Các công trình nghiên cứu, phát triển và ứng dụng
Công nghệ Thông tin và Truyền thông
https://www.ictmag.vn/cntt-tt
Tập 2022, Số 1, Tháng 6
Số đặc biệt CITA 2022

MỤC LỤC

Các công trình nghiên cứu, phát triển và ứng dụng
Công nghệ Thông tin và Truyền thông
https://www.ictmag.vn/cntt-tt
Tập 2022, Số 1, Tháng 6
Số đặc biệt CITA 2022

TABLE OF CONTENT

Tổng quan ứng dụng học máy trong dự đoán nguy cơ đa di truyền hướng tới y học cá thể hóa

Trịnh Thị Xuân[1], Tạ Văn Nhân[2], Hoàng Đỗ Thanh Tùng[3], Trương Nam Hải[4], Trần Đăng Hưng[5]
[1] Khoa Công nghệ thông tin, Đại học Mở Hà Nội
[2] Công ty TNHH LOBI Việt Nam, Hà Nội
[3] Phòng Nghiên cứu hệ thống và quản lý, Viện Công nghệ Thông Tin, VAST
[4] Phòng Kỹ thuật di truyền, Viện Công nghệ Sinh học, VAST
[5] Khoa Công nghệ thông tin, Đại học Sư phạm Hà Nội

Tác giả liên hệ: Trịnh Thị Xuân, Email: trinhxuan@gmail.com
Ngày nhận bài: 06/01/2022, ngày sửa chữa: 21/04/2022, ngày duyệt đăng: 10/05/2022
Định danh DOI: 10.32913/mic-ict-research-vn.v2022.n1.1027

Tóm tắt: Trong thời gian gần đây, Điểm nguy cơ đa di truyền (Polygenic Risk Score - PRS) được xem như một công cụ tiềm năng cho y học chính xác dựa trên các biến dị di truyền phổ biến có đóng góp từ nhỏ tới vừa đối với nguy cơ mắc bệnh di truyền, nhưng tổng gộp các biến dị này lại có thể nâng cao giá trị dự đoán bệnh trong quần thể. Đã có nhiều phương pháp học máy được đưa ra nhằm cải tiến khả năng dự đoán của PRS cũng như những nỗ lực để đưa PRS vào ứng dụng trong lâm sàng. Mặc dù vậy, việc lựa chọn phương pháp một cách hệ thống và những ứng dụng của PRS vẫn chưa thực sự rõ ràng. Vì vậy, trong bài báo tổng quan này, chúng tôi cung cấp một cái nhìn tổng quan về điểm nguy cơ đa di truyền và các nghiên cứu cải tiến sử dụng học máy nhằm nâng cao khả năng áp dụng trong lâm sàng của PRS.

Từ khóa: *Bệnh phổ biến, điểm nguy cơ đa di truyền, GWAS, SNP, mảng SNP, học máy.*

Title: **An Overview of Machine Learning Applications in Polygenic Risk Prediction Towards Personalized Medicine**

Abstract: In recent times, the Polygenic Risk Score (PRS) has been considered as a potential tool for precision medicine based on common genetic variants with small to moderate contributions to the genetic disease risk, but the aggregation of these variations can enhance the predictive value of disease in the population. Many machine learning methods have been proposed to improve the predictive ability of PRS as well as efforts to bring PRS into clinical application. However, the systematic selection of methods and applications of PRS are still not really clear. Therefore, in this review, we provide an overview of polygenic risk scores and innovative studies using machine learning to improve the clinical applicability of PRS.

Keywords: *Common diseases, polygenic risk scores, GWAS, SNP, SNP arrays, machine learning.*

I. GIỚI THIỆU

Với sự phát triển của công nghệ sinh học, các nhà khoa học có thể dựa vào dữ liệu DNA để dự đoán nguy cơ mắc bệnh ở người. Ngoại trừ các đột biến xôma, dữ liệu DNA không thay đổi trong suốt thời gian sống của chúng ta. Vì vậy, nguy cơ mắc bệnh liên quan đến gen của một người có thể xuất hiện ngay từ khi người đó sinh ra. Điều đó cho thấy việc xác định nguy cơ mắc bệnh về gen có ý nghĩa rất lớn trong y tế dự phòng. Chẳng hạn, một nghiên cứu vào năm 2017 ước lượng rằng có khoảng 72% phụ nữ thừa hưởng đột biến gen *BRCA1* và khoảng 69% phụ nữ thừa hưởng đột biến gen *BRCA2* sẽ phát triển ung thư trước tuổi 80 [1]. Việc phát hiện một bệnh nhân mang các biến dị có hại sẽ hỗ trợ bác sỹ đưa ra lời khuyên giúp thay đổi lối sống theo hướng tích cực hoặc đưa ra các can thiệp phòng ngừa tùy theo mức độ nguy cơ.

Với các bệnh do một gen quy định, việc ước lượng nguy cơ mắc bệnh của một người có thể chỉ là tìm kiếm các biến dị có hại trên một gen nào đó. Các nghiên cứu phân tích liên kết di truyền (Genetic Linkage Analysis) đã được phát triển từ lâu để xác định vị trí của gen bệnh dựa trên mối liên kết của nó với các vị trí đánh dấu di truyền trên

nhiễm sắc thể. Phương pháp này đã rất thành công khi tìm ra các đột biến của một số bệnh đơn gen như múa giật [2, 3] hay ung thư vú [4]. Tuy nhiên, phân tích liên kết chưa cho thấy sự hiệu quả đối với các bệnh đa di truyền và phổ biến. Từ năm 2005, các nghiên cứu tương quan toàn hệ gen (Genome-Wide Association Studies - GWAS) đã bắt đầu tìm kiếm các đa hình đơn nucleotit (Single-Nucleotide Polymorphism, SNP) phổ biến (tần số alen phụ, minor allele frequency, MAF $\geq$ 1%) [5–7]. GWAS ngày càng được tạo điều kiện thuận lợi nhờ sự phát triển của các mảng SNP (SNP array) với chi phí tương đối thấp, các mảng này có thể chứa từ 200.000 đến 2.000.000 SNP [8]. Ngoài ra, quá trình phân tích tổng hợp các nghiên cứu tương quan toàn hệ gen riêng lẻ cũng góp phần tạo ra các bộ dữ liệu GWAS ngày càng lớn hơn. Các SNP bị thiếu trên mảng SNP sẽ được bổ sung thông qua quá trình suy diễn thống kê từ các SNP đã được quan sát và các kiểu gen đơn bội (Haplotype) của dữ liệu tham chiếu Chính nhờ quá trình này mà ta có được các tập dữ liệu GWAS lớn như UK Biobank: 96 triệu biến dị cho gần 500 nghìn cá thể mà ban đầu chỉ có khoảng 800 nghìn SNP được xác định bởi kiểu gen [9]. Mặc dù vậy, GWAS vẫn tỏ ra hạn chế trong dự đoán các bệnh đa di truyền, cho dù sử dụng các SNP có tương quan cao với bệnh thì dự đoán cũng không thực sự chính xác [10, 11]. Để nâng cao hiệu suất dự đoán, các nghiên cứu tập trung vào việc lựa chọn tập SNP đặc trưng, thực sự ảnh hưởng đến bệnh. Các phương pháp truyền thống thường dựa vào đặc điểm sinh học và thống kê, điển hình như loại bỏ các SNP do mất cân bằng liên kết hoặc do di truyền cùng nhau mà không ảnh hưởng đến bệnh, giữ lại các SNP có tương quan cao với bệnh. Với một bệnh cụ thể, mặc dù mỗi SNP đặc trưng chỉ ảnh hưởng một phần nhỏ tới bệnh, nhưng kết hợp chúng với nhau có thể giải thích một phần đáng kể tỷ lệ mắc bệnh trong quần thể [8, 12–14].

Để lựa chọn tập SNP đặc trưng, ngoài các phương pháp dựa vào đặc điểm sinh học, các phương pháp áp dụng học máy cũng đang nở rộ trong thời gian gần đây. Cụ thể, ta có thể đưa vấn đề về bài toán lựa chọn đặc trưng [15–17], đặc biệt là các phương pháp lựa chọn đặc trưng áp dụng cho dữ liệu chiều cao có số đặc trưng lớn hơn nhiều số mẫu. Đã có nhiều nhóm phương pháp khác nhau được sử dụng như lọc bằng cách đặt ngưỡng [18], loại bỏ đặc trưng ngay trong quá trình học [19–21], cũng như tính đến các ảnh hưởng phi tuyến [22]. Các phương pháp có thể kết hợp một cách khéo léo trong mô hình dự đoán nguy cơ đa di truyền (Polygenic Risk Scores - PRS) [23] để dự đoán nguy cơ mắc bệnh trong một nhóm thuần tập đích sử dụng thống kê tóm tắt GWAS của một nhóm thuần tập cơ sở, độc lập với nhóm thuần tập đích [24, 25]. Sự đa dạng trong các phương pháp góp phần giúp các nhà khoa học có một số nghiên cứu quy mô trong những năm gần đây với nhiều hơn các bệnh di truyền phức tạp, cùng với đó là các tập dữ liệu ngày càng lớn. Với hiệu suất mô hình dự đoán ngày càng được cải thiện, điểm nguy cơ đa di truyền đang góp phần vào nỗ lực phân tầng nguy cơ di truyền và có triển vọng áp dụng rộng rãi trong lâm sàng.

Trong các phần tiếp theo của bài báo chúng tôi sẽ trình bày về các kiến thức cơ sở trong Phần II, về tiền xử lý dữ liệu trong Phần III. Các cải tiến sử dụng học máy để nâng cao khả năng áp dụng trong lâm sàng của PRS được trình bày ở Phần IV. Phần V đề cập đến các xu hướng cải tiến hiệu năng cho mô hình dự đoán nguy cơ đa di truyền trong tương lai. Cuối cùng, chúng tôi kết luận bài báo ở Phần VI.

II. KIẾN THỨC CƠ SỞ

1. Nghiên cứu tương quan toàn hệ gen (GWAS)

Nghiên cứu tương quan toàn hệ gen (GWAS) là một phương pháp sàng lọc nhanh các chỉ thị phân tử trên toàn bộ DNA hoặc bộ gen của nhiều người, mục đích để tìm các biến dị di truyền liên quan đến một căn bệnh cụ thể hoặc một tính trạng mà ta quan tâm. Để thực hiện một nghiên cứu tương quan toàn hệ gen, các nhà nghiên cứu sử dụng hai nhóm người: nhóm thuần tập ca bệnh (case) và thuần tập đối chứng (control). DNA thu được bằng cách tách triết mẫu máu hoặc tế bào niêm mạc miệng của những người tham gia được đặt trên các mảng SNP và được quét trên các máy tự động. Các máy này nhanh chóng khảo sát bộ gen của mỗi người để tìm ra các SNP. Nếu SNP nào được tìm thấy ở những người mắc bệnh với tần suất cao hơn so với những người không mắc bệnh, thì SNP đó được cho là có mối tương quan đến bệnh. Các SNP này có thể đóng vai trò là những điểm đánh dấu di truyền (makers) chỉ dẫn đến khu vực chứa gen bệnh [1].

Do phải tuân theo quy tắc bảo mật của từng dự án nghiên cứu và các chính sách bảo mật thông tin cá nhân mà việc truy cập vào dữ liệu GWAS có hai mức độ:

- Mức độ thống kê tóm tắt: bao gồm các giá trị đại diện cho nhiều điểm dữ liệu, chẳng hạn như mức độ ảnh hưởng và P-value cho sự tương quan giữa mỗi SNP với một kiểu hình mà ta quan tâm. Chú ý rằng, mức độ ảnh hưởng được biểu diễn bằng tỷ lệ chênh lệch (Odds Ratio, viết tắt là OR) với tính trạng rời rạc, và được biểu diễn bằng chỉ số β với tính trạng liên tục. Ngoài ra, dữ liệu còn bao gồm các alen, tên SNP và vị trí của chúng trên nhiễm sắc thể.
- Mức độ cá thể: bao gồm định danh của cá thể, bố, mẹ, và phả hệ của cá thể. Hơn nữa, dữ liệu cũng cung cấp các thông tin về giới tính, kiểu hình, các alen, vị

[1] https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet

trí của các SNP trên nhiễm sắc thể, khoảng cách di truyền cũng như các hiệp biến. Dữ liệu GWAS ở mức độ cá thể thường được lưu dưới dạng các tệp định dạng PLINK [26, 27].

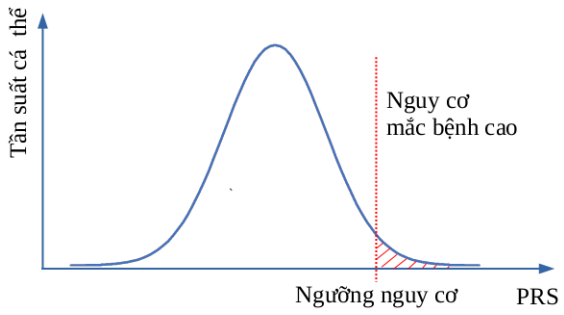
Dữ liệu thống kê tóm tắt GWAS có thể được tra cứu dễ dàng với nhiều tính trạng khác nhau trong Danh mục NHGRI-EBI về các nghiên cứu tương quan toàn hệ gen [28] [2]. Ngược lại, các dữ liệu GWAS ở mức độ cá thể cần được cấp phép để truy cập như tài nguyên trên dbGaP [29] hay UK Biobank [9].

2. Tổng quan về điểm nguy cơ đa di truyền (PRS)

Điểm nguy cơ đa di truyền (Polygenic Risk Score, PRS) được tính bằng tổng điểm có trọng số của các alen nguy cơ với trọng số dựa trên các mức độ ảnh hưởng từ GWAS [24]. Công thức mặc định để tính PRS trong PLINK [26] là:

$$PRS_j = \frac{\sum_i^N S_i.G_{i,j}}{P.M_j}$$

trong đó mức độ ảnh hưởng của SNP thứ i là S_i; số các alen ảnh hưởng của SNP thứ i được quan sát trong mẫu j là $G_{i,j}$; đơn bội của mẫu là P (thường là 2 cho người); tổng số SNP của mẫu j là N; tổng số SNP không thiếu được quan sát ở mẫu j là M_j. Nếu mẫu j có một kiểu gen thiếu SNP thứ i thì tần số alen phụ của quần thể được nhân với đơn bội ($MAF_i.P$) được sử dụng thay thế $G_{i,j}$.



Hình 1. **Biểu đồ phân phối của điểm nguy cơ đa di truyền**. Những cá thể có PRS nằm trong ngưỡng nguy cơ ở phía đuôi phải của phân phối có nguy cơ mắc bệnh đa di truyền cao.

Điểm nguy cơ đa di truyền (PRS) đang được coi là độ đo tương đối để mọi người có thể tìm hiểu về nguy cơ phát triển bệnh của họ. Ta có thể biểu diễn các giá trị PRS dưới dạng biểu đồ phân phối tần suất mà một cá thể có nguy cơ mắc bệnh cao sẽ nằm trong ngưỡng nguy cơ ở đuôi phải của phân phối (Xem hình 1). Tuy nhiên, độ chính xác của các dự đoán nguy cơ mắc bệnh đối với các nhóm người là khác nhau do nó phụ thuộc vào sự đầy đủ của dữ liệu GWAS. Tính đến năm 2018, phần lớn các nghiên cứu tương quan toàn hệ gen đã được thực hiện với tỷ lệ số lượng cá thể dựa trên sắc tộc bao gồm 78% người châu Âu, 10% người châu Á, 2% người châu Phi, 1% người gốc Tây Ban Nha và tất cả các sắc tộc khác chỉ chiếm nhỏ hơn 1% GWAS [30]. Ngoài ra, PRS cũng chỉ ra đóng góp của các yếu tố đa di truyền đối với một kiểu hình mà dữ liệu GWAS không phát hiện được. Vào năm 2009, một GWAS cho bệnh tâm thần phân liệt do Purcell và các đồng nghiệp thực hiện đã tìm thấy chỉ một SNP tương quan với kiểu hình, mặc dù bệnh này được biết đến như là một bệnh có tính di truyền cao. Tuy nhiên, bằng cách xây dựng một PRS sử dụng các kết quả GWAS và kiểm tra điểm nguy cơ đa di truyền trong tập dữ liệu độc lập khác, Purcell đã chứng minh rằng có một đóng góp đa di truyền tới bệnh tâm thần phân liệt [31]. Do đó, phân tích đa di truyền cho thấy dữ liệu GWAS ở giai đoạn đầu là chưa đủ mạnh, mà chúng ta cần có cỡ mẫu lớn hơn [32]. Mục đích khác của việc sử dụng PRS là để kiểm tra mối tương quan giữa các SNP với một kiểu hình khác với kiểu hình được sử dụng trong GWAS. Kỹ thuật này cho phép các nhà nghiên cứu chứng minh rằng có sự đóng góp đa di truyền chung giữa hai tính trạng. Cụ thể, Purcell đã chứng minh rằng có sự đóng góp di truyền chung giữa bệnh tâm thần phân liệt và rối loạn lưỡng cực (xem Hình 2).

Nếu như nghiên cứu của Purcell đã chỉ ra một số đặc điểm cơ bản là nền móng của PRS thì gần đây một số nghiên cứu lớn đã cải thiện một cách đáng kể độ tin cậy của kết quả dự đoán. Năm 2018, Amit V. Khera và các đồng nghiệp đã đã sử dụng dữ liệu di truyền toàn bộ hệ gen và các phương pháp suy diễn thống kê để đánh giá hàng triệu biến dị di truyền phổ biến liên quan đến 5 bệnh phổ biến: động mạch vành, rung nhĩ, tiểu đường loại 2, viêm ruột, và ung thư vú. Đối với mỗi bệnh, họ đã áp dụng một thuật toán tính điểm nguy cơ đa di truyền từ tất cả các biến dị để phản ánh tính nhạy cảm di truyền của một người đối với những bệnh này [33]. Năm 2019, một nghiên cứu với 272 tác giả và tổ chức sử dụng dữ liệu GWAS về ung thư vú được coi là lớn nhất với tập hợp từ 69 nghiên cứu gồm 94,075 mẫu bệnh và 75,017 mẫu đối chứng. Kết quả đã tìm ra được 313 SNP cho tính toán PRS tốt nhất với AUC đạt 63%, độ tin cậy 95%. Ngoài ra, nhóm tác giả đã phân tầng nguy cơ mắc bệnh dựa trên cả lịch sử gia đình và các yếu tố nguy cơ khác [34]. Một số nghiên cứu đột phá trong lịch sử phát triển của phương pháp tính điểm nguy cơ đa di truyền ở trên được tóm tắt trong Bảng I.

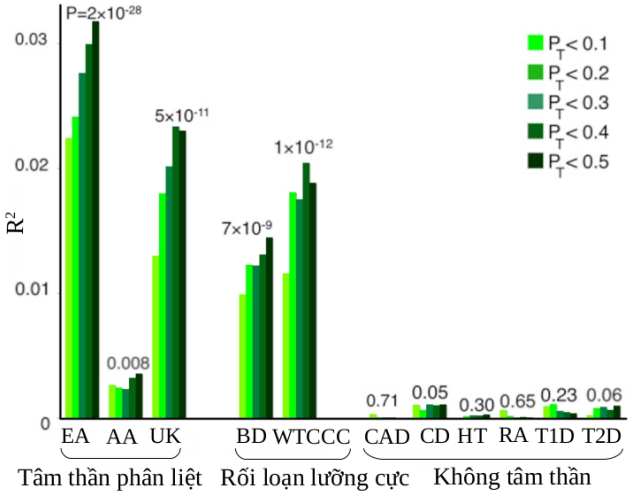
Để xây dựng mô hình tính điểm nguy cơ đa di truyền ta cần hai loại dữ liệu:

- Dữ liệu cơ sở (base data): bao gồm các thống kê tóm tắt của GWAS (ví dụ, β, OR, P-values) của tương quan kiểu gen-kiểu hình tại một biến dị di truyền (SNP).
- Dữ liệu đích (target data): kiểu gene, kiểu hình của

[2] https://www.ebi.ac.uk/gwas/

Bảng I
CÁC NGHIÊN CỨU ĐỘT PHÁ TRONG TÍNH TOÁN PRS.

STT	Tác giả	Năm	Nội dung
1	S. M. Purcell	2009	Xác định ảnh hưởng đa di truyền chung giữa bệnh tâm thần phân liệt và rối loạn lưỡng cực [31].
2	A. V. Khera	2018	Đánh giá hàng triệu biến dị phổ biến liên quan đến năm bệnh đa di truyền phổ biến [33].
3	N. Mavaddat	2019	Phân tầng nguy cơ ung thư vú dựa trên dữ liệu lớn của mẫu bệnh và mẫu đối chứng [34].



Hình 2. **Các thành phần đa di truyền với các mẫu độc lập của bệnh tâm thần phân liệt và rối loạn lưỡng cực**. Có sự đóng góp đa di truyền chung giữa bệnh tâm thần phân liệt và rối loạn lưỡng cực. Hình ảnh được tham khảo từ nghiên cứu của Hiệp hội Tâm thần Quốc tế [31].

các cá thể.

Một bệnh phát sinh ngoài nguyên nhân do di truyền thì còn có các nguyên nhân khác như môi trường, lối sống... Do đó bệnh không chỉ được dự đoán hoàn toàn nhờ các yếu tố di truyền, mà các yếu tố này chỉ chiếm một tỷ lệ nhất định h^2 (cũng được gọi là hệ số di truyền SNP: h^2_{SNP}, xem định nghĩa ở hộp 1). Vì việc xác định tập biến dị ảnh hưởng hoặc ước lượng mức độ ảnh hưởng có thể chưa chính xác, cũng như có sự khác biệt giữa các mẫu cơ sở và các mẫu đích mà tỷ lệ dự đoán của PRS nhỏ hơn h^2_{SNP}. Để đảm bảo mô hình PRS có được hiệu suất dự đoán cao ta cần kiểm tra kỹ lưỡng và tiền xử lý dữ liệu theo các tiêu chí sau:

- Các mẫu trong dữ liệu cơ sở và dữ liệu đích cần độc lập với nhau.
- Dữ liệu cơ sở và dữ liệu đích có các vị trí SNP được xây dựng từ cùng một hệ gen tham chiếu.

Hộp 1 | Một số khái niệm cơ bản

Alen nguy cơ (Risk Allele): alen của SNP làm tăng nguy cơ mắc bệnh. Một alen ảnh hưởng chỉ là alen có tương quan với bệnh và có thể tăng hoặc giảm nguy cơ.
Tần số alen (Allele Frequency): tỷ lệ các nhiễm sắc thể mang alen đó trong một quần thể.
Tần số alen phụ (MAF): tần số mà alen phổ biến thứ hai xuất hiện trong một quần thể nhất định.
Biến dị nhân quả (Causal Variant): biến dị tại một vị trí xác định trong nhiễm sắc thể có tương quan mạnh với kiểu hình, các biến dị này thực sự có ảnh hưởng sinh học đến kiểu hình [35].
Mức độ ảnh hưởng (Effect Size): mức độ tương quan giữa tần số alen của biến dị với một bệnh nào đó.
Hệ số di truyền (Heritability): mức độ giải thích của các yếu tố di truyền với kiểu hình trong một quần thể (ký hiệu: h^2 hoặc h^2_{SNP}) [36].
Mất cân bằng liên kết (Linkage Disequilibrium, LD): thước đo sự kết hợp không ngẫu nhiên giữa các alen ở các vị trí khác nhau trên cùng một nhiễm sắc thể trong một quần thể nhất định. Các SNP có trong LD khi tần số tương quan của các alen của chúng cao hơn mong đợi [27].
Phân tầng quần thể (Population Stratification): sự hiện diện của nhiều quần thể con (ví dụ: các cá nhân có nguồn gốc dân tộc khác nhau) trong một nghiên cứu. Sự phân tầng quần thể có thể dẫn đến các tương quan dương tính giả.
ROC (Receiver Operating Characteristic Curve): đường cong đặc tính biểu diễn tương quan giữa dương tính giả và dương tính thật với các ngưỡng nào đó [37, 38].
AUC (Area Under the Curve): diện tích dưới đường ROC, giá trị AUC nằm trong khoảng (0, 1), giá trị này càng lớn thì hiệu quả của mô hình càng cao [39].
Precision: tỉ lệ số điểm dương tính thật trên số điểm được phân loại là dương tính (dương tính thật + dương tính giả).
Recall: tỉ lệ số điểm dương tính thật trên số điểm thực sự là dương tính (dương tính thật + âm tính giả).
Overfitting: hiện tượng mô hình tìm được quá khớp với dữ liệu huấn luyện dẫn đến độ chính xác của dự đoán không còn tốt trên dữ liệu kiểm thử. Vì tần số alen có thể khác nhau giữa các quần thể con nên sự phân tầng quần thể có thể dẫn đến các tương quan dương tính giả.

- Thực hiện đầy đủ quá trình kiểm soát chất lượng cho dữ liệu cơ sở và dữ liệu đích.

Ngoài ra, phân tầng quần thể, yếu tố gây nhiễu trong GWAS, lại có thể được sử dụng để tính toán PRS tốt hơn.

3. Ứng dụng của PRS trong công nghiệp

PRS có thể được ứng dụng để dự đoán nguy cơ mắc bệnh trong công nghiệp từ mẫu nước bọt hoặc mẫu máu sử dụng công nghệ định kiểu gen (Genotyping) không tốn kém. Mặc dù vậy, PRS không bao giờ có thể dự đoán chắc chắn về tình trạng của các bệnh phổ biến phức tạp vì các yếu tố di truyền chỉ đóng góp một phần nguy cơ và PRS cũng chỉ nắm bắt được một phần các đóng góp di truyền đó.Tuy nhiên, cũng giống như y học lâm sàng sử dụng vô số các biện pháp dự đoán khác, PRS đóng một vai trò đáng kể như là một phần của các thuật toán dự đoán đa biến [40].

Bảng II
10 BỆNH HOẶC TÌNH TRẠNG Y TẾ ĐƯỢC FDA THÔNG QUA CHO CÁC XÉT NGHIỆM VỀ NGUY CƠ MẮC BỆNH DI TRUYỀN CỦA 23ANDME.

Bệnh	Mô tả
Parkinson	Rối loạn hệ thống thần kinh ảnh hưởng đến chuyển động
Alzheimer khởi phát muộn	Chứng rối loạn não tiến triển phá hủy trí nhớ và kỹ năng tư duy
Celiac	Rối loạn dẫn đến không thể tiêu hóa gluten
Thiếu Alpha-1 antitrypsin	Rối loạn làm tăng nguy cơ mắc bệnh phổi và gan
Loạn trương lực cơ nguyên phát khởi phát sớm	Rối loạn vận động liên quan đến các cơn co thắt cơ không tự chủ và các cử động mất kiểm soát khác
Thiếu yếu tố XI	Rối loạn đông máu
Gaucher loại 1	Rối loạn cơ quan và mô
Thiếu hụt Glucose-6-Phosphate Dehydrogenase	Còn được gọi là G6PD, một tình trạng hồng cầu
Huyết sắc tố di truyền	Rối loạn quá tải sắt
Máu khó đông di truyền	Rối loạn đông máu

Năm 2017, Cơ quan Quản lý Thực phẩm và Dược phẩm Hoa Kỳ (U.S. Food and Drug Administration - FDA) đã cho phép tiếp thị các xét nghiệm nguy cơ sức khỏe di truyền (Genetic Health Risk, GHR) của dịch vụ hệ gen cá nhân 23andMe đối với 10 bệnh hoặc tình trạng y tế (Xem Bảng II). Đây là các xét nghiệm đầu tiên được FDA cho phép cung cấp thông tin về khuynh hướng di truyền của một cá nhân đối với một số bệnh hoặc tình trạng y tế nhất định [3]. Các xét nghiệm GHR của 23andMe hoạt động bằng cách phân lập DNA từ mẫu nước bọt, sau đó được kiểm thử với hơn 500,000 biến dị di truyền. Sự hiện diện hoặc vắng mặt của một số biến dị này có liên quan đến việc tăng nguy cơ phát triển của bệnh.

Tại hội nghị South by Southwest (SXSW) 2019, 23andMe đã thông báo họ sẽ cung cấp phân tích nguy cơ dựa trên di truyền với bệnh tiểu đường loại 2 cho hàng triệu khách hàng. Thông tin di truyền từ các cá nhân với dữ liệu liên quan đến sức khỏe cũng hỗ trợ việc xây dựng các mô hình thống kê cho tính toán PRS và dự đoán các tính trạng và các tình trạng y tế khác nhau từ một DNA cá nhân. Đặc biệt trong báo cáo về bệnh tiểu đường loại 2 vào năm 2022, 1.1 triệu khách hàng của 23andMe đã đồng ý tham gia nghiên cứu giúp tạo ra cơ sở dữ liệu kiểu gen-kiểu hình lớn với 11,999 biến dị di truyền. Nhờ đó, các nhà nghiên cứu của 23andMe đã nâng cao được độ chính xác trong tính toán PRS với hơn 1000 biến dị liên quan đến bệnh [4].

III. TIỀN XỬ LÝ DỮ LIỆU

1. Kiểm soát chất lượng (Quality Control, QC)

Độ chính xác dự đoán của PRS phụ thuộc lớn vào chất lượng của dữ liệu cơ sở và dữ liệu đích. Do đó, đã có nhiều nghiên cứu cho quá trình kiểm soát chất lượng (Quality Control, QC). Cả hai tập dữ liệu thường được tiến hành QC với các tiêu chuẩn QC chung của GWAS [27, 41, 42]. Vào năm 2020, Shing Wan Choi và các đồng nghiệp đã chia ra ba cấp QC liên quan đến từng loại dữ liệu [43]. Dựa vào các bước QC từ những nghiên cứu trước đây, chúng tôi đã sắp xếp lại để đưa ra một quy trình kiểm soát chất lượng rõ ràng và đầy đủ (Xem Hình 3).

- **QC chỉ liên quan đến dữ liệu cơ sở**: gồm kiểm tra hệ số di truyền và xác định các alen ảnh hưởng.
- **QC chỉ liên quan đến dữ liệu đích**: cỡ mẫu được khuyến nghị lớn hơn hoặc bằng 100 [44] và thận trọng với những phân tích sử dụng dữ liệu cơ sở có h^2_{SNP} thấp và cỡ mẫu dữ liệu đích nhỏ. Mặt khác, cần kiểm soát chất lượng trên một số tiêu chí như: tỉ lệ kiểu gen (Genotyping Rate > 0.99), tỷ lệ mẫu thiếu (Sample Missingless < 0.02), cân bằng Hardy-Weinberg (P > 10^{-6}), loại bỏ những cá thể có hệ số F (đại diện cho tỷ lệ dị hợp tử) lớn hoặc nhỏ hơn 3 lần độ lệch chuẩn so với trung bình, loại bỏ các SNP tương quan với $r^2 > 0.25$.
- **QC tiêu chuẩn**: áp dụng cho cả dữ liệu cơ sở và dữ liệu đích. Do dữ liệu lớn, ta cần đảm bảo dữ liệu không bị lỗi trong quá trình tải hoặc chuyển tệp. Các dữ liệu cần được QC theo các tiêu chí chung cho dữ liệu GWAS như: tần số alen phụ (MAF > 1%, hoặc MAF > 5% nếu số lượng mẫu đích nhỏ hơn 1000), điểm thông tin (Info Score) bổ sung (Imputation) lớn

[3] https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-direct-consumer-tests-provide-genetic-risk-information-certain-conditions

[4] https://blog.23andme.com/23andme-research/screening-for-t2d/

hơn 0.8. Ngoài ra, ta cần loại bỏ các cá thể có sự khác biệt giới tính, xử lý các SNP không khớp, loại bỏ các SNP trùng lặp, loại bỏ các mẫu trùng lặp hoặc các mẫu có quan hệ họ hàng.

Dữ liệu cơ sở và dữ liệu đích sau khi được QC riêng sẽ tham gia quá trình QC tiêu chuẩn để tạo ra dữ liệu cuối cùng phục vụ cho tính toán điểm nguy cơ đa di truyền.

2. Kiểm soát mất cân bằng liên kết (Linkage Disequilibrium, LD)

Sau quá trình QC trên dữ liệu cơ sở và dữ liệu đích và định dạng các tập dữ liệu một cách phù hợp, ta cần lựa chọn một tập hợp các SNP ảnh hưởng trên toàn hệ gen để phục vụ tính toán PRS. Công việc này khá phức tạp do mối tương quan giữa các SNP được sinh ra từ sự mất cân bằng liên kết. Ta có thể lựa chọn tập hợp các SNP ảnh hưởng thông qua:

- Tính toán mất cân bằng liên kết (Linkage Disequilibrium, LD).
- Thu hẹp mức độ ảnh hưởng được ước lượng bởi GWAS (Shrinkage).

Phương pháp truyền thống thường được sử dụng là "Nhóm và Đặt ngưỡng" ("Clumping and Thresholding", hay còn gọi là "C+T"). Các kiểm định mối tương quan trong GWAS được thực hiện từng SNP một, trong khi cấu trúc tương quan trên toàn bộ hệ gen rất phức tạp, nên việc ước lượng các tác động di truyền độc lập là một vấn đề thách thức. Các SNP có thể được "Nhóm" ("Clumping", C) bởi công cụ PLINK để chọn ra các SNP có mối tương quan thấp với nhau. Trước tiên, Clumping chọn ra một SNP đặc trưng được gọi là SNP chỉ mục (SNP index) và tính toán mối tương quan giữa SNP này với các SNP gần đó (trong một khoảng cách di truyền ví dụ 250kb). Sau đó nó loại bỏ các SNP gần đó nếu mối tương quan giữa chúng lớn hơn một ngưỡng nhất định (ví dụ r^2 = 0.2, [32]). Như vậy, bước Clumping giúp loại bỏ dữ liệu dư thừa do mất cân bằng liên kết (LD) gây ra. Tập hợp SNP của GWAS tương quan với kiểu hình dưới một ngưỡng P-value nào đó sẽ được lựa chọn để tính toán PRS, phương pháp này còn được gọi là "Đặt ngưỡng" ("Thresholding", T). Vì không thể xác định trước ngưỡng P-value tối ưu, nên nhiều tập SNP với nhiều ngưỡng P-value khác nhau sẽ được lần lượt được đưa vào mô hình huấn luyện. Mô hình này còn có sự đóng góp của các hiệp biến và các thành phần chính của tập đích được tính toán dựa trên phân tầng quần thể. Ngưỡng P-value nào cho ra được độ chính xác dự đoán cao nhất thì tập hợp các SNP tương ứng với ngưỡng đó sẽ được lựa chọn cho tính toán PRS [24, 31, 45].

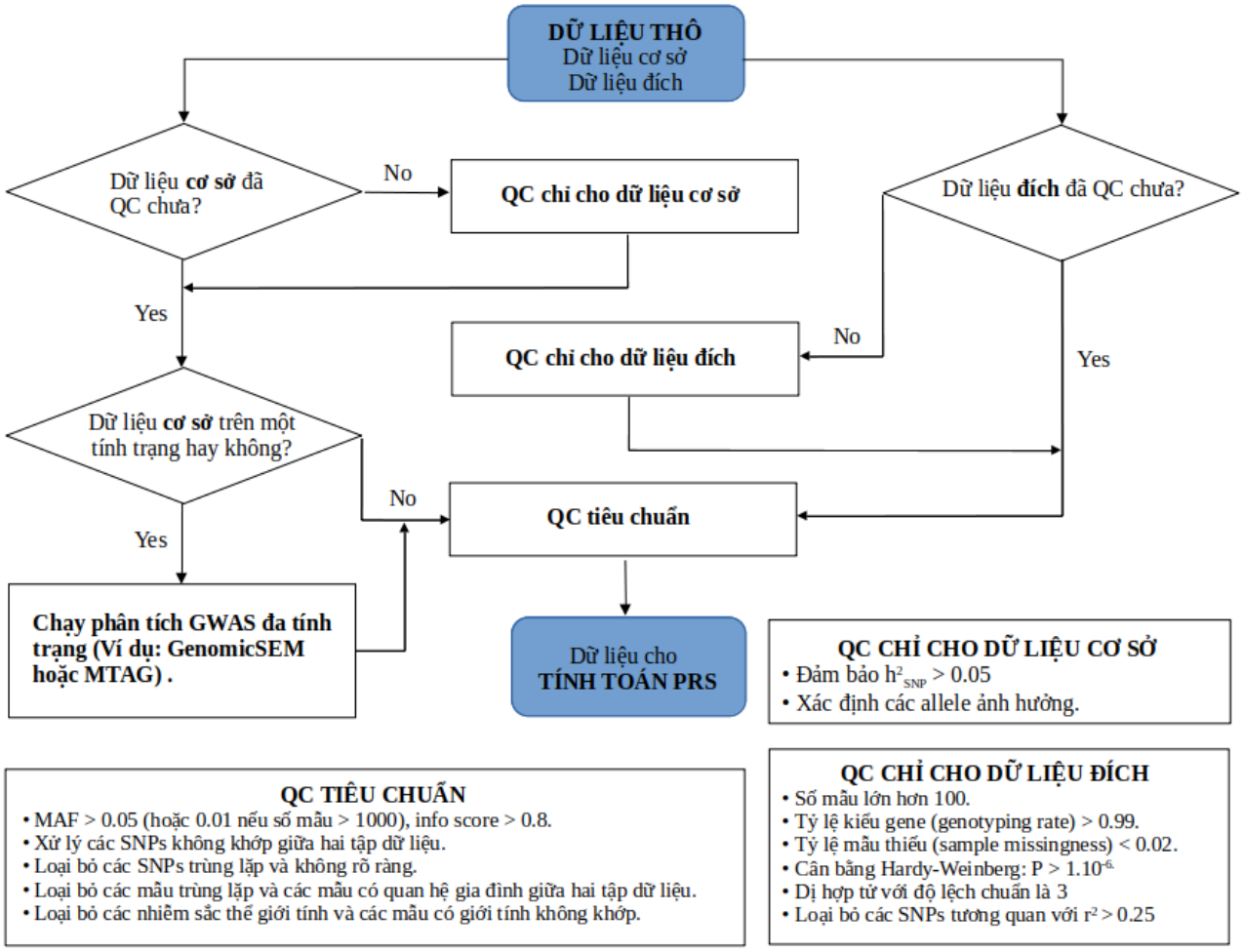
Ưu điểm của phương pháp "Nhóm và Đặt ngưỡng" là tốc độ tính toán nhanh do các tập dữ liệu tương ứng với các ngưỡng khác nhau có dạng ma trận (các hàng là các cá thể, các cột là các SNP) được chuyển về một vec tơ mà mỗi phần tử là điểm nguy cơ đa di truyền của từng cá thể. Tuy nhiên, nhược điểm lớn của phương pháp này là việc lựa chọn các ngưỡng hoàn toàn dựa trên kinh nghiệm, vì vậy khó tìm ra một cách đặt ngưỡng nhất quán cho các nghiên cứu khác nhau. Hơn nữa, phương pháp vẫn có thể bỏ sót các SNP ngoài phạm vi của ngưỡng nhưng thực sự ảnh hưởng đến bệnh.

Thay vì dựa vào kinh nghiệm để lựa chọn các tập SNP, một loạt các phương pháp thu gọn sử dụng học máy đã được phát triển để có thể loại bỏ các SNP không ảnh hưởng đến kiểu hình ngay trong quá trình học. Ngoài ra, các phương pháp mới nhất đi sâu vào việc tìm hiểu các đặc điểm sinh học kết hợp với học máy để giữ lại những SNP thực sự có ảnh hưởng đến bệnh cho dù chúng có thể không độc lập với nhau. Một số phương pháp cũng góp phần cải thiện kiểm soát mất cân bằng liên kết (Xem Phần IV).

IV. NHỮNG CẢI TIẾN SỬ DỤNG HỌC MÁY ĐỂ NÂNG CAO KHẢ NĂNG ÁP DỤNG TRONG LÂM SÀNG CỦA DỰ ĐOÁN NGUY CƠ ĐA DI TRUYỀN

Mặc dù PRS được tạo ra từ các nghiên cứu hệ gen quy mô lớn, nhưng do sự phức tạp của yếu tố đa di truyền kết hợp với các yếu tố môi trường mà việc áp dụng lâm sàng ở quy mô rộng của PRS đối mặt với những thách thức đáng kể. Để nâng cao khả năng sử dụng của PRS trong lâm sàng các nhà khoa học đã đưa ra một số cách tiếp cận khác nhau. Cách tiếp cận thứ nhất, dự đoán chỉ dựa trên các SNP có tương quan cao với bệnh đã biết trong dữ liệu GWAS. Cách tiếp cận thứ hai, các phương pháp chọn biến và học máy được phát triển để thu hẹp tập SNP [46] sử dụng các kỹ thuật điều chỉnh/thu hẹp trong thống kê như LASSO hoặc hồi quy Ridge (Ridge Regression) [47], hoặc sử dụng cách tiếp cận Bayes thông qua việc xác định phân phối [46, 48]. Cách tiếp cận thứ ba, mô hình dự đoán tính đến tỷ lệ của các biến dị nhân quả trong tổng số biến dị di truyền như phương pháp LDpred. Hiệu năng của các mô hình được đánh giá trong Nghiên cứu Dịch tễ học Di truyền về Sức khỏe Người trưởng thành và Lão hóa (the Genetic Epidemiology Research on Adult Health and Aging, GERA) [49] được giới thiệu ở Bảng III.

Hiện nay, các phương pháp chú trọng hơn đến việc xác định các biến dị ảnh hưởng thực sự đến kiểu hình và tìm cách đánh trọng số phù hợp cho các loại biến dị như: cây hồi quy tăng cường gradient và điều chỉnh mất cân bằng liên kết (GrabBLD) [50], dự đoán di truyền đa biến với ngưỡng trơn [51], xác định các điểm đánh dấu di truyền [52]. Mặt khác, điểm nguy cơ đa di truyền thường cung cấp một độ đo tương đối về nguy cơ được đánh giá ở cấp độ một nhóm người chứ không phải ở cấp độ cá nhân [53]



Hình 3. **Sơ đồ kiểm soát chất lượng cho tính toán PRS**. Trước tiên, dữ liệu đích và dữ liệu cơ sở được tiến hành QC riêng, nếu dữ liệu cơ sở được xây dựng trên nhiều tính trạng thì ta cần chạy phân tích GWAS đa tính trạng. Sau đó, cả hai tập dữ liệu được QC tiêu chuẩn để tạo ra một tập dữ liệu phục vụ cho tính toán PRS.

nên việc áp dụng PRS trong lâm sàng như một công cụ dự đoán khả năng mắc bệnh của cá nhân gặp những khó khăn nhất định. Để giải quyết vấn đề này, phương pháp ước lượng giới hạn tin cậy đã cho phép tính toán giá trị xác suất có điều kiện của tình trạng bệnh trên từng cá nhân [53]. Đặc biệt, phương pháp học sâu cũng cho thấy sức mạnh trong việc dự đoán nguy cơ mắc bệnh đa di truyền khi so sánh với một loạt các phương pháp học máy trước đó [54]. Trong khuôn khổ bài báo, chúng tôi sẽ giới thiệu chi tiết hơn một vài phương pháp điển hình gần đây sử dụng học máy được liệt kê trong Bảng IV.

1. Mô hình hồi quy logistic phạt (Penalized Logistic Regression, PLR)

Nhóm nghiên cứu của Florian Privé đã đưa ra mô hình hồi quy logistic phạt (Penalized Logistic Regression, PLR) để cải thiện hiệu quả cho tính toán PRS [55]. Trong mô hình này, hàm điều chỉnh L2 ("Ridge") có tác dụng thu nhỏ các hệ số và hàm điều chỉnh L1 ("LASSO") đưa các một phần các hệ số về giá trị 0 và có thể được sử dụng để chọn biến ngay trong quá trình học. Kết hợp giữa các hàm điều chỉnh L1 và L2 ("Elastic-Net") rất hiệu quả trong trường hợp số lượng SNP lớn hơn rất nhiều số lượng mẫu.

Cụ thể, bài toán được đưa về ước lượng các hệ số β_0, β để cực tiểu hóa hàm tổn thất được điều chỉnh

$$L(\lambda,\alpha) = -\sum_{i=1}^{n}(y_i log(z_i) + (1-y_i)log(1-z_i)) + \lambda((1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1)$$

trong đó $z_i = 1/(1+e^{-(\beta_0 + x_i^T\beta)})$, x biểu diễn kiểu gen và các hiệp biến (ví dụ: các thành phần chính, tuổi, giới tính), y là tình trạng bệnh, λ và α là hai siêu tham số điều chỉnh.

Bảng III
SO SÁNH HIỆU NĂNG CỦA CÁC PHƯƠNG PHÁP THUỘC BA CÁCH TIẾP CẬN KHÁC NHAU TRONG TÍNH TOÁN PRS VỚI DỮ LIỆU GERA. BẢNG SO SÁNH ĐƯỢC THAM KHẢO TỪ NGHIÊN CỨU CỦA M. THOMAS VÀ CÁC ĐỒNG NGHIỆP [49].

Các phương pháp tính PRS		**Số biến dị**	**AUC (95% CI)**
Cách tiếp cận 1: Các biến dị GWAS đã biết			
Các biến dị đã biết		140	0.629 (0.613–0.645)
Cách tiếp cận 2: Lựa chọn SNP và Học máy			
Ridge		10000	0.633 (0.617–0.648)
Lasso		10000	0.629 (0.601–0.646)
Elastic Net		10000	0.630 (0.612–0.641)
XGBoost		10000	0.629 (0.614–0.643)
Cách tiếp cận 3: LDpred			
LDpred	ρ = 1	1180765	0.620 (0.603–0.637)
	ρ = 0.3	1180765	0.625 (0.608–0.642)
	ρ = 0.1	1180765	0.628 (0.611–0.645)
	ρ = 0.03	1180765	0.635 (0.619–0.651)
	ρ = 0.01	1180765	0.646 (0.630–0.662)
	ρ = 0.005	1180765	0.649 (0.633–0.664)
	ρ = 0.003	1180765	0.654 (0.639–0.669)
	ρ = 0.001	1180765	0.643 (0.628–0.658)
Với LDpred, ρ là tỉ lệ các biến dị nhân quả trong tổng số biến dị của bệnh ung thư đại trực tràng (Accurate colorectal cancer, CRC).			

Dữ liệu được sử dụng là kiểu gen thực tế của các cá thể người châu Âu từ nghiên cứu dựa trên thuần tập ca bệnh và thuần tập đối chứng cho bệnh Celiac [56]. Để loại bỏ sự phân tầng di truyền gây ra bởi phân tầng quần thể, các tác giả chỉ sử dụng 15,155 cá thể với 4,496 mẫu ca bệnh và 10,659 mẫu đối chứng sau khi lọc; cả hai thuần tập này chứa 281,122 SNP.

Nhóm nghiên cứu đã xây dựng ba kịch bản mô phỏng để xét đến sự ảnh hưởng của các yếu tố đa di truyền cũng như cỡ mẫu đến kết quả dự đoán. PLR cho thấy hiệu quả dự đoán tốt hơn "C+T", giá trị AUC (xem hộp 1) cao hơn, trong hầu hết các kịch bản ngoại trừ một vài trường hợp yếu tố đa di truyền cao và hệ số di truyền thấp. Ngoài ra, nhóm nghiên cứu cũng đã phát triển hai gói công cụ trên R là bigstatsr và bigsnpr. Vấn đề bộ nhớ được giải quyết bằng cách sử dụng định dạng dữ liệu được lưu trữ dưới dạng tệp nhị phân. Các chức năng được cung cấp trong các gói công cụ này đều được song song hóa. Trong khi gói bigstatsr cung cấp các thuật toán tiêu chuẩn trong phân tích thống kê thì gói bigsnpr được xây dựng trên gói bigstatsr cung cấp các thuật toán dành riêng cho GWAS.

2. Cây hồi quy tăng cường gradient và điều chỉnh mất cân bằng liên kết (Gradient Boosted and LD, GraBLD)

Một phương pháp mới sử dụng các kỹ thuật học máy dựa trên cây hồi quy tăng cường gradient và điều chỉnh mất cân bằng liên kết (Gradient Boosted and LD, GraBLD) để tăng hiệu quả dự đoán nguy cơ đa di truyền. Đây là lần đầu tiên mà cây hồi quy tăng cường Gradient được sử dụng để tối ưu hóa trọng số của SNP kết hợp với việc điều chỉnh các vùng gây ra bởi sự mất cân bằng liên kết [50]. Một số kế quả khả quan của GraBLD áp dụng với các bộ dữ liệu tương ứng với các tính trạng khác nhau được liệt kê dưới đây:

- Dữ liệu thống kê tóm tắt từ Điều tra di truyền các tính trạng nhân trắc học (Genetic Investigation of Anthropometric Traits, GIANT) cho tính trạng chiều cao và tính trạng trọng lượng cơ thể trên bình phương chiều cao (Body Mass Index, BMI) trong nghiên cứu của UK Biobank [57] (130 nghìn cá thể; 1.98 triệu SNP): GrabBLD giải thích được tương ứng 46.9% và 32.7% của toàn bộ phương sai đa di truyền.
- Dữ liệu thống kê tóm tắt từ phân tích tổng hợp và sao chép di truyền bệnh tiểu đường (Diabetes Genetics Replication And Meta-analysis, DIAGRAM) trong nghiên cứu UK Biobank: diện tích dưới đường cong đặc tính (AUC) là 0.602.

GraBLD vượt trội hơn các phương pháp tính điểm truyền thống khi áp dụng để dự đoán chiều cao ($p < 2.2 * 10^{-16}$) và BMI ($p < 1.57 * 10^{-4}$), và tương đương với LDpred đối với bệnh tiểu đường. Về mặt phương pháp, cây hồi quy tăng cường gradient là phương pháp mạnh mẽ và linh hoạt kết hợp các phân loại yếu để tạo ra một phương pháp học mạnh cho dự đoán đầu ra là biến liên tục. GraBLD có khả năng cải thiện các trọng số SNP trong điểm nguy cơ đa di truyền mà không yêu cầu kiểu gen ở mức độ cá thể vì nó có thể mô hình hóa các mối quan hệ phi tuyến mà không cần lựa chọn đặc trưng.

3. Dự đoán di truyền đa biến với ngưỡng trơn (Smooth-Threshold Multivariate Genetic Prediction, STMGP)

Nghiên cứu của Yuta Takahashi và các đồng nghiệp vào năm 2020 đã đưa ra một thuật toán dự đoán mới, dự đoán di truyền đa biến với ngưỡng trơn (Smooth-Threshold Multivariate Genetic Prediction, STMGP), thuật toán đã cải thiện độ chính xác trong việc dự đoán các kiểu hình bệnh thần kinh đa di truyền bằng cách giảm overfitting (Xem định nghĩa trong hộp 1) thông qua việc lựa chọn các biến dị và xây dựng mô hình hồi quy phạt [51]. Mô hình dự đoán sử dụng tập huấn luyện gồm 3685 người ở quận Miyagi, tập kiểm thử độc lập gồm 3048 người ở quận Iwate tại Nhật Bản. Trong đó, kiểu hình là các triệu chứng trầm cảm và các kiểu hình được mô phỏng với độ phức tạp khác nhau và sự phân bố mức độ ảnh hưởng khác nhau của các alen nguy cơ (xem định nghiã trong hộp 1).

Nghiên cứu đã chỉ ra nguyên nhân khiến các nghiên cứu di truyền trước đây bị hạn chế trong việc dự đoán các kiểu hình của bệnh tâm thần bằng cách xem xét hai loại biến dị chủ yếu (Xem Hình 4):

Bảng IV
CÁC NGHIÊN CỨU NỔI BẬT GẦN ĐÂY SỬ DỤNG HỌC MÁY ĐỂ CẢI THIỆN ĐỘ CHÍNH XÁC CỦA DỰ ĐOÁN BỆNH ĐA DI TRUYỀN.

STT	Tên nghiên cứu	Tác giả	Năm	Phương pháp	Nội dung chính
1	Machine-learning heuristic to improve gene score prediction of polygenic traits	G. Paré	2017	GraBLD	Sử dụng cây hồi quy tăng cường gradient và điều chỉnh mất cân bằng liên kết để tăng hiệu quả dự đoán nguy cơ mắc bệnh đa di truyền.
2	Efficient Implementation of Penalized Regression for Genetic Risk Prediction	F. Privé	2019	PLR	Sử dụng mô hình hồi quy logistic phạt với dữ liệu kiểu gen-kiểu hình và các hiệp biến để chọn các đặc trưng ngay trong quá trình học.
3	Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes	Y. Takahashi	2020	STMGP	Phương pháp giúp giữ lại các biến dị nhạy thực sự có ảnh hưởng đến bệnh nhưng có tương quan với nhau, và loại bỏ các biến dị rỗng bằng việc đánh trọng số các tập biến dị và xây dựng hồi quy ridge.
4	Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction	J. Sun	2021	MCCP	Dựa vào độ đo NCM được áp dụng trên tập hiệu chỉnh để tìm xác suất mà một cá thể thuộc về một lớp, từ đó xác định được giới hạn tin cậy của một dự đoán.
5	Deep neural network improves the estimation of polygenic risk scores for breast cancer	A. Badré	2021	DNN	Mô hình được chứng minh là hoạt động tốt hơn các kỹ thuật học máy và các thuật toán thống kê truyền thống như BLUP, BayesA và Ldpred nhờ tạo ra được hai phân phối chuẩn của hai lớp có trung bình khác biệt và tính đến các ảnh hưởng phi tuyến.
6	Improving the Utility of Polygenic Risk Scores as a Biomarker for Alzheimer's Disease	D. Vlachakis	2021	Biomarker	Xây dựng một quy trình lọc và đánh trọng số dữ liệu SNP, tạo tiền đề cho việc áp dụng các kỹ thuật học máy trong việc nâng cao độ chính xác của dự đoán AD, cũng như các bệnh đa di truyền khác.

- Biến dị nhạy thực sự: là các biến dị ảnh hưởng đến bệnh. Trong đó, có các biến dị nhạy thực sự độc lập với nhau (màu cam) và các biến dị nhạy thực sự tương quan với nhau (màu vàng). Các biến dị này làm tăng độ chính xác của dự đoán.
- Biến dị rỗng (màu xanh): là các biến dị không ảnh hưởng đến bệnh. Các biến dị rỗng làm giảm độ chính xác của dự đoán.

Do khả năng giới hạn trong việc phân biệt các biến dị nhạy thực sự với các biến dị rỗng, nên mô hình có thể khớp một số lượng lớn nhiễu là các biến dị rỗng, kết quả dẫn đến overfitting (Xem định nghĩa trong hộp 1). Đây chính là vấn đề của các nghiên cứu về PRS trước đó khi nó ước lượng quá mức một số lượng lớn các biến dị rỗng kết hợp với các biến dị được nhóm trong quá trình Clumping (Xem Phần III.2), cũng như loại trừ các biến dị thực sự nhạy tương quan với nhau nhưng có liên quan đến bệnh.

STMGP xây dựng mô hình dự đoán dựa trên tập các biến dị được lựa chọn bởi các ngưỡng P-value của GWAS tương tự như PRS. Tuy nhiên, STMGP khắc phục được các vấn đề của PRS ở hai khía cạnh. Thứ nhất, STMGP có thể tránh overfitting bởi việc đánh trọng số các biến dị, trong đó các biến dị nhạy thực sự có trọng số cao hơn các biến dị rỗng. Thứ hai, STMGP sử dụng cả các biến dị nhạy thực sự tương quan với nhau mà có thể đóng góp vào độ chính xác dự đoán bởi việc xây dựng hồi quy Ridge tổng quát.

4. Ước lượng giới hạn tin cậy của dự đoán nguy cơ đa di truyền

Để dự đoán tình trạng bệnh cho từng cá nhân Jiangming Sun và đồng nghiệp đã giới thiệu một kỹ thuật học máy, MCCP (Mondrian Cross-Conformal Prediction), để tìm ra giá trị xác suất có điều kiện về tình trạng bệnh cho từng cá nhân thông qua ước lượng giới hạn tin cậy của dự đoán PRS tới nguy cơ mắc bệnh [53]. MCCP xây dựng độ đo NCM (Nonconformity Measures) nhờ các giá trị được dự đoán trên tập hiệu chỉnh bằng mô hình được huấn luyện trên tập huấn luyện (độc lập và cùng phân phối với tập hiệu chỉnh); tập huấn luyện và tập hiệu chỉnh đều chứa thông tin kiểu hình.

$$NCM_y = -y * d(x_i)$$

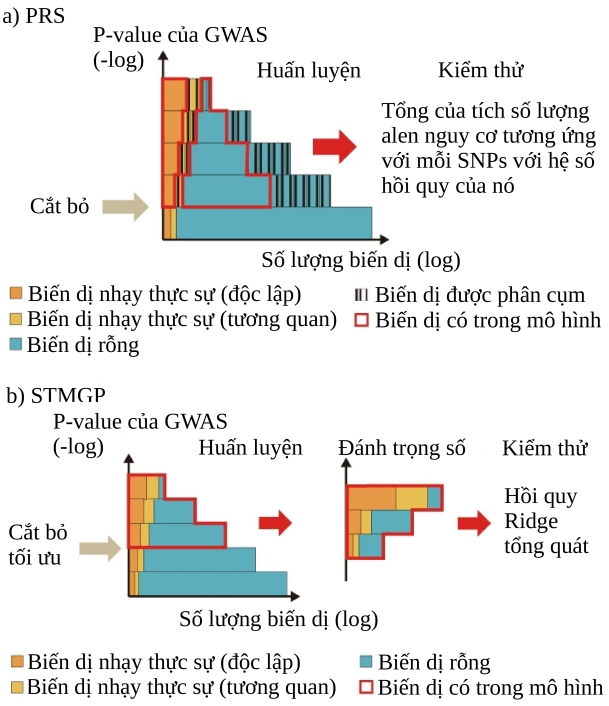
trong đó y là các lớp khác không, chẳng hạn (1, -1), và $d(x_i)$ là giá trị quyết định nhận được từ hàm quyết định của mô hình được học từ dữ liệu huấn luyện. Xác suất mà cá thể i thuộc về lớp y được tính như sau:

$$p_y^i = \frac{|\{j = 1, ..., N_{hc_y} : y_i = y, NCM_j \geq NCM_i\}|}{\{N_{hc_y} + 1 : y_i = y\}}$$

trong đó N_{hc_y} là cỡ mẫu của lớp y trong tập hiệu chỉnh. Với một sai số kỳ vọng $\alpha \in [0, 1]$ kết quả của vùng dự đoán được xác định dưới đây:

$$\Gamma^\alpha = \{y \in Y : p_y > \alpha\}$$

trong đó Y là tập hợp các lớp, p_y là giá trị xác suất khi một cá thể thuộc về lớp y, vùng dự đoán Γ^α là một tập



Hình 4. **So sánh giữa PRS và STMGP**. a) PRS bị overfitting do mô hình khớp với một số lượng lớn nhiễu là các biến dị rỗng (màu xanh) và các biến dị được nhóm bởi Clumping. Trong khi nó chỉ giữ lại các biến dị nhạy thực sự độc lập (màu cam) và loại bỏ các biến dị nhạy thực sự có tương quan với nhau (màu vàng) nhưng lại đóng góp vào độ chính xác của mô hình. b) STMGP không những tránh được overfitting do việc đánh trọng số các biến dị, mà còn giữ lại được các biến dị nhạy thực sự có tương quan với nhau nhưng ảnh hưởng đến bệnh nhờ mô hình hồi quy ridge tổng quát. Hình ảnh được tham khảo từ nghiên cứu của Yuta Takahashi và các đồng nghiệp [51].

hợp có thể rỗng hoặc chứa một hoặc chứa hai lớp. Độ tin cậy của MCCP được định nghĩa như sau:

$$sup\{1-\alpha : \Gamma^{\alpha} \leq 1\}$$

là giá trị lớn nhất của $1-\alpha$ với Γ^{α} là một lớp duy nhất, chẳng hạn lớp bệnh hoặc lớp đối chứng. Như vậy, với một cá thể có trạng thái được dự đoán, MCCP có thể ước lượng giới hạn tin cậy của dự đoán đó.

Nhóm nghiên cứu đã áp dụng MCCP đối với bệnh động mạch vành (CAD), đái tháo đường týp 2 (T2D), bệnh viêm ruột (IBD) và ung thư vú (*BRCA*) sử dụng tài nguyên của UK Biobank [9, 57] và hai bộ dữ liệu bổ sung dựa trên quần thể, mẫu bệnh tâm thần phân liệt (SCZ) trong nghiên cứu tâm thần tích hợp (iPSYCH) [58] và mẫu T2D trong nghiên cứu chế độ ăn kiêng Malmo và ung thư (MDC) [59]. Kết quả cho thấy ở cấp độ cá nhân, MCCP báo cáo xác suất dự đoán được hiệu chỉnh tốt, ước tính có hệ thống giới hạn tin cậy của của PRS trong dự đoán nguy cơ mắc bệnh phức tạp ở người. Ở cấp độ nhóm, MCCP vượt trội hơn các phương pháp tiêu chuẩn trong việc phân tầng chính xác các cá nhân thành các nhóm nguy cơ.

5. Mạng nơ-ron sâu (Deep Neural Network, DNN)

Vào năm 2021, Adrien Badré và đồng nghiệp đã so sánh một loạt các mô hình tính toán để ước tính PRS cho bệnh ung thư vú. Một mạng nơ-ron sâu (Deep Neural Network, DNN) được chứng minh là hoạt động tốt hơn các kỹ thuật học máy và các thuật toán thống kê truyền thống như BLUP, BayesA và LDpred [54]. Về hiệu quả của mô hình, diện tích dưới đường cong đặc tính (AUC) là 67,4% đối với DNN, 64,2% đối với BLUP, 64,5% đối với BayesA, và 62,4% đối với LDpred. Ưu điểm lớn nhất của DNN là có thể tách quần thể ca bệnh thành hai quần thể: một quần thể con có nguy cơ di truyền cao với PRS trung bình cao hơn đáng kể so với quần thể đối chứng, một quần thể con có nguy cơ bình thường với PRS trung bình tương tự như quần thể đối chứng. Nguyên nhân là vì PRS do DNN tạo ra tuân theo hai phân phối chuẩn với các trung bình khác nhau rõ rệt. Ngược lại, BLUP, BayesA và LPpred đều tạo ra PRS tuân theo chỉ một phân phối chuẩn trong quần thể ca bệnh. Điều này cho phép DNN đạt được recall 18,8% ở precision 90% (Xem định nghĩa recall và precision ở hộp 1) trong thuần tập kiểm thử. Mặt khác, mô hình DNN còn xác định được các biến dị đặc trưng mà chỉ được gán giá trị P-value không đáng kể bởi GWAS, các biến dị này có thể tương quan với kiểu hình thông qua các mối quan hệ phi tuyến. Đây cũng là ưu điểm lớn của các mạng học sâu so với các phương pháp học máy truyền thống.

6. Xác định các điểm đánh dấu di truyền

Để cải thiện độ đo PRS, các nhà khoa học đã đề xuất một quy trình mới áp dụng cho bệnh Alzheimer (AD) nhằm nâng cao độ chính xác dự đoán từ các yếu tố đa di truyền [52]. Nhóm tác giả lạc quan rằng các phân tích dự kiến trong tương lai của họ sẽ chứng minh cách tiếp cận này có thể nâng cao đáng kể hiệu suất của PRS và khả năng ứng dụng của nó trong lâm sàng. Các bước của quy trình được tóm tắt như sau:

- Bước 1: Xem xét mạng quan hệ giữa các SNP trong các khu vực gần gen AD. Có thể có các SNP không ảnh hưởng đến kiểu hình đi kèm với các SNP đặc chưng vì chúng được liên kết hoặc di truyền cùng nhau.
- Bước 2: Đánh trọng số các SNP cao hơn trong các gen mã hóa vì những thay đổi mà chúng mang lại cho chuỗi polypeptit có thể có tác động sửa đổi sau dịch mã.
- Bước 3: Đánh trọng số cao hơn cho các SNP tương quan với các dạng nghiêm trọng hơn của kiểu hình bắt nguồn từ các nghiên cứu GWAS khác nhau.

Các nhà khoa học đã phát triển một tập dữ liệu tích hợp với tất cả các gen liên quan đến bệnh Alzheimer. Các SNP từ các nghiên cứu được ánh xạ vào 1241 gen liên quan đến AD mà được lựa chọn theo qui trình. Kết quả cuối cùng cho ra ba nhóm SNP. Một nhóm bao gồm các SNP được xác định trong cả tập dữ liệu bệnh Alzheimer và bệnh Alzheimer khởi phát muộn mà không được thưởng hay phạt. Một nhóm bao gồm các SNP chỉ có trong các nghiên cứu GWAS về AD và liên tiếp được thưởng (vì chúng thuộc nhóm biểu hiện lâm sàng nặng). Một nhóm bao gồm các các SNP chỉ có trong dữ liệu AD khởi phát muộn và bị phạt vì chúng được liên kết với một biểu hiện nhẹ hơn của AD. Phương pháp này có thể mở đường cho các ứng dụng mới của học máy hệ gen vào dữ liệu GWAS trong nỗ lực xác định các điểm đánh dấu di truyền cho các bệnh phức tạp.

V. CÁC XU HƯỚNG CẢI TIẾN HIỆU NĂNG CHO MÔ HÌNH DỰ ĐOÁN NGUY CƠ ĐA DI TRUYỀN TRONG TƯƠNG LAI

1. Mở rộng nghiên cứu tương quan toàn hệ gen

Với sự phát triển mạnh mẽ của các nghiên cứu tương quan toàn hệ gen, GWAS sẽ không chỉ tập trung vào nhóm người châu Âu mà đang mở rộng ra các nhóm người khác trên toàn thế giới [60–66]. Điều này làm cho dữ liệu GWAS trở nên đáng tin cậy hơn khi sử dụng để dự đoán nguy cơ đa di truyền. Hơn nữa, các cơ sở dữ liệu như HapMap [67], dự án 1000 hệ gen [68] vẫn đang tiếp tục được hoàn thiện giúp cho quá trình suy diễn thống kê bổ sung các SNP bị thiếu trong các bộ dữ liệu từ các nghiên cứu riêng lẻ trở nên chính xác hơn [69, 70]. Các phân tích tổng hợp (meta analysis) kết hợp các nghiên cứu riêng lẻ với nhau sẽ tạo nên những bộ dữ liệu lớn bao gồm nhiều tính trạng khác nhau [71–74]. Đặc biệt, xu hướng kết hợp dữ liệu GWAS với dữ liệu giải trình tự thế hệ mới đang trở nên phổ biến khi mà giá thành giải trình tự ngày càng rẻ [75, 76]. Điều này góp phần tạo ra một số lượng SNP lớn cần thiết mà trước đây chưa bao giờ có.

Như vậy, nhờ sự phát triển mạnh mẽ các nghiên cứu mà các bộ dữ liệu GWAS ngày càng đầy đủ và đáng tin cậy. Đây là điều kiện cần để thúc đẩy các nghiên cứu dự đoán nguy cơ đa di truyền cho nhiều tính trạng mới với độ chính xác cao hơn trước, từ đó giúp mở rộng phạm vi áp dụng của PRS trong lâm sàng.

2. Phát triển mô hình dự đoán nguy cơ mắc bệnh

Để hướng tới y học chính xác, các mô hình dự đoán nguy cơ đa di truyền đang dần được phát triển để có thể xác định nguy cơ mắc bệnh của người trong một khung thời gian cụ thể. Đại diện cho hướng phát triển này là mô hình tính toán nguy cơ tuyệt đối được giới thiệu bởi Nilanjan Chatterjee và các cộng sự vào năm 2016 [77]. Từ đó đến nay đã có một số nghiên cứu về nguy cơ tuyệt đối cho các bệnh khác nhau như: động mạch vành [78], ung thư vú [79], ung thư phổi [80], ung thư đại trực tràng [81], ung thư tuyến tiền liệt [82]. Để cải thiện hiệu năng dự đoán, ta có thể bổ sung các yếu tố được tổng hợp từ các nghiên cứu dịch tễ chất lượng cao với cỡ mẫu lớn bao gồm: biến dị di truyền, yếu tố nguy cơ từ môi trường, các dấu hiệu sinh học về phơi nhiễm, trong đó có tính đến sự tương tác giữa các yếu tố nguy cơ. Đặc biệt để dự báo nguy cơ phát triển của bệnh trong một khoảng thời gian cụ thể ta cần thêm phân phối của các yếu tố nguy cơ như độ tuổi khởi phát bệnh, tỷ lệ tử vong trong quần thể nghiên cứu. Khi đó mô hình có khả năng tính toán xác suất của một cá thể không có triệu chứng sẽ phát triển bệnh trong một khoảng thời gian nào đó.

Việc phát triển mô hình dự đoán nguy cơ đa di truyền nhằm xác định nguy cơ tuyệt đối có tiềm năng lớn trong tương lai, nó có thể được áp dụng rộng rãi trong lâm sàng khi khả năng tiếp cận các dữ liệu lâm sàng và dữ liệu dịch tễ trở nên dễ dàng hơn.

VI. KẾT LUẬN

Trong bài báo này chúng tôi đã giới thiệu một cách tổng quan về dự đoán nguy cơ đa di truyền từ quá trình phát triển của GWAS cho đến các nghiên cứu cải tiến phương pháp tính toán PRS liên quan đến học máy.

Chúng tôi đã phân tích và sắp xếp lại các bước kiểm soát chất lượng để đưa ra một quy trình tiền xử lý dữ liệu rõ ràng và đầy đủ. Đây là quy trình rất quan trọng vì nó ảnh hưởng trực tiếp đến hiệu suất của mô hình khi mà dữ liệu phục vụ cho quá trình huấn luyện mô hình thường đến từ nhiều nghiên cứu khác nhau. Mặt khác, chúng tôi cũng thấy được những thách thức mà các nhà khoa học gặp phải khi tìm cách áp dụng PRS vào lâm sàng. Vì vậy, những nhược điểm của PRS và các phương pháp ngày càng tốt hơn giúp giải quyết những nhược điểm đó cũng được liệt kê một cách khá đầy đủ. Theo đó, một loạt các nghiên cứu cải tiến điển hình liên quan đến học máy trong thời gian gần đây được lựa chọn và trình bày tóm tắt một cách có hệ thống.

Với đa dạng các nghiên cứu theo nhiều hướng tiếp cận khác nhau của học máy đã góp phần nâng cao độ chính xác của các dự đoán dựa vào PRS cho các bệnh phổ biến. Các tác giả không những tập trung vào loại bỏ các SNP là nhiễu do mất cân bằng liên kết mà còn xác định được chính xác hơn ảnh hưởng của các SNP với kiểu hình. Các phương pháp xác định có thể tìm được các ảnh hưởng tuyến tính như tương quan giữa SNP với kiểu hình, cũng có thể tìm được các ảnh hưởng phi tuyến. Độ chính xác của PRS

được cải thiện góp phần nâng cao độ chính xác của tính toán nguy cơ tuyệt đối giúp phân tầng nguy cơ di truyền trong lâm sàng.

Chúng tôi hy vọng cái nhìn tổng quan về PRS, nhất là trong thời gian gần đây, sẽ giúp các nhà nghiên cứu dễ dàng hơn trong việc tiếp cận, cũng như tìm ra những hướng nghiên cứu cải tiến mới có giá trị nhằm mở rộng phạm vi ứng dụng của PRS trong lâm sàng.

LỜI CẢM ƠN

Toàn bộ nghiên cứu được tài trợ bởi quỹ Nghiên cứu và Ứng dụng LB.Sci của Công ty TNHH LOBI Việt Nam.

TÀI LIỆU THAM KHẢO

[1] Kuchenbaecker, K.B. et al. (2017) "Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers", *JAMA*, 317(23), pp. 2402–2416. doi:10.1001/jama.2017.7112.

[2] Wexler, N.S. et al. (1987) "Homozygotes for Huntington's disease", *Nature*, 326(6109), pp. 194–197. doi:10.1038/326194a0.

[3] Gusella, J.F. (1989) "Location cloning strategy for characterizing genetic defects in Huntington's disease and Alzheimer's disease", *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 3(9), pp. 2036–2041. doi:10.1096/fasebj.3.9.2568302.

[4] Ford, D. et al. (1998) "Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families", *The American Journal of Human Genetics*, 62(3), pp. 676–689. doi:10.1086/301749.

[5] Klein, R.J. et al. (2005) "Complement factor H polymorphism in age-related macular degeneration", Science (New York, N.Y.), 308(5720), pp. 385–389. doi:10.1126/science.1109557.

[6] Burton, P.R. et al. (2007) "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls", *Nature*, 447(7145), pp. 661–678. doi:10.1038/nature05911.

[7] Loos, R.J.F. (2020) "15 years of genome-wide association studies and no signs of slowing down", *Nature Communications*, 11(1), p. 5900. doi:10.1038/s41467-020-19653-5.

[8] Visscher, P.M. et al. (2017) "10 Years of GWAS Discovery: Biology, Function, and Translation", *American Journal of Human Genetics*, 101(1), pp. 5–22. doi:10.1016/j.ajhg.2017.06.005.

[9] Bycroft, C. et al. (2018) "The UK Biobank resource with deep phenotyping and genomic data", *Nature*, 562(7726), pp. 203–209. doi:10.1038/s41586-018-0579-z.

[10] Pepe, M.S. et al. (2004) "Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker", *American Journal of Epidemiology*, 159(9), pp. 882–890. doi:10.1093/aje/kwh101.

[11] Jakobsdottir, J. et al. (2009) "Interpretation of Genetic Association Studies: Markers with Replicated Highly Significant Odds Ratios May Be Poor Classifiers", *PLOS Genetics*, 5(2), p. e1000337. doi:10.1371/journal.pgen.1000337.

[12] Wray, N.R., Goddard, M.E. and Visscher, P.M. (2007) "Prediction of individual genetic risk to disease from genome-wide association studies", *Genome Research*, 17(10), pp. 1520–1528. doi:10.1101/gr.6665407.

[13] Manolio, T.A. et al. (2009) "Finding the missing heritability of complex diseases", *Nature*, 461(7265), pp. 747–753. doi:10.1038/nature08494

[14] Cecile, A., Janssens, J.W. and Joyner, M.J. (2019) "Polygenic Risk Scores That Predict Common Diseases Using Millions of Single Nucleotide Polymorphisms: Is More, Better?", *Clinical Chemistry*, 65(5), pp. 609–611. doi:10.1373/clinchem.2018.296103.

[15] Sha, Z., Hu, T. and Chen, Y. (2021) "Feature Selection for Polygenic Risk Scores using Genetic Algorithm and Network Science", in *2021 IEEE Congress on Evolutionary Computation (CEC)*, pp. 802–808. doi:10.1109/CEC45853.2021.9504993.

[16] Klinger, J.E. et al. (2021) *Interaction-based feature selection algorithm outperforms polygenic risk score in predicting Parkinson's Disease status*. medRxiv, p. 2021.07.20.21260848. doi:10.1101/2021.07.20.21260848.

[17] Kulm, S., Mezey, J. and Elemento, O. (2022) Benchmarking Polygenic Risk Score Model Assumptions: towards more accurate risk assessment. *bioRxiv*, p. 2022.02.18.480983. doi:10.1101/2022.02.18.480983.

[18] Privé, F. et al. (2019) "Making the Most of Clumping and Thresholding for Polygenic Scores", *The American Journal of Human Genetics*, 105(6), pp. 1213–1221. doi:10.1016/j.ajhg.2019.11.001.

[19] Hahn, G. et al. (2021) "A fast and efficient smoothing approach to Lasso regression and an application in statistical genetics: polygenic risk scores for chronic obstructive pulmonary disease (COPD)", *Statistics and Computing*, 31(3), p. 35. doi:10.1007/s11222-021-10010-0.

[20] Pattee, J. and Pan, W. (2020) "Penalized regression and model selection methods for polygenic scores on summary statistics", *PLOS Computational Biology*, 16(10), p. e1008271. doi:10.1371/journal.pcbi.1008271.

[21] Dickson, S.P. et al. (2021) "GenoRisk: A polygenic risk score for Alzheimer's disease", *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 7(1), p. e12211.

[22] Peng, J. et al. (2021) A Deep Learning-based Genome-wide Polygenic Risk Score for Common Diseases Identifies Individuals with Risk. *medRxiv*, p. 2021.11.17.21265352. doi:10.1101/2021.11.17.21265352.

[23] Zhao, B. and Zou, F. (2021) "On polygenic risk scores for complex traits prediction", *Biometrics* [Preprint]. doi:10.1111/biom.13466.

[24] Euesden, J., Lewis, C.M. and O'Reilly, P.F. (2015a) "PRSice: Polygenic Risk Score software", *Bioinformatics (Oxford, England)*, 31(9), pp. 1466–1468. doi:10.1093/bioinformatics/btu848.

[25] Uffelmann, E. et al. (2021) "Genome-wide association studies", *Nature Reviews Methods Primers*, 1(1), pp. 1–21. doi:10.1038/s43586-021-00056-9.

[26] Purcell, S. et al. (2007) "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses", *American Journal of Human Genetics*, 81(3), pp. 559–575.

[27] Marees, A.T. et al. (2018) "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis", *International Journal of Methods in Psychiatric Research*, 27(2), p. e1608. doi:10.1002/mpr.1608.

[28] Buniello, A. et al. (2019) "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019", *Nucleic Acids Research*, 47(D1), pp. D1005–D1012. doi:10.1093/nar/gky1120.

[29] Tryka, K.A. et al. (2014) "NCBI's Database of Genotypes and Phenotypes: dbGaP", *Nucleic Acids Research*, 42(D1), pp. D975–D979. doi:10.1093/nar/gkt1211.

[30] Sirugo, G., Williams, S.M. and Tishkoff, S.A. (2019) "The Missing Diversity in Human Genetic Studies", *Cell*, 177(1), pp. 26–31. doi:10.1016/j.cell.2019.02.048.

[31] Purcell, S.M. et al. (2009) "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder", *Nature*, 460(7256), pp. 748–752. doi:10.1038/nature08185.

[32] Wray, N.R. et al. (2014) "Research review: Polygenic methods and their application to psychiatric traits", *Journal of Child Psychology and Psychiatry*, and Allied Disciplines, 55(10), pp. 1068–1087. doi:10.1111/jcpp.12295.

[33] Khera, A.V. et al. (2018) "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations", *Nature Genetics*, 50(9), pp. 1219–1224. doi:10.1038/s41588-018-0183-z.

[34] Mavaddat, N. et al. (2019) "Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes", *American Journal of Human Genetics*, 104(1), pp. 21–34. doi:10.1016/j.ajhg.2018.11.002.

[35] Hormozdiari, F. et al. (2015) "Identification of causal genes for complex traits", *Bioinformatics*, 31(12), pp. i206–i213. doi:10.1093/bioinformatics/btv240.

[36] Visscher, P.M., Hill, W.G. and Wray, N.R. (2008) "Heritability in the genomics era — concepts and misconceptions", *Nature Reviews Genetics*, 9(4), pp. 255–266. doi:10.1038/nrg2322.

[37] Lusted, L.B. (1971) "Signal detectability and medical decision-making", *Science (New York, N.Y.), 171(3977)*, pp. 1217–1219. doi:10.1126/science.171.3977.1217.

[38] Fawcett, T. (2006) "An introduction to ROC analysis", *Pattern Recognition Letters*, 27(8), pp. 861–874. doi:10.1016/j.patrec.2005.10.010.

[39] Hanley, J.A. and McNeil, B.J. (1982) "The meaning and use of the area under a receiver operating characteristic (ROC) curve", *Radiology*, 143(1), pp. 29–36. doi:10.1148/radiology.143.1.7063747.

[40] Wray, N.R. et al. (2021) "From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer", *JAMA Psychiatry*, 78(1), pp. 101–109. doi:10.1001/jamapsychiatry.2020.3049.

[41] Anderson, C.A. et al. (2010) "Data quality control in genetic case-control association studies", *Nature Protocols*, 5(9), pp. 1564–1573. doi:10.1038/nprot.2010.116.

[42] Coleman, J.R.I. et al. (2016) "Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray", *Briefings in Functional Genomics*, 15(4), pp. 298–304. doi:10.1093/bfgp/elv037.

[43] Choi, S.W., Mak, T.S.-H. and O'Reilly, P.F. (2020) "Tutorial: a guide to performing polygenic risk score analyses", *Nature Protocols*, 15(9), pp. 2759–2772. doi:10.1038/s41596-020-0353-1.

[44] Han, B. and Eskin, E. (2011) "Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies", *American Journal of Human Genetics*, 88(5), pp. 586–598. doi:10.1016/j.ajhg.2011.04.014.

[45] Allan, B.L. (1987) "Calculating medication error rates", *American Journal of Hospital Pharmacy*, 44(5), pp. 1044, 1046.

[46] Ge, T. et al. (2019) "Polygenic prediction via Bayesian regression and continuous shrinkage priors", *Nature Communications*, 10(1), p. 1776. doi:10.1038/s41467-019-09718-5.

[47] Mak, T.S.H. et al. (2017) "Polygenic scores via penalized regression on summary statistics", *Genetic Epidemiology*, 41(6), pp. 469–480. doi:10.1002/gepi.22050.

[48] Newcombe, P.J. et al. (2019) "A flexible and parallelizable approach to genome-wide polygenic risk scores", *Genetic Epidemiology*, 43(7), pp. 730–741. doi:10.1002/gepi.22245.

[49] Thomas, M. et al. (2020) "Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk", *The American Journal of Human Genetics*, 107(3), pp. 432–444. doi:10.1016/j.ajhg.2020.07.006.

[50] Paré, G., Mao, S. and Deng, W.Q. (2017) "A machine-learning heuristic to improve gene score prediction of polygenic traits", *Scientific Reports*, 7(1), p. 12665. doi:10.1038/s41598-017-13056-1.

[51] Takahashi, Y. et al. (2020) "Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes", *Translational Psychiatry*, 10(1), pp. 1–11. doi:10.1038/s41398-020-00957-5.

[52] Vlachakis, D. et al. (2021) "Improving the Utility of Polygenic Risk Scores as a Biomarker for Alzheimer's Disease", *Cells*, 10(7), p. 1627. doi:10.3390/cells10071627.

[53] Sun, J. et al. (2021) "Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction", *Nature Communications*, 12(1), p. 5276. doi:10.1038/s41467-021-25014-7.

[54] Badré, A. et al. (2021) "Deep neural network improves the estimation of polygenic risk scores for breast cancer", *Journal of Human Genetics*, 66(4), pp. 359–369. doi:10.1038/s10038-020-00832-7.

[55] Privé, F., Aschard, H. and Blum, M.G.B. (2019) "Efficient Implementation of Penalized Regression for Genetic Risk Prediction", *Genetics*, 212(1), pp. 65–74. doi:10.1534/genetics.119.302019.

[56] Dubois, P.C.A. et al. (2010) "Multiple common variants for celiac disease influencing immune gene expression", *Nature Genetics*, 42(4), pp. 295–302. doi:10.1038/ng.543.

[57] Sudlow, C. et al. (2015) "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age", *PLOS Medicine*, 12(3), p. e1001779. doi:10.1371/journal.pmed.1001779.

[58] Pedersen, C.B. et al. (2018) "The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders", *Molecular Psychiatry*, 23(1), pp. 6–14. doi:10.1038/mp.2017.196.

[59] Berglund, G. et al. (1993) "The Malmo Diet and Cancer Study. Design and feasibility", *Journal of Internal Medicine*, 233(1), pp. 45–51. doi:10.1111/j.1365-2796.1993.tb00647.x.

[60] Stevenson, A. et al. (2019) "Neuropsychiatric Genetics of African Populations-Psychosis (NeuroGAP-Psychosis): a case-control study protocol and GWAS in Ethiopia, Kenya, South Africa and Uganda", *BMJ Open*, 9(2), p. e025469. doi:10.1136/bmjopen-2018-025469.

[61] Wang, Y.-F. et al. (2021) "Multi-ancestral GWAS identifies shared and Asian-specific loci for SLE and links type III interferon signaling and lysosomal function to the disease", *Arthritis & Rheumatology (Hoboken, N.J.)* [Preprint]. doi:10.1002/art.42021.

[62] Swart, Y. et al. (2022) GWAS in the southern African context. *bioRxiv*. doi:10.1101/2022.02.16.480704.

[63] Shen, H. et al. (2020) "Polygenic prediction and GWAS of depression, PTSD, and suicidal ideation/self-harm in a Peruvian cohort", *Neuropsychopharmacology*, 45(10), pp. 1595–1602. doi:10.1038/s41386-020-0603-5.

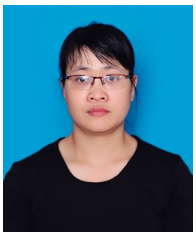
[64] Cardona Tobar, K.M. et al. (2020) "Genome-wide association studies in sheep from Latin America. Review", *Revista mexicana de ciencias pecuarias*, 11(3), pp. 859–883. doi:10.22319/rmcp.v11i3.5372.

[65] Yang, Z. et al. (2021) "Genome-wide association study reveals genetic variations associated with ocean acidification resilience in Yesso scallop Patinopecten yessoensis", *Aquatic Toxicology*, 240, p. 105963. doi:10.1016/j.aquatox.2021.105963.

[66] Peng, W. et al. (2021) "Identification of growth-related SNP and genes in the genome of the Pacific abalone (Haliotis discus hannai) using GWAS", *Aquaculture*, 541, p. 736820. doi:10.1016/j.aquaculture.2021.736820.
[67] Gibbs, R.A. et al. (2003) "The International HapMap Project", *Nature* [Preprint]. doi:10.1038/nature02168.
[68] Siva, N. (2008) "1000 genomes project", *Nature Biotechnology*, 26(3), pp. 256–257.
[69] Jadhav, A., Pramod, D. and Ramanathan, K. (2019) "Comparison of Performance of Data Imputation Methods for Numeric Dataset", *Applied Artificial Intelligence*, 33(10), pp. 913–933. doi:10.1080/08839514.2019.1637138.
[70] Austin, P.C. et al. (2021) "Missing Data in Clinical Research: A Tutorial on Multiple Imputation", *Canadian Journal of Cardiology*, 37(9), pp. 1322–1331. doi:10.1016/j.cjca.2020.11.010.
[71] Choquet, H. et al. (2021) "A large multiethnic GWAS meta-analysis of cataract identifies new risk loci and sex-specific effects", *Nature Communications*, 12(1), p. 3595. doi:10.1038/s41467-021-23873-8.
[72] Powell, V. et al. (2021) "Investigating regions of shared genetic variation in attention deficit/hyperactivity disorder and major depressive disorder: a GWAS meta-analysis", *Scientific Reports*, 11(1), p. 7353. doi:10.1038/s41598-021-86802-1.
[73] Levey, D.F. et al. (2020) GWAS of Depression Phenotypes in the Million Veteran Program and Meta-analysis in More than 1.2 Million Participants Yields 178 Independent Risk Loci. *medRxiv*, p. 2020.05.18.20100685. doi:10.1101/2020.05.18.20100685.
[74] Taherkhani, L. et al. (2022) "The Candidate Chromosomal Regions Responsible for Milk Yield of Cow: A GWAS Meta-Analysis", *Animals*, 12(5), p. 582. doi:10.3390/ani12050582.
[75] Li, J.H. et al. (2021) "Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays", *Genome Research*, 31(4), pp. 529–537. doi:10.1101/gr.266486.120.
[76] Huang, J. et al. (2022) "A Next Generation Sequencing-Based Protocol for Screening of Variants of Concern in Autism Spectrum Disorder", *Cells*, 11(1), p. 10. doi:10.3390/cells11010010.
[77] Chatterjee, N., Shi, J. and García-Closas, M. (2016) "Developing and evaluating polygenic risk prediction models for stratified disease prevention", *Nature Reviews Genetics*, 17(7), pp. 392–406. doi:10.1038/nrg.2016.27.
[78] Natarajan, P. (2018) "Polygenic Risk Scoring for Coronary Heart Disease", *Journal of the American College of Cardiology*, 72(16), pp. 1894–1897. doi:10.1016/j.jacc.2018.08.1041.
[79] Zhang, X. et al. (2018) "Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: A nested case–control study", *PLOS Medicine*, 15(9), p. e1002644. doi:10.1371/journal.pmed.1002644.
[80] Hung, R.J. et al. (2021) "Assessing Lung Cancer Absolute Risk Trajectory Based on a Polygenic Risk Model", *Cancer Research*, 81(6), pp. 1607–1615. doi:10.1158/0008-5472.CAN-20-1237.
[81] Carr, P.R. et al. (2020) "Estimation of Absolute Risk of Colorectal Cancer Based on Healthy Lifestyle, Genetic Risk, and Colonoscopy Status in a Population-Based Study", *Gastroenterology*, 159(1), pp. 129-138.e9. doi:10.1053/j.gastro.2020.03.016.
[82] Darst, B.F. et al. (2021) "Combined Effect of a Polygenic Risk Score and Rare Genetic Variants on Prostate Cancer Risk", *European Urology*, 80(2), pp. 134–138. doi:10.1016/j.eururo.2021.04.013.

SƠ LƯỢC VỀ TÁC GIẢ

Trịnh Thị Xuân



Tốt nghiệp Thạc sỹ ngành Công nghệ Thông tin năm 2007 tại Trường đại học Công nghệ, Đại học Quốc Gia Hà Nội.
Hiện là giảng viên, phó bộ môn cơ sở, khoa công nghệ thông tin, Trường đại học Mở Hà Nội.
Lĩnh vực nghiên cứu: khai phá dữ liệu, tin sinh học.
Email: trinhxuan@hou.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 TNHH LOBI Việt Nam.
Lĩnh vực nghiên cứu: giải trình tự hệ gen người, dự đoán nguy cơ mắc bệnh đa di truyền, dược học hệ gen, mô hình Markov ẩn và mạng Bayes trong tin sinh học.
Email: nhanta@lobi.vn

Hoàng Đỗ Thanh Tùng

Tốt nghiệp Tiến sĩ ngành Khoa học Máy tính tại trường đại học quốc gia Chungbuk, Hàn Quốc.
Hiện là Trưởng phòng Nghiên cứu hệ thống và quản lý, viện Công nghệ Thông tin, VAST.
Lĩnh vực nghiên cứu: các giải pháp lưu trữ, nén, biến đổi, đánh chỉ số và các mô hình cơ sở dữ liệu sinh tinh học nhằm tăng hiệu quả lưu trữ, khai thác thông tin sinh học như về nguồn gốc giống loài, bệnh, thuốc v.v..
Email: tunghdt@gmail.com

Trương Nam Hải

Năm 1988 nhận bằng Tiến sĩ Sinh học phân tử tại Liên Xô. Năm 2004 được phong PGS ngành Sinh học. Năm 2012 được phong GS ngành Sinh học. Nguyên Viện trưởng Viện Công nghệ sinh học và là NCV Cao cấp.
Hiện công tác tại phòng Kỹ thuật di truyền, viện Công nghệ Sinh học, VAST.
Lĩnh vực nghiên cứu: nghiên cứu tạo các protein tái tổ hợp sử dụng trong y dược, nghiên cứu tạo các bộ sinh phẩm chẩn đoán bệnh cho người, động vật, nghiên cứu tạo vaccine bằng protein tái tổ hợp, nghiên cứu gen, loài trong quần thể/khu hệ vi sinh vật bằng kỹ thuật metagenomics.
Email: tnhai@ibt.ac.vn

Trần Đăng Hưng

Tốt nghiệp Thạc sỹ ngành Khoa học máy tính tại trường Đại học Công nghệ, ĐHQGHN năm 2006. Tốt nghiệp Tiến sỹ ngành khai phá tri thức năm 2009 tại viện Khoa học và Công nghệ tiên tiến Nhật Bản (JAIST).
Hiện là Trưởng khoa khoa Công nghệ Thông tin, Trường Đại học Sư phạm Hà Nội.
Lĩnh vực nghiên cứu: khai phá dữ liệu sinh học phân tử, khai phá mạng sinh học, học máy.
Email: hungdt@hnue.edu.vn

Khởi tạo trọng số cho mạng nơ ron tích chập số phức sử dụng giải thuật di truyền

Phạm Minh Tuấn[1], Nguyễn An Hưng[1]

[1] Trường Đại học Bách khoa – Đại học Đà Nẵng, 54 Nguyễn Lương Bằng, Đà Nẵng, Việt Nam

Tác giả liên hệ: Phạm Minh Tuấn, pmtuan@dut.udn.vn

Ngày nhận bài: 01/05/2022, ngày sửa chữa: 01/06/2022, ngày duyệt đăng: 30/06/2022

Định danh DOI: 10.32913/mic-ict-research-vn.v2022.n1.1054

Tóm tắt: Khởi tạo trọng số sử dụng phương pháp Glorot là một trong những phương pháp hữu hiệu cho mạng nơ ron tích chập (CNN). Ngoài ra những nghiên cứu trước đây cũng đã sử dụng phương pháp này cho CNN số phức. Tuy nhiên, không có nghiên cứu nào khẳng định được phương pháp khởi tạo trọng số Glorot có hiệu quả tốt nhất cho CNN số phức. Nghiên cứu này tập trung vào việc tối ưu CNN số phức dựa trên cách thức khởi tạo trọng số kết nối của các tế bào nơ ron. Bài báo này đề xuất sử dụng giải thuật di truyền (GA) để tìm tham số tối ưu cho phân bố Rayleigh để khởi tạo trọng số trước khi huấn luyện CNN số phức. Thực nghiệm cho thấy GA đã tìm ra tham số tốt hơn khi sử dụng phương pháp Glorot như các nghiên cứu trước đây.

Từ khóa: *Mạng nơ ron tích chập, số phức, giải thuật di truyền, phương pháp Glorot.*

Title: **Weight Initialization for Complex Number Convolutional Neural Networks Using Genetic Algorithms**

Abstract: Weight initialization using Glorot method is one of the effective methods for convolutional neural networks (CNN). In addition, previous studies have also used this method for complex number CNNs. However, there is no research to confirm that Glorot weight initialization method has the best effect for complex number CNN. This study focuses on optimizing complex number CNNs based on how to initialize connection weights of neurons. This paper proposes to use genetic algorithm (GA) to find optimal parameters for Rayleigh distribution to initialize weights before training complex number CNN. Experiments show that GA has found better parameters when using the Glorot method as in previous studies

Keywords: *Convolutional neural networks, complex number, genetic algorithms, glorot method.*

I. MỞ ĐẦU

Mạng nơron tích chập số phức [1, 2] (Complex Valued Convolutional Neural Networks – CNN số phức) là trong những phương pháp hiệu quả hiện nay trong nhận dạng [3], xử lý nhiễu ảnh [4, 5] hay xử lý tín hiệu [6]. Có 2 hướng tiếp cận chính đối với tế bào nơ ron trong mô hình CNN số phức [7]. Hướng thứ nhất là tách riêng thành 2 tế bào nơ ron riêng biệt, nơ ron phụ trách tính toán phần thực và nơ ron phụ trách tính toán phần ảo. Hướng thực hiện này có thể giúp cho các nghiên cứu có kế thừa mô hình CNN số thực một cách dễ dàng nhưng nó làm mất đi tính chất đặc biệt của số phức là tích 2 số phức hoạt động như một phép biến đổi hình học trong không gian 2 chiều. Hướng tiếp cận thứ 2 là áp dụng số phức cho từng nơ ron trong mạng CNN số phức [8]. Các tín hiệu đầu vào, đầu ra và trọng số kết nối giữa các lớp trong mạng CNN số phức đều sử dụng số phức. Hướng tiếp cận này cho thấy sự hiệu quả trong các bài toán có tín hiệu đầu vào hay đầu ra là số phức như kết hợp với phương pháp biến đổi Fourier nhanh (Fast Fourier Transform – FFT) để khử nhiễu hay nhận dạng tín hiệu, hình ảnh [7].

Bài báo này tập trung vào phân tích hướng tiếp cận thứ 2 là áp dụng số phức vào từ nơ ron trong mạng CNN số phức. Để huấn luyện 1 mạng CNN số phức, các phương pháp trước đây chú trọng việc xây dựng cấu trúc mạng, việc khởi tạo các trọng số bằng số phức ít được chú ý tới. Một số phương pháp [9, 10] có đề xuất sử dụng phương pháp Glorot [11] trong việc khởi tạo trọng số của các mạng CNN số thực để áp dụng cho CNN số phức. Tuy nhiên các nghiên cứu trước đây chưa chứng minh được việc khởi tạo như vậy là tối ưu nhất.

Nghiên cứu này đề xuất phương pháp khởi tạo trọng số cho mạng CNN số phức sử dụng giải thuật di truyền [12]. Bài báo này gồm 5 mục, mở đầu nêu bối cảnh và lý do lựa

chọn phương pháp đề xuất. Tiếp theo bài báo giới thiệu các nghiên cứu liên quan tới mạng CNN số phức và giải thuật di truyền. Tiếp đó là giải pháp đề xuất và thực nghiệm và đánh giá kết quả. Cuối cùng là kết luận.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Nghiên cứu này chú trọng việc khởi tạo trọng số cho mạng CNN số phức sử dụng giải thuật di truyền. Vì vậy trong mục này sẽ giới thiệu mạng CNN số phức, trong đó có cách thức khởi tạo trước đây. Tiếp theo, mục này sẽ giới thiệu giải thuật di truyền trong các bài toán tối ưu các tham số trong các mô hình học máy, đặc biệt là mạng nơ ron nhân tạo.

1. Mạng nơ ron tích chập số phức

Số phức là số trong không gian $\mathbf{C}$ được biểu diễn dưới dạng:

$$z = x + \mathbf{i}y \in \mathbf{C} \tag{1}$$

Với $i^2 = -1$ là số ảo và $x, y \in \mathbf{R}$ là các số thực. Một hàm f trong không gian $\mathbf{C}$ được xác định bởi:

$$f : \mathbf{C} \to \mathbf{C} \tag{2}$$

$$f(z) = u(z) + \mathbf{i}v(z) \tag{3}$$

với $u, v : \mathbf{R} \to \mathbf{R}$ là các hàm được xác định trong không gian số thực.

Mạng nơ ron số phức [13][14] được triển khai tượng tự như mạng nơ ron số thực bằng cách định nghĩa lại cách nhân đối với trọng số $\mathbf{w} = w_r + \mathbf{i}w_i$ và tín hiệu đầu vào $\mathbf{x} = x_r + \mathbf{i}x_i \in \mathbf{C}$ của từng tế bào nơ ron.

$$\begin{aligned} \mathbf{wx} &= (w_r + \mathbf{i}w_i)(x_r + \mathbf{i}x_i) \\ &= (w_r x_r - w_i x_i) + \mathbf{i}(w_r x_i + w_i x_r) \in \mathbf{C} \end{aligned} \tag{4}$$

Đối với, mạng CNN số phức [8], ta coi như các trọng số là một ma trận số phức $W = W_r + iW_i$ và lớp đầu vào cũng là một ma trận số phức $X = X_r + iX_i$ với $W_r, W_i \in R^{S_1 * S_2}$ và $X_r, X_i \in R^{L_1 * L_2}$ lần lượt là các ma trận số thực, $S_1 * S_2$ là kích thước của ma trận trọng số và $L_1 * L_2$ là kích thước của ma trận đầu vào. Lúc này tích chập của 2 ma trận trọng số và ma trận đầu vào là một ma trận kích thước $L_1 * L_2$ với các phần tử $Y_{k,l}$ được xác định bởi công thức:

$$\begin{aligned} Y_{k,l} &= \sum_{x=1}^{S_1}\sum_{y=1}^{S_2} W_{x,y} X_{k+x,l+y} \\ &= \sum_{x=1}^{S_1}\sum_{y=1}^{S_2} (W_{x,y_r} X_{k+x,l+y_r} - W_{x,y_i} X_{k+x,l+y_i}) \\ &+ i \sum_{x=1}^{S_1}\sum_{y=1}^{S_2} (W_{x,y_r} X_{k+x,l+y_i} + W_{x,y_i} X_{k+x,l+y_r}) \end{aligned} \tag{5}$$

Việc khởi tạo ma trận trọng số trước khi huấn luyện thường được thực hiện dựa trên phương pháp của Glorot [12] nhằm xác định độ lệch chuẩn của các phần tử trong ma trận trọng số trong lúc khởi tạo. Gọi $w = w_r + iw_i$ là một phần tử bất kỳ trong ma trận trọng số, ta có thể biểu diễn w thành:

$$w = |w|e^{i\Theta} \tag{6}$$

Với Θ và $|w|$ lần lượt là góc và độ lớn của w. Phương sai của w và $|w|$ được tính bởi:

$$Var(w) = E(|w|^2) - (E(w))^2 \tag{7}$$

$$Var(|w|) = E(|w|^2) - (E(|w|))^2 \tag{8}$$

Khi ta cho giá trị kỳ vọng của w là $E(w) = 0$, từ (7) và (8) ta có:

$$Var(w) = Var(|w|) - (E(|w|))^2 \tag{9}$$

Dựa trên phân bố Rayleigh với tham số σ ta có:

$$E(|w|) = \sigma\sqrt{\frac{\pi}{2}} \tag{10}$$

và

$$Var(|w|) = \frac{4-\pi}{2}\sigma^2 \tag{11}$$

Suy ra:

$$Var(w) = \frac{4-\pi}{2}\sigma^2 + \left(\sigma\sqrt{\frac{\pi}{2}}\right)^2 = 2\sigma^2 \tag{12}$$

Theo như Glorot thì ta cần $Var(w) = \dfrac{2}{n_{in} + n_{out}}$, với n_{in}, n_{out} lần lượt là số lượng tín hiệu đầu vào và đầu ra của tế bào nơ ron. Vì vậy $\sigma = \sqrt{\dfrac{1}{n_{in} + n_{out}}}$ là giá trị tham số của phân bố Rayleigh trong các nghiên cứu trước đây. Tuy nhiên việc khởi tạo trọng số đựa trên Glorot đối với CNN số phức có thực sự là phương án tối ưu hay chưa thì chưa được sáng tỏ. Vì thế, bài báo này đề xuất sử dụng giải thuật di truyền nhằm tìm ra phương pháp khởi tạo trọng số cho CNN số phức.

2. Giải thuật di truyền

Giải thuật di truyền (Genetic Algorithm – GA) [15] là một phương pháp tối ưu tổ hợp dựa trên cơ chế di truyền của sinh vật. Giải thuật di truyền thường được sử dụng để tìm ra giải pháp tối ưu cho các bài toán NP khó. Giải thuật di truyền được mô tả như sau:

Bước 1: Tạo tập nghiệm ngẫu nhiên

Bước 2: Đánh giá các nghiệm hiện có

Bước 3: Nếu nghiệm tốt nhất thỏa mãn yêu cầu thì nhảy đến Bước 7

Bước 4: Lựa chọn các nghiệm tốt để nhân bản và các nghiệm xấu để loại bỏ

Bước 5: Lai ghép các cặp nghiệm để tạo ra nghiệm mới

Bước 6: Đột biến một số nghiệm trong tập nghiệm.

Bước 7: In ra kết quả tốt nhất

Để hiểu rõ hơn về giải thuật di truyền, ta lấy ví dụ sử dụng giải thuật di truyền để giải bài toán tìm nghiệm của phương trình:

$$x^2 - 12x + 36 = 0 \tag{13}$$

Bước 1: Thiết kế bộ gen là các số trong hệ nhị phân. Phương án nghiệm của bài toán là 1 số khi đổi qua cơ số thập phân của bộ gen. Ví dụ, ta có thể tạo ngẫu nhiên tập nghiệm ban đầu như sau: 0100, 1011, 0101, 1010.

Bước 2: Cần 1 tiêu chí đánh giá các nghiệm hiện có. Ta có thể chọn hàm đánh giá $f(x) = |x^2 - 12x + 36|$ vì $f(x)$ càng nhỏ thể hiện phương án nghiệm càng gần với kết quả chính xác. Khi đã có hàm đánh giá, ta tính giá trị f đối với tất cả các phương án nghiệm đang có. Ví dụ: $f(0100) = 4, f(1011) = 25, f(0101) = 1, f(1010) = 16$.

Bước 3: Kiểm tra kết thúc việc thực thi hay không. Giả sử tiêu chí kết thúc thuật toán $f(x) = 0$ thì chưa có nghiệm nào thỏa mãn tiêu chí kết thúc.

Bước 4: Lựa chọn các nghiệm tốt để nhân bản và các nghiệm xấu để loại bỏ. Ví dụ: nghiệm 1011 có $f(1011) = 25$ bị loại bỏ và nghiệm 0101 có$f(0101) = 1$ được nhân bản. Sau khi kết thúc bước này, tập hợp nghiệm sẽ là: 0101, 0101, 0100, 1010.

Bước 5: Lai ghép. Có nhiều hình thức lai ghép như lai ghép 1 điểm, lai ghép nhiều điểm hay lai ghép đồng nhất. Giả sử ta chọn lai ghép 1 điểm tại vị trí $k = 1$ thì có nghĩa là bít đầu nghiệm thứ nhất sẽ ghép với 3 bít sau của nghiệm thứ hai và ngược lại. Trong quá trình lai ghép theo cách thức trên, nếu cặp nghiệm 0101 và 1010 được lai ghép thì kết quả sẽ được 2 nghiệm mới là 1|101 và 0|010. Ta thay thế 2 nghiệm được lai ghép thành 2 nghiệm mới trong tập nghiệm.

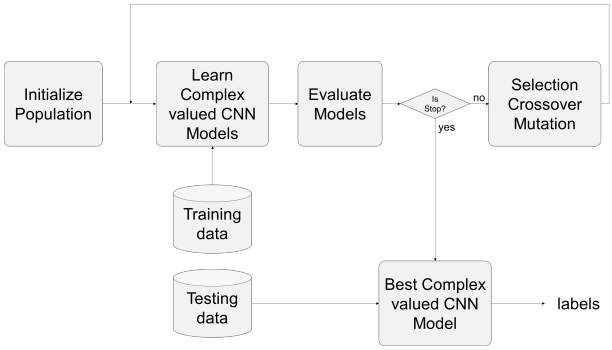
Bước 6: Đột biến. Tại một nghiệm bất kỳ, đổi giá trị 1 thành 0 hoặc 0 thành 1 ở tại 1 bít nào đó. Ví dụ 0101 được chọn và đột biến thành 010**0**, hoặc 0010 được chọn và đột biến thành 0**1**10.

Quá trình từ **Bước 2** đến **Bước 6** sẽ được lặp đi lặp lại đến khi tìm được cá thể thỏa mãn tiêu chí của bài toán. Trong ví dụ trên thì nghiệm 0110 chính là nghiệm tốt nhất của bài toán.

Việc áp dụng GA trong việc tối ưu các tham số hay cấu trúc mô hình học máy được thực hiện rất nhiều trong nghiên cứu trước đây, đặc biệt là đối với các mô hình mạng nơ ron nhân tạo. Whitley Darrell [Whitley 1990] đề xuất sử dụng GA nhằm tối ưu cách thức kết nối giữa các tế bào nơ ron trong mô hình ANN. Cook [16] đề xuất phương pháp ứng dụng GA nhằm tối ưu các tham số trong mô hình ANN. Karegowda [17] đề xuất phương pháp tối ưu trong số kết nối trong mô hình ANN sử dụng GA. Aszemi [18] đề xuất phương pháp tối ưu tham số cho mô hình CNN sử dụng GA. Chung [19] đề xuất phương pháp sử dụng GA trong mô hình CNN nhiều kênh (Multi-channel CNN).

III. PHƯƠNG PHÁP ĐỀ XUẤT

Bài báo này đề xuất mô hình huấn luyện CNN số phức kết hợp với GA nhằm tối ưu tham số σ của phân bố Rayleigh trong quá trình khởi tạo trọng số đựa trên Glorot đối với CNN số phức. Hình 1 thể hiện mô hình khởi tạo trọng số cho CNN số phức sử dụng GA.



Hình 1. Mô hình khởi tạo trọng số cho CNN số phức sử dụng GA.

Các bước cụ thể trong mô hình được trình bày như sau:

Bước 1: Khởi tạo tập $n_{GA} = 50$ các cá thể.

Mục đích của GA là nhằm tối ưu tham số σ của phân bố Rayleigh trong quá trình khởi tạo trọng số đựa trên Glorot đối với CNN số phức. Bài báo này đề xuất:

$$\sigma = \left(\frac{1}{n_{in} + n_{out}}\right)^{\rho} \tag{14}$$

với $\rho \in (0, 1)$ là tham số cần tìm trong quá trình tối ưu hàm đánh giá $f(CNN_{complex}(\rho))$. Trong đó, $CNN_{complex}(\rho)$ là mô hình CNN số phức với khởi tạo trọng số theo phân bố Rayleigh. Mỗi cá thể đại diện cho ρ và được biểu diễn bằng 1 số ở hệ nhị phân 32 bit.

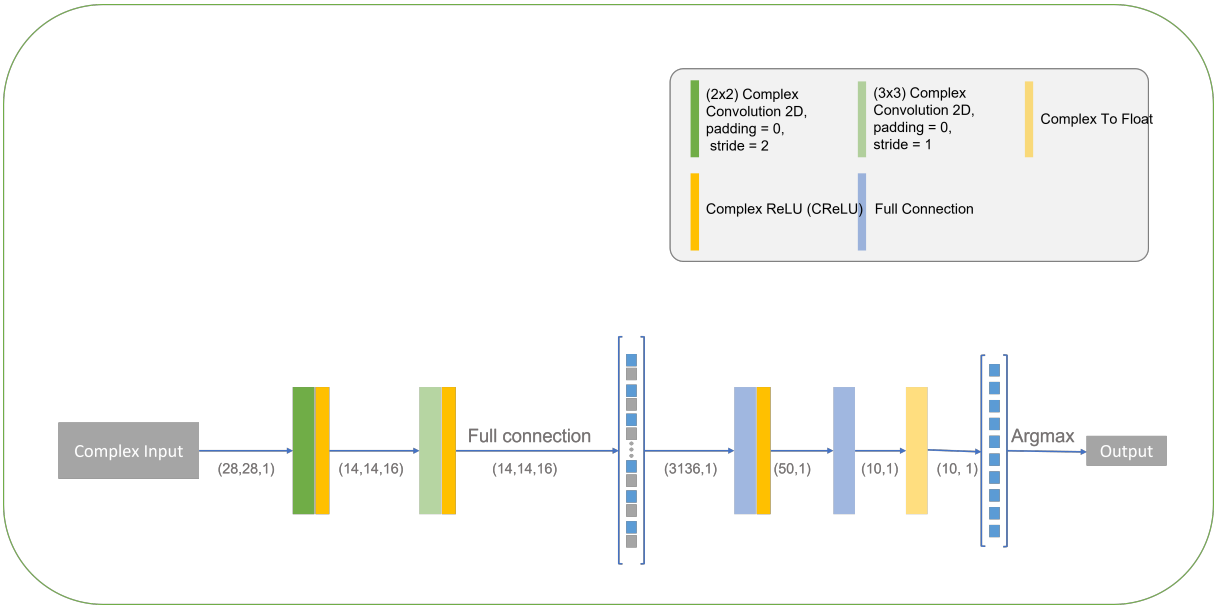
$$\overline{b_1 b_2 ... b_{32}}, b_i \in \{0, 1\} \tag{15}$$

và

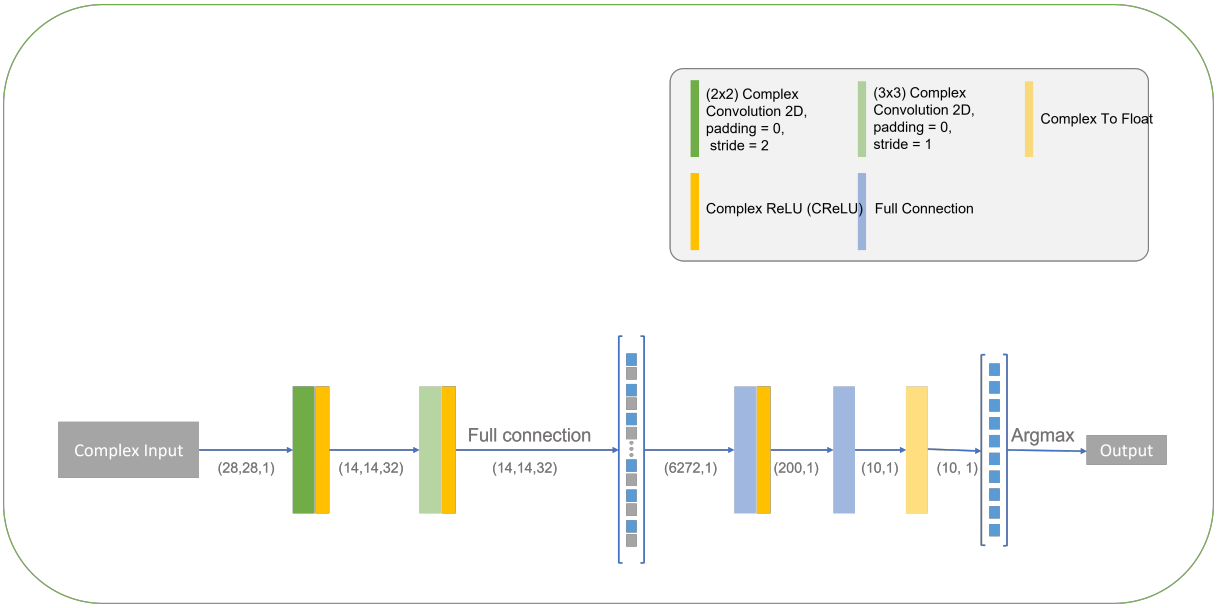
$$\rho = \sum_{i=1}^{32} b_i 2^{-i} \tag{16}$$

Bước 2: Ứng với mỗi cá thể ρ, nghiên cứu xây dựng mô hình CNN số phức $CNN_{complex}(\rho)$ sử dụng dữ liệu huấn luyện. Bài báo sử dụng 2 mô hình CNN số phức tại Hình 2 và 3 để đánh giá mức độ tin cậy của GA.

Bước 3: Đánh giá các mô hình CNN số phức bằng cách tính độ chính xác của việc nhận dạng dựa trên tập huấn luyện. Nếu độ chính xác của mô hình tốt nhất đủ lớn, quá



Hình 2. Mô hình CNN số phức với số lượng nơ ron nhỏ



Hình 3. Mô hình CNN số phức với số lượng nơ ron lớn

trình GA sẽ kết thúc và đưa ra mô hình CNN tốt nhất đó cũng như tham số ρ, ngược lại thì tiếp tục Bước 4.

Bước 4: Thực hiện các toán tử của GA là lựa chọn, lại ghép và đột biến. Việc lựa chọn được thực hiện bằng cách chọn $p_{selection} = 0.8$ các cá thể tốt nhất để giữ lại cho thế hệ sau,$1 - p_{selection}$ còn lại sẽ bị loại bỏ và thay thế bằng cách nhân bản ngẫu nghiên các cá thể trong thế hệ hiện tại. Việc lai ghép được thực hiện với tỉ lệ $p_{crossover} = 0.3$ số cặp các cá thể và sử dụng phương pháp lai ghép đồng nhất (uniform crossover). Quá trình đột biến được tiến hành ngẫu nhiên với tỉ lệ $p_{mutation} = 0.01$.

Việc đánh giá được dựa trên kết quả nhận dạng đối với tập kiểm thử trên mô hình CNN số phức tốt nhất. Kết quả nhận dạng của phương pháp đề xuất được kỳ vọng sẽ tốt hơn phương pháp sử dụng khởi tạo trọng số đựa trên Glorot đối với CNN số phức bởi vì phương pháp Glorot là phương pháp sử dụng tham số σ với công thức cố định.

$$\sigma = \sqrt{\frac{1}{n_{in} + n_{out}}} \tag{17}$$

Trong khi phương pháp đề xuất là phương pháp tìm kiếm

tối ưu đối với tham số σ không cố định tùy thuộc vào $\rho \in (0, 1)$.

$$\sigma = \left(\frac{1}{n_{in} + n_{out}}\right)^{\rho} \tag{18}$$

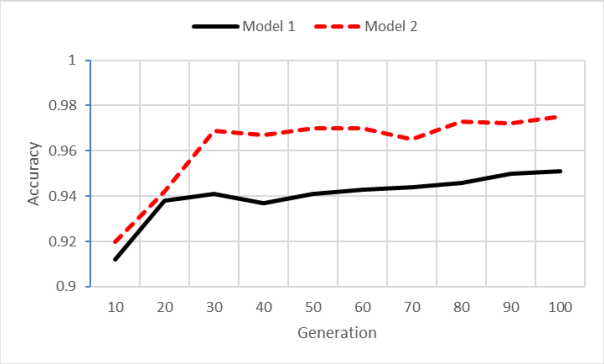
Vì thế, phương pháp đề xuất sẽ có kết quả không thể thấp hơn phương pháp dựa trên Glorot.

Tuy nhiên, về mặt thời gian để tìm ra tham số tối ưu của phương pháp đề xuất thì rõ ràng phải thực hiện rất nhiều lần việc huấn luyện mạng ứng với từng tham số ρ trong giải thuật di truyền. Bài báo này chỉ chú trọng việc chứng minh rằng phương pháp khởi tạo trọng số dựa trên Glorot chưa phải là phương pháp tốt nhất, nên không chú trọng đến việc tối ưu về mặt thời gian huấn luyện.

IV. KẾT QUẢ THỰC NGHIỆM

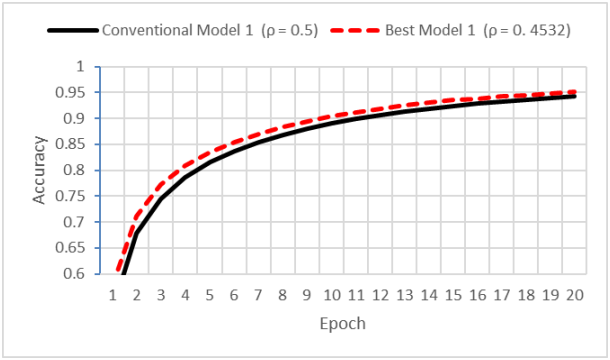
Để đánh giá mô hình đề xuất, nghiên cứu đã sử dụng dữ liệu MNIST Handwritten Digit Database [20] để sinh ngẫu nhiên 1000 mẫu cho tập huấn luyện và 5000 mẫu cho tập kiểm thử. Nghiên cứu sử dụng GA để tìm tham số ρ ứng với 2 mô hình CNN số phức như ở Hình 2 (Model 1) và Hình 3 (Model 2).

Hình 4 là kết quả nhận dạng đúng của 2 mô hình CNN số phức qua các thế hệ GA. Hình 4 cho thấy kết quả nhận dạng đúng của cả 2 mô hình CNN số phức tăng dần qua từng thế hệ trong GA.

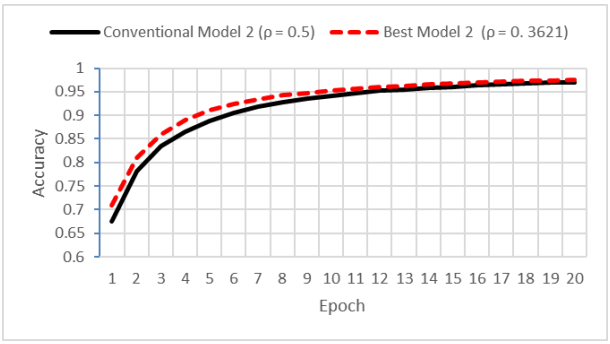


Hình 4. Kết quả nhận dạng đúng của 2 mô hình CNN số phức qua các thế hệ GA

Hình 5 và Hình 6 là kết quả so sánh tỉ lệ nhận dạng đúng biến thiên theo từng epoch giữa mô hình CNN số phức tốt nhất do GA đưa ra và mô hình CNN số phức trước đây ($\rho = 0.5$). Kết quả so sánh cho thấy Mô hình CNN số phức tốt nhất do GA đưa ra đều tốt hơn mô hình CNN số phức với tham số được đề xuất bởi Glorot. Điều đó cho thấy việc khởi tạo trọng số sử dụng phương pháp Glorot chưa phải phù hợp nhất cho mô hình CNN số phức. Cả hai mô hình tốt nhất đều có khuynh hướng $\rho_{GA} < \rho_{Glorot}$.



Hình 5. Tỉ lệ nhận dạng đúng biến thiên theo epoch đối với mô hình CNN số phức (Model 1) với tham sốρ được đề xuất bởi phương pháp trước đây và tham số tốt nhất do GA tìm được



Hình 6. Tỉ lệ nhận dạng đúng biến thiên theo epoch đối với mô hình CNN số phức (Model 2) với tham số ρ được đề xuất bởi phương pháp trước đây và tham số tốt nhất do GA tìm được

V. KẾT LUẬN

Bài báo này đã đề xuất mô hình sử dụng thuật toán di truyền để tìm tham số tối ưu cho việc khởi tạo trọng số trong mô hình mạng nơ ron tích chập số phức. Kết quả thực nghiệm đã cho thấy việc khởi tạo với tham số được đề xuất bởi mô hình đề xuất tốt hơn so với việc khởi tạo bằng phương pháp Glorot. Việc sử dụng thuật toán di truyền để tìm tham số bắt buộc phải xây dựng rất nhiều mô hình mạng nơ ron tích chập số phức nên việc tìm ra mô hình tối ưu nhất là điều có thể lường trước được. Tuy nhiên, việc thực nghiệm với thuật toán di truyền này đã chứng minh được rằng phương pháp khởi tạo trọng số Glorot chưa phù hợp với mô hình số phức. Hướng phát triển trong thời gian tới của nghiên cứu này là tìm ra được công thức phù hợp hơn cho tham số khởi tạo trọng số của mô hình mạng nơ ron số phức để thay thế cho phương pháp Glorot. Tiếp đến đưa ra các mô hình toán học để chứng minh tính đúng đắn của công thức đã tìm thấy. Cuối cùng là thử nghiệm trên các dữ liệu lớn hơn để kiểm chứng tính đúng đắn đó.

TÀI LIỆU THAM KHẢO

[1] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, "Complex-valued convolutional neural network and its application in polarimetric sar image classification," *IEEE Transactions on Geoscience and Remote Sensing 55.12*, pp. 7177–7188, 2017.

[2] N. Guberman, "On complex valued convolutional neural networks," *arXiv preprint*, 2016.

[3] C.-A. Popa, "Complex-valued convolutional neural networks for real-valued image classification," *2017 International Joint Conference on Neural Networks (IJCNN). IEEE*, 2017.

[4] S. Rawat, K.P.S.Rana, and V. Kumar, "A novel complex-valued convolutional neural network for medical image denoising," *Biomedical Signal Processing and Control 69*, p. 102859, 2021.

[5] Y. Quan, Y. Chen, Y. Shao, H. Teng, Y. Xu, and H. Ji, "Image denoising using complex-valued deep cnn," *Pattern Recognition 111*, p. 107639, 2021.

[6] J. Gao, B. Deng, Y. Qin, H. Wang, and X. Li, "Enhanced radar imaging using a complex-valued convolutional neural network," *IEEE Geoscience and Remote Sensing Letters 16.1*, 2018.

[7] M. T. Pham, V. Q. Nguyen, C. D. Hoang, H. L. Vo, D. K. Phan, and A. H. Nguyen, "Efficient complex valued neural network with fourier transform on image denoising." *The 5th International Conference on Future Networks & Distributed Systems*, 2021.

[8] Bassey, Joshua, L. Qian, , and X. Li, "A survey of complex-valued neural networks," *arXiv preprint*, 2021.

[9] H. Pratt, B. Williams, and F. C. . Y. Zheng, "Fcnn: Fourier convolutional neural networks," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham,*, 2017.

[10] D. A. B. O. S. F. B. J.-Y. S. M. Cord, "Complex-valued neural networks for fully-temporal micro-doppler classification," *2019 20th International Radar Symposium (IRS). IEEE*, 2019.

[11] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks." *Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings,*, 2011.

[12] W. Darrell, T. Starkweather, , and C. Bogart., "Genetic algorithms and neural networks: Optimizing connections and connectivity," *Parallel computing 14.3*, pp. 347–361, 1990.

[13] A. Hirose, "Complex-valued neural networks," *Springer Science & Business Media*, vol. 400, 2012.

[14] Virtue, Patrick, X. Y. Stella, , and M. Lustig, "Better than real: Complex-valued neural nets for mri fingerprinting." *2017 IEEE international conference on image processing (ICIP).*, 2017.

[15] Mirjalili and Seyedali., "Genetic algorithm," *Evolutionary algorithms and neural networks. Springer, Cham*, pp. 43–55, 2019.

[16] C. D. F., C. T. Ragsdale, , and R. L. Major, "Combining a neural network with a genetic algorithm for process parameter optimization." *Engineering applications of artificial intelligence 13.4*, pp. 391–396, 2000.

[17] K. A. Gowda, A. S. Manjunath, , and M. A. Jayaram., "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima indians diabetes," *International Journal on Soft Computing 2.2*, pp. 15–23, 2011.

[18] A. N. Mohd, , and P. D. D. Dominic, "Hyperparameter optimization in convolutional neural network using genetic algorithms." *Int. J. Adv. Comput. Sci. Appl 10.6*, pp. 269–278, 2019.

[19] C. Hyejung, , and K. shik Shin, "Genetic algorithm-optimized multi-channel convolutional neural network for stock market prediction," *Neural Computing and Applications 32.12*, pp. 7897–7914, 2020.

[20] Deng and Li, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE signal processing magazine 29.6*, pp. 141–142, 2012.

SƠ LƯỢC VỀ TÁC GIẢ

Phạm Minh Tuấn

Nhận học vị tiến sĩ chuyên ngành Khoa học tính toán tại đại học Nagoya, Nhật Bản năm 2012. Từ năm 2012 là giảng viên khoa Công nghệ thông tin tại trường Đại học Bách Khoa - Đại Học Đà Nẵng. Từ năm 2015 là Trưởng bộ môn Mạng và truyền thông, khoa CNTT, trường Đại học Bách khoa - Đại học Đà Nẵng. Từ năm 2020 là Phó trưởng Khoa Công nghệ thông tin, trường Đại học Bách khoa - Đại học Đà Nẵng.

Lĩnh vực Nghiên cứu: trí tuệ nhân tạo, tính toán mềm, đại số hình học.

Email: pmtuan@dut.udn.vn

Nguyễn An Hưng

Là sinh viên khoa Công nghệ thông tin, trường Đại học Bách Khoa - Đại Học Đà Nẵng.

Lĩnh vực Nghiên cứu: trí tuệ nhân tạo, xử lý dữ liệu, toán học và khoa học máy tính, khoa học dữ liệu trong sinh học.

Email: 102210009@dut.udn.vn

Thuật toán Ro-CS giải bài toán lập lịch với tài nguyên giới hạn

Nguyễn Thế Lộc[1], Đặng Quốc Hữu[2]
[1] Trường Đại học Sư phạm Hà Nội, Việt Nam
[2] Trường Đại học Thương mại, Hà Nội, Việt Nam
Tác giả liên hệ: Đặng Quốc Hữu, huudq@tmu.edu.vn
Ngày nhận bài: 01/05/2022, ngày sửa chữa: 01/06/2022, ngày duyệt đăng: 30/06/2022
Định danh DOI: 10.32913/mic-ict-research-vn.v2022.n1.1055

Tóm tắt: Bài báo đề xuất phương pháp tìm lời giải cho Bài toán MS-RCPSP (Multi Skill-Resource Constrained Project Scheduling Problem). MS-RCPSP đã được chứng minh là bài toán NP-Khó, do vậy cần sử dụng các phương pháp tính toán tiến hóa, cận tối ưu nhằm tìm được lời giải phù hợp trong thời gian chấp nhận được. Thuật toán đề xuất là thuật toán lai ghép giữa thuật toán Cuckoo Search (CS) và kỹ thuật Rotate giúp mở rộng không gian tìm kiếm sau mỗi thế hệ tiến hóa, nhằm tăng khả năng tìm được lời giải tốt hơn. Thuật toán mới gọi là Ro-CS có khả năng áp dụng để tìm lời giải cho bài toán MS-RCPSP. Để kiểm chứng thuật toán Ro-CS, bài báo tiến hành thực nghiệm trên bộ dữ liệu chuẩn iMOPSE, kết quả thực nghiệm được tổng hợp, đánh giá, so sánh, phân tích cho thấy tính hiệu quả của thuật toán đề xuất.

Từ khóa: *Tính toán tiến hóa, thuật toán cận tối ưu, bài toán MS-RCPSP.*

Title: **The Ro-CS Algorithm Solving the MS-RCPSP Problem**

Abstract: This paper proposes a method to solve the MS-RCPSP problem (Multi Skill-Resource Constrained Project Scheduling Problem). MS-RCPSP is an NP-Difficult, so it is necessary to use evolutionary, optimization methods to find a suitable solution in an acceptable time. The proposed algorithm is a hybrid be-tween Cuckoo Search (CS) algorithm and Rotate technique to expand the investi-gation space after each evolutionary generation, to increase the possibility of find-ing a better solution. The new algorithm called Ro-CS is suitable to find a feasible solution for the MS-RCPSP. The proposed algorithm has been experimented on the iMOPSE standard dataset, and the results are synthesized, evaluated, com-pared, and analyzed to denote the effectiveness of the proposed algorithm.

Keywords: *Evolutionary computing, optimization, MS-RCPSP problem.*

I. GIỚI THIỆU

Bài toán lập lịch thực hiện dự án với tài nguyên giới hạn RCPSP (Resource Con-strained Project Scheduling Problem)[1, 2], là bài toán đặt ra mục tiêu tìm phương án lịch biểu để bố trí các tài nguyên cho trước thực hiện và hoàn thành mọi công việc (tác vụ) trong một dự án với thời gian ngắn nhất hoặc chi phí thấp nhất hoặc cân bằng cả hai yếu tố thời gian và chi phí. Thông thường, số lượng tác vụ cần thực hiện của dự án lớn hơn nhiều so với số lượng tài nguyên có thể sử dụng. Bài toán MS-RCPSP (Multi Skill - RCPSP)[1–4] mở rộng từ RCPSP bằng việc bổ sung thêm ràng buộc mới, gần với các dự án trong thực tế hơn, đó là: mỗi tài nguyên có thể có nhiều kỹ năng khác nhau và mỗi kỹ năng có mức/bậc kỹ năng nhất định. Mỗi tác vụ sẽ có yêu cầu về tài nguyên thực hiện cần đáp ứng loại kỹ năng cụ thể và mức kỹ năng nhất định. Khi đó, các tài nguyên có kỹ năng đúng theo yêu cầu của tác vụ và mức kỹ năng lớn hơn hoặc bằng so với yêu cầu sẽ có năng lực để thực hiện tác vụ. MS-RCPSP đã được chứng minh thuộc lớp NP-Khó nên không thể tìm ra lời giải tối ưu trong thời gian đa thức, thay vì đó, nhiều nghiên cứu đã tập trung vào việc tìm các phương pháp giải tiến hóa, cận tối ưu để tìm gia lời giải chấp nhận được và đủ tốt trong thời gian phù hợp.

Phần tiếp theo của bài báo có cấu trúc như sau: Phần II giới thiệu một số công trình nghiên cứu có liên quan đến bài toán MS-RCPSP. Phần III mô tả phát biểu toán học của bài toán. Phần IV trình bày thuật toán đề xuất Ro-CS để tìm lời giải cận tối ưu cho bài toán MS-RCPSP. Phần V mô tả các thực nghiệm để kiểm chứng thuật toán, phân tính, đánh giá chất lượng lời giải. Phần VI tóm tắt những kết quả chính của bài báo và hướng nghiên cứu sẽ tiến hành trong tương lai.

II. NHỮNG CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Bài toán lập lịch với tài nguyên giới hạn và đa kỹ năng MS-RCPSP có nhiều ứng dụng trong khoa học và thực tiễn nên đã được nghiên cứu rộng rãi bởi nhiều nhà khoa học và đã có nhiều công bố khoa học liên quan đến bài toán này. Myszkowski và cộng sự [5] có nhiều công bố về bài toán này, nhóm tác giả đã đề xuất các phương pháp tìm lời giải cho bài toán MS-RCPSP dựa trên các thuật toán như: thuật toán Đàn kiến [4]; thuật toán GA [5, 14] và một số thuật toán khác. Một đóng góp quan trọng của Myszkowski và cộng sự là đã xây dựng và công bố bộ dữ liệu chuẩn iMOPSE [5, 6] sử dụng cho để thực nghiệm các thuật toán cho bài toán này. Hosseinian [7] và cộng sự cũng có nhiều nghiên cứu về bài toán MS-RCPSP. Năm 2018, Hosseinian và Baradaran công bố kết quả nghiên cứu [7] về bài toán Multi-mode Multi-skilled Resource-Constrained Project Scheduling Problem (MMSRCPSP), một biến thể của bài toán MS-RCPSP trong đó bổ sung thêm ràng buộc rằng mỗi tác vụ chỉ có thể được thực thi ở một vài chế độ (mode) định trước và không thể thay đổi mode một khi quá trình thực thi đã được bắt đầu. Để giải quyết bài toán, Hosseinian và cộng sự đã sử dụng thuật toán GA cổ điển kết hợp với phương pháp ra quyết định dựa trên độ đo thông tin Shanon-entropy để lựa chọn cá thể cho thế hệ kế tiếp. Năm 2019 nhóm tác giả này công bố bài báo [8] trong đó sử dụng thuật toán Dandelion Algorithm để giải quyết bài toán MS-RCPSP, trong bài này nhóm tác giả lựa chọn thực nghiệm trên iMOPSE [5, 6, 14].

Javanmard và cộng sự sử dụng 2 thuật toán tiến hóa truyền thống là GA và PSO để lên phương án lịch biểu cho các tài nguyên đa kỹ năng (trên thực tế là các công nhân và kỹ sư) thực hiện một dự án trong ngành hóa chất với mục tiêu tối thiểu hóa tổng chi phí của dự án [9]. Hai thuật toán đề xuất được kiểm chứng trên bộ dữ liệu PSPLIB, vốn không hoàn toàn thích hợp với những bài toán lập lịch có mục tiêu là tối thiểu hóa chi phí.

H. Davari-Ardakani [10] đề xuất một biến thể đa mục tiêu của bài toán MS-RCPSP nhằm tối thiểu hóa hai đại lượng là thời gian và chi phí thực hiện của dự án. Dừng lại ở việc phát biểu bài toán, H. Davari-Ardakani và đồng nghiệp [11] chưa đề xuất được giải pháp mới mà chỉ tiến hành thực nghiệm với những giải pháp đã có như Max-min, vì vậy bài toán có thể coi là vẫn còn để mở.

III. PHÁT BIỂU BÀI TOÁN

Là bài toán mở rộng từ bài toán RCPSP, MS-RCPSP [1–3, 12] bổ sung thêm ràng buộc về kỹ năng của tài nguyên thực hiện: mỗi tài nguyên có nhiều kỹ năng (skill) khác nhau và mỗi kỹ năng có một mức/bậc (level) cụ thể. Để thực hiện được tác vụ, tài nguyên cần đáp ứng về kỹ năng và mức kỹ năng theo yêu cầu của tác vụ. Các ràng buộc của bài toán quy định, tài nguyên có cùng loại kỹ năng và có mức kỹ năng lớn hơn hoặc bằng mức kỹ năng mà tác vụ yêu cầu thì có thể thực hiện được tác vụ đó.

Phát biểu toán học của bài toán MS-RCPSP

Các ký hiệu:

- C_i: các tác vụ cần thực hiện và hoàn thành trước tác vụ i;
- S: tập kỹ năng của tất cả các tài nguyên; S^i: tập các kỹ năng của tài nguyên i, $S^i \subseteq S$;
- S_i: kỹ năng thứ i;
- t_j: thời gian thực hiện tác vụ j;
- L: các tài nguyên có khả năng tái sử dụng được dùng trong dự án;
- L^k: các tài nguyên đáp ứng yêu cầu của tác vụ k; $L^k \subseteq L$;
- L_i: tài nguyên thứ i;
- W: các tác vụ cần thực hiện của dự án;
- W^k: tập tác vụ phù hợp với năng lực của tài nguyên k , $W^k \subseteq W$;
- W_i: tác vụ thứ i ;
- r^i: tập kỹ năng được yêu cầu để thực hiện tác vụ i;
- B_k, E_k: thời điểm bắt đầu và kết thúc tác vụ k ;
- $A_(u, v)^t$: có giá trị bằng 1 cho biết tài nguyên v; đang thực hiện tác vụ u tại thời điểm t và ngược lại;
- h_i: mức kỹ năng của kỹ năng (skill) i;
- g_i: loại kỹ năng i; m: makespan của cá thể (lịch biểu);
- P: lịch biểu thỏa mãn các ràng buộc của bài toán;
- $P_(all)$: tập mọi lịch biểu;
- $f(P)$: hàm mục tiêu, trả về makespan của lịch biểu P;
- z: số lượng tài nguyên có thể sử dụng của dự án;
- n: số lượng tác vụ cần thực hiện của dự án;

Lời giải của bài toán MS-RCPSP làm tìm ra một lịch biểu P để:

$$f(P) \rightarrow min \tag{1}$$

Một lịch biểu đúng cần đáp ứng các ràng buộc sau:

$$S^k \neq \oslash \quad \forall L_k \in L \tag{2}$$

$$t_j \geq 0 \quad \forall W_j \in W \tag{3}$$

$$E_j \geq 0 \quad \forall W_j \in W \tag{4}$$

$$E_i \leq E_j - t_j \quad \forall W_j \in W, j \neq 1, W_i \in C_j \tag{5}$$

$$\forall W_i \in W^k \quad \exists S_q \in S^k : g_(S_q) = g_(r^i) \; v \; h_(S_q) \geq h_(r^i) \tag{6}$$

$$\forall L_k \in L, \forall q \in m : \sum_{i=1}^{n} A_{i,k}^q \leq 1 \tag{7}$$

$$\forall W_j \in W \exists! q \in [0{-}m], !L_k \in L : A_{j,k}^q = 1; với \; A_{j,k}^q \in \{0; 1\} \tag{8}$$

Các ràng buộc về kỹ năng của tài nguyên được yêu cầu để thực hiện một tác vụ được mô tả dưới dạng một ma trận như ví dụ được minh họa trong Hình 1 dưới đây. Trong Hình 1 , ký hiệu: $S_(i.j)$ nghĩa là kỹ năng S_i và mức kỹ năng j.

IV. THUẬT TOÁN ĐỀ XUẤT

Thuật toán 1: fRotate.

```
Dữ liệu vào: currentBest (các thể tốt nhất sau mỗi thế hệ)..
Dữ liệu ra: lịch biểu sau khi được hiệu chỉnh.
begin
    makespan = f(currentBest);
    newbest = currentBest;
    L_b ← lastResource (newbest); //tài nguyên hoàn thành sau cùng
    W_b ← {các task vụ được thực hiện bởi L_b};
    for i = size(W_b^R) downto 1 do
        freepos ← { Vị trí tài nguyên rỗi };
        if freepos > 0 then
            newbest = { ;
              W_i thực hiện tại vị trí freepos;
              Tính toán lại vị trí thực hiện của các task khác;
            }
            newmakespan = f(newbest); // tính toán lại makespan
            if newmakespan < makespan then
                makespan = newmakespan;
                return newbest; // dịch chuyển thành công, trả về kết quả và dừng hàm
            end
        end
    end
    return newbest;
end
```

Là bài toán thuộc lớp NP-Khó, để tìm lời giải cho MS-RCPSP, cần sử dụng các phương pháp tìm lời giải gần đúng. Ro-CS là lai giữa phương pháp gần đúng Cuckoo Search và kỹ thuật Rotate.

1. Thuật toán Cuckoo Search (CS)

Thuật toán Cuckoo Search(CS) [12, 13] được Yang và Deb đưa ra vào năm 2009, đến nay CS đã được nhiều nhà khoa học nghiên cứu và áp dụng cho các bài toán tối ưu khác nhau, trong đó có bài toán MS-RCPSP. Thuật toán này được phát triển dựa trên đặc tính sinh hoạt của loài chim Cuckoo và hành vi bay của một số loài chim và ruồi giấm - gọi là Lévy flight.

CS phù hợp với các bài toán tối ưu tính toán trên dữ liệu lớn bằng cách tạo ra sự cân bằng trong các kỹ thuật tìm kiếm nghiệm toàn cục và tìm kiếm cục bộ nhằm tạo ra các nghiệm tốt hơn ở thế hệ tiếp theo. Mục tiêu của thuật toán là sinh ra những cá thể mới có tiềm năng tốt hơn thay

Thuật toán 2: Ro-CS.

```
Dữ liệu vào: dataset, t_max : số thế hệ tiến hóa
Dữ liệu ra: cá thể tốt nhất và makespan
begin
    Load and Valid dataset;
    t = 0;
    P_all = Initial population;
    f = {Calculate the fitness, makespan, bestnest};
    while t < t_max do
        P_new = {Tạo cá thể mới bằng Lévy Flight} ( theo công thức 10);
        P_i = {Lấy ngẫu nhiên một cá thể từ P_all };
        if f(P_new) < f(Pi) then
            P_i = P_new;
        end
        P^pa = pa% tổ kém nhất từ P_all;
        for j = 1 to size(P^pa) do
            P_new = Tạo cá thể mới bằng Lévy Flight(theo công thức 9);
            if f(P_new) < f(P_j^pa) then
                P_j^pa = P_new;
            end
        end
        P_all = {Thay thế P^pa vào P_all};
        f = {Calculate the fitness, bestnest};
        bestnest = fRotate(bestnest);
        makespan = f(bestnest);
        t = t+1;
    end
    return {bestnest, makespan};
end
```

thế các cá thể (nghiệm) có chất lượng kém. Việc tạo ra cá thể mới áp dụng kỹ thuật Lévy Flight với các phép Tìm kiếm toàn cục (Global Search) và Tìm kiếm cục bộ (Local Search).

Thế hệ tiếp theo tạo ra bằng phép tìm kiếm cục bộ, áp dụng công thức 9:

$$x_i^{t+1} = x_i^t + \alpha s \bigotimes H(pa - e) \bigotimes (x_j^t - x_k^t) \qquad (9)$$

Thế hệ tiếp theo tạo ra bằng phép tìm kiếm toàn cục, áp dụng công thức 10:

$$x_i^{t+1} = x_i^t + \alpha L(s, \lambda) \qquad (10)$$

với:

$$L(s, \lambda) = \frac{\lambda \Gamma(\lambda) sin(\pi\lambda/2)}{\pi} \frac{1}{s^{1+\lambda}}, s >> s_0 > 0 \qquad (11)$$

trong đó:

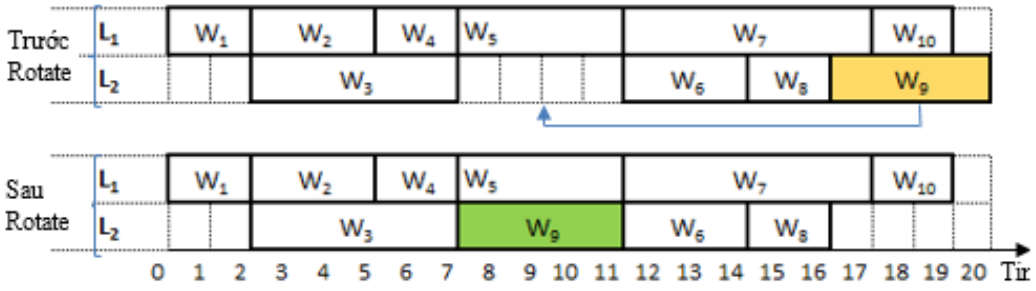
- λ: là một hằng số, $0 < \lambda \leq 2$
- $\Gamma(\lambda)$: là hàm trả lại giá trị $(\lambda$-1)!

Thuật toán CS thực hiện bằng cách tiến hóa quần thể qua nhiều thế hệ khác nhau. Ở mỗi thế hệ, CS thực hiện tạo một cá thể mới bằng kỹ thuật tìm kiếm toàn cục (Glob-al Search). Sau khi tính toán, tiến hóa, thuật toán sẽ thay thế một số cá thể kém nhất bằng các cá thể mới, các cá thể

✓ Có thể thiết lập
× Không thể thiết lập

Tài nguyên		Tác vụ			
		$S_{2.2}$	$S_{3.1}$	$S_{2.2}$	$S_{1.1}$
		W_1	W_2	W_3	W_4
$S_{1.3}, S_{2.2}$	L_1	✓	×	✓	✓
$S_{2.1}, S_{3.2}$	L_2	×	✓	×	×
$S_{1.2}, S_{2.1}$	L_3	×	×	×	✓
$S_{2.2}, S_{3.3}$	L_4	✓	✓	✓	×

Hình 1. Tài nguyên thực hiện tác vụ.



Hình 2. Biểu đồ Gantt khi áp dụng phương pháp Rotate.

mới tạo ra bằng kỹ thuật tìm kiếm cục bộ (Local Search). Kết quả của thuật toán là một cá thể (lịch biểu) với thời gian thực hiện tối thiểu.

2. Kỹ thuật Rotate

Rotate là kỹ thuật xoay vòng nhằm hạn chế tối đa khoảng thời gian rỗi của tài nguyên, dẫn đến giảm tổng thời gian thực hiện dự án. Các bước thực hiện cụ thể như sau:

- **Bước 1**: Tìm tài nguyên sử dụng nhiều nhất trong Cbest, gọi là Lb. Đây là tài nguyên kết thúc công việc sau cùng trong các tài nguyên sử dụng để thực hiện dự án.
- **Bước 2**: Duyệt các tác vụ được thực hiện bởi Lb theo thứ tự ngược với thứ tự thực hiện từ phải qua trái trên tài nguyên, tức là tác vụ nào thực hiện sau sẽ xem xét trước.
  W_b^R: tập tác vụ vụ đang được thực hiện bởi L_b
- **Bước 3**: Với mỗi tác vụ $W_i \in W_b^R (i = sizeof(W_b^R) downto 1)$, thực hiện:
  - *Bước 3.1*: Tìm khoảng thời gian rỗi trên tài nguyên
  - *Bước 3.2*: Chuyển thứ tự thực hiện của tác vụ hiện tại về vị trí của tài nguyên rỗi nếu không vi phạm ràng buộc về thứ tự ưu tiên.
  - *Bước 3.3*: Điều chỉnh thời gian bắt đầu và kết thúc thực hiện của các tác vụ có vị trí thực hiện sau vị trí mới của tác vụ hiện tại
  - *Bước 3.4*: Tính toán lại makespan của dự án, nếu tạo ra cá thể tốt hơn thì dừng thuật toán.

Về bản chất, phương pháp Rotate là cách tái sắp xếp thứ tự thực hiện các tác vụ trên tài nguyên hoàn thành công việc sau cùng mà không vi phạm các ràng buộc về thứ tự ưu tiên giữa các tác vụ, đồng thời hạn chế tối đa thời gian tài nguyên rỗi. Việc này giúp giảm tổng thời gian thực hiện dự án đáng kể.

Hình 2 là ví dụ về một dự án với 10 tác vụ và 2 tài nguyên. Trước khi áp dụng kỹ thuật Rotate, thời gian thực hiện tác vụ là 20, sau khi áp dụng kỹ thuật Roate, thời gian thực hiện dự án giảm xuống còn 10.

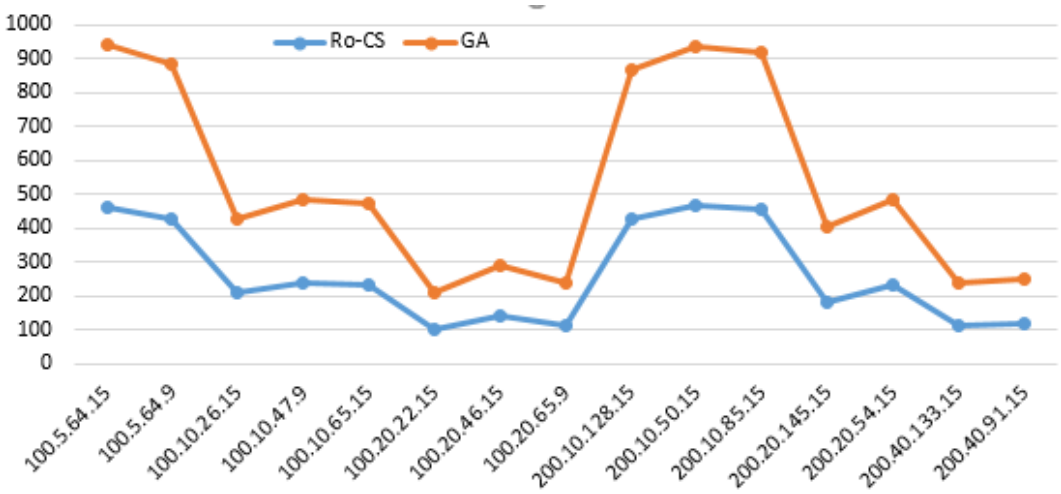
Phương pháp Rotate được trình bày chi tiết trong **Thuật toán 1**.

với:

- f: hàm tính makespan của một lịch biểu
- size: hàm tính số phần tử của một tập/mảng
- lastResource: hàm trả về tài nguyên kết thúc thực hiện sau cùng

Bảng I
KẾT QUẢ THỰC NGHIỆM TRÊN BỘ DỮ LIỆU IMOPSE

Dataset	GA			Ro-CS		
	Best	**Avg**	**Std**	**Best**	**Avg**	**Std**
100.5.64.15	478	487	8.2	462	467	2.5
100.5.64.9	453	457	3.2	430	434	2.1
100.10.26.15	221	224	2.1	209	213	1.9
100.10.47.9	247	253	5.9	237	240	1.3
100.10.65.15	240	244	3.2	233	236	2.2
100.20.22.15	109	115	5.2	101	109	5.2
100.20.46.15	145	150	4.1	143	152	7.2
100.20.65.9	124	131	6.5	115	127	6.3
200.10.128.15	440	446	5.7	428	435	4.1
200.10.50.15	468	471	2.9	465	467	0.5
200.10.85.15	465	469	4	456	458	0.3
200.20.145.15	222	229	6.1	183	192	3.8
200.20.54.15	249	255	5.1	235	240	2.7
200.40.133.15	126	134	7.4	115	127	9
200.40.91.15	133	139	5.8	119	128	7.1



Hình 3. So sánh kết quả thực nghiệm của giá trị BEST.

3. Thuật toán Ro-CS

Áp dụng kỹ thuật Rotate vào thuật toán CS ta được thuật toán 2 Ro-CS dưới đây.

với:

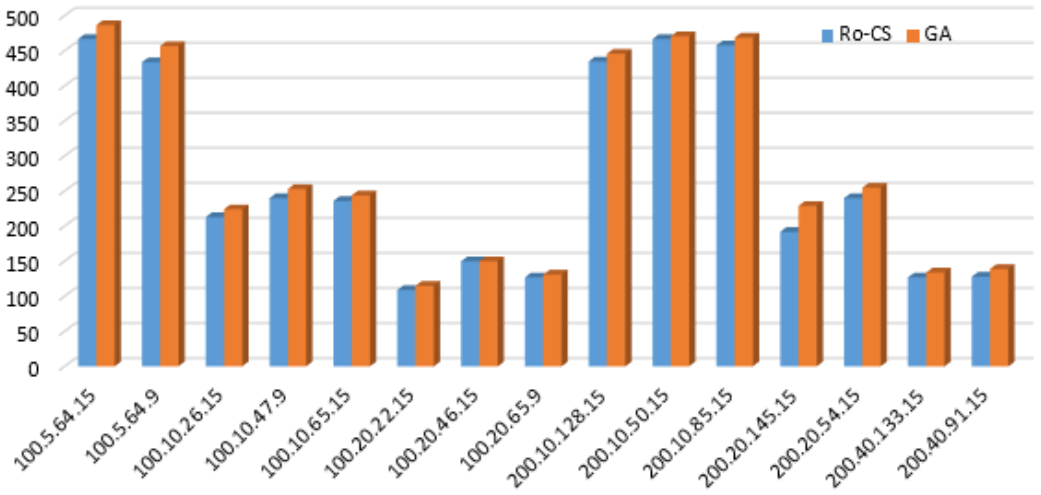
- f: hàm tính makespan của một lịch biểu
- fRotate: là hàm áp dụng kỹ thuật Rotate.

V. KẾT QUẢ THỰC NGHIỆM

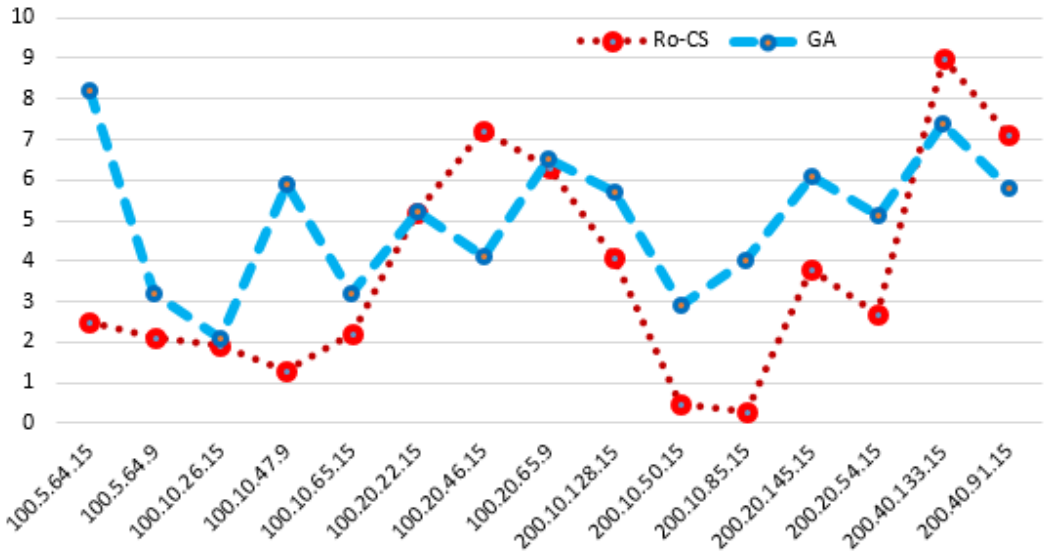
Để kiểm chứng thuật toán đề xuất Ro-CS, bài báo tiến hành thực nghiệm trên bộ dữ liệu iMOPSE [5,6,14] (là bộ dữ liệu chuẩn quốc tế được phát triển riêng cho bài toán MS-RCPSP. Bộ dữ liệu iMOPSE được thể hiện như trong Bảng II dưới đây.

Trong đó:

- Tasks: là số tác vụ cần thực hiện của dự án.
- Resources: là số tài nguyên có thể sử dụng để thực hiện các tác vụ, các tài nguyên này có khả năng tái sử dụng, nghĩa là sau khi hoàn thành một tác vụ, có thể sử dụng tài nguyên đó để thực hiện tác vụ khác.
- Precedence Relations: là tổng số ràng buộc ưu tiên giữa các tác vụ trong bài toán, ràng buộc ưu tiên chỉ ra thứ tự thực hiện của các tác vụ.
- Skills: số kỹ năng của tài nguyên thực hiện, mỗi tài



Hình 4. So sánh kết quả thực nghiệm của giá trị AVG.



Hình 5. So sánh kết quả thực nghiệm của giá trị STD.

nguyên có thể có nhiều kỹ năng khác nhau, mỗi kỹ năng có một mức kỹ năng cụ thể.

1. Tham số thực nghiệm

Thực nghiệm Ro-CS được cài đặt với các tham số đầu vào như sau:

- Dữ liệu: 15 file dữ liệu trong bộ dữ liệu iMOPSE như trong Bảng I.
- Khởi tạo quần thể ban đầu với 100 cá thể;
- Số thế hệ tiến hóa: 50.000;
- Số lần chạy thực nghiệm trên mỗi bộ dữ liệu: 25;
- Môi trường thực nghiệm: Thực nghiệm được tiến hành trên môi trường Matlab 2014
- Máy tính thực nghiệm: CPU Intel Core i5, tốc độ 2.2 GHz, RAM 6GB, hệ điều hành Windows 10.

2. Tham số thực nghiệm

Kết quả thực nghiệm với bộ dữ liệu iMOPSE [5,6] được trình bày trong Bảng I.

Trong Bảng I, dữ liệu cột GA được lấy từ phần thực nghiệm của nhóm tác giải Myszkowsky với công cụ GARunner [14] do chính nhóm tác giải công bố.

Các thực nghiệm được triển khai trên 02 thuật toán GA và Ro-CS sử dụng bộ dữ liệu iMOPSE. Bài báo đã tiến hành thực nghiệm các thuật toán một cách độc lập với các tham số đầu vào cài đặt như trong mục v. Kết quả thực

Bảng II
BỘ DỮ LIỆU IMOPSE

Name	Tasks	Resources	Precedence Relations	Skills
100.5.64.15	100	5	64	15
100.5.64.9	100	5	64	9
100.10.26.15	100	10	26	15
100.10.47.9	100	10	47	9
100.10.65.15	100	10	65	15
100.20.22.15	100	20	22	15
100.20.46.15	100	20	46	15
100.20.65.9	100	20	65	9
200.10.128.15	200	10	128	15
200.10.50.15	200	10	50	15
200.10.85.15	200	10	85	15
200.20.145.15	200	20	145	15
200.20.54.15	200	20	54	9
200.40.133.15	200	40	133	15
200.40.91.15	200	40	91	9

nghiệm được trình bày trong Bảng II chỉ ra thuật toán Ro-CS có chất lượng lời giải tốt hơn thuật toán GA trong mọi trường hợp. Cụ thể:

- Với giá trị tốt nhất (BEST) Ro-CS tốt hơn GA từ 0.6% đến 17.6%. So sánh chi tiết được thể hiện trong Hình 3.
- Với giá trị trung bình (AVG) Ro-CS tốt hơn GA từ 0% đến 16.2%. So sánh chi tiết được thể hiện trong Hình 4.
- Tổng độ lệch chuẩn của GA 75.4, của Ro-CS 56.2, điều này cho thấy Ro-CS hoạt động ổn định hơn. So sánh chi tiết được thể hiện trong Hình 5.

Như vậy, thuật toán mới Ro-CS với việc bổ sung kỹ thuật Rotate mang lại quả tốt hơn so với GA.

VI. KẾT LUẬN

Bài báo này đã trình bày phát biểu toán học cụ thể về bài toán MS-RCPSP thông qua các ràng buộc, đây là bài toán có nhiều ứng dụng trong khoa học và thực tế. Để tìm lời giải cho bài toán MS-RCPSP, bài báo đã đề xuất một thuật toán mới là Ro-CS, đây là thuật toán lai giữa thuật toán di truyền và kỹ thuật Rotate nhằm mở rộng không gian tìm kiếm sau mỗi thế hệ tiến hóa. Để kiểm chứng tính hiệu quả của thuật toán đề xuất, bài báo đã tiến hành thực nghiệm trên bộ dữ liệu iMOPSE (bộ dữ liệu chuẩn được phát triển để tiến hành các thực nghiệm riêng cho bài toán MS-RCPSP). Kết quả thực nghiệm đã được tổng hợp, so sánh với GA trên các giá trị tốt nhất (BEST), giá trị bình quân (AVG) và giá trị độ lệch chuẩn (STD). Thông qua việc so sánh cho thấy, Ro-CS mang lại hiệu quả tốt hơn GA với cùng bài toán và cùng bộ dữ liệu thực nghiệm, cụ thể: giá trị BEST tốt hơn từ 0.6% đến 17.6%, giá trị AVG tốt hơn đến 16.2 và tổng STD tốt hơn 19.2.

Ngoài việc sử dụng tìm lời giải cho bài toán MS-RCPSP, thuật toán Ro-CS còn có khả năng áp dụng tốt cho các bài toán lập lịch khác. Trong thời gian tới, tác giả sẽ tiếp tục nghiên cứu và cải tiến thuật toán dựa trên các phương pháp gần đúng khác, sử dụng bước di chuyển ngẫu nhiên dựa trên xác suất Gauss, Cauchy, ... nhằm nâng cao chất lượng lời giải.

TÀI LIỆU THAM KHẢO

[1] A. Christian, S. Demassey, E. Néron, “Resource Constrained Project Scheduling: Models, Algorithms, Extensions and Applications”, *ISBN 978-1-84821-034-9*, 2008.

[2] R. Klein: “Scheduling of Resource-Constrained Projects”, *Springer Science & Business Media*, Vol. 10, 2012.

[3] Huu Dang Quoc, Loc Nguyen The, Cuong Nguyen Doan, "Naixue Xiong: Effective Evolutionary Algorithm for Solving the Real-Resource-Constrained Scheduling Problem," *Journal of Advanced Transportation*, vol. 2020, Article ID 8897710, 11 pages, 2020

[4] P.B. Myszkowski, M. Skowroński, L. Olech, K. Oślizło, "Hybrid Ant Colony Optimization in solving Multi–Skill Resource–Constrained Project Scheduling Problem", *Soft Computing Journal*, Volume 19, Issue 12, pp.3599–3619, 2015.

[5] P.B. Myszkowski, M. Laszczyk, I. Nikulin, M. Skowro, “iMOPSE: a library for bicriteria optimization in Multi-Skill Resource-Constrained Project Scheduling Problem”, *Soft Computing Journal*, 23: 32397, 2019.

[6] P.B. Myszkowski, M.E. Skowronski, K.Sikora, “A new benchmark dataset for Multi-Skill Resource-Constrained Project Scheduling Problem”, *Computer Science and Information Systems*, ACSIS, Vol. 5, pp. 129–138, 2015. DOI: 10.15439/2015F273.

[7] A.H. Hosseinian, V. Baradaran, "An Evolutionary Algorithm Based on a Hybrid Multi-Attribute Decision Making Method for the Multi-Mode Multi-Skilled Resource-Constrained Project Scheduling Problem.", *Journal of Optimization in Industrial Engineering*, 12.2, pp. 155-178,2019.

[8] A.H. Hosseinian, V. Baradaran, "Detecting communities of workforces for the multi-skill resource-constrained project scheduling problem: A dandelion solution approach.", *Journal of Industrial and Systems Engineering*, pp. 72-99, 12.2019.

[9] S. Javanmard, B. Afshar-Nadjafi, S.T. Niaki, "Preemptive multi-skilled resource investment project scheduling problem: Mathematical modeling and solution approaches.", *Computers & Chemical Engineering*, 96, pp. 55-68, 2017.

[10] H. Najafzad, H. Davari-Ardakani, R. Nemati-Lafmejani, "Multi-skill project scheduling problem under time-of-use electricity tariffs and shift differential payments.", *Energy Journal*, vol. 168, pp. 619-636, Elsevier,2019.

[11] Huu Dang Quoc, Loc Nguyen The, Cuong Nguyen Doan, Toan Phan Thanh, Naixue Xiong, “Intelligent Differential Evolution Scheme for Network Resources in IoT”, *Scientific Programming (IF:1.28, Q3)*, Volume 2020, Article ID 8860384 | 12, 2020. DOI: 10.1155/2020/8860384

[12] X.S. Yang, “Nature-Inspired Metaheuristic Algorithms”, *Luniver Press*, ISBN-13: 978-1-905986-28-6, 2010.

[13] X.S. Yang, S. Deb, "Cuckoo search via Lévy flights", *Proc. of World Congress on Nature & Biologically Inspired Computing (NaBIC 2009)*, USA, pp. 210-214, 2009.
[14] Bộ dữ liệu iMOPSE và công cụ GARunner: http://imopse.ii.pwr.wroc.pl/download.html

SƠ LƯỢC VỀ TÁC GIẢ

Nguyễn Thế Lộc

Tốt nghiệp đại học khoa Toán Tin, đại học Sư phạm Hà Nội năm 1993, tốt nghiệp thạc sĩ CNTT tại trường đại học Bách khoa Hà Nội, nhận bằng tiến sỹ tại Viện Nghiên cứu Khoa học Công nghệ Nhật Bản JAIST năm 2007. Hiện đang công tác tại trường Đại học Sư phạm Hà Nội.
Lĩnh vực nghiên cứu: mạng máy tính và truyền thông, xử lý song song và phân tán.
Điện thoại : 0988.765.837.
E-mail: locnt@hnue.edu.vn

Đặng Quốc Hữu

Tốt nghiệp đại học và thạc sĩ khoa CNTT, Đại học Quốc Gia Hà Nội năm 2006 và 2015, nhận bằng Tiến sĩ tại Viện Khoa học và Công nghệ Quân sự năm 2022. Hiện đang công tác tại đại học Thương Mại.
Lĩnh vực nghiên cứu: tính toán tối ưu, tính toán tiến hóa, xử lý song song và phân tán.
Điện thoại : 0988710727.
E-mail: huudq@tmu.edu.vn

Một mô hình tìm kiếm ảnh dựa trên cấu trúc R-Tree kết hợp KD-Tree Random Forest

Lê Mạnh Thạnh[1], Lê Thị Vĩnh Thanh[1,4], Lương Thị Thanh Xuân[1,5], Nguyễn Thị Định[1,3], Văn Thế Thành[2]
[1] Trường Đại học Khoa học, Đại học Huế
[2] Trường Đại học Sư phạm TP. HCM
[3] Khoa Công nghệ Thông tin, Trường ĐH Công nghiệp Thực phẩm TP.HCM
[4] Trường Đại học Bà Rịa Vũng Tàu
[5] Trường THCS-THPT Nguyễn Văn Khải, Đồng Tháp

Tác giả liên hệ: Lê Mạnh Thạnh, lmthanh@hueuni.edu.vn
Ngày nhận bài: 15/04/2022, ngày sửa chữa: 01/06/2022, ngày duyệt đăng: 30/06/2022
Định danh DOI: 10.32913/mic-ict-research-vn.v2022.n1.1053

Tóm tắt: Nâng cao hiệu suất truy vấn ảnh là vấn đề được quan tâm trong lĩnh vực thị giác máy tính. Trong bài báo này, một phương pháp kết hợp cấu trúc R-Tree và KD-Tree Random Forest được thực hiện nhằm nâng cao hiệu suất phân lớp và tìm kiếm ảnh. Với mỗi ảnh đầu vào được phân lớp bằng KD-Tree Random Forest; sau đó, kỹ thuật gom cụm trên R-Tree được thực hiện và tìm kiếm tập ảnh tương tự. Để thực hiện bài toán này, phương pháp xây dựng KD-Tree Random Forest kết hợp với cấu trúc R-Tree cùng với các thuật toán xây dựng rừng ngẫu nhiên, phân lớp và tìm kiếm ảnh được đề xuất. Trên cơ sở đó, một mô hình tìm kiếm ảnh dựa trên sự kết hợp KD-Tree Random Forest và R-Tree được đề xuất. Thực nghiệm được thực hiện trên các bộ ảnh COREL và Caltech101 nhằm để đánh giá tính khả thi, hiệu quả của mô hình; đồng thời so sánh với các công trình khác thực nghiệm trên cùng bộ dữ liệu và so sánh với kết quả truy vấn ảnh khi sử dụng một cấu trúc riêng lẻ. Theo kết quả thực nghiệm cho thấy phương pháp đề xuất của chúng tôi là hiệu quả và có thể áp dụng được cho các hệ truy vấn ảnh với nhiều bộ dữ liệu khác nhau.

Từ khóa: *KD-Tree, Random Forest, R-Tree, gom cụm, phân lớp, tìm kiếm ảnh*

Title: **A Model of Combining R-Tree and KD-Tree Random Forest for Content-based Image Retrieval**

Abstract: Improving image query performance is a matter of interest in the field of computer vision. In this paper, a method combining R-Tree and KD-Tree Random Forest structures is implemented to enhance the performance of im-age classification and retrieval. For each input image is classified by KD-Tree Random Forest; then, the clustering technique on R-Tree is performed and a similar image set is retrieved. To implement this problem, the method of construction KD-Tree Random Forest combined with the R-Tree structure with the algorithms for building a random forest, classification, and image retrieval. On that basis, an image retrieval model based on the combination of KD-Tree Random Forest and R-Tree is proposed. Experiments were performed on the COREL and Caltech101 image sets to evaluate the feasibility and effectiveness of the model; at the same time experiment is compared with other works on the same data set and compared with image query results when using an individual structure. The experimental results show that our proposed method is effective and can be applied to image retrieval systems for many sets of images in many different fields

Keywords: *KD-Tree, R-Tree, Random Forest, Clustering, Image Classification, Image Retrieval*

I. GIỚI THIỆU

Phương pháp kết hợp nhiều kỹ thuật học máy nhằm nâng cao hiệu suất tìm kiếm ảnh là vấn đề cần được quan tâm trong lĩnh vực truy vấn ảnh và thị giác máy tính. Bên cạnh đó, dữ liệu đa phương tiện tăng nhanh theo thời gian về số lượng và thể loại là thách thức cho vấn đề lưu trữ, hiệu suất và thời gian tìm kiếm. Vì vậy, việc tích hợp một số kỹ thuật và phương pháp cho một hệ truy vấn ảnh nhằm nâng cao hiệu suất, ổn định thời gian tìm kiếm cũng như tối ưu hóa không gian lưu trữ là cần thiết [1].

Không gian lưu trữ ảnh số là vấn đề cần được quan tâm và tối ưu nhằm đáp ứng thực trạng gia tăng ảnh số trong bối cảnh hiện nay. Để giải quyết vấn đề này, một số cấu trúc dữ liệu như S-Tree [2], R-Tree [3], v.v. nhằm tối ưu không gian lưu trữ ảnh số đã được thực hiện đã mang lại những kết quả khả quan. Trong bài báo này, một cấu trúc

R-Tree được ứng dụng cho quá trình gom cụm dữ liệu thành các cụm tương đồng và áp dụng cho bài toán tìm kiếm ảnh được đề xuất. Bên cạnh đó, phân lớp dữ liệu hình ảnh là giai đoạn đầu, đồng thời góp phần nâng cao hiệu suất cho bài toán tìm kiếm ảnh tương tự. Hiện nay, có nhiều kỹ thuật phân lớp hình ảnh được đánh giá với hiệu suất phân lớp khá cao như CNN, DNN, k-NN, v.v. Tuy nhiên, hướng tiếp cận mô hình phân lớp bằng KD-Tree là một cải tiến trên KD-Tree đã mang lại hiệu suất phân lớp ổn định, đồng thời áp dụng cho bài toán tìm kiếm ảnh hiệu quả [16, 26]. Do đó, trong bài báo này một phương pháp phân lớp ảnh số bằng KD-Tree Random Forest được thực hiện dựa trên cơ sở xây dựng nhiều cấu trúc KD-Tree đồng thời để hình thành quá trình phân lớp ảnh bằng rừng ngẫu nhiên. Bên cạnh đó, gom cụm trên R-Tree là một phương pháp nâng cao hiệu quả tìm kiếm ảnh [15]. Vì vậy, mục tiêu của bài báo này là thực hiện sự kết hợp những điểm mạnh của cấu trúc KD-Tree và R-Tree nhằm nâng cao hiệu suất tìm kiếm ảnh cần được thực hiện.

Đóng góp của bài báo gồm: (1) Đề xuất phương pháp kết hợp cấu trúc R-Tree và KD-Tree Random Forest cho bài toán tìm kiếm ảnh; (2) Đề xuất mô hình tìm kiếm ảnh dựa trên kỹ thuật phân lớp KD-Tree Random Forest và gom cụm R-Tree; (3) Đánh giá, so sánh hiệu suất truy vấn ảnh trên các phương pháp khác cùng bộ ảnh thực nghiệm COREL [4], Clatech101 [5].

Cấu trúc của bài báo gồm: phần II trình bày các công trình liên quan về gom cụm trên R-Tree, phân lớp ảnh bằng KD-Tree và rừng ngẫu nhiên; phần III đề xuất mô hình tìm kiếm ảnh dựa trên sự kết hợp của cấu trúc KD-Tree Random Forest và R-Tree; phần IV trình bày cấu trúc và phương pháp xây dựng KD-Tree Random Forest, xây dựng và tìm kiếm trên R-Tree cho bài toán tìm kiếm ảnh; phần V thực nghiệm và đánh giá kết quả trên bộ ảnh COREL, Calterch101; kết luận và hướng phát triển được trình bày ở phần VI.

II. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Phân lớp và gom cụm luôn được sử dụng cho bài toán tìm kiếm ảnh vì sự ảnh hưởng kết quả của các quá trình này đến kết quả tìm kiếm ảnh là khá lớn. Hiện nay, có nhiều công trình tìm kiếm ảnh đã sử dụng các kỹ thuật phân lớp và gom cụm như CNN, DNN, k-NN, k-Means, v.v kết hợp với cấu trúc R-Tree, KD-Tree đã mang lại hiệu suất cao tiêu biểu là:

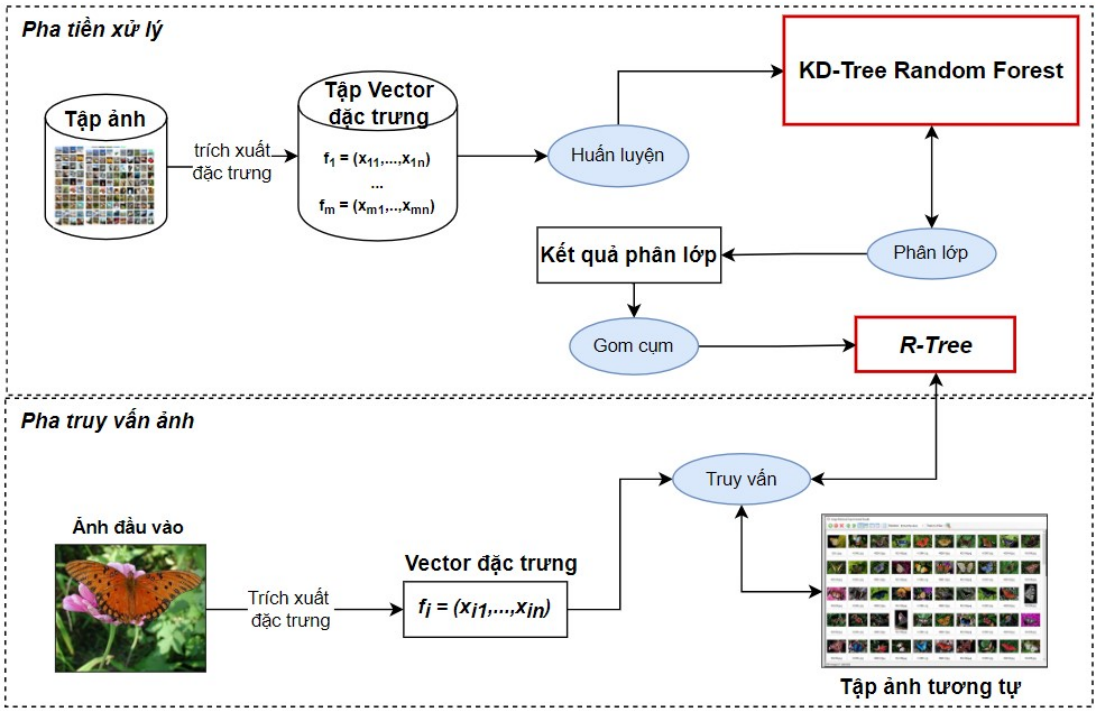
Yuqian Zhang và cộng sự (2016) [6] đã sử dụng phương pháp phân chia vùng ảnh để nhận diện khuôn mặt sử dụng cấu trúc KD-Tree. Tại pha huấn luyện, mỗi ảnh được chia thành nhiều vùng; cấu trúc KD-Tree được xây dựng dựa trên tập ảnh phân vùng nhằm phân lớp và lưu trữ tập ảnh huấn luyện. Tại pha kiểm thử, mỗi ảnh được chia thành nhiều vùng và cấu trúc KD-Tree được sử dụng trong quá trình tìm kiếm theo k láng giềng gần nhất bằng cách đối sánh các vùng ảnh tìm kiếm với ảnh gốc bằng thuật toán k-NN. Công trình này được đánh giá là sự kết hợp giữa cấu trúc KD-Tree và thuật toán k-NN đã mang lại hiệu suất cao hơn so với chỉ dùng một kỹ thuật đơn lẻ.

Fatin Zaklouta và cộng sự (2011) [17], đã thực hiện đánh giá hiệu suất phân lớp ảnh đèn giao thông bằng KD-Tree và rừng ngẫu nhiên để phân loại biển báo giao thông bằng cách sử dụng đặc trưng biểu đồ kích thước khác nhau bằng đặc trưng HOG (Histogram of Oriented Gradients) và phép biến đổi khoảng cách (Distance Transforms). Các tác giả sử dụng tập dữ liệu ảnh biển báo giao thông của Đức gồm 43 lớp và hơn 50.000 hình ảnh. Trong bài báo này, sự kết hợp của bộ phân loại KD-Tree với đặc trưng HOG và Distance Transforms thì tỷ lệ phân loại lên tới 0.97 và 0.8180. Kết quả này được đánh giá là khá cao và có khả năng mở rộng cho các hệ phân loại ảnh khác.

Weishi Man và cộng sự (2018) [18] đã thực hiện một cải tiến trong thuật toán tách nút rừng ngẫu nhiên để nâng cao hiệu suất phân lớp hình ảnh bằng cách kết hợp mô hình phân tách thuộc tính rừng ngẫu nhiên ID3 và CART. Thứ nhất, tính hợp lệ của mô hình Kim tự tháp không gian được xác minh dựa trên mô hình túi từ (Bag of Words). Thứ hai, thuật toán tách nút của rừng ngẫu nhiên được cải tiến. Cuối cùng, hiệu quả và độ chính xác của thuật toán đề xuất được minh chứng thông qua so sánh thực nghiệm trên bộ dữ liệu hình ảnh Caltech101 với hiệu suất cao, điều này cho thấy phân lớp ảnh bằng rừng ngẫu nhiên là một phương pháp đúng đắn, hiệu quả.

Vanitha J., Senthilmurugan M. (2018) [7] đã đề xuất một cấu trúc cây chỉ mục SR-Tree (Sphere Rectangle Tree) là sự kết hợp giữa cây SS-Tree và cây R-Tree để lưu trữ dữ liệu hình ảnh dạng chỉ mục. Một hệ thống truy vấn ảnh dựa trên nội dung sử dụng cấu trúc cây SR-Tree được thực nghiệm trên các bộ ảnh với kết quả được đánh giá là khả quan. Cùng thời điểm này, Arpitha Y.C và cộng sự (2018) [8] đã đề xuất một cách tiếp cận bài toán xử lý hình ảnh đa chiều bằng cách sử dụng cấu trúc R-Tree, trong đó các đối tượng hình ảnh đa chiều được lưu trữ theo phương pháp chỉ mục. Trên cơ sở này, một phương pháp tìm kiếm các đối tượng trong một vùng cho trước và tìm kiếm láng giềng gần nhất dựa trên cây R-Tree được thực hiện đã mang lại kết quả tốt.

Bên cạnh đó, Xinlu Wang và cộng sự (2019) [9] đã đề xuất một phương pháp truy vấn động trên cấu trúc R-Tree, các tâm cụm động được tối ưu hóa dựa trên cấu trúc R-Tree. Bằng cách chọn một tâm cụm tối ưu hóa, phương pháp này tổ chức lưu dữ liệu cùng một không gian con thành cùng cây con và xây dựng một cây chỉ mục R-Tree. Kết quả thực nghiệm cho thấy phương pháp được đề xuất cải thiện tính



Hình 1. Mô hình truy vấn ảnh dựa trên R-Tree kết hợp KD-Tree Random Forest

ổn định của hệ thống và hiệu quả cao, đồng thời ứng dụng cho hệ thống tra cứu thông tin bệnh viện.

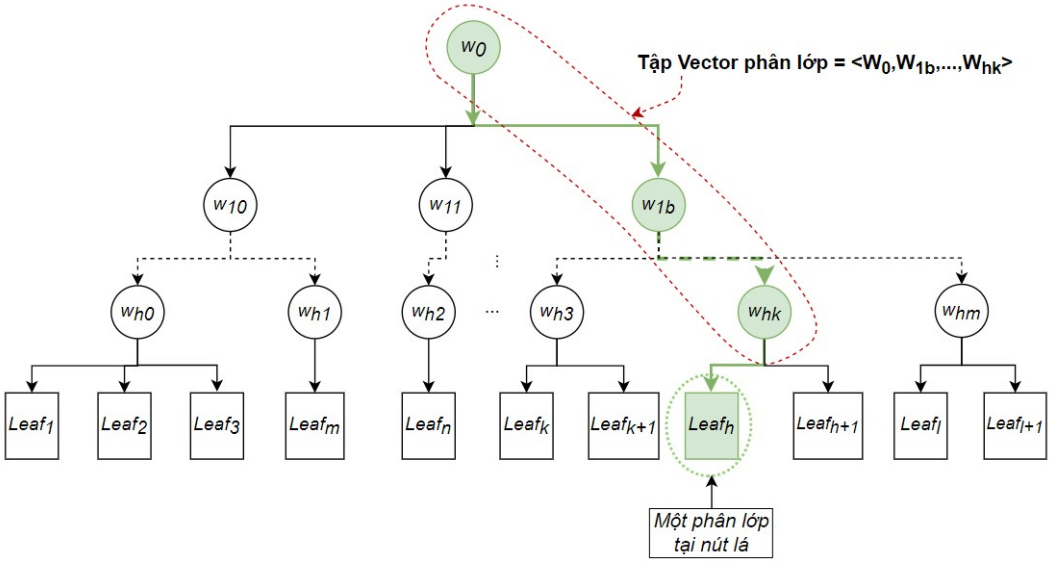
Trên cơ sở các công trình nghiên cứu liên quan, nhóm tác giả Nguyễn Thị Định và cộng sự [16, 26] đã thực hiện một phương pháp phân lớp dữ liệu bằng KD-Tree và áp dụng cho bài toán tìm kiếm ảnh tương tự. Trong những công trình này, nhóm tác giả đã thực nghiệm trên các bộ ảnh COREL, Wang, image CLEF với hiệu suất khá cao. Quá trình xây dựng KD-Tree theo hướng tiếp cận phân lớp dữ liệu để hình thành các phân lớp tại nút lá. Đồng thời nhóm tác giả Lê Thị Vĩnh Thanh và cộng sự [15] đã thực hiện một phương pháp gom cụm trên R-Tree cho bài toán tìm kiếm ảnh đã mang lại hiệu suất cao khi so sánh với các công trình khác cùng lĩnh vực. Nhóm tác giả đã thực hiện cải tiến cấu trúc R-Tree nhằm nâng cao khả năng mở rộng không gian lưu trữ, tối ưu thời gian tìm kiếm.

Từ những công trình đã khảo sát cho thấy các loại đặc trưng cơ bản như màu sắc, hình dạng, kết cấu, đường viền được sử dụng khá phổ biến và mang lại hiệu suất cao. Tuy nhiên các công trình này chưa quan tâm đến vấn đề tối ưu khả lưu trữ. Vì vậy, trong bài báo này một số đặc trưng sau khi trích xuất véc-tơ được phân lớp bằng KD-Tree Random Forest, sau đó gom cụm trên R-Tree để ổn định thời gian tìm kiếm và tăng khả năng lưu trữ đáp ứng nhu cầu ngày càng gia tăng dữ liệu ảnh số. Phương pháp kết hợp giữa kỹ thuật phân lớp ảnh bằng rừng ngẫu nhiên trên cơ sở xây dựng nhiều KD-Tree đồng thời; sau đó tiếp tục gom cụm trên R-Tree đã mang lại hiệu suất tìm kiếm cao, thời gian tìm kiếm ổn định và áp dụng cho các bộ ảnh có tính chất khác nhau với hiệu quả cao.

III. MÔ HÌNH TÌM KIẾM ẢNH DỰA TRÊN R-TREE KẾT HỢP KD-TREE RANDOM FOREST

Để minh chứng cho cơ sở lý thuyết đề xuất, một mô hình tìm kiếm ảnh kết hợp cấu trúc R-Tree và KD-Tree minh họa bởi Hình 1. Trong đó, cấu trúc KD-Tree được sử dụng cho quá trình phân lớp hình ảnh, từ phân lớp để thực hiện gom cụm trên R-Tree để trích xuất tập ảnh tương tự tại nút lá và được kế thừa từ công trình [15, 16]. Để thực hiện mô hình này, các thuật toán đề xuất gồm: (1) thuật toán huấn luyện KD-Tree; (2) thuật toán xây dựng KD-Tree Random Forest; (3) thuật toán phân lớp ảnh bằng KD-Tree Random Forest; (4) thuật toán gom cụm bằng R-Tree; (5) thuật toán tìm kiếm ảnh tương tự trên R-Tree.

Mô hình truy vấn ảnh dựa trên cấu trúc R-Tree kết hợp KD-Tree Random Forest gồm pha tiền xử lý và pha truy vấn ảnh với các bước như sau: (1) Trích xuất đặc trưng cho tập dữ liệu thực nghiệm thành tập véc-tơ đặc trưng; (2) Xây dựng và huấn luyện cấu trúc KD-Tree Random Forest từ tập véc-tơ đặc trưng theo phương pháp Bagging; (3) Phân lớp ảnh bằng cấu trúc KD-Tree Random Forest theo cơ chế phiếu bầu để trích xuất tên phân lớp ảnh; (4) Gom cụm dữ liệu trên R-Tree đã xây dựng sau khi thực hiện phân lớp;



Hình 2. Cấu trúc KD-Tree phân lớp

(5) Trích xuất véc-tơ đặc trưng cho ảnh đầu vào; (6) Truy vấn trên R-Tree để trích xuất tập ảnh tương tự với ảnh đầu vào.

IV. CẤU TRÚC KD-TREE RANDOM FOREST VÀ R-TREE

Thuật toán 1: Huấn luyện KD-Tree

Input:Bộ Véc-tơ InitWeight, chiều cao h, b nhánh
Output: Trained-KD-Tree
Function: **TKDT** (InitWeight, h, b)
begin
    Init-KD-Tree = ∅;
    while $P_j < P_i$ **do**
        Xây dựng Init-KD-Tree với bộ Véc-tơ InitWeight, chiều cao h, b nhánh;
        Gán nhãn cho các nút lá trên Init-KD-Tree;
        Pi = Tính hiệu suất phân lớp trên Init-KD-Tree;
        Xác định đường đi sai của những véc-tơ f_k trên Init-KD-Tree;
        Xác định đường đi đúng của những véc-tơ f_k bị sai đường đi;
        Tìm vị trí $Node_i$ làm sai đường đi của vectơ f_k;
        TrainVéc-tơ = Điều chỉnh véc-tơ lưu tại $Node_i$ để f_k đúng đường đi;
        Tạo lại Trained KD-Tree theo bộ TrainVéc-tơ;
        Gán nhãn cho các nút lá trên Trained-KD-Tree;
        P_j= Tính hiệu suất phân lớp sau khi điều chỉnh véc-tơ tại Nodei;
    end
    Véc-tơ = TrainVéc-tơ; **return** Trained-KD-Tree;
end

Cấu trúc dữ liệu lưu trữ là đối tượng giúp tối ưu hóa không gian lưu trữ, giảm thời gian tìm kiếm và ảnh hưởng trực tiếp đến hiệu suất truy vấn ảnh. Trong bài báo này, một phương pháp kết hợp giữa hai cấu trúc R-Tree và KD-Tree Random Forest để thực hiện gom cụm và phân lớp cho bài toán tìm kiếm ảnh được thực hiện.

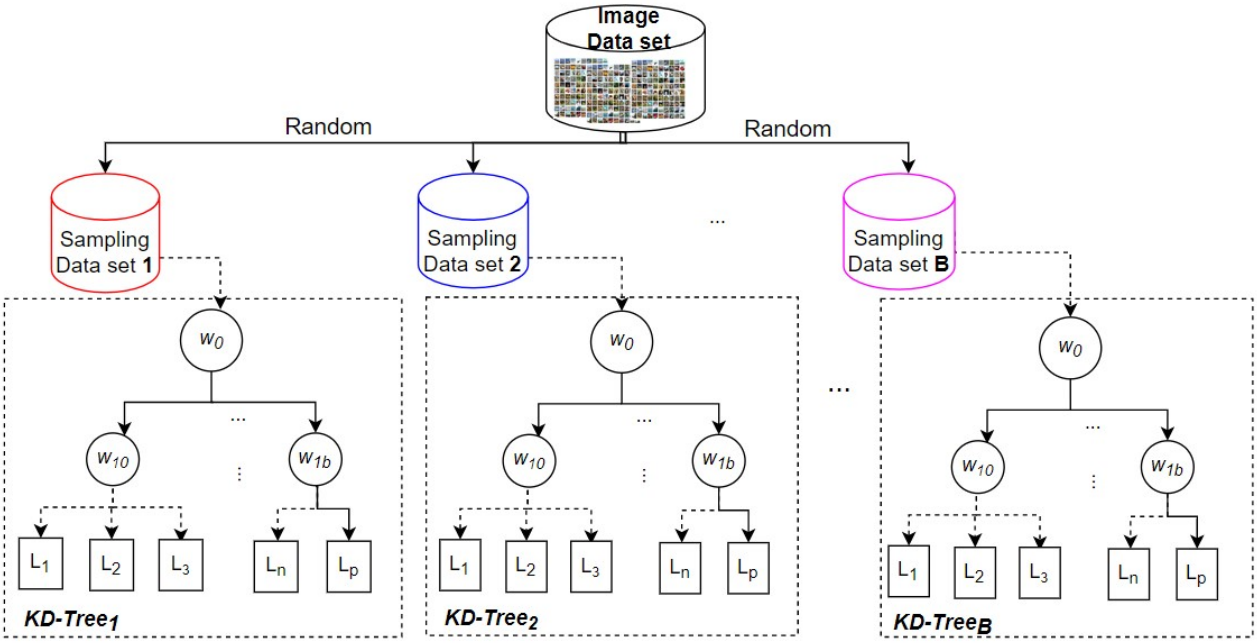
1. Cấu trúc KD-Tree cho phân lớp hình ảnh

Dựa trên tính chất cây KD-Tree nguyên thủy [10] và kế thừa từ công trình [16, 26], trong bài báo này cấu trúc KD-Tree phân lớp được xây dựng là một cây đa nhánh cân bằng, gồm một nút gốc lưu trữ véc-tơ phân lớp (w_0); tập nút trong Node, mỗi nút trong lưu trữ một véc-tơ phân lớp ($w_l k$) và tập nút lá Leaf, mỗi nút lá chứa nhiều véc-tơ hình ảnh $F = f_1, f_2, \ldots, f_k$. Sau khi xây dựng cấu trúc KD-Tree phân lớp tại mỗi nút lá là một phân lớp chứa tập véc-tơ hình ảnh thuộc nhiều phân lớp khác nhau. Cấu trúc KD-Tree đóng vai trò phân lớp cho ảnh đầu vào từ nút gốc đến nút lá và quá trình phân lớp trên KD-Tree được minh họa như Hình 2.

Độ phức tạp của thuật toán TKDT (Huấn luyện KD-Tree) là $O(p*h*n)$. Trong đó p là số lần điều chỉnh véc-tơ trọng số; h là chiều cao cây (hằng số); n là số phần tử tham gia vào quá trình xây dựng cây. Quá trình huấn luyện cấu trúc KD-Tree được thực hiện thông qua số lần cập nhật trọng số để tạo cây. Do đó độ phức tạp của thuật toán TKDT là độ phức tạp của số lần tạo cây. Vì vậy, độ phức tạp của thuật toán TKDT là $O(p*n)$.

2. Xây dựng cấu trúc KD-Tree Random Forest bằng kỹ thuật Bagging

Để nâng cao hiệu suất phân lớp cho ảnh đầu vào, trong bài báo này một cấu trúc KD-Tree Random Forest được



Hình 3. Minh họa KD-Tree Random Forest

xây dựng với kỹ thuật Bagging được đề xuất. Trong đó, KD-Tree Random Forest là cấu trúc gồm nhiều KD-Tree được xây dựng độc lập và huấn luyện để hình thành mô hình phân lớp bằng rừng ngẫu nhiên.

Rừng ngẫu nhiên là một trong những kỹ thuật phân lớp hình ảnh hiệu quả bởi sự kết hợp các kết quả đánh giá khách quan từ nhiều mô hình phân lớp độc lập. Quá trình xây dựng rừng ngẫu nhiên kết hợp với kỹ thuật Bagging được nhiều công trình đã công bố với kết quả phân lớp ảnh khả quan [19, 20, 21]. Trong bài báo này, một phương pháp kết hợp rừng ngẫu nhiên trong đó mỗi bộ phân lớp là một cấu trúc KD-Tree được xây dựng với kỹ thuật Bagging nhằm nâng cao hiệu suất phân lớp hình ảnh đầu vào. Rừng ngẫu nhiên được ứng dụng khá nhiều vào các công trình phân loại dữ liệu, phân lớp hình ảnh [18, 23] bởi hiệu suất phân loại đối tượng cao hơn một số kỹ thuật học máy đơn lẻ như k-NN, CNN, ANN, v.v. Tuy nhiên chi phí thời gian xây dựng và huấn luyện mô hình phân loại bằng rừng ngẫu nhiên là khá lớn. Trong bài báo này, trên cơ sở xây dựng KD-Tree Random Forest nhằm phân loại đối tượng hình ảnh bằng một mô hình kết hợp để thực hiện phân loại hình ảnh qua nhiều cấu trúc KD-Tree độc lập, rừng ngẫu nhiên cần được xây dựng. Kết quả phân lớp hình ảnh bằng mô hình rừng ngẫu nhiên dựa trên KD-Tree Random Forest bằng kỹ thuật Bagging áp dụng cho bài toán tìm kiếm ảnh được đánh giá là hiệu quả.

Trong công trình [16, 26] việc phân lớp ảnh bằng cấu trúc KD-Tree là thực hiện phân lớp cho ảnh đầu vào bằng một cấu trúc KD-Tree nên hiệu suất phân lớp chưa thực sự khách quan. Do đó, kỹ thuật phân lớp ảnh bằng nhiều cây KD-Tree để đảm bảo tính khách quan cho người dùng và nhằm nâng cao hiệu suất phân lớp là thực sự cần thiết. Để thực hiện điều này, rừng ngẫu nhiên được xây dựng bởi nhiều cây KD-Tree để thực hiện phân lớp đồng thời cho ảnh đầu vào. Sau khi thực hiện phân lớp trên từng cây KD-Tree, phương pháp bỏ phiếu bầu để lấy số đông cho kết quả phân lớp bằng rừng ngẫu nhiên.

Thuật toán 2: Xây dựng KD-Tree Random Forest

Input:Traning set S, Number of KD-Tree B;
Output: KD-Tree Forest
Function: **RF-KDT** (S,B)
begin
  foreach $x \in \mathcal{X}$ **do**
    S_i = a bootstrap sample from S;
    h_i = Create KD-Tree(S_i, W_{kt}, h, b);
    KD-TreeForest = KD-TreeForest;
    $\cup\ h_i$;
  end
  return KD-Tree Forest;
end

Trong bài báo này, rừng ngẫu nhiên được xây dựng là tập hợp nhiều cây KD-Tree nhằm thực hiện phân lớp nhiều lần cho một ảnh đầu vào (I) dựa trên các mô hình phân lớp riêng biệt và thực hiện song song. Minh họa rừng ngẫu nhiên KD-Tree RandomForest và kỹ thuật Bagging được minh họa trong hình 3 gồm các bước: (1) Xây dựng bộ Data Radom (Data$_i$); (2) Xây dựng KD-Tree cho từng bộ Data$_i$; (3) Tập hợp B cây KD-Tree hình thành KD-Tree Random Forest.

Độ phức tạp của thuật toán RF-KDT (Xây dựng KD-Tree Random Fores) là *B* O(k*h)*; trong đó *B* là số cây KD-Tree cần xây dựng và *O(P)* (*P = k*h*) là độ phức tạp cho xây dựng một cây KD-Tree có *k* phần tử và chiều cao *h*. Vậy độ phức tạp của thuật toán RF-KDT là *O(B*P)*, vì *B* là hằng số nên độ phức tạp của thuật toán RF-KDT là *O(P)*.

3. Phân lớp ảnh bằng KD-Tree Random Forest theo cơ chế phiếu bầu

Sự kết hợp nhiều cấu trúc cây phân lớp KD-Tree để xây dựng rừng ngẫu nhiên nhằm thực hiện phân loại đối tượng qua nhiều cấu trúc KD-Tree khác nhau. Mỗi hình ảnh được phân lớp bằng nhiều cây KD-Tree sẽ có nhiều hơn một kết quả nhãn lớp. Do đó, kết quả phân loại được thực hiện thông qua cơ chế phiếu bầu để mang lại hiệu quả phân lớp tốt hơn cho tập ảnh đơn đối tượng cũng như ảnh đa đối tượng [21, 22]. Trong bài báo này, một phương pháp phân lớp bằng KD-Tree Random Forest theo cơ chế phiếu bầu được đề xuất và thực hiện. Một ảnh đầu vào được thực hiện phân lớp bởi nhiều cây KD-Tree, kết quả phân lớp trên mỗi cây là một phân lớp cho ảnh đầu vào. Do đó, cần thực hiện bỏ phiếu bầu để tìm số phân lớp giống nhau là nhiều nhất cho ảnh I bởi rừng ngẫu nhiên. Sau khi thực hiện phiếu bầu thì phân lớp cho ảnh I là phân lớp có phiếu bầu nhiều nhất. Công thức phiếu bầu để chọn phân lớp cuối cùng (Final Class) cho ảnh đầu vào I là:

FinalClass(I) = max(number of voting I in Random Forest)

Trong trường hợp ảnh đầu vào thuộc *m* phân lớp khác nhau, kỹ thuật phiếu bầu sẽ được thực hiện lấy kết hợp *m* phân lớp có số phiếu bầu cao nhất.

Thuật toán 3: Phân lớp ảnh bằng KD-Tree Random Forest theo cơ chế phiếu bầu

```
Input:vetor f_i of image I
Output: Final class name of image I
Function: CKDFV (f_i, KD-Tree Forest)
begin
    Khởi tạo N KD-Tree cho rừng ngẫu nhiên (KD-Tree Forest);
    foreach KD-Tree[i] in KD-Tree Forest do
        Phân lớp ảnh I theo từng KD-Tree[i];
    end
    FinalClass(I) = Phân lớp được chọn theo phiếu bầu nhiều nhất;
    return FinalClass(I);
end
```

Độ phức tạp của thuật toán CKDFV là *O(N*h*k)*. Trong đó *B* là số cây KD-Tree trong KD-Tree Forest; *h* là chiều cao cây KD-Tree và k là số nhánh tối đa trên KD-Tree. Vì *B, h, k* là những hằng số nên đặt *T = B*h*k*. Vậy độ phức tạp của thuật toán CKDFV là *O(T)*, *T* là hằng số.

4. Cấu trúc R-Tree gom cụm dữ liệu

Quá trình gom cụm các véc-tơ đặc trưng dựa trên cấu trúc R^S-Tree [13]. Trong công trình này, tác giả đề xuất một cấu trúc cây để gom cụm và lưu trữ các véc-tơ đa chiều: R^S-Tree, một cấu trúc cải tiến của R-Tree, là cây đa nhánh cân bằng ứng dụng cho bài toán tìm kiếm ảnh tương tự. Cây R^S-Tree là cây phân hoạch dữ liệu không gian bao gồm: một nút gốc, một tập nút trong và một tập nút lá. R^S-Tree được đề xuất nhằm lưu trữ các phần tử hình ảnh trong không gian d chiều, mỗi phần tử $spED$ là một khối cầu có tâm $\overrightarrow{c}_{sp}$ và bán kính r_{sp} chứa véc-tơ đặc trưng ảnh $\overrightarrow{f_I} = (v_{I1}, v_{I2}, ..., v_{Id})$, với v_{Ii} là đặc trưng cấp thấp thứ i của hình ảnh I. Các nút trên R^S bao gồm: Nút trong S_{node}: là một bộ ⟨MBS,p⟩, trong đó MBS là một khối cầu có tâm $\overrightarrow{c}_{node}$ và bán kính r_{node}, p là con trỏ liên kết đến các nút con. Khối cầu này bao phủ các khối cầu của các nút thuộc nhánh cây con. Mỗi nút trong S_{node} có số phần tử tối thiểu là 2 và tối đa là N. Nút lá S_{leaf}: là bộ $<MBS, element>$, trong đó MBS là một khối cầu có tâm $\overrightarrow{c}_{leaf}$ và bán kính r_{leaf} chứa một tập thực thể element với mỗi thực thể $spED$ là một bộ $<MBS, oid>$ trong đó MBS là khối cầu có tâm $\overrightarrow{c}_{sp}$ và bán kính r_{sp} chứa không gian đối tượng, oid là định danh đối tượng $\overrightarrow{f_I} = (v_{I1}, v_{I2}, ..., v_{Id})$. Mỗi nút lá S_{leaf} có số phần tử tối đa là M và số phần tử tối thiểu là $m(1 < m < M/2)$.

5. Xây dựng cấu trúc R^S-Tree

Khi một phần tử $spED = <MBS(\overrightarrow{c}_{sp}, r_{sp}); f>$ được thêm vào cây R^S-Tree, việc chèn được thực hiện bắt đầu từ nút gốc, lần lượt duyệt qua tất cả các nút con của nút gốc và đi theo nhánh có khoảng cách $d_{\min}\{d_E(S_{node}.\overrightarrow{c}_{node}, spED.\overrightarrow{c}_{sp}), j = 1..N\}$, cho đến khi tìm được nút lá. Sau đó, phần tử $spED_i$ được phân phối vào nút lá hiện hành thỏa điều kiện $d_E(spED.\overrightarrow{c}_{sp}, S_{leaf}.\overrightarrow{c}_{leaf})<\theta$. Ngược lại, một nút lá mới $S_{newleaf}$ được tạo cùng cha với nút lá hiện hành để lưu phần tử $spED$. Quá trình xây dựng cấu trúc R^S-Tree theo các khối cầu nhằm giảm không gian lưu trữ so với phương pháp xây dựng theo khối hình chữ nhật, đây là một cải tiến của bài báo sử dụng cấu trúc R^S-Tree cho tìm kiếm ảnh. Thuật toán xây dựng R^S-Tree được trình bày như trong **Thuật toán 4**.

Gọi n là số phần tử của tập dữ liệu, M là số phần tử tối đa trong một nút của R^S-Tree. Thuật toán CRST lần lượt thực hiện duyệt từ nút gốc đến nút lá, mỗi lần duyệt qua

Thuật toán 4: Xây dựng R^S-Tree

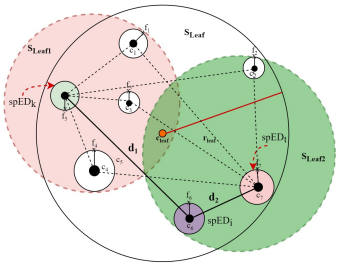
Input: Nút S_N, ngưỡng M và phần tử dữ liệu S_{spED}
Output: Cây R^S-Tree sau khi thêm phần tử
Function: **CRST** (S_N,$spED$)
begin
  if (S_N *không phải nút lá*) **then**
    Chọn nhánh gần với phần tử được thêm vào theo độ do Euclidean cho đến khi gặp được nút lá hiện hành S_{Lcrt};
  else
    if (S_{Lcrt} *có số phần tử nhỏ hơn* M) **then**
      Tính khoảng cách Euclidean $d = Euclidean(S_{spED}.\vec{c}, S_{Lcrt}.\vec{c}) + S_{spED}.r$;
      if ($d < \theta$) **then**
        Thêm phần tử S_{spED} vào nút lá S_{Lcrt};
        Cập nhật tâm cụm;
      else
        Tạo một nút lá mới S_{Lnew} cùng cha với nút lá hiện hành;
        Thêm phần tử S_{spED} vào nút lá mới S_{Lnew};
        Cập nhật tâm cụm;
      end
    else
      Thực hiện tách nút lá S_{leafk};
    end
  end
  return R^S-Tree;
end

M phần tử, mỗi lần duyệt thuật toán CRST thực hiện phép cập nhật tâm và tách nút từ nút lá đến nút gốc. M là hằng số. Do đó, thuật toán CRST có độ phức tạp là $O(logn)^3$.

6. Cải tiến tách nút trên R^S-Tree

Thuật toán tách nút trong công trình [13], việc phân hoạch tập các phần tử về hai nút con được thực hiện bằng cách lựa chọn tuần tự ngẫu nhiên trong tập $M-1$ phần tử để phân phối lần lượt vào hai nút con. Trong bài báo này, thuật toán tách nút được cải tiến như sau: tập $M-1$ phần tử sẽ được tính độ đo sai biệt so với hai phần tử được chọn làm hai tâm của hai nút mới. Trong quá trình tách nút lá thành hai nút con, việc lựa chọn phần tử $spED_i$ phân phối vào hai nút lá mới cho phù hợp được thực hiện dựa theo độ sai biệt (ký hiệu là d_{Dif}) được minh họa như trong Hình 4, phần tử có độ đo sai biệt lớn sẽ được chọn để phân phối trước.

Trong Hình 4, độ đo sai biệt (d_{Dif}) của phần tử $spED_i$ so với hai phần tử $spED_k$ và $spED_t$ là hai phần tử được chọn làm tâm của hai nút lá mới S_{Leaf1} và S_{Leaf2} được tính như sau: $d_{Dif} = |d_1 - d_2|$, trong đó $d_1 = d_E(spED_i.\vec{c}_{spi}, spED_k.\vec{c}_{spk}), d_2 = d_E(spED_i.\vec{c}_{spi}, spED_t.\vec{c}_{spt})$. Độ sai biệt càng lớn thì khả năng xác định phần tử thuộc về một trong hai nút càng cao.



Hình 4. Mô tả thuật toán tách nút dựa vào độ đo sai biệt

Để thực hiện tách nút lá thành hai nút con, hai khối cầu S_{spEDk}, S_{spEDt} xa nhau nhất trong nút tách sẽ được chọn để làm tâm khởi đầu hai nút con. Việc chọn hai tâm được thực hiện theo **Thuật toán 5**.

Thuật toán 5: Lựa chọn hai phần tử làm tâm hai nút con

Input: Nút lá S_{Leaf}
Output: Hai khối cầu phần tử $spED_k$, $spED_t$
Function: **SelectedSpEDs** (S_{Leaf})
begin
  Khởi tạo $d_{crmax} = 0$;
  Khởi tạo $spED_k$ = null;
  Khởi tạo $spED_t$ = null;
  for *Duyệt các phần tử trong nút lá* S_{Leaf} **do**
    Tính khoảng cách Euclidean;
    Chọn phần từ xa tâm nút lá S_{Leaf} nhất gán giá trị cho phần tử $spED_k$;
  end
  $S_{Leaf} = S_{Leaf}$-$spED_k$;
  for *Duyệt các phần tử trong nút lá* S_{Leaf} **do**
    Chọn phần tử xa phần tử $spED_k$ gán cho $spED_t$;
  end
  return {$spED_k$,$spED_t$};
end

Để phân phối các phần tử vào hai nút lá được tối ưu, các phần tử trong danh sách nút lá cần tách được sắp xếp theo độ đo theo độ sai biệt, thuật toán được trình bày như trong **Thuật toán 6**.

Sau khi lựa chọn được hai phần tử khối cầu làm tâm của hai nút lá được tách ra, các phần tử còn lại lần lượt được phân phối vào hai nút đã tách. Quá trình phân phối được thực hiện quy tắc chọn nhánh tốt nhất. Thuật toán phân phối các phần tử được trình bày như trong **Thuật toán 7**.

Nút S_{Leaf} bị tràn được xử lý bằng cách tạo hai nút mới S_{Leaf1}, S_{Leaf2}, cùng mức với S_{Leaf}, phân vùng $M+1$ phần tử vào hai nút. Khi quá trình phân tách xảy ra có thể lan truyền đến nút trong, thực hiện tách nút trong. Quá trình này có thể lan truyền đến nút gốc, khi nút gốc đầy một nút gốc mới sẽ được tạo ra. Thuật toán tách nút lá được trình

Thuật toán 6: Sắp xếp các phần tử theo độ đo sai biệt

Input: Ba nút lá S_{Leaf}, S_{Leaf1}, S_{Leaf2}
Output: Danh sách các phần tử sau khi sắp xếp S_{Lsort}
Function: **SortListSpED** (S_{Leaf},S_{Leaf1},S_{Leaf2})
begin
  for *Duyệt các phần tử trong nút lá* S_{Leaf} **do**
    Tính các khoảng cách Euclidean
    d_1=$Euclidean(S_{Leaf1}.\overrightarrow{c}, S_{Leaf}.SpED[i].\overrightarrow{c})$;
    d_2=$Euclidean(S_{Leaf2}.\overrightarrow{c}, S_{Leaf}.SpED[i].\overrightarrow{c})$;
    Tính độ đo sai biệt; $d_i = |d_1 - d_2|$;
    Đưa các phần tử vào danh sách $List_{SL}$;
    $List_{SL}.Add(S_{Leaf}.SpED[i], d_i)$;
  end
  Sắp xếp các phần tử theo độ đo sai biệt; $S_{Lsort} = List_{SL}.Sort$;
  return S_{Lsort};
end

Thuật toán 7: Phân phối các phần tử vào hai nút lá

Input: Ba nút lá cần tách S_{Leaf}, S_{Leaf1}, S_{Leaf2}
Output: Các nút sau khi phân phối xong
Function: **DistributeSpED** (S_{Leaf},S_{Leaf1},S_{Leaf2})
begin
  Khởi tạo
  $S_{temp} = SortListSpED(S_{Leaf}, S_{Leaf1}, S_{Leaf2})$;
  while $(S_{temp} == \emptyset || S_{Leaf1}.size! = M - m + 1 || S_{Leaf2}.size! = M - m + 1)$ **do**
    Tính khoảng cách Euclidean đến tâm của hai nút lá S_{Leaf1}, S_{Leaf2};
    if $(Euclidean(S_{Leaf1}.\overrightarrow{c}, spED.\overrightarrow{c} < Euclidean(S_{Leaf2}.\overrightarrow{c}, spED.\overrightarrow{c}))$ **then**
      Đưa phần tử spED vào nút lá S_{Leaf1};
      Cập nhật tâm;
    else
      Đưa phần tử spED vào nút lá S_{Leaf1};
      Cập nhật tâm;
      $S_{temp} = S_{temp} - \{spED\}$;
    end
  end
  if $(S_{Leaf1}.size == M - m + 1)$ **then**
    $S_{Leaf2} = S_{Leaf2} \cup S_{temp}$;
  else
    $S_{Leaf1} = S_{Leaf1} \cup S_{temp}$;
  end
  return $\{S_{Leaf1}, S_{Leaf2}\}$;
end

bày như **Thuật toán 8**.

Gọi n là số phần tử của tập dữ liệu, M là số phần tử tối đa trong một nút R^S-Tree. Khi thực hiện tách nút, trong trường hợp xấu nhất, Thuật toán SplitLeafRST phải tách từ nút lá đến nút gốc. Mỗi lần tách nút, Thuật toán SplitLeafRST phải thực hiện M phép so sánh để phân bố về k-cụm. Mặt khác, mỗi lần thực hiện tách nút, trong trường hợp xấu nhất Thuật toán SplitLeafRST phải gọi đệ

Thuật toán 8: Thuật toán tách nút lá

Input: Nút lá cần tách S_{Leaf}
Output: Cây R^S-Tree sau khi tách
Function: **SplitLeafRST** (S_{Leaf})
begin
  Tạo nút lá S_{Leaf1}; Tạo nút lá S_{Leaf2};
  Chọn hai phần tử xa nhau nhất trong nút lá S_{Leaf};
  SelectedSpEDs(S_{Leaf});
  Đưa $spED_k$ vào nút lá S_{Leaf1};
  Cập nhật tâm và bán kính;
  Đưa $spED_t$ vào nút lá S_{Leaf2};
  Cập nhật tâm và bán kính;
  Xóa hai phần tử này ra khỏi S_{Leaf};
  Gọi hàm sắp xếp các phần tử theo độ đo sai biệt;
  S_{Lsort} = SortListSpED(S_{Leaf},S_{Leaf1},S_{Leaf2});
  Gọi hàm phân phối các phần tử vào hai nút;
  DistributeSpED(S_{Lsort},S_{Leaf1},S_{Leaf2});
  S_{Lpr} = S_{Leaf}.parent;
  if (*Số phần tử nút cha* S_{Lpr} *nhỏ hơn* N) **then**
    Cập nhật tâm nút cha;
  else
    Tách nút cha;
  end
end

quy từ nút lá đến nút gốc. M là hằng số. Do đó, độ phức tạp của Thuật toán SplitLeafRST là $O(logn)^2$.

7. Tìm kiếm ảnh tương tự dựa trên R-KD-Tree

Từ cây phân cụm dữ liệu không gian R^S-Tree đã tạo, một thuật toán tra cứu ảnh tương tự theo nội dung dựa trên cây R^S-Tree được đề xuất. Quá trình tìm kiếm ảnh tương tự được thực hiện trên cây R^S-Tree và được mô tả như **Thuật toán 9**.

Thuật toán 9: Thuật toán tìm kiếm ảnh trên R-KD-Tree

Input: Đặc trưng ảnh truy vấn $spED$, R^S-Tree
Output: Tập ảnh tương tự SI
Function: **RSIR** (S_{Nr}, $spED$)
begin
  $S_{Node} = S_{Nr}$;
  if (S_{Node} *trỏ tới phần tử null*) **then**
    return *null*;
  else
    if (S_{Nk} *không phải lá nút lá*) **then**
      Đi theo nhánh gần nhất với phần tử truy vấn ký hiệu S_{Nk};
      RSIR (S_{Nk}, $spED$);
    else
      SI = {Các phần tử ảnh tương tự trong nút lá hiện hành S_{Lcrt}};
    end
  end
  return SI;
end

Bảng I
MÔ TẢ CÁC THAM SỐ THỰC NGHIỆM

Tham số	COREL	Caltech-101
M	120	100
m	1	1
N	20	25
θ	0.078	0.085
$Topk$	80	50-100
Thời gian xây dựng cây R^S-Tree (giờ)	0.41	2.79

Gọi n là số phần tử của tập dữ liệu, M là số phần tử tối đa trong một nút của R^S-Tree. Thuật toán RSIR lần lượt duyệt qua các nút từ gốc đến lá, hơn nữa vì cây R^S-Tree là cây cân bằng nên thuật toán RSIR duyệt qua chiều cao h của cây. Mỗi lần duyệt, thuật toán RSIR phải so sánh với M phần tử của mỗi nút. M là hằng số. Do đó, độ phức tạp của thuật toán RSIR là $O(logn)$.

V. THỰC NGHIỆM

1. Mô tả dữ liệu thực nghiệm

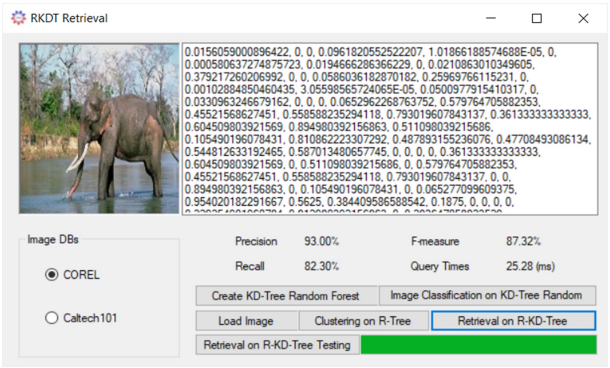
Để đánh giá hiệu năng của các hệ thống truy vấn ảnh, một số bộ dữ liệu ảnh tiêu chuẩn được sử dụng cho quá trình thực nghiệm gồm COREL, Caltech101. Bộ ảnh COREL gồm 1,000 ảnh JPEG chia thành 10 chủ đề khác nhau. Mỗi chủ đề gồm 100 ảnh. Bộ ảnh Caltech-101 gồm 9,144 ảnh JPEG chia thành 102 chủ đề. Mỗi chủ đề chứa từ 40 đến 800 ảnh. Mỗi ảnh có kích thước từ 200 – 300 pixels. Mỗi hình ảnh được trích xuất thành một véc-tơ đặc trưng có 81 thành phần. Quá trình trích xuất véc-tơ đặc trưng được kế thừa từ công trình [16]. Dữ liệu thực nghiệm trong bài báo này không có ràng buộc thêm so với các công trình khác thực nghiệm trên cùng bộ ảnh.

2. Thực nghiệm và đánh giá

Trong bài báo này, các tham số M, m, N, θ được tiến hành thực nghiệm để lựa chọn giá trị nhằm nâng cao độ chính xác truy vấn. Gọi $nMax_{inclass}$ là số phần tử tối đa trong một phân lớp. Chúng tôi thực nghiệm chọn $M \in [nMax_{inclass} - \alpha, nMax_{inclass} + \alpha]$, α là hằng số; $N \in [2, M]$. Ngưỡng θ để đánh giá độ tương tự các phần tử thuộc một cụm. Số phần tử của mỗi phân lớp xấp xỉ 10% của bộ dữ liệu. Do đó, chúng tôi thực nghiệm giá trị $\theta \in [0.1 - \mu, 0.1 + \mu], \mu \in [0, 0.05]$ là hệ số học. Thông qua quá trình thực nghiệm, các tham số tối ưu được lựa chọn sao cho hệ thống đạt được độ chính xác tốt nhất được trình bày như Bảng I.

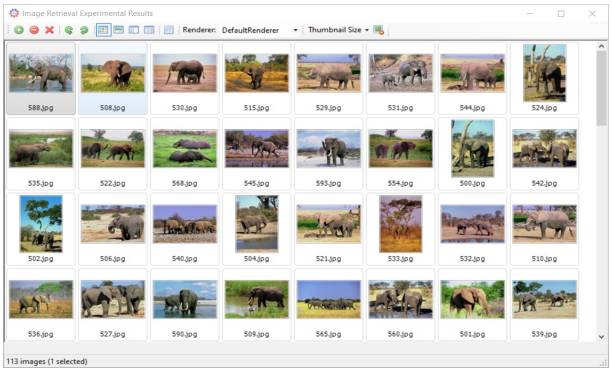
Việc lựa chọn tham số trong quá trình thực nghiệm phụ thuộc vào tập dữ liệu bao gồm: (1) số các phần tử dữ liệu mẫu trong một phân lớp; (2) số lượng phần tử trong tập dữ liệu ảnh. Số phần tử trong một phân lớp càng lớn thì tham số M càng lớn. Số lượng phần tử trong tập dữ liệu lớn thì N càng lớn nhằm hạn chế chiều cao của cây giúp giảm thời gian tìm kiếm. Ngưỡng θ phụ thuộc vào độ tương tự của các phần tử trong một phân lớp, nếu các phần tử trong một phân lớp càng gần nhau thì ngưỡng θ càng bé. $TopK$ càng lớn độ chính xác càng thấp, độ phủ càng cao. Tham số $TopK$ được lựa chọn sao cho độ đo dung hòa đạt tốt nhất.

Thực nghiệm tìm kiếm ảnh tương tự dựa trên cấu trúc R-Tree kết hợp KD-Tree Random Forest được minh họa bởi hình 5. Hệ truy vấn ảnh RKDT dựa trên kết quả phân lớp ảnh đầu vào (Load Image) bằng KD-Tree Random Forest (Image Classification on KD-Tree Random Forest), sau khi đã xây dựng và huấn luyện mô hình phân lớp bằng rừng ngẫu nhiên là nhiều cấu trúc KD-Tree (Create KD-Tree Randm Forest). Sau khi thực hiện phân lớp cho ảnh đầu vào là quá trình gom cụm bằng R-Tree (Clustering on R-Tree). Dựa trên kết quả gom cụm từ R-Tree để tìm kiếm và trích xuất tập ảnh tương tự với ảnh đầu vào sau khi trích xuất véc-tơ đặc trưng.

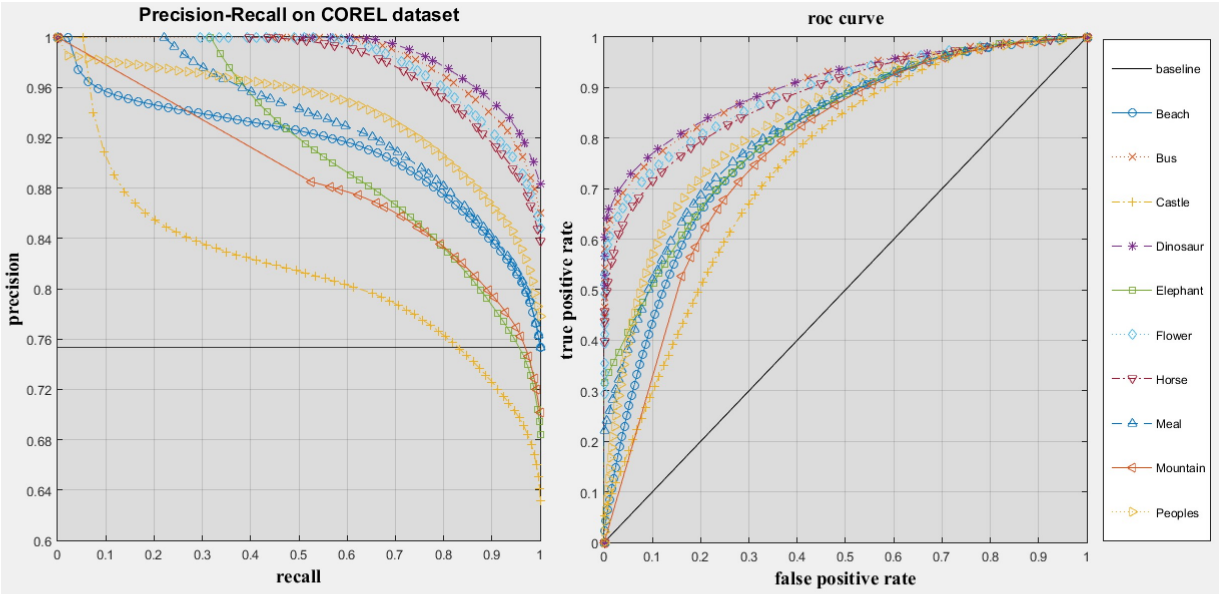


Hình 5. Hệ truy vấn ảnh RKDT

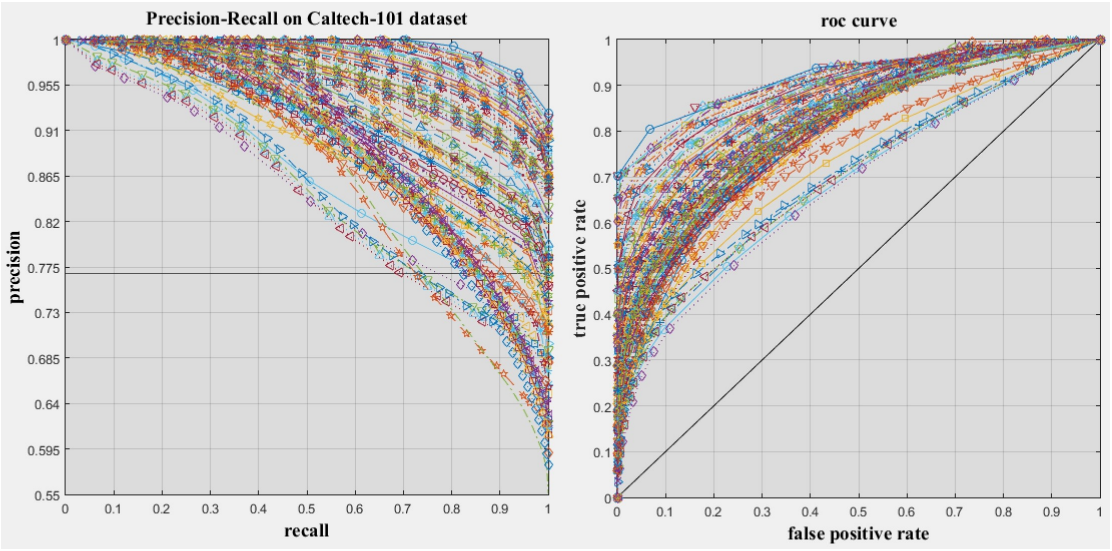
Kết quả tìm kiếm tập ảnh tương tự với ảnh đầu vào (ảnh 588.jpg bộ COREL) được minh họa bởi Hình 6.



Hình 6. Kết quả truy vấn ảnh 588.jpg ($COREL$)



Hình 7. Biểu đồ Precision, Recall và ROC bộ ảnh COREL



Hình 8. Biểu đồ Precision, Recall và ROC bộ ảnh Caltech101

Bảng II
HIỆU SUẤT PHÂN LỚP ẢNH BẰNG
KD-TREE RANDOM FOREST

Bộ dữ liệu	Số ảnh	Hiệu suất phân lớp
COREL	1,000	0.8972
Caltech101	9,144	0.7705

Sau khi xây dựng và huấn luyện mô hình phân lớp ảnh bằng KD-Tree Random Forest kết quả thực nghiệm phân lớp bằng rừng ngẫu nhiên được trình bày trong Bảng II.

Kết quả thực nghiệm tìm kiếm ảnh trên các bộ ảnh COREL, Caltech101 với độ chính xác (*Precision*), độ phủ (*Recall*) và độ dung hòa (*F-measure*), thời gian truy vấn trung bình của các bộ ảnh (*Time query*) tính theo mili giây (*ms*) được trình bày trong Bảng III. Kết quả này cho thấy hiệu suất tìm kiếm ảnh trên các bộ ảnh khi kết hợp hai cấu trúc này là cao hơn so với các công trình sử dụng riêng lẻ từng cấu trúc trong công trình [15, 16].

Kết quả thực nghiệm trên các bộ ảnh COREL, Caltech101 được minh họa bằng biểu đồ ROC thực hiện trên Matlab 2015 được trình bày như Hình 7 – Hình 8.

Bảng IV, trình bày so sánh thực nghiệm tìm kiếm ảnh hệ truy vấn RKDT trên từng bộ dữ liệu với các công trình

Bảng III
HIỆU SUẤT TRUY VẤN ẢNH TRÊN HỆ **RKDT**

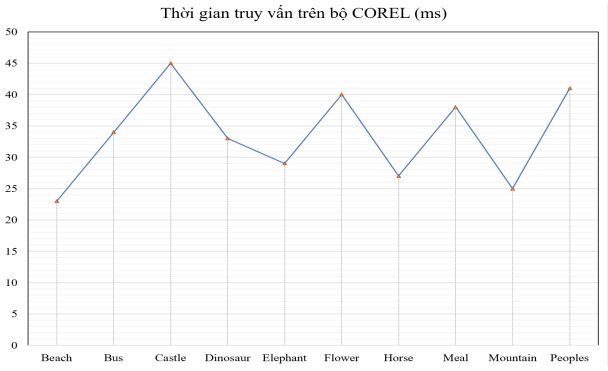
Bộ ảnh	Precision	Recall	F-measure	Query time
COREL	**0.8101**	0.7344	0.7703	33.16
Caltech101	**0.7354**	0.7020	0.7183	98.12

Bảng IV
SO SÁNH HIỆU SUẤT TRUY VẤN VỚI CÁC PHƯƠNG PHÁP KHÁC TRÊN CÙNG BỘ ẢNH

Phương pháp	Tập ảnh	Precision
B-SHIFT (2016), [11]	COREL	0.7200
k-NN KD-Tree (2019), [12]	COREL	0.6430
ORB 8-D with MPL (2020), [24]	COREL	0.6656
RKDT	**COREL**	**0.8101**
BDIPs, GLCM - SVM (2020), [13]	Caltech101	0.6314
MTSD, (2020), [14]	Caltech101	0.6942
Feature Detector (2021), [25]	Caltech101	0.6200
RKDT	**Caltech101**	**0.7354**

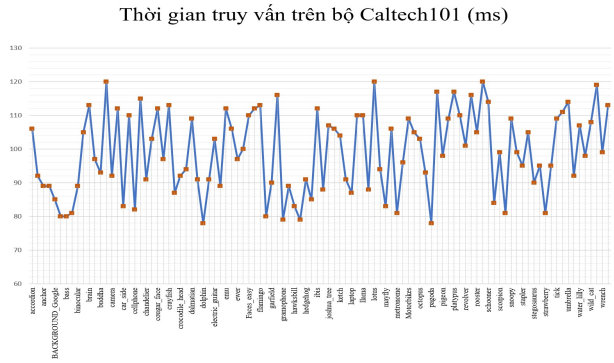
khác là cao hơn bởi các yếu tố: (1) hệ truy vấn RKDT đã kết hợp được cấu trúc R-Tree với KD-Tree nên phát triển những ưu điểm từ hai cấu trúc này; (2) Quá trình phân lớp trên KD-Tree Random For-est trước khi gom cụm trên R-Tree đã góp phần nâng cao hiệu suất truy vấn ảnh. Đồng thời, kết quả thực nghiệm bộ ảnh COREL cho thấy sự kết hợp cấu trúc KD-Tree và R-Tree đã mang lại hiệu suất tìm kiếm ảnh cao hơn so với các công trình đã công bố bởi chúng tôi đã công bố trước đây khi chỉ dùng một cấu trúc KD-Tree hoặc R-Tree riêng lẻ [15, 16].

Hình 9 và Hình 10 biểu diễn hiệu suất thời gian truy vấn trung bình cho từng phân lớp trên các bộ ảnh COREL và Caltech101.



Hình 9. Thời gian truy vấn trung bình của bộ ảnh COREL

Trong bài báo này, kết quả hiệu suất truy vấn ảnh được thực hiện so sánh với một số công trình khác trên cùng bộ ảnh thực nghiệm nhưng không so sánh chi phí thời gian huấn luyện mô hình phân lớp vì cấu hình máy tính không đồng nhất. Đồng thời, chi phí huấn luyện mô hình phân lớp KD-Tree Random Forest chỉ thực hiện một lần và sau đó được thực thi nhiều lần với bộ trọng số đã huấn



Hình 10. Thời gian truy vấn trung bình của bộ ảnh Caltech101

luyện. Kết quả thời gian truy vấn ảnh trung bình trên các bộ ảnh COREL, Cal-tech101 cho thấy phương pháp phân lớp ảnh bằng KD-Tree Random Forest là hoàn toàn khả thi và hiệu quả. Từ kết quả phân lớp này thực hiện gom cụm trên R-Tree đã mang lại hiệu suất truy vấn ảnh được đánh giá là khá cao.

VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài báo này, một mô hình truy vấn ảnh có sự kết hợp cấu trúc R-Tree với KD-Tree Random Forest được đề xuất và thực nghiệm trên các bộ ảnh COREL, Caltech101. Sự kết hợp hai cấu trúc được xây dựng theo hướng tiếp cận gom cụm và phân lớp đã góp phần nâng cao hiệu suất tìm kiếm ảnh vì phát huy được những ưu điểm của từng cấu trúc và bổ trợ cho nhau. Thực nghiệm truy vấn ảnh trên các bộ ảnh COREL, Caltech101 với độ chính xác trung bình tương ứng là 0.8101, 0.7354. Trong công trình này đã cải tiến nâng cao hiệu suất cho quá trình phân lớp bằng KD-Tree Random Forest trước khi gom cụm và tìm kiếm trên R-Tree. Định hướng phát triển tiếp theo của chúng tôi là từ kết quả phân lớp bằng KD-Tree Random Forest kết hợp cải tiến R-Tree và đồ thị làng giềng để nhằm thực hiện truy vấn trên ontology nhằm nâng cao hiệu suất cho bài toán tìm kiếm ảnh theo ngữ nghĩa và áp dụng cho nhiều bộ dữ liệu nhằm đa dạng tập ảnh thực nghiệm.

LỜI CẢM ƠN

Chúng tôi xin trân trọng cảm ơn Khoa Công nghệ thông tin – Trường Đại học Khoa học - Đại học Huế, nhóm nghiên cứu SBIR-HCM đã góp ý chuyên môn cho nghiên cứu này. Chúng tôi xin trân trọng cảm ơn Trường Đại học Sư phạm TP.HCM đã tạo điều kiện về cơ sở vất chất giúp chúng tôi hoàn thành bài nghiên cứu này.

TÀI LIỆU THAM KHẢO

[1] A Patrizio, "Data center explorer", *Network World*, https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html, (31/3/2022).

[2] Le, Thanh Manh, et al, "Image retrieval system based on EMD similarity measure and S-tree", *In: Intelligent Technologies and Engineering Systems, Springer, New York, NY*, pp. 139-146, 2013.

[3] Chandresh Pratap Singh, "R-Tree implementation of image databases", *Sig-nal and Image Processing: An International Journal (SIPIJ)*, vol. 18.4, 2011.

[4] Corel 1k Database: http://wang.ist.psu.edu/docs/related/, (21/2/2022)

[5] Caltech101: http://www.vision.caltech.edu/Images Dataset/-Caltech101, (21/2/2022).

[6] Zhang, Yuqian, et al, "Fast face sketch synthesis via kd-tree search", *European Conference on Computer Vision. Springer, Cham*, 2016.

[7] Vanitha J., Senthilmurugan M, "Multidimensional Image Indexing Using SR-Tree Algorithm for Content-Based Image Retrieval System", *In: Shetty N., Patnaik L., Prasad N., Nalini N. (eds) Emerging Research in Computing, In-formation, Communication and Applications*, pp. 81-88, 2018.

[8] Arpitha Y.C, Bhargavi A.G, Mahima K.T, Nagavi K.K, Meenakshi H.N, "A Navigation Supporting System Using R-Tree", *Published by AIJR Publisher in Proceedings of the 3rd National Conference on Image Processing, Compu-ting, Communication, Networking and Data Analytics*, 2018.

[9] Xinlu Wang, Weiming Meng, Mingchuan Zhang, "A novel information re-trieval method based on R-tree index for smart hospital information system", *International Journal of Advanced Computer Research*, vol. 9.41, 2019.

[10] Bentley, Jon Louis, "Multidimensional binary search trees used for associative searching", *Communications of the ACM*, vol. 18.9, pp. 509-517, 1975.

[11] Douik, Ali, Mehrez Abdellaoui, and Leila Kabbai, "Content based image re-trieval using local and global features descriptor", *2016 2nd international conference on advanced Technologies for Signal and Image Processing (ATSIP) IEEE*, pp. 151-154, 2016.

[12] Abdulkadhem, Abdulkadhem Abdulkareem, and Tawfiq A. Al-Assadi, "Pro-posed a Content-Based Image Retrieval System Based on the Shape and Tex-ture Features", *Int. J. Innovat. Technol. Explor*, vol. 8, pp. 2189, 2019.

[13] Rajesh, M. B., Sathiamoorthy, S., "Co-occurrence of Edges and Valleys with Support Vector Machine for Content Based Image retrieval", *In 2020 Fourth International Conference on Computing Methodologies and Communication*, pp. 465-470, 2020.

[14] Sathiamoorthy, S., Natarajan, M., "An efficient content based image retrieval using enhanced multi-trend structure descriptor", *SN Applied Sciences*, vol. 2.2, pp. 1-20, 2020.

[15] Lê Thị Vĩnh Thanh, Văn Thế Thành, Lê Mạnh Thạnh, "Một phương pháp tìm kiếm ảnh hiệu quả dựa trên cấu trúc R-Tree", *Kỷ yếu Hội thảo Quốc gia về Công nghệ thông tin và ứng dụng trong các lĩnh vực (CITA2021), Đại học Đà Nẵng, Nhà xuất bản Đà Nẵng, ISBN: 978-604-84-5998-7*, trang 259-271, 2021.

[16] Nguyễn Thị Định, Văn Thế Thành, Lê Mạnh Thạnh, "Phân lớp ảnh bằng cây KD-Tree cho bài toán tìm kiếm ảnh tương tự", *Chuyên san Các công trình nghiên cứu, phát triển và ứng dụng Công nghệ thông tin và Truyền thông*, Tập 2021, Số 1, 2021.

[17] Zaklouta, F., Stanciulescu, B., and Hamdoun, O., "Traffic sign classification us-ing kd trees and random forests", *In The 2011 international joint conference on neural networks, IEEE*, pp. 2151-2155, 2011.

[18] Man, W., Ji, Y., and Zhang, Z., "Image classification based on improved random forest algorithm", *In 2018 IEEE 3rd International Conference on Cloud Com-puting and Big Data Analysis (ICCCBDA)*, pp. 346-350, 2018.

[19] Amini, S., Homayouni, S., Safari, A., and Darvishsefat, A, "Object-based classi-fication of hyperspectral data using Random Forest algorithm", *Geo-spatial in-formation science*, vol. 21.2, pp. 127-138, 2018.

[20] Jain, V., and Phophalia, A, "M-ary Random Forest-A new multidimensional par-titioning approach to Random Forest", *Multimedia Tools and Applications*, vol. 80.28, pp. 35217-35238, 2021.

[21] Kabari, L. G., and Onwuka, U. C, "Comparison of Bagging and Voting Ensemble Machine Learning Algorithm as a Classifier", *International Journals of Ad-vanced Research in Computer Science and Software Engineering*, vol. 9.3, pp. 19-23, 2019.

[22] Zhang, J., and Shi, H, "Kd-tree based efficient ensemble classification algorithm for imbalanced learning", *In 2019 international conference on machine learn-ing, big data and business intelligence (MLBDBI), IEEE*, pp. 203-207, 2019.

[23] Wang, A., Wang, Y., and Chen, Y, "Hyperspectral image classification based on convolutional neural network and random forest", *Remote sensing letters*, vol. 10.11, pp. 1086-1094, 2019.

[24] Chhabra, P., Garg, N. K., and Kumar, M, "Content-based image retrieval system using ORB and SIFT features", *Neural Computing and Applications*, vol. 32.7, pp. 2725-2733, 2020.

[25] Ahmed, K. T., Aslam, S., Afzal, H., Iqbal, S., Mehmood, A., and Choi, G. S, "Symmetric Image Contents Analysis and Retrieval Using Decimation", *Pattern Analysis, Orientation, and Features Fusion. IEEE Access*, vol. 9, pp. 57215-57242, 2021.

[26] Nguyễn Thị Định, Thế Thành Văn, Mạnh Thạnh Lê, "Một phương pháp phân lớp dựa trên cấu trúc KD-Tree cho bài toán tìm kiếm ảnh theo ngữ nghĩa", *Kỹ yếu hội nghị quốc gia lần thứ 14 về nghiên cứu cơ bản và ứng dụng công nghệ thông tin, Fair2021, TP. Hồ Chí Minh, 23/12/2021*. DOI: 10.15625/vap.2021.0075. (2021).

[27] Dinh, N. T., and Le, T. M, "An Improvement Method of Kd-Tree Using k-Means and k-NN for Semantic-Based Image Retrieval System", *In World Conference on Information Systems and Technologies, Springer, Cham*, pp. 177-187, 2022.

[28] Thanh, L. T. V., and Thanh, L. M, "Semantic-Based Image Retrieval Using-Tree and Neighbor Graph", *In World Conference on Information Systems and Technologies . Springer, Cham*, pp. 165-176, 2022.

SƠ LƯỢC VỀ TÁC GIẢ

Lê Mạnh Thạnh

Sinh năm 1953, nhận bằng Tiến sĩ ngành khoa học máy tính tại Đại học Budapest (ELTE), Hungary vào năm 1994. Nhận hàm Phó giáo sư tại trường Đại học Khoa học, Đại học Huế, Việt Nam vào năm 2004. Lĩnh vực nghiên cứu quan tâm bao gồm cơ sở dữ liệu, cơ sở tri thức và lập trình logic. Email: lmthanh@hueuni.edu.vn

Lê Thị Vĩnh Thanh

Sinh năm 1983, tốt nghiệp ngành Sư phạm tin học Trường Đại học Sư phạm TP.HCM vào năm 2006, nhận bằng Thạc sĩ ngành Hệ thống thông tin tại Trường Đại học Khoa học Tự nhiên TP.HCM vào năm 2014. Hiện là nghiên cứu sinh ngành Khoa học máy tính Trường Đại học Khoa học, Đại học Huế. Lĩnh vực nghiên cứu: xử lý ảnh, tìm kiếm ảnh và cơ sở dữ liệu. Email: thanhltv@bvu.edu.vn

Lương Thị Thanh Xuân

Sinh năm 1987, tốt nghiệp ngành Sư phạm Tin học Trường Đại học Đồng Tháp năm 2008. Hiện đang học Thạc sĩ ngành Khoa học máy tính khóa 2020 – 2022 tại trường Đại học Khoa học – Đại học Huế. Lĩnh vực nghiên cứu quan tâm là tìm kiếm ảnh. Email: lttxuan.c23nvk.dtp@moet.edu.vn

Nguyễn Thị Định

Sinh năm 1983, tốt nghiệp ngành Sư phạm tin học Trường Đại học Sư phạm TP.HCM vào năm 2006, nhận bằng Thạc sĩ ngành Truyền số liệu và mạng máy tính tại Học viện Công nghệ Bưu chính viễn thông TP.HCM vào năm 2011. Hiện là nghiên cứu sinh ngành Khoa học máy tính Trường Đại học Khoa học, Đại học Huế. Lĩnh vực nghiên cứu quan tâm gồm xử lý ảnh, tìm kiếm ảnh và cơ sở dữ liệu. Email: dinhnt@hufi.edu.vn

Văn Thế Thành

Sinh năm 1979, tốt nghiệp chuyên ngành Toán tin tại Đại học Khoa học Tự nhiên - Đại học Quốc gia TP.HCM vào năm 2001, nhận bằng Thạc sĩ Khoa học Máy tính tại Đại học Quốc gia TP.HCM vào năm 2008. Năm 2016, nhận bằng Tiến sĩ Khoa học Máy tính tại Đại học Khoa học, Đại học Huế. Lĩnh vực nghiên cứu quan tâm bao gồm xử lý ảnh, khai thác dữ liệu ảnh và tìm kiếm ảnh. Email: thanhvt@hufi.edu.vn

Automatic Classification of Software Requirements in Vietnamese Based on Machine Learning Techniques

Thi Nham Cao[1], Dai Tho Dang[2], Thi My Hanh Le[3], Nguyen Thanh Binh[2]

[1]University Of Economics, The University Of Danang, Danang
[2]Vietnam - Korea University of Information and Communication Technology, The University of Danang, Danang
[3]University of Science and Technology, The University of Danang, Danang

Correspondence: Thanh Binh Nguyen, Email: ntbinh@vku.udn.vn
Communication: received 01/05/2022, revised 01/06/2022, accepted 30/06/2022
DOI: 10.32913/mic-ict-research.v2022.n1.1051

***Abstract*: Requirements engineering is often the first stage in the software process to understand the problem statement. Finding mistakes earlier in requirements helps reduce the development cost. One activity contributing to defining clear, complete and precise requirements is classifying requirement items in the specification. This paper presents a classification approach of functional and non-functional requirements in Vietnamese using different supervised machine learning techniques. Five supervised machine learning algorithms, including Naïve Bayes (NB), Support Vector Machine (SVM), Logistics Regression (LR), Multi-layer Perceptron Neural Network (MLP), and FastText, are implemented, trained, tested and compared using a dataset. The experimental results show that NB is the best model in terms of accuracy.**

Keywords: *Software engineering, requirements engineering, requirements classification, machine learning.*

I. Introduction

Requirement Engineering (RE) is a process of eliciting imprecise, incomplete needs and wishes of the potential users of the software, analyzing, validating, and translating them into complete, precise, and formal specifications. This activity is generally the first step in the software process. Misunderstanding requirements could be very serious. Because fixing a requirement mistake after delivery may cost up to 100 times the cost of fixing an implementation mistake. Therefore, the importance of RE is enormous to develop software and reduce errors at a very early stage of the development process.

Software requirements are often classified into two types: functional and non-functional. Functional requirements are statements of services or functions of software, while non-functional requirements are constraints on services or functions, such as performance, usability, operating platform, and software process... This classification impacts how requirements are processed during analysis, and validation activities. The requirement classification is often done manually in the software process. Thus it could require many developers' efforts, and they may fail to identify the critical requirements.

Recently, several studies have been conducted in this area to reduce efforts and help to improve the quality of the software requirements. One of them is automatically distinguishing functional from non-functional requirements [1-5, 13] or identifying non-functional [9, 12] or categorizing non-functional requirements [6, 7, 8] or prioritizing requirements [10, 11]. These works concentrate mostly on studying requirement specifications in English by using various machine learning techniques. The results are quite interesting. However, automatically classifying software requirements in Vietnamese has not been focused in the literature. Besides, no dataset for this problem has been built yet.

That is why we are motivated to study the application of machine learning algorithms in classifying software requirements in Vietnamese into functional and non-functional categories and evaluating their performance. In this paper, we first build a dataset consisting of functional and non-functional requirements, then propose a methodology for classifying requirements, select different machine learning algorithms for experimentation, and evaluate the experimental results.

The paper is organized as follows. Section I introduces the motivation of the study. The related works are presented in Section II. The methodology is described in Section III. Section IV is reserved for experimentation. The results are discussed in Section V. Finally, the paper finishes with the

conclusion and future works.

II. Related works

The requirements classification is often manually done late in the software process. Thus, automatic classification of software requirements into functional and non-functional requirements is helpful. This task-based on machine learning has been examined in English requirements.

In [1], the authors used the SVM for requirements classification. The study combines SVM and linguistic features as functional and non-functional requirements. It reaches a recall and precision of about 92% for both classes. This approach was developed and expanded to 17 linguistic features [3].

Another approach for preprocessing requirements was proposed in [2]. It standardizes and normalizes requirements before applying classification algorithms. The experimental results show that using this approach improves the efficiency of algorithms Latent Dirichlet Allocation, Biterm Topic Modeling, K-means, and Naïve Bayes in classifying requirements.

Software requirements usually vary in wording and style. Thus, the efficiency of machine learning algorithms reduces in the case used for unseen projects. To deal with the problem, using the transfer learning capabilities of BERT for requirements classification is proposed in [4]. This approach was used for different tasks in the domain of requirements classification. Results are similar or better (F1-scores up to 94%) on both seen and unseen projects for classifying functional and non-functional requirements.

In [5], the authors examined the problem of combining feature extraction techniques and machine learning algorithms for classifying software requirements. The feature extraction techniques are Bag of Words and Term Frequency - Inverse Document. The algorithms used for classification are LR, SVM, NB, and k-Nearest Neighbors. The experimental results show that the combination of Term Frequency - Inverse Document and LR has the best efficiency, with an F-measure of 91%.

In Vietnam, the software industry has been developing quickly in recent years. The total revenue of the software industry reached 5.44 billion U.S. dollars in 2020 [14]. However, the automatic classification of Vietnamese software requirements has not been investigated in the literature.

III. Methodology

This section presents the details of our proposed methodology for automatically classifying Vietnamese software requirements. The classification of requirements is performed in four phases. As illustrated in Figure 1, software requirements are pre-processed, and then features are extracted. Finally, five machine learning algorithms are performed and evaluated.

1. Data pre-processing

The most important phase in any data-driven project is obtaining quality data. Without these pre-processing steps, the results of a project can easily be biased or wholly misunderstood. This phase is the process of normalizing data and removing elements that are not meant for text classification. In this paper, we realize the following works for data pre-processing:

- *Removing HTML tags.* To build the dataset for this study, requirements were collected from websites. They may contain HTML tags; thus, it is necessary to remove these tags. Because HTML tags could lead to a bad result in text classification, we use regex in Python to remove these tags.
- *Normalizing Unicode encoding.* In Vietnamese, there are several forms of Vietnamese encoding, but the most popular ones are the combination Unicode and precomposed Unicode. Using different encoding standards in the same texts can lead to a problem: when putting a word into a training model, it will interpret that word as different words even though humans see it as the same. Our solution for this case is converting texts to the precomposed Unicode standard.
- *Standardizing typing styles.* Different typing styles could lead to different words in text processing. For instance, “òa” and “oà” are different. We develop a Python function to convert texts to the same typing style.
- *Converting to lower texts.* This work is essential because this feature does not contribute to the text classification result. Converting texts to lower cases also reduces features and increases the accuracy of the model.
- *Deleting special characters in texts.* Punctuations, numbers, and other special characters do not contribute to the result of text classification, so we remove them.
- *Removing stop words.* These words do not contribute to the classification result, thus we also develop a Python function to remove them from texts.
- *Tokenizing words.* There are currently several Vietnamese natural language processing toolkits such as VnCoreNLP, underthesea,... In this study, we use Python programming languages to develop models, therefore underthesea library, which is a suite of open-source Python modules, was used to tokenize the Vietnamese words.

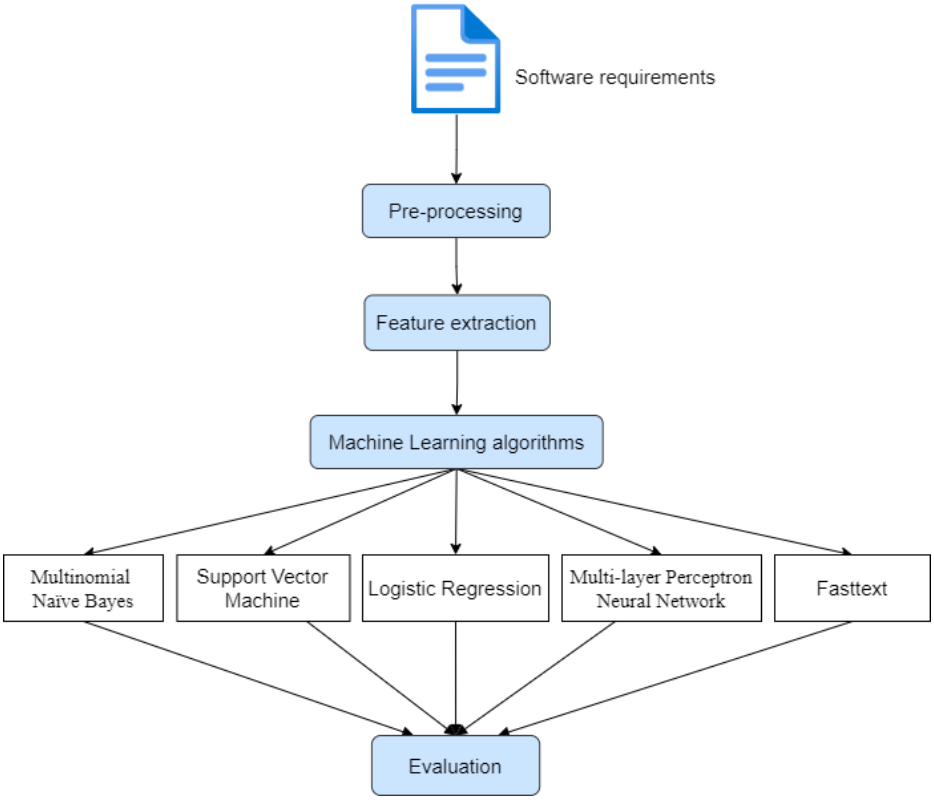


Figure 1. Schema of classifying software requirements in Vietnamese

2. Feature extraction

Machine learning techniques work on numeric features; therefore, we need to present them in numbers to be understood by the algorithms to deal with textual data. There are currently three popular ways to extract features from text documents: Bag of Words, Term Frequency - Inverse Document Frequency, and Word Counts. This uses word counts techniques with CountVectorizer of Scikitlearn library to convert texts to vector representations.

3. Classification

Supervised machine learning algorithms are used to classify software requirements into functional or non-functional categories. This study applies five algorithms as follows.

- *Naïve Bayes.* It is a probabilistic learning strategy often utilized in natural language processing. NB computes each tag's probability and gives the tag the highest probability as the outcome.
- *Support Vector Machine.* SVM is a supervised machine learning algorithm used for classification and regression challenges. Each data item is plotted as a point in n-dimensional. The value of each feature is the value of a specific coordinate. Then, the category is done by finding the hyperplane that differentiates well the two classes.
- *Logistics Regression.* LR is a classification algorithm that assigns observations to a discrete set of classes. It is the baseline supervised machine learning algorithm close to neural networks.
- *Multi-layer Perceptron Neural Network.* MLP includes at least three layers: an input layer, a hidden layer, and an output layer. MLP uses a supervised learning technique named backpropagation for training. Its multiple layers and nonlinear activation distinguish MLP from a linear perceptron.
- *FastText.* FastText is a method for encoding words as numeric vectors. It is much faster and easier to maintain than deep neural networks like BERT. Pretrained FastText embedding is utilized for solve problems, for example, named entity recognition or text classification.

4. Evaluation

To evaluate the ability to distinguish functional requirements and non-functional requirements, we use the Precision, Recall, and F1-score metrics. Before we get into precision and recall, we present True positive, True

TABLE I
EXPERIMENTAL RESULTS OF FIVE ALGORITHMS

Algorithms	Precision	Recall	F1-score	Training time (seconds)
Naïve Bayes	0.929	0.935	0.932	0.0172
Logistics regression	0.907	0.935	0.932	0.0259
Support Vector Machine	0.895	0.903	0.888	0.0168
Multi-layer Perceptron Neural Network	0.578	0.578	0.578	0.0153
FastText	0.755	0.755	0.755	1.5640

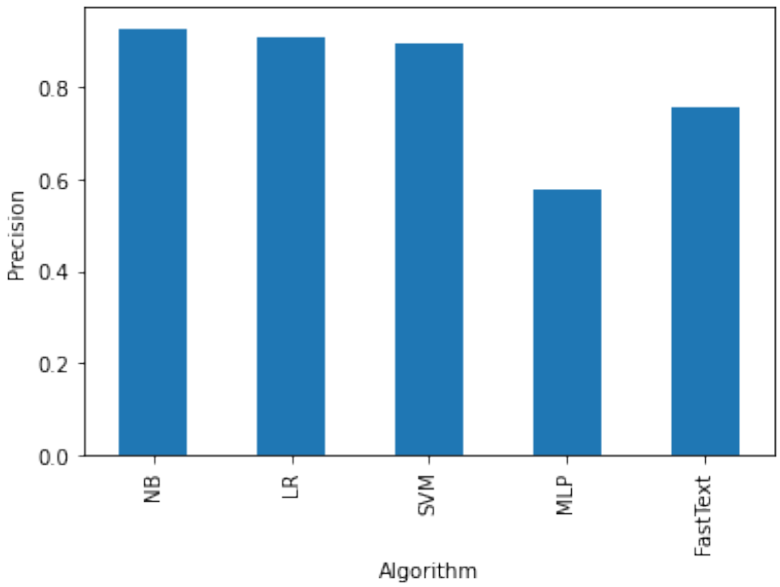


Figure 2. Algorithms comparison in terms of precision

negative, False positive, and False negative outcome classifications. True positives are understood positive results that the algorithm predicted correctly. True negatives are negative results that the algorithm predicted correctly. False positives are positive results that the algorithm predicted incorrectly. False negatives are negative results that the algorithm predicted incorrectly. By TP, TN, FP, and FN, we denote True positives, True negatives, False positives, and False negatives, respectively.

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It is determined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall is the ratio of correctly predicted positive observations to all observations in an actual class. Recall is computed as the following:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

F1-score is the weighted average of Precision and Recall. It is calculated as follows:

$$F_1 - Score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{3}$$

IV. EXPERIMENT AND ANALYSIS

1. Dataset

Since there has not been an available software requirement dataset in Vietnamese yet, we firstly built a testing sample dataset. It contains 222 requirements collected from four different software projects on the Internet. The requirements are in the form of a sentence or a short paragraph, such as "Sản phẩm phải có bảng màu và phông chữ nhất quán". In this dataset, we labeled 131 requirements as functional and 91 requirements as non-functional.

2. Experimental setups

This study conducts experiments with the requirements classification in Vietnamese with algorithms NB, SVM,

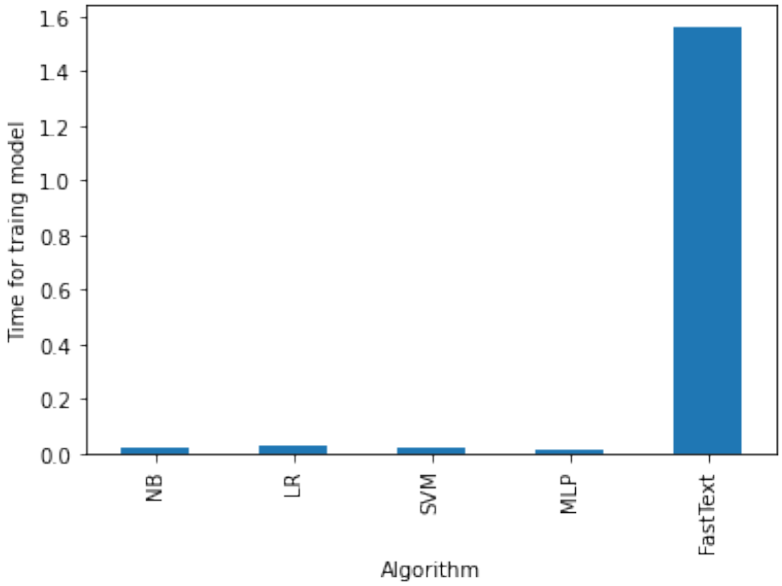


Figure 3. Algorithms comparison in terms of time for training model

LR, MLP, and FastText. These five algorithms all execute on the same data set that we built. We used the underthesea library to pre-process Vietnamese texts and the Scikit-learn library for five algorithms.

3. Experimental results

In this section, we evaluate the experimental results of five algorithms in the classification of requirements in Vietnamese. Table 1 shows the performance of each classification method.

Figure 2 shows a significant difference in precision value between the algorithms. NB has the best performance with 92.9% precision, and MLP is the worst with only 57.8%.

In terms of time for training models with the same training dataset size, FastText takes much more time than the other models, with 1.0629 seconds, while those of the others are only 0.01 – 0.03 seconds. We can see this difference clearly in Figure 3. Besides, the precision of FastText is only 75.5%.

We consider the remaining four models. The training time of MLP is the smallest with 0.0172 seconds, the second is SVM with 0.0168 seconds, the third is NB with 0.0172 seconds, and the fourth is LR with 0.0259 seconds. The difference between these fourth training times is not significant. Besides, NB has the best performance in all algorithms with 92.9% precision. We can assume that NB is best in classifying software requirements written in Vietnamese in both performance and time of model training.

V. Discussion

This section provides a discussion of the results in terms of the performance and how the results of this study can be useful for industry practitioners. In this era of software development, requirement analysis could generate a large amount of unstructured software specification documents during the elicitation process. The work in this paper is concerned with the development of models to classify functional and non-functional requirements in Vietnamese. The results of this study could help researchers as well as practitioners from the software industry find the type of requirements based on their description in the early phases of software development. It is a factor that contributes to software quality.

The experimental results show that NB and LR are the best supervised machine learning methods for classifying software requirements in Vietnamese into functional and non-functional categories. The efficiency of these two algorithms can be further improved by more suitable data representation.

Determining whether these two algorithms are efficient for Vietnamese texts in other fields requires much research with datasets in each field. It is also an issue that we will delve into in future research.

VI. Conclusion

Classification of software requirements into functional and non-functional is critical in the software process. In this study, we have presented a machine learning approach to classify functional and non-functional requirements in

Vietnamese in order to support the requirement elicitation process. We first built a dataset containing functional and non-functional requirements and then labeled them. After pre-processing the requirements, we performed five machine learning algorithms, including NB, SVM, LR, MLP, and FastText for the experiment. The experimental results show that the NB algorithm gives the best performance for distinguishing the requirements.

REFERENCES

[1] Zijad Kurtanovic, Walid Maalej, "Automatically Classifying Functional and Non-Functional Requirements Using Supervised Machine Learning," *IEEE International Conference on Requirements Engineering*, 2017, pp. 490-495.

[2] Zahra Shakeri Hossein Abad, Oliver Karras, Parisa Ghazi, Martin Glinz, Guenther Ruhe, Kurt Schneider, "What Works Better? A Study of Classifying Requirements," *25th IEEE International Requirements Engineering Conference*, 2017, pp. 496-501.

[3] Fabiano Dalpiaz, Davide Dell'Anna, Fatma Başak Aydemir, Sercan Çevikol, "Requirements Classification with Interpretable Machine Learning and Dependency Parsing," *IEEE International Conference on Requirements Engineering*, 2019, pp. 142-152.

[4] Tobias Hey, Jan Keim, Anne Koziolek, Walter F. Tichy, "NoRBERT- Transfer Learning for Requirements Classification," *IEEE 28th International Requirements Engineering Conference*, 2020.

[5] Edna Dias Canedo, Bruno Cordeiro Mendes, "Software Requirements Classification Using Machine Learning Algorithms," *Entropy*, vol. 22, 2020.

[6] Rajesh Kumar Gnanasekaran, Suranjan Chakraborty, Josh Dehlinger, Lin Deng, "Using recurrent neural networks for classification of natural language-based non-functional requirements," *Proceedings of REFSQ-2021 Workshops*, 2021.

[7] Rajni Jindal, Ruchika Malhotra, Abha Jain, Ankita Bansal, "Mining Non-Functional Requirements using Machine Learning Techniques," *e-Informatica Software Engineering Journal*, vol. 15, 2021.

[8] Nouf Rahimi, Fathy Eassa, Lamiaa Elrefaei, "An Ensemble Machine Learning Technique for Functional Requirement Classification," *Symmetry*, vol. 12 (10), 2019.

[9] Agustin Casamayor, Daniela Godoy, Marcelo Campo, "Identification of non-functional requirements in textual specifications: A semi-supervised learning approach," *Information and Software Technology*, vol. 52 (4), 2010.

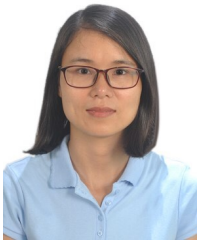
[10] Anna Perini, Angelo Susi, Paolo Avesani, "A Machine Learning Approach to Software Requirements Prioritization," *IEEE Transactions on Software Engineering*, vol. 9 (4), 2013.

[11] Lorijn van Rooijen, Frederik Simon Baumer, Marie Christin Platenius,Michaela Geierhos, Heiko Hamann, Gregor Engels, "From User Demand to Software Service: Using Machine Learning to Automate the Requirements Specification Process," *IEEE 25th International Requirements Engineering Conference Workshops*, 2017, pp. 379-385.

[12] Tetsuo Tamai, Taichi Anzai, "Quality Requirements Analysis with Machine Learning," *Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering*, 2018, pp. 241-248.

[13] Law Foong Li, Nicholas Chia Jin-An, Zarinah Mohd Kasirun, "An Empirical Comparison of Machine Learning Algorithms for Classification of Software Requirements," *International Journal of Advanced Computer Science and Applications*, vol. 10, 2019.

[14] Minh-Ngoc Nguyen, "Revenue of software industry in Vietnam 2016-2020", *Statista*, Jan 24, 2022 https://www.statista.com/statistics/1193528/vietnam-software-industry-revenue/



Thi Nham Cao graduated master program of software engineering at VNU University of Engineering and Technology in 2010. Now she is working at Faculty of Statistics and Informatics of Danang University of Economics. Her reseach interests are Machine Learning, Explainable Artificial Intelligence.

Dai Tho Dang received the Master of Computer Sciences from the Nice Sophia Antipolis, France, in 2010, and the Ph.D. degree from Yeungnam University, Republic of Korea, in cooperation with the Wrocław University of Science and Technology, Poland, in 2020. His research interests include Collective Intelligence, Evolutionary Computation, Deep Learning, and NLP. He is an active reviewer of IEEE Transactions on Cybernetics and Applied Intelligence.

Le Thi My Hanh is currently a lecturer of the Information Technology Faculty, University of Science and Technology, Danang, Vietnam. She gained M.Sc. degree in 2004 and the Ph.D. degree in Computer Science at the University of Danang in 2016. Her research interests are about software engineering and more generally application of heuristic techniques to problems in software engineering.
Email: ltmhanh@dut.udn.vn

Thanh Binh Nguyen graduated in Information Technology from the University of Danang - University of Science and Technology in 1997. He received PhD. degree in Information Technology at Grenoble Institute of Technology, France in 2004. He has been qualified as Associate Professor since 2013. He is currently working at the University of Danang - Vietnam-Korea University of Information and Communication Technology. His research interests include software engineering and software quality.
Email: ntbinh@vku.udn.vn