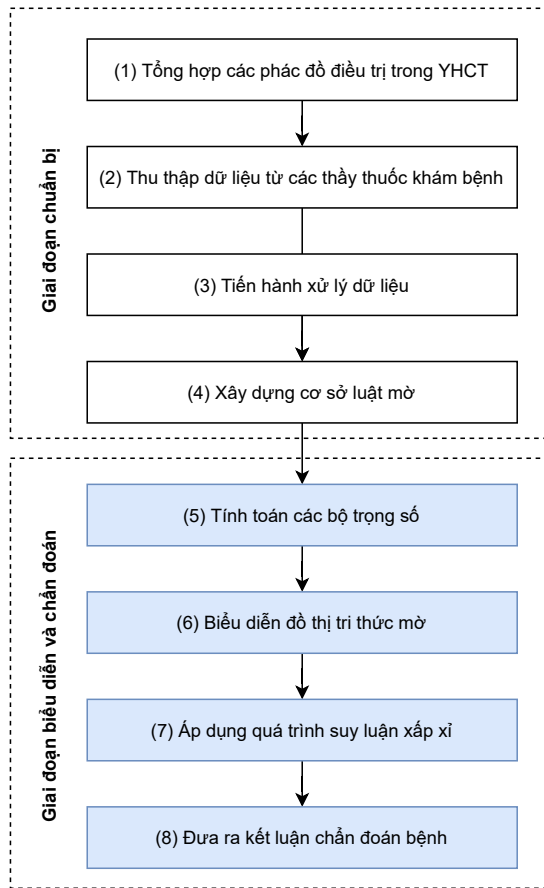
Đầu ra: Kết luận chẩn đoán bệnh của thầy thuốc đối với bệnh nhân mới dựa trên cơ sở luật mờ đã có.

2. Xây dựng đồ thị tri thức mờ dạng cặp cho bài toán chẩn đoán bệnh trong y học cổ truyền ở Việt Nam

Trong phần này, chúng tôi trình bày một số nội dung chủ yếu liên quan đến việc xây dựng đồ thị tri thức mờ dạng cặp (FKG-Pairs) cho bài toán chẩn đoán bệnh trong y học cổ truyền ở Việt Nam, cụ thể là: Đề xuất mô hình cho bài toán chẩn đoán bệnh trong y học cổ truyền ở Việt Nam; trình bày các bước thực hiện chẩn đoán bệnh theo mô hình đề xuất; và đưa ra ví dụ số minh họa cho mô hình đề xuất.

a) Mô hình đề xuất cho bài toán chẩn đoán bệnh trong y học cổ truyền ở Việt Nam

Mô hình đề xuất cho bài toán chẩn đoán bệnh trong y học cổ truyền ở Việt Nam gồm có tám bước được chia thành hai giai đoạn (giai đoạn chuẩn bị; giai đoạn biểu diễn và chẩn đoán) như Hình 1.



Hình 1. Mô hình đề xuất cho bài toán chẩn đoán bệnh trong y học cổ truyền ở Việt Nam.

Như đã nêu trong phần phát biểu bài toán, giả sử chúng ta có một cơ sở luật mờ của bài toán chẩn đoán bệnh trong y học cổ truyền sau khi áp dụng quá trình tiền xử lý đối với các phác đồ điều trị trong y học cổ truyền và tập dữ liệu kết quả đo, xét nghiệm được thu thập bởi các thầy thuốc.

Trong phạm vi nghiên cứu của bài báo này, chúng tôi tập trung vào các quá trình biểu diễn đồ thị tri thức mờ và suy luận xấp xỉ để đưa ra kết luận chẩn đoán bệnh của thầy thuốc đối với người bệnh.

b) Các bước thực hiện

Từ cơ sở luật mờ nêu tại Bảng I, một số bước dưới đây sẽ được thực hiện để tính toán các bộ trọng số, biểu diễn đồ thị tri thức mờ dạng cặp, áp dụng quá trình suy luận xấp xỉ, và đưa ra kết luận chẩn đoán bệnh.

Bước 1. Tính các bộ trọng số $\tilde{A}, \tilde{B}$

Bộ trọng số $\tilde{A}$ của đồ thị tri thức mờ dạng cặp là các trọng số các cạnh nối giữa các nhãn ngôn ngữ của các thuộc tính trong luật t (R_t). Các trọng số này được tính bởi công thức số (3) dưới đây:

$$\tilde{A}_{ij...k}^t = \frac{|S_i \rightarrow S_j \rightarrow \dots \rightarrow S_{k+1} \text{ trong luật } t|}{|R|} \quad (3)$$

trong đó $t = \overline{1, n}, 1 \leq i < j < \dots < k \leq m - 1$.

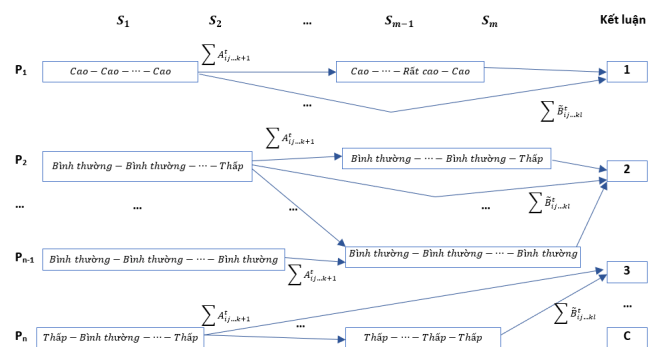
Bộ trọng số $\tilde{B}$ của đồ thị tri thức mờ dạng cặp là các trọng số các cạnh nối giữa nhãn của các cặp thuộc tính với nhãn đầu ra trong luật t (R_t). Chúng được tính bởi công thức số (4) dưới đây.

$$\tilde{B}_{ij...kl}^t = \left(\sum \tilde{A}_{ij...k+1}^t \right) \times \min \left(\frac{|S_i \rightarrow l \text{ trong luật } t|}{|R|}, \frac{|S_j \rightarrow l \text{ trong luật } t|}{|R|}, \dots, \frac{|S_k \rightarrow l \text{ trong luật } t|}{|R|} \right) \quad (4)$$

trong đó $t = \overline{1, n}, 1 \leq i < j < \dots < k \leq m - 1, l = \overline{1, C}$

Bước 2. Biểu diễn đồ thị tri thức mờ dạng cặp.

Từ cơ sở luật mờ nêu tại Bảng I, chúng tôi biểu diễn một đồ thị tri thức mờ dạng cặp như Hình 2 và lưu trữ nó trong một ma trận kề.



Hình 2. Biểu diễn đồ thị tri thức mờ dạng cặp.

Ở đây, mỗi cặp $(S_i, S_j, \dots, S_k)$ tạo một siêu đỉnh gồm các nhân ngôn ngữ của k thuộc tính kết hợp lại và được gọi là cặp k. Ví dụ, hai thuộc tính kết hợp lại gọi là cặp đôi, ba thuộc tính kết hợp thành cặp ba,...

Sau đó, chúng tôi biểu diễn các cung nối giữa các siêu đỉnh trong luật R_t , và các cung nối giữa siêu đỉnh với nhân đầu ra của luật R_t .

Bước 3. Áp dụng quá trình suy luận xấp xỉ và đưa ra kết quả chẩn đoán bệnh.

Ở bước này, chúng tôi đã cải tiến và áp dụng thuật toán FISA trong [4] để suy luận xấp xỉ cho trường hợp cặp k nhằm đưa ra kết luận chẩn đoán bệnh của các bệnh nhân mới. Đầu tiên, chúng tôi tính toán tổng trọng số của các cạnh $(\tilde{C})$ từ các siêu đỉnh đến nhân đầu ra bằng cách sử dụng công thức số (5) dưới đây.

$$\tilde{C}_{ij..kl}^t = \sum_l \tilde{B}_{ij..kl}^t \quad (5)$$

trong đó $t = \overline{1, n}, l = \overline{1, C}, 1 \leq i < j < \dots \leq k$

Tiếp theo, chúng tôi áp dụng phép toán Max-Min để tính các giá trị $(\tilde{D})$ bằng cách sử dụng công thức số (6) dưới đây.

$$\tilde{D}_l = \text{Max}_{1 \leq i < j < \dots \leq k} (\tilde{C}_{ij..kl}) + \text{Min}_{1 \leq i < j < \dots \leq k} (\tilde{C}_{ij..kl}) \quad (6)$$

Cuối cùng, nhân đầu ra của bản ghi mới (hay kết luận chẩn đoán của thầy thuốc đối với người bệnh) được đưa ra bằng cách sử dụng phép toán Max trong công thức số (7) dưới đây.

$$\text{Label} = p \text{ If } \tilde{D}_p = \text{Max}_{l=\overline{1, C}} (\tilde{D}_l) \quad (7)$$

Các bước thực hiện được trình bày chi tiết trong lưu đồ Thuật toán .

Độ phức tạp của thuật toán KFG dạng cặp được đánh giá là $O(nCm^k)$.

IV. KẾT QUẢ THỰC NGHIỆM

1. Thử nghiệm trên bộ dữ liệu mô phỏng

Giả sử rằng chúng ta nhận được một danh sách gồm sáu bệnh nhân $\{BN_1, BN_2, BN_3, BN_4, BN_5, BN_6\}$, mỗi bệnh nhân đều được ghi nhận kết quả đo, xét nghiệm và được thể hiện qua các triệu chứng $\{S_1, S_2, S_3, S_4, S_5, S_6\}$. Các bác sĩ đã xem xét và đưa ra kết luận chẩn đoán bệnh của từng người bệnh thông qua nhân ngôn ngữ *Bình thường*, *TSG*, *TSG nặng* tương ứng với các nhân 0, 1, 2. Sau giai đoạn chuẩn bị, chúng ta nhận được một cơ sở luật mờ như Bảng II.

Ngoài ra, trong giai đoạn biểu diễn và chẩn đoán bệnh dựa vào đồ thị tri thức mờ dạng cặp, nhóm tác giả nhận

Thuật toán 1: FKG dạng cặp.

```

1 Dữ liệu vào: Tập dữ liệu kiểm thử, m: Số thuộc tính của
   mỗi luật, n: Số mẫu của tập dữ liệu, C: Số nhân của
   mỗi thuộc tính.
2 Dữ liệu ra: Nhân của mẫu mới.
3 begin
4   Nhập các giá trị ;
5   Nhận tập dữ liệu kiểm thử;
6   Tiến hành mờ hóa tập dữ liệu kiểm thử ;
7   for  $i = 1$  to  $m$  do
8     for  $t = 1$  to  $n$  do
9       for  $l = 1$  to  $C$  do
10        while  $1 \leq i < j < \dots \leq k$  do
11          Tính  $\tilde{C}_{ij..kl}^t = \sum_l \tilde{B}_{ij..kl}^t$ ; Tính
             $\tilde{D}_l = \text{Max}_{1 \leq i < j < \dots \leq k} (\tilde{C}_{ij..kl}) +$ 
             $\text{Min}_{1 \leq i < j < \dots \leq k} (\tilde{C}_{ij..kl})$  ;
12        end
13        Xác định nhân của mẫu t:  $\text{Label} =$ 
             $p \text{ If } \tilde{D}_p = \text{Max}_{l=\overline{1, C}} (\tilde{D}_l)$ 
14        end
15        Nhận nhân của mẫu t và lặp lại các bước từ
            5 đến 12 để tìm nhân của các mẫu khác
            cho đến khi kết thúc.
16      end
17    end
18 end

```

Bảng II
CƠ SỞ LUẬT MỜ CỦA SÁU BỆNH NHÂN DỰA
VÀO KẾT LUẬN CHẨN ĐOÁN BỆNH CỦA BÁC SĨ

	S_1	S_2	S_3	S_4	S_5	S_6	Kết luận
BN_1	Rất cao	Bình thường	Bình thường	Bình thường	Cao	Cao	2
BN_2	Cao	Bình thường	Bình thường	Bình thường	Bình thường	Bình thường	1
BN_3	Cao	Bình thường	Bình thường	Bình thường	Cao	Cao	2
BN_4	Bình thường	Bình thường	Bình thường	Bình thường	Bình thường	Bình thường	0
BN_5	Cao	Bình thường	Bình thường	Bình thường	Bình thường	Bình thường	1
BN_6	Rất cao	Bình thường	Bình thường	Bình thường	Bình thường	Bình thường	2

được thêm hồ sơ bệnh án của bệnh nhân mới. Hồ sơ bệnh nhân mới được biểu diễn dạng luật mờ như sau:

IF S_1 là *Cao*, S_2 là *Cao*, S_3 là *Bình thường*, S_4 là *Bình thường*, S_5 là *Cao*, S_6 là *Cao* THEN Kết luận =?

Yêu cầu: Đưa ra kết luận chẩn đoán bệnh đối với bệnh nhân mới dựa trên cơ sở luật mờ đã nêu tại Bảng II mà không phải hỏi thêm ý kiến của các bác sĩ. Để giải quyết yêu cầu của bài toán là đưa ra kết luận chẩn đoán bệnh đối với bệnh nhân mới, chúng tôi đã áp dụng mô hình đề xuất như đã được trình bày tại mục B (Phần III). Dưới đây, nhóm tác giả mô tả các bước thực hiện với trường hợp FKG-Pairs với $k=3$ (cặp 3) như sau:

Bước 1: Tính các bộ trọng số $\tilde{A}, \tilde{B}$.

Trước hết, bộ trọng số $\tilde{A}$ gồm các cạnh nối giữa các nhân

Bảng III
KẾT QUẢ TÍNH BỘ TRỌNG SỐ $\tilde{A}$

$\tilde{A}$	BN_1	BN_2	BN_3	BN_4	BN_5	BN_6
$\tilde{A}_{1234}$	1/6	1/3	1/6	1/6	1/3	1/6
$\tilde{A}_{1235}$	1/6	1/3	1/6	1/6	1/3	1/6
$\tilde{A}_{1236}$	1/6	1/3	1/6	1/6	1/3	1/6
$\tilde{A}_{1245}$	1/6	1/3	1/6	1/6	1/3	1/6
$\tilde{A}_{1246}$	1/6	1/3	1/6	1/6	1/3	1/6
$\tilde{A}_{1256}$	1/6	1/3	1/6	1/6	1/3	1/6
$\tilde{A}_{1345}$	1/6	1/3	1/6	1/6	1/3	1/6
$\tilde{A}_{1346}$	1/6	1/3	1/6	1/6	1/3	1/6
$\tilde{A}_{1356}$	1/6	1/3	1/6	1/6	1/3	1/6
$\tilde{A}_{1456}$	1/6	1/3	1/6	1/6	1/3	1/6
$\tilde{A}_{2345}$	1/3	1/3	1/3	1/6	1/3	1/6
$\tilde{A}_{2346}$	1/3	1/3	1/3	1/6	1/3	1/6
$\tilde{A}_{2356}$	1/3	1/3	1/3	1/6	1/3	1/6
$\tilde{A}_{2456}$	1/3	1/3	1/3	1/6	1/3	1/6
$\tilde{A}_{3456}$	1/3	1/3	1/3	1/6	1/3	1/6

Bảng IV
KẾT QUẢ TÍNH BỘ TRỌNG SỐ $\tilde{B}$

$\tilde{B}$	BN_1	BN_2	BN_3	BN_4	BN_5	BN_6
$\tilde{B}_{1234}$	10/9	5/3	5/9	5/12	5/3	5/12
$\tilde{B}_{1245}$	10/9	5/3	5/9	5/12	5/3	5/12
$\tilde{B}_{1256}$	10/9	5/3	5/9	5/12	5/3	5/12
$\tilde{B}_{1264}$	10/9	5/3	5/9	5/12	5/3	5/12
$\tilde{B}_{1345}$	10/9	5/3	5/9	5/12	5/3	5/12
$\tilde{B}_{1356}$	10/9	5/3	5/9	5/12	5/3	5/12
$\tilde{B}_{1364}$	10/9	5/3	5/9	5/12	5/3	5/12
$\tilde{B}_{1456}$	10/9	5/3	5/9	5/12	5/3	5/12
$\tilde{B}_{1464}$	10/9	5/3	5/9	5/12	5/3	5/12
$\tilde{B}_{1564}$	10/9	5/3	5/9	5/12	5/3	5/12
$\tilde{B}_{2345}$	10/9	5/3	10/9	5/12	5/3	5/12
$\tilde{B}_{2356}$	10/9	5/3	10/9	5/12	5/3	5/12
$\tilde{B}_{2364}$	10/9	5/3	10/9	5/12	5/3	5/12
$\tilde{B}_{2456}$	10/9	5/3	10/9	5/12	5/3	5/12
$\tilde{B}_{2464}$	10/9	5/3	10/9	5/12	5/3	5/12
$\tilde{B}_{2564}$	10/9	5/3	10/9	5/12	5/3	5/12
$\tilde{B}_{3456}$	10/9	5/3	10/9	5/12	5/3	5/12
$\tilde{B}_{3464}$	10/9	5/3	10/9	5/12	5/3	5/12
$\tilde{B}_{3564}$	10/9	5/3	10/9	5/12	5/3	5/12
$\tilde{B}_{4564}$	10/9	5/3	10/9	5/12	5/3	5/12

ngôn ngữ của các triệu chứng trong hồ sơ bệnh nhân t (P_t) được tính bởi công thức số (3).

Chẳng hạn, xét trong BN_1 trọng số $\tilde{A}$ được tính như sau:

$$\begin{aligned}\tilde{A}_{1234}^1 &= \frac{[Rất\ cao \rightarrow B.thường \rightarrow B.thường \rightarrow Cao]}{6} = \frac{1}{6}, \\ \tilde{A}_{1235}^1 &= \frac{[Rất\ cao \rightarrow B.thường \rightarrow B.thường \rightarrow Cao]}{6} = \frac{1}{6}, \\ \tilde{A}_{1236}^1 &= \frac{[Rất\ cao \rightarrow B.thường \rightarrow B.thường \rightarrow B.thường]}{6} = \frac{1}{6}\end{aligned}$$

Tương tự, ta có:

$$\begin{aligned}\tilde{A}_{1245}^1 &= \tilde{A}_{1246}^1 = \tilde{A}_{1256}^1 = \tilde{A}_{1345}^1 = \tilde{A}_{1346}^1 = \tilde{A}_{1356}^1 = \\ \tilde{A}_{1456}^1 &= \frac{1}{6} \\ \tilde{A}_{2345}^1 &= \tilde{A}_{2346}^1 = \tilde{A}_{2356}^1 = \tilde{A}_{3456}^1 = \frac{1}{3}\end{aligned}$$

Bằng cách áp dụng công thức số (3) tính tương tự, ta có kết quả tính toán bộ trọng số $\tilde{A}$ được hiển thị trong Bảng III. Tiếp theo, chúng tôi áp dụng công thức số (4) để tìm bộ trọng số các cạnh nối giữa nhân của các cặp thuộc tính với nhân đầu ra kết luận chẩn đoán bệnh nhân t (P_t). Kết quả tính toán bộ trọng số $\tilde{B}$ được hiển thị ở Bảng IV

Bước 2. Biểu diễn đồ thị tri thức mờ dạng cặp.

Ở bước này, một đồ thị tri thức mờ dạng cặp 3 được biểu diễn từ cơ sở luật mờ ở Bảng II và các bộ trọng số $\tilde{A}$, $\tilde{B}$ đã tính ở Bước 1. Hình 3 minh họa đồ thị tri thức mờ dạng cặp 3 với trọng số.

Bước 3. Áp dụng quá trình suy luận xấp xỉ và đưa ra kết quả chẩn đoán bệnh.

Dựa vào kết quả tính bộ trọng số $\tilde{B}$, chúng tôi đưa ra kết luận chẩn đoán bệnh của bệnh nhân mới dựa trên cơ sở luật mờ đã nêu tại Bảng II. Đầu tiên, các giá trị ($\tilde{C}$) và ($\tilde{D}$) của hồ sơ bệnh của bệnh nhân mới được tính toán, suy luận xấp xỉ bằng cách sử dụng công thức số (5) ta tính được các giá trị ($\tilde{C}$):

$$\begin{aligned}\tilde{C}_{1230} &= \tilde{C}_{1240} = \tilde{C}_{1250} = \tilde{C}_{1260} = \tilde{C}_{1340} = \tilde{C}_{1350} = \tilde{C}_{1360} = \\ \tilde{C}_{1450} &= \tilde{C}_{1460} = \tilde{C}_{1560} = \tilde{C}_{2340} = \tilde{C}_{2350} = \tilde{C}_{2360} = \tilde{C}_{2450} = \\ \tilde{C}_{2460} &= \tilde{C}_{2560} = \tilde{C}_{3450} = \tilde{C}_{3460} = \tilde{C}_{3560} = \tilde{C}_{4560} = 0.\end{aligned}$$

$$\begin{aligned}\tilde{C}_{1231} &= \tilde{C}_{1241} = \tilde{C}_{1251} = \tilde{C}_{1261} = \tilde{C}_{1341} = \tilde{C}_{1351} = \tilde{C}_{1361} = \\ \tilde{C}_{1451} &= \tilde{C}_{1461} = \tilde{C}_{1561} = \tilde{C}_{2341} = \tilde{C}_{2351} = \tilde{C}_{2361} = \tilde{C}_{2451} = \\ \tilde{C}_{2461} &= \tilde{C}_{2561} = \tilde{C}_{3451} = \tilde{C}_{3461} = \tilde{C}_{3561} = \tilde{C}_{4561} = 0.\end{aligned}$$

$$\begin{aligned}\tilde{C}_{1232} &= \tilde{C}_{1242} = \tilde{C}_{1252} = \tilde{C}_{1262} = \tilde{C}_{2342} = \tilde{C}_{2352} = \tilde{C}_{2362} = \\ \tilde{C}_{2452} &= \tilde{C}_{2462} = \tilde{C}_{2562} = 0.\end{aligned}$$

$$\tilde{C}_{1342} = \tilde{C}_{1352} = \tilde{C}_{1362} = \tilde{C}_{1452} = \tilde{C}_{1462} = \tilde{C}_{1562} = 5/9$$

$$\tilde{C}_{3452} = \tilde{C}_{3462} = \tilde{C}_{3562} = \tilde{C}_{4562} = 10/9$$

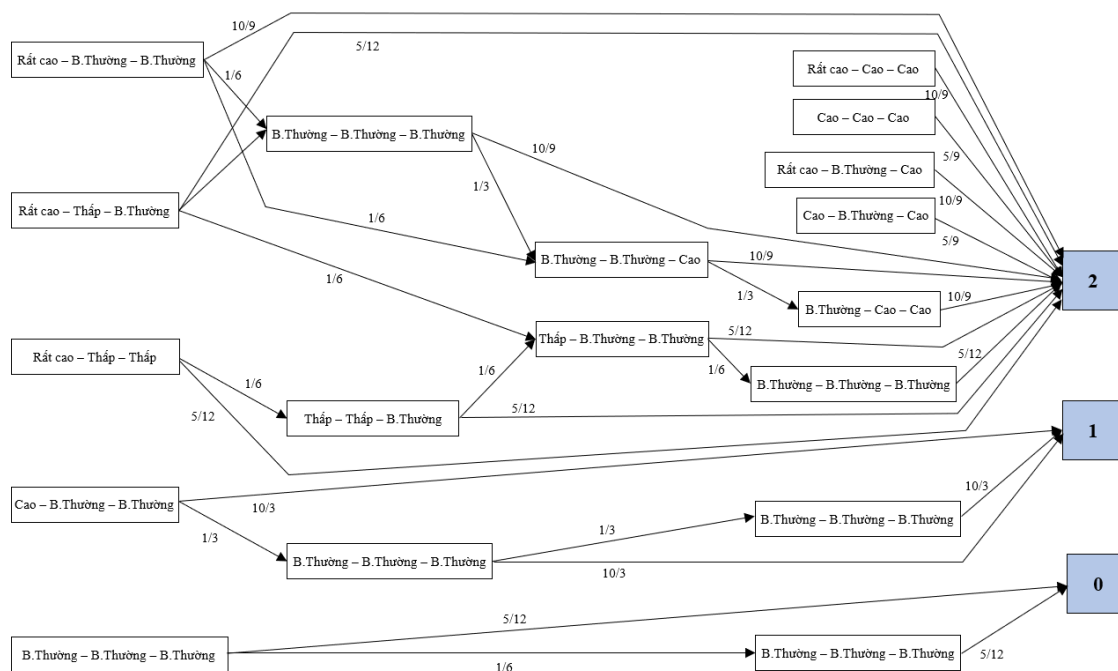
Giá trị ($\tilde{D}$) được tính bằng cách áp dụng công thức (6), ta có: $\tilde{D}_0 = \tilde{D}_1 = 0, \tilde{D}_2 = 10/9$. Do đó, sử dụng phép toán Max trong công thức số (7) ta thu được nhân đầu ra của hồ sơ bệnh nhân là 2 vì $\tilde{D}_2$ lớn nhất (nghĩa là bệnh nhân được chẩn đoán bị tiền sản giật nặng).

Tóm lại, qua kết quả ví dụ thực nghiệm trên dữ liệu mô phỏng, có thể thấy rằng mô hình đề xuất áp dụng cho bài toán chẩn đoán bệnh trong y học cổ truyền ở Việt Nam hoàn toàn có thể tin cậy. Điều này khẳng định rằng mô hình đề xuất trong bài báo này có thể tiếp tục nghiên cứu để ứng dụng chẩn đoán bệnh bằng cách kết hợp giữa y học cổ truyền với y học hiện đại tại các bệnh viện, cơ sở khám bệnh, đặc biệt tại các bệnh viện y học cổ truyền và cơ sở khám bệnh ở địa phương.

2. Thực nghiệm trên bộ dữ liệu thực tế

a) Kích bản thực hiện

Trong phần này, nhóm tác giả tiến hành thực nghiệm với bộ dữ liệu về các sản phụ được bác sĩ theo dõi và chỉ định làm xét nghiệm để chẩn đoán liệu sản phụ có bị tiền sản giật hay không? Từ đó bác sĩ có thể đưa ra chỉ định yêu cầu theo dõi chặt chẽ đối với sản phụ trong suốt thời gian



Hình 3. Biểu diễn đồ thị tri thức mở dạng cặp.

còn lại của thai kỳ để hạn chế tối đa biến chứng đối với sản phụ và thai nhi. Tập dữ liệu (BN-TienSanGiat) được sử dụng để thực nghiệm gồm 199 sản phụ đến khám định kỳ theo dõi thai nhi với 19 thông số kiểm tra, theo dõi được trình bày chi tiết trong Bảng V (trong đó, theo kết luận chẩn đoán có 114 sản phụ thai bình thường, 57 sản phụ có dấu hiệu tiền sản giật và 28 sản phụ bị tiền sản giật nặng).

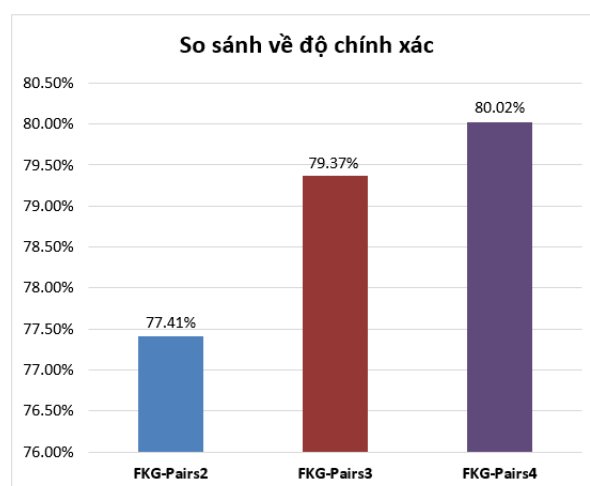
Để tiến hành thực nghiệm, chúng tôi sử dụng ngôn ngữ Matlab 2014 đã được cài đặt trên máy tính xách tay (DELL Intel(R) Core (TM) i5-7200U CPU @ 2,50GHz 2,70 GHz) và thực nghiệm theo phương pháp k-fold (với $k = 5$) trên bộ dữ liệu BN-TienSanGiat. Lý do chúng tôi lựa chọn chia bộ dữ liệu theo phương pháp k-fold là vì bộ dữ liệu BN-TienSanGiat tương đối nhỏ (199 bản ghi). Mục tiêu của thực nghiệm là nhằm kiểm chứng khả năng áp dụng đồ thị tri thức mở dạng cặp để hỗ trợ chẩn đoán dấu hiệu tiền sản giật của các sản phụ trong y học cổ truyền.

b) Kết quả thực hiện

Sau khi tiến hành thực hiện, kết quả về độ chính xác (Accuracy) và thời gian tính toán (Computational Time) của các phương pháp FKG-Pairs khác nhau (bao gồm FKG-Pairs2, FKG-Pairs3, và FKG-Pairs4) trong việc chẩn đoán bệnh tiền sản giật đối với sản phụ được trình bày trong Bảng VI.

Hình 4 chứng tỏ rằng phương pháp FKG-Pairs4 tốt nhất về mặt độ chính xác (80.02%), nhưng thời gian cao gấp khoảng bốn lần so FKG-Pairs2 và gấp đôi so với FKG-

Pairs3. Trong khi đó, Hình 5 cho thấy rằng phương pháp FKG-Pairs2 có thời gian tính toán thấp nhất (*chỉ 0,0156s*) nhưng độ chính xác lại kém so với FKG-Pairs3 và FKG-Pairs4 lần lượt là 1,96% và 2,61%. Điều này có thể nhận thấy rằng, suy luận và chẩn đoán bệnh dựa trên đồ thị tri thức mở dạng cặp (FKG-Pairs) về mặt tham số độ chính xác đã được cải thiện một chút khi tăng cặp k , trong khi tham số thời gian tính toán lại tăng rất nhanh.



Hình 4. Biểu đồ so sánh độ chính xác của các phương pháp.

Tóm lại, kết quả thực nghiệm đã chứng minh rằng khả năng áp dụng đồ thị tri thức mở dạng cặp để hỗ trợ chẩn

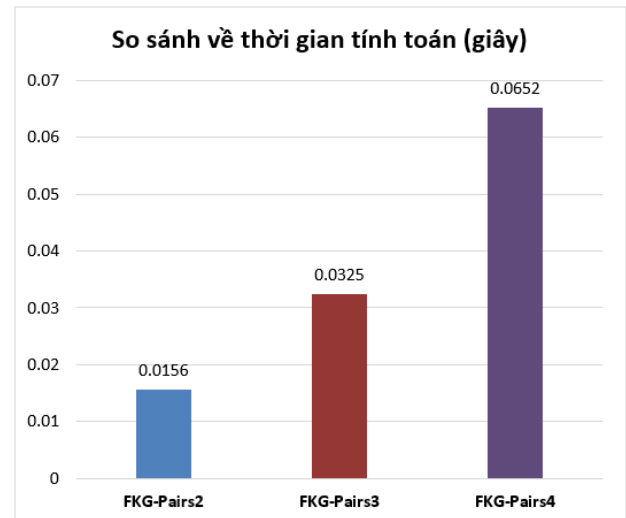
Bảng V
DANH SÁCH CÁC THUỘC TÍNH TRONG
TẬP DỮ LIỆU BỆNH NHÂN TIỀN SẢN GIẬT

STT	Tên thuộc tính	Miền giá trị
1	Tuổi của sản phụ	18 - 66
2	Nghề nghiệp của sản phụ	Công nhân, Nông dân CCVC, Tự do,...
3	Tuổi của thai nhi	15 - 40
4	Số lần mang thai	0 - 9
5	Chiều cao của sản phụ	1,40m - 1,90m
6	Cân nặng của sản phụ	45 - 95
7	Huyết áp trên	90 - 129
8	Huyết áp dưới	60 - 84
9	Chỉ số HGB	120 - 160
10	Chỉ số PLT	150 - 450
11	Chỉ số URE	2,5 - 6,7
12	Chỉ số CREATEIN	50,4 - 98,1
13	Chỉ số ACID URIC	150 - 350
14	Chỉ số ALT	< 31/37°C
15	Chỉ số AST	< 31/37°C
16	Chỉ số PROTEIN toàn phần	64 - 83
17	Chỉ số ALBUMIN	35 - 52
18	Chỉ số LDH	< 247
19	Chỉ số PROTEIN niệu	0,1 - 0,25
Kết quả đầu ra	Nhân (kết luận chẩn đoán)	0: Bình thường 1: Tiền sản giật 2: Tiền sản giật nặng

Bảng VI
KẾT QUẢ SO SÁNH ĐỘ CHÍNH XÁC VÀ THỜI
GIAN THỰC HIỆN CỦA CÁC PHƯƠNG PHÁP

Phương pháp	Độ chính xác (%)	Thời gian (giây)
FKG-Pairs2	77,41%	0,0156
FKG-Pairs3	79,37%	0,0325
FKG-Pairs4	80,02%	0,0652

đoán đầu hiệu tiền sản giật của các sản phụ trong y học cổ truyền là hoàn toàn khả thi. Độ chính xác của phương pháp cặp tư (FKG-Pairs4) đạt 80,02%, cao hơn một chút so với các phương pháp cặp đôi (FKG-Pairs2) và cặp ba (FKG-Pairs3) lần lượt là 2,61% và 1,96%. Trong khi thời gian tính toán của phương pháp FKG-Pairs4 cao hơn gấp gần 4,18 lần so với FKG-Pairs2 (và 2,01 lần với FKG-Pairs3).



Hình 5. Biểu đồ so sánh thời gian tính toán của các phương pháp.

V. KẾT LUẬN

Bài báo này đã trình bày một phương pháp khám và chẩn đoán bệnh hoàn toàn mới dựa trên đồ thị tri thức mờ dạng cặp (*FKG-Pairs*). Một mô hình áp dụng cho bài toán chẩn đoán bệnh trong y học cổ truyền ở Việt Nam dựa trên đồ thị tri thức mờ dạng cặp đã được đề xuất gồm với tám bước chia thành hai giai đoạn: (i) chuẩn bị dữ liệu; (ii) biểu diễn đồ thị và chẩn đoán bệnh. Đặc biệt, nhóm tác giả đã cải tiến các công thức tính bộ trọng số trong mô hình M-CFIS-FKG từ cặp đơn thành dạng cặp k . Điều này góp phần thúc đẩy ứng dụng FKG-Pairs vào trong các bài toán khám và chẩn đoán bệnh trong y học cổ truyền. Ngoài ra, với cách tiếp cận sử dụng *FKG-Pairs* cũng góp phần làm tăng độ chính xác trong quá trình suy luận để tìm nhân đầu ra. Một số hướng nghiên cứu tiếp theo liên quan đến FKG-Pairs được khuyến nghị, cụ thể là: (1) xác định giá trị ngưỡng k phù hợp (ở đó $k = \overline{1, m}$ để có thể áp dụng FKG-Pairs cho các bài toán trợ giúp ra quyết định dựa trên tập dữ liệu đầu vào có nhiều thuộc tính; (2) tìm giá trị hợp lý nhất k^* (ở đó $k^* = \overline{1, k}$) để ứng dụng trong các bài toán trợ giúp ra quyết định dựa trên tập dữ liệu lớn góp phần làm tăng hiệu quả về mặt thời gian tính toán (*computational time*); (3) nghiên cứu tích hợp học tăng cường (*reinforcement learning*) vào đồ thị tri thức mờ để xử lý bài toán trợ giúp ra quyết định trong các trường hợp cực đoan (*extreme cases*).

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Quỹ Phát triển khoa học và công nghệ Quốc gia (NAFOSTED) trong đề tài mã số 102.05-2019.316.

TÀI LIỆU THAM KHẢO

- [1] W. H. Organization *et al.*, *WHO traditional medicine strategy: 2014-2023*. World Health Organization, 2013.
- [2] T. chí y học cổ truyền Viện Y Dược cổ truyền dân tộc, “Y học cổ truyền là gì? Nguyên tắc chữa bệnh và địa chỉ khám, chữa uy tín [đăng tải ngày 23/6/2020]”, tại địa chỉ: <https://www.tapchihoccotruyen.com/y-hoc-co-truyen.html>, [truy cập ngày 20/5/2021].
- [3] P. A. Bonatti, S. Decker, A. Polleres, and V. Presutti, “Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371),” in *Dagstuhl Reports*, vol. 8, no. 9. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.
- [4] L. T. H. Lan, T. M. Tuan, T. T. Ngan, N. L. Giang, V. T. N. Ngoc, P. Van Hai *et al.*, “A new complex fuzzy inference system with fuzzy knowledge graph and extensions in decision making,” *IEEE Access*, vol. 8, pp. 164 899–164 921, 2020.
- [5] Đảng Cộng Việt, *Nghiên cứu chuẩn đoán và xử trí tiền sản giật tại Bệnh viện Phụ sản Trung ương*, Luận văn thạc sĩ, 2020, Đại học Y Hà Nội.
- [6] H. Van Pham and L. K. Cu, “Intelligent rule-based support model using log files in big data for optimized service call center schedule,” in *Intelligent Computing in Engineering*. Springer, 2020, pp. 931–942.
- [7] S. Meier, B. Gebel-Sauer, and P. Schubert, “Knowledge graph for the visualisation of crm objects in a social network of business objects (sonbo): Development of the sonbo visualiser,” *Procedia Computer Science*, vol. 181, pp. 448–456, 2021.
- [8] C. K. Long, H. Q. Trung, T. N. Thang, N. T. Dong, and P. Van Hai, “A knowledge graph approach for the detection of digital human profiles in big data,” *Journal of Science and Technology: Issue on Information and Communications Technology*, vol. 19, no. 6.2, pp. 6–15, 2021.
- [9] T. T. Ngan, L. T. H. Lan, T. M. Tuan, N. H. Minh *et al.*, “Colorectal cancer diagnosis with complex fuzzy inference system,” in *Frontiers in Intelligent Computing: Theory and Applications*. Springer, 2020, pp. 11–20.
- [10] J. Liu, F. Schmid, K. Li, and W. Zheng, “A knowledge graph-based approach for exploring railway operational accidents,” *Reliability Engineering & System Safety*, vol. 207, p. 107352, 2021.
- [11] X. Wang and D. Calvanese, “Editorial for special issue of journal of big data research on “big data meets knowledge graphs,”” 2021.
- [12] L. Zadeh, “Fuzzy sets, information and control,” vol. 8, no. 3, pp. 338–353, 1965.
- [13] L. C. Ortega, L. D. Otero, and C. Otero, “Fuzzy inference system framework to prioritize the deployment of resources in low visibility traffic conditions,” *IEEE Access*, vol. 7, pp. 174 368–174 379, 2019.
- [14] T. M. Tuan, L. T. H. Lan, S.-Y. Chou, T. T. Ngan, L. H. Son, N. L. Giang, M. Ali *et al.*, “M-cfis-r: Mamdani complex fuzzy inference system with rule reduction using complex fuzzy measures in granular computing,” *Mathematics*, vol. 8, no. 5, p. 707, 2020.
- [15] L. A. Zadeh, “Approximate reasoning based on fuzzy logic,” in *Proceedings of the 6th international joint conference on Artificial intelligence-Volume 2*, 1979, pp. 1004–1010.
- [16] S. Moussa and S. B. H. Kacem, “Symbolic approximate reasoning with fuzzy and multi-valued knowledge,” *Procedia computer science*, vol. 112, pp. 800–810, 2017.
- [17] T. T. Sơn, “Lập luận xấp xỉ với giá trị của biến ngôn ngữ,” *Tạp chí Tin học và Điều khiển*, vol. 15, pp. 6–10, 1999.
- [18] N. Đức Thọ và cộng sự, “Ứng dụng Đồ thị tri thức, tri thức suy diễn và học máy trong chuyển đổi số,” pp. 453–458, 2020.
- [19] S. Jin, J. Peng, Z. Li, and Q. Shen, “Bidirectional approximate reasoning-based approach for decision support,” *Information Sciences*, vol. 506, pp. 99–112, 2020.
- [20] J. Xiong and J. Wu, “Construction of approximate reasoning model for dynamic cps network and system parameter identification,” *Computer Communications*, vol. 154, pp. 180–187, 2020.

SƠ LƯỢC VỀ TÁC GIẢ

Cù Kim Long



Nhận bằng Thạc sĩ Công nghệ thông tin năm 2012 tại Đại học Quốc gia Hà Nội (VNU). Nghiên cứu sinh từ năm 2018 tại Đại học Bách khoa Hà Nội.

Hiện là Phó Giám đốc Trung tâm Công nghệ thông tin - Bộ Khoa học và Công nghệ, thành viên Ban kỹ thuật TCVN/JTC1 “Công nghệ thông tin”.

Lĩnh vực nghiên cứu: OPS, MIS, Enterprise Architecture, Cyber Security, e-Government, Industry 4.0, Smart City, IoT, Big Data, Trí tuệ nhân tạo, Evaluation, DSS, Picture Fuzzy Set, Knowledge Graph, Fuzzy Knowledge Graph, Reinforcement Learning, ...

Email: longck.2006@gmail.com

Trần Mạnh Tuấn



Nhận bằng thạc sĩ Công nghệ thông tin tại trường Đại học Thái Nguyên và bằng tiến sĩ Cơ sở toán học cho tin học tại Học viện Khoa học Công nghệ, Viện hàn lâm Khoa học và Công nghệ Việt Nam năm 2017.

Hiện là Phó trưởng bộ môn Hệ thống thông tin, Khoa Công nghệ thông tin, Trường Đại học Thủy lợi.

Lĩnh vực nghiên cứu: ứng dụng trí tuệ nhân tạo, học máy, dữ liệu lớn, logics mờ.

Email: tmtuan@tlu.edu.vn

Lê Hoàng Sơn



Nhận bằng tiến sĩ Toán - Tin tại trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội (ĐHQGHN) năm 2013. Được phong hàm Phó Giáo sư ngành Công nghệ Thông tin năm 2017.

Hiện là Trưởng phòng Đa phương tiện và Thực tại ảo, Viện Công nghệ Thông tin, ĐHQGHN.

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, khai phá dữ liệu, tính toán mềm, hệ tư vấn và hệ thống thông tin địa lý, ứng dụng trong y tế và mạng viễn thông. Một số ứng dụng đã triển khai gần đây: Hệ thống hỗ trợ chẩn đoán bệnh từ ảnh y tế với mô hình học sâu và đồ thị tri thức, các hệ thống bán giám sát trong phát hiện đối tượng trên ảnh viễn thám, hệ tư vấn lớn (Mega RS) với các kỹ thuật trích rút thông tin trên bệnh án điện tử, các phần mềm phân tích biểu cảm khuôn mặt, và hỗ trợ ra quyết định động.

Email: sonlh@vnu.edu.vn

Lương Thị Hồng Lan



Nhận bằng thạc sĩ về SPKT Tin học tại trường Đại học Bách Khoa Hà nội; bảo vệ luận án tiến sĩ chuyên ngành Khoa học máy tính tại Học viện Khoa học Công nghệ, Viện hàn lâm Khoa học và Công nghệ Việt Nam năm 2021.

Hiện là giảng viên bộ môn Hệ thống thông tin, khoa Công nghệ thông tin, trường Đại học Thủy lợi.

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, tính toán mềm, hệ mờ, logic mờ, khai phá dữ liệu.
Email: lanlhbk@tlu.edu.vn

Phạm Minh Chuẩn



Nhận bằng thạc sĩ Toán tin ứng dụng và bằng tiến sĩ Hệ thống thông tin tại trường Đại học Bách Khoa Hà nội năm 2018.

Hiện là Trưởng bộ môn Khoa học máy tính, khoa Công nghệ thông tin, trường Đại học Sư phạm Kỹ thuật Hưng Yên.

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, khai phá dữ liệu, học máy, hệ tư vấn.
Email: chuanpm@gmail.com

Nguyễn Thọ Thông



Nhận bằng tiến sĩ chuyên ngành Hệ thống thông tin tại trường Đại học Công nghệ, Đại học quốc gia Hà Nội (ĐHQGHN) năm 2021.

Hiện là giảng viên của khoa Công nghệ thông tin, trường Đại học Thủy Lợi.

Lĩnh vực nghiên cứu: Ra quyết định đa tiêu chí, hệ thống thông tin, hệ tư vấn và trí tuệ nhân tạo.

Email: thongnt89@tlu.edu.vn

Phạm Văn Hải



Nhận bằng tiến sĩ Công nghệ thông tin tại trường Đại học Ritsumeikan Nhật Bản năm 2013. Được phong hàm Phó Giáo sư năm 2017.

Hiện là giảng viên cao cấp trường Công nghệ Thông tin - Truyền thông, Đại học Bách khoa Hà Nội.

Lĩnh vực nghiên cứu: Ứng dụng trí tuệ nhân tạo (chuyển đổi số, cách mạng công nghiệp 4.0), học máy, tính toán mềm, phân

tích dữ liệu lớn, ứng dụng hệ mờ, logics, hệ thống thông minh lai, tích hợp hệ thống, Big Data; ...

Email: haipy@soict.hust.edu.vn

Xây dựng hệ thống nhận dạng hành động sử dụng cảm biến gia tốc và các phương pháp học máy trên vi điều khiển hiệu năng thấp

Đào Tô Hiệu^{1,2}, Ngô Văn Công^{2,3}, Nguyễn Quang Huy², Trần Đức Nghĩa², Trần Đức Tân^{4,*}

¹ Trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên, Thái Nguyên

² Học Viện Khoa học và Công nghệ (VAST), Hà Nội

³ Tổng cục Tiêu chuẩn Đo lường Chất lượng, Hà Nội

⁴ Trường Đại học Phenikaa, Hà Nội

Tác giả liên hệ: Trần Đức Tân, tan.tranduc@phenikaa-uni.edu.vn

Ngày nhận bài: 19/08/2021, ngày sửa chữa: 15/11/2021, ngày duyệt đăng: 21/11/2021

Định danh DOI: 10.32913/mic-ict-research-vn.v2021.n2.1004

Tóm tắt: Hệ thống ghi nhận hoạt động của con người sử dụng cảm biến giúp hỗ trợ đánh giá tình trạng sức khỏe. Yêu cầu đặt ra là hiệu suất hoạt động cao, giá cả rẻ và hoạt động theo thời gian thực ngay trên các vi điều khiển có hiệu năng thấp. Từ dữ liệu thu thập của các hành động cần trích xuất tính chất đặc trưng điển hình. Những đặc trưng này được huấn luyện để phân loại hành động. Một thiết bị gắn trên eo người được lập trình kết nối hệ thống thông qua WiFi và cho phép phân loại hành động người theo thời gian thực, đồng thời người dùng có thể giám sát hoạt động và quản lý dữ liệu trên máy tính hoặc điện thoại thông minh.

Từ khóa: Phân loại, hành động, học máy, web server, cảm biến gia tốc, ứng dụng, thu thập dữ liệu.

Title: Building Human Activity Recognition System using Accelerometers and Machine Learning Methods on Low-Performance Microcontrollers

Abstract: The system recognizes human activities using sensors to help assess health status. The requirements are high performance, low price and real-time operation even on low-performance microcontrollers. From the collected data of the activities, it is necessary to extract the typical feature. These features are trained to classify activities. A programmed waist-mounted device connects to the system via WiFi and allows real-time classification of human activities, and users can monitor activities and manage data on smart devices.

Keywords: Classification, activity, machine learning, web server, acceleration sensor, application, data collection.

I. GIỚI THIỆU

Trong vài thập kỷ qua, những nghiên cứu về hành động của con người phát triển rất mạnh mẽ và có nhiều ứng dụng vào cuộc sống. Những nhu cầu ứng dụng dựa trên phân loại hành động như chăm sóc sức khỏe [1], [2], phát hiện tai nạn cho người già [3], [4], công nghiệp [5], an ninh [6], ... Mô hình phân loại hành động cần những thông tin về hành vi của người và các hoạt động tương tác, từ đó cho phép xây dựng hệ thống hỗ trợ người dùng một cách chủ động trong công việc [7].

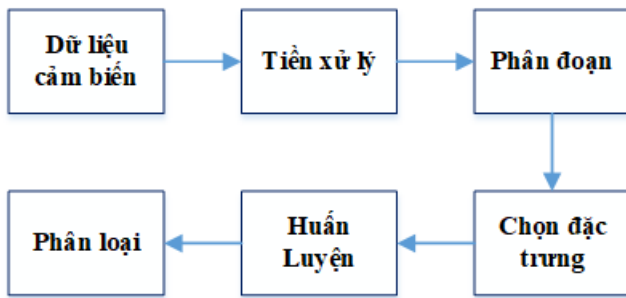
Với cách tiếp cận dựa trên cảm biến mang trên người đã mở ra nhiều ứng dụng tiềm năng trong phân loại hành động [8-13]. Cách tiếp cận này không cần thiết lập trước các điều kiện và không phụ thuộc vào không gian, thời gian

sử dụng. Do đó, nhóm nghiên cứu hướng đến ứng dụng kỹ thuật phân loại hành vi người để hướng đến xây dựng một hệ thống chăm sóc khỏe cho người già, người bệnh mới hồi phục sau tai nạn, ... thông qua hệ thống theo dõi hành vi được gắn trực tiếp lên người dùng. Bởi vậy, nghiên cứu này sẽ tập trung vào việc khai thác khả năng phân loại theo thời gian thực các hành động cơ bản của con người giúp cho các bác sĩ có thể giám sát hoạt động của người bệnh và từ đó có cơ sở dữ liệu để chẩn đoán, hỗ trợ chăm sóc sức khỏe từ xa. Bên cạnh đó, nhóm nghiên cứu đề xuất sử dụng phân đoạn dữ liệu theo thời gian (cửa sổ trượt) có chiều rộng cố định n-giây. Các đặc trưng sẽ được trích xuất trên mỗi cửa sổ trượt (Sliding Window) và dùng làm dữ liệu đầu vào cho quá trình huấn luyện, phân loại hành động. Yêu cầu đặt ra với hệ thống là hiệu suất hoạt động

cao, giá rẻ và hoạt động theo thời gian thực ngay trên các vi điều khiển có hiệu năng thấp. Để thực hiện được mục tiêu như vậy, thiết bị cần được thiết kế gọn nhẹ, tất cả các bước gồm: 1/ thu thập dữ liệu, 2/ tiền xử lý, 3/ trích chọn đặc trưng, 4/ phân loại trực tiếp trên vi điều khiển, 5/ truyền không dây tới server phải thật gọn nhẹ và hiệu quả để có thể đáp ứng yêu cầu thời gian thực ngay trên các dòng vi điều khiển hiệu năng thấp (MCU ESP8266).

II. PHƯƠNG PHÁP NGHIÊN CỨU

Thiết bị giám sát được gắn trực tiếp trên eo người và thu thập dữ liệu gia tốc theo ba trục Ax, Ay và Az. Dữ liệu được phân đoạn theo phương pháp cửa sổ trượt. Sau đó, mỗi vectơ đặc trưng được trích xuất từ mỗi cửa sổ dữ liệu và được huấn luyện bởi thuật toán học máy là cây quyết định [14] để nâng cao hiệu suất phân loại. Để xây dựng mô hình phân loại hành động, bài báo nghiên cứu 2 đặc trưng của dữ liệu gia tốc: giá trị trung bình (mean) và độ lệch chuẩn (SD). Cấu trúc mô hình phân loại được thể hiện trong Hình 1.



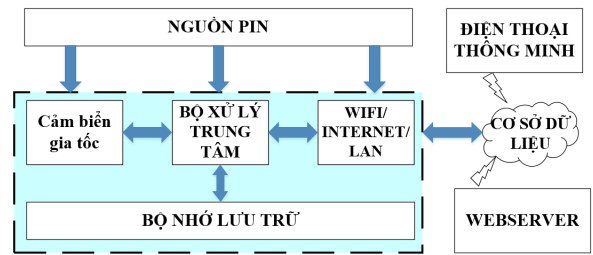
Hình 1. Cấu trúc mô hình phân loại hành động

1. Hệ thống thu thập dữ liệu

Cảm biến gia tốc được kết nối với vi điều khiển ESP8266 thông qua giao tiếp I2C để thu nhận dữ liệu gia tốc theo ba trục Ax, Ay, Az. Dữ liệu này được ghi log trên thẻ nhớ (bộ nhớ ngoài), đồng thời được truyền về Server thông qua WIFI (INTERNET/LAN) theo giao thức HTTP (Hypertext Transfer Protocol) và phương thức GET. Trong thiết kế này, thiết bị thu thập sử dụng pin lithium 3.7V-850mAh. Cấu trúc hệ thống được minh họa trong Hình 2.

2. Dữ liệu

Thiết bị được đeo trên người như trong Hình 3. Sau khi thu thập, dữ liệu này được tiền xử lý để loại bỏ các tín hiệu bị mất trong quá trình truyền, loại bỏ nhiễu và được gán nhãn theo năm lớp, bao gồm các hành động: nằm, ngồi, đứng, đi bộ và chạy bộ (Bảng I). Tần số lấy mẫu dữ liệu cảm biến được sử dụng là 1Hz.



Hình 2. Sơ đồ khối hệ thống phân loại hành vi

Bảng I
ĐỊNH NGHĨA CÁC HÀNH ĐỘNG

HÀNH ĐỘNG	ĐỊNH NGHĨA
Nằm	Toàn thân người trên mặt nền trong tư thế nghỉ.
Ngồi	Đầu và lưng không chạm nền và đôi chân nghỉ.
Đứng	Người thẳng, đôi chân không di chuyển.
Đi bộ	Người thẳng và di chuyển chậm đến vị trí khác nhau, có ít nhất 1 chân chạm đất.
Chạy bộ	Người thẳng và di chuyển nhanh đến vị trí khác, 2 chân không chạm đất khi chạy tại 1 thời điểm.



Hình 3. Vị trí đặt thiết bị giám sát

Gia tốc trên từng trục được tính theo:

$$A_i = \frac{\left(\frac{Sam_i}{1024} * R - O_i \right)}{S_i} \quad (1)$$

A_i : Giá trị gia tốc theo trục i (X, Y, Z);

Sam_i : Giá trị sau khi lấy mẫu trục i;

R: Tham chiếu điện áp;

O_i : Phần bù;

S_i : Độ nhảy của cảm biến gia tốc trên trục i .

Dữ liệu đo được từ các hành động được phân đoạn để tăng hiệu suất nhận dạng hành động. Trong nghiên cứu này, nhóm nghiên cứu sử dụng phương pháp cửa sổ trượt có kích thước cố định theo thời gian là 6 giây. Nếu kích thước cửa sổ quá ngắn, tỷ lệ trùng lặp trong dữ liệu thu thập tăng lên. Tuy nhiên nếu chiều rộng cửa sổ thời gian quá dài thì sẽ tăng xác suất có nhiều hành động xuất hiện trong một cửa sổ trượt [13].

Bảng II
SỐ LƯỢNG CỦA SỐ DỮ LIỆU

HÀNH ĐỘNG	CỬA SỐ TRƯỢT
Nằm	540
Ngồi	572
Đứng	700
Đi bộ	728
Chạy bộ	505
Tổng	3045

Do vậy, bộ dữ liệu thu thập được phân đoạn thành 3045 cửa sổ (Bảng II) với kích thước mỗi cửa sổ là 6 giây (s). Mẫu bản ghi đoạn dữ liệu cho mỗi cửa sổ có độ dài 6s được thể hiện trong Bảng III. Giá trị gia tốc sử dụng đơn vị là g với ($g = 9.8m/s^2$).

Bảng III
SỐ QUAN SÁT HÀNH ĐỘNG THU THẬP TỪ EO

X(g)	Y(g)	Z(g)	THỜI GIAN
0,14	0,03	1,03	2021-08-03 09:44:30
0,14	0,04	1,02	2021-08-03 09:44:31
0,15	0,03	1,03	2021-08-03 09:44:32
0,14	0,03	1,03	2021-08-03 09:44:33
0,14	0,04	1,02	2021-08-03 09:44:34
0,14	0,04	1,03	2021-08-03 09:44:35

3. Trích chọn đặc trưng

Để phân loại hành động đạt hiệu suất cao, việc lựa chọn đặc trưng từ dữ liệu các hoạt động rất quan trọng. Các đặc trưng được sử dụng trong nghiên cứu này: Mean và SD, Các đặc trưng này được trích xuất theo công thức (2), (3) với đầu vào là giá trị cảm biến thu thập tương ứng trên từng trục X, Y và Z với đơn vị là g .

$$Mean \quad \mu(X_j) = \frac{1}{N} \sum_{i=1}^N x_i \quad (2)$$

$$SD \quad \sigma(X_j) = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad (3)$$

Trong đó: X_j là bản ghi thứ j và x_i là mẫu thứ i trong bản ghi X_j ; N là số bản ghi; $\mu(X_j)$ là giá trị Mean của X_j .

Dữ liệu tổng hợp tại Bảng IV cho thấy đặc trưng Mean của 3 hành động đứng, đi bộ và chạy bộ có xu hướng dao động quanh một giá trị trung tâm theo trục X là -0,93g. Bên cạnh đó, hành động nằm và ngồi (trạng thái nghỉ) có giá trị đặc trưng Mean tập trung trên trục Y, Z khác biệt so với đứng, đi bộ chạy bộ (trạng thái vận động). Ví dụ: hành động nằm và ngồi có đặc trưng Mean trên trục Y nhỏ nhất là 0,37g trong khi với hành động đứng, đi bộ và chạy bộ có đặc trưng Mean trên trục Y lớn nhất là 0,085g. Tiếp theo, giữa hành động nằm và ngồi có giá trị các trục khác biệt trông thấy. Ví dụ: nằm có đặc trưng Mean trên trục Z là 0,501g còn đối với ngồi là -0,105g. Do đó, Mean có thể phân biệt người đó ở trạng thái nằm hay ngồi và phân biệt với những trạng thái còn lại.

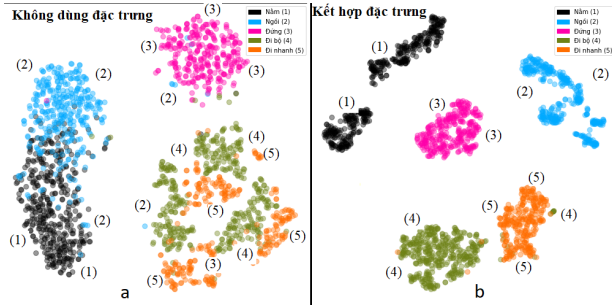
Bảng IV
CÁC ĐẶC TRƯNG DỮ LIỆU GIA TỐC THEO TRỤC X, Y VÀ Z

Đặc trưng	Trục	nằm	Ngồi	Đứng	Đi bộ	Chạy bộ
Mean	X(g)	-0,064	-0,874	-0,934	-0,928	-0,932
	Y(g)	0,823	0,37	0,085	0,072	0,004
	Z(g)	0,501	-0,105	-0,024	0,03	-0,032
SD	X(g)	0,018	0,011	0,0085	0,184	0,488
	Y(g)	0,022	0,023	0,025	0,206	0,42
	Z(g)	0,021	0,045	0,03	0,111	0,209

Khác với đặc trưng Mean, trạng thái động (đi bộ, chạy bộ) có giá trị đặc trưng SD khác biệt rõ ràng so với trạng thái tĩnh (đứng, ngồi, nằm). Ví dụ: giá trị đặc trưng của người khi đang di chuyển $SD_{min} = 0,111g$ trong khi ở trạng thái không di chuyển thì đặc trưng $SD_{max} = 0,045g$. Bên cạnh đó, đặc trưng SD có chênh lệch rõ rệt giữa hai hành động đi bộ và chạy bộ. Ví dụ: Theo trục X, đi bộ có $SD = 0,184g$ so với chạy bộ có $SD = 0,488g$. Do đó, đặc trưng có thể dùng để phân biệt một người ở trạng thái tĩnh hay động và người đó đi bộ hay chạy bộ.

4. Phân loại hành động

Để giảm số lượng kích thước dữ liệu giúp cho việc quan sát dữ liệu trực quan hơn, nhóm nghiên cứu sử dụng phương pháp t-SNE (t-Distributed Stochastic Neighbor Embedding) [15]. Thuật toán này sẽ ánh xạ từng đặc trưng đã được trích xuất trong không gian 3 chiều (X, Y, Z) sang không gian 2 chiều mà không sai lệch tính chất đặc trưng. Nếu không dùng đặc trưng, các hành động trùng lặp gần như không thể phân biệt (Hình 4).

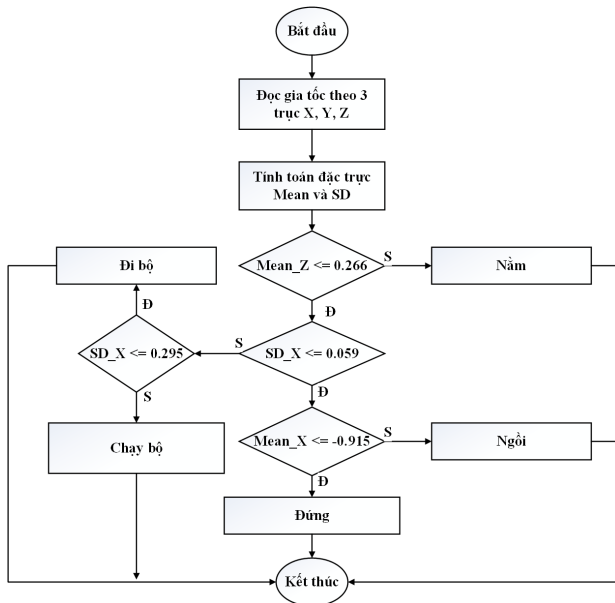


Hình 4. Ảnh xạ bộ dữ liệu trước và sau khi sử dụng đặc trưng

Với việc kết hợp hai đặc trưng SD và Mean, việc phân loại hành động đã cho đạt kết quả tốt. Trong Hình 4 - a, sự trùng lặp nhỏ xuất hiện giữa hành động đi bộ (khu vực màu xanh lá cây) và chạy bộ (khu vực màu hồng). Tuy nhiên ranh giới giữa các hành động được xác định khá rõ rệt so với khi chưa dùng đặc trưng.

5. Thuật toán phân loại

Đặc trưng Mean và SD sau khi được trích xuất từ tập dữ liệu sẽ được huấn luyện bởi thuật toán học máy cây quyết định và từ đó xây dựng thuật toán phân loại hành động (Hình 5).

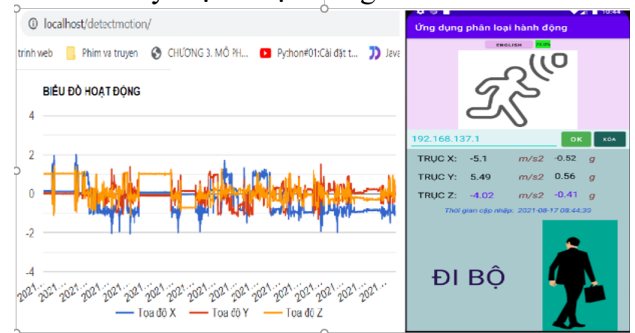


Hình 5. Thuật toán phân loại hành động người

Thiết bị sau khi được nhúng thuật toán sẽ thực hiện quá trình phân loại hành động dựa theo dữ liệu gia tốc theo thời gian. Cứ sau mỗi 6 giây dữ liệu, thiết bị sẽ phân loại hành động.

6. WebServer và ứng dụng phân loại hành động

Việc lưu trữ dữ liệu trên bộ nhớ mở rộng của thiết bị gây khó khăn trong việc quan sát dữ liệu thu thập và truy cập dữ liệu. Do đó, thiết bị được kết nối đến hệ thống WebServer (hình 6), dữ liệu sẽ được lưu trữ trên Server và có thể quan sát trên Website hay điện thoại thông minh.



Hình 6. Giao diện WebServer và phần mềm giám sát

Giao diện của ứng dụng trên điện thoại thông minh được xây dựng giúp người sử dụng dễ dàng quan sát tham số thu thập từ cảm biến theo đơn vị gia tốc g hoặc m/s² và thời gian cập nhập dữ liệu từ cảm biến. Bên cạnh đó, chức năng hiển thị dung lượng pin giúp người dùng biết khi nào thiết bị cần sạc pin. Người dùng có thể nhấn vào nút chức năng English/Tiếng việt để chuyển đổi giao diện ngôn ngữ hiển thị.

III. KẾT QUẢ VÀ ĐÁNH GIÁ

1. Phương pháp đánh giá

Hiệu năng của mô hình sử dụng được đánh giá theo ba thông số sau: độ chính xác (Accuracy), độ nhạy (Sensitivity) và độ chỉ rõ (Specify). Các giá trị trên được xác định dựa trên ma trận nhầm lẫn (Confusion Matrix) [16] cho từng hành động:

$$acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

$$sen = \frac{TP}{TP + FN} \quad (5)$$

$$spe = \frac{TN}{TN + FP} \quad (6)$$

Trong đó, TP (True Positive): Số lượng hành động cần phân loại được phân loại đúng so với quan sát. FN (False Negative): Số lượng hành động cần phân loại bị phân loại thành các hành động khác. FP (False Positive): Số lượng độ hành động khác bị phân loại thành hành động cần phân loại. TN (True Negative): Số lượng hành động khác được phân loại đúng so với quan sát thực tế.

2. Đánh giá mô hình phân loại

Trong nghiên cứu này, chúng tôi đánh giá mô hình đề xuất với những mô hình phân loại như: Gradient Boosted Decision Tree (GBDT) [17], Support Vector Machine [18], Random Forest [19], k-Nearest Neighbor [20].

Bảng V
SO SÁNH KHI ÁP DỤNG MỘT SỐ THUẬT TOÁN THÔNG DỤNG

Đặc trưng	DT	GBDT	SVM	RF	KNN
Mean	85%	92,7%	91%	93,4%	92%
SD	80,3%	84,6%	88%	84%	83%
Kết hợp	99,3%	99,6%	100%	99,6%	100%

Đặc trưng dữ liệu được huấn luyện và đánh giá trong Bảng V. Hiệu suất cao nhất khi sử dụng mô hình phân loại GBDT với đặc trưng Mean. Độ chính xác thấp nhất của các mô hình phân loại là 85% với mô hình cây quyết định (DT) và đặc trưng SD. Tuy nhiên, sự kết hợp giữa 2 đặc trưng SD và Mean đã cho kết quả tăng đến 99.3%.

Bảng VI
MA TRẬN NHẦM LẤN TRÊN DỮ LIỆU THU THẬP

Hành Động	Hành động dự đoán				
	Nằm	Ngồi	Đứng	Đi bộ	Chạy bộ
Nằm	278	0	0	0	0
Ngồi	0	228	0	0	0
Đứng	0	0	215	1	0
Đi bộ	0	0	1	286	4
Chạy bộ	0	0	0	4	198

Kết quả được thể hiện chi tiết của mô hình cây quyết định được trình bày trong Bảng VI. Thống kê số trường hợp hành động được phân loại chính xác và số trường hợp nhầm lẫn với các hành động khác cho thấy các mẫu thử nghiệm về hành động được phân loại tốt. Tuy có 4/202 mẫu chạy bộ, 5/291 mẫu đi bộ và 1/216 mẫu đứng phân loại sai nhưng tỷ lệ sai trong khoảng 0,5% - 2% là chấp nhận được.

Bảng VII thống kê độ chính xác (acc), độ nhạy (sen), độ chỉ rõ (spe) khi phân loại hành động nằm và ngồi đạt 100%. Những hành động còn lại có độ chính xác, độ chỉ rõ và độ nhạy lớn hơn 98%.

Nhìn chung với bộ dữ liệu thu thập được, các chỉ số đạt được trên 98% với các hành động. Điều này cho thấy mô hình phân loại cây quyết định thực sự có hiệu quả rất cao do dữ liệu các hành động được thực hiện lấy mẫu là thuần nhất, ít có sự chồng chéo giữa các hành động được lấy mẫu và không có hành động khác xen vào gây nhiễu.

Bảng VII
CHỈ SỐ CHÍNH XÁC MÔ HÌNH CÂY QUYẾT ĐỊNH

Chỉ số	Nằm	Ngồi	Đứng	Đi bộ	Chạy bộ
acc	100%	100%	99,83%	99,18%	99,34%
sen	100%	100%	99,54%	98,28%	98,02%
spe	100%	100%	99,90%	99,46%	99,60%

3. Kết quả thực nghiệm

Quá trình thực nghiệm phân loại hành động trên thiết bị thực được tiến hành như sau: mỗi hành động được kiểm tra trong khoảng thời gian 5-10 phút, quá trình này diễn ra liên tục và có sự thay đổi hành động sát với thực tế trong quá trình thực nghiệm. Ví dụ chuyển từ ngồi sang đi bộ, chuyển từ đi bộ sang nằm.

Bảng VIII
MA TRẬN NHẦM LẤN KHI THỰC NGHIỆM

Hành Động	Hành động dự đoán				
	Nằm	Ngồi	Đứng	Đi bộ	Chạy bộ
nằm	65	0	0	0	0
Ngồi	3	60	5	0	0
Đứng	0	5	54	2	0
Đi bộ	0	3	7	60	3
Chạy bộ	0	0	0	9	55

Kết quả thực nghiệm phân loại các hành động trên thiết bị gắn tại eo người được tổng hợp trên ma trận chồng chéo (Bảng VIII). Hành động nằm được phân loại với độ chính xác rất cao đạt 96,71%, tương đối ít hành động bị phân loại nhầm với độ nhạy thấp nhất là 82,19%. Hiệu suất khi phân loại từng hành động trên thiết bị theo thời gian thực được mô tả tại Bảng IX. Hiệu suất tổng thể khi phân loại trên thiết bị có độ chính xác đạt trên 91%. Trong đó, độ chính xác rất cao đạt được 98,94% với hành động nằm, tiếp theo là hành động chạy bộ với 96,08%. Hành động đứng và ngồi có độ chính xác thấp hơn một chút nhưng cao hơn so với hành động đi bộ.

Bảng IX
HIỆU SUẤT MÔ HÌNH KHI THỰC NGHIỆM

Chỉ số	Nằm	Ngồi	Đứng	Đi bộ	Chạy bộ
acc	96,71%	94,23%	94,53%	91,59%	96,08%
sen	100%	85,71%	84,38%	82,19%	85,94%
spe	95,82%	96,69%	97,17%	94,35%	98,76%

IV. THẢO LUẬN

Thực nghiệm phân loại theo thời gian thực có độ chính xác giảm xuống do dữ liệu gia tốc thu thập không thuần

nhất. Trong một phân đoạn dữ liệu theo thời gian có thể bị lẫn vài giây của hành động trước đó hoặc kế tiếp gây khó khăn trong việc phân loại. Ví dụ: thời gian diễn ra hành động đứng có thể tồn tại thời gian chuyển từ hành động đi bộ sang hành động đứng và ngược lại nhưng điều này không bao giờ xảy ra với bộ dữ liệu được huấn luyện trên máy tính do đã được nhóm theo các hành động. Bên cạnh đó nhiều tích lũy có thể xảy ra khi chuyển trạng thái tình sang động với tốc độ nhanh. So sánh kết quả nghiên cứu của chúng tôi với kết quả nghiên cứu của Nishkam Ravi [8] và Ling Bao [9] được trình bày trong Bảng X.

Bảng X
SO SÁNH VỚI CÁC MÔ HÌNH PHÂN LOẠI NAVI VÀ BAO

Hành động	L. Bao	N. Ravi	Nghiên cứu này
Nằm	95%		99,00%
Ngồi	94,8%	96,6%	94,89%
Đứng	95,7%	91,8%	93,99%
Đi bộ	89,7%	99,5%	92,52%
Chạy bộ	87,7%	84,8%	96,12%
Tổng thể	92,58%	94,54%	95,3%

Bên cạnh đó, nhiều nghiên cứu lựa chọn khá nhiều đặc điểm để nâng cao tỷ lệ phân loại hoạt động. Ví dụ: Nghiên cứu sử dụng 43 đặc trưng của C. Y. Min [10], Catal [12], Van der Maaten [15]; nghiên cứu sử dụng đặc trưng của Vavoulas [11]. Trong tài liệu [12], Catal đã cách kết hợp rất nhiều đặc trưng và cho kết quả là 91,6%. Mặc dù họ đã phân loại nhiều hoạt động hơn (sáu hoạt động) so với nghiên cứu của chúng tôi, nhưng họ đã thử nghiệm một thuật toán phức tạp hơn khi kết hợp nhiều thuật toán phân loại khác nhau. Bên cạnh đó, việc sử dụng 43 đặc trưng khác nhau khiến giá thành và chi phí thiết bị tăng cao hơn rất nhiều và có thể khó áp dụng đại trà. Tuy nhiên với kết quả thực nghiệm thu được mô hình phân loại có khả năng ứng dụng tốt bởi nhiều lần xảy ra trong thực tế.

Để kiểm chứng hiệu quả của mô hình đề xuất, nhóm nghiên cứu áp dụng mô hình phân loại trên tập dữ liệu WISDM [23], [24] được thu thập từ 36 người tình nguyện viên với các hoạt động thông thường gồm đi bộ (đi bộ bình thường, lên/xuống cầu thang), chạy bộ, ngồi và đứng. Bộ dữ liệu công khai này được lấy mẫu với tần số 20Hz. Tỷ lệ dữ liệu đào tạo và dữ liệu kiểm tra là 6/4. Số mẫu trên một phân đoạn dữ liệu là 120 mẫu tương ứng với tần số lấy mẫu 20Hz (Bảng XI)

Kết quả phân loại hành động trên phần dữ liệu kiểm tra được tổng hợp trong Bảng XII và XIII. Kết quả này có sự tương đồng với kết quả thực nghiệm (Bảng VIII) về sự nhầm lẫn giữa các hành vi với nhau, phù hợp với phân tích phía trên của chúng tôi. Độ chính xác tổng thể khi phân loại trên tập dữ liệu WISDM (98,4%) giảm xuống so với khi

Bảng XI
KẾT QUẢ PHÂN LOẠI TRÊN TẬP DỮ LIỆU WISDM

Hành động	Hành động dự đoán			
	Ngồi	Đứng	Đi bộ	Chạy bộ
Ngồi	169	9	4	0
Đứng	1	142	5	0
Đi bộ	0	1	2029	47
Chạy bộ	0	0	47	1039

phân loại trên tập dữ liệu thu thập (99,3%) nhưng cao hơn khi tiến hành thực nghiệm (95,26%). Cụ thể, hành động ngồi có 9 mẫu bị phân loại thành hành động đứng và 9 mẫu bị phân loại thành đi bộ; hành động đi bộ có số lượng mẫu nhiều nhất (2076 mẫu) nhưng chỉ có 47 mẫu bị phân loại thành chạy bộ. Bên cạnh đó tỷ lệ hành động ở trạng thái di chuyển bị phân loại thành hành động ở trạng thái di chuyển là nhỏ với 1 mẫu hành động đi bộ bị phân loại thành đứng.

Bảng XII
KẾT QUẢ PHÂN LOẠI TẬP DỮ LIỆU WISDM

Hành động	Số lượng mẫu	Cửa sổ dữ liệu	Dữ liệu đào tạo	Dữ liệu kiểm tra
Ngồi	59939	457	275	182
Đứng	48395	371	223	148
Đi bộ	647696	5193	3116	2007
Chạy bộ	342177	2716	1630	1086
Tổng	1098207	8737	5244	3493

Các chỉ số đạt được khi áp dụng mô hình đề xuất trên tập dữ liệu WISDM là khá tốt trên 92%. Độ nhạy thấp nhất với hành động ngồi (92,86%) và cao nhất với hành động đi bộ (97,69%). Bên cạnh đó, hành động đứng và chạy bộ có độ nhạy tương đương với 96%. Độ chỉ rõ trên 98% với các hành động ngồi, đứng và chạy bộ trong khi hành động đi bộ có độ chỉ rõ đạt 96%.

Bảng XIII
HIỆU SUẤT PHÂN LOẠI TRÊN TẬP WISDM

Chỉ số	Ngồi	Đứng	Đi bộ	Chạy bộ	Tổng thể
acc	99,59%	99,53%	97,01%	97,29%	98,36%
sen	92,86%	95,95%	97,69%	95,67%	95,54%
spe	99,97%	99,69%	96,02%	98,03%	98,43%

Có thể thấy rằng, việc giám sát hành vi người có thể chia thành hai giải pháp là dùng 1/ camera và 2/ cảm biến gắn người. Mỗi giải pháp này đều có những ưu và nhược điểm khác nhau. Việc lựa chọn giải pháp sẽ tùy thuộc vào mục đích sử dụng. Những nghiên cứu hiện nay về xử lý ảnh

đã đạt được nhiều thành tựu rất lớn khi có thể phân loại rất nhiều hành vi với độ chính xác mà một trong những bộ dữ liệu mà các nhóm nghiên cứu thường sử dụng đó là HMDB51 dataset gồm 6766 video clips và 51 lớp hành động. Cụ thể, trong một nghiên cứu mới năm 2021 [25], độ chính xác trung bình đạt được là 95,4% cho 51 lớp hành động. Giải pháp dùng camera sẽ thể hiện nhiều bất lợi về môi trường hoạt động (ví dụ: bị gián đoạn, che khuất khi người cần giám sát di chuyển từ phòng tới khu vệ sinh, hành lang không trang bị camera,...), độ phức tạp cao khi định danh nếu giám sát nhiều người cùng lúc (điều không gặp phải nếu dùng thiết bị gắn người), sự riêng tư của người sử dụng... Ngoài ra, đây sẽ còn là một hệ thống phức tạp và khó thành công về mặt thương mại, giá cao hơn nhiều so với giá của thiết bị sử dụng vi điều khiển và cảm biến gia tốc giá thành rẻ. Nghiên cứu của chúng tôi hướng đến chăm sóc sức khỏe cho người già, người hồi phục sau phẫu thuật hay tai nạn. Do đó, số lớp hành vi liên quan tới vận động cơ bản là đủ cho loại ứng dụng này nhưng lại cần phân loại được theo thời gian thực và giá thành rẻ. Do đó, giải pháp dùng xử lý ảnh chưa phù hợp theo mục tiêu của ứng dụng này.

Mô hình của chúng tôi để có thể được áp dụng cho những đối tượng thực hiện công việc đặc biệt nguy hiểm như lính cứu hỏa thì sẽ phát triển thêm để nhận một số dạng hành động khác như: bò, đi lên, đi xuống, trạng thái sinh tồn, định vị [21,22]. Bởi lẽ, hoạt động của các đối tượng này thường diễn ra với cường độ cao hơn trong điều kiện phức tạp như cháy lớn, khói dày đặc.

V. KẾT LUẬN

Trong nghiên cứu này, nhóm nghiên cứu đã xây dựng một hệ thống phân loại hành động với chi phí thấp, thời gian tính toán nhỏ thời gian thực. Kết quả thử nghiệm với cửa sổ trượt có kích thước 6 giây cho thấy khả năng phân loại theo mô hình cây quyết định có độ chính xác cao hơn GBT, SVM, KNN. Đối với nghiên cứu sâu hơn, chúng tôi tiếp tục phát triển và cải thiện những kết quả này thông qua các thí nghiệm trên các hoạt động phức tạp hơn để kiểm tra độ chính xác của mô hình đề xuất. Bên cạnh đó, ngoài các đặc trưng thống kê cửa sổ thời gian, nghiên cứu có thể mở rộng các đặc trưng khác trong miền tần số hoặc kết hợp thông tin trên cả hai miền thời gian và tần số trong quá trình trích chọn đặc trưng. Hiện nay, xu hướng thực thi học máy trên các bộ vi điều khiển hiệu năng thấp (TinyML [26]) đang dần trở nên thịnh hành trên thế giới và nghiên cứu này cũng góp phần thực hiện TinyML cho các ứng dụng giám sát hành vi phục vụ chăm sóc sức khỏe từ xa.

LỜI CẢM ƠN

Trần Đức Nghĩa được hỗ trợ bởi Quỹ Đổi mới sáng tạo Vingroup (VinIF), Viện Nghiên cứu Dữ liệu lớn (VINBIG-DATA) trong khuôn khổ chương trình học bổng Sau Tiến sĩ trong nước (mã số học bổng VINIF.2021.STS.21).

TÀI LIỆU THAM KHẢO

- [1] Pham, Minh, Yehenew Mengistu, Ha Manh Do, and Weihua Sheng. "Cloud-based smart home environment (CoSHE) for home healthcare.", *IEEE International Conference on Automation Science and Engineering*, pp. 483-488, 2016.
- [2] Peetoom, Kirsten KB, et al. "Literature review on monitoring technologies and their outcomes in independently living elderly people.", *Disability and Rehabilitation: Assistive Technology* 10.4, 271-294, 2015.
- [3] Tan, Tran Duc, and Nguyen Van Tinh. "Reliable fall detection system using an 3-DOF accelerometer and cascade posture recognitions.", *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1-6, 2014.
- [4] Van Thanh, Pham, Duc-Tan Tran, Dinh-Chinh Nguyen, Nguyen Duc Anh, Dang Nhu Dinh, S. El-Rabaie, and Kumbesan Sandrasegaran. "Development of a real-time, simple and high-accuracy fall detection system for elderly using 3-DOF accelerometers.", *Arabian Journal for Science and Engineering*, vol 44, no. 4, pp. 3329-3342, 2019
- [5] Stiefmeier, Thomas, Daniel Roggen, Georg Ogris, Paul Lukowicz, and Gerhard Tröster. "Wearable activity tracking in car manufacturing.", *IEEE Pervasive Computing*, vol 8 , no. 2, pp.42-50, 2018.
- [6] Taha, Ahmed, Hala H. Zayed, M. E. Khalifa, and El-Sayed M. El-Horbaty. "Human activity recognition for surveillance applications." In *Proceedings of the 7th International Conference on Information Technology*, pp. 577-586. 2015.
- [7] Putra, I. P. E. S., James Brusey, Elena Gaura, and Rein Vesilo. "An event-triggered machine learning approach for accelerometer-based fall detection.", *Sensors*, vol 18, no. 1, 2018.
- [8] Ravi, Nishkam, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. "Activity recognition from accelerometer data.", *Aaai*, vol. 5, pp. 1541-1546, 2005.
- [9] Bao, Ling, and Stephen S. Intille. "Activity recognition from user-annotated acceleration data.", *International conference on pervasive computing*, pp. 1-17, 2004.
- [10] Min, Ya, Yin Yin Htay, and Khin Khin Oo. "Comparing the performance of machine learning algorithms for human activities recognition using wisdm dataset.", *International Journal of Computer (IJC)*, vol 38, no. 1, pp. 61-72, 2020.
- [11] Vavoulas, George, Charikleia Chatzaki, Thodoris Malliotakis, Matthew Padiaditis, and Manolis Tsiknakis. "The mobiact dataset: Recognition of activities of daily living using smartphones.", *International Conference on Information and Communication Technologies for Ageing Well and e-Health*, vol. 2, pp. 143-151. 2016.
- [12] Catal, Cagatay, Selin Tufekci, Elif Pirmit, and Guner Kocabag. "On the use of ensemble of classifiers for accelerometer-based activity recognition.", *Applied Soft Computing*, vol 37, pp. 1018-1022, 2015.
- [13] Pham, Van Thanh, Duc Anh Nguyen, Nhu Dinh Dang, Hong Hai Pham, Van An Tran, Kumbesan Sandrasegaran, and Duc-Tan Tran. "Highly accurate step counting at various walking states using low-cost inertial measurement unit support indoor positioning system.", *Sensors*, vol 18, no. 10, 2018.
- [14] Kamiński, Bogumił, Michał Jakubczyk, and Przemysław Szufel. "A framework for sensitivity analysis of decision

- trees.", *Central European journal of operations research*, vol 26, no. 1, pp. 135-159, 2018.
- [15] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE.", *Journal of machine learning research*, vol 9, no. 11, 2008.
- [16] Wang, Jun, Zhitao He, Guoqiang Zheng, Song Gao, and Kaixuan Zhao. "Development and validation of an ensemble classifier for real-time recognition of cow behavior patterns from accelerometer data and location data.", *PloS one*, vol 13, no. 9, 2018.
- [17] Alcolea, Adrián, and Javier Resano. "FPGA Accelerator for Gradient Boosting Decision Trees.", *Electronics*, vol 10, no. 3, 2021.
- [18] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20, no. 3 (1995): 273-297.
- [19] Shiri, Jalal. "Improving the performance of the mass transfer-based reference evapotranspiration estimation approaches through a coupled wavelet-random forest methodology.", *Journal of Hydrology* 561, pp. 737-750, 2018.
- [20] Hattori, Kazuo, and Masahito Takahashi. "A new edited k-nearest neighbor rule in the pattern classification problem.", *Pattern Recognition*, vol 33, no. 3, pp. 521-528, 2000.
- [21] Pham, Van Thanh, Quang Bon Le, Duc Anh Nguyen, Nhu Dinh Dang, Huu Tue Huynh, and Duc Tan Tran. "Multi-sensor data fusion in a real-time support system for on-duty firefighters.", *Sensors*, vol 19, no. 21, 2019.
- [22] Tan, Tran Duc, L. M. Ha, N. T. Long, N. D. Duc, and N. P. Thuy. "Integration of inertial navigation system and global positioning system: Performance analysis and measurements.", *International Conference on Intelligent and Advanced Systems*, pp. 1047-1050, 2007.
- [23] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers.", *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74-82, 2011.
- [24] Weiss, Gary Mitchell, and Jeffrey Lockhart. "The impact of personalization on smartphone-based activity recognition.", *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.
- [25] Varshney, Neeraj, and Brijesh Bakariya. "Deep convolutional neural model for human activities recognition in a sequence of video by combining multiple CNN streams.", *Multimedia Tools and Applications*, pp. 1-13, 2021.
- [26] Dutta, Lachit, and Swapna Bharali. "TinyML Meets IoT: A Comprehensive Survey.", *Internet of Things*, 2021.

SƠ LƯỢC VỀ TÁC GIẢ

Đào Tô Hiệu



Tốt nghiệp thạc sĩ chuyên ngành Kỹ thuật điện tử tại Đại học Bách Khoa Hà Nội. Đang làm nghiên cứu sinh tại Viện Hàn lâm Khoa học và Công nghệ Việt Nam (VAST).
Hiện là giảng viên trường Đại học Công nghệ Thông tin và Truyền thông, Đại học Thái Nguyên.
Email: dthieu@ictu.edu.vn

Ngô Văn Công



Tốt nghiệp trường Đại học kỹ thuật Lê Quý Đôn năm 2012, nhận bằng thạc sĩ năm 2015.
Hiện là nghiên cứu sinh năm nhất tại Học viện Khoa học và Công nghệ.
Lĩnh vực nghiên cứu: Điều khiển và tự động hóa; Tiêu chuẩn - Đo lường - Chất lượng.
Email: ngovancong@tcvn.gov.vn

Nguyễn Quang Huy



Sinh năm 1990, tốt nghiệp thạc sĩ ngành Công nghệ thông tin tại Học viện Công nghệ Bưu chính Viễn thông.
Hiện công tác tại Viện Công nghệ Thông tin (IOIT), Viện Hàn lâm Khoa học và Công nghệ Việt Nam (VAST).
Lĩnh vực nghiên cứu: Phân tích dữ liệu, mạng cảm biến không dây và các ứng dụng.
Email: nqhuy@ioit.ac.vn

Trần Đức Nghĩa



Nhận bằng tiến sĩ tại Université Paris Descartes (Paris, Pháp) năm 2018.
Lĩnh vực nghiên cứu: Phổ EPR, Internet vạn vật, ứng dụng học máy, cảm biến gia tốc.
Email: nghiatd@ioit.ac.vn

Trần Đức Tân



Sinh năm 1980 tại Bắc Ninh. Được phong PGS năm 2013.
Hiện là Phó Trưởng khoa Điện – Điện tử, Trường Đại học PHENIKAA.
Lĩnh vực nghiên cứu: Kỹ thuật điện tử - xử lý tín hiệu.
Email: tan.tranduc@phenikaa-uni.edu.vn

A Reliable Numerical Scheme for a Computer Virus Propagation Model

Pham Hoai Thu¹, Hoang Manh Tuan²

¹ Faculty of Information Security, People's Security Academy, Hanoi, Vietnam ² Department of Mathematics, FPT University, Hoa Lac Hi-Tech Park, Km29 Thang Long Blvd, Hanoi, Vietnam

Correspondence: Hoang Manh Tuan, tuanhm14@fe.edu.vn; hmtuan01121990@gmail.com

Date received: 28/08/2021, Date revised: 29/10/2021, Date accepted: 15/11/2021

Digital Object Identifier: 10.32913/mic-ict-research-vn.v2021.n2.1007

Abstract: In this paper, we develop the Mickens' methodology to construct a dynamically consistent Non-Standard Finite Difference (NSFD) scheme for a recognized computer virus propagation model. It is proved that the constructed NSFD scheme correctly preserves essential mathematical features of the continuous-time model, which are positivity, boundedness and asymptotic stability. As an important consequence, we obtain an effective numerical scheme that has the ability to provide reliable approximations, meanwhile, some typical standard finite difference schemes fail to preserve the essential properties of the computer virus propagation model; hence, they can generate numerical approximations which are not only negative but also unstable. Finally, a set of numerical experiments is performed to support the theoretical results as well as to demonstrate the advantage of the NSFD scheme over standard ones. As we expected, there is a good agreement between the numerical results and theoretical assertions.

Keywords: *Computer virus, NSFD scheme, positivity, stability, dynamics consistency*

Title: Một phương pháp hiệu quả mô phỏng số mô hình lan truyền virus máy tính

Abstract: Trong bài báo này, chúng tôi phát triển phương pháp luận của Mickens để xây dựng lược đồ sai phân phi tiêu chuẩn tương thích động lực với một mô hình lan truyền virus máy tính. Các phân tích toán học chỉ ra rằng lược đồ sai phân phi tiêu chuẩn bảo toàn các tính chất quan trọng của mô hình liên tục, bao gồm tính dương, tính bị chặn và tính chất ổn định tiệm cận. Trong khi đó, một số lược đồ sai phân tiêu chuẩn không thể bảo toàn các tính chất quan trọng của mô hình liên tục. Do đó, chúng có thể sinh ra các lời giải xấp xỉ âm và không ổn định. Các mô phỏng số được thực hiện để hỗ trợ và minh họa cho các kết quả lý thuyết cũng như để chứng minh ưu thế của lược đồ sai phân phi tiêu chuẩn so với các lược đồ sai phân tiêu chuẩn. Các mô phỏng số chỉ ra rằng các kết quả số là tương thích với các kết quả lý thuyết.

Keywords: *Computer virus, NSFD scheme, positivity, stability, dynamics consistency.*

I. INTRODUCTION

The study of numerical methods for solving differential equations has played an important role not only in theory but also in practice. Over the past of several decades, a large number of numerical methods have been proposed and developed by mathematicians and engineers (see, for example, [1–3]). The concept of Non-Standard Finite Difference schemes was first introduced by Mickens to overcome drawbacks of Standard Finite Difference (SFD) schemes [4–8]. One of the main advantages of NSFD schemes is that they can preserve essential mathematical features of differential equations models (positivity, stability, monotonicity, periodicity, physical properties, etc.) regardless of chosen step sizes. Nowadays, NSFD schemes have been widely used

as a powerful and effective approach to solve differential equation models that describe important phenomena and processes arising in science and technology. Good reviews of NSFD schemes can be found in [7, 9, 10]. Also, we refer the readers to [11–15] for some recent notable results on NSFD schemes for numerical approximations of differential equation models. Recently, we have successfully developed the Mickens' methodology to formulate NSFD schemes for mathematical models arising in real-world situations [16–22].

Nowadays, computer viruses have been one of the biggest threats to Internet security as well as to work and daily life. Therefore, effective strategies and measures for preventing and controlling computer viruses are always needed. For

this purpose, many mathematical models, based on the high similarity between computer viruses and biological ones, have been constructed and developed to study the spread of computer viruses (see, for example, [23–28] and references therein). From the analysis of these models, we can propose effective strategies and measures for preventing and controlling computer viruses.

Motivated by the importance of computer virus propagation models and their numerical approximations, in this work we consider a recognized computer virus propagation model proposed by Gan et al. in a previous work [25]. The mathematical model and its dynamics will be provided in Section II. Although this computer virus propagation model has many useful applications in information technology and computer science and it has been studied and developed at some levels (see, for example, [23, 24]), to the best of our knowledge, dynamically consistent discrete-time models have not been constructed yet. Therefore, our objective is to construct a discrete-time model preserving essential qualitative features of the continuous-time model for all finite step sizes. For this purpose, the Mickens' methodology is applied to propose an appropriate NSFD scheme for the continuous-time model. After that, the positivity, boundedness and asymptotic stability of the proposed NSFD scheme are established. As a result, we obtain a dynamically consistent NSFD scheme that is able to provide reliable numerical approximations for the computer virus propagation model. In addition to the theoretical results, a set of numerical experiments is performed to support the theoretical results and to show advantages of the NSFD scheme over some typical SFD schemes. As will be seen in Section IV, the standard Euler and second-order Runge-Kutta (RK2) schemes fail to correctly preserve the dynamical properties of the continuous model for some given step sizes. More clearly, they generate numerical approximations that are not only negative but also unstable. Conversely, numerical approximations generated by the NSFD scheme faithfully reflect the dynamics of the continuous model regardless of chosen step sizes.

The plan of the paper is as follows:

The mathematical model of the computer virus propagation model is provided in Section II. The NSFD scheme is proposed and analyzed in Section III. Section IV reports some numerical experiments. Some conclusions and discussions are presented in the last section.

II. MATHEMATICAL MODEL AND ITS DYNAMICS

In this section, we briefly recall from [25] the mathematical model of the computer virus propagation model and its dynamics. In this model, at any time, a computer

is called internal or external depending on it is accessed to the Internet or not. All internal computers are classified into two compartments:

- 1) S compartment: The set of all uninfected internal computers or susceptible internal computers.
- 2) I compartment: The set of all infected internal computers.

Besides, all external computers are categorized into a single compartment:

- 1) E compartment: The set of all external computers.

Let $S(t)$, $I(t)$, and $E(t)$ be the the average numbers of susceptible, infected and external computers at time t , respectively. Based on some reasonable hypotheses (see the hypotheses $(A_1) - (A_7)$ in [25]), the following differential equation model was proposed:

$$\begin{aligned}\dot{S} &= \gamma_2 I + \eta_2 E - \mu S - \beta SI - \gamma_1 S, \\ \dot{I} &= \beta SI - \mu I - \gamma_1 I - \gamma_2 I + \eta_1 E, \\ \dot{E} &= \delta + \gamma_1 S + \gamma_1 I - \mu E - \eta_1 E - \eta_2 E,\end{aligned}\quad (1)$$

where all the parameters are assumed to be positive. We refer the readers to [25] for more details of this model. It was proved in [25] that:

- 1) The model (1) has no virus-free equilibrium point.
- 2) The model (1) possesses a unique viral equilibrium point $E_V^* = (S^*, I^*, E^*)$, where

$$\begin{aligned}E^* &= \frac{\delta(\mu + \gamma_1)}{\mu(\mu + \gamma_1 + \eta_1 + \eta_2)}, \\ I^* &= \frac{\beta w - (\mu + \gamma_1 + \gamma_2)}{2\beta} \\ &\quad + \frac{\sqrt{[\beta w - (\mu + \gamma_1 + \gamma_2)]^2 + 4\beta\eta_1 E^*}}{2\beta}, \\ S^* &= \frac{\delta}{\mu} - E^* - I^*,\end{aligned}\quad (2)$$

with

$$w := \frac{\delta(\gamma_1 + \gamma_2)}{\mu(\mu + \gamma_1 + \eta_1 + \eta_2)}.$$

- 3) The viral equilibrium point E_V^* is not only locally asymptotically stable but globally asymptotically stable with respect to the positively invariant set

$$\Omega := \left\{ (S, I, E) | S, I, E \geq 0, S + I + E \leq \frac{\delta}{\mu} \right\}.$$

III. THE CONSTRUCTION OF NSFD SCHEME

Consider the continuous-time model (1) on a finite interval $[0, T]$ and discretize this interval by a uniform mesh

$$0 = t_0 < t_1 < \dots < t_i < \dots < t_{M-1} < t_M = T,$$

where $t_{n+1} - t_n = \Delta t$ for $i = 0, 1, \dots, M - 1$. Let (S_n, I_n, R_n) ($n = 1, 2, \dots, M$) be the intended approximations for $(S(t_n), I(t_n), R(t_n))$, respectively. Using the Mickens' methodology [4–8], we discretize the model (1) as follows

$$\begin{aligned}\dot{S}(t_n) &\approx \frac{S_{n+1} - S_n}{\varphi(\Delta t)}, \\ \dot{I}(t_n) &\approx \frac{I_{n+1} - I_n}{\varphi(\Delta t)}, \\ \dot{E}(t_n) &\approx \frac{E_{n+1} - E_n}{\varphi(\Delta t)},\end{aligned}\quad (3)$$

and

$$\begin{aligned}&\gamma_2 I(t_n) + \eta_2 E(t_n) - \mu S(t_n) - \beta S(t_n) I(t_n) - \gamma_1 S(t_n) \\ &\approx \gamma_2 I_n + \eta_2 E_n - \mu S_{n+1} - \beta S_{n+1} I_n - \gamma_1 S_{n+1}, \\ &\beta S(t_n) I(t_n) - \mu I(t_n) - \gamma_1 I(t_n) - \gamma_2 I(t_n) + \eta_1 E(t_n) \\ &\approx \beta S_{n+1} I_n - \mu I_{n+1} - \gamma_1 I_{n+1} - \gamma_2 I_n + \eta_1 E_n, \\ &\delta + \gamma_1 S(t_n) + \gamma_1 I(t_n) - \mu E(t_n) - \eta_1 E(t_n) - \eta_2 E(t_n) \\ &\approx \delta + \gamma_1 S_{n+1} + \gamma_1 I_{n+1} - \mu E_{n+1} - \eta_1 E_n - \eta_2 E_n,\end{aligned}\quad (4)$$

where $\varphi(\Delta t)$ is a positive function satisfying $\varphi(\Delta t) = \Delta t + O(\Delta t^2)$ as $\Delta t \rightarrow 0$. For convenience, the variable Δt will be omitted in some presentations. The discretization process (3)-(4) leads to the following NSFD scheme for the model (1)

$$\begin{aligned}\frac{S_{n+1} - S_n}{\varphi(\Delta t)} &= \gamma_2 I_n + \eta_2 E_n - \mu S_{n+1} \\ &\quad - \beta S_{n+1} I_n - \gamma_1 S_{n+1}, \\ \frac{I_{n+1} - I_n}{\varphi(\Delta t)} &= \beta S_{n+1} I_n - \mu I_{n+1} - \gamma_1 I_{n+1} \\ &\quad - \gamma_2 I_n + \eta_1 E_n, \\ \frac{E_{n+1} - E_n}{\varphi(\Delta t)} &= \delta + \gamma_1 S_{n+1} + \gamma_1 I_{n+1} - \mu E_{n+1} \\ &\quad - \eta_1 E_{n+1} - \eta_2 E_n.\end{aligned}\quad (5)$$

We now investigate dynamical properties of the NSFD model (5). Let us define a sequence $\{N_n\}_{n=0}^\infty$ with

$$N_n := S_n + I_n + R_n, \quad n \geq 0.$$

Theorem 1 (Positivity and boundedness): Let $\varphi(\Delta t)$ be a function satisfying

$$\varphi(\Delta t) < \Delta t^* := \min \left\{ \frac{1}{\gamma_2}, \frac{1}{\eta_1 + \eta_2} \right\} \text{ for all } h > 0. \quad (6)$$

Then, we have the following assertions:

- 1) The set $\mathbb{R}_3^+ = \{(S, I, R) | S, I, E \geq 0\}$ is a positively invariant set of the NSFD scheme (5), i.e., $S_{n+1}, I_{n+1}, E_{n+1} \geq 0$ whenever $S_n, I_n, E_n \geq 0$.
- 2) The sequence $\{N_n\}_{n=0}^\infty$ monotonically converges to $N^* := \delta/\mu$ as $n \rightarrow \infty$.

Proof: Proof of Part 1. First, we rewrite the system (5) in the explicit form:

$$\begin{aligned}S_{n+1} &= \frac{S_n + \varphi\gamma_2 I_n + \varphi\eta_2 E_n}{1 + \varphi(\mu + \beta I_n + \gamma_1)}, \\ I_{n+1} &= \frac{(1 - \varphi\gamma_2)I_n + \varphi\beta S_{n+1} I_n + \varphi\eta_1 E_n}{1 + \varphi(\mu + \gamma_1)}, \\ E_{n+1} &= \frac{\varphi\delta + \varphi\gamma_1(S_{n+1} + I_{n+1}) + [1 - \varphi(\eta_1 + \eta_2)]E_n}{1 + \varphi\mu}.\end{aligned}\quad (7)$$

The condition (6) implies that $1 - \varphi\gamma_2 > 0$ and $1 - \varphi(\eta_1 + \eta_2) > 0$. So, we deduce that $S_{n+1}, I_{n+1}, E_{n+1} \geq 0$ whenever $S_n, I_n, E_n \geq 0$. The proof of this part is complete.

Proof of Part 2. Adding side-by-side the 1st, 2nd and 3rd equations of the system (5) we obtain

$$\frac{N_{n+1} - N_n}{\varphi} = \delta - \mu N_{n+1}, \quad (8)$$

which implies that

$$N_{n+1} = \left(\frac{\delta}{\mu} - N_0 \right) \left(\frac{1}{1 + \varphi\mu} \right)^n + \frac{\delta}{\mu}.$$

Consequently, the monotone convergence of the sequence $\{N_n\}$ is proved by using the fact that $1/(1 + \varphi\mu) \in (0, 1)$. The proof of this part is completed.

Next, to determine the set of equilibria of the model (5), we need to solve the following system

$$S_{n+1} = S_n, \quad I_{n+1} = I_n, \quad E_{n+1} = E_n.$$

It is easy to verify that equilibrium points of the models (5) and (1) are identical. Hence, the NSFD model (5) only possesses a unique viral (positive) equilibrium point E_V^* , which is given by (2).

Theorem 2 (Asymptotic stability): Let $\varphi(\Delta t)$ be a function satisfying the condition (6). Then, the viral equilibrium point E_V^* is a locally asymptotically stable equilibrium point of the NSFD model (5).

Proof: First, using the relation $S_n = N_n - I_n - E_n$ and the equation (8) we transform the system (5) to the form

$$\begin{aligned}I_{n+1} &= \frac{(1 - \varphi\gamma_2)I_n + \varphi\beta(N_{n+1} - I_{n+1} - E_{n+1})I_n + \varphi\eta_1 E_n}{1 + \varphi(\mu + \gamma_1)}, \\ E_{n+1} &= \frac{\varphi\delta + \varphi\gamma_1(N_{n+1} - E_{n+1}) + [1 - \varphi(\eta_1 + \eta_2)]E_n}{1 + \varphi\mu}, \\ N_{n+1} &= N_n + \varphi\delta - \varphi\mu N_{n+1}.\end{aligned}\quad (9)$$

Then, the viral equilibrium point $E_V^* = (S^*, I^*, E^*)$ of (5) is transformed to $E_{new}^* = (I^*, E^*, \delta/\mu)$. We need to show that

E_{new}^* is a locally asymptotically stable equilibrium point of the model (9). Indeed, let us denote

$$\begin{aligned}\lambda_1 &:= \frac{1 - \varphi\gamma_2 + \varphi\beta S^*}{\left[1 + \frac{\varphi\beta I^*}{1 + \varphi(\mu + \gamma_1)}\right] \left[1 + \varphi(\mu + \gamma_1)\right]}, \\ \lambda_2 &:= \frac{1 - \varphi(\eta_1 + \eta_2)}{1 + \varphi(\mu + \gamma_1)}, \\ \lambda_3 &:= \frac{1}{1 + \varphi\mu}.\end{aligned}\quad (10)$$

Then, the Jacobian matrix of the system (9) at E_{new}^* is given by

$$J(E_{new}^*) = \begin{pmatrix} \lambda_1 & \frac{\partial I_{n+1}}{\partial E_n}(E_{new}^*) & \frac{\partial I_{n+1}}{\partial N_n}(E_{new}^*) \\ 0 & \lambda_2 & \frac{\partial E_{n+1}}{\partial N_n}(E_{new}^*) \\ 0 & 0 & \lambda_3 \end{pmatrix}. \quad (11)$$

Consequently, λ_1, λ_2 and λ_3 are all the eigenvalues of $J(E^*)$. It is clear that $|\lambda_3| < 1$. On the other hand, the condition (6) implies that $|\lambda_2| < 1$. We now show that $|\lambda_1| < 1$. Indeed, it follows from (6) that $\lambda_1 > 0$. Moreover, from the second equation of the system (1) we obtain

$$1 - \lambda_1 = \varphi \frac{\mu + \gamma_1 + \gamma_2 + \beta I^* - \beta S^*}{1 + \varphi(\mu + \gamma_1) + \varphi\beta I^*} = \varphi \frac{\beta I^* + \eta_1 E^* / I^*}{1 + \varphi(\mu + \gamma_1) + \varphi\beta I^*},$$

which means that $\lambda_1 < 0$. Therefore, $0 < \lambda_1 < 1$.

So, all the eigenvalues of $J(E_{new}^*)$ are inside the unit ball. By the Lyapunov indirect method [29, 30] we obtain the local asymptotic stability of E_{new}^* . The proof is completed.

Remark 1:

- 1) Theorems 1 and 2 mean that the NSFD scheme (5) preserves the positivity, boundedness and asymptotic stability of the model (1) for all finite step sizes $\Delta t > 0$.
- 2) There are many denominator functions $\varphi(\Delta t)$ satisfying the condition (6), a typical function is (see [4–8])

$$\varphi(\Delta t) = \frac{1 - e^{-\tau\Delta t}}{\tau}, \quad \tau > \frac{1}{\Delta t^*}.$$

General families of denominator functions satisfying (6) can be found in [18].

- 3) Numerical examples in Section IV show that the NSFD scheme (5) preserves not only local asymptotic stability but also global asymptotic stability of the model (1).

IV. NUMERICAL EXPERIMENTS

In this section, we reports some numerical examples to support the theoretical results as well as to show the

advantage of the NSFD scheme over standard ones. For this purpose, we consider the model (1) with the following data.

Table I
PARAMETERS USED IN NUMERICAL EXAMPLES

Case	β	δ	μ	η_1	η_2	γ_1	γ_2
1	0.05	0.8	0.01	0.75	0.85	0.6	0.8
2	0.05	0.02	0.01	0.8	0.9	0.6	0.8

Example 1: Consider the model (1) with the parameters given by Case 1 in Table I. In this case, the viral equilibrium point $E_V^* = (19.56, 38.36, 22.08)$ is asymptotically stable. Numerical approximations generated by the standard Euler scheme, RK2 scheme and NSFD scheme are depicted in Figures 1-4, respectively. It is clear that the Euler and RK2 schemes fail to preserve the asymptotic stability of the continuous model, meanwhile, the NSFD scheme correctly preserves the dynamics of the continuous model. Furthermore, the dynamics of NSFD scheme does not depend on the chosen step sizes.

Example 2: Consider the model (1) with the parameters given by Case 2 in Table I. In this case, the viral equilibrium point $E_V^* = (1.16, 0.31, 0.52)$ is asymptotically stable. As shown in Figure 5, the Euler and RK2 schemes destroy not only the positivity but also asymptotic stability, meanwhile, the NSFD scheme preserve all the mathematical features of the continuous model.

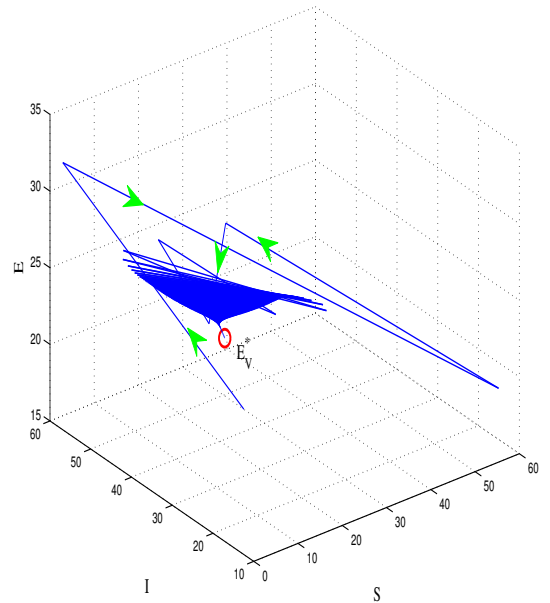


Figure 1. The phase space generated by the Euler scheme for $(S(0), I(0), E(0)) = (30, 45, 15)$ and $\Delta t = 0.8$ after 1000 iterations.

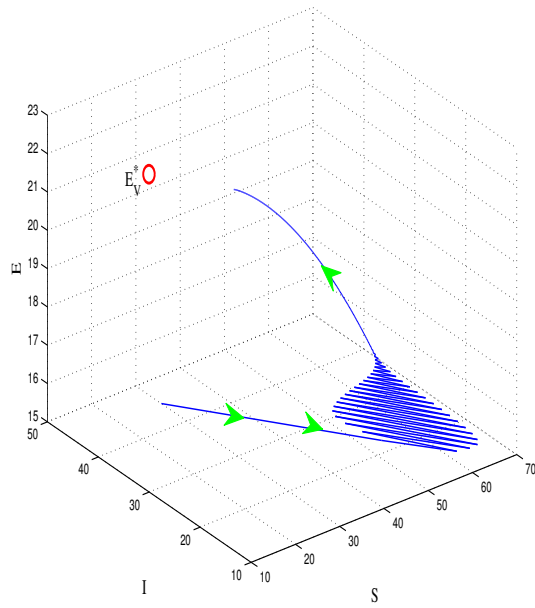


Figure 2. The phase space generated by the RK2 scheme for $(S(0), I(0), E(0)) = (30, 45, 15)$ and $\Delta t = 0.9$ after 1000 iterations.

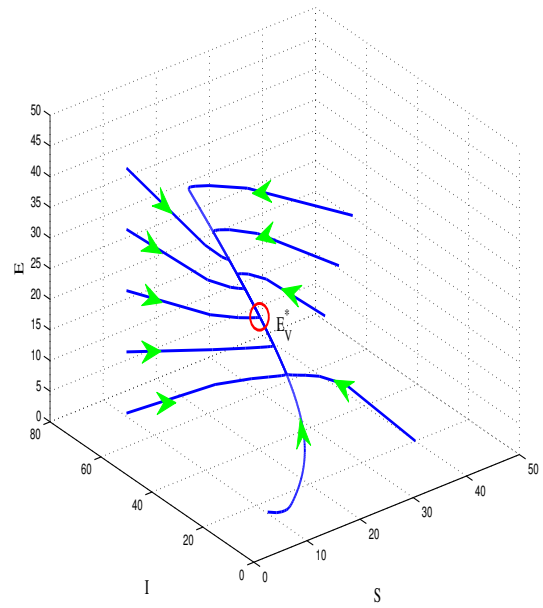


Figure 4. The phase spaces generated by the NSFD scheme for $\Delta t = 2$ and $\varphi(\Delta t) = \frac{1 - e^{-1.6\Delta t}}{1.6}$ after 1000 iterations.

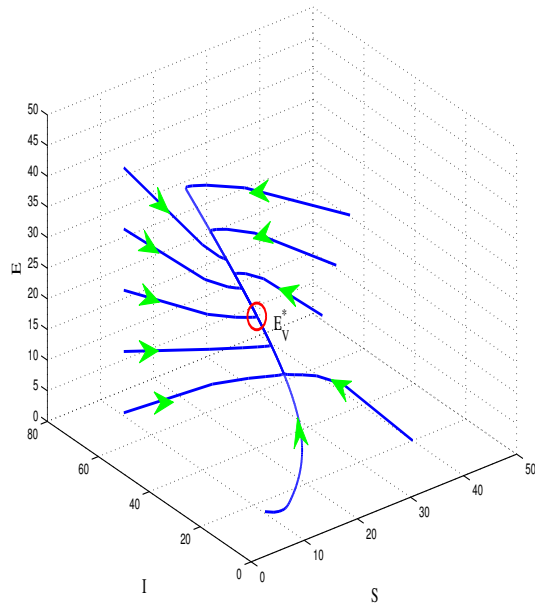


Figure 3. The phase spaces generated by the NSFD scheme for $\Delta t = 1$ and $\varphi(\Delta t) = \frac{1 - e^{-1.6\Delta t}}{1.6}$ after 1000 iterations. Each blue curve represents a phase space corresponding to a specific initial data, the red circle represents the viral equilibrium point and the green arrows show the evolution of the model.

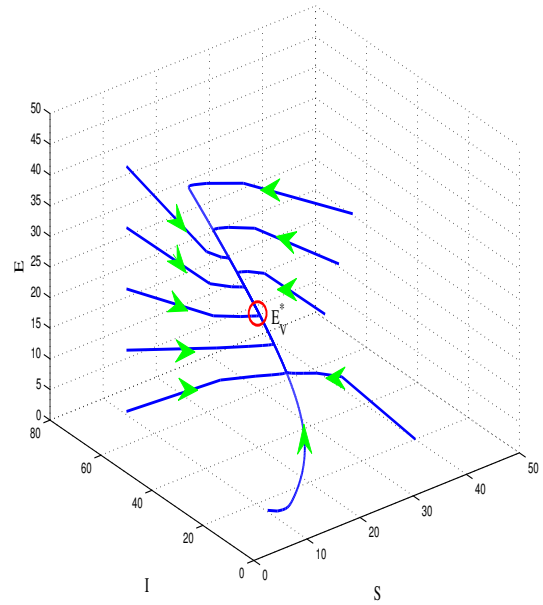


Figure 5. The phase spaces generated by the Euler scheme with $h = 0.87$ after 100 iterations, the RK2 scheme with $h = 0.88$ after 100 iterations and the NSFD scheme with $\Delta t = 1$ and $\varphi(\Delta t) = \frac{1 - e^{-1.8\Delta t}}{1.8}$ after 100 iterations.

From Examples 1 and 2, we see that the standard Euler

and RK2 scheme cannot preserve the positivity and stability of the continuous model for all finite step sizes. A similar comment for the classical Runge-Kutta schemes can be found in some of our previous works [16–22]. Consequently, the advantage of the NSFD scheme over the standard Euler and RK2 schemes is that they can correctly preserve the dynamics of the continuous model regardless of the chosen step sizes. Hence, the NSFD scheme is appropriate and effective to simulate dynamics of the differential equation model over long time periods.

V. CONCLUSIONS AND DISCUSSIONS

In this paper, we have successfully developed the Mickens' methodology to construct a reliable NSFD scheme for the computer virus propagation model (1). The theoretical assertions have been supported by a set of numerical examples. It was proved that the constructed NSFD scheme preserves the positivity, boundedness and asymptotic stability of the continuous model for all finite step sizes. As a result, it provided reliable approximations, meanwhile, some typical standard schemes failed to preserve the essential mathematical features of the continuous model; consequently, they generated numerical solutions which are not only unstable but also negative.

In the near future, we will extend the approach and obtained results in this paper to construct dynamically consistent NSFD schemes for ordinary and fractional-order differential equation models arising in science and technology. In addition, the construction of high-order NSFD schemes for the computer virus propagation models will be considered.

REFERENCES

- [1] U. M. Ascher and L. R. Petzold, *Computer methods for ordinary differential equations and differential-algebraic equations*. Siam, 1998, vol. 61.
- [2] G. Wanner and E. Hairer, *Solving ordinary differential equations II*. Springer Berlin Heidelberg, 1996, vol. 375.
- [3] G. Wanner, E. Hairer, and S. P. Norsett, *Solving ordinary differential equations I*. Springer Berlin Heidelberg, 1990.
- [4] R. E. Mickens, *Nonstandard Finite Difference Schemes: Methodology and Applications*. World Scientific, 2020.
- [5] —, *Advances in the applications of nonstandard finite difference schemes*. World Scientific, 2005.
- [6] —, "Nonstandard finite difference schemes for differential equations," *Journal of Difference Equations and Applications*, vol. 8, no. 9, pp. 823–847, 2002.
- [7] —, *Applications of nonstandard finite difference schemes*. World Scientific, 2000.
- [8] —, *Nonstandard finite difference models of differential equations*. world scientific, 1994.
- [9] K. Patidar, "Nonstandard finite difference methods: recent trends and further developments," *Journal of Difference Equations and Applications*, vol. 22, no. 6, pp. 817–849, 2016.
- [10] K. C. Patidar, "On the use of nonstandard finite difference methods," *Journal of Difference Equations and Applications*, vol. 11, no. 8, pp. 735–758, 2005.
- [11] O. Adekanye and T. Washington, "Nonstandard finite difference scheme for a tacoma narrows bridge model," *Applied Mathematical Modelling*, vol. 62, pp. 223–236, 2018.
- [12] A. Korpusik, "A nonstandard finite difference scheme for a basic model of cellular immune response to viral infection," *Communications in Nonlinear Science and Numerical Simulation*, vol. 43, pp. 369–384, 2017.
- [13] J. Martín-Vaquero, A. Martín del Rey, A. Encinas, J. Hernández Guillén, A. Queiruga-Dios, and G. Rodríguez Sánchez, "Higher-order nonstandard finite difference schemes for a mseir model for a malware propagation," *Journal of Computational and Applied Mathematics*, vol. 317, pp. 146–156, 2017.
- [14] Y. Yang, J. Zhou, X. Ma, and T. Zhang, "Nonstandard finite difference scheme for a diffusive within-host virus dynamics model with both virus-to-cell and cell-to-cell transmissions," *Computers and Mathematics with Applications*, vol. 72, no. 4, pp. 1013–1020, 2016.
- [15] J. Cresson and F. Pierret, "Non standard finite difference scheme preserving dynamical properties," *Journal of Computational and Applied Mathematics*, vol. 303, pp. 15–30, 2016.
- [16] M. T. Hoang, "Reliable approximations for a hepatitis b virus model by nonstandard numerical schemes," *Mathematics and Computers in Simulation*, vol. 193, pp. 32–56, 2022.
- [17] Q. A. Dang and M. T. Hoang, "Positivity and global stability preserving nsfd schemes for a mixing propagation model of computer viruses," *Journal of Computational and Applied Mathematics*, vol. 374, p. 112753, 2020.
- [18] —, "Positive and elementary stable explicit nonstandard runge-kutta methods for a class of autonomous dynamical systems," *International Journal of Computer Mathematics*, vol. 97, no. 10, pp. 2036–2054, 2020.
- [19] —, "Nonstandard finite difference schemes for a general predator-prey system," *Journal of Computational Science*, vol. 36, p. 101015, 2019.
- [20] —, "Complete global stability of a metapopulation model and its dynamically consistent discrete models," *Qualitative theory of dynamical systems*, vol. 18, no. 2, pp. 461–475, 2019.
- [21] M. T. Hoang and A. Nagy, "Uniform asymptotic stability of a logistic model with feedback control of fractional order and nonstandard finite difference schemes," *Chaos, Solitons & Fractals*, vol. 123, pp. 24–34, 2019.
- [22] Q. A. Dang and M. T. Hoang, "Lyapunov direct method for investigating stability of nonstandard finite difference schemes for metapopulation models," *Journal of Difference Equations and Applications*, vol. 24, no. 1, pp. 15–47, 2018.
- [23] C. Gan, X. Yang, and Q. Zhu, "Propagation of computer virus under the influences of infected external computers and removable storage media," *Nonlinear Dynamics*, vol. 78, no. 2, pp. 1349–1356, 2014.
- [24] C. Gan, X. Yang, W. Liu, Q. Zhu, J. Jin, and L. He, "Propagation of computer virus both across the internet and external computers: a complex-network approach," *Communications in Nonlinear Science and Numerical Simulation*, vol. 19, no. 8, pp. 2785–2792, 2014.
- [25] C. Gan, X. Yang, Q. Zhu, J. Jin, and L. He, "The spread of computer virus under the effect of external computers," *Nonlinear Dynamics*, vol. 73, no. 3, pp. 1615–1620, 2013.
- [26] C. Gan, X. Yang, W. Liu, Q. Zhu, and X. Zhang, "An epidemic model of computer viruses with vaccination and generalized nonlinear incidence rate," *Applied Mathematics and Computation*, vol. 222, pp. 265–274, 2013.

- [27] J. Ren, X. Yang, Q. Zhu, L.-X. Yang, and C. Zhang, "A novel computer virus model and its dynamics," *Nonlinear Analysis: Real World Applications*, vol. 13, no. 1, pp. 376–384, 2012.
- [28] J. R. C. Piqueira and V. O. Araujo, "A modified epidemiological model for computer viruses," *Applied Mathematics and Computation*, vol. 213, no. 2, pp. 355–360, 2009.
- [29] S. Elaydi, *An Introduction to Difference Equations*. Springer, 2005.
- [30] H. K. Khalil, *Nonlinear systems third edition*, 2002, vol. 115.



Hoang Manh Tuan received the Ph.D. degree in Applied Mathematics from Graduate University of Science and Technology, Vietnam Academy of Science and Technology (VAST) in 2021, the M.S in Applied Mathematics in 2015 and the B.S degree in Mathematics in 2012 from VNU University of Science. Currently, he is a lecturer-

researcher at the Department of Mathematics, FPT University. His research interests are the qualitative theory and numerical analysis of differential equations and mathematical methods in information technology.

Email: tuanhm@fpt.edu.vn; hmtuan01121990@gmail.com



Pham Hoai Thu received the M.S in Information Systems from Posts and Telecommunications Institute of Technology in 2019 and the B.S in Informatics from People's Security Academy in 2014. Currently, she is a lecturer at the Faculty of Information Security, People's Security Academy. Her research interests include mathematical

foundations for computer science and database.

Email: phamthuhvan@gmail.com

Nâng cao hiệu quả tra cứu ảnh nhãn hiệu sử dụng cây quyết định và phản hồi liên quan

Nguyễn Việt Anh¹, Vũ Văn Hiệu¹, Nguyễn Thị Trang², Lê Ngọc Minh¹

¹ Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

² Trường Đại học Hạ Long

Tác giả liên hệ: Nguyễn Việt Anh, anhnv@ioit.ac.vn

Ngày nhận bài: 16/09/2021, ngày sửa chữa: 30/10/2021, ngày duyệt đăng: 18/11/2021

Định danh DOI: 10.32913/mic-ict-research-vn.v2021.n2.1012

Tóm tắt: Tra cứu ảnh nhãn hiệu là công việc quan trọng để bảo vệ bản quyền của nhãn hiệu đã đăng ký. Nhãn hiệu phân biệt ký hiệu hoặc thông điệp của một cá nhân, tổ chức thương mại hoặc pháp nhân được ủy quyền để nhận ra các sản phẩm hoặc dịch vụ từ một nguồn duy nhất. Thực tế, việc xác định sự tương đồng về mặt thị giác bởi con người và xác định sự tương đồng giữa các ảnh nhãn hiệu trên máy tính vẫn là một nhiệm vụ đầy thách thức. Trong bài báo này, chúng tôi đề xuất một phương pháp để tra cứu ảnh nhãn hiệu dựa trên đặc điểm màu sắc, hình dạng và kết cấu của ảnh nhãn hiệu. Các đặc trưng này được kết hợp sử dụng để xây dựng một bộ độ đo khoảng cách có trọng số được sử dụng để đo độ tương tự và làm kết quả để xây dựng bộ phân lớp sử dụng cây quyết định. Để cải thiện kết quả phân lớp, chúng tôi đề xuất sử dụng phản hồi liên quan để tăng cường số mẫu huấn luyện, hiệu chỉnh trọng số tương tự và truy vấn theo quan tâm của người dùng. Kết quả thực nghiệm trên tập dữ liệu Flickr Logos 27 [1] chứng minh tính hiệu quả của phương pháp đề xuất.

Từ khóa: Nhãn hiệu, tra cứu ảnh dựa vào nội dung, phản hồi liên quan, cây quyết định.

Title: Improving Image Trademark Retrieval Effectiveness using Decision Trees and Relevance Feedback.

Abstract: Trademark images searching and comparing play important role in copyright protection of registered brands. The registered brands can help to distinguish the symbols or messages of individuals, commercial or authorized legal entities so that their original products or services can be easily recognized. In practices, identifying the visual similarity of trademark images by human is still a very challenging task. To deal with the problem, we propose in this paper a method for trademark images retrieval based on their colors, shapes and texture features. Specifically, we propose a similarity measure that take into account these features with weights. A classification model is built based on this similarity measure and the decision tree method. To improve the classifying results, the relevant feedback technique is used to improve the selection of training examples and to adjust of the similarity weights according to user's needs. Experimental results on the Flickr Logos 27 dataset [1] prove the effectiveness of our proposed method.

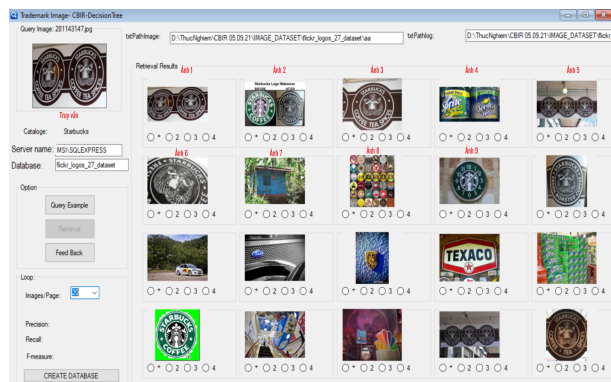
Keywords: Trademarks, content-based image retrieval, relevance feedback, decision trees.

I. GIỚI THIỆU

Nhãn hiệu là một biểu tượng để nhận biết hoặc văn bản liên quan để xác định sản phẩm hoặc dịch vụ của một cá nhân, một tổ chức kinh doanh hoặc một tổ chức pháp nhân với những sản phẩm hoặc dịch vụ của họ. Các nhãn hiệu đã đăng ký được xem như một dạng tài sản hợp pháp và cần được bảo vệ khỏi vi phạm bản quyền và vi phạm nhãn hiệu. Để bảo vệ và hợp pháp hóa nhãn hiệu của mình, chủ sở hữu phải đăng ký nhãn hiệu của họ tại các cơ quan cấp bằng sáng chế ở nhiều quốc gia. Hơn 100 triệu công ty được biết là đang tồn tại trên thị trường toàn cầu và nhiều công ty trong số đó sở hữu ít nhất một nhãn hiệu đã đăng

ký. Theo Tổ chức Sở hữu Trí tuệ thế giới [2], 3 triệu đăng ký nhãn hiệu tồn tại trên toàn thế giới và các đơn đăng ký nhãn hiệu tiếp tục tăng với tỷ lệ 6-8% trong những năm gần đây.

Khi áp dụng một nhãn hiệu mới, cần phải đảm bảo rằng nhãn hiệu mới không bắt chước hoặc đủ khác biệt với các nhãn hiệu hiện có. Ở hầu hết các nước phát triển, các tổ chức như văn phòng sáng chế có trách nhiệm bảo vệ nhãn hiệu khỏi bị xâm phạm. Để tránh các vi phạm khác nhau, họ loại trừ việc đăng ký các nhãn hiệu gần như trùng lặp hoặc bị bắt chước có chủ ý bằng cách kiểm tra nhãn hiệu trong cơ sở dữ liệu hoặc bằng cách sử dụng hệ thống theo



Hình 1. Ví dụ nhãn hiệu tương tự

cách thủ công. Số lượng đăng ký ô ạt đã làm hoạt động thủ công và tự động bị quá tải và làm giảm chất lượng dịch vụ của các văn phòng cấp bằng sáng chế, điều này để lại một kẽ hở cho các hành vi vi phạm nhãn hiệu. Điều tồi tệ hơn, hai nhãn hiệu tương tự được đăng ký nhằm sẽ làm tăng sự phức tạp của việc xử lý tranh chấp pháp lý giữa các chủ sở hữu. Để giảm bớt gánh nặng cho các văn phòng cấp bằng sáng chế, cần phải có một hệ thống tra cứu nhãn hiệu tự động mạnh với các kỹ thuật phân tích ảnh thông minh.

Thực tế, việc tra cứu tất cả các điểm tương đồng về nhãn hiệu một cách hiệu quả là một thách thức vì:

- Sự tương đồng về nội dung nhưng về mặt ngữ nghĩa lại khác nhau, như Ảnh 4 có mức tương tự cao hơn Ảnh 5, Ảnh 6 đối với ảnh truy vấn như trong Hình 1.
- Sự giống nhau mang tính chủ quan chủ yếu do thiếu các tiêu chí rõ ràng để quyết định sự giống nhau. Sự giống nhau về mặt hình ảnh, đặc biệt là trong trường hợp tương tự về nhãn hiệu, bị ảnh hưởng rất nhiều bởi nhiều khía cạnh bao gồm nền tảng giáo dục, tôn giáo, sở thích, v.v.

Một yếu tố quan trọng khác ảnh hưởng đến sự tương đồng là thực tế là số lượng nhãn hiệu hiện có là rất lớn và gia tăng nhanh chóng, đặt ra thách thức cho việc tạo ra nhãn hiệu mới, khác biệt đáng kể để thể hiện cùng một nội dung. Điều này, theo thời gian, có thể dẫn đến sự thay đổi trong việc quyết định sự giống nhau vì chúng ta có thể hết cách để diễn đạt một ý nghĩa.

- Các phương pháp tra cứu ảnh hiện nay, chủ yếu được điều chỉnh để xác định sự giống nhau về các đặc trưng liên quan đến đối tượng, không phải là giải pháp tối ưu cho các vấn đề tra cứu nhãn hiệu, vì hình ảnh nhãn hiệu chủ yếu kết hợp thông tin trừu tượng với nhiều biến đổi và chi tiết khác nhau. Trên thực tế, các hệ thống tra cứu nhãn hiệu cần được trang bị các khả năng trực quan cấp cao như phân nhóm trực quan,

nhận dạng đối tượng, hiểu cảnh/nội dung, v.v. để có thể xử lý các trường hợp như trong Hình 2.



Hình 2. Ví dụ về cách các nguyên tắc Gestalt ảnh hưởng đến nhận thức về nhãn hiệu

1. Các đóng góp chính của nghiên cứu

Trong bài báo này, chúng tôi tập trung vào xác định tương tự về nhãn hiệu trong kết quả tra cứu về mặt tương tự hình ảnh bao gồm các khía cạnh màu sắc, hình dạng và kết cấu và bỏ qua các điểm tương tự về khái niệm/ngữ nghĩa trong các lần phản hồi liên quan). Chúng tôi sử dụng kiến thức phản hồi phân loại liên quan vào xây dựng cây quyết định. Dựa trên các cây quyết định, các suy luận được đưa ra về những ảnh nào mà người dùng muốn xem nhất trong lần lặp lại tra cứu tiếp theo. Chúng tôi chứng minh độ chính xác tra cứu tăng lên sau một hoặc hai lần lặp, yêu cầu người dùng chỉ cung cấp phản hồi về một số ít hình ảnh. Để cải thiện nhanh kết quả tra cứu, chúng tôi đề xuất người dùng chọn các ảnh hiển thị thuộc nhóm liên quan với nhau (gồm 4 nhóm: liên quan với ảnh truy vấn và ba nhóm khác nhau). Chúng tôi cũng cho thấy rằng ứng dụng tra cứu của chúng tôi đủ nhanh để sử dụng trực tuyến.

2. Cấu trúc bài báo

Phần hai chúng tôi giới thiệu các nghiên cứu liên quan đến tra cứu ảnh nhãn hiệu. Phần ba chúng tôi trình bày phương pháp đề xuất. Phần bốn chúng tôi chứng minh tính hiệu quả của phương pháp tra cứu nhãn hiệu trên tập dữ liệu với các phương pháp khác. Cuối cùng, Phần năm là kết luận và hướng nghiên cứu tiếp theo.

II. NGHIÊN CỨU LIÊN QUAN

Trong phần này, chúng tôi thảo luận về các cách tiếp cận hiện tại đối với đánh giá sự tương tự nhãn hiệu, bao gồm cả các phương pháp thủ công.

1. Kiểm tra sự tương tự của nhãn hiệu bằng phương pháp thủ công

Cho đến nay hầu hết các cơ quan cấp bằng sáng chế vẫn dựa vào nỗ lực thủ công để đánh giá tính tương tự của nhãn hiệu. Trong cách tiếp cận hoàn toàn thủ công, trước tiên con người ghi nhớ một nhãn hiệu và sau đó đọc lướt qua toàn bộ bộ sưu tập nhãn hiệu để hy vọng phát hiện ra những điểm tương đồng. Trong cách tiếp cận bán thủ công, trước tiên con người gắn nhãn các nhãn hiệu, truy xuất các nhãn hiệu có cùng nhãn và kiểm tra trực quan sự giống nhau giữa các nhãn hiệu đã tra cứu. Khi được truy vấn về sự tương tự nhãn hiệu, trước tiên nhãn hiệu đó được gắn nhãn các loại, sau đó các nhãn hiệu được đưa vào các danh mục để được truy xuất và sự tương tự được đánh giá bởi một chuyên gia bằng cách sử dụng phương pháp kiểm tra trực quan.

Hệ thống phân loại với cách tiếp cận thủ công, có một số nhược điểm: (i) Quá trình phân loại mang tính chủ quan vì chi tiết của việc gán nhãn có thể phụ thuộc vào người quan sát. (ii) Các danh mục được cố định và không thể mở rộng hay thay đổi. (iii) Không thể mô tả bằng lời tất cả nội dung của nhãn hiệu.

Tóm lại, các phương pháp tiếp cận thủ công, ngay cả các hệ thống ghi nhãn phức tạp, (i) đòi hỏi thời gian lớn (trong trường hợp hoàn toàn thủ công, phải mất 3-4 ngày để một chuyên gia là con người kiểm tra bằng mắt một nhãn hiệu trong số khoảng 1 hàng triệu nhãn hiệu), (ii) khá dễ xảy ra lỗi vì yếu tố con người có liên quan trong quá trình này, (iii) không đáp ứng thực tế đối với hệ thống nhãn hiệu đang phát triển nhanh chóng với rất nhiều các nhãn hiệu mới. Do đó, việc tự động hóa đánh giá tính tương tự nhãn hiệu là rất cần thiết.

2. Kiểm tra sự tương tự của nhãn hiệu bằng các phương pháp tự động

Mặc dù hiện tại các phòng cấp bằng sáng chế vẫn dựa vào các phương pháp thủ công là chủ yếu, các nghiên cứu đã nghiên cứu các phương pháp hoàn toàn tự động để tra cứu sự tương tự nhãn hiệu trong khoảng hai thập kỷ qua.

Trước tiên phải kể đến các kỹ thuật áp dụng đặc trưng toàn cục mức thấp, bao gồm vectơ đặc trưng [3], miêu tả Fourier [4], các moment ảnh [4, 5], cũng như các đặc trưng hình dạng đơn giản hơn và tính toán nhanh như tỷ lệ hình, tính lặp, phép lọc, biến đổi tâm [4], mức xám [6], biểu đồ hướng gradient [5], wavelets [7]. Ngoài các đặc điểm liên quan đến hình dạng và kết cấu, các phương pháp tiếp cận dựa trên đặc điểm màu sắc cũng đã được áp dụng để truy xuất nhãn hiệu [8, 9].

Jiang và cộng sự [10] chỉ ra rằng các bộ mô tả nói trên không kết hợp thông tin hình học của các đối tượng được

trích rút. Các bộ mô tả này sẽ không tốt trong trường hợp các nhãn hiệu khớp với nhau ở từng bộ phận hoặc các nhãn hiệu không liên quan dẫn đến các bộ mô tả toàn cục. Để cải thiện kết quả tra cứu, các kết hợp khác nhau của các đặc trưng này đã được áp dụng. Mặc dù có bằng chứng tương phản [11], việc tích hợp hiệu quả nhiều đặc trưng đã được chứng minh là cải thiện hiệu suất tra cứu [5].

Để cải thiện kết quả truy xuất và so khớp từng phần, một cách tiếp cận là phân đoạn nhãn hiệu thành một số đối tượng và đối sánh nhãn hiệu bằng cách so sánh các phần mô tả của chúng [12, 13].

Mô tả nhãn hiệu với các đặc điểm toàn cục được trích rút từ toàn bộ nhãn hiệu hoặc các bộ phận của nó là hiệu quả về thời gian và bộ nhớ. Tuy nhiên, những phương pháp này bỏ qua thông tin cục bộ, thông tin này có thể quan trọng trong việc giải quyết các vấn đề vi phạm một phần. Để bao gồm thông tin cục bộ giải quyết một phần đối sánh, các phương pháp dựa trên điểm chính như SIFT [14, 15], góc Harris [9], v.v. đã được thử nghiệm.

3. Các nghiên cứu liên quan: Phát hiện và xác định nhãn hiệu

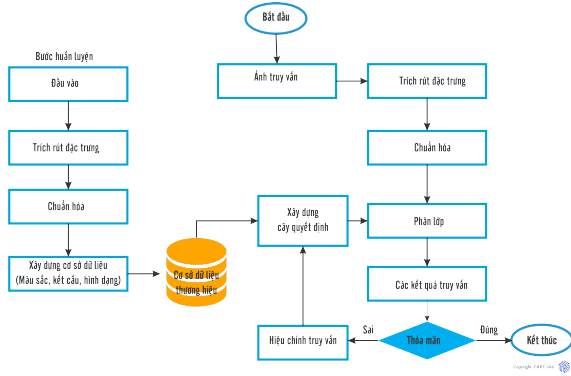
Phát hiện và xác định nhãn hiệu là hai vấn đề liên quan đến việc tra cứu nhãn hiệu. Phát hiện nhãn hiệu là bài toán tìm kiếm tất cả các nhãn hiệu trong cảnh. Mặt khác, xác định nhãn hiệu quan tâm đến việc tìm một nhãn hiệu cụ thể trong ảnh, Kesidis và cộng sự [16] khảo sát rất chi tiết về những vấn đề này.

Kesidis và cộng sự [16] chỉ ra sự khác biệt giữa sự giống nhau và sự trùng khớp là rất quan trọng đối với việc tra cứu nhãn hiệu, vì hầu hết các phương pháp tra cứu ảnh được thiết kế để đối sánh chính xác hơn là phát hiện sự giống nhau. Ví dụ, các phương pháp dựa trên điểm khóa dựa vào việc phát hiện và khớp các điểm khóa giống nhau. Tuy nhiên, trong một vấn đề tương tự, hai nhãn hiệu có thể không sở hữu chung bất kỳ điểm chính nào.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Phát biểu bài toán tra cứu nhãn hiệu như sau: Cho một cơ sở dữ liệu ảnh gồm M ảnh. Mỗi ảnh I trong cơ sở dữ liệu và truy vấn Q là nhãn hiệu cần được tra cứu để tìm ra sự tương tự với các nhãn hiệu khác được miêu tả bởi tập T đặc trưng ($t = 1, \dots, T$), mỗi ảnh I đồng nhất với một tập đặc trưng trong cơ sở dữ liệu đặc trưng $\mathbb{E}$. Độ tương tự tương ứng của truy vấn Q với ảnh I trong cơ sở dữ liệu đặc trưng là $D_Q^t(I)$, $t = 1, \dots, T$. Một tập T độ đo khoảng cách của mỗi ảnh I và truy vấn Q được xác định là $D_Q(I) = \{D_Q^t(I)\} = \{D^t(Q^t, I^t)\}_{1 \leq t \leq T}$. Sau khi tính độ tương tự của mỗi ảnh I với truy vấn Q tạo thành không gian tương tự $\mathbb{S}_Q = \{(I, D_Q(I)) / I \in \mathbb{E}\}$

Như vậy, bài toán tra cứu ảnh nhãn hiệu được xác định như việc tìm các ảnh phù hợp trên một cơ sở dữ liệu đặc trưng ảnh. Với cách tiếp cận sử dụng phản hồi liên quan trên bộ phân lớp cây quyết định, mô hình hệ thống đề xuất (RFTrademarkTree) được miêu tả như Hình 3.



Hình 3. Sơ đồ hệ thống đề xuất RFTrademarkTree

Sau khi các đặc trưng được trích rút miêu tả màu sắc, hình dạng và kết cấu sẽ được chuẩn hóa. Phép chuẩn hóa thường được sử dụng là chuẩn hoá *min-max* hoặc chuẩn hoá Gauss.

1. Chuẩn hóa Gauss

Chuẩn hóa Gauss hay gọi là 3σ , chuẩn hóa này khá hiệu quả trong nghiên cứu như [17, 18]. Chuẩn hoá thành phần thứ k của bộ đặc trưng t_{ij} như sau:

$$\{t_{ijk}\} \mapsto \{\tilde{t}_{ijk}\}, \forall k, \tilde{t}_{ijk} = \frac{t_{ijk} - \mu_k}{3\sigma_k}, \quad (1)$$

trong đó $\mu_k \stackrel{\text{def}}{=} \text{mean}(\{t_{ijk}\})$, $\sigma_k \stackrel{\text{def}}{=} \text{var}(\{t_{ijk}\})$.

Giá trị khoảng cách tương ứng của các đặc trưng riêng được chuẩn hóa như sau:

$$\tilde{D}^{t_{ij}}(I_x, I_y) = \frac{D^{t_{ij}}(I_x, I_y) - \mu_{t_{ij}}}{K\sigma_{t_{ij}}}, \quad (2)$$

trong đó phân bố chuẩn $\sigma_{t_{ij}}$ theo bộ đặc trưng t_{ij} là

$$\sigma_{t_{ij}} = \frac{2}{N(N-1)} \sum_{x=1}^{N-1} \sum_{y=x+1}^N [D^{t_{ij}}(I_x, I_y) - \mu_{t_{ij}}]^2, \quad (3)$$

và khoảng cách trung bình $\mu_{t_{ij}}$ giữa các cặp ảnh theo bộ đặc trưng t_{ij} :

$$\mu_{t_{ij}} = \frac{2}{N(N-1)} \sum_{x=1}^{N-1} \sum_{y=x+1}^N D^{t_{ij}}(I_x, I_y). \quad (4)$$

2. Sử dụng cây quyết định và phản hồi liên quan để nâng cao hiệu quả tra cứu nhãn hiệu

Tra cứu nhãn hiệu sử dụng phản hồi liên quan giống như bộ máy phân lớp lấy đầu vào là ảnh truy vấn và danh sách các kết quả đã được người dùng đánh dấu liên quan hoặc lựa chọn phân lớp vào tập huấn luyện. Chu kỳ được lặp lại và tập huấn luyện được bổ sung theo lựa chọn của người dùng. (Lưu ý trong lần lặp đầu tiên, không tồn tại thông tin phản hồi nào như trong Hình 4.)

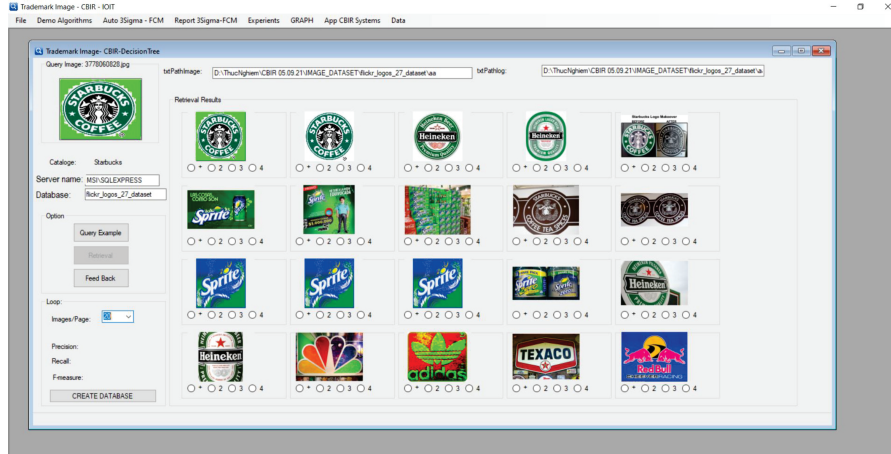
Chúng tôi xem nhiệm vụ phản hồi liên quan như giải pháp xác định dữ liệu huấn luyện (ảnh truy vấn và các ảnh được đánh dấu). Một khái niệm được suy ra từ dữ liệu này và các mẫu huấn luyện được tăng cường sau các lần phản hồi liên quan). Bài toán tra cứu ảnh nhãn hiệu được xem như một bài toán phân lớp: có liên quan và không liên quan.

Ảnh truy vấn tự động được gắn nhãn có liên quan, là tiêu chuẩn về mức độ liên quan để so sánh với các ảnh khác trong cơ sở dữ liệu. K ảnh được phân loại theo mức độ liên quan được thực hiện bởi người dùng. Với các ảnh trả về trong các lần phản hồi có liên quan đến ảnh truy vấn người dùng có thể đưa vào lớp cùng với ảnh truy vấn (lớp đầu + trên Hình 4), các ảnh không liên quan đến ảnh truy vấn, người dùng tự phán xét và đưa vào các lớp 2, 3 hoặc 4. Do đó, một nhóm các mẫu huấn luyện được thiết lập (K ảnh được đánh dấu, cộng với truy vấn). Sau đó, chúng tôi sử dụng một kỹ thuật để học từ dữ liệu huấn luyện này và tạo ra một tập các ảnh được tra cứu cho người dùng bằng cách sử dụng mô hình đã học. Thuật toán được sử dụng phân lớp là cây quyết định phản hồi về mức độ liên quan trên tập kết quả trả về.

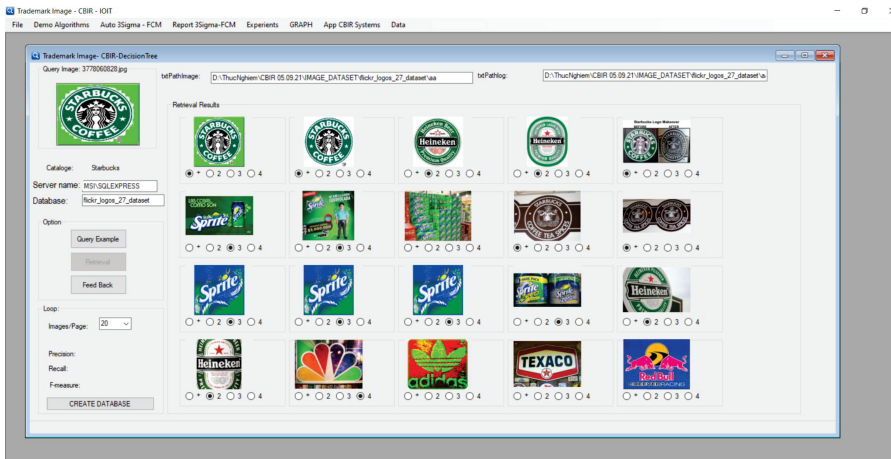
Trong lần lặp đầu tiên, không có thông tin phản hồi nào tồn tại, vì vậy bộ tra cứu thực hiện tra cứu lấy kết quả trả về là K lần cận gần nhất. Người dùng sau đó đánh dấu các ảnh đã tra cứu là có liên quan hoặc không liên quan theo quan tâm. Phản hồi này được chuyển tiếp trở lại hệ thống và lần lặp thứ hai bắt đầu. Trong lần lặp thứ hai, thuật toán được miêu tả với K ảnh có nhãn tương ứng với K mẫu huấn luyện như trong Hình 5, chúng tôi tạo ra một cây quyết định thông qua C4.5. [19].

Cây quyết định là một phương pháp để phân vùng đệ quy một không gian đặc trưng sao cho mỗi phân vùng được gắn nhãn bởi một giá trị lớp duy nhất. Thuật toán tiếp tục thực hiện cắt giảm cho đến khi tất cả các mẫu trong một phân vùng thuộc cùng một lớp; phân vùng sau đó được gắn nhãn với giá trị lớp. Thuật toán sử dụng thông tin phản hồi và phân lớp sử dụng cây quyết định được trình bày trong Thuật toán 2 được trình bày trong mục tiếp theo.

Cụ thể, cây được sử dụng để xác định mỗi vectơ đặc trưng ảnh tới một nút lá có liên quan đến lớp hoặc không



Hình 4. Kết quả trong lần lặp đầu tiên



Hình 5. Lựa chọn gắn lớp liên quan và không liên quan

liên quan. Khi tất cả các mẫu trong cơ sở dữ liệu đã lọc qua cây quyết định, tất cả các mẫu được lọc thành một lá có lớp "có liên quan" được tập hợp thành một danh sách. Từ danh sách này, các mẫu gần nhất với truy vấn được lấy ra bằng cách thực hiện một K không trọng số láng giềng gần nhất trong danh sách.

Trong lần lặp tiếp theo, hoạt động của người tra cứu giống với hoạt động ở lần lặp thứ hai, ngoại trừ việc bây giờ có các ảnh kết quả là phản hồi của người dùng mà từ đó hệ thống sẽ đưa ra cây quyết định (Đánh dấu mức độ liên quan của phản hồi được giữ lại từ lần lặp lại tiếp theo. Việc tăng cường tập huấn luyện dừng lại nếu tập ảnh kết quả phù hợp với yêu cầu người dùng). Thông tin phản hồi được sử dụng để hiệu chỉnh trọng số độ tương tự và truy vấn được trình bày trong Thuật toán 1.

Tương tự, mỗi lần lặp lại cho phép tra cứu học từ ảnh ở với lần lặp trước. Quá trình này tiếp tục cho đến khi người dùng hài lòng với kết quả hoặc cho đến khi người dùng kết

thúc tra cứu. Hình 6 là kết quả tra cứu sau khi sử dụng bộ phân lớp cây quyết định và hiệu chỉnh trọng số.

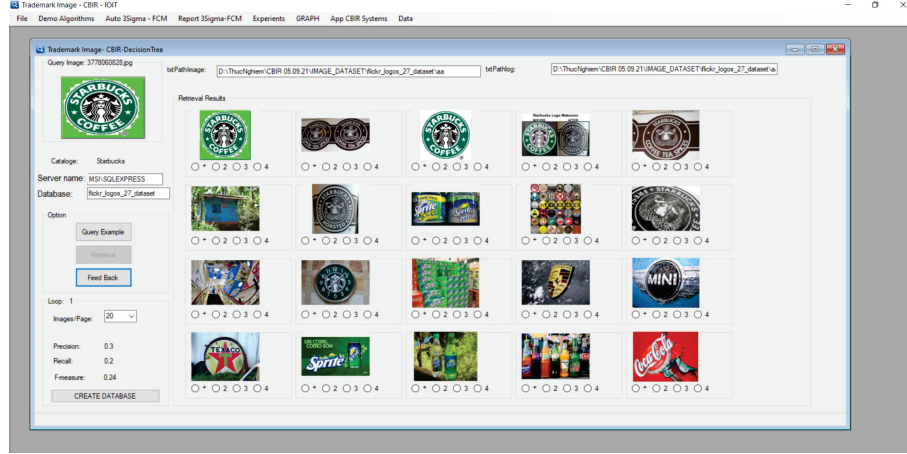
3. Hiệu chỉnh trọng số tính độ tương tự và hiệu chỉnh truy vấn

Công thức tính trọng số được điều chỉnh cho từng đặc trưng theo Công thức (5) và (6), truy vấn được hiệu chỉnh sau lần lặp đầu tiên theo Công thức (7) được xác định như sau:

- Tính giá trị trọng số

$$w_t^{(l+1)} = \max(w_t^{(l)} + \Delta w_t^{(l)}, 0) \quad (5)$$

trong đó l là ký hiệu lần lặp, $\Delta w_t^{(l)} = f\left(\text{score}(I_t^l), \sigma_{t, D_Q^{(l)}}^{(l), NB^+}, \sigma_{t, \|I_t^l\|}^{(l), NB^+}\right) \in (-\infty, +\infty)$,



Hình 6. Kết quả phân lớp và hiệu chỉnh ở lần lặp thứ hai.

hàm f có thể chọn chẳng hạn:

$$f(score(I_i^t), \sigma_{t, D_Q^t(I)}, \sigma_{t, ||I_i^t||}) = \frac{score(I_i^t)}{(\sigma_{t, D_Q^t(I)} \times \sigma_{t, ||I_i^t||})} \quad (6)$$

Truy vấn mới ứng với bộ đặc trưng t được xác định lại như sau:

$$Q_t' \stackrel{\text{def}}{=} \frac{1}{\#listNB_t^*} \sum_{I_i^t \in listNB_t^*} I_i^t \quad (7)$$

trong đó, $listNB_t^{* \text{def}} I_i^t \in listNB_t / |I_i^t - \bar{Q}_t| \leq 3\bar{\sigma}_t$ có sự đồng nhất cao về đặc trưng, khi $\#listNB_t \geq 2$ và $\#listNB_t^* \geq 1$ (đảm bảo có sự thay đổi).

Thuật toán 1 thực hiện hiệu chỉnh trọng số tính độ tương tự.

Bộ truy xuất cây (RFTree) hoạt động như sau.

IV. THỰC NGHIỆM

1. Cơ sở dữ liệu

Tập dữ liệu FlickrLogos-27 [1] là một tập dữ liệu logo có chú thích tương ứng với 27 lớp logo/nhãn hiệu (30 hình ảnh cho mỗi lớp), gồm 4207 ảnh. Chúng tôi chọn ngẫu nhiên, bộ truy vấn bao gồm 270 ảnh. Có năm ảnh cho mỗi lớp trong số 27 lớp được chú thích, gồm 135 ảnh có chứa biểu tượng. Ngoài ra, bộ truy vấn bổ sung 135 ảnh Flickr không được mô tả bất kỳ lớp nào, cho tổng cộng 270 hình ảnh thử nghiệm.

Hình 7 cho thấy các hình ảnh mẫu về nhãn hiệu được sử dụng trong quá trình thực nghiệm. Các đặc trưng mức thấp của ảnh như hình dạng, màu sắc và kết cấu thể hiện nội dung trực quan của ảnh. Việc trích rút đặc trưng hình

Thuật toán 1: Hiệu chỉnh trọng số độ đo khoảng cách.

```

1 Dữ liệu vào:
2  $\{\tilde{E}_i^t\}_{1 \leq i \leq N}$ ; // CSDL đã chuẩn hoá
3 Vector đặc trưng mức thấp  $Q$  của ảnh truy vấn.  $Lmax$  số lần lặp phản hồi,  $K$ : số ảnh thỏa mãn trả về.
4 Dữ liệu ra: Tập trọng số  $\{w_t\}_{1 \leq t \leq T}$ 
5 begin
6   (1) Khởi tạo  $\forall t = \overline{1, T}$ ,  $w_t = \frac{1}{T}$ ;
7   (2) Chuẩn đặc trưng ảnh truy vấn  $Q$  theo chuẩn hóa  $3\sigma$ ;
8   for  $l = 1$  to  $Lmax$  do
9     for Each  $I_i \in \{I_i\}_{1 \leq i \leq N}$ ,  $\forall t = \overline{1, T}$  do
10      (i) Chuẩn hoá độ đo khoảng cách cho bộ đặc trưng  $t$  theo Công thức (2);
11      (ii) Hiệu chỉnh trọng số dựa trên tập ảnh phản hồi liên quan  $listNB_t^+$ ;
12      for each  $I \in NB^{(l)}$ ,  $\forall t = \overline{1, T}$  do
13        if  $I \in listNB_t^+$  then
14          Hiệu chỉnh  $w_t$  theo Công thức (5) và (6);
15        end
16      end
17      (iii) Chuẩn hóa lại trọng số  $w_t$ ,  $\forall t = \overline{1, T}$ ,
18         $w_t = w_t / \sum_{t=1}^T w_t$ ;
19    end
20 end
21 Return: Kết thúc thu được  $\{w_t\}_{1 \leq t \leq T}$ 

```

dạng, màu sắc và kết cấu được thực hiện với kết quả là vectơ đặc trưng ảnh nhãn hiệu.

2. Các miêu tả ảnh

Trong thực nghiệm chúng tôi biểu diễn ảnh với ba miêu tả đặc trưng màu sắc, kết cấu và hình dạng với tám bộ đặc trưng mức thấp (xem Bảng I). Những đặc trưng này

Thuật toán 2: Thuật toán phản hồi liên quan sử dụng cây quyết định (RFTrademarkTree)

```

1 Dữ liệu vào:  $\{E_i^t\}_{1 \leq i \leq N}$ ; // CSDL đã chuẩn hoá
2  $\{\sigma_{i,c}^{(D)}\}_{1 \leq c \leq C_i^{(D)}}$ ; // Độ lệch chuẩn theo đặc
   trung và khoảng cách
3 Dữ liệu ra: Ảnh hoặc tập ảnh thoả mãn người dùng
4 begin
5   Khởi tạo:
6   (i)  $NB^+ = listNB^+ = \{Q\}$  gán nhãn 0;
7   (ii)  $NB^- = listNB^- = \{I_j\}$  gán nhãn 1; // ( $I_j$ 
   là ảnh người dùng lựa chọn không
   cùng khái niệm với ảnh truy vấn)
8 ; while người dùng chưa thoả mãn do
9   (1) Xây dựng cây quyết định
    $DTRF(Train_{(E,Q)})$ ; //  $E, Q$  Tính độ
   tương tự của đặc trưng truy vấn
   với mỗi đặc trưng trong CSDL đặc
   trưng ảnh huấn luyện
10  (2) Xây dựng tập dữ liệu kiểm tra  $Test_{(E,Q)}$ ;
11  (3) Dự báo phân lớp cho mỗi ảnh thuộc tập
    $Test$ ;
12  (4) Sắp xếp các ảnh trong tập kiểm tra theo phân
   lớp;
13  (5) Lấy ra tập ảnh có độ tương tự nhất;
14  begin
15    (i)  $NB \leftarrow$  Danh sách Top  $k$  ảnh trong tập
    $Test$ ;
16    (ii) Người dùng đánh giá các ảnh theo tiêu
   chí liên quan đưa vào tập  $NB^+$  và không
   liên quan đưa vào tập  $NB^-$ ;
17  end
18  (6) Cập nhật tập huấn luyện:
19  begin
20    (i)  $listNB^+ = listNB^+ \cup NB^+$ ;
21    (ii)  $listNB^- = listNB^- \cup NB^-$ ;
22  end
23  (7) Hiệu chỉnh trọng số và truy vấn theo Thuật
   toán 1;
24  (8) Xác định lại tập  $Test$  theo hiệu chỉnh;
25 end

```

có kết quả rất tốt và sử dụng trong nhiều nghiên cứu liên quan. Các kỹ thuật biểu diễn đặc trưng này được miêu tả trong [20–26]. Sau khi trích rút đặc trưng, mỗi chiều của đặc trưng được chuẩn hoá vào phạm vi $[-1, 1]$ sử dụng phương pháp chuẩn trình bày trong Mục III.1.

3. Các phương pháp cơ sở

Để chứng tỏ kỹ thuật đề xuất hiệu quả, chúng tôi so sánh với bốn kỹ thuật được đề xuất trước đó gồm:

- Kỹ thuật đề xuất bởi Zhang và cộng sự [28], (gọi tắt là Zhang) sử dụng phương pháp học máy SVM, chúng tôi đã cài đặt lại kỹ thuật này và so sánh hiệu năng trên tập dữ liệu [1].
- Kỹ thuật đề xuất bởi Su và cộng sự [29], (gọi tắt là Su) sử dụng phản hồi về mức độ liên quan, dựa trên



Hình 7. Nhãn hiệu mẫu từ Cơ sở dữ liệu FlickrLogos-27

Bảng I
CÁC MIÊU TẢ ẢNH

Miêu tả	Loại	Số chiều	Hàm đo
Lược đồ HSV [20]	Màu	32	$L1$
Tự tương quan màu [21]	Màu	64	$L2$
Các mô men màu [22]	Màu	6	$L2$
Miêu tả đồng mức (GLCM)[27]	Kết cấu	4	$L2$
Các phép lọc Gabor [23]	Kết cấu	48	$L2$
Mô men Wavelet [24]	Kết cấu	40	$L2$
Mẫu nhị phân cục bộ (LBP) [25]	Hình dạng, kết cấu	255	$L2$
Đặc trưng HOG [26]	Hình dạng	3780	$L2$

việc khai thác các mẫu và hiệu chỉnh truy vấn, thực hiện trên tập dữ liệu trên tập dữ liệu [1].

- Kỹ thuật đề xuất bởi Zhao và cộng sự [30] (gọi tắt là Zhao). Nghiên cứu này có sử dụng phản hồi liên quan, phân lớp SVM sử dụng hàm nhân RBF $K(x_i, x_j) = \exp(-\lambda \|x_i - x_j\|^2)$ và được cài đặt lại sử dụng thư viện LibSVM¹, thực hiện trên tập dữ liệu trên tập dữ liệu [1].
- Kỹ thuật đề xuất bởi Tripathi và cộng sự [31] (gọi tắt là Tripathi). Kỹ thuật này sử dụng bước: Bước 1, kết quả tra cứu nhãn hiệu tương tự dựa vào đặc trưng bất biến Zernike. Bước 2, đo độ tương tự của truy vấn với tập kết quả ở Bước 1 sử dụng các đặc trưng SIFT và SURF. Kỹ thuật này cũng được cài đặt lại và sử dụng tập dữ liệu [1].

4. Đánh giá hiệu năng

Để thống nhất và công bằng trong đánh giá độ chính xác, chúng tôi sử dụng độ đo chính (*precision* (Pr)). Giả sử

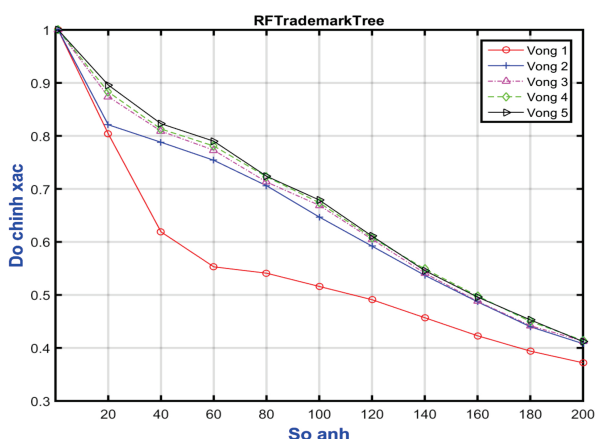
¹<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

số các ảnh liên quan với một truy vấn cụ thể là A , số các ảnh không liên quan cho một truy vấn cụ thể là B , tổng số các ảnh được tra cứu là D và số các ảnh liên quan được tra cứu là E . Độ chính xác là tỉ số giữa số các ảnh liên quan được tra cứu với tổng số các ảnh được tra cứu: $Pr = \frac{E}{D}$.

5. Các kết quả thực nghiệm và luận giải

Hệ thống tra cứu nhãn hiệu được chúng tôi xây dựng (gọi tắt là RFTrademarkTree) có giao diện như trong các Hình 4, Hình 5 và Hình 6. Đầu tiên, hệ thống thiết lập k ảnh có liên quan/tương tự được trả về dựa trên sự tương tự. Từ kết quả trả về, chúng tôi thu kiến thức phản hồi người dùng (phản hồi liên quan). Phản hồi của người dùng được thực hiện dựa trên nhận định thuộc cùng lớp với ảnh truy vấn (người dùng chọn +) hoặc khác lớp với nhau và khác ảnh truy vấn (chọn 2, 3 hoặc 4).

Các ảnh trong tập được phản hồi bởi người dùng sử dụng làm tập huấn luyện, các ảnh trong phần còn lại trong cơ sở dữ liệu được sử dụng làm tập kiểm tra. Quá trình được lặp đi lặp lại đến khi các hình ảnh đầu ra cuối thỏa mãn người dùng nhất. Hình 8 cho thấy hiệu quả của phản hồi liên quan và hiệu chỉnh trọng số sau mỗi lần lặp (từ vòng lặp thứ nhất đến vòng lặp thứ năm) trong 20, 40, 60, 80, 100, 120, 140, 160, 180, 200 ảnh tương tự nhất được trả về trên màn hình kết quả.

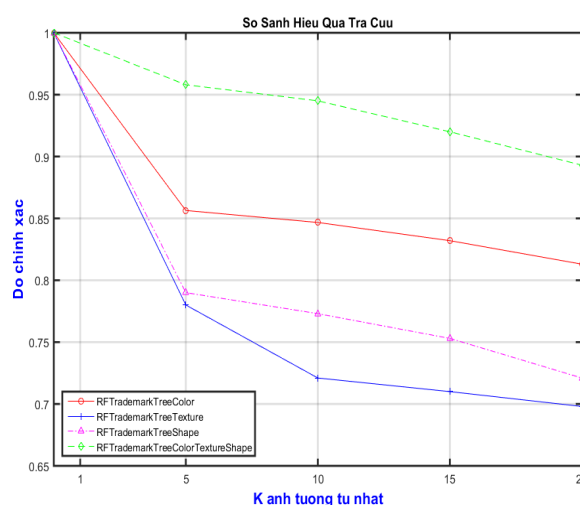


Hình 8. Biểu đồ độ chính xác từng vòng lặp trên số ảnh trả về

Để minh họa kết quả tra cứu nhãn hiệu, Bảng II minh họa kết quả tra cứu của nhãn hiệu tương ứng với các kết quả tìm được sau 5 vòng lặp phản hồi liên quan.

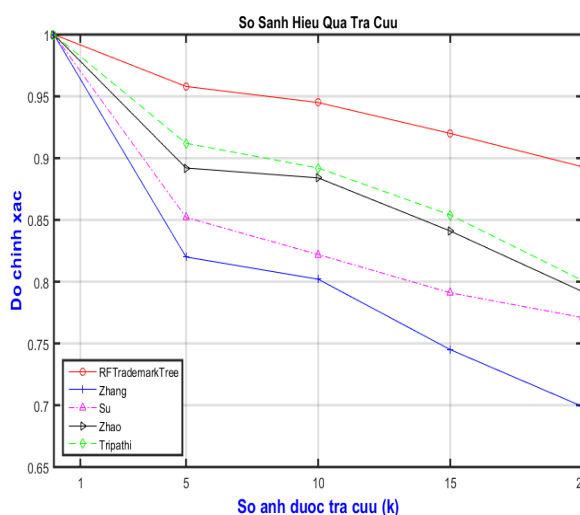
Hệ thống đề xuất sử dụng các đặc trưng miêu tả màu sắc ([21], [22],[20]), kết cấu ([24], [23], [27]) và hình dạng ([26], [25]) cho thấy hiệu quả của sự kết hợp này. Hình 9 so sánh độ chính xác hiệu quả khi kết hợp đặc trưng màu sắc, kết cấu và hình dạng (RFTrademarkTreeColorTextureShape) so các kết quả nếu chỉ sử dụng đơn lẻ đặc trưng

màu sắc (RFTrademarkTreeColor), kết cấu (RFTrademarkTreeTexture) hoặc hình dạng (RFTrademarkTreeShape)



Hình 9. Biểu đồ so sánh hiệu quả sử dụng đặc trưng

Kỹ thuật đề xuất được so sánh độ chính xác tương ứng trên 5, 10, 15 và 20 ảnh tương tự nhất sử dụng cùng một truy vấn được trả về trên giao diện tra cứu cho cả bốn thuật đề xuất bởi Zhang, Su, Zhao và Tripathi. So sánh hiệu quả của kỹ thuật đề xuất được biểu diễn bởi Hình 10.



Hình 10. Biểu đồ so sánh độ chính xác kỹ thuật đề xuất với bốn kỹ thuật

V. KẾT LUẬN VÀ NGHIÊN CỨU TƯƠNG LAI

Trong hệ thống tra cứu nhãn hiệu được đề xuất, mô hình RFTrademarkTree được kết hợp với việc hiệu chỉnh truy vấn. Trong lần lặp lại đầu tiên của phản hồi liên quan,

Bảng II
KẾT QUẢ TRA CỨU ẢNH NHÃN HIỆU ‘3193687005.JPG’

Ảnh truy vấn	Hình ảnh đầu ra có liên quan				
					
					

thông tin phản hồi từ người dùng được cung cấp làm đầu vào cho cây quyết định và được huấn luyện dựa trên các đặc trưng có liên quan và không liên quan. Trong các lần lặp lại phản hồi liên quan tiếp, các biểu diễn đặc trưng này được hiệu chỉnh trong mô hình cây quyết định. Điều này đã cải thiện hiệu suất tra cứu do hiểu biết thông tin về nhu cầu của người dùng nên mang lại ảnh gần với truy vấn và di chuyển ra xa các ảnh không liên quan. Để giảm không gian tìm kiếm, do đó, thời gian cần thiết để thực thi nên cần các kỹ thuật tối ưu. Công việc nghiên cứu trong tương lai hướng đến việc bổ sung thuật toán tối ưu để giảm không gian tìm kiếm để tăng kết quả liên quan.

Kỹ thuật đề xuất bởi Zhang [28] (2001) và Zhao [30] (2012) chưa đạt hiệu quả cao do mỗi chỉ sử dụng SVM để phân lớp, trong khi tập huấn luyện trong bài toán tra cứu ảnh là rất nhỏ và phụ thuộc vào lựa chọn của người dùng trong kết quả trả về. Kỹ thuật đề xuất bởi Su [29] (2010) đã sử dụng hiệu chỉnh trọng số để phù hợp với nhu cầu người dùng, tuy nhiên nghiên cứu này chưa có kết quả tốt do chưa giảm tập kết quả để hiển danh sách ảnh tương tự nhất so với các phương pháp có sử dụng bộ phân lớp. Kỹ thuật đề xuất bởi Tripathi và cộng sự [31] (2021) không chú trọng vào việc tìm kiếm nhãn hiệu tương tự trong màu sắc, chú trọng cao vào tìm kiếm sự tương tự về hình dạng nhãn hiệu nên chưa đạt được kết quả tốt.

Kỹ thuật đề xuất bởi chúng tôi đã có kết quả tốt hơn do kết hợp đặc trưng màu sắc kết cấu, hình dạng. Ngoài ra, kỹ thuật đề xuất sử dụng hiệu chỉnh trọng số để phù hợp với yêu cầu người dùng và sử dụng cây quyết định trong phân loại nhãn hiệu. Trong nghiên cứu tương lai, chúng tôi sẽ thử nghiệm tra cứu nhãn hiệu có sử dụng đặc trưng được trích rút trong học sâu để cải thiện tốt hơn kết quả tra cứu.

LỜI CẢM ƠN

Chúng tôi xin cảm ơn đề tài mã số CS21.13 của Viện CNTT, Viện Hàn Lâm Khoa học và Công nghệ Việt Nam đã hỗ trợ nghiên cứu này.

TÀI LIỆU THAM KHẢO

- [1] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis, “Scalable triangulation-based logo recognition,” in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, 2011, pp. 1–7.
- [2] W. I. P. Organization, *WIPO Patent Report-Statistics on Worldwide Patent Activities (2007)*. WIPO, 2007, vol. 931.
- [3] T. Kato, “Database architecture for content-based image retrieval,” in *image storage and retrieval systems*, vol. 1662. International Society for Optics and Photonics, 1992, pp. 112–123.
- [4] J. P. Eakins, J. D. Edwards, K. J. Riley, and P. L. Rosin, “Comparison of the effectiveness of alternative feature sets in shape retrieval of multicomponent images,” in *Storage and Retrieval for Media Databases 2001*, vol. 4315. International Society for Optics and Photonics, 2001, pp. 196–207.
- [5] A. K. Jain and A. Vailaya, “Shape-based retrieval: A case study with trademark image databases,” *Pattern recognition*, vol. 31, no. 9, pp. 1369–1390, 1998.
- [6] J.-K. Wu, C.-P. Lam, B. M. Mehtre, Y. J. Gao, and A. D. Narasimhalu, “Content-based retrieval for trademark registration,” in *Representation and Retrieval of Visual Media in Multimedia Systems*. Springer, 1996, pp. 69–91.
- [7] G. Ciocca and R. Schettini, “Content-based similarity retrieval of trademarks using relevance feedback,” *Pattern Recognition*, vol. 34, no. 8, pp. 1639–1655, 2001.
- [8] M. Rusinol, D. Aldavert, D. Karatzas, R. Toledo, and J. Lladós, “Interactive trademark image retrieval by fusing semantic and visual content,” in *European Conference on Information Retrieval*. Springer, 2011, pp. 314–325.
- [9] A. Zeggari, F. Hachouf, and S. Foufou, “Trademarks recognition based on local regions similarities,” in *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*. IEEE, 2010, pp. 37–40.
- [10] H. Jiang, C.-W. Ngo, and H.-K. Tan, “Gestalt-based feature similarity measure in trademark database,” *Pattern recognition*, vol. 39, no. 5, pp. 988–1001, 2006.
- [11] J. P. Eakins, K. J. Riley, and J. D. Edwards, “Shape feature matching for trademark image retrieval,” in *International conference on Image and video retrieval*. Springer, 2003, pp. 28–38.
- [12] S. Alwis and J. Austin, “Trademark image retrieval using multiple features,” in *Challenge of Image Retrieval*, 1999, pp. 1–11.
- [13] W. H. Leung and T. Chen, “Trademark retrieval using contour-skeleton stroke classification,” in *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 2.

- IEEE, 2002, pp. 517–520.
- [14] P. Kochakornjarupong, “Trademark image retrieval by local features,” Ph.D. dissertation, University of Glasgow, 2011.
- [15] C.-H. Wei, Y. Li, W.-Y. Chau, and C.-T. Li, “Trademark image retrieval using synthetic features for describing global shape and interior structure,” *Pattern Recognition*, vol. 42, no. 3, pp. 386–394, 2009.
- [16] A. Kesidis and D. Karatzas, “Logo and trademark recognition,” in *Handbook of Document Image Processing and Recognition*, 2014, pp. 591–646.
- [17] Y. Rui, T. S. Huang, and S. Mehrotra, “Content-based image retrieval with relevance feedback in mars,” in *Image Processing, 1997. Proceedings., International Conference on*, vol. 2. IEEE, 1997, pp. 815–818.
- [18] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, “Relevance feedback: a power tool for interactive content-based image retrieval,” *IEEE Transactions on circuits and systems for video technology*, vol. 8, no. 5, pp. 644–655, 1998.
- [19] S. L. Salzberg, “C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993,” 1994.
- [20] M. J. Swain and D. H. Ballard, “Color indexing,” *International journal of computer vision*, vol. 7, no. 1, pp. 11–32, 1991.
- [21] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, “Image indexing using color correlograms,” in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. IEEE, 1997, pp. 762–768.
- [22] M. A. Stricker and M. Orengo, “Similarity of color images,” in *IS&T/SPIE’s Symposium on Electronic Imaging: Science & Technology*. International Society for Optics and Photonics, 1995, pp. 381–392.
- [23] T. S. Lee, “Image representation using 2d gabor wavelets,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 18, no. 10, pp. 959–971, 1996.
- [24] P. Hiremath, S. Shivashankar, and J. Pujari, “Wavelet based features for color texture classification with application to cbir,” *International Journal of Computer Science and Network Security*, vol. 6, no. 9A, pp. 124–133, 2006.
- [25] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.
- [26] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, vol. 1. Ieee, 2005, pp. 886–893.
- [27] B. Pathak and D. Barooah, “Texture analysis based on the gray-level co-occurrence matrix considering possible orientations,” *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 2, no. 9, pp. 4206–4212, 2013.
- [28] L. Zhang, F. Lin, and B. Zhang, “Support vector machine learning for image retrieval,” in *Image Processing, 2001. Proceedings. 2001 International Conference on*, vol. 2. IEEE, 2001, pp. 721–724.
- [29] J.-H. Su, W.-J. Huang, S. Y. Philip, and V. S. Tseng, “Efficient relevance feedback for content-based image retrieval by mining user navigation patterns,” *IEEE transactions on knowledge and data engineering*, vol. 23, no. 3, pp. 360–372, 2010.
- [30] L. Zhao, J. Tang, X. Yu, Y. Li, S. Mi, and C. Zhang, “Content-based remote sensing image retrieval using image multi-feature combination and svm-based relevance feedback,” in *Recent Advances in Computer Science and Information Engineering*. Springer, 2012, pp. 761–767.

- [31] R. Tripathi and S. Agrawal, “A shape and texture features fusion to retrieve similar trademark image material,” in *IOP Conference Series: Materials Science and Engineering*, vol. 1116, no. 1. IOP Publishing, 2021, p. 012026.

SƠ LƯỢC VỀ TÁC GIẢ

Nguyễn Việt Anh



Tốt nghiệp Đại học Kyoto, nhận bằng tiến sĩ năm 2012.

Hiện là Nghiên cứu viên chính của Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Lĩnh vực nghiên cứu: Học máy, khai phá dữ liệu và phân tích mạng xã hội.

Email: anhnv@ioit.ac.vn

Vũ Văn Hiệu



Nhận bằng tiến sĩ Cơ sở toán học cho tin học tại Học Viện Khoa học Công nghệ, Viện hàn lâm Khoa học và Công nghệ Việt Nam năm 2017.

Hiện là nghiên cứu viên tại Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Lĩnh vực nghiên cứu: Xử lý ảnh và nhận dạng, ứng dụng trí tuệ nhân tạo, học máy.

Email: vvhiu@ioit.ac.vn

Nguyễn Thị Trang



Tốt nghiệp trường Đại học Khoa học- Đại học Thái Nguyên năm 2019, đang làm Nghiên cứu sinh tại Học viện Khoa học Công nghệ - Viện hàn lâm Khoa học và Công nghệ Việt Nam.

Hiện là Phó Giám đốc trung tâm ngoại ngữ - tin học, trường Đại học Hạ Long.

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, học sâu. Một số ứng dụng đã triển khai gần đây: Phần mềm quản lý trung tâm ngoại ngữ - tin học, phần mềm học trực tuyến.

Email: nguyenthitrang.daihochalong@gmail.com

Phạm Ngọc Minh



Nhận bằng tiến sĩ Kỹ thuật điều khiển và tự động hóa tại Học Viện Khoa học Công nghệ, Viện hàn lâm Khoa học và Công nghệ Việt Nam năm 2021.

Hiện là nghiên cứu viên chính tại Phòng Điều khiển và Hệ thống nhúng, Viện Công nghệ thông tin.

Lĩnh vực nghiên cứu: Hệ thống nhúng, điều khiển quá trình, mạng truyền thông công nghiệp, mạng không dây băng thông

rộng, điều khiển robot, hình ảnh tiên tiến.

Email: pnminh@ioit.ac.vn

Quá trình tiến hóa sinh kháng thuốc của vi-rút HIV với cây đột biến di truyền theo mô hình Markov

Nguyễn Văn Thế¹, Tạ Văn Nhân², Nguyễn Thị Kim Duyên¹, Trịnh Mai Phương¹, Nguyễn Thị Hồng Minh¹

¹ Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội, Hà Nội

² Công ty LOBI Việt Nam, Hà Nội

Tác giả liên hệ: Tạ Văn Nhân, tavannhan@gmail.com

Ngày nhận bài: 23/09/2021, ngày sửa chữa: 29/10/2021, ngày duyệt đăng: 15/11/2021

Định danh DOI: 10.32913/mic-ict-research-vn.v2021.n2.1014

Tóm tắt: Trong bài báo này, chúng tôi dự đoán quá trình tiến hóa của vi-rút HIV qua 14 đột biến kháng thuốc trong phác đồ điều trị sử dụng thuốc Efavirenz bằng mô hình Markov ẩn và cây đột biến di truyền. Với dữ liệu mới gồm 396 bệnh nhân trên cơ sở dữ liệu kháng thuốc HIV của trường đại học Stanford, chúng tôi tiến hành kiểm định giả thiết và nhận thấy dữ liệu phù hợp để đưa vào mô hình tính toán. Phần thực nghiệm cho thấy thuật toán EM dùng để ước lượng và tối ưu tham số khi áp dụng vào mô hình có tốc độ hội tụ nhanh. Hơn nữa, dựa vào các tham số sau khi tối ưu, chúng tôi cũng xác định được thứ tự xuất hiện của các đột biến theo thời gian trong đó đột biến K103N xuất hiện sớm nhất.

Từ khóa: Mô hình Markov ẩn, thuật toán EM, cây đột biến di truyền, thuốc Efavirenz.

Title: Evolution of Drug Resistance of HIV Virus with Mutagenetic Tree according to Markov Model

Abstract: In this paper, we predict the evolution of 14 mutations associated with the HIV resistance to Efavirenz using mutagenetic tree hidden Markov model. With new data of 396 patients from the HIV drug resistance database by Stanford University, we test statistical assumptions and found this data set significant for further modeling. Model results show that applying EM algorithm to estimate and optimize parameters has fast convergence. Furthermore, based on optimized parameters, we determine the occurrence order of mutations over time in which the K103N mutation appeared earliest.

Keywords: Hidden markov model, EM algorithm, mutagenetic tree, efavirenz.

I. MỞ ĐẦU

Luận thuyết trung tâm của Crick đã cho thấy một sơ đồ tổng hợp nên Protein bắt đầu bằng quá trình phiên mã từ DNA thành RNA [1]. Không lâu sau đó, các nhà khoa học đã khám phá ra một quá trình mà ban đầu tưởng như mâu thuẫn với giáo điều của luận thuyết trung tâm đó là quá trình phiên mã ngược (Retrotranscription) [2]. Đối với quá trình này, các RNA thông tin (mRNA) được dùng làm khuôn để tạo ra sợi đơn DNA bổ sung (cDNA) giống như bản khuôn đã phiên mã ra nó. Một số loại vi-rút có khả năng phiên mã ngược (Retroviruses) do sở hữu men phiên mã ngược (Reverse Transcriptase, viết tắt là RT). Chẳng hạn như vi-rút T-lymphotropic gây bệnh bạch cầu ở người (HTLV) [3], hay vi-rút gây suy giảm miễn dịch ở người (Human Immunodeficiency Virus, viết tắt là HIV) [4]. Hướng tiếp cận chính để điều trị bệnh gây ra bởi các Retroviruses là phát triển các loại thuốc gây ức chế men

phiên mã ngược của chúng. Trong bài báo này, chúng tôi tập trung nghiên cứu đối với vi-rút HIV, một loại vi-rút nguy hiểm phá hủy tế bào lympho CD4+, làm giảm khả năng miễn dịch qua trung gian tế bào, tăng nguy cơ nhiễm trùng và ung thư ở người mang vi-rút. Hiện tại, có rất nhiều loại thuốc khác nhau được sử dụng kết hợp trong các phác đồ điều trị HIV. Trong đó có hai nhóm chính là: (i) nhóm ức chế phiên mã ngược nucleoside (Nucleoside Reverse Transcriptase Inhibitors, viết tắt là NRTIs); và (ii) nhóm ức chế phiên mã ngược không Nucleoside (Non-Nucleoside Reverse Transcriptase Inhibitors, viết tắt là NNRTIs)¹. Mặc dù đã trải qua nhiều năm với vô vàn nỗ lực, nhưng cho đến nay loài người vẫn chưa tìm ra phương pháp đặc trị vi-rút HIV. Nguyên nhân gây khó khăn được xác định là trong quá trình điều trị, vi-rút xuất hiện những đột biến theo thời

¹ <https://www.msmanuals.com/professional/infectious-diseases/human-immunodeficiency-virus-hiv/drug-treatment-of-hiv-infection>

gian có khả năng kháng lại các thuốc ức chế Protein.

Để quá trình điều trị HIV hiệu quả hơn, các nhà khoa học đã tiến hành các nghiên cứu nhằm dự đoán các thời điểm xảy ra các đột biến kháng thuốc trong các phác đồ sử dụng thuốc khác nhau trên các dữ liệu thu được từ quá trình điều trị. Nghiên cứu của Niko Beerenwinkel và các cộng sự vào năm 2007 đã chỉ ra quá trình tiến hóa của vi-rút HIV qua 7 đột biến kháng thuốc điển hình [5]. Phương pháp chính dựa vào cây đột biến di truyền và mô hình Markov ẩn. Ngoài ra, việc xây dựng cây đột biến cũng là một quá trình quan trọng để xác định quá trình tiến hóa của vi-rút. Dữ liệu cho phân tích khi đó gồm 163 bệnh nhân với 3350 nhân bản trình tự (Clones) được lấy từ ba nghiên cứu về liệu pháp kết hợp sử dụng thuốc Efavirenz (DMP 266-003, -004, -005) [6]. Tiếp tục những kết quả nghiên cứu của Niko Beerenwinkel, với mục tiêu thực nghiệm trên các bộ dữ liệu mới làm cơ sở cho phát triển phương pháp, trong nghiên cứu này, chúng tôi tiến hành dự đoán quá trình tiến hóa của vi-rút HIV qua 14 đột biến kháng thuốc với dữ liệu gồm 396 bệnh nhân được thu thập trên cơ sở dữ liệu kháng thuốc HIV của trường Đại học Stanford². Với dữ liệu từ năm 2005 đến nay, nhiều đột biến kháng thuốc mới xuất hiện cũng như nhiều loại thuốc mới ra đời. Để thực hiện mô hình, chúng tôi đã tiến hành kiểm định lại ba giả thiết của Niko Beerenwinkel và nhận thấy chúng vẫn đúng với bộ dữ liệu lớn hơn và mới hơn. Mặc dù, quá trình tiến hóa của vi-rút HIV vẫn được phân tích dựa trên việc sử dụng thuốc Efavirenz của nhóm NNRTIs, tuy nhiên với bộ dữ liệu mới này, cây đột biến được xây dựng lại với các nhánh lớn hơn, số các phụ thuộc giữa các nút đại diện cho các đột biến cũng nhiều hơn. Với giả thiết các đột biến xuất hiện theo thời gian tuân theo quá trình Poisson, chúng tôi cũng đã đưa ra dự đoán các thời điểm xuất hiện các đột biến kháng thuốc trong các phác đồ điều trị có sử dụng Efavirenz. Trong các phần tiếp theo của bài báo, chúng tôi sẽ trình bày về bộ dữ liệu HIV cho nghiên cứu trong mục II, về mô hình Markov trong mục III. Các nội dung tính toán thực nghiệm được trình bày trong mục IV, các kết quả được bàn luận trong mục V, cuối cùng mục VI là kết luận.

II. DỮ LIỆU

1. Dữ liệu HIV

Dữ liệu về tập hợp trình tự các nhân bản của HIV theo thời gian thu được từ 25 nghiên cứu lâm sàng với các phác đồ điều trị kết hợp trong đó có sử dụng thuốc Efavirenz. Các nghiên cứu này được thực hiện trong thời gian từ năm 1998 tới 2018 với 416 bệnh nhân, thông tin của mỗi bệnh nhân gồm các nhân bản trình tự Protein (Clones) tại các thời điểm khác nhau, ở mỗi Clones đã xác định vị trí của

các đột biến. Bộ dữ liệu này được công bố trên cơ sở dữ liệu về kháng thuốc với HIV của đại học Stanford. Sau quá trình loại bỏ các quan sát không đủ thông tin, chúng tôi giữ lại thông tin của 396 bệnh nhân. Những mô tả chi tiết về mẫu, khuếch tán RNA, nhân bản hay giải trình tự có thể xem tại [7]. Tất cả các kết quả nghiên cứu đưa vào phân tích đều của các bệnh nhân được điều trị bằng Efavirenz.

Nghiên cứu của Bachelier và cộng sự vào năm 2001 đã xác định được dãy biến đổi của các axit amin trong quá trình phiên mã ngược của HIV có ảnh hưởng tới việc kháng thuốc Efavirenz. Họ đã chỉ ra được 2 đường chuyển hóa thay thế để kháng lại Efavirenz, một là chuyển hóa bắt đầu từ đột biến K103N (đường 103) và còn lại là đường chuyển hóa từ đột biến Y188L (đường 188). Chúng tôi thực hiện đánh giá cả 2 đường này và sử dụng để xây dựng cây đột biến. Trong số các đột biến kháng thuốc Efavirenz, để đảm bảo thời gian thực hiện tính toán, trước mắt chúng tôi giới hạn chỉ lấy các đột biến có tần suất xuất hiện lớn hơn 2%.

Sau quá trình tiền xử lý dữ liệu, kết quả thu được là dãy 14 đột biến với tần suất xuất hiện như sau: K103N (52,4%), L100I (9,1%), N348I (8,4%), Y181C (7,8%), G190A (7,7%), H221Y (6,0%), G190S (5,5%), P225H (4,6%), Y188L (4,6%), V108I (4,3%), A98G (3,8%), K101E (3,6%), V179D (3,1%), V106I (2,4%).

Ở đây chúng tôi giải thích thêm về mẫu đột biến có trong bộ dữ liệu HIV. Lấy ví dụ dữ liệu của bệnh nhân mã LB4-P142, tại thời điểm trước khi điều trị (thời điểm 1996-12), có ba Clone mang đột biến K103N. Đến mốc thời điểm tiếp theo (1997-01), đột biến V106I được phát hiện thêm trong quần thể vi-rút của bệnh nhân. Những dữ liệu này là căn cứ cho việc kiểm chứng các giả thiết theo đề xuất của Niko Beerenwinke, và sẽ trình bày ở những phần tiếp theo của bài báo.

2. Kiểm định thống kê

Để mô hình hóa quá trình tiến hóa tích lũy, Niko Beerenwinke đã đưa ra 3 giả thiết [5]:

(A1) Những đột biến thay thế không xảy ra độc lập. Con đường tiến hóa của vi-rút kháng thuốc được xem xét với các đột biến có tính duy trì.

(A2) Sự tồn tại của các đột biến trong quần thể là vĩnh viễn, tức là những biến đổi đã diễn ra sẽ luôn được duy trì, không thể đảo ngược hoặc mất đi.

(A3) Tại mỗi thời điểm, quần thể vi-rút bị chi phối bởi một chủng duy nhất và các Clones là độc lập với nhau.

Trong 3 giả thiết trên, có thể kiểm chứng bằng thống kê đối với giả thiết (A2) và (A3). Chúng tôi đã thực hiện kiểm chứng hai giả thiết này với bộ dữ liệu đã được tiền xử lý bằng phương pháp kiểm tra ngẫu nhiên hóa (Randomization Test) [8]. Cụ thể như sau:

²<https://hivdb.stanford.edu/cgi-bin/RTIPairs.cgi>

a) Giả thiết (A2) nêu lên tính chất không thể đảo ngược của các thay thế, tức khi đột biến đã xuất hiện trong Clone k của bệnh nhân i tại 1 thời điểm thì đột biến đó vẫn xuất hiện tại tất cả các Clones sau đó, chứ không mất đi. Giả thiết được kiểm tra bằng cách theo dõi sự thay đổi tần suất alen đột biến cho mỗi bệnh nhân theo thời gian.

Kí hiệu $[N] = \{1, 2, \dots, N\}$ là tập các bệnh nhân, $[M] = \{1, 2, \dots, M\}$ là tập các đột biến được xem xét, $[K_{ij}]$ là tập K_{ij} Clones quan sát được của bệnh nhân i ($i = 1, \dots, N$) tại thời điểm j ($j = 1, \dots, J_i$). Hàm chỉ $y_{ijkm} \in \{0, 1\}$ thể hiện sự hiện diện hay không của đột biến $m \in M$ trong Clone $k \in [K_{ij}]$ của bệnh nhân i tại thời điểm j . Khi đó, tần suất alen đột biến $m \in [M]$ cho mỗi bệnh nhân $i = 1, \dots, N$ theo thời gian $j = 1, \dots, J_i$ xác định bởi công thức

$$f_{ijm} = \frac{1}{K_{ij}} \sum_{k \in [K_{ij}]} y_{ijkm} \quad (1)$$

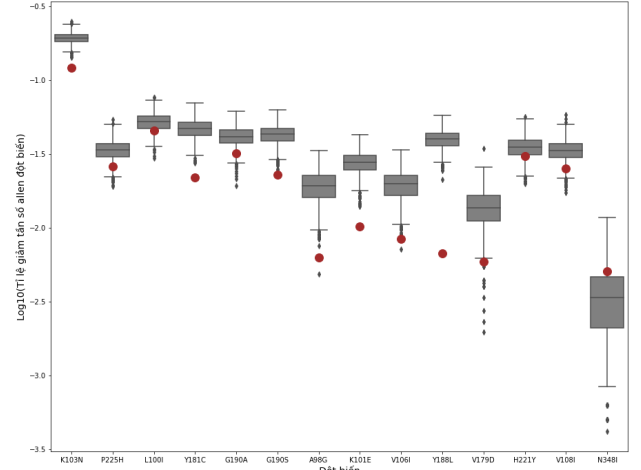
Với mỗi đột biến m , giá trị kiểm định thống kê A_m xác định mức độ tần suất alen của nó giảm từ một thời điểm này cho tới thời điểm khác. Với I là hàm chỉ, thì

$$A_m = \frac{1}{N} \sum_{i=1}^N \frac{1}{J_i - 1} \sum_{j=1}^{J_i-1} I\{f_{i,j,m} > f_{i,j+1,m}\} \quad (2)$$

Sử dụng phương pháp kiểm định ngẫu nhiên (Randomization Test) với giả thuyết không (Null Hypothesis) là xác suất của việc tăng và giảm tần số alen là bằng nhau. Phương pháp kiểm định ngẫu nhiên thực hiện bằng cách xáo trộn trình tự (Resampling) trong quần thể Clones. Tại mỗi lần Resampling dữ liệu, ta tính được một giá trị kiểm định thống kê. Với giả thiết sự thay thế không thể đảo ngược (A2), tỉ lệ % của việc giảm tần suất đột biến chiếm số lượng rất nhỏ trong quần thể. Để giả thiết này đúng, giá trị kiểm định thống kê của dữ liệu ban đầu cần phải nhỏ hơn các giá trị kiểm định thống kê sau khi Resampling dữ liệu.

Trong tính toán của chúng tôi, sử dụng Randomization Test với 1000 lần. Giá trị p_{value} được tính là tỉ lệ trên 1000 lần của số lần giá trị kiểm định thống kê với dữ liệu ngẫu nhiên mà nhỏ hơn giá trị kiểm định thống kê với dữ liệu ban đầu (Hình 1). Với hầu hết các đột biến, giá trị này nhỏ đáng kể ($p_{K103N} < 0,001$, $p_{Y181C} < 0,001$, $p_{K103N} < 0,001$, $p_{G190S} < 0,001$, $p_{Y188L} < 0,001$, $p_{A98G} = 0,001$, $p_{V106I} = 0,004$, $p_{V179D} = 0,017$, $p_{V108I} = 0,073$, $p_{G190A} = 0,078$, $p_{225H} = 0,084$), bên cạnh một số ngoại lệ ($p_{L100I} = 0,191$, $p_{H221Y} = 0,216$, $p_{N348I} = 0,841$). Chúng tôi đánh giá rằng giả thiết (A2) về sự thay thế không thể đảo ngược là hợp lý đối với phần lớn dữ liệu được sử dụng trong nghiên cứu này.

b) Giả thiết (A3) sẽ được kiểm định bằng cách xét sự đa dạng di truyền học của các Clones tại mỗi thời điểm $j \in [T_i]$ của bệnh nhân i , đa dạng càng nhỏ sẽ càng khẳng



Hình 1. Sự đột biến đã xuất hiện được duy trì theo thời gian. Với mỗi đột biến, chấm tròn biểu thị số lần giảm tần suất trong quần thể theo thời gian. Biểu đồ hộp biểu diễn phân phối của giá trị thống kê tạo bởi kiểm định ngẫu nhiên 1000 lần.

định được các Clones không sinh ra chủng loài mới mà bảo tồn gene từ một chủng chính.

Để đo sự đa dạng di truyền học giữa hai Clone c và c' có hay không đột biến m kí hiệu tương ứng là c_m và c'_m , khoảng cách Hamming được sử dụng theo công thức

$$D_H(c, c') = \sum_{m \in [M]} I\{c_m \neq c'_m\} \quad (3)$$

Độ đa dạng của bộ mẫu gồm $c_1, \dots, c_K$ gồm K Clones là giá trị kỳ vọng của khoảng cách Hamming giữa 2 Clones bất kỳ trong mẫu đó,

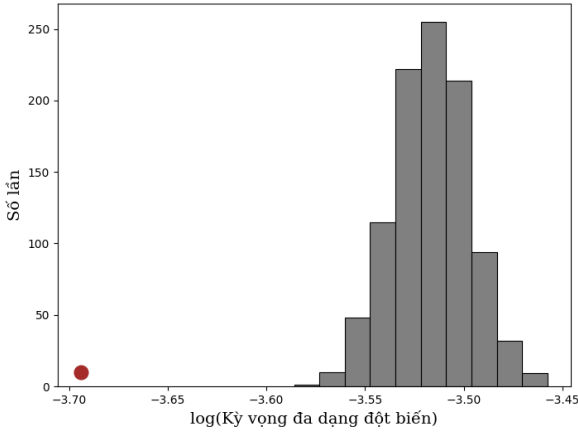
$$D_H(c_1, \dots, c_K) = \frac{2}{K(K-1)} \sum_{k < k'} D_H(c_k, c_{k'}) \quad (4)$$

Đại lượng thống kê để đưa vào kiểm định là kỳ vọng độ đa dạng của toàn bộ mẫu, tính theo

$$B = \frac{1}{N} \sum_{i=1}^N \frac{1}{J_i} \sum_{j=1}^{J_i} D_H(y_{ij1}, \dots, y_{ijK_{ij}}) \quad (5)$$

Đặt giả thuyết bác bỏ là tại từng thời điểm, đều có thể quan sát được đa dạng di truyền học đầy đủ tính trên tất cả Clones của bệnh nhân. Giá trị của nó được ước lượng bằng cách tính kỳ vọng đa dạng di truyền học sau 1000 lần hoán đổi ngẫu nhiên các Clones vào thời điểm khác nhau. Một quan sát với B nhỏ hơn giá trị kỳ vọng này sẽ có mức đa dạng nhỏ hơn đa dạng di truyền học đầy đủ và bác bỏ được giả thuyết trên nếu tỉ lệ nhỏ hơn là đáng kể ($p_{value} < 5\%$).

Kết quả tính toán cho thấy trên bộ dữ liệu hiện tại, đa dạng di truyền học tính theo từng thời điểm nhỏ hơn rất



Hình 2. Kết quả giá trị logarit đa dạng đột biến của dữ liệu mẫu (hình tròn) và phân phối tần suất sau khi thực hiện hoán đổi 1000 lần các Clones của một bệnh nhân vào thời điểm ngẫu nhiên (biểu đồ tần suất).

hiều so với đa dạng của toàn bộ mẫu. Chúng tôi cũng quan sát được từ tính toán, các Clones tại mỗi thời điểm chỉ khác biệt bởi 1 trên 4900 đột biến được xét. Trong khi đó, kết quả trung bình sau khi ngẫu nhiên hóa là 1 trên 3300 đột biến với kết quả kiểm định p_{value} rất nhỏ. Như vậy, có thể kết luận rằng việc phát sinh chủng mới trong các Clones rất nhỏ và xảy ra giữa các thời điểm chứ không phải tại mỗi thời điểm. Từ đó khẳng định được rằng giả thiết (A3) là có ý nghĩa trên tập các nhân bản HIV được sử dụng trong nghiên cứu này.

III. MÔ HÌNH MARKOV

1. Cây đột biến

Với tập đột biến được xem xét $[M] = \{1, 2, \dots, M\}$, xét một cây có hướng T trên tập hợp đỉnh $V = \{0\} \cup [M]$ với gốc tại 0. Ứng với mỗi đỉnh $m \in V$ của T , xét một biến nhị phân ngẫu nhiên X_m trong đó mỗi tương quan của các biến ngẫu nhiên này phụ thuộc vào hình dạng của cây T và $\Pr(X_0 = 1) = 1$. Theo cách định nghĩa này, cây T cảm sinh một mô hình đồ thị có hướng cho phân bố đồng thời của véc-tơ ngẫu nhiên $X = (X_1, \dots, X_M)$. Cụ thể, phân bố của X cảm sinh từ mô hình này xác định bởi

$$\Pr(X = x) = \prod_{m \in [M]} \Pr(X_m = x_m | X_{pa(m)} = x_{pa(m)}) \quad (6)$$

trong đó $x = (x_1, \dots, x_M) \in \mathcal{I}$ với $\mathcal{I} = \{0, 1\}^M$ là không gian trạng thái của véc-tơ ngẫu nhiên X , $x_0 = 1$ và $pa(m)$ là đỉnh cha của m trong cây T . Ví dụ

- Nếu T là đồ thị đường từ $0 \rightarrow 1 \rightarrow 2 \rightarrow \dots \rightarrow M$, thì

$$\Pr(X = x) = \prod_{m=1}^M \Pr(X_m = x_m | X_{m-1} = x_{m-1}).$$

- Nếu T là một đồ thị sao với gốc tại 0 thì

$$\Pr(X = x) = \prod_{m \in [M]} \Pr(X_m = x_m | X_0 = 1).$$

Từ công thức (6), phân phối của X được xác định bởi các ma trận chuyển trạng thái có dạng

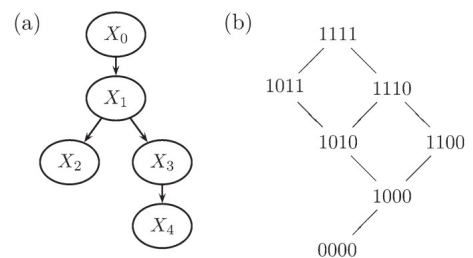
$$\vartheta^m = \begin{matrix} & 0 & 1 \\ \begin{matrix} 0 \\ 1 \end{matrix} & \begin{pmatrix} 1 & 0 \\ 1 - \vartheta_{11}^m & \vartheta_{11}^m \end{pmatrix} \end{matrix}$$

trong đó $\vartheta_{ab}^m = \Pr(X_m = b | X_{pa(m)} = a) \in [0, 1]$.

Phần còn lại của mục này mô tả việc áp dụng mô hình trên cho các sự kiện đột biến, gọi tắt là đột biến. Cụ thể, xét tập M đột biến $\{1, 2, \dots, M\}$, với mỗi đột biến $m \in [M]$, biến nhị phân ngẫu nhiên X_m với hai giá trị 1 và 0 lần lượt biểu thị sự xuất hiện và không xuất hiện của đột biến m . Một dãy đột biến $x \in \mathcal{I}$ được gọi là *tương thích* với cây T nếu trạng thái x có thể xuất hiện với xác suất khác 0 trong mô hình phân bố đồng thời cảm sinh bởi T . Do đó, x tương thích với T nếu tồn tại tham số $\vartheta = (\vartheta_{11}^1, \dots, \vartheta_{11}^M) \in [0, 1]^M$ sao cho

$$\Pr(X = x) = \prod_{m \in [M]} \vartheta_{pa(m), x_m}^m > 0.$$

Kí hiệu $C(T)$ là tập hợp các trạng thái tương thích với cây T , tập hợp này tạo nên một mạng tinh thể $(C(T), \vee, \wedge)$ trong đó $\vee$ và $\wedge$ lần lượt là các toán tử max và min theo thành phần (xem [9, Lemma 14.3]). Biểu đồ Hasse trong hình vẽ dưới đây mô tả các con đường (dãy đột biến) được nhắc tới ở trên bằng việc tương ứng mỗi con đường tiến hóa với một đường đi từ trạng thái tự nhiên $(0, \dots, 0) \in C(T)$ đến trạng thái $(1, \dots, 1) \in C(T)$ khi tất cả các đột biến đều đã xuất hiện.



Hình 3. (a) Cây đột biến và (b) mạng tinh thể cảm sinh của các trạng thái tương thích.

Trong mô hình cây đột biến theo thời gian, các đột biến xuất hiện theo quá trình Poisson độc lập. Nếu λ_m là tỷ lệ

của quá trình này trên cạnh $pa(m) \rightarrow m$, xác suất để đột biến m xuất hiện trong khoảng thời gian Δt là

$$\theta_{11}^m = \Pr(X_m = 1 | X_{pa(m)} = 1) = 1 - e^{-\lambda_m \Delta t}.$$

Cây đột biến có thể được tái thiết lập từ dữ liệu bằng việc giải bài toán nhánh cực đại trong đồ thị đầy đủ trên tập đỉnh V bằng một tổ hợp thuật toán hiệu quả, xem thêm [9, 10] để biết chi tiết về thuật toán. Trong bài báo này, chúng tôi sử dụng thuật toán nêu trên để thiết lập hình dạng của T .

2. Mô hình Markov ẩn

Giả sử ta có thể quan sát được chuỗi đột biến vi-rút của một bệnh nhân nhiều hơn một thời điểm. Gọi X_{jm} là biến ngẫu nhiên chỉ sự xuất hiện của đột biến m tại thời điểm t_j với $j = 1, \dots, J$ trong quần thể vi-rút của bệnh nhân. Ta giả thiết rằng quá trình tiến hóa bắt đầu tại thời điểm 0 ở trạng thái tự nhiên - không có đột biến nào. Do đó, $X_{1m} = 0$ với mọi $m \in [M]$ với $t_1 = 0$.

Sự phát triển của đột biến m tại thời điểm t_j với $j \geq 2$, mã hóa bởi biến ngẫu nhiên X_{jm} , lúc này phụ thuộc vào trạng thái thời điểm trước đó $X_{j-1,m}$, cũng như trạng thái hiện tại của đột biến cha $X_{j,pa(m)}$. Sự phụ thuộc này nảy sinh từ tính chất: sự hiện diện của đột biến m tại thời điểm t_j là kết quả của quá trình phát triển qua cạnh $pa(m) \rightarrow m$ tại thời điểm t_j hoặc từ tính không thể đảo ngược của nó trong quần thể vi-rút và do đó phụ thuộc vào sự hiện diện của nó tại thời điểm t_{j-1} . Cấu trúc phụ thuộc giữa các đột biến $(X_{jm} | j = 1, \dots, J, m \in [M])$ có thể được biểu thị dưới một đồ thị có hướng không có chu trình (Xem Hình 4).

Ma trận chuyển của mô hình xích Markov này có dạng

$$\theta_j(\lambda_m) = \begin{matrix} & \begin{matrix} 0 & 1 \end{matrix} \\ \begin{matrix} 00 \\ 01 \\ 10 \\ 11 \end{matrix} & \begin{pmatrix} 1 & 0 \\ e^{-\lambda_m(t_j - t_{j-1})} & 1 - e^{-\lambda_m(t_j - t_{j-1})} \\ * & * \\ 0 & 1 \end{pmatrix} \end{matrix} \quad (7)$$

trong đó các hàng được đánh chỉ số bởi cặp $(m, pa(m)) \in \{0, 1\}^2$. Vị trí đánh dấu $*$ chỉ các vị trí không cần phải xem xét vì không có đột biến m nào xuất hiện trước đột biến mẹ $pa(m)$ của nó. Với ma trận này, ta định nghĩa mô hình cây Markov đột biến như là họ các phân bố đồng thời có dạng

$$\begin{aligned} \Pr(X_{jm} = x_{jm}, j = 1, \dots, J, m \in [M]) \\ = \prod_{j=1}^J \prod_{m \in [M]} \theta_j(\lambda_m)_{(x_{j-1,m}, x_{jpa(m)}), x_{jm}}, \end{aligned}$$

trong đó $x_{j0} = 1$ và $t_0 = 0$. Điều này dẫn tới

$$\begin{aligned} \Pr(X_{jm} = x_{jm} | X_{j-1,m} = x_{j-1,m}, X_{jpa(m)} = x_{jpa(m)}) \\ = \theta_j(\lambda_m)_{(x_{j-1,m}, x_{jpa(m)}), x_{jm}}. \end{aligned}$$

3. Mô hình Markov ẩn cho cây đột biến

Chuyển qua trường hợp các mẫu Clones quan sát được tại các thời điểm khác nhau, mô hình hóa dữ liệu này bằng việc giả sử các Clones là các sao chép lỗi của một chuỗi đột biến ẩn mà chuỗi đột biến này tiến hóa dựa theo một cây đột biến. Như vậy, X_{jm} lúc này là một biến nhị phân ẩn. Dữ liệu quan sát được là ví dụ của các biến nhị phân ngẫu nhiên $Y_{jkm}, k \in [K_j]$ chỉ đột biến m có hiện diện trong đột biến thứ k lấy mẫu từ quần thể vi-rút tại thời điểm t_j hay không. Các Clones là độc lập có điều kiện với điều kiện $(X_j), j = 1, \dots, J$ cho trước. Mô hình đồ thị thu được xem như là một mô hình Markov ẩn cho cây đột biến (Mtree-HMM), một ví dụ xem ở Hình 4.

Sử dụng các kí hiệu $\varepsilon^+ = (\varepsilon_1^+, \dots, \varepsilon_M^+) \in [0, 1]^M$ và $\varepsilon^- = (\varepsilon_1^-, \dots, \varepsilon_M^-) \in [0, 1]^M$ là các véc-tơ tham số chứa xác suất quan sát phải dương tính giả và âm tính giả của từng đột biến. Tỷ lệ dương tính giả và âm tính giả cho biết sự khác biệt so với trạng thái quần thể có thể phát sinh từ các đột biến trong phản ứng PCR. Do đó, các tham số này định lượng kỳ vọng đa dạng di truyền của một quần thể vi-rút. Với điều kiện trạng thái X_{jm} , xác suất quan sát được đột biến m trong Clone k tại thời điểm t_j là

$$\theta'(\varepsilon_m^+, \varepsilon_m^-) = \begin{pmatrix} 0 & 1 \\ 1 - \varepsilon_m^+ & \varepsilon_m^+ \\ \varepsilon_m^- & 1 - \varepsilon_m^- \end{pmatrix}$$

Các yếu tố của ma trận này là xác suất có điều kiện

$$\theta'(\varepsilon_m^+, \varepsilon_m^-) = \Pr(Y_{jkm} = y_{jkm} | X_{jm} = x_{jm}).$$

Do đó, các Clones khác nhau $Y_{jk} = (Y_{jk1}, \dots, Y_{jkM})$, với $k \in [K_j]$ được mô hình thành các biến ngẫu nhiên độc lập cùng phân bố. Đặt

$$Y = (Y_{jkm} | j = 1, \dots, J, k \in [K_j], m \in [M])$$

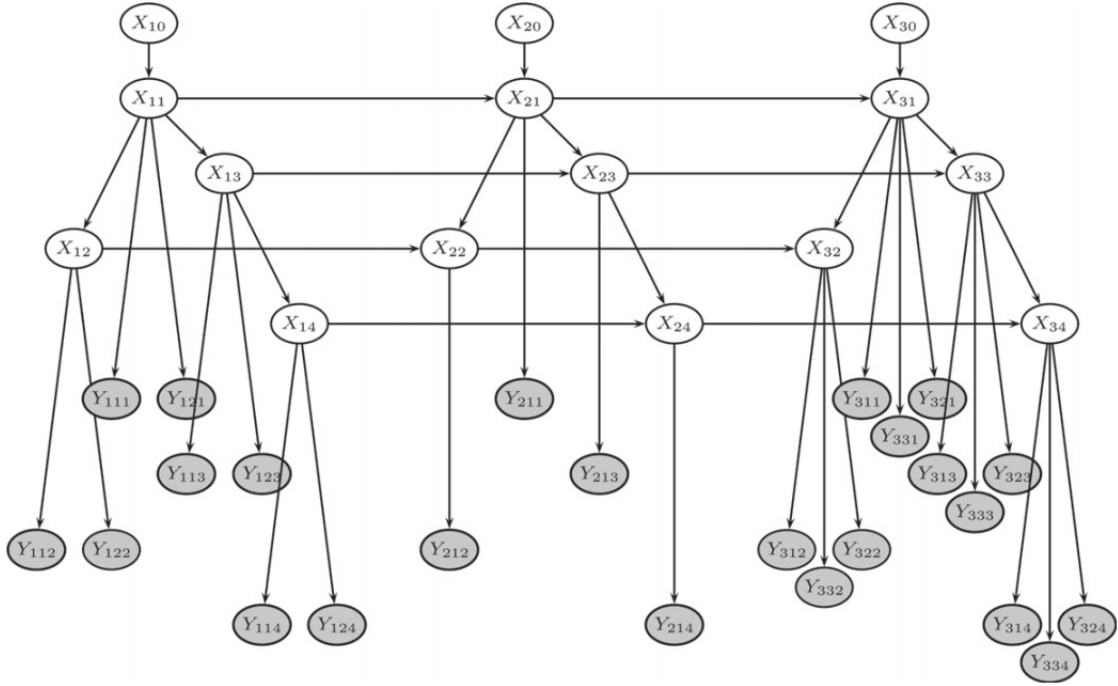
là véc tơ biểu thị tất cả dãy quan sát Clones. Mô hình Markov ẩn cho cây đột biến là họ các phân bố đồng thời của Y được cho bởi

$$\begin{aligned} \Pr(Y = y) = \sum_{x_1 \in C(T)} \dots \sum_{x_J \in C(T)} \prod_{m \in [M]} \prod_{j=1}^J \\ \left(\theta_j(\lambda_m)_{(x_{j-1,m}, x_{jpa(m)}), x_{jm}} \prod_{k \in [K_j]} \theta'(\varepsilon_m^+, \varepsilon_m^-)_{x_{jm}, y_{jkm}} \right), \end{aligned}$$

trong đó phép lấy tổng được thực hiện với tất cả các trạng thái ẩn của mô hình. Cấu trúc đồ thị của mô hình Markov ẩn cho cây đột biến được biểu diễn ở Hình 4.

4. Tính toán tham số

Với mỗi bệnh nhân $i \in [N] = 1, 2, \dots, N$, ta có các quan sát tại các thời điểm $t_{i1}, t_{i2}, \dots, t_{iJ_i}$. Gọi X_{ijm} là biến ngẫu



Hình 4. Đồ thị không có chu trình mô tả mô hình Markov ẩn cho cây đột biến theo thời gian với 3 thời điểm. Các đỉnh trắng ứng với biến ngẫu nhiên ẩn và các đỉnh xám là các biến quan sát được [5].

nhiên chỉ sự xuất hiện của đột biến m trong quần thể vi-rút của bệnh nhân i tại thời điểm t_{ij} . Biến ngẫu nhiên Y_{ijkm} chỉ sự xuất hiện của đột biến m trong Clone $k \in [K_{ij}]$ của bệnh nhân i tại thời điểm t_{ij} . Kí hiệu ma trận chuyển (7) tương ứng với bệnh nhân i bởi $\theta_{ij}(\lambda_m)$, chẳng hạn

$$\theta_{ij}(\lambda_m)_{01,0} = e^{-\lambda_m(t_{ij}-t_{i,j-1})}.$$

Giả sử các bệnh nhân độc lập với nhau và dữ liệu của mỗi bệnh nhân dựa theo một mô hình cây Markov ẩn trên một cây T cố định. Khi đó, kết quả của mô hình cho quan sát

$$Y = (Y_{ijkm} | i \in [N], j = 1, \dots, J_i, k \in [K_{ij}], m \in [M])$$

thực hiện tại các thời điểm $\{t_{ij} | i \in [N], j = 1, \dots, J_i\}$ là các tham số $\lambda = (\lambda_1, \dots, \lambda_n)$, $\varepsilon^+ = (\varepsilon_1^+, \dots, \varepsilon_m^+)$, và $\varepsilon^- = (\varepsilon_1^-, \dots, \varepsilon_m^-)$ làm cực đại hàm Likelihood

$$L_{\text{obs}}(\lambda, \varepsilon^+, \varepsilon^-) = \sum_{x_{11} \in C(T)} \dots \sum_{x_{NJ_N} \in C(T)} \prod_{i \in [N]} \prod_{m \in [M]} \prod_{j=1}^{J_i} \left(\theta_{ij}(\lambda_m)_{x_{i,j-1,m}, x_{ijpa(m)}} \prod_{k \in [K_{ij}]} \theta'(\varepsilon^+, \varepsilon^-)_{x_{ijm}, y_{ijkm}} \right)$$

trong đó $x_{i0m} = 1$ và $t_{i0} = 0$ với mọi $i \in [N], m \in [M]$. Vì việc tính toán trên hàm Likelihood khó khăn hơn, nên hàm Log-Likelihood sẽ được sử dụng thay thế.

Gọi $\{x_{ijm}\}$ là giá trị của các biến ẩn $\{X_{ijm}\}$ tương thích với mô hình cây đột biến ẩn. Tính chất sau được suy ra trực tiếp từ mô hình cây Markov ẩn.

$$x_{ijm} = \begin{cases} 1 & \text{nếu } x_{i,j-1,m} = 1, \\ 0 & \text{nếu } x_{ijpa(m)} = 0. \end{cases}$$

Gọi I là hàm đặc trưng (hàm chỉ) nhận giá trị 0 và 1, đặt

$$\begin{aligned} \chi_{ijm}(a) &= I\{x_{i,j-1,m} = 0, x_{ijpa(m)} = 1, x_{ijm} = a\}, \\ \chi'_{ijkm}(a, b) &= I\{x_{ijm} = a, y_{ijkm} = b\}, \end{aligned}$$

với $a, b = 0, 1$. Khi đó, hàm Log-Likelihood $\ell_{\text{hid}}(\lambda, \varepsilon^+, \varepsilon^-)$ của mô hình ẩn xác định bởi

$$\begin{aligned} \ell_{\text{hid}}(\lambda, \varepsilon^+, \varepsilon^-) &= \sum_{i \in [N]} \sum_{m \in [M]} \sum_{j=1}^{J_i} \left\{ -\chi_{ijm}(0) \lambda_m (t_{ij} - t_{i,j-1}) \right. \\ &\quad \left. + \chi_{ijm}(1) \log(1 - e^{\lambda_m(t_{ij}-t_{i,j-1})}) \right. \\ &\quad \left. + \sum_{k \in [K_{ij}]} \left[\chi'_{ijkm}(0, 0) \log(1 - \varepsilon_m^+) + \chi'_{ijkm}(0, 1) \log \varepsilon_m^+ \right. \right. \\ &\quad \left. \left. + \chi'_{ijkm}(1, 0) \log \varepsilon_m^- + \chi'_{ijkm}(1, 1) \log(1 - \varepsilon_m^-) \right] \right\}. \end{aligned}$$

Để giải bài toán tối ưu này và tìm được ước lượng hợp lý cực đại, chúng tôi sử dụng thuật toán EM hay Forward-Backward của mô hình Markov ẩn. Chi tiết của thuật toán này có thể tham khảo ở nhiều tài liệu, chẳng hạn xem [11].

IV. TÍNH TOÁN THỰC NGHIỆM

Nhắc lại, chúng tôi thực hiện tính toán thực nghiệm trên bộ dữ liệu đã làm sạch và kiểm tra các giả thiết bao gồm 14 đột biến đối với vi-rút HIV. Theo cách dựng cây đột biến đã trình bày trong mục III.A với, nút gốc là K103N và Y188C được giả thiết là xuất hiện ngay từ đầu (chỉ có trạng thái 1) trong các tập đột biến có thể có (compatible set). Mặt khác, các nút này không có nút cha mẹ, do đó việc tính toán tham số λ_m , dương tính giả (ε_m^+), và âm tính giả (ε_m^-) được thực hiện dựa trên thống kê từ bộ dữ liệu ban đầu. Tham số của các nút còn lại được tối ưu nhờ quá trình học của thuật toán EM gồm hai bước Expectation (E-step) và Maximization (M-step) [12]. Trong bước E-step ta cần tính được xác suất chuyển trạng thái $u_{ijm,a}$ thông qua hàm GST, xác suất xuất hiện quan sát $u_{m,ab}$ thông qua hàm GSE, trong đó a, b nhận các giá trị 0, 1.

$$\text{GST} : u_{ijm,a} = \Pr(x_{i,j-1,m} = 0, x_{ijpa(m)} = 1, x_{ijm} = a | Y)$$

$$\text{GSE} : u_{m,ab} = \sum_{i \in [N]} \sum_{j=1}^{J_i} \sum_{k \in [K_{ij}]} \Pr(x_{ijm} = a, y_{ijk} = b | Y)$$

Bước M-step được thực hiện nhờ vòng lặp để khép sai số các tham số λ_m , ε_m^+ , ε_m^- , thuật toán dừng lại khi sai số giữa hai bước của λ_m nhỏ hơn $5 \cdot 10^{-4}$ (Xem Thuật toán 1).

Để đánh giá hiệu quả của mô hình, chúng tôi dựa trên một số tiêu chí gồm độ chính xác, độ hội tụ, và thời gian tính toán. Độ chính xác của mô hình được đánh giá bằng xác suất dương tính giả, xác suất âm tính giả giữa các đột biến dự đoán và các đột biến quan sát trong dữ liệu. Độ hội tụ của mô hình được xem xét thông qua số vòng lặp của thuật toán tương ứng với mỗi đột biến. Cuối cùng, thời gian tính toán được đo cho mỗi đột biến khi thực nghiệm trên hệ thống với cấu hình máy tính core i7, 4 CPUs, 8GB RAM, ngôn ngữ Python.

V. KẾT QUẢ

1. Cây đột biến

Sử dụng thuật toán tái xây dựng cây đột biến từ dữ liệu cắt ngang chúng tôi thu được cây đột biến với 15 đỉnh (Xem Hình 5). Nhánh một gồm 3 đột biến là K103N, L100I, P225H; nhánh hai gồm 11 đột biến là Y188C, G190S, G190A, V108I, H221Y, A98G, K101E, N348I, V106I, Y188L, và V179D.

2. Quá trình tiến hóa của vi-rút HIV

Sau khi tối ưu tham số với hai cây có các nút gốc là K103N và Y188C ta thu được các giá trị λ_m tương ứng với từng đột biến. Vì giá trị λ_m mang ý nghĩa là số lần xuất hiện đột biến trong khoảng thời gian một tuần nên giá trị này càng lớn thì đột biến xuất hiện càng sớm. Kết

Thuật toán 1:

Hidden Markov Model Maximization
HMMM($cSet$, $mSet$, $rTree$)

```

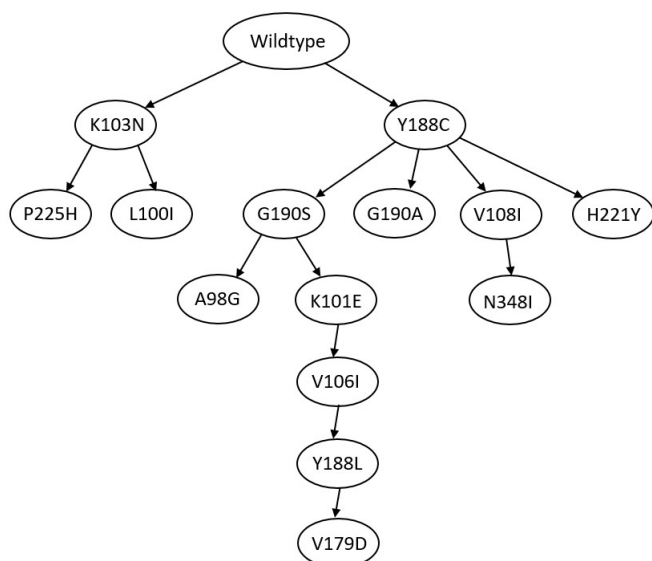
1  Dữ liệu vào:
2     $cSet$ : Compatible Set,
3     $mSet$ : Mutation Set,
4     $rTree$ : Reconstruction Tree.
5  Dữ liệu ra:
6     $pr$ : Parameters.
7  Khởi tạo:
8     $\lambda_m = [0.1] * |mSet|$ 
9     $\varepsilon_m^+ = 0.01$  // Dương tính giả
10    $\varepsilon_m^- = 0.01$  // Âm tính giả
11 Chuẩn bị trước:
12   GEP: Function to Get Emission Probability
13   GST: Function to Get Sum of Transition Probability
14 begin
15   for  $m = mSet[1]$  to  $mSet[-1]$  do
16      $pr = \text{empty}$ ;
17     while  $\text{error of } \lambda_m \geq 5e - 04$  do
18        $u_{m,00} = \text{GEP}(cSet, rTree, \lambda_m, 0, 0, \varepsilon_m^+, \varepsilon_m^-)$ ;
19        $u_{m,01} = \text{GEP}(cSet, rTree, \lambda_m, 0, 1, \varepsilon_m^+, \varepsilon_m^-)$ ;
20        $u_{m,11} = \text{GEP}(cSet, rTree, \lambda_m, 1, 1, \varepsilon_m^+, \varepsilon_m^-)$ ;
21        $u_{m,10} = \text{GEP}(cSet, rTree, \lambda_m, 1, 0, \varepsilon_m^+, \varepsilon_m^-)$ ;
22        $\varepsilon_m^- = \frac{u_{m,01}}{u_{m,00} + u_{m,01}}$ ;
23        $\varepsilon_m^+ = \frac{u_{m,10}}{u_{m,10} + u_{m,11}}$ ;
24        $u_{ijm,0} = \text{GST}(cSet, rTree, \lambda_m, 0, \varepsilon_m^+, \varepsilon_m^-)$ ;
25        $u_{ijm,1} = \text{GST}(cSet, rTree, \lambda_m, 1, \varepsilon_m^+, \varepsilon_m^-)$ ;
26        $\lambda_m = \sum_{i \in [N]} \sum_{j=1}^{J_i} \frac{u_{ijm,1}}{\Delta_{i,j}(u_{ijm,0} + u_{ijm,1})}$ 
27     end
28     append  $(\lambda_m, \varepsilon_m^+, \varepsilon_m^-)$  to  $pr$ 
29   end
30   return  $pr$ 
31 end

```

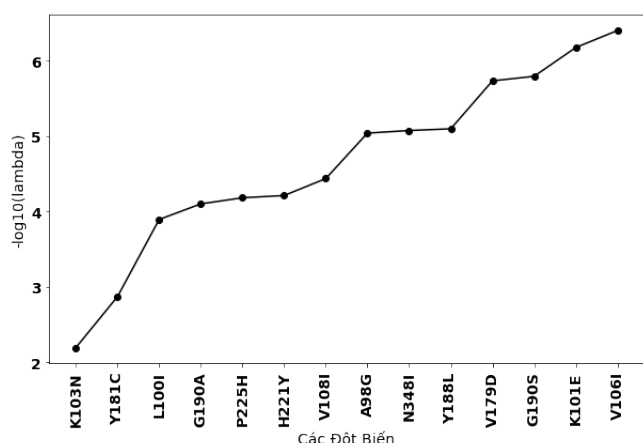
quả cho thấy đột biến K103N xuất hiện sớm nhất ở tuần $\lambda_{K103N}^{-1} = 157$, tiếp theo là các đột biến Y181C, L100I, G190A, P225H, H221Y, V108I, A98G, N348I, Y188L, V179D, G190S, K101E, V106I (Xem Hình 6).

3. Hiệu quả của mô hình

Dương tính giả (ε_m^+) xuất hiện khi một đột biến được dự đoán là tồn tại trong quần thể virus, nhưng thực tế lại không tìm thấy ở những nhân bản trình tự Protein (Clones). Ngược lại, âm tính giả (ε_m^-) xuất hiện khi một đột biến được dự đoán là không tồn tại trong quần thể nhưng lại tìm thấy trong các Clones quan sát. Theo kết quả tái xây dựng cây đột biến (xem phần V.1) ta thu được hai nút gốc là K103N và Y188C của hai nhánh. Với hai đột biến đóng vai trò hai nút gốc này, các tham số được tính dựa trên thống kê dữ liệu ban đầu nên dương tính giả khá cao, xấp xỉ 0.5. Trong các nghiên cứu tiếp theo, chúng tôi có thể cải thiện kết quả này bằng cách sử dụng phương pháp mặt cắt ngang để ước lượng các tham số khởi tạo cho hai nút gốc. Các



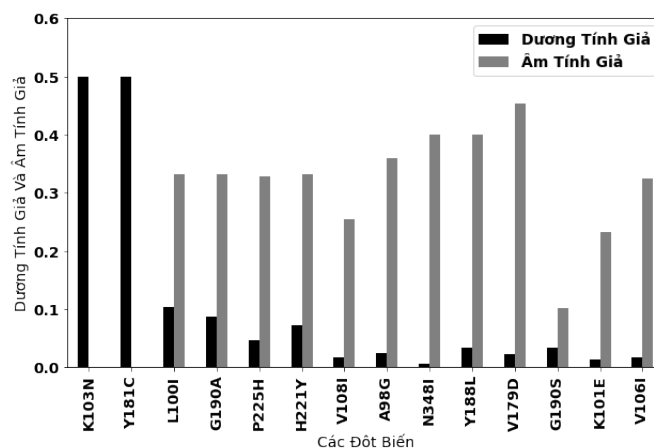
Hình 5. Mô hình cây đột biến của sự xuất hiện kháng thuốc Efavirenz trong HIV. Các đỉnh được đặt tên theo các thay thế axit amin.



Hình 6. Quá trình tiến hóa của vi-rút HIV qua 14 đột biến kháng thuốc. Trục hoành biểu diễn tên của các đột biến. Trục tung biểu diễn $-\log_{10}(\lambda)$. Chỉ số trên trục tung càng lớn thì đột biến xuất hiện càng muộn trong quá trình điều trị.

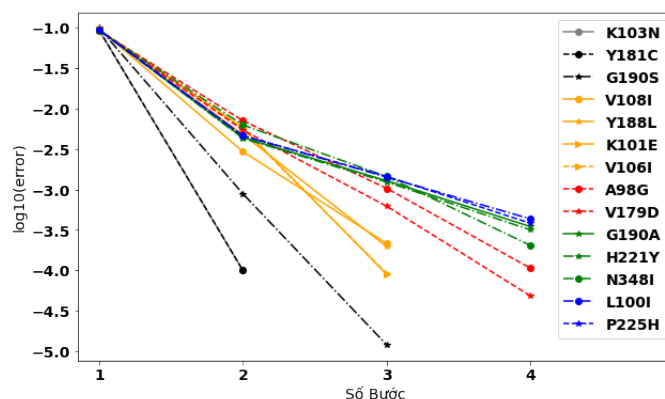
đột biến còn lại của hai nhánh từ L100I đến V106I có xác suất dương tính giả khá nhỏ (từ 0,5% tới 10,4%) so với xác suất âm tính giả (từ 10,2% đến 45,3%) (Xem Hình 7). Dương tính giả nhỏ chứng tỏ giả thiết A2 về sự không thể đảo ngược của đột biến có độ chính xác cao.

Trong bài báo này, phương pháp ước lượng và tối ưu tham số EM vào mô hình Markov ẩn của cây đột biến gene cho tốc độ hội tụ khá cao. Ngoài hai nút gốc là K103N và Y188C được tính toán không dựa vào vòng lặp thì các tham số λ_m của các đột biến còn lại hội tụ chỉ từ 3 đến 4 bước với ngưỡng sai số là 5.10^{-4} (Xem Hình 8). Với hai cây con được tạo ra từ việc xây dựng lại cây đột biến, nhánh một



Hình 7. Tỷ lệ dương tính giả và âm tính giả của dự đoán ứng với các đột biến. Hai đột biến đóng vai trò gốc ở hai nhánh của cây đột biến là K103N và Y181C có tỷ lệ dương tính giả lớn nhất, xấp xỉ 0.5. Ngược lại, các đột biến khác đều có tỷ lệ dương tính giả rất nhỏ so với tỷ lệ âm tính giả.

gồm 3 nút với thời gian tính toán chỉ từ 1,15 đến 2,89 giây với từng đột biến, nhánh hai gồm 11 nút có thời gian tính toán từ 168,1 đến 551,82 giây cho mỗi đột biến (xem bảng I).



Hình 8. Độ hội tụ của thuật toán khi thực hiện vòng lặp cho các đột biến khác nhau với ngưỡng sai số giữa hai bước nhỏ hơn 5.10^{-4} . Trục hoành biểu diễn số bước lặp, trục tung biểu diễn $\log_{10}(\text{error})$.

VI. KẾT LUẬN

Với mô hình Markov ẩn và cây đột biến di truyền áp dụng cho dữ liệu kháng thuốc HIV được cập nhật mới nhất, chúng tôi đã tìm được quá trình tiến hóa của vi-rút HIV qua 14 đột biến kháng thuốc Efavirenz. Thời điểm xuất hiện đột biến được dự đoán thông qua tham số của quá trình poisson. Kết quả cho thấy đột biến K103N xuất hiện sớm nhất và V106I xuất hiện muộn nhất. Mặc dù vậy, do sự kết hợp của thuốc ngày càng phức tạp, nhiều thuốc

Bảng I
THỜI GIAN TÍNH TOÁN CHO MỖI ĐỘT BIẾN VỚI SỐ VÒNG
LẬP VÀ SAI SỐ TƯƠNG ỨNG. NHÁNH 1 CHỈ GỒM BA
ĐỘT BIẾN K103N, L100I, VÀ P225H NÊN THỜI GIAN
TÍNH TOÁN NHANH HƠN CÁC ĐỘT BIẾN CỦA NHÁNH 2.

Đột biến	λ	Số bước	Sai số	Thời gian (s)
K103N	0,0063689	2	0	2,891
Y181C	0,0013390	2	0	168,1
L100I	0,0001269	4	0,00043	1,167
G190A	0,0000788	4	0,00035	551,783
P225H	0,0000654	4	0,00038	1,151
H221Y	0,0000610	4	0,00032	551,818
V108I	0,0000364	3	0,00022	360,737
A98G	0,0000091	4	0,00011	550,339
N348I	0,0000084	4	0,0002	528,277
Y188L	0,0000080	3	0,00019	363,145
V179D	0,0000018	4	0,00005	463,144
G190S	0,0000016	3	0,00001	278,808
K101E	0,0000007	3	0,00009	330,593
V106I	0,0000004	3	0,00009	358,929

được sử dụng trước khi sử dụng Efavirenz trong các phác đồ điều trị nên thời điểm xuất hiện của các đột biến được dự đoán là khá muộn. Để kết quả đạt độ chính xác cao hơn chúng tôi cần tiếp tục tinh sạch dữ liệu làm cho dữ liệu phù hợp nhất với mô hình tính toán.

Thời gian tính toán vẫn đảm bảo không quá lớn với cây đột biến di truyền được xây dựng lại gồm nhiều nút và số lượng ảnh hưởng giữa các nút tăng lên đáng kể. Tuy nhiên, các tham số được tính toán với các nút gốc K103N và Y181C dựa trên thống kê dữ liệu là lý do dẫn đến các đột biến này hầu như chưa tham gia vào quá trình tối ưu tham số. Do đó, chúng cần được tính toán kỹ lưỡng hơn dựa trên việc xây dựng lại cây đột biến di truyền.

Quá trình ước lượng và tối ưu tham số nhờ thuật toán EM tỏ ra khá hiệu quả với mô hình và dữ liệu trong bài báo. Cụ thể, với sai khác các giá trị tham số giữa hai vòng lặp liên tiếp là 5.10^{-4} , thuật toán hội tụ chỉ từ 3 đến 4 bước với mỗi đột biến. Đồng thời, xác suất dương tính giả của các đột biến trong quá trình tiến hóa cũng khá thấp, điều này có ý nghĩa lớn trong lâm sàng khi ta không cần lựa chọn thuốc thay thế cho các đột biến giả trong quá trình điều trị.

Cho rằng các nghiên cứu liên quan tới quá trình tiến hóa của các chủng vi-rút, đặc biệt với các vi-rút chứa men phiên mã ngược là rất có ý nghĩa vì sự cấp thiết trong điều trị một khi vi rút tấn công tế bào cơ thể người. Do đó, chúng tôi sẽ tiếp tục phát triển nghiên cứu này với những bộ dữ liệu mới hơn, lớn hơn. Bên cạnh mục tiêu hoàn thiện mô hình cây đột biến, nghiên cứu sự tiến hóa của vi-rút là việc áp dụng phương pháp với các loại vi-rút mới, biến đổi nhanh.

LỜI CẢM ƠN

Nghiên cứu được hoàn thành bởi nhóm nghiên cứu gồm các học viên và giảng viên chương trình đào tạo Thạc sĩ Khoa học Dữ liệu của Khoa Toán-Cơ-Tin học, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội với hai tác giả chính là Nguyễn Văn Thế và Tạ Văn Nhân. Nghiên cứu được hỗ trợ bởi VinIF trong khuôn khổ chương trình hợp tác đào tạo thạc sĩ Khoa học dữ liệu với khoa Toán-Cơ-Tin học, trường Đại học Khoa học Tự nhiên, Đại học quốc gia Hà Nội. Nguyễn Văn Thế được tài trợ bởi Tập đoàn Vingroup – Công ty CP và hỗ trợ bởi chương trình học bổng đào tạo thạc sĩ, tiến sĩ trong nước của Quỹ Đổi mới sáng tạo Vingroup (VINIF), Viện Nghiên cứu Dữ liệu lớn (VinBigdata), mã số VINIF.2020.ThS.KHTN.05.

TÀI LIỆU THAM KHẢO

- [1] F. Crick, "Central Dogma of Molecular Biology," *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970.
- [2] H. M. Temin and D. Baltimore, "RNA-Directed DNA Synthesis and RNA Tumor Viruses," in *Advances in Virus Research*, K. M. Smith, M. A. Lauffer, and F. B. Bang, Eds. Academic Press, Jan. 1972, vol. 17, pp. 129–186.
- [3] R. C. Gallo, "Summary of Recent Observations on the Molecular Biology of RNA Tumor Viruses and Attempts at Application to Human Leukemia," *American Journal of Clinical Pathology*, vol. 60, no. 1, pp. 80–87, Jul. 1973.
- [4] B. D. Preston, B. J. Poiesz, and L. A. Loeb, "Fidelity of HIV-1 reverse transcriptase," *Science (New York, N.Y.)*, vol. 242, no. 4882, pp. 1168–1171, Nov. 1988.
- [5] N. Beerenwinkel and M. Drton, "A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data," *Biostatistics (Oxford, England)*, vol. 8, no. 1, pp. 53–71, Jan. 2007.
- [6] S.-Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer, "Human immunodeficiency virus reverse transcriptase and protease sequence database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 298–303, Jan. 2003.
- [7] L. Bachelier, S. Jeffrey, G. Hanna, R. D'Aquila, L. Wallace, K. Logue, B. Cordova, K. Hertogs, B. Larder, R. Buckery, D. Baker, K. Gallagher, H. Scarnati, R. Tritch, and C. Rizzo, "Genotypic correlates of phenotypic resistance to efavirenz in virus isolates from patients failing nonnucleoside reverse transcriptase inhibitor therapy," *Journal of Virology*, vol. 75, no. 11, pp. 4999–5008, Jun. 2001.
- [8] C. E. Lunneborg, "Random assignment of available cases: Bootstrap standard errors and confidence intervals," *Psychological Methods*, vol. 6, no. 4, pp. 402–412, 2001.
- [9] N. Beerenwinkel and M. Drton, *Mutagenetic Tree Models*. Cambridge University Press, 2005.
- [10] R. Desper, F. Jiang, O.-P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer, "Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data," *Journal of Computational Biology*, vol. 6, no. 1, pp. 37–51, Jan. 1999, publisher: Mary Ann Liebert, Inc., publishers.
- [11] O. Cappé, "Online EM Algorithm for Hidden Markov Models," *Journal of Computational and Graphical Statistics*, vol. 20, no. 3, pp. 728–749, 2011.
- [12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

SƠ LƯỢC VỀ TÁC GIẢ

Nguyễn Văn Thế



Nhận bằng cử nhân khoa học tài năng ngành Toán học năm 2020.

Hiện là học viên chương trình đào tạo thạc sĩ Khoa học dữ liệu, trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội.

Các lĩnh vực nghiên cứu: Tổ hợp, lý thuyết đồ thị, Giải tích ma trận, thuật toán, tin sinh học.

Email: nguyenvanthe@hus.edu.vn

Trịnh Mai Phương



Hiện là học viên chương trình thạc sĩ Khoa học dữ liệu, trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội.

Lĩnh vực nghiên cứu: Dự báo trong kinh tế và kinh doanh, khai phá dữ liệu lớn, học máy có giám sát trong tin sinh học.

Email: trinhmaiphuong_ch2020@hus.edu.vn

Tạ Văn Nhân



Nhận bằng thạc sĩ ngành Khoa học dữ liệu năm 2021, trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội.

Hiện là chuyên viên tin sinh học tại công ty LOBI Việt Nam.

Lĩnh vực nghiên cứu: Giải trình tự hệ gen người, được học hệ gen, mô hình Markov ẩn và mạng Bayes trong tin sinh học.

Email: tavannhan@gmail.com

Nguyễn Thị Hồng Minh



Nhận học vị tiến sĩ Toán - Tin năm 2001, học hàm Phó giáo sư ngành Công nghệ thông tin năm 2018.

Hiện là giảng viên cao cấp, trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội.

Lĩnh vực nghiên cứu: Độ phức tạp thuật toán, tính toán song song, khai phá dữ liệu, tính toán mềm, tin sinh học.

Email: minhnh@hus.edu.vn

Nguyễn Thị Kim Duyên



Nhận bằng cử nhân khoa học ngành Toán-tin năm 2010.

Hiện là học viên chương trình đào tạo thạc sĩ Khoa học dữ liệu, trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội.

Lĩnh vực nghiên cứu: Điểm nguy cơ đa di truyền, áp dụng mô hình học máy trong tin sinh học.

Email: nguyenthikimduyen_ch2020@hus.edu.vn

Ngân sách tối thiểu phát hiện nguồn thông tin sai lệch trên mạng xã hội trực tuyến, đảm bảo đạt ít nhất một ngưỡng cho trước

Phạm Văn Dũng^{1,3}, Nguyễn Thị Tuyết Trinh², Vũ Chí Quang³, Hà Thị Hồng Vân⁴, Nguyễn Việt Anh⁵

¹ Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Hà Nội

² Học viện Y Dược học cổ truyền Việt Nam, Hà Nội

³ Học viện An ninh nhân dân, Bộ Công an, Hà Nội

⁴ Viện nghiên cứu, phát triển KTNV và kiểm định an ninh thiết bị kỹ thuật, Cục KTNV, Bộ Công An, Hà Nội

⁵ Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Hà Nội

Tác giả liên hệ: Phạm Văn Dũng, pvdungc500@gmail.com

Ngày nhận bài: 20/09/2021, ngày sửa chữa: 29/10/2021, ngày duyệt đăng: 15/11/2021

Định danh DOI: 10.32913/mic-ict-research-vn.v2021.n2.1015

Tóm tắt: Phát hiện nguồn phát tán thông tin sai lệch trên mạng xã hội trực tuyến đóng vai trò quan trọng trong việc hạn chế hành vi sai trái trên mạng. Các nghiên cứu gần đây cho thấy, phương pháp đặt máy giám sát có thể phát hiện nguồn thông tin sai lệch. Tuy nhiên, không thể đặt máy giám sát đối với tất cả người dùng mạng vì ngân sách hạn chế.

Trong bài báo này, một mạng xã hội được biểu diễn bởi đồ thị có hướng, mỗi người dùng là một nút trên đồ thị và phát tán thông tin trên đồ thị theo mô hình Bậc độc lập. Trên mô hình này, giả sử biết trước tập nút nghi ngờ sẽ phát tán thông tin sai lệch, chúng tôi đề xuất tìm tập nút nhỏ nhất để đặt giám sát sao cho số nút bị phát hiện đạt ít nhất một ngưỡng cho trước. Ba thuật toán xấp xỉ được đề xuất, bao gồm: Tham lam, phát hiện thông tin sai lệch dựa trên tập mẫu phát hiện và phát hiện thông tin sai lệch dựa trên tập mẫu phát hiện quan trọng. Các thử nghiệm được thực hiện trên bộ dữ liệu của mạng xã hội thực cho thấy các thuật toán của chúng tôi đề xuất vượt trội hơn các thuật toán khác cả về hiệu suất và thời gian thực hiện.

Từ khóa: Tối ưu hóa, mạng xã hội trực tuyến, phát hiện thông tin sai lệch.

Title: Minimum Budget for Misinformation Source Detection in Online Social Networks with Guaranteed to Reach at Least a Given Threshold

Abstract: Detecting the source of misinformation on social media online plays an important role in curbing online misconduct. Recent studies show that the monitoring method can detect the source of false information. However, it is not possible to set the monitor for all network users because of the limited budget. In this paper, a social network is represented by a directed graph, each user is a node on the graph and propagates information on the graph according to the Degree of Independence model. On this model, assuming that we know in advance the set of suspected nodes will spread false information, we propose to find the smallest set of nodes to set the monitoring so that the number of detected nodes reaches at least a given threshold. Three approximation algorithms are proposed, including: Greedy, Misinformation detection based on detection sample set and Misinformation detection based on important detection sample set. Tests performed on real social network datasets show that our proposed algorithms outperform other base algorithms both in terms of performance and execution time.

Keywords: Optimization, online social network, misinformation detection.

I. GIỚI THIỆU

Ngày nay, Mạng xã hội (MXH) đã và đang trở thành một nền tảng quan trọng trong truyền thông số, có ảnh hưởng không nhỏ đến người dùng. Bên cạnh những lợi ích to lớn thì MXH cũng cho phép phát tán thông tin sai lệch (TTSL)

ảnh hưởng đáng kể đến kinh tế và chính trị, v.v.. [1–3]. Do đó, phát hiện TTSL trước khi nó gây ra hậu quả nghiêm trọng là điều hết sức cần thiết. Một số nghiên cứu đã chỉ ra rằng TTSL có thể được phát hiện bằng học máy dựa trên đặc điểm thời gian, cấu trúc, ngôn ngữ, nội dung bài đăng, v.v..[4–6]. Một số khác đề xuất phát hiện TTSL bằng cách

đặt giám sát tại một số nút quan trọng, như phát hiện dịch bệnh [7], TTSL lệch với kích thước tối thiểu [8, 9], phát hiện TTSL bằng tiếp cận heuristic [10], v.v.. Những nghiên cứu về phát hiện thông tin là tiền đề quan trọng để giải quyết bài toán phát hiện và ngăn chặn TTSL trên MXH [11, 11–15].

Mặc dù đã có nhiều nghiên cứu đặt máy giám sát để phát hiện TTSL, tuy nhiên việc phát hiện TTSL trên mạng có hàng trăm nghìn người dùng là không khả thi và không biết có thể đặt bao nhiêu máy giám sát để có thể phát hiện được TTSL và cũng không thể đặt giám sát với tất cả người dùng mạng. Trong bài báo này, chúng tôi nghiên cứu tìm ra tập các nút nhỏ nhất để đặt giám sát sao cho các nút bị phát hiện dự kiến (chức năng phát hiện) đạt ít nhất một ngưỡng cho trước trên mô hình Bậc độ lập IC (Independent Cascade) [16], bài toán được gọi là: Ngân sách tối thiểu phát hiện nguồn thông tin sai lệch MBD (Minimum Budget for Misinformation source Detection). Những đóng góp của bài báo như sau:

- Bài báo chỉ ra rằng MBD là bài toán NP-khó trong trường hợp đạt xấp xỉ $(1 - \epsilon) \ln n$. Hàm phát hiện có tính chất đơn điệu và submodular và tính toán hàm phát hiện là #P-khó.
- Bài báo đề xuất ba thuật toán xấp xỉ, bao gồm: Thuật toán tham lam Greedy; Phát hiện dựa trên tập mẫu SMD (Sampling based Misinformation Detection) và Phát hiện dựa trên tập mẫu quan trọng ISMD (Important Sampling based Misinformation Detection).

Thực nghiệm được thực hiện trên bộ dữ liệu MXH thực. Kết quả cho thấy các thuật toán được đề xuất mới vượt trội hơn các thuật toán khác cả về hiệu suất và thời gian thực hiện. Bài báo được trình bày bởi 05 phần chính: Phần I - Giới thiệu, Phần II - Mô hình và định nghĩa bài toán, Phần III - Đề xuất thuật toán, Phần IV - Thực nghiệm và đánh giá kết quả, Phần V - Kết luận.

II. MÔ HÌNH VÀ ĐỊNH NGHĨA BÀI TOÁN

Trong phần này, bài báo giới thiệu mô hình IC [16], đây là mô hình được sử dụng phổ biến trong các nghiên cứu bài toán phát tán thông tin trên MXH và định nghĩa bài toán MBD trên mô hình này, các ký hiệu sử dụng trong bài báo được thể hiện trong Bảng I.

1. Mô hình bài toán

Trong bài báo này, chúng tôi sử dụng mô hình IC để mô tả phát tán thông tin trên MXH. Đặc trưng chính của mô hình này là quá trình phát tán thông tin dọc theo các cạnh của đồ thị một cách độc lập với nhau. Trong mô hình IC, mỗi cạnh $(u, v) \in E$ được gán một xác suất ảnh hưởng (Influence Probability) $p(u, v) \in [0, 1]$ biểu diễn mức độ

Bảng I
BẢNG KÝ HIỆU ĐẶC BIỆT

Ký hiệu	Diễn giải
n, m	số nút và số cạnh của đồ thị G
$N_{in}(v), N_{out}(v)$	tập nút vào và ra của v
S	tập nút bị nghi ngờ phát tán TTSL
A	tập nút đặt giám sát
$\mathbb{D}(A)$	hàm ảnh hưởng của tập nút A .
$\hat{\mathbb{D}}(A)$	hàm ước lượng của $\mathbb{D}(A)$
$Cov(A, R_j)$	$= \min\{1, A \cap R_j \}$
$Cov_{\mathcal{R}}(A)$	$= \sum_{R_j \in \mathcal{R}} Cov(A, R_j)$, số tập DS trong $\mathcal{R}$ được bao phủ bởi A
$N_i(\delta, \epsilon)$	$\frac{(2 + \frac{2}{3}\epsilon)n}{\epsilon^2(\gamma - \epsilon\gamma)} \ln(\binom{n}{i}/\delta)$, số tập DS tại bước i

ảnh hưởng của nút u với nút v . Nếu $(u, v) \notin E$, thì $p(u, v) = 0$ và $D^t(G, S)$ là tập các nút bị kích hoạt bởi S tại thời điểm t . Quá trình phát tán theo các bước thời gian t rồi rạc và kết thúc khi sau mỗi bước không có nút nào được kích hoạt thêm. Nghĩa là, $D^t(G, S) = D^{t-1}(G, S)$.

- Tại thời điểm $t = 0$, tất cả các nút trong tập nguồn $S = D^0(G, S)$ đều có trạng thái kích hoạt.

- Tại thời điểm $t \geq 1$, mỗi nút $u \in D^{t-1}(G, S)$ có một cơ hội duy nhất kích hoạt đến nút $v \in N_{out}(u)$ với xác suất thành công là $p(u, v)$. Biến cố này có thể được thực hiện bằng cách áp dụng phép thử Bernoulli (Phép tung đồng xu độc lập) với xác suất thành công là $p(u, v)$. Nếu thành công ta thêm v vào tập $D^t(G, S)$ và nói rằng u kích hoạt v tại thời điểm t . Nếu nhiều nút kích hoạt v tại thời điểm t , kết quả tương tự xảy ra, v được thêm vào tập $D^t(G, S)$. Một nút ở trạng thái kích hoạt, nó sẽ giữ nguyên trạng thái.

2. Định nghĩa bài toán

Để phát hiện thông tin sai lệch, chúng tôi đề xuất phương pháp tìm ra tập các nút A để đặt máy giám sát, sao cho xác suất phát hiện thông tin sai lệch đạt ngưỡng γ . Gọi tập $S \subseteq V$ là tập các nút bị nghi ngờ phát tán thông tin sai lệch, tức là các nút có khả năng là nguồn gây ra thông tin sai lệch. Mỗi nút được phát hiện là một nguồn phát tán thông tin sai lệch với xác suất $\rho(u) \geq 0$.

Mô hình IC tương đương với mô hình cạnh trực tuyến [16]. Theo đó, chúng ta có thể tạo ra một đồ thị mẫu g từ đồ thị ban đầu G được ký hiệu là $g \sim G$ bằng cách chọn từng cạnh $e = (u, v) \in E$ độc lập, với xác suất chọn cạnh là $p(u, v)$ và không chọn cạnh là $1 - p(u, v)$ Xác suất tạo đồ thị mẫu g từ đồ thị G là:

$$\Pr[g \sim G] = \prod_{e \in E(g)} p(u, v) \prod_{e \in E \setminus E(g)} (1 - p(u, v)) \quad (1)$$

Trong đó, $E(g)$ là tập cạnh của đồ thị mẫu g . Nếu đặt thiết bị giám sát tại nút v , nó sẽ phát hiện TTSL từ các nút được kết nối với nó. Gọi $d_g(u, v)$, là khoảng cách từ u đến

v , thì phải mất $d_g(u, v)$ bước để phát hiện thông tin từ u . Xác suất để tập A phát hiện ra thông tin từ nút u là:

$$\mathbb{D}(A, u) = \sum_{g \sim G} \Pr[g \sim G] R(A, g, u) \quad (2)$$

với:

$$R(A, g, u) = \begin{cases} 1, & \text{nếu } d_g(u, A) < \infty, \\ 0, & \text{ngược lại} \end{cases} \quad (3)$$

Trong đó, $d_g(u, A) = \min_{v \in A} d(u, v)$, vì xác suất phát hiện nút u là nguồn phát tán thông tin sai lệch là $\rho(u)$, nên xác suất phát hiện của giám sát đặt tại các nút của tập A như sau (hàm mục tiêu):

$$\mathbb{D}(A) = \sum_{u \in S} \rho(u) \sum_{g \sim G} \Pr[g \sim G] R(A, g, u) \quad (4)$$

Bài toán Ngân sách tối thiểu để phát hiện nguồn thông tin sai lệch MBD được định nghĩa như sau:

Định nghĩa 1 (MBD): Một MXH cho bởi đồ thị $G(V, E)$ theo mô hình IC, Tập $S \subseteq V$ là tập các nút được nghi ngờ là nguồn phát tán thông tin sai lệch và mỗi nút $u \in S$ có xác suất $\rho(u) \geq 0$ là nguồn thông tin sai lệch. Cho ngưỡng phát hiện thông tin sai lệch $\gamma > 0$, bài toán đặt ra là tìm tập nhỏ nhất các nút $A \subseteq V$ để đặt giám sát sao cho $\mathbb{D}(A) \geq \gamma$.

Khi các nút có cùng xác suất là nguồn thông tin sai lệch thì giá hàm phát hiện của tập A bằng giá trị ảnh hưởng của A trên đồ thị ngược lại [9]. Vì vậy, tính toán Hàm ảnh hưởng là #P-khó [17], suy ra tính toán Hàm phát hiện $\mathbb{D}(A)$ cũng là #P-khó. Định lý 1 trong [18] đã chứng minh rằng: MBD chỉ có thể đạt xấp xỉ $(1 - \epsilon) \ln n$ trừ khi $NP \in DTIME(n^{O(\log \log n)})$.

III. ĐỀ XUẤT THUẬT TOÁN

Trong phần này, bài báo ước tính hàm phát hiện trên mô hình IC và đề xuất các thuật toán xấp xỉ, bao gồm: Greedy, SMD và ISMD cho bài toán MBD.

1. Ước tính giá trị hàm phát hiện (hàm mục tiêu)

Phương pháp ước tính hàm phát hiện $\mathbb{D}(\cdot)$ dựa trên tập mẫu phát hiện DS (Detection Sampling) được thể hiện qua định nghĩa sau:

Định nghĩa 2 (DS): Cho đồ thị $G = (V, E)$ trên mô hình IC, đặt $\rho(S) = \sum_{u \in S} \rho(u)$. Gọi R_j là bộ mẫu phát hiện ngẫu nhiên trên đồ thị G , R_j được tạo như sau:

- 1) Chọn nút nguồn $u \in V$ với xác suất $\frac{\rho(u)}{\rho(S)}$.
- 2) Tạo một đồ thị mẫu g từ G , và trả về tập R_j là các nút có thể bắt đầu từ u trong g .

Nút u trong định nghĩa trên được gọi là nguồn của R_j , ký hiệu $src(R_j) = u$. Ω là không gian xác suất của các bộ

DS, trong đó xác suất tạo R_j với nút nguồn u (ký hiệu là $R_j(u)$) được tính là:

$$\Pr[R_j(u) \sim \Omega] = \frac{\rho(u)}{\rho(S)} \cdot \sum_{g \sim G: R_j(g, u)=1} \Pr[g \sim G] \quad (5)$$

Về cơ bản, vai trò của tập DS tương tự như tập RR (Reachable Reverse) được thiết lập trong việc ước tính hàm phát tán [19–23]. Một biến ngẫu nhiên $X_j(A)$ được định nghĩa như sau:

$$X_j(A) = \begin{cases} 1, & \text{Nếu } R_j \cap A \neq \emptyset \\ 0, & \text{ngược lại} \end{cases} \quad (6)$$

Tương tự bổ đề 2 trong [19], với mọi tập nút $A \in V$, và $A \subseteq V$, gọi $\rho(S) = \sum_{u \in S} \rho(u)$ ta có:

$$\mathbb{D}(A) = \rho(S) \cdot \mathbb{E}[X_j(A)] \quad (7)$$

2. Thuật toán Greedy

Thuật toán Greedy chọn nút u cho vào tập A nếu mức tăng lớn nhất trong việc giảm hiệu suất trong mỗi bước, được định nghĩa như sau:

$$\delta(A, u) = \min(\mathbb{D}(A \cup \{u\}), \gamma) - \mathbb{D}(A) \quad (8)$$

cho đến khi $\mathbb{D}(A) \geq \gamma - \epsilon$, tuy nhiên, không thể áp dụng trực tiếp thuật toán cho MXH vì tính toán hàm phát hiện là #P-khó. Vì vậy, bài báo sử dụng mô phỏng Monte-Carlo (MC) để ước tính hàm phát hiện dựa trên xấp xỉ trung bình mẫu của các lần mô phỏng.

Thuật toán 1: Thuật toán Greedy

Input: Đồ thị $G(V, E)$, tập nguồn $S \subseteq V$, ngưỡng γ

Output: Tập nút đặt giám sát A

1. $A \leftarrow \emptyset$;
 2. **while** $\mathbb{D}(A) < \gamma - \epsilon$ **do**
 3. $u \leftarrow \arg \max_{v \in V \setminus S} (\min(\mathbb{D}(A \cup \{v\}), \gamma) - \mathbb{D}(A))$;
 4. $A \leftarrow A \cup \{u\}$;
 5. **end**
 6. **return** A .
-

Theo kết quả chứng minh của bổ đề 2 trong [24], thuật toán Greedy cho tỷ lệ xấp xỉ $(1 - \ln \frac{\gamma}{\epsilon})$ với $\epsilon \in (0, \gamma)$. Gọi R là thời gian mô phỏng MC, độ phức tạp của Greedy là $O(Rnk)$, trong đó k là số vòng lặp trong thuật toán, n là số đỉnh của đồ thị.

3. Thuật toán dựa trên tập mẫu phát hiện - SMD

Thuật toán này kết hợp hai kỹ thuật: (1) tạo ra một bộ các DS đủ lớn để ước tính chức năng phát hiện bằng cách áp dụng lý thuyết Martingale [25] và (2) sử dụng thuật toán Greedy để tìm ra tập A đủ tốt đảm bảo chất lượng lời giải.

Ký hiệu $Cov_{\mathcal{R}}(A) = \sum_{R_j \in \mathcal{R}} \min\{1, |A \cap R_j|\}$ là số bộ DS trong $\mathcal{R}$ được phủ bởi A . Chúng ta thu được ước lượng của $\mathbb{D}(A)$ từ $\mathcal{R}$ như sau:

vì số lượng bộ DS trong $\mathcal{R}$ được bao phủ bởi A . Từ Bổ đề 2 trong [18], chúng ta ước tính $\mathbb{D}(A)$ từ $\mathcal{R}$ như sau:

$$\hat{\mathbb{D}}_{\mathcal{R}}(A) = \frac{\rho(S)}{|\mathcal{R}|} Cov_{\mathcal{R}}(A) \quad (9)$$

Vì $Cov_{\mathcal{R}}()$ là một hàm đơn điệu và submodular, $\mathbb{D}(\cdot)$ cũng có tính chất tương tự. Tập mẫu DS được tạo ra bằng thuật toán 2. Đầu tiên, chọn một nút nguồn u với xác suất $\frac{\rho(u)}{\rho(S)}$ (dòng 1). Sau đó sử dụng hàng đợi Q để lưu trữ các nút đã truy cập. Trong mỗi bước, thuật toán chọn một nút u trong Q và thêm nó vào R_j . Sau đó, chọn từng nút lân cận v với xác suất $p(u, v)$ theo mô hình cạnh trực tuyến (dòng 8) và đưa vào Q . Quá trình này lặp lại cho đến khi Q rỗng.

Thuật toán 2: Thuật toán tạo mẫu phát hiện DS

Input: Đồ thị $G(V, E)$, tập nguồn $S \subseteq V$

Output: tập nút phát hiện R_j

1. Chọn nút $u \in V$ với xác suất $\Pr[u] = \frac{\rho(u)}{\rho(S)}$;
 2. Queue $Q \leftarrow \{u\}$;
 3. **while** Q chưa rỗng **do**
 4. $u \leftarrow Q.pop()$;
 5. $R_j \leftarrow R_j \cup \{u\}$;
 6. **foreach** $v \in N_{out}(u) \setminus R_j$ **do**
 7. **if** $v \notin Q$ **then**
 8. Chọn nút v với xác suất $p(u, v)$;
 9. **if** (v được chọn) **then**
 10. $Q.push(v)$;
 11. **end**
 12. **end**
 13. **end**
 14. **end**
 15. **return** R_j .
-

Thuật toán SMD được thực hiện như sau: Đầu tiên, thuật toán tạo ra tập $\mathcal{R}$ chứa $N = \frac{(2+\frac{2}{3}\epsilon)\rho(S)}{\epsilon^2(\gamma-\epsilon\gamma)} \ln(n/\delta)$ tập DS đảm bảo xấp xỉ (δ, ϵ) cho giải pháp tối ưu A^* , nghĩa là:

$$\Pr[(1+\epsilon)\mathbb{D}_{\mathcal{R}}(A^*) \geq \hat{\mathbb{D}}_{\mathcal{R}}(A^*) \geq (1-\epsilon)\mathbb{D}_{\mathcal{R}}(A^*)] \geq 1 - \delta \quad (10)$$

bằng cách áp dụng lý thuyết martingale [25]. Trong mỗi vòng lặp chọn nút u có giá trị tăng lớn nhất của hàm phát hiện $\delta(A, v)$ như sau:

$$\delta(A, v) = \hat{\mathbb{D}}(A \cup \{u\}) - \hat{\mathbb{D}}(A) \quad (11)$$

Nếu giải pháp hiện tại A đạt giá trị $\hat{\mathbb{D}}(A) \geq (\gamma - \epsilon\gamma) - \epsilon$ thì thuật toán trả về A . Mặt khác, kiểm tra số lượng mẫu hiện tại có đủ cho lần lặp tiếp theo không? Nếu số lượng mẫu vẫn đủ thì di chuyển vào các vòng lặp tiếp theo. Nếu

số lượng mẫu không đủ, sẽ tạo thêm $(N_i - N)$ bộ mẫu mới (dòng 13) và đặt lại giải pháp hiện tại A , tức là $A \leftarrow \emptyset$ (dòng 15). Thuật toán di chuyển đến bước tiếp theo và chỉ kết thúc khi nó đáp ứng được điều kiện $\hat{\mathbb{D}}(A) \geq (\gamma - \epsilon\gamma) - \epsilon$.

Thuật toán 3: Thuật toán phát hiện SMD

Input: $G(V, E)$, $S \subseteq V$, γ , tham số $\epsilon, \delta \in (0, 1)$

Output: Tập nút đặt giám sát A

1. ; $N \leftarrow \frac{(2+\frac{2}{3}\epsilon)\rho(S)}{\epsilon^2(\gamma-\epsilon\gamma)} \ln(n/\delta)$;
 2. Tạo tập $\mathcal{R}$ chứa N tập DS bởi thuật toán 2
 3. $A \leftarrow \emptyset$;
 4. **while** **True** **do**
 5. $u \leftarrow \arg \max_{v \in V \setminus A} (\hat{\mathbb{D}}(A \cup v) - \hat{\mathbb{D}}(A))$
 6. $A \leftarrow A \cup \{u\}$;
 7. **if** $\hat{\mathbb{D}}(A) \geq (\gamma - \epsilon\gamma) - \epsilon$ **then**
 8. **return** A ;
 9. **else**
 10. $i \leftarrow |A| + 1$;
 11. $N_i \leftarrow \frac{(2+\frac{2}{3}\epsilon)\rho(S)}{\epsilon^2(\gamma-\epsilon\gamma)} \ln(\binom{n}{i}/\delta)$;
 12. **if** $N < N_i$ **then**
 13. Tạo thêm $N_i - N$ tập DS đưa vào tập $\mathcal{R}$;
 14. $N \leftarrow N_i$;
 15. $A \leftarrow \emptyset$;
 16. **end**
 17. **end**
 18. **end**
 19. **return** A .
-

4. Thuật toán dựa trên tập mẫu quan trọng - ISMD

Ý tưởng chính của thuật toán là sử dụng các mẫu phát hiện quan trọng và lý thuyết martingale để ước tính hàm phát hiện. Trong thuật toán 4, để tạo tập IDS, đầu tiên chọn nút nguồn IDS (dòng 1). Sau đó, tính toán xác suất $\Pr[E_i], i = 1, \dots, l(u)$ và chọn một nút ra u_i trong $N_{out}(u)$ với xác suất $\frac{\Pr[E_i]}{\phi(u)}$ (dòng 3). Điều này sẽ đảm bảo rằng ít nhất một trong các hàng xóm của u sẽ được kích hoạt. Các nút $v_1, v_2, \dots, v_{j-1}$ vẫn không bị kích hoạt và các nút $v_{i+1}, \dots, v_l$ sau đó được kích hoạt độc lập với xác suất $p(u, v_j)$ (dòng 7). Phần còn lại của thuật toán này tương tự như Thuật toán 2.

Chi tiết của ISMD được trình bày trong Thuật toán 5. Đầu tiên, ISMD tạo bộ IDS thay vì DS (dòng 2) và sử dụng chúng để ước tính chức năng phát hiện. Thứ hai, trong mỗi trình lặp, SMD cần $\frac{(2+\frac{2}{3}\epsilon)\rho(S)}{\epsilon^2(\gamma-\epsilon\gamma)} \ln(\binom{n}{i}/\delta)$ bộ DS nhưng ISMD chỉ cần $\frac{q(2+\frac{2}{3}\epsilon)\rho(S)}{\epsilon^2(\gamma-\epsilon\gamma)} \ln(\binom{n}{i}/\delta)$ tập IDS, với $(q < 1)$. Gọi M là thời gian tạo ra một mẫu, Thuật toán 5 có độ phức tạp thời gian là $O\left(q i_{max} \rho(S) \ln(\binom{n}{i_{max}}/\delta) \epsilon^{-2} M\right)$.

Thuật toán 4: Tạo mẫu phát hiện quan trọng IDS

Input: Đồ thị $G = (V, E)$, tập nguồn $S \subseteq V$

Output: Tập mẫu phát hiện quan trọng R_j ;

1. Chọn nút $u \in V$ với xác suất $\Pr[u] = \frac{\varphi(u)\rho(u)}{\Phi}$;
2. Tính $\Pr[E_i], i = 1 \dots l(u)$;
3. Chọn nút $v_i \in N_{out}(u)$ với xác suất $\frac{\Pr[E_i]}{\varphi(u)}$;
4. $R_j \leftarrow \{u, v_i\}$;
5. Queue $Q \leftarrow \{v_i\}$;
6. **for** $j = i + 1$ **to** l **do**
7. Chọn nút v_j với xác suất $p(u, v_j)$;
8. **if** (v_j được chọn) **then**
9. $Q.push(v_j), R_j \leftarrow R_j \cup \{v_j\}$;
10. **end**
11. **end**
12. **while** Q chưa rỗng **do**
13. $u \leftarrow Q.pop()$;
14. **foreach** $v \in N_{out}(u) \setminus R_j$ **do**
15. **if** $v \notin Q$ **then**
16. Chọn nút v với xác suất $p(u, v)$;
17. **if** (v được chọn) **then**
18. $Q.push(v)$;
19. $R_j \leftarrow R_j \cup \{v\}$;
20. **end**
21. **end**
22. **end**
23. **end**
24. **return** R_j .

Các thuật toán SMD và ISMD cung cấp cùng một kết quả lý thuyết, tuy nhiên, tổng số mẫu ISMD sử dụng thấp hơn SMD và thời gian chạy của ISMD ít hơn thời gian của SMD. Nhận xét này phù hợp với kết quả thực nghiệm trên các tập dữ liệu trong Phần V.

IV. KẾT QUẢ THỰC NGHIỆM

Để đánh giá toàn diện các thuật toán được đề xuất, thực nghiệm tiến hành so sánh các thuật toán đề xuất là Greedy, SMD, ISMD với nhau và so sánh với các thuật toán cơ sở phổ biến là Deegree và Pagerank. Ngoài ra, SMD và ISMD được so sánh với thuật toán OPIM [22], thuật toán lấy mẫu RR mới nhất cho bài toán Tối đa hóa ảnh hưởng. Dữ liệu được lấy từ web: [http://snap.stanford.edu/data/] bao gồm các OSN có quy mô từ hàng nghìn đến hàng triệu cạnh, cụ thể là: Email-Eu-Core [26, 27], Wiki-Vote [28, 29], CA-HepPh [27], Email-Eu-All [27]. Các thuật toán được cài đặt bằng ngôn ngữ Python trên máy tính có cấu hình: CPU Intel Core i7 – 8550U 1,8Ghz, RAM 8GB DDR4 2400MHz, hệ điều hành Linux.

Thuật toán 5: Thuật toán phát hiện ISMD

Input: $G(V, E)$, $S \subseteq V$, ngưỡng $\gamma, \epsilon, \delta \in (0, 1)$

Output: Tập nút đặt giám sát A ;

1. $N \leftarrow \frac{q(2+\frac{2}{3}\epsilon)\rho(S)}{\epsilon^2(\gamma-\epsilon\gamma)} \ln(n/\delta)$;
2. Tạo tập $\mathcal{R}$ chứa N tập mẫu phát hiện IDS;
3. $A \leftarrow \emptyset$;
4. **while** True **do**
5. $u \leftarrow \arg \max_{v \in V \setminus A} (\hat{\mathbb{D}}(A \cup v) - \hat{\mathbb{D}}(A))$;
6. $A \leftarrow A \cup \{u\}$;
7. **if** $\hat{\mathbb{D}}(A) \geq \gamma - \epsilon\gamma - \epsilon$ **then**
8. **return** A ;
9. **else**
10. **end**
11. $i \leftarrow |A| + 1$;
12. $N_i \leftarrow \frac{q(2+\frac{2}{3}\epsilon)\rho(S)}{\epsilon^2(\gamma-\epsilon\gamma)} \ln(\binom{n}{i}/\delta)$;
13. **if** $N < N_i$ **then**
14. Tạo thêm $N_i - N$ tập IDS thêm vào tập $\mathcal{R}$;
15. $N \leftarrow N_i$;
16. $A \leftarrow \emptyset$;
17. **end**
18. **end**
19. **return** A .

Bảng II
DỮ LIỆU THỰC NGHIỆM

Bộ dữ liệu	Nút	Cạnh	Kiểu	Bậc TB
Email-Eu-Core	1,005	25,571	Có hướng	25,44
Wiki-Vote	7,115	103,689	Có hướng	14,57
CA-HepPh	12,008	118,521	Vô hướng	9,87
Email-Eu-All	265,214	420,045	Có hướng	1,58

1. Cài đặt thực nghiệm

Thực nghiệm dựa trên mô hình Trivalency [6, 16, 23, 30] để chọn trọng số của các cạnh. Xác suất ảnh hưởng được chọn ngẫu nhiên từ tập được xác định trước, trong thực nghiệm này, ta chọn $p(u, v) \in \{0, 001, 0, 01, 0, 1\}$. Ý tưởng của mô hình này là các cạnh có trọng số 0,001 thì nút đầu được coi là có mức độ ảnh hưởng thấp, 0,01 tương ứng với mức ảnh hưởng trung bình và 0,1 ảnh hưởng mức cao. Tham số đầu vào được chọn như sau: $\delta = 1/n$ theo các nghiên cứu [20–23]. Các nút nghi ngờ được chọn ngẫu nhiên với kích thước $n/2$ và xác suất $\rho(u)$ được chọn ngẫu nhiên trong $[0, 1]$. Ngưỡng γ và tham số ϵ được chọn tùy thuộc vào quy mô của mạng. Ký hiệu $\Psi = \gamma/\rho(S)$ phản ánh mối quan hệ giữa γ và $\rho(S)$. Giá trị của các tham số này được mô tả trong Bảng III. Trong thuật toán cơ sở, phương pháp Monte Carlo được sử dụng 10.000 lần để ước tính hàm phát hiện. Đối với mỗi thuật toán, chạy 10 lần để lấy kết quả trung bình. Thời gian chạy của mỗi thuật toán

được giới hạn trong vòng 24 giờ.

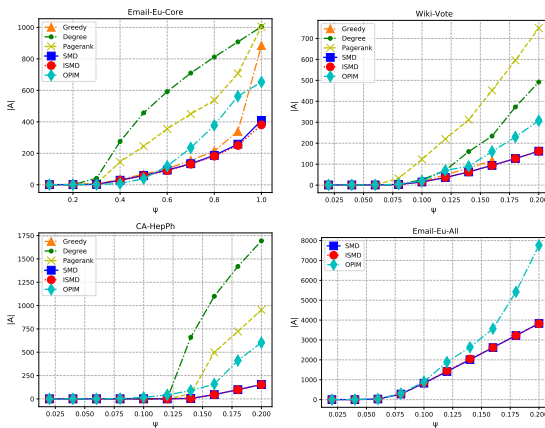
Bảng III
THAM SỐ ĐẦU VÀO CHO CÁC MẠNG

Bộ dữ liệu	$\rho(S)$	Ψ	ϵ
Email-Eu-Core	249,33	1,0	0,01
Wiki-Vote	1784,78	0,2	0,01
CA-HepPh	3009,29	0,2	0,01
Email-Eu-All	171217	0,1	0,1

2. Đánh giá kết quả

a) So sánh hiệu suất thuật toán

Hiệu suất của các thuật toán được xác định bằng kích thước của tập nút đặt giám sát. Thuật toán tốt hơn trả về tập nút đặt giám sát kích thước nhỏ hơn. Hình 1 cho thấy SMD và ISMD có cùng hiệu suất vượt trội hơn nhiều các thuật toán khác trong mọi trường hợp, giá trị của Ψ càng lớn thì khoảng cách vượt càng lớn. Cụ thể, với cùng giá trị Ψ , SMD và ISMD tốt hơn OPIM tới 3,9 lần, tốt hơn Greedy 2,3 lần. Các thuật toán tác giả đề xuất cũng tốt hơn nhiều lần so với các thuật toán cơ sở là Degree và Pagerank. Điều này chứng tỏ rằng thuật toán SMD và ISMD có hiệu suất tốt hơn các thuật toán khác. Ngoài ra, việc ước lượng giá trị hàm phát hiện bằng các bộ mẫu DS và IDS cho kết quả tốt hơn so với mô phỏng MC trong thuật toán Greedy.

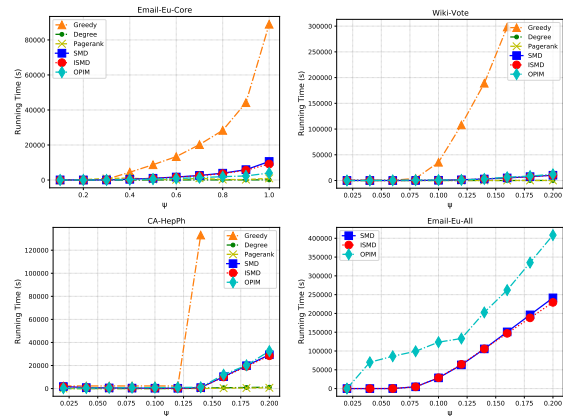


Hình 1. So sánh hiệu suất thuật toán

b) So sánh thời gian chạy thuật toán

Trong thực nghiệm này, thời gian được tính bằng giây (s). Hình 2 cho thấy, thời gian chạy của thuật toán SMD và ISMD đều vượt trội hơn đáng kể so với các thuật toán còn lại. Trong đó SMD, ISMD nhanh hơn OPIM tới 1,5 lần trên hầu hết các mạng, OPIM chỉ cho thời gian tốt hơn trên mạng Email-Eu-Core. Điều này là do OPIM mất nhiều thời gian để tìm kiếm nhị phân cho các giải pháp. So với Greedy, SMD nhanh hơn Greedy tới 10,2 lần và ISMD

nhanh hơn Greedy tới 12,4 lần. Đối với các mạng lớn như Email-Eu All Greedy không thể hoàn thành trong thời gian giới hạn (24 giờ) trong khi các thuật toán SMD và ISMD vẫn hoạt động và cho kết quả tốt. Điều này cho thấy việc ước lượng hàm phát hiện bằng DS và IDS nhanh hơn so với việc sử dụng phương pháp mô phỏng Monte Carlo truyền thống của Greedy. So sánh riêng SMD và ISMD thì thời gian chạy trung bình của ISMD nhanh hơn SMD tới 1,4, nguyên nhân chính là do số lượng mẫu yêu cầu của ISMD thấp hơn so với SMD. Các thuật toán cơ sở có thời gian chạy nhỏ nhất vì chúng là các thuật toán heuristic đơn giản với độ phức tạp thấp.



Hình 2. So sánh thời gian chạy thuật toán

V. KẾT LUẬN

Trong bài báo này, bài toán Ngân sách tối thiểu phát hiện nguồn thông tin sai lệch MBD được nghiên cứu nhằm mục đích tìm ra tập hợp các nút nhỏ nhất để đặt giám sát sao cho phát hiện được thông tin sai lệch từ các nút bị nghi ngờ, chức năng phát hiện dự kiến đảm bảo đạt ít nhất một ngưỡng cho trước $\gamma > 0$. Trên mô hình IC, bài báo đề xuất 03 thuật toán xấp xỉ để giải quyết bài toán MB, bao gồm: Greedy, SMD và ISMD. Thực nghiệm cho thấy thuật toán SMD và ISMD vượt trội hơn các thuật toán khác. Tuy nhiên, thời gian thực hiện thuật toán chưa thật sự tốt để áp dụng cho các mạng có quy mô hàng triệu đỉnh và cạnh. Thời gian tới, chúng tôi nghiên cứu cải thiện thời gian chạy của các thuật toán để có thể áp dụng cho các mạng lớn hơn.

LỜI CẢM ƠN

Công trình này được hỗ trợ bởi Viện Hàn lâm Khoa học và Công nghệ Việt Nam, mã đề tài: VAST01.05 / 21-22.

TÀI LIỆU THAM KHẢO

- [1] "Misinformation on social media led to pune violence: Minister," in <https://www.ndtv.com/mumbai-news/misinformation-on-social-media-led-to-pune-violence-minister-1795562>, 2018.

- [2] P. Domm, "False rumor of explosion at white house causes stocks to briefly plunge; ap confirms its twitter feed was hacked," in Available: <https://www.cnn.com/id/100646197>, 2013.
- [3] V. Luckerson, "Fear, misinformation, and social media complicate ebola fight," in <http://time.com/3479254/ebola-social-media/>, 2014.
- [4] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, 2013, pp. 1103–1108. [Online]. Available: <https://doi.org/10.1109/ICDM.2013.61>
- [5] J. Ma, W. Gao, and K. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 2017, pp. 708–717. [Online]. Available: <https://doi.org/10.18653/v1/P17-1066>
- [6] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, 2010, pp. 88–97. [Online]. Available: <https://doi.org/10.1109/ICDM.2010.118>
- [7] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst, "Patterns of cascading behavior in large blog graphs," in *Proceedings of the Seventh SIAM International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA*, 2007, pp. 551–556. [Online]. Available: <https://doi.org/10.1137/1.9781611972771.60>
- [8] H. Zhang, A. Kuhnle, H. Zhang, and M. T. Thai, "Detecting misinformation in online social networks before it is too late," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016*, 2016, pp. 541–548. [Online]. Available: <https://doi.org/10.1109/ASONAM.2016.7752288>
- [9] H. Zhang, H. Zhang, X. Li, and M. T. Thai, "Limiting the spread of misinformation while effectively raising awareness in social networks," in *Computational Social Networks - 4th International Conference, CSoNet 2015, Beijing, China, August 4-6, 2015, Proceedings*, 2015, pp. 35–47. [Online]. Available: https://doi.org/10.1007/978-3-319-21786-4_4
- [10] H. Zhang, M. A. Alim, X. Li, M. T. Thai, and H. T. Nguyen, "Misinformation in online social networks: Detect them all with a limited budget," *ACM Trans. Inf. Syst.*, vol. 34, no. 3, pp. 18:1–18:24, 2016. [Online]. Available: <http://doi.acm.org/10.1145/2885494>
- [11] C. V. Pham, H. M. Dinh, H. D. Nguyen, H. T. Dang, and H. X. Hoang, "Limiting the spread of epidemics within time constraint on online social networks," in *Proceedings of the Eighth International Symposium on Information and Communication Technology, Nha Trang City, Viet Nam, December 7-8, 2017*, 2017, pp. 262–269. [Online]. Available: <https://doi.org/10.1145/3155133.3155157>
- [12] D. V. Pham, G. L. Nguyen, T. N. Nguyen, C. V. Pham, and A. V. Nguyen, "Multi-topic misinformation blocking with budget constraint on online social networks," *IEEE Access*, vol. 8, pp. 78 879–78 889, 2020.
- [13] E. B. Khalil, B. N. Dilkina, and L. Song, "Scalable diffusion-aware optimization of network topology," in *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 2014, pp. 1226–1235. [Online]. Available: <http://doi.acm.org/10.1145/2623330.2623704>
- [14] C. V. Pham, M. T. Thai, H. V. Duong, B. Q. Bui, and H. X. Hoang, "Maximizing misinformation restriction within time and budget constraints," *J. Comb. Optim.*, vol. 35, no. 4, pp. 1202–1240, 2018. [Online]. Available: <https://doi.org/10.1007/s10878-018-0252-3>
- [15] H. T. Nguyen, A. Cano, V. Tam, and T. N. Dinh, "Blocking self-avoiding walks stops cyber-epidemics: A scalable gpu-based approach," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2019.
- [16] D. Kempe, J. M. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, 2003, pp. 137–146. [Online]. Available: <http://doi.acm.org/10.1145/956750.956769>
- [17] W. Chen, A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rincón, X. Sun, Y. Wang, W. Wei, and Y. Yuan, "Influence maximization in social networks when negative opinions may emerge and propagate," in *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA*, 2011, pp. 379–390. [Online]. Available: <https://doi.org/10.1137/1.9781611972818.33>
- [18] C. V. Pham, D. V. Pham, B. Q. Bui, and A. V. Nguyen, "Minimum budget for misinformation detection in online social networks with provable guarantees," *Optimization Letters*, pp. 1–30, 2021.
- [19] C. Borgs, M. Brautbar, J. T. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, 2014, pp. 946–957. [Online]. Available: <https://doi.org/10.1137/1.9781611973402.70>
- [20] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: near-optimal time complexity meets practical efficiency," in *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, 2014, pp. 75–86. [Online]. Available: <http://doi.acm.org/10.1145/2588555.2593670>
- [21] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, 2015, pp. 1539–1554. [Online]. Available: <http://doi.acm.org/10.1145/2723372.2723734>
- [22] G. Das, C. M. Jermaine, and P. A. Bernstein, Eds., *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*. ACM, 2018. [Online]. Available: <https://doi.org/10.1145/3183713>
- [23] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks," in *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, 2016, pp. 695–710. [Online]. Available: <http://doi.acm.org/10.1145/2882903.2915207>
- [24] A. Goyal, F. Bonchi, L. V. S. Lakshmanan, and S. Venkatasubramanian, "On minimizing budget and time in influence propagation over social networks," *Social Netw. Analys. Mining*, vol. 3, no. 2, pp. 179–192, 2013. [Online]. Available: <https://doi.org/10.1007/s13278-012-0062-z>
- [25] F. R. K. Chung and L. Lu, "Survey: Concentration inequalities and martingale inequalities: A survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, 2006. [Online]. Available: <https://doi.org/10.1080/15427951.2006.10129115>
- [26] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich,

“Local higher-order graph clustering,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, 2017, pp. 555–564. [Online]. Available: <https://doi.org/10.1145/3097983.3098069>

- [27] J. Leskovec, J. M. Kleinberg, and C. Faloutsos, “Graph evolution: Densification and shrinking diameters,” *TKDD*, vol. 1, no. 1, p. 2, 2007. [Online]. Available: <https://doi.org/10.1145/1217299.1217301>
- [28] J. Leskovec, D. P. Huttenlocher, and J. M. Kleinberg, “Signed networks in social media,” in *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010*, 2010, pp. 1361–1370. [Online]. Available: <https://doi.org/10.1145/1753326.1753532>
- [29] —, “Predicting positive and negative links in online social networks,” in *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, 2010, pp. 641–650. [Online]. Available: <https://doi.org/10.1145/1772690.1772756>
- [30] Y. Zhang and B. A. Prakash, “Data-aware vaccine allocation over large networks,” *TKDD*, vol. 10, no. 2, pp. 20:1–20:32, 2015. [Online]. Available: <http://doi.acm.org/10.1145/2803176>

SƠ LƯỢC VỀ TÁC GIẢ

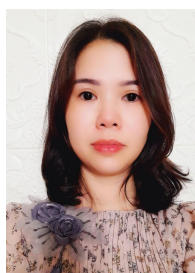
Phạm Văn Dũng



Nhận bằng thạc sĩ Công nghệ thông tin năm 2012 tại Học viện Kỹ thuật Quân sự. Đang làm nghiên cứu sinh tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Hiện là giảng viên Khoa An ninh thông tin, Học viện An ninh nhân dân. Lĩnh vực nghiên cứu: Tối ưu, phân tích mạng xã hội, an ninh mạng và phòng chống tội phạm sử dụng công nghệ cao.

Email: pvdungc500@gmail.com

Nguyễn Thị Tuyết Trinh



Nhận bằng thạc sĩ Công nghệ thông tin năm 2012 tại Học viện Kỹ thuật quân sự. Hiện là giảng viên bộ môn Toán tin, chuyên viên phòng Công nghệ thông tin, Học viện Y Dược học cổ truyền Việt Nam. Lĩnh vực nghiên cứu: Lý thuyết đồ thị, cơ sở dữ liệu, phân tích thiết kế hệ thống an toàn thông tin, IoT, Big Data. Email: trinhnt83@gmail.com

Vũ Chí Quang



Nhận bằng thạc sĩ Công nghệ thông tin năm 2008 tại Đại học Công nghệ - ĐHQGHN. Đang làm nghiên cứu sinh tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Hiện là giảng viên chính Khoa An ninh thông tin, Học viện An ninh nhân dân.

Lĩnh vực nghiên cứu: Lý thuyết đồ thị, tối ưu hóa, cơ sở dữ liệu, IoT, Big Data, an

toàn hệ thống thông tin, an ninh mạng.

Email: quangvc.hvan@gmail.com

Hà Thị Hồng Vân



Nhận bằng thạc sĩ Quản lý nhà nước về an ninh và trật tự an toàn xã hội tại Học viện Cảnh sát năm 2013.

Hiện là Phó trưởng phòng công nghệ và giải pháp phần mềm tại Viện Công nghệ thông tin thuộc VHL Khoa học và Công nghệ Việt Nam.

Lĩnh vực nghiên cứu: Quản lý nhà nước về an ninh mạng.

Email: vanha2809@gmail.com

Nguyễn Việt Anh



Nhận bằng Tiến sĩ tại Đại học Kyoto, Nhật Bản năm 2012.

Hiện là nghiên cứu viên cao cấp tại Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Lĩnh vực nghiên cứu: Machine learning, graph mining, and social network analysis.

Email: Email:anhnv@ioit.ac.vn

Searching 3D Motion Patterns of Vietnamese Traditional Dances

Le Thi Thanh Van¹, Vu Ngoc Quang¹, Pham Thanh Huyen², Ma Thi Chau¹, Le Thanh Ha¹

¹ HMI Laboratory, University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam

² Faculty of Information technology, Ha Long University, Quang Ninh, Vietnam

Correspondence: Le Thi Thanh Van, lvank52thb@live.com

Date received: 23/09/2021, Date revised: 29/10/2021, Date accepted: 15/11/2021

Digital Object Identifier: 10.32913/mic-ict-research-vn.v2021.n2.1020

Abstract: Vietnam has many traditional dances in old theatres such as Xoan singing, “tuồng” or “chè”. They all urgently need to be preserved in digital formats, especially in 3D motion capture format for dances. In digital formats, they bring many values such as the ability to automatically classify and search for content of dances’ movement. In this paper, we propose a system for 3D movement search of Cheo dance’s postures and gestures. The system applies sliding window technique, Dynamic Time Warping algorithm and a novel feature selection method named CheoAngle. Results show that the proposed system reach good scores in several metrics. We also compare CheoAngle with other feature selection methods for 3D movement and show that CheoAngle give the best results.

Keywords: *Motion search, motion recognition, similarity matching, dynamic time warping*

Title: **Tìm kiếm mẫu chuyển động 3D của múa truyền thống Việt Nam**

Abstract: Việt Nam có nhiều thể loại múa truyền thống trong nghệ thuật sân khấu cổ truyền như hát Xoan, tuồng, chèo. Tất cả những hình thức nghệ thuật này đều cần được bảo tồn ở dạng số hóa, đặc biệt là dạng chuyển động 3D cho múa. Trong dạng số hóa, chúng có nhiều giá trị như khả năng phân loại tự động và tìm kiếm nội dung chuyển động múa. Trong bài nghiên cứu này, chúng tôi đề xuất hệ thống tìm kiếm chuyển động 3D cho các tư thế và động tác của múa Chèo. Hệ thống áp dụng kỹ thuật sliding window, thuật toán Dynamic Time Warping và phương pháp chọn đặc trưng mới (tên là CheoAngle). Những kết quả thực nghiệm cho thấy hệ thống đề xuất cho kết quả tốt ở một vài phép đo. Chúng tôi cũng so sánh CheoAngle với những phương pháp chọn đặc trưng khác cho chuyển động 3D và chỉ ra rằng CheoAngle cho kết quả tốt nhất.

Keywords: *Tìm kiếm chuyển động, nhận diện chuyển động, so sánh tương đồng, dynamic time warping*

I. INTRODUCTION

Vietnamese traditional dance is one of basic types in Vietnamese long-established theatres [1–3]. They combined many factors, which consists of music, songs, art and literary screenplay, in order to represent a character’s personalities, psychological states... “chè” (Cheo dance) is a particular-popular traditional dance form whose postures and motions mimicked life’s activities, plants and animals [1].

Need of digitalizing 3D motion for the Vietnamese traditional dance is one of necessary tasks. The traditional dances in 3D format help researchers to analyze and check basic postures (each posture is an unmoving position of body parts at a time (see Figure 1, 2 and 3) at different views. Thus, teaching dance also become more easily and feasibly while combining 3D technology. However, until



Figure 1. Arm cluster in Cheo dance. From left to right, 5 basic postures include: “múa chạy đàn”, “dâng hoa”, “dâng rượu”, “múa cướp bông”, “tấu nhạc”. Source: disanso.vn

now, we do not have any system which supports to manage 3D dataset of the traditional dance efficiently. To manage the data, the system must have operations such as search, edition, insertion and deletion. Accordingly, search is one of the most important operations. Indeed, this paper suggests a way to store 3D data of the traditional dance, define 3D basic movements of the traditional dance and build



Figure 2. Leg cluster in Cheo dance. From left to right, 5 basic postures include: “chữ V”, “chữ chi”, “quả trám”, “chữ đinh”, “đệm gót”. Source: disanso.vn



Figure 3. Sit cluster in Cheo dance. From left to right, 5 basic postures include: “chân chống chân quỳ”, “hai đầu gối cùng quỳ”, “ngồi hai chân co về một bên”, “duỗi thẳng hai chân”, “hai chân bắt chéo”. Source: disanso.vn

an algorithm for searching 3D movement of traditional dances in order to support making a system which manage traditional dance data in 3D space.

Cheo dance has the most systematic teaching [4] in types of Vietnamese traditional dance. It divides postures into 3 clusters: arm (see Figure 1), leg (see Figure 2) and sit (see Figure 3). Each cluster has 5 basic postures and their variants. For example, in the posture “dâng hoa” [4], both arms put on two body sides and move the same direction. Accordingly, any motion satisfying these two criterions will belong to the posture “dâng hoa” (see Figure 4). This motion is called as a variant of the posture “dâng hoa”. However, each posture has several special gestures which often mimic daily activities. For instance, gesture “gánh” of posture “tấu nhạc” [4] imitates an action “carry rice” of farmers. This gesture requires a performer to raise up one hand (at shoulder-equal position) and put down the other while arms and legs are swinging up and down. Additionally, combination of 3 clusters is in need of attention. At the same time, a performer only can use one leg posture or one sit posture. Unlike leg and sit postures, one arm posture can combine any leg or sit posture concurrently. To simplify, the leg and sit cluster is grouped into the same leg-sit cluster. It is worth noting that Cheo dance has basic postures (or movements that position of body features is changed from time to time) more complex than other traditional dances. For instance, in Hat Xoan Phu Tho songs [3], the palm moves in/out or ups/downs; the leg moves horizontally or around.

From above observations, in this paper, we have chosen 3D dataset of Cheo dance [5] as a representative in order to build a proposal algorithm. It was captured from an Xsens MVN motion capture system (a product of Xsens

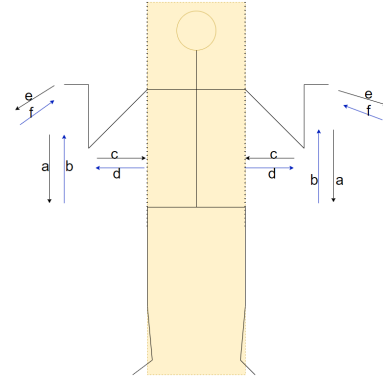


Figure 4. Variants of posture “dâng hoa”. A variant is a motion of two arms in the same direction – namely a, b, c, d, e, f – at two body sides in the extent which is the non-yellow region. The same motion is displayed by the same arrow color (black or blue).

Technologies B.V.) at format BVH (biovision hierarchy animation file). This file stores 2 parts: hierarchy of joints of a human skeleton and their 3D coordinates’ joints of the skeleton in each frame. Here, BVH file displays each frame as one time and uniquely interprets motion of one performer.

Problem of dance detection and recognition at 3D space has been studied in a long time. These works constructed algorithms for little gestures [6–13]. In fact, dance genres have no much gestures. So that, designing new algorithms for gestures recognition becomes more easily. For instance, Björn Rennhak and colleagues suggested a technique “turning motions” [6] which would detect poses in order to interpolate trajectory from state to state (in skills of Learning from Observation). This technique was applied for a traditional Japanese folk dance which only consists of 8 simple gestures which were captured by Vicon Motion iQ System equipment with 38 markers. The technique “turning motions” found out intersection points of forearms after projecting them orthogonally from 3D to 2D space. In spite of focusing parts of causing motion, this method was not likely to classify for same-position motions. Moreover, deep learning approach is also applied for dance dataset [8, 9, 12, 13]. This approach tends to 2 branches: 3D reconstruction from 2D images/videos, 3D dataset. However, to use deep learning algorithms, researchers must collect a huge number of samples for training process. Indeed, this paper suggests a way to store 3D data of the traditional dance, define 3D basic movement of traditional dances in order to support making a system which manages traditional dance dataset in 3D space.

II. RELATED WORKS AND METHODS

1. Problem

Problem of searching 3D motion patterns of Vietnamese traditional dances must find out each pattern's start point and end point and classify patterns in 3D space. Searching these points can be solved by many techniques. However, sliding window technique (see II.3.a) is likely to scan almost appearance of patterns. So that, we are interested in the classification step. In fact, pattern classification in 3D motion is considered as a problem of comparing similarity level between two motions. This problem is defined as below:

A given pattern set $A = \{A_1, A_2, \dots, A_n\}$, where A_i is a pattern, and a movement data X . The problem requires to search patterns A_i in X .

Here, a pattern is a basic posture or its certain variant at 3D space while a movement data can be a dance or compound of several postures or variants in 3D space.

2. Related works

In 3D movement, motions are analyzed and stored as human skeleton joints in the three-dimensional space in time order. A pattern is a 3D motion segment which shows a complete meaning action. To search for patterns in 3D movements, we need to detect and classify patterns.

a) Pattern detection

Patterns detection is an important step of searching patterns in time series. Approach Time series segmentation [14] has already used in a long time due to its superior features. Three main approaches consist of sliding window, top-down, bottom-up. The result of sliding window is segments of data which can contain patterns for the pattern classification step. We call these segments are proposals.

b) Pattern classification

Patterns classification is to compare predefined patterns to a proposal and choose the pattern which matches the best with the proposal. To do this, we need a technique to compare similarity between them. According to Anna Sebernegg [15], there are many approaches for this problem, such as time alignment, trajectory-based and rule-based. However, time alignment is the most popular due to its feasibility. Several representative algorithms with this approach are Dynamic Time Warping (DTW), hidden Markov model (HMM), etc...

c) Feature extraction

Feature extraction of the 3D motion data affects an algorithm's performance significantly. In 2017, Yusuke Goutsu, Wataru Takano and Yoshihiko Nakamura found two

matters with 3D motion features [16]. Firstly, local features, which got from some parts of the body, would give better results than global features, which collected the whole body. Secondly, the local feature often attaches a certain aim of an action. In 2019, Sheng Li and colleagues realized that only some joints of a human skeleton impacted a motion [17]. Then, this idea was applied in a 3D compression algorithm (MJS).

3. Related methods

a) Sliding window

Sliding window technique uses a window (or kernel) to slide through positions on an array. In this paper, the technique will be used for creating a list of proposals for each posture group.

b) Dynamic Time Warping algorithm (DTW)

DTW algorithm [18–21] are used to compare a similarity ability between a pattern and a proposal. In DTW, the series X and Y can be shrunk according to the following fixed rules:

- The first positions of 2 series are mapped
- The last positions of 2 series are mapped
- A position at X can be mapped into multiple-position at Y , and vice versa
- All positions must have their mapping
- Do not exist values such that $i < j$, $k < l$ and X_i is mapped to Y_l , X_j is mapped to Y_k . From the above rules, we have a distance between two series in the DTW algorithm:

$$DTW(X, Y) = \begin{cases} 0 & \text{if the length of } X \text{ and } Y \text{ are equal,} \\ \infty & \text{if the length of } X \text{ or } Y \text{ is } 0, \\ dist(X_1, Y_1) + \min(& \\ \quad DTW(Rest(X), Y), DTW(X, Rest(Y)), & \\ \quad DTW(Rest(X), Rest(Y)) & \\ \text{other cases} & \end{cases} \quad (1)$$

In equation (1), $dist(X_1, Y_1)$ is the distance between X_1 and Y_1 . It can be calculated by using the Euclidean distance, the Manhattan distance, etc. We choose Euclidean distance in this paper. We use a modification for the algorithm that is to use a fixed window length w such that X_i only maps to Y_j if $|i - j| < w$. In this way, it is likely to reduce significantly the running time of the DTW algorithm while the result is not affected much.

III. PROPOSED SYSTEM OF SEARCHING 3D MOTION PATTERNS

A dance is a consecutive sequence of several basic postures and their variants. In this paper, we consider each



Figure 5. Some pattern instances of ARM2. From left to right, pattern instances include posture “dâng hoa (cao)”, posture “dâng hoa (hạ)”, gesture “bay”, gesture “chống sườn”, gesture “đưa thoi”.

posture group consists of a basic posture and its whole variants (including named gestures of the posture). It means that each posture group at 3D space is called as a pattern. For instance, in Cheo dance, we have 5 patterns in the arm cluster: “múa chạy đàn” (ARM1), “dâng hoa” (ARM2), “dâng rượu” (ARM3), “múa cướp bông” (ARM4) and “tấu nhạc” (ARM5); 10 patterns in leg-sit cluster: “chữ V” (LEG1), “chữ chi” (LEG2), “quả trám” (LEG3), “chữ đỉnh” (LEG4), “đệm gót” (LEG5), “chân chống chân quỳ” (SIT1), “hai đầu gối cùng quỳ” (SIT2), “ngồi hai chân co về một bên” (SIT3), “đuỗi thẳng hai chân” (SIT4) and “hai chân bắt chéo” (SIT5). Now, as a case in point, let’s look at the dance “Múa phối kết hợp”. This dance uses 4 arm patterns, 5 leg patterns and 1 sit patterns with the different contribution (see Table I and II). In addition, each representative of a pattern is called as a pattern instance or instance. Thus, each pattern has one or more pattern instances (see Figure 5). A 3D movement is a dance, mixing between basic postures (or its variants) or a complex gesture at 3D space.

We can now define the problem of searching 3D motion patterns of Vietnamese traditional dances. Its input consists of a 3D movement and a set of pattern instances. Its output is a list of recognized patterns and their positions in the input. Our proposed system stores a list of pattern instances of basic postures and a list of angles for each pattern (called a list of pattern angles). Accordingly, the system will use those to recognize patterns through 2 main steps (Figure 6): pattern detection and pattern classification.

1. Choose a list of angles as features of a pattern (CheoAngle)

It is easy to see that angles calculation from joints will help us skip a data normalization step. The fact that Cheo data use many joints to display human motion in 3D space. As joints’ coordinates are local, we must convert these local coordinates into global ones through tool “bvh converter”¹. Then, we choose 19 joints for calculation (see Figure 7). In addition, angles of each joint at 2-axels motion data do not depend on these series’ coordinate system. Moreover, to decrease noises, with patterns of arm groups, we only consider the above half of the body (from hips to head);

¹<https://github.com/tekulvw/bvh-converter>

with patterns of leg and sit groups, we focus on the under half of the body (from hips to feet). Besides, based on the idea of a robot arm in kinematics [22], in the arm cluster, the line created from two shoulders is a base, each wrist is an end effector; in the leg-sit cluster, the line created from two hips are a base, each ankle is an end effector. Focusing on these joints, we will build a list of angles for each group basing on their impacts. Moreover, we also can get additional angles from other joints. These angles are calculated following the formula

$$\theta = \arccos\left(\frac{\vec{AB} \cdot \vec{A'B'}}{|\vec{AB}| \cdot |\vec{A'B'}|}\right) \quad (2)$$

, where $\vec{AB}$ and $\vec{A'B'}$ are two vectors.

To facilitate for analysis, we use joint name in the vector notion. We assume that vector A_B connects joint A to B; vector C_D connects joint C to D. Accordingly, angle $\angle(A_B, C_D)$ is constructed by the two vectors A_B and C_D . For instance, the arm pattern ARM5 (group “tấu nhạc”) requires to raise up one hand and put down the other (see Figure 8). A list of angles for ARM5 needs satisfying criterions: features of the posture “tấu nhạc”, differentiation between the posture “tấu nhạc” and the other ones. The posture “tấu nhạc” is different from the other ones on rules: pattern ARM1’s attributes are put hands on the chest at the middle of body and move them the same direction; hands in pattern ARM2 must be put along the body and displaced in a similar manner (see Figure 4); in pattern ARM3, hands are in the same body side and move in the similar manner; pattern ARM 4’s attributes are that hands in the same extent and move in the opposite direction. Thus, we have the list of angles for ARM5 comprises $\angle(LSHOULDER_LWRIST, RSHOULDER_RWRIST)$, $\angle(LWRIST_RWRIST, LSHOULDER_RSHOULDER)$ and $\angle(LELBOW_LWRIST, RELBOW_RWRIST)$. In the first angle, $LSHOULDER_LWRIST$ and $RSHOULDER_RWRIST$ show one arm goes up and the other goes down. The other angles fulfill the second criterion.

2. Detect patterns

This section will give details about the generation of proposals for each instance. Here, this section consists of

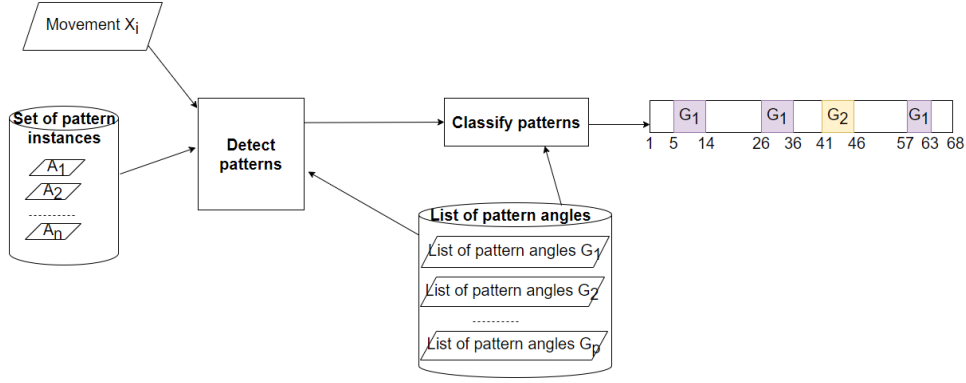


Figure 6. The system architecture. Accordingly, movement X_i is a whole dance or an excerpt of a dance at 3D space in time order (called time series); output is a list of pattern names ($G_1, G_2, \dots$) and their positions on a time series (in this figure, its length is 68); a set of pattern instances ($A_1, A_2, \dots, A_n$) is different samples of each posture group (called pattern); a list of pattern angles has angles which are calculated for a pattern.

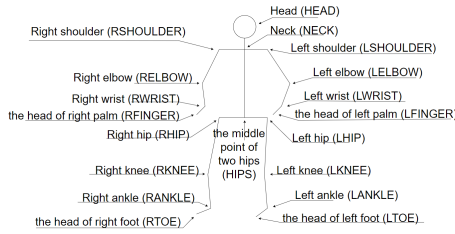


Figure 7. Joints of a human skeleton are used in the proposed system.

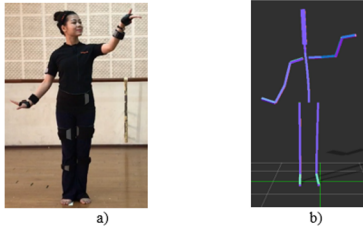


Figure 8. The posture “tầu nhac”: a) 2D space b) 3D space. Source: disanso.vn

3 main steps: make a list of proposals, filter proposals, normalize proposals.

a) Make a list of proposals

The system makes a list of proposals for each instance A_j , which appears in the movement X_i , through 3 main steps (see Figure 9): extract feature, general proposal $proposal_k$, measure similarity. Steps are described more details as below:

- Extract feature: for both movement X_i and instance A_j , we extract features by calculating angles (using equation (2)) in the same list of pattern angles. Take the pattern instance “gánh” (has 76 frames) of pattern ARM5 and dance “múa phối kết hợp” (has 6906

frames) as an example. From the list of angles for ARM5 (see III.1), we calculate value of 3 angles and push into an array at each frame of this instance. As an effect, we obtain the instance feature as a 76×3 matrix. In like manner, we also possess the movement feature as a 6906×3 matrix.

- General proposal $proposal_k$: proposals $proposal_k$ will be created by using the sliding window technique with a proper window size following the length of the instance and a sliding step a . In this case, the sliding step’s default is 5 (see Figure 10). If the instance length is smaller than the default value, the sliding step will be the instance length.
- Measure similarity: the system sequentially compares features at each frame in A_j to features at each frame in a proposal $proposal_k$ (use equation (1)). The system returns a cost $cost_{A_j,k}$ of the optimal path between A_j and $proposal_k$ (called the similarity cost). If the similarity cost is 0, A_j is like $proposal_k$. Otherwise, the more the cost is, the more they are different.

b) Filter proposals

Filter proposals in the same instance A_j which has the length n_{A_j} , coefficient $\beta_{A_j} > 0$ and the number of angles n_θ as steps below:

- If the instance is a posture which has mostly no change of joints’ position, $thresh\theta := n_{A_j} * \alpha$ where α is constant. Otherwise, $thresh\theta := n_{A_j} * \beta_{A_j} * n_\theta$.
- Assume that A_j has a set of proposals $proposal_k$ and $cost_{A_j,k}$ is the similarity cost between A_j and $proposal_k$. We choose $proposal_k$ if $cost_{A_j,k} < thresh\theta$.

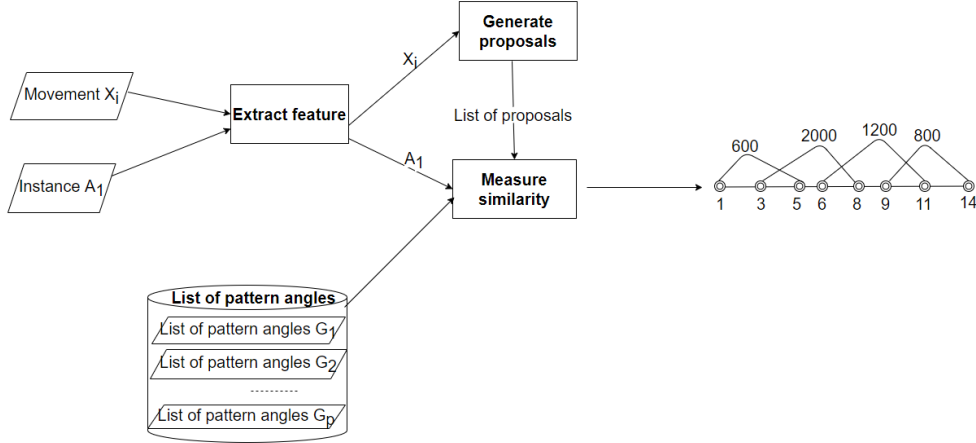


Figure 9. The workflow of the step Make a list of proposals for a pattern instance A_1 .

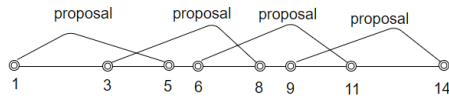


Figure 10. Several proposals are generated with the length 5 and sliding step 3

c) Normalize proposals

The system is only interested in proposals of instances in the same pattern. Two tasks need to do sequentially as:

- Merge proposals, which overlap each other in the same instance, into the unique one. As an illustration, in the Figure 10, we assume that the first and second and forth proposal are chosen. Since the first and second proposal is overlapped each other, we have to unify them into a unique one, which is a segment from frame 1 to 8. This proposal is written as $\langle 1, 8 \rangle$.
- Merge proposals, which duplicate and overlap each other of instances in the same pattern, into discrete ones. For example, pattern ARM1 has instances A_1 , A_2 and A_3 . Accordingly, A_1 consists of proposals: $\langle 1, 4 \rangle$, $\langle 6, 8 \rangle$, $\langle 12, 18 \rangle$; A_2 comprises proposals: $\langle 4, 8 \rangle$, $\langle 12, 15 \rangle$; A_3 encompasses proposals: $\langle 1, 8 \rangle$. After process, ARM1 will have discrete proposals: $\langle 1, 8 \rangle$, $\langle 12, 18 \rangle$.

3. Classify patterns

The system will do 2 steps: reorder proposals and attach a label to each pattern.

a) Reorder proposals

The system performs to reorder proposals in according to the following steps:

- The system lists the head and tail positions of the whole proposals of groups (called original proposals) in order to put on a list L.
- Duplicated elements in L are removed and the remain of L is ordered increasingly before putting on a list L2.
- Based on L2, create all possible proposals. From that, choose proposals such that they are related to at least an original proposal. Finally, we have a list of new proposals.

b) Attach each label to a pattern

Assume that there is a set of n patterns $\{group_k\}_n$ which has the same proposal SG . In pattern $group_k$, compare instances of A_j the pattern to SG , then calculate the similarity: $sim_{group_k} = \max\{sim_{group_k, A_j}\}$ such that $sim_{group_k, A_j} = \frac{n_{optimalSG}}{l_{SG}}$, where $n_{optimalSG}$ is the number of different positions of SG on the optimal path, l_{SG} is the length of SG . The label of SG is $\max\{sim_{group_k}\}$. In a particular case of choosing the label between the first arm group (posture “múa chạy đàn”) and the forth arm group (posture “cướp bông”), the system calculates angles in the same list of angles (called ARM1_4) which is different with ones of arm group 1 and 4. This procedure is also applied for arm group 3 and 4. For leg-sit cluster, at one time, only one group can appear. To enforce this rule, the system must classify labels which appear leg and sit groups in the same proposal.

IV. EXPERIMENTS AND RESULTS

1. Dataset

We use 3D movement data of Cheo dance [5] for experiments. This 3D dataset is a part of ANIAGE (High Dimensional Heterogeneous Data based Animation

Techniques for Southeast Asian Intangible Cultural Heritage Digital Content)², conducted by HMI Laboratory³. The dataset is in BVH format that stores human skeleton motion at 3D space. We use this dataset to construct a pattern database of 70 patterns and a test dataset of 7 files. This test dataset is called HMI-3DCheo.

For patterns, basic postures are represented by 5 frames while their variants have variable lengths. The two reasons for this rule. First, a posture is body features not vary over time. It means that a posture only needs 1 frame. However, the DTW algorithm calculates total cost of the optimal path for two time series. In equation (1), this algorithm must get values of 3 neighbors for the considered element. It means that each series needs at least 3 frames. To avoid noise, we choose the length as 5 frames. Second, a variant is a meaning motion in dance. In fact, gestures have 3 stages: start, main, end. The stage “main” is very important because it is likely to decide the gesture detection. Hence, we tried to get this stage for each instance as possible as we could. As an evidence, gesture “bay” mimics a flying action of a bird (see Figure 5). It takes 130 frames to describe the flying action.

HMI-3DCheo consists of “múa phối kết hợp” (6906 frames), compound of gestures “quay lưu không nam và nữ” (3149 frames), compound of gestures “đi nữ lệch và nam ngang” (2744 frames), gesture “đi quả trám” (1037 frames), gesture “lão say” (1604 frames), excerpt “Sứ Vân giả dại” (“xe tơ – dệt vải”, 6218 frames), “Thị Kính hát sù rầu ba than” (6804 frames). Distribution of patterns in the test data is shown in the Table I and II.

2. Metrics

We use precision, recall and F1-score to evaluate experimental outcomes. They are calculated for an instance A_j [?] as belows:

$$precision_{A_j} = \frac{TruePred_{A_j}}{TotalPred_{A_j}} \quad (3)$$

$$recall_{A_j} = \frac{TruePred_{A_j}}{Real_{A_j}} \quad (4)$$

$$F1 - score_{A_j} = \frac{2 * precision_{A_j} * recall_{A_j}}{precision_{A_j} + recall_{A_j}} \quad (5)$$

where $TruePred_{A_j}$ is the length of true A_j predicted segments on the whole test data; $TotalPred_{A_j}$ is the total of A_j predicted segments on the whole test data; $Real_{A_j}$ is the total of real A_j segments on the whole test data.

²<https://www.euh2020aniage.org/>

³<https://hmiuet.wordpress.com/projects/>

3. Comparison methods

Since each 3D movement of the dance is represented by a time series, we need to compare with other methods for time series such as edit distance on real sequence (EDR) [21, 23–25] and Longest common subsequence (LCSS) [20, 21, 26]. Some system constraints must be modified for these methods. In the step Make a list of proposals (see III.2.a), cost matrices of EDR and LCSS are edited. Assume that we have a proposal $X = \{x_i\}_{i=1}^m$ and an instance $A = \{y_j\}_{j=1}^n$. Accordingly, $x_i = \{\theta_k\}_{k=1}^p$ and $y_j = \{\omega_k\}_{k=1}^p$ are features (each feature consists of p angles). The condition for x_i and y_j being alike in LCSS is $\sqrt{\sum_{k=1}^p (\theta_k - \omega_k)^2} \leq \frac{\phi * \pi}{180} * \sqrt{n}$ and not being alike in EDR is $\sqrt{\sum_{k=1}^p (\theta_k - \omega_k)^2} > \frac{\phi * \pi}{180} * \sqrt{n}$, where ϕ is the acceptable angle deviation. At the method LCSS, the longest common substring in matrix T is named ω . At the method EDR, the least edition operation on X and A in matrix T is named n_{EDR} . In step Filter proposals (see III.2.b), LCSS only selects proposals which satisfy $n_{\omega} > 0.4 * n$ while EDR filter following the condition: $n_{EDR} < 0.4 * \max(n, m)$. Consequently, the step Attach each label to a pattern (see III.3.b) is modified as $sim_A^X = \frac{n_{\omega}}{n}$ at LCSS and $sim_A^X = \max(n, m)$ at EDR. In addition, in the EDR method, label of SG is $\min\{sim_A^X\}$.

On the other hand, we cannot apply a deep learning model for this problem. The reason is too little training data. In fact, we have only 70 instances of 15 patterns for training process, which means an average of 4.6 instances per pattern. As a result, a training model would likely overfit the small training dataset.

4. Results and discussion

We performed experiments on the dataset HMI-3DCheo. This paper pays attention on the efficiency of DTW algorithm and the feature selection technique CheoAngle.

The Table V and IV show experimental results in order to compare values among methods. Generally, the precision and recall averages of DTW better than ones of LCSS and EDR. In the arm cluster, recall averages of DTW, LCSS and EDR are 49.78%, 46.74%, 35.82%. In leg-sit cluster, precision averages of DTW, LCSS and EDR are 66.36%, 61.47%, 60.82%; their recall averages are 62.55%, 54.99%, 52.72%.

To prove the efficiency of CheoAngle, this paper bases on two comparison methods: ShengLi [17] and JointRelativeDist [27]. The method ShengLi considers joints which cause a movement. Accordingly, the paper only applies this way to build a list of angles for each cluster (arm, leg, sit) while CheoAngle makes a list of angles for each posture

Table I
DISTRIBUTION OF PATTERNS OF ARM GROUP FOR TEST DATA

Pattern Test data	ARM1	ARM2	ARM3	ARM4	ARM5
múa phối kết hợp (6906 frames)	35.9%	23.8%	10.0%	0.0%	21.4%
quay lưu không nam và nữ (3149 frames)	0.0%	100.0%	0.0%	0.0%	0.0%
đi nữ lệch và nam ngang (2744 frames)	36.0%	64.0%	0.0%	0.0%	0.0%
đi quả trám (1037 frames)	0.0%	100.0%	0.0%	0.0%	0.0%
lão say (1604 frames)	0.0%	12.0%	32.4%	0.0%	29.3%
xe tơ – dệt vải (6218 frames)	16.3%	22.0%	23.5%	9.2%	26.2%
Thị Kính hát sử rầu ba than (6804 frames)	23.2%	32.1%	38.9%	2.2%	1.5%
HMI-3DCheo	21.3%	39.8%	18.7%	2.5%	12.9%

Table II
DISTRIBUTION OF PATTERNS OF LEG AND SIT GROUP FOR TEST DATA

Pattern Test data	LEG1	LEG2	LEG3	LEG4	LEG5	SIT1	SIT2	SIT3	SIT4	SIT5
múa phối kết hợp (6906 frames)	37.0%	37.2%	2.8%	7.1%	6.0%	9.7%	0.0%	0.0%	0.0%	0.0%
quay lưu không nam và nữ (3149 frames)	20.9%	9.1%	14.1%	27.9%	23.2%	0.0%	0.0%	0.0%	0.0%	0.0%
đi nữ lệch và nam ngang (2744 frames)	27.2%	71.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
đi quả trám (1037 frames)	7.1%	24.4%	34.8%	0.0%	11.2%	0.0%	0.0%	0.0%	0.0%	0.0%
lão say (1604 frames)	14.7%	20.2%	30.2%	31.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
xe tơ – dệt vải (6218 frames)	11.8%	6.8%	0.06%	0.1%	5.5%	0.8%	5.9%	51.5%	16.1%	0.2%
Thị Kính hát sử rầu ba than (6804 frames)	22.6%	56.3%	1.5%	0.0%	0.9%	0.7%	16.4%	0.0%	0.0%	0.0%
HMI-3DCheo	23.0%	33.9%	5.6%	6.6%	5.9%	2.7%	5.2%	11.2%	3.5%	0.05%

Table III
TRIAL OUTCOMES ON THE ARM CLUSTER

Methods	Metrics	ARM1	ARM2	ARM3	ARM4	ARM5
DTW	Precision (%)	65.7	85.6	51.2	1.6	66.7
	Recall (%)	75.6	72.3	29.4	4.1	67.5
	F1-score (%)	70.3	78.4	37.3	2.3	67.1
LCSS	Precision (%)	75.4	71.1	63.1	0.0	58.3
	Recall (%)	74.3	28.0	49.9	0.0	81.5
	F1-score (%)	74.8	40.2	55.7	0.0	68.0
EDR	Precision (%)	91.6	60.2	74.8	0.0	65.3
	Recall (%)	70.2	22.8	35.6	0.0	50.5
	F1-score (%)	79.5	33.1	48.3	0.0	57.0

group. The method JointRelativeDist relatively calculates a distance between a joint pair of 2 skeletons. From this idea, we construct a way to choose features for a certain skeleton. Assume that $P_1 = \{u_k\}_{k=1}^P$ is a considered skeleton. The Euclidean distance between two points $u_k = (u_k^x, u_k^y, u_k^z)$ and $u_o = (u_o^x, u_o^y, u_o^z)$ on P_1 is:

$$D_{P_1}(u_k, u_o) = \sqrt{(u_k^x - u_o^x)^2 + (u_k^y - u_o^y)^2 + (u_k^z - u_o^z)^2} \quad (6)$$

From the equation (6), we describe a feature of the skeleton P_1 whose is obtained from the method JointRelativeDist is:

$$F_{P_1} = \{D_{P_1}(u_k, u_o)\}_{i=1}^q \quad (7)$$

where q is the number of point-need-calculated pairs of the skeleton. Here, the paper defines joint pairs for arm,

Table IV
TRIAL OUTCOMES ON THE LEG-SIT CLUSTER

Methods	Metrics	LEG1	LEG2	LEG3	LEG4	LEG5	SIT1	SIT2	SIT3	SIT4	SIT5
DTW	Precision (%)	83.8	72.5	34.1	53.2	35.4	66.4	74.6	99.0	96.5	48.1
	Recall (%)	44.7	41.0	30.0	69.1	34.8	94.4	23.4	93.5	94.6	100.0
	F1-score (%)	58.3	52.4	31.9	60.1	35.1	78.0	35.7	96.2	95.5	65.0
LCSS	Precision (%)	81.0	56.4	25.0	56.8	34.9	82.0	74.5	95.1	86.8	22.2
	Recall (%)	65.0	87.0	1.8	41.0	33.5	88.6	7.7	97.8	96.7	30.8
	F1-score (%)	72.1	68.4	3.4	47.6	34.1	85.2	14.0	96.4	91.5	25.8
EDR	Precision (%)	82.4	63.2	62.5	36.5	32.3	90.2	29.3	97.7	91.9	22.2
	Recall (%)	54.5	83.6	11.2	39.8	33.9	80.9	1.6	91.1	99.8	30.8
	F1-score (%)	65.6	72.0	19.0	38.1	33.1	85.3	3.1	94.3	95.7	25.8

leg and sit cluster instead of posture groups.

To compare the proposed method CheoAngle with ShengLi and JointRelativeDist following F1-score, we can see that this way is efficient on arm and leg-sit cluster. In the arm cluster (Table V), values F1-score of patterns ARM1, ARM2, ARM3 and ARM5 are higher than ones of ShengLi and JointRelativeDist. In particular, the pattern ARM4 has the highest result at JointRelativeDist. In the leg-sit cluster (Table VI), 8/10 patterns at CheoAngle have F1-score values higher than ones of the method remains. In particular, F1-score values of the pattern ARM1 and ARM2 at ShengLi is higher than others.

On the contrary, the result for pattern ARM4 is low for all methods. There might be several reasons for this result. First, the number of instances for pattern ARM4 are very small compared to others (see Table I and II). Meanwhile, pattern ARM4 has many variants in which hands are extended and moved in opposite direction. These invariants are also similar to variant of pattern ARM1 and ARM3. As a result, it is very easy to mistake an instance of ARM4 for ARM1 and ARM3. Second, the contribution of this pattern in the test dataset is very little (see Table I and II). Moreover, the appearance of pattern ARM4 in the test dataset often complicated cases while the ARM4 instances are basic samples of the basic posture “múa cướp bông”. As an example, in the excerpt “Súy Vân giả dại”, action “xâu kim” (thread needle), at the middle of the body, uses a ARM5 variant (prepare for needle thread) and a ARM4 variant (threading needle). And so, the system can consider this ARM4 variant as a ARM1 case. This means that if ARM4 predictions are incorrect, ARM1 and ARM3 outcomes also are affected seriously. Similarly, we can explain the unlogic results of leg-sit patterns. Moreover, we have difficulty distinguishing leg-sit patterns. It is because their attributes are not much

dissimilar. To illustrate, we analyze LEG2 and LEG5. In gesture “chân chữ chi” (LEG2), a performer must lift the heels (LEG5).

V. CONCLUSIONS

Results at Section IV proves the efficiency of applying several techniques. The technique sliding window supports to search patterns on the global scope. The algorithm DTW makes patterns classification to become simple. In addition, the way to select need-to-calculated angles of motions also contributes an important role to predictions as well as the data scale.

Besides advantages, the proposed algorithm has a small group of limitations. The first limitation is the dependency on pattern instances’ length. Instances have short or long length depends on dance rhythm and skills of the dancer. Thus, the pattern’s length affects the proposals filter. The second limitation is the window size and the stride. If the stride is short, it takes a long time to run the algorithm. Otherwise, the long stride will skip patterns that we need to search. Meanwhile, the shorter window size is, the more difficult the system classify patterns due to getting only a part of information in the pattern. In contradiction, the longer window size is, the more easily the system predicts pattern incorrectly.

In the future, we will optimize lists of pattern angles and complete instance and test dataset. For pattern angles optimization, we improve techniques. In classification process, a group having few highlight features is easy to predict incorrect patterns. Moreover, several patterns barely have no movement, but some joints change positions very slightly. This will cause the difficulty for comparison between groups which are in the same posture. Thus, we will add the information of rotation degree of joints in order to

Table V
FEATURES COMPARISON ON THE ARM CLUSTER

Methods	Metrics	ARM1	ARM2	ARM3	ARM4	ARM5
CheoAngle	Precision (%)	65.7	85.6	51.2	1.6	66.7
	Recall (%)	75.6	72.3	29.4	4.1	67.5
	F1-score (%)	70.3	78.4	37.3	2.3	67.1
ShengLi	Precision (%)	61.0	93.8	27.1	0.0	44.2
	Recall (%)	76.5	46.4	30.4	0.0	78.0
	F1-score (%)	67.9	62.1	28.7	0.0	56.4
JointRelativeDist	Precision (%)	84.7	96.1	8.9	11.8	46.4
	Recall (%)	48.6	38.8	4.9	33.3	42.8
	F1-score (%)	61.8	55.3	6.3	17.4	44.5

Table VI
FEATURES COMPARISON ON THE LEG-SIT CLUSTER

Methods	Metrics	LEG1	LEG2	LEG3	LEG4	LEG5	SIT1	SIT2	SIT3	SIT4	SIT5
CheoAngle	Precision (%)	83.8	72.5	34.1	53.2	35.4	66.4	74.6	99.0	96.5	48.1
	Recall (%)	44.7	41.0	30.0	69.1	34.8	94.4	23.4	93.5	94.6	100.0
	F1-score (%)	58.3	52.4	31.9	60.1	35.1	78.0	35.7	96.2	95.5	65.0
ShengLi	Precision (%)	87.9	69.1	59.8	36.2	61.8	25.1	5.8	95.6	87.5	0.0
	Recall (%)	58.1	70.7	3.5	32.4	21.4	94.4	11.2	16.2	73.0	0.0
	F1-score (%)	69.9	69.9	6.6	34.2	31.8	39.6	7.6	27.7	79.6	0.0
JointRelativeDist	Precision (%)	77.0	56.9	8.2	24.3	59.2	7.5	0.0	0.0	0.0	0.0
	Recall (%)	62.7	11.1	0.8	2.7	4.6	87.7	0.0	0.0	0.0	0.0
	F1-score (%)	69.1	18.6	1.5	4.9	8.6	13.8	0.0	0.0	0.0	0.0

complement patterns comparison. For the instance dataset, we must add almost representatives for patterns: ARM2, ARM4, LEG1, SIT2, SIT4, SIT5. For the test dataset, we must give dances or compounds of basic postures for patterns: ARM2, ARM3, ARM4, LEG3, LEG4, LEG5, SIT1, SIT2, SIT3, SIT4, SIT5.

REFERENCES

- [1] Lê Ngọc Canh, *Nghệ thuật múa Chèo*. Nhà xuất bản Sân khấu, 2003.
- [2] Nhà hát tuồng Việt Nam, “Múa tuồng (vũ đạo),” 2020. [Online]. Available: <http://vietnamtuongtheatre.com/mua-tuong-vu-dao/>
- [3] Nguyễn Bích Thủy, “Nghệ thuật trình diễn hát xoan phú thọ,” 2018. [Online]. Available: http://svhttdl.phutho.gov.vn/tin/nghe-thuat-trinh-dien-hat-xoan-phu-tho_1010.html
- [4] Trần Thị Ngọc, *Giáo trình múa Chèo*. Trường đại học Sân khấu và Điện ảnh Hà Nội, 1998.
- [5] HMI Laboratory, VNU University of Engineering and Technology, “3d dataset for the traditional cheo dance,” 2021, accessed: 2021-11-16. [Online]. Available: <http://disanso.vn/vidman>
- [6] B. Renhak, T. Shiratori, S. Kudoh, P. Vinayavekhin, and K. Ikeuchi, “Detecting dance motion structure using body components and turning motions,” *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2264–2269, 2010.
- [7] E. Protopapadakis, A. Grammatikopoulou, A. Doulamis, and N. Grammalidis, “Folk dance pattern recognition over depth images acquired via kinect sensor,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-2/W3, pp. 587–593, 2017. [Online]. Available: <https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W3/587/2017/>
- [8] Z. Li, “Three-dimensional diffusion model in sports dance video human skeleton detection and extraction,” *Advances in Mathematical Physics*, vol. 2021, pp. 1–11, sep 2021. [Online]. Available: <https://doi.org/10.1155/2021/3772358>
- [9] X. Hu and N. Ahuja, “Unsupervised 3d pose estimation for hierarchical dance video recognition,” *CoRR*, vol. abs/2109.09166, 2021. [Online]. Available: <https://arxiv.org/abs/2109.09166>
- [10] F. A. Damastuti, A. K. Nurindiyani, and D. Pramadihanto, “Gesture 3d modeling for traditional javanese dance,” in *2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*, 2018, pp. 69–73.
- [11] J. Sutopo, M. K. A. Ghani, B. M. Aboobaidar, and Zulhawati, “Dance gesture recognition using space component and effort component of laban movement analysis,” *International Journal of Scientific & Technology Research*, vol. 9, no. 2, 2020.
- [12] N. Jain, V. Bansal, D. Virmani, V. Gupta, L. Salas-Morera, and L. Garcia-Hernandez, “An enhanced deep convolutional neural network for classifying indian classical dance forms,” *Applied Sciences*, vol. 11, no. 14, 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/14/6253>

- [13] E. Protopapadakis, A. Voulodimos, A. Doulamis, S. Camarinopoulos, N. Doulamis, and G. Miaoulis, "Dance pose identification from motion capture data: A comparison of classifiers," *Technologies*, vol. 6, no. 1, 2018. [Online]. Available: <https://www.mdpi.com/2227-7080/6/1/31>
- [14] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "Segmenting time series: A survey and novel approach," in *Data mining in time series databases*. World Scientific, 2004, pp. 1–21.
- [15] A. S. . P. K. . H. Kaufmann, *Motion Similarity Modeling A State of the Art Report*, 8 2020.
- [16] G. Yusuke, T. Wataru, and N. Yoshihiko, "Classification of multi-class daily human motion using discriminative body parts and sentence descriptions," *Int. J. Comput. Vision*, vol. 126, no. 5, p. 495–514, May 2018. [Online]. Available: <https://doi.org/10.1007/s11263-017-1053-3>
- [17] S. Li, J. Tingting, T. Yonghong, and H. Tiejun, "3d human skeleton data compression for action recognition," in *2019 IEEE Visual Communications and Image Processing (VCIP)*, Dec 2019, pp. 1–4.
- [18] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *KDD Workshop*, 1994, pp. 359–370.
- [19] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, no. 3, pp. 358–386, mar 2005. [Online]. Available: <https://link.springer.com/article/10.1007/s10115-004-0154-9>
- [20] M. Vlachos, D. Gunopulos, and G. Kollios, "Discovering similar multidimensional trajectories," *Proceedings 18th International Conference on Data Engineering*, pp. 673–684, 2002.
- [21] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. J. Keogh, "Indexing multi-dimensional time-series with support for multiple distance measures," in *KDD '03*, 2003.
- [22] H. Václav, "Robot kinematics," 2005. [Online]. Available: <http://people.ciirc.cvut.cz/~hlavac/TeachPresEn/51Robotics/11KinematicsRobot.pdf>
- [23] L. C. . M. T. Özsu , Vincent Oria, "Robust and fast similarity search for moving object trajectories," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 491–502, 2005.
- [24] L. Allison, "Dynamic programming algorithm (dpa) for edit-distance," Faculty of Information Technology (Clayton), Monash University, Australia, 1999. [Online]. Available: <http://users.monash.edu/~lloyd/tildeAlgDS/Dynamic/Edit/>
- [25] L. V. I. and I. V., "Binary codes capable of correcting deletions, insertions and reversals," *SPhD*, vol. 10, p. 707, 1966. [Online]. Available: <https://ui.adsabs.harvard.edu/abs/1966SPhD...10..707L/abstract>
- [26] G.-M. I. of Electronics and C. Science, "Longest common subsequences," France, 1998. [Online]. Available: <https://www-igm.univ-mlv.fr/~lecroq/seqcomp/node4.html>
- [27] J. K. T. Tang, H. Leung, T. Komura, and H. P. H. Shum, "Emulating human perception of motion similarity," *Comput. Animat. Virtual Worlds*, vol. 19, no. 3–4, p. 211–221, sep 2008.



Le Thi Thanh Van received Master degree of Computer Science from University of Engineering and Technology, Vietnam National University, Hanoi in 2021. Her research includes computer vision, machine learning.
Email: lvank52thb@live.com



Vu Ngoc Quang received Master degree of Computer Science from Japan Advanced Institute of Science and Technology, Japan in 2016. He is working as Research Assistant in HMI Laboratory, University of Engineering and Technology, Vietnam National University, Hanoi.

His research includes machine learning and brain-computer interface.

Email: quang.vn@outlook.com



Pham Thanh Huyen received Master's degree in Computer Science at Hanoi National University of Education in 2010. She is a Ph.D student in the University of Engineering and Technology, Vietnam National University, Hanoi. She is a lecturer of Ha Long University in Quang Ninh province.

Her research includes database, data mining, machine learning.

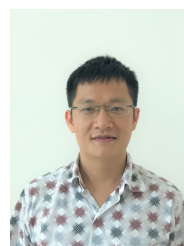
Email: phamthanhuyen@daihochalong.edu.vn



Ma Thi Chau received Ph.D degree from University of Engineering and Technology, Vietnam National University, Hanoi in 2014.

Her research includes Computer Graphics and 3D Reconstruction.

Email: chaumt@vnu.edu.vn



Le Thanh Ha received Ph.D degree from Korea University, Korea in 2010.

He is the head of HMI Laboratory, University of Engineering and Technology, Vietnam National University, Hanoi.

His research includes Video Coding and Communication, Image Processing, Computer Vision and Satellite Image Processing.

ing.

Email: ltha@vnu.edu.vn

THÔNG TIN DÀNH CHO TÁC GIẢ

Các công trình nghiên cứu, phát triển và ứng dụng Công nghệ Thông tin và Truyền thông - gọi tắt là Công nghệ TT'TT, phiên bản tiếng Anh là ICT Research – là chuyên san khoa học của Tạp chí Thông tin và Truyền thông, do Bộ Thông tin và Truyền thông xuất bản từ năm 1999. Chuyên san Công nghệ TT'TT có phản biện, mở trực tuyến (open access) và được phát hành hàng năm với Ấn phẩm tiếng Việt (ISSN: 1859-3526) có 2 số và Ấn phẩm tiếng Anh (ISSN: 1859-3534) – Research and Development on Information and Communication Technology (ICT Research) – có 2 số.

Chuyên san Công nghệ TT'TT là một diễn đàn dành cho các nhà nghiên cứu và các chuyên gia phổ biến các ý tưởng mới và sáng tạo trong các lĩnh vực của Công nghệ Thông tin, Truyền thông và Điện tử ở Việt Nam và trên toàn thế giới. Hội đồng biên tập của Chuyên san CNTT-TT bao gồm các nhà khoa học uy tín đến từ các đơn vị nghiên cứu ở Việt Nam và các quốc gia khác.

HƯỚNG DẪN DÀNH CHO CÁC TÁC GIẢ

Các tác giả nên tìm hiểu cách viết bài báo kỹ thuật, chẳng hạn như các hướng dẫn trong tài liệu “How to write for technical periodicals & conferences” của Viện Kỹ sư Điện, Điện tử Hoa Kỳ (IEEE).

Chuyên san Công nghệ TT'TT chỉ nhận các bản thảo được nộp trực tuyến trên trang web của Chuyên san Công nghệ TT'TT tại <http://ictmag.vn>. Tác giả cần chế bản bản thảo theo mẫu mà Chuyên san Công nghệ TT'TT hướng dẫn.

Các tác giả phải chịu trách nhiệm trong việc đảm bảo rằng bản thảo này chưa từng được công bố trước đó, cũng như hiện đang không trong quá trình phản biện hoặc chuẩn bị xuất bản ở nơi khác. Tác giả liên hệ cũng phải chịu trách nhiệm đảm bảo sự đồng ý của tất cả các đồng tác giả khi nộp bài cho Chuyên san Công nghệ TT'TT. Bản thảo cần được nộp dưới dạng PDF. Để hỗ trợ phản biện kín hai chiều, tác giả cần nộp hai phiên bản của bản thảo, một phiên bản có đầy đủ thông tin của tất cả các tác giả, một không có thông tin gì.

Các bản thảo nộp lại sau khi chỉnh sửa cũng cần tuân theo các chỉ dẫn như của bản thảo nộp lần đầu, nhưng cần kèm theo một văn bản giải trình các ý kiến của phản biện, trong đó giải thích các nội dung được chỉnh sửa và/hoặc trả lời các ý kiến của các phản biện. Sau khi một bản thảo được chính thức chấp nhận xuất bản, các tác giả cần phải chuẩn bị bản thảo cuối cùng bằng LATEX theo mẫu trình bày cho IEEE Transactions.

ĐẠO ĐỨC TRONG XUẤT BẢN

Chuyên san Công nghệ TT'TT cam kết tuân thủ các quy định, tiêu chuẩn và hướng dẫn được thiết lập bởi COPE (Committee on Publication Ethics) trong việc đảm bảo tính liêm chính của tạp chí cũng như xử lý các trường hợp vi phạm. Tất cả các bên tham gia (Biên tập viên, Tác giả, Phản biện và Nhà xuất bản) cần phải hiểu và tuân thủ các quy định, tiêu chuẩn và hướng dẫn được công bố trên trang web của COPE.

Ban biên tập của Chuyên san Công nghệ TT'TT sẽ sàng lọc tất cả các bản thảo trước khi công bố. Tất cả các bản thảo sẽ được kiểm tra về mức độ tương tự với các công trình khác sử dụng các công cụ trực tuyến có được. Mọi hình thức đạo văn không được chấp nhận và bản thảo sẽ bị từ chối.

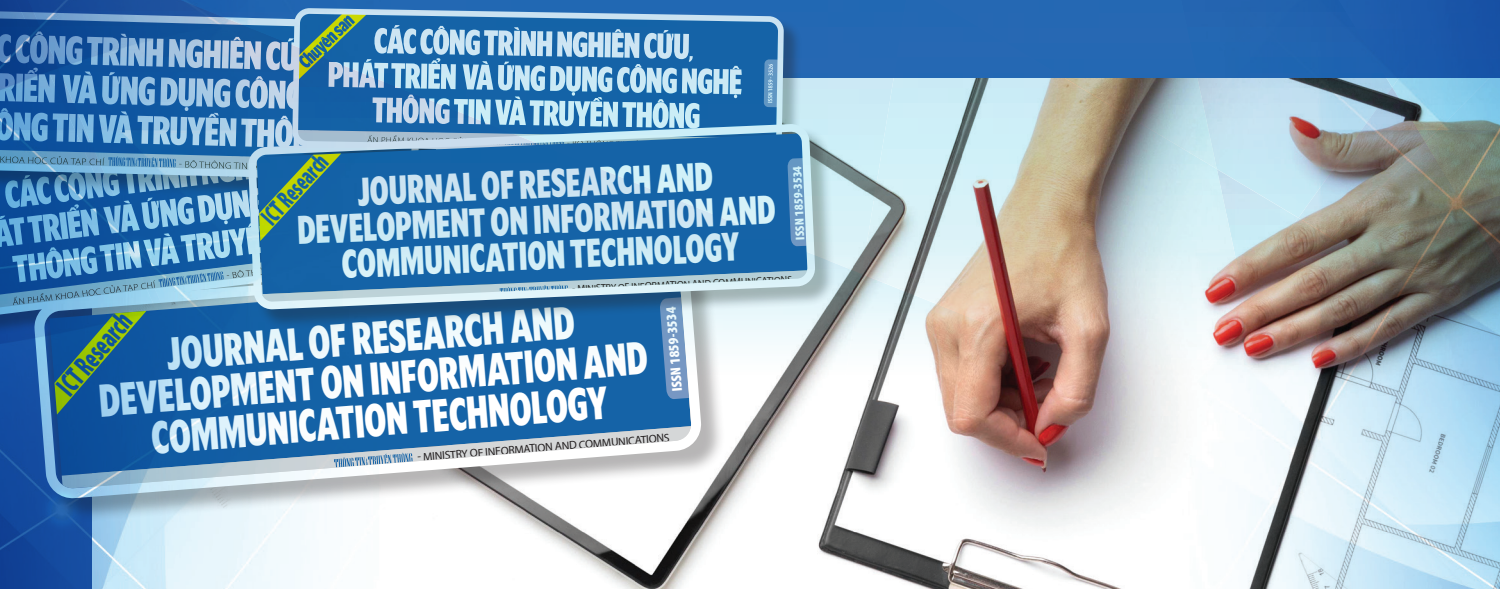
Các kết luận phản biện gồm có: (A) Chấp nhận, (B1) Chấp nhận với yêu cầu chỉnh sửa nhỏ, (B2) Chỉnh sửa và nộp lại (bản thảo cần phải viết lại với các thay đổi lớn và sau đó cần qua vòng phản biện lần thứ hai), và (C) Từ chối.

Mỗi vòng phản biện có thể cần khoảng hai tháng. Trong trường hợp các tác giả được Biên tập viên yêu cầu chỉnh sửa bản thảo, phiên bản được chỉnh sửa cần được nộp trong vòng hai tháng đối với yêu cầu chỉnh sửa lớn và một tháng với yêu cầu chỉnh sửa nhỏ. Tác giả nào cần nhiều thời gian hơn xin vui lòng liên hệ với Biên tập viên.

Mỗi bài báo được chấp nhận sẽ được xuất bản trực tuyến ngay sau đó, với số DOI, sau khi được biên tập và chế bản theo các quy tắc biên tập của Chuyên san Công nghệ TT'TT. Đồng thời, khi có đủ một số lượng bài phù hợp, Chuyên san Công nghệ TT'TT sẽ tập hợp lại và xuất bản một số in, kèm với mục lục các bài báo.

THÔNG TIN VỀ BẢN QUYỀN

Khi bài báo được chấp nhận, bản quyền của bài báo sẽ tự động được chuyển giao cho Tạp chí Công nghệ Thông tin và Truyền thông, là tổ chức chịu trách nhiệm xuất bản Chuyên san Công nghệ TT'TT, với đảm bảo rằng mọi nội dung của bài báo được chấp nhận xuất bản sẽ được công bố miễn phí trên trang web của Chuyên san Công nghệ TT'TT. Việc chuyển giao này cho phép bên xuất bản có toàn quyền giải quyết bất kỳ khiếu nại nào về việc sử dụng trái phép. Các tác giả có quyền đăng tải lại hay sử dụng lại bài báo được xuất bản dưới bất kỳ hình thức nào cho bất kỳ mục đích nào, miễn là có trích dẫn đầy đủ tới bài báo gốc xuất bản trên Chuyên san Công nghệ TT'TT. Việc nộp bài báo được xem như là các tác giả đã đọc và đồng ý với chính sách bản quyền này.



CHƯƠNG TRÌNH HỖ TRỢ NCS, CHỦ NHIỆM ĐỀ TÀI CÔNG BỐ KẾT QUẢ NGHIÊN CỨU KHOA HỌC

Các công trình nghiên cứu, phát triển và ứng dụng công nghệ thông tin và truyền thông (Công nghệ TTTT) là ấn phẩm khoa học hàng đầu trong lĩnh vực CNTT&TT, có trong danh sách tính điểm ở mức cao của Hội đồng Giáo sư Nhà nước với cả ấn phẩm tiếng Việt và tiếng Anh.

Tiếp nối truyền thống 22 năm và hướng tới hội nhập quốc tế, Chuyên san triển khai **chương trình hỗ trợ** các nghiên cứu sinh (NCS), chủ nhiệm đề tài các cấp công bố kết quả nghiên cứu khoa học **hoàn toàn miễn phí** đến hết năm 2022, cụ thể như sau:

- Đảm bảo tiến độ: có kết quả phản biện kín **trong vòng 1 tháng** kể từ khi nộp bài đúng mẫu trực tuyến trên trang <https://ictmag.vn>
- Hỗ trợ hoàn thiện bài báo chế bản bằng LATEX với template được cung cấp sẵn trên website;

- Hỗ trợ chuyển ngữ sang tiếng Anh: bài nộp tiếng Việt được xử lý bình thường. Khi được chấp thuận đăng, nếu tác giả có nhu cầu, sẽ được hỗ trợ chuyển ngữ để chuyển sang đăng trên Chuyên san tiếng Anh – **Journal of Research and Development on Information and Communication (ICT Research)**;

Hội đồng biên tập Chuyên san cũng sẵn sàng tham gia, tài trợ hội nghị, hội thảo khoa học của các đơn vị nhằm cùng Ban Tổ chức phát hiện các nội dung nghiên cứu có khả năng phát triển tốt có thể công bố trên Chuyên san. Những kết quả nghiên cứu có thể chưa mới trên thế giới nhưng **triển khai thực sự, có tác dụng lan tỏa ở Việt Nam** cũng được khuyến khích.

Mọi thông tin chi tiết xin liên hệ: TS. Nguyễn Quang Hưng, Thư ký HĐBT; Tạp chí Thông tin và Truyền thông; Tầng 5, số 115 Trần Duy Hưng, Q. Cầu Giấy, HN. ĐT: (024) 3773 7136; website: <https://ictmag.vn>; e-mail: chuyensanbcvt@mic.gov.vn