ÁP DỤNG HỌC MÁY ĐỂ NÂNG CAO ĐỘ CHÍNH XÁC CHO DỰ ĐOÁN NGUY CƠ ĐA DI TRUYỀN VỚI DỮ LIỆU RỐI LOẠN PHỔ TỰ KỶ

APPLYING MACHINE LEARNING TO IMPROVE THE ACCURACY OF POLYGENIC RISK SCORES WITH AUTISM SPECTRUM DISORDER DATA

Trịnh Thị Xuân*[*], *Lê Thị Thanh Thuỳ*[†], *Tạ Văn Nhân*[‡], *Hoàng Đỗ Thanh Tùng*[§], *Trương Nam Hải*[¶], *Trần Đăng Hưng

Ngày tòa soạn nhận được bài báo: 03/11/2021
Ngày nhận kết quả phản biện đánh giá: 05/05/2022
Ngày bài báo được duyệt đăng: 26/05/2022

Tóm tắt: *Điểm nguy cơ đa di truyền (polygenic risk scores, PRS) là một giá trị ước lượng tương đối nguy cơ mắc bệnh dựa vào việc xác định tập hợp các biến dị di truyền ảnh hưởng. Trong những năm gần đây, đã có nhiều cố gắng đưa tính toán PRS ứng dụng vào lâm sàng, tuy nhiên việc lựa chọn các biến dị di truyền ảnh hưởng đến bệnh có độ chính xác chưa cao dẫn đến hiệu quả mô hình chưa đạt kỳ vọng. Trong nghiên cứu này, chúng tôi đã thực nghiệm các mô hình khác nhau để chọn ra tập hợp các biến dị cho giá trị dự đoán tốt nhất. Dữ liệu được sử dụng là dữ liệu trong các nghiên cứu tương quan toàn hệ gen (Genome-Wide Association Studies, GWAS) về rối loạn phổ tự kỷ (Autism Spectrum Disorder, ASD). Tập hợp các biến dị ban đầu được thu gọn bằng phương pháp nhóm và đặt ngưỡng (Clumping and Thresholding, «C + T»), hồi quy logistic phạt (Penalized Logistic Regression, PLR), và loại bỏ đặc trưng đệ quy dựa trên máy vec-tơ tựa (Support Vector Machine Recursive Feature Selection, SVM-RFE). Kết quả cho thấy phương pháp SVM-RFE đưa ra được một tập SNPs mà mô hình dự đoán đạt hiệu năng tốt nhất.*

Từ khóa: *Bệnh đa di truyền, điểm nguy cơ đa di truyền, GWAS, SNPs, mảng SNP, học máy, bệnh tự kỷ.*

Abstract: *Polygenic risk scores (PRS) are relative estimation values of disease risk based on identification of effect variant set. In recent years, there have been many attempts to apply PRS calculation to clinical practice, however, selection of genetic variants affecting*

---

* Khoa Công nghệ thông tin, Trường Đại học Mở Hà Nội

† Khoa Công nghệ thông tin, Trường Đại học Mở Hà Nội

‡ Công ty TNHH LOBI Việt Nam

§ Phòng Nghiên cứu hệ thống và quản lý, Viện Công nghệ Thông Tin, VAST

¶ Phòng Kỹ thuật di truyền, Viện Công nghệ Sinh học, VAST

diseases has not been accurate, leading to the model's performance not yet reached hope. In this study, we have implemented different models to choose the set of variants giving the best prediction. The data used were taken from Genome-Wide Association Studies (GWAS) of Autism Spectrum Disorder (ASD). Original set of variants was reduced by Clumping and Thresholding ("C + T"), Penalized Logistic Regression (PLR), and Recursive Feature Elimination based on Support Vector Machine (SVM-RFE). As a result, the SVM-RFE method gives a set of SNPs that the prediction model has the best performance.

Keywords: *Complex diseases, polygenic risk scores, GWAS, SNPs, SNP arrays, machine learning, autism.*

I. Đặt vấn đề

Hiện nay, các ứng dụng lâm sàng về dự đoán nguy cơ mắc bệnh di truyền thường tập trung vào các bệnh đơn gen hiếm gặp với nguy cơ cao mặc dù phần lớn nguy cơ mắc bệnh có bản chất là đa gen. Lý do là vì độ chính xác trong các dự đoán bệnh đa gen chưa cao bằng độ chính xác trong các dự đoán bệnh đơn gen [1]. Từ thực tế đó, các nhà khoa học đã có khá nhiều các nghiên cứu cải tiến dự đoán nguy cơ đa di truyền với mục đích đưa phương pháp này vào thực hành lâm sàng. Trước tiên, các phương pháp được phát triển để thu hẹp tập SNPs [2], [3], [4] sử dụng các kỹ thuật điều chỉnh/thu hẹp trong thống kê như LASSO hoặc hồi quy ridge (ridge regression) [3], hoặc sử dụng cách tiếp cận Bayes thông qua việc xác định phân phối [2], [4], [5]. Sau đó, 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) [6], dự đoán di truyền đa biến với ngưỡng trơn [7], xác định các điểm đánh dấu di truyền [8].

Tiếp nối nhóm phương pháp cải thiện độ chính xác trong việc xác định các biến dị ảnh hưởng thực sự đến bệnh, chúng tôi đã áp dụng một số phương pháp thuộc các nhóm rút gọn đặc trưng khác nhau như hồi quy logisic phạt (Penalized Logistic Regression, PLR) và loại bỏ đặc trưng đệ quy dựa trên máy vec-tơ tựa (Support Vector Machine Recursive Feature Selection, SVM-RFE). Dữ liệu được sử dụng là các mẫu rối loạn phổ tự kỷ (Autism Spectrum Disorder, ASD) bao gồm dữ liệu kiểu gen toàn hệ gen được truy xuất từ dữ liệu trao đổi tài nguyên di truyền bệnh tự kỷ (Autism Genetic Resource Exchange, AGRE) [9]. Dữ liệu sau quá trình QC được huấn luyện bằng các mô hình học máy khác nhau. Kết quả sau khi so sánh mô hình mô hình "Nhóm và đặt ngưỡng" ("C + T"), PLR, và SVM-RFE cho thấy mô hình sử dụng SVM-RFE cho hiệu năng tốt nhất với 100 SNPs.

Trong các phần tiếp theo của bài báo chúng tôi sẽ trình bày về tiền xử lý dữ liệu trong mục II, phương pháp trong mục III. Kết quả so sánh giữa các mô hình được trình bày ở mục IV. Cuối cùng, chúng tôi kết luận bài báo trong mục V.

II. Cơ sở lý thuyết

2.1. Dữ liệu

2.1.1. Dữ liệu cơ sở

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).

Ở đây, chúng tôi sử dụng dữ liệu thống kê tóm tắt gồm 9,112,386 SNPs được xây dựng trên hệ gen hg19 [10]. Trong đó, điểm thông tin của quá trình suy diễn thống kê (imputation information score) IN F OR > 0.7; tần số alen phụ (Minor Allele Frequency, MAF) > 0.01; hệ số di truyền h2G = 0.118 > 0.5. Dữ liệu này đảm bảo tuân thủ các tiêu chuẩn để đưa vào tính toán điểm nguy cơ đa di truyền[**]

2.1.2. Dữ liệu đích

Dữ liệu đích là dữ liệu GWAS ở cấp 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 trí của các SNPs 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 [11].

Chúng tôi sử dụng dữ liệu kiểu gen - kiểu hình của các mẫu rối loạn phổ tự kỷ (autism spectrum disorder, ASD)[††] được truy xuất từ dữ liệu trao đổi tài nguyên di truyền bệnh tự kỷ (Autism Genetic Resource Exchange, AGRE) [9]. Dữ liệu gồm ba tệp định dạng PLINK *.fam, *.bim, và tệp nhị phân *.bed với hệ gen tham chiếu hg17. Trong đó có 399,147 biến dị; 2,883 mẫu với 1,816 nam và 1,066 nữ, 1 cá thể chưa rõ (nhãn 1 được gán cho nam, nhãn 2 được gán cho nữ); 2,879 cá thể có kiểu hình, 4 cá thể không có kiểu hình (nhãn 1 được gán cho mẫu đối chứng, nhãn 2 được gán cho mẫu bệnh).

2.2. Kiểm soát chất lượng (QC)

Độ chính xác dự đoán của PRS phụ thuộc lớn vào chất lượng của dữ liệu cơ sở và dữ liệu đích. Cả hai tập dữ liệu thường được tiến hành QC với các tiêu chuẩn QC chung của GWAS [12], [13], [14], và QC cho từng loại dữ liệu [15].

1) QC dữ liệu cơ sở: Chúng tôi tiến hành QC tiêu chuẩn cho dữ liệu cơ sở với IN F O > 0.8, kiểm tra các SNPs trùng lặp, và loại bỏ các SNPs không rõ ràng. Sau quá trình này, dữ liệu cơ sở còn lại 7,301,379 biến dị.

2) QC dữ liệu đích: Dữ liệu đích được chuyển tọa độ từ hệ gen tham chiếu hg17 sang hệ gen tham chiếu hg19, có 34 biến dị không khớp tọa độ hoặc đã bị loại bỏ. Chúng tôi thực hiện QC tiêu chuẩn với tần số alen phụ M AF > 0.01; ngưỡng p-value từ kiểm định χ^2 hoặc kiểm định Fisher cho cân bằng Hardy-Weinberg hwe = 10^{-6}; loại bỏ các biến dị và cá thể có tỷ lệ kiểu gen bị thiếu với ngưỡng geno = 0.01, mind = 0.01. Ngoài ra, quá trình pruning được thực hiện để giữ lại các SNPs có tương quan thấp $r^2 < 0.25$. Trên thực tế, tỷ lệ dị hợp tử cao có thể do chất lượng mẫu thấp, tỷ lệ này thấp có thể do ảnh hưởng của giao phối cận huyết, vì vậy 74 cá thể được lọc ra để dữ liệu đạt được tỷ lệ dị hợp tử tốt nhất. Tiếp theo, 134,126 SNPs không khớp của dữ liệu đích so với dữ liệu cơ sở cũng được xác định nhờ phương pháp đảo ngược sợi DNA. Ngoài

** https://ipsych.dk/en/research/downloads/data-download-agreement-ipsych-pgc-asd-nov2017/thank-you/

†† https://figshare.com/articles/dataset/Autism_GWAS_data/14253230

ra, dữ liệu không bao gồm 26 cá thể có sai khác về giới tính sinh học và 1,446 cá thể có quan hệ gần. Dữ liệu đích sau QC bao gồm 264,987 biến dị; 1,138 mẫu, trong đó có 142 mẫu bệnh, 996 mẫu đối chứng.

III. Phương pháp nghiên cứu

3.1. Tính điểm nguy cơ đa di truyền (Polygenic Risk Score, 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 [16]. Công thức mặc định để tính PRS trong PLINK [11] 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 SNPs của mẫu j là N; tổng số SNPs không thiếu được quan sát 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 ($M AF_i.P$) được sử dụng thay thế $G_{i,j}$.

3.2. Tính toán phân tầng quần thể

Sự phân tầng quần thể có thể được hiểu là sự hiện diện của nhiều quần thể con trong dữ liệu, ví dụ các cá nhân có nguồn gốc dân tộc khác nhau. Vì mức độ ảnh hưởng ứng với tần số alen phụ có thể khác nhau đối với các quần thể khác nhau nên việc tính toán PRS cho các cá thể trở nên không chính xác. Do đó các thành phần chính (Principal Components, PCs) đại diện cho phân tầng quần thể được đưa vào mô hình dự đoán để giảm sự sai lệch của dữ liệu GWAS. Tuy nhiên, vấn đề nằm ở chỗ xác định số lượng thành phần chính. Theo kinh nghiệm, các nhà nghiên cứu thường chọn số PCs là 10 [17], [14]. Một cách khác để chọn số lượng PCs thích hợp là thực hiện GWAS trên kiểu hình đang nghiên cứu với số lượng PCs khác nhau. Sau đó, phân tích hồi quy điểm mất cân bằng liên kết (LD Score regression, LDSC) có thể được thực hiện trên tập hợp các thống kê tóm tắt GWAS, số PCs mà cấu trúc quần thể được kiểm soát chính xác nhất là số PCs mà hệ số tự do của LDSC gần 1 nhất [15].

Trong bài báo này, chúng tôi đề xuất phương pháp xác định số lượng thành phần chính bằng thuật toán phân cụm k-means clustering, một thuật toán học không giám sát.

Giả thuyết 1. *Số cụm tối ưu, sau khi huấn luyện* mô hình dựa trên chính đặc điểm của dữ liệu, tương đương với số lượng các PCs.

Các PCs với số lượng khác nhau được đưa vào huấn luyện, số lượng PCs cho kết quả dự đoán tốt nhất chính bằng số cụm tối ưu đã chứng minh cho giả thuyết trên.

3.3. Phương pháp nhóm và đặt ngưỡng ("C+T")

Phương pháp truyền thống thường được sử dụng là "Nhóm và Đặt ngưỡng" ("Clumping and Thresholding", hay còn gọi là "C+T"). Các SNPs được "Nhóm" ("Clumping", C) bởi công cụ PLINK để chọn ra các SNPs 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 SNPs gần đó. Ở đây, chúng tôi duyệt qua tất cả các

SNPs, coi chúng đều là các SNPs chỉ mục (clump − p1 = 1). Sau đó nó loại bỏ các SNPs với khoảng cách di truyền clump − kb = 250 (kb) nếu mối tương quan giữa chúng $r^2 > 0.1$ [19]. Như vậy, bước Clumping giúp loại bỏ dữ liệu dư thừa do mất cân bằng liên kết (LD) gây ra. Sau quá trình "Clumping" ta lựa chọn được 67,188 SNPs cho tính toán PRS.

Phương pháp "Đặt ngưỡng" ("Thresholding") chọn ra tập hợp SNPs của GWAS tương quan với kiểu hình dưới các ngưỡng P-value khác nhau trong các bán đoạn (0, 10^{-5}], (0, 10^{-3}], (0, 0.0225], (0, 0.05], (0, 0.1], (0, 0.2], (0, 0.3], (0, 0.4], (0, 0.5]. Các tập SNPs tương ứng với các ngưỡng sẽ được sử dụng để tính toán PRS.

Ngoài ra, mô hình dự đoán còn có sự đóng góp của hiệp biến giới tính và các thành phần chính của tập đích được tính toán dựa trên phân tầng quần thể. Mô hình ban đầu (mô hình 1), chưa tính đến dữ liệu kiểu gen, được coi như mô hình vô hiệu (null) để so sánh độ chính xác dự đoán với các mô hình có tính đến các tập SNPs với các ngưỡng P-value khác nhau (mô hình 2).

$$log(\frac{\hat{p}}{1-\hat{p}}) = \gamma * PC + \sigma * SEX \quad (1)$$

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta * X + \gamma * PC + \sigma * SEX \quad (2)$$

trong đó $\hat{p} = P(y = 1)$ với y là kiểu hình của bệnh ($y \in \{1, 2\}$), PC là ma trận mà các cột là các thành phần chính, SEX là hiệp biến giới tính, X là vec tơ mà mỗi thành phần là điểm nguy cơ đa di truyền tương ứng của một cá thể được tính tương ứng với một tập SNPs nào đó. Vec tơ X tương ứng với P-value nào cho ra được độ chính xác dự đoán của mô hình cao nhất và cao hơn độ chính xác của mô hình vô hiệu thì tập hợp các SNPs tương ứng với ngưỡng đó sẽ được xác định là có ảnh hưởng đến bệnh.

3.1. Phương pháp sử dụng hồi quy logistic phạt (Penalized Logistic Regression, PLR)

Mô hình hồi quy logistic phạt (Penalized Lo-gistic Regression, PLR) [20] chứa hai hàm điều chỉnh. Hàm điều chỉnh L2 ("Ridge") có tác dụng thu nhỏ các hệ số và hàm điều chỉnh L1 ("LASSO" [21]) 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" [22]) rất hiệu quả trong trường hợp số lượng SNPs 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 $z_i = 1/(1 + \exp(-(\beta_0 + x_i^T \beta)))$, x biểu diễn kiểu gen và các hiệp biến (các thành phần chính và giới tính), y là tình trạng bệnh, λ và α là hai siêu tham số điều chỉnh.

3.2. Phương pháp loại bỏ đặc trưng đệ quy dựa trên máy vec-tơ tựa (Support Vector Machine Recursive Feature Elimination, SVM-RFE)

Mục đích chính của SVM-RFE là tính toán các trọng số được xếp hạng với

tất cả các đặc trưng và sắp xếp các đặc trưng theo các vec-tơ trọng số. SVM-RFE là quá trình lặp của việc loại bỏ ngược các đặc trưng [23].

• Sử dụng tập dữ liệu hiện tại để huấn luyện mô hình phân loại.

• Tính toán trọng số cho tất cả các đặc trưng.

• Xóa các đặc trưng với trọng số nhỏ nhất.

Chi tiết thuật toán SVM-RFE được đưa ra bởi Isabelle Guyon và các đồng nghiệp trong một nghiên cứu chọn gen cho phân loại ung thư [24].

3.3. Đánh giá hiệu năng của mô hình

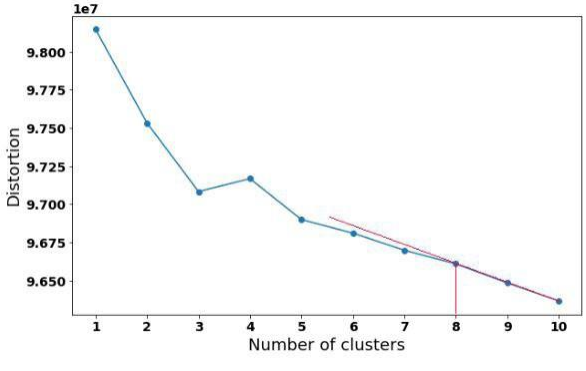
Để đánh giá hiệu năng của mô hình ta sử dụng đường cong đặc tính (Receiver Operating Characteristic Curve, ROC) biểu diễn tương quan giữa dương tính giả và dương tính thật với các ngưỡng nào đó. Hiệu năng mô hình được đánh giá thông qua giá trị diện tích dưới đường ROC (Area Under the Curve, AUC), giá trị này nằm trong khoảng (0, 1), AUC càng lớn thì hiệu năng của mô hình càng cao.

Mặt khác, nhằm đảm bảo các mô hình khác nhau được đánh giá trên cùng một tập kiểm thử, dữ liệu ban đầu được chia ngẫu nhiên thành hai tập, tập huấn luyện với 80% dữ liệu, tập kiểm thử với 20% dữ liệu. Với mục đích giảm sự quá khớp (overfiting), kỹ thuật đánh gía chéo k lần (k-fold cross validation) được áp dụng với tập huấn luyện, đây ta chọn k = 5. Với một mô hình nhất định, các siêu tham số tối ưu được xác định tương ứng với AUC trung bình cao nhất khi mô hình khớp với 5 tập dữ liệu khác nhau. Trong bài báo này, tập SNPs sau khi được thu gọn nhờ mô hình PLR sẽ được tính toán PRS, sau đó được huấn luyện và đánh giá tương tự như phương pháp "C+T". Điều này cho thấy rõ ràng hiệu năng của hai mô hình khi được đánh giá với cùng một phương pháp đánh giá trên cùng một tập kiểm thử.

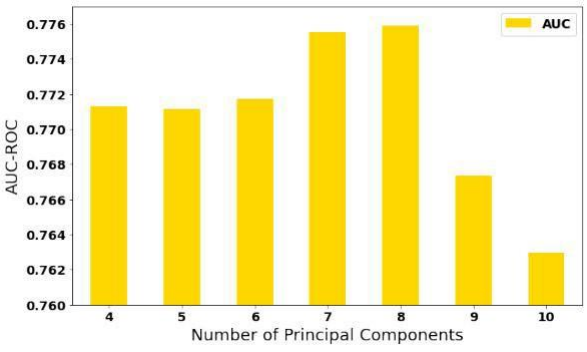
IV. Kết quả

4.1. Phân tầng quần thể

Phương sai của tâm các cụm (distortion) được tính toán khi số cụm tăng từ 1 đến 10. Từ cụm thứ 8 trở đi, phương sai của tâm các cụm giảm tuyến tính. Theo phương pháp khuỷu tay, ta chọn số cụm tối ưu bằng 8 (xem hình 1). Ta tiếp tục kiểm tra giả thuyết 1 bằng thực nghiệm với phương pháp "C + T". Số lượng PCs thay đổi từ 4 đến 10 được đưa vào mô hình 2. Với 8 PCs, AUC lớn nhất đạt 0.776 (Xem hình 2). Điều này chứng tỏ với dữ liệu hiện tại và phương pháp "C + T", ta có thể lựa chọn số lượng PCs chính bằng số cụm tối ưu của dữ liệu kiểu gen.



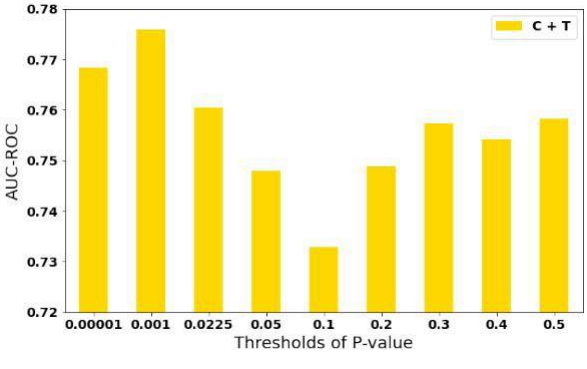
Hình 1. Xác định số cụm tối ưu. Số cụm tối ưu là 8 khi phương sai tâm của các cụm giảm tuyến tính.



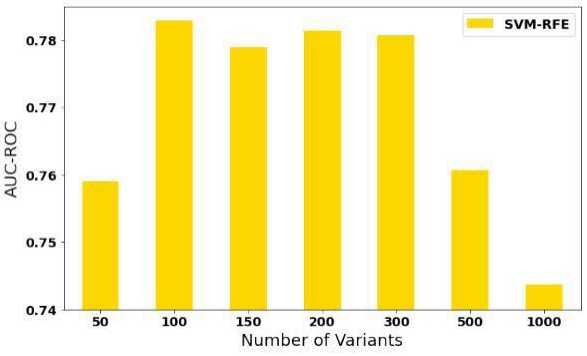
Hình 2. Hiệu năng của mô hình tương ứng với số thành phần chính. Với số thành phần chính là 8, AUC của mô hình "C+T" cao nhất đạt 0.776.

4.2. So sánh hiệu năng của các mô hình

Đối với phương pháp «C + T», mô hình đạt AUC lớn nhất bằng 0.776 với tập hợp 262 SNPs có P −values ≤ 0.001 (Xem hình 3). Với phương pháp PLR, ta chọn được 215 SNPs sau khi rút gọn đặc trưng, AUC của mô hình đặt 0.75, thấp hơn so với mô hình "C + T". Khi các đặc trưng được lựa trọn bằng mô hình SVM-RFE, AUC đạt giá trị lớn nhất là 0.783 với 100 SNPs (Xem hình 4). So sánh ba mô hình "C + T", PLR, và SVM-RFE cho thấy phương pháp SVM-RFE có thể tìm được tập hợp các SNPs với giá trị AUC cao nhất (Xem hình 5).



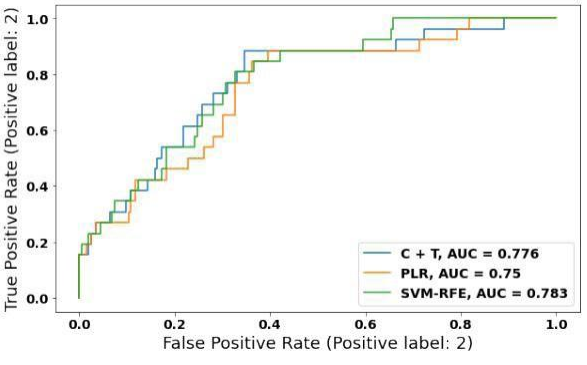
Hình 3. Hiệu năng của mô hình "C + T". ACU đạt 0.776 tại 262 SNPs với P − value ≤ 0.001.



Hình 4. Hiệu năng của mô hình SVM-RFE. AUC đạt 0.783 với 100 SNPs.

V. Kết luận

Nhằm nâng cao độ chính xác cho dự đoán nguy cơ đa di truyền của rối loạn phổ tự kỷ, chúng tôi đã tiến hành một cách đầy đủ các bước QC dữ liệu theo các nghiên cứu trước đây cũng như sử dụng các mô hình học máy khác nhau để lựa chọn được tập SNPs cho kết qủa dự đoán tốt nhất.



Hình 5. So sánh hiệu năng của các mô hình "C + T", PLR, và SVM-RFE. Mô hình SVM-RFE cho giá trị AUC cao nhất là 0.783

Mô hình truyền thống "C + T" vẫn cho thấy sự đơn giản nhưng hiệu quả hơn mô hình PLR trong trường hợp áp dụng với dữ liệu của bài báo. Tuy nhiên, mô hình PLR cho phép ta chọn biến một cách tự động ngay trong quá trình huấn luyện mô hình. Do đó, kết quả của PLR không phụ thuộc nhiều vào kinh nghiệm như

việc lựa chọn các ngưỡng P − values của phương pháp "C + T". Đặc biệt, phương pháp chọn biến thông qua việc xếp hạng các đặc trưng của SVM-RFE giúp ta thu được tập hợp SNPs cho dự đoán tốt nhất.

Từ quá trình thực nghiệm tính toán PRS sử dụng các mô hình học máy, một số gợi mở cũng như phương pháp mới sẽ được tiếp tục cải thiện để tăng độ chính xác cho dự đoán nguy cơ đa di truyền và mở rộng phạm vi ứng dụng của PRS trong lâm sàng.

Lời cảm ơn:

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

Tài liệu tham khảo:

[1]. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, P. Natarajan, E. S. Lander, S. A. Lubitz, P. T. Ellinor, and S. Kathiresan, "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations," *Nature Genetics*, vol. 50, no. 9, pp. 1219–1224, Sep. 2018.

[2]. J. Vilhjálmsson, J. Yang, H. K. Finucane, A. Gu-sev, S. Lindstrom,¨ S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, T. Hayeck, H.-H. Won, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Vari-ants in Breast Cancer (DRIVE) study, S. Kathiresan, M. Pato, C. Pato, R. Tamimi, E. Stahl, N. Zaitlen, B. Pasaniuc, G. Belbin, E. E. Kenny, M. H. Schierup, P.De Jager, N. A. Patsopoulos, S. McCarroll, M. Daly, S. Purcell, D. Chasman, B. Neale, M. Goddard, P. M. Visscher, P. Kraft, N. Patterson, and A. L. Price, "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores**,**" *American Journal of Human Genetics*, vol. 97, no. 4, pp. 576–592, Oct. 2015.

[3]. T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham, "Polygenic scores via penalized regression on summary statistics," *Genetic Epidemiology*, vol. 41, no. 6, pp. 469–480, Sep. 2017.

[4]. T. Ge, C.-Y. Chen, Y. Ni, Y.-C. A. Feng, and J. W. Smoller, "Polygenic prediction via Bayesian regression and continuous shrinkage priors," *Nature Communica-tions*, vol. 10, no. 1, p. 1776, Apr. 2019.

[5]. P. J. Newcombe, C. P. Nelson, N. J. Samani, and F. Dudbridge, "A flexible and parallelizable approach to genome-wide polygenic risk scores," *Genetic Epidemiology*, vol. 43, no. 7, pp. 730–741, 2019.

[6]. G. Paré, S. Mao, and W. Q. Deng, "A machine-learning heuristic to improve gene score prediction of polygenic traits," *Scientific Reports*, vol. 7, no. 1, p. 12665, Oct. 2017.

[7]. Y. Takahashi, M. Ueki, G. Tamiya, S. Ogishima, K. Ki-noshita, A. Hozawa, N. Minegishi, F. Nagami, K. Fuku-moto, K. Otsuka, K. Tanno, K. Sakata, A. Shimizu, M.Sasaki, K. Sobue, S. Kure, M. Yamamoto, and H. Tomita, "Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic predic-tion of polygenic psychiatric phenotypes," *Translational Psychiatry*, vol. 10, no. 1, pp. 1–11, Aug. 2020.

[8]. Vlachakis, E. Papakonstantinou, R. Sagar, F. Ba-copoulou, T. Exarchos, P. Kourouthanassis, V. Kary-otis, P. Vlamos, C. Lyketsos, D. Avramopoulos, and V. Mahairaki, "Improving the Utility of Polygenic Risk Scores as a Biomarker for Alzheimer's Disease," *Cells*, vol. 10, no. 7, p. 1627, Jun. 2021.

[9]. H. Geschwind, J. Sowinski, C. Lord, P. Iversen, J. Shestack, P. Jones, L. Ducat, and S. J. Spence, "The Autism Genetic Resource Exchange: A Resource for the Study of Autism

and Related Neuropsychiatric Conditions," *American Journal of Human Genetics*, vol. 69, no. 2, pp. 463–466, Aug. 2001.

[10]. J. Grove, S. Ripke, T. D. Als, M. Mattheisen, R. K. Wal-ters, H. Won, J. Pallesen, E. Agerbo, O. A. Andreassen, R. Anney, S. Awashti, R. Belliveau, F. Bettella, J. D. Buxbaum, J. Bybjerg-Grauholm, M. Bækvad-Hansen, F. Cerrato, K. Chambert, J. H. Christensen, C. Church-house, K. Dellenvall, D. Demontis, S. De Rubeis, B. Devlin, S. Djurovic, A. L. Dumont, J. I. Goldstein, B. S. Hansen, M. E. Hauberg, M. V. Hollegaard, S. Hope, D. P. Howrigan, H. Huang, C. M. Hultman, I. Klei, J. Maller, J. Martin, A. R. Martin, J. L. Moran, I.Nyegaard, T. Nærland, D. S. Palmer, A. Palotie, C.B. Pedersen, M. G. Pedersen, T. dPoterba, J. B.Poulsen, B. S. Pourcain, P. Qvist, K. Rehnstrom,¨ A. Re-ichenberg, J. Reichert, E. B. Robinson, K. Roeder, P. Roussos, E. Saemundsen, S. Sandin, F. K. Satter-strom, G. Davey Smith, H. Stefansson, S. Steinberg, C. R. Stevens, P. F. Sullivan, P. Turley, G. B. Walters, X. Xu, K. Stefansson, D. H. Geschwind, M. Nordentoft, D. M. Hougaard, T. Werge, O. Mors, P. B. Mortensen, B. M. Neale, M. J. Daly, and A. D. Børglum, "Identification of common genetic risk variants for autism spectrum disorder," *Nature Genetics*, vol. 51, no. 3, pp. 431–444, Mar. 2019.

[11]. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. de Bakker, M. Daly, and P. Sham, "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, Sep. 2007.

[12]. C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," *Nature Protocols*, vol. 5, no. 9, pp. 1564–1573, Sep. 2010.

[13]. J. R. I. Coleman, J. Euesden, H. Patel, A. A. Folarin, S. Newhouse, and G. Breen, "Quality control, imputa-tion and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray," *Briefings in Functional Genomics*, vol. 15, no. 4, pp. 298–304, Jul. 2016.

[14]. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, and E. M. Derks, "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis," *International Journal of Methods in Psychiatric Research*, vol. 27, no. 2, p. e1608, Feb. 2018.

[15]. S. W. Choi, T. S.-H. Mak, and P. F. O'Reilly, "Tutorial: a guide to performing polygenic risk score analyses," *Nature Protocols*, vol. 15, no. 9, pp. 2759–2772, Sep. 2020.

[16]. J. Euesden, C. M. Lewis, and P. F. O'Reilly, "PRSice: Polygenic Risk Score software," *Bioinformat-ics*, vol. 31, no. 9, pp. 1466–1468, May 2015.

[17]. H. Zhao, N. Mitra, P. A. Kanetsky, K. L. Nathanson, and T. R. Rebbeck, "A Practical Approach to Adjusting for Population Stratification in Genome-wide Association Studies: Principal Components And Propensity Scores (PCAPS)," *Statistical applications in genetics and molecular biology*, vol. 17, no. 6, pp. /j/sagmb.2018.17. issue–6/sagmb–2017–0054/sagmb– 2017–0054.xml, Dec. 2018.

[18]. B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale, "LD Score regression distinguishes confounding from polygenicity in genome-wide asso-ciation studies," *Nature Genetics*, vol. 47, no. 3, pp. 291–295, Mar. 2015.

[19]. N. R. Wray, S. H. Lee, D. Mehta, A. A. E. Vinkhuyzen, F. Dudbridge, and C. M. Middeldorp, "Research re-view: Polygenic methods and their application to psy-chiatric traits," *Journal of Child Psychology and Psy-chiatry, and Allied Disciplines*, vol. 55, no. 10, pp. 1068–1087, Oct. 2014.

[20]. Privé, H. Aschard, and M. G. B. Blum, "Efficient Implementation of Penalized Regression for Genetic Risk Prediction," *Genetics*, vol. 212, no. 1, pp. 65–74, May 2019.

[21]. R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[22]. H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[23]. M.-L. Huang, Y.-H. Hung, W. M. Lee, R. K. Li, and B.-R. Jiang, "SVM-RFE Based Feature Selection and Taguchi Parameters Optimization for Multiclass SVM Classifier," *The Scientific World Journal*, vol. 2014, p. 795624, 2014.

[24]. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vec-tor Machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, Jan. 2002.

Địa chỉ tác giả: Khoa Công nghệ thông tin, Trường Đại học Mở Hà Nội

Email: trinhxuan@hou.edu.vn