Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 9-10/7/2015
DOI: 10.15625/vap.2015.000141

CẢI TIẾN PHƯƠNG PHÁP RỪNG NGẪU NHIÊN CÓ ĐIỀU HƯỚNG ĐỂ ÁP DỤNG CHO DỮ LIỆU SNP

Hoàng Thị Hà [1], Nguyễn Thanh Tùng[2]

[1]Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam
[2]Khoa Công nghệ thông tin, Trường Đại học Thủy lợi

htha@vnua.edu.vn, tungnt@tlu.edu.vn

TÓM TẮT *- Rừng ngẫu nhiên có hiệu quả với dữ liệu có số chiều vừa phải, khi số chiều lớn hơn thì vẫn hạn chế. Deng và Runger đã đề xuất phương pháp rừng ngẫu nhiên có điều hướng (GRRF, Pattern Recognition-2013) ưu tiên để chọn đặc trưng, tuy nhiên vẫn kém hiệu quả với các tập dữ liệu có số chiều rất lớn mà số mẫu ít, chẳng hạn dữ liệu đa hình đơn nucleotide* SNP *(Single Nucleotide Polymorphism) trên quy mô toàn bộ hệ gien. Trong bài báo này, chúng tôi đề xuất phương pháp đánh trọng số đặc trưng mới thay cho cách đánh trọng số của GRRF. Kết quả thực nghiệm trên 2 tập dữ liệu Parkinson (408.803 SNPs) và Alzheimer (380.157 SNPs) cho thấy phương pháp cải tiến này có hiệu quả hơn hẳn GRRF và các phương pháp hiện thời.*

Từ khóa *- Dữ liệu chiều cao, máy học, khai phá dữ liệu, rừng ngẫu nhiên*

I. ĐẶT VẤN ĐỀ

Đa hình đơn nucleotide (Single Nucleotide Polymorphism, SNP) là những biến thể trình tự DNA xảy ra khi một đơn nucleotide (A, T, C, hoặc G) trong trình tự bộ gien bị thay đổi và là loại biến thể di truyền phổ biến tạo nên sự khác biệt chủ yếu giữa các cá thể cùng loài. Kết quả của bản đồ gien người cho biết, đối với loài người, hơn 99% trình tự ADN là giống nhau, sự khác biệt chỉ chiếm nhỏ hơn 1%, trong đó các SNP chiếm phần lớn sự khác biệt. Vì vậy, trong y sinh, dữ liệu SNP có vai trò quan trọng trong chẩn đoán bệnh tật, sự kháng thuốc, những phản ứng khác nhau trong quá trình điều trị… [***1***] [***2***].

Những nghiên cứu liên kết mức toàn bộ hệ gen (Genome-wide association studies – GWAS) là một tiếp cận chuẩn để xác định được nhiều biến dị gien dẫn tới một số bệnh phức tạp [***3***]. Tuy nhiên, xét trên quy mô toàn bộ hệ gien số lượng SNP là vô cùng lớn. Dữ liệu SNP cần kiểm tra là dữ liệu về hàng trăm ngàn SNP được lấy mẫu từ vài nghìn thậm chí chỉ vài trăm cá thể, trong đó có rất nhiều các SNP không liên quan tới một loại bệnh cụ thể [***2***]. Do đó, dữ liệu SNP có số lượng thuộc tính lớn hơn nhiều so với dung lượng mẫu và chứa nhiều nhiễu. Vì vậy, việc xác định những nhóm SNP có ảnh hưởng lớn tới bệnh là một bài toán khó. Các phương pháp học máy hiện nay dựa trên hai lớp giải thuật tiêu biểu là máy học véc tơ hỗ trợ của Vapnik (SVM) [***4***] và rừng ngẫu nhiên của Breiman (RF) [***5***] đã khá thành công trong bài toán trích chọn đặc trưng và phân lớp các dữ liệu sinh học, nhưng khi áp dụng trên các tập dữ liệu SNP đối chứng (case-control) trên toàn hệ gien lại cho kết quả không tốt.

Rừng ngẫu nhiên (Random Forest – RF) [***5***] cho độ chính xác phân lớp cao khi so sánh với các thuật toán học có giám sát hiện nay bao gồm Boosting, Baging, các láng giềng gần nhất (Nearest neighbors), SVM, Neural Network, C45,... [***6***]. Tuy nhiên, tiếp cận cài đặt RF ban đầu chỉ cho kết quả tốt trên các dữ liệu có số chiều vừa phải và giảm đáng kể hiệu năng khi xử lý bài toán có số chiều rất cao, nhiều nhiễu, dung lượng mẫu ít và bài toán phân tích dữ liệu SNP trên toàn hệ gien là một trường hợp cụ thể. Bureau và cộng sự cho biết RF đạt kết quả tốt trên dữ liệu SNP đối chứng (case-control) với cỡ chỉ 42 SNPs [*7*]. RF cũng có thể áp dụng trên các tập dữ liệu giả lập với số lượng SNP không quá 1000 [***8***]. Nguyên nhân chính là trong quá trình xây dựng cây quyết định, tại mỗi nút, RF dùng phương pháp chọn ngẫu nhiên một tập con thuộc tính (có kích thước *mtry*) từ tập thuộc tính ban đầu để tìm thuộc tính phân hoạch tốt nhất phân tách nút. Do đó, khi xử lý với các dữ liệu nhiều nhiễu như SNP, RF có thể lựa chọn ngẫu nhiên nhiều SNP nhiễu vào không gian con thuộc tính dùng cho việc tách nút khi dựng cây, nên khả năng dự đoán của RF giảm sút. Vì vậy, RF nguyên bản ít khi được sử dụng cho phân tích dữ liệu SNP chiều cao mức toàn hệ gien.

Gần đây, Deng và Runger đã đề xuất phương pháp rừng ngẫu nhiên có điều hướng (Guided Regularized Random Forests-GRRF) [***9***] ưu tiên để trích chọn đặc trưng, giúp cải thiện quá trình lựa chọn thuộc tính và phân lớp khi xử lý dữ liệu chiều cao, nhiều nhiễu. Kết quả thực nghiệm của Deng và Runger trên các bộ dữ liệu gien cho thấy tiếp cận GRRF đạt kết quả phân lớp tốt hơn RF ban đầu và cho kết quả trích chọn đặc trưng tốt hơn một số phương pháp đã được biết đến như: LASSO, varSelRF, RRF [***9***]. Tuy nhiên, GRRF dựa vào độ đo sự quan trọng thuộc tính từ RF nguyên bản để tạo trọng số cho các thuộc tính. Theo phân tích của Kim và đồng nghiệp [***10***], Strobl và đồng nghiệp [***11***,***12***,***13***,***14***], RF có lỗi bias trong quá trình lựa chọn thuộc tính khi có xu hướng lựa chọn những thuộc tính chứa nhiều giá trị (multi-valued), chứa nhiều dữ liệu trống (missing value) nhưng chúng không tốt cho quá trình phân hoạch, do đó GRRF bị giảm đáng kể độ chính xác khi phân lớp với dữ liệu nhiều nhiễu và gặp hạn chế lớn khi phân tích dữ liệu SNP trên toàn hệ gien.

Bài báo này đề xuất một phương pháp đánh trọng số thuộc tính mới thay cho cách đánh trọng số của GRRF. Kết quả thực nghiệm trên các bộ dữ liệu SNP ở mức toàn hệ gien cho thấy tiếp cận này giúp RF đạt độ chính xác phân lớp tốt hơn hẳn GRRF, SVM và một số phương pháp cải tiến của RF trong thời gian gần đây.

Phần II trình bày tóm tắt tiếp cận GRRF và phân tích ưu nhược điểm của phương pháp này. Phần III trình bày phương pháp đề xuất cải tiến cho GRRF. Kết quả thực nghiệm sẽ được trình bày trong phần IV. Phần V kết luận và hướng phát triển.

II. RỪNG NGẪU NHIÊN CÓ ĐIỀU HƯỚNG

Mục này trình bày tóm tắt rừng ngẫu nhiên có điều hướng (GRRF), phân tích hướng tiếp cận GRRF cho bài toán phân tích dữ liệu SNP mức toàn hệ gien.

A. Độ đo sự quan trọng thuộc tính

Rừng ngẫu nhiên [***5***] gồm một tổ hợp các cây quyết định không cắt nhánh. Mỗi cây quyết định được xây dựng bởi thuật toán CART [***15***] trên tập mẫu bootstrap (lấy mẫu ngẫu nhiên có hoàn lại) từ tập dữ liệu ban đầu. Tại mỗi nút, một phân hoạch tốt nhất được thực hiện dựa trên thông tin trong một không gian con các thuộc tính được chọn ngẫu nhiên từ không gian thuộc tính ban đầu. RF tổng hợp kết quả dự đoán của các cây quyết định làm kết quả cuối cùng.

Ưu điểm của RF là xây dựng cây không thực hiện việc cắt nhánh từ các tập dữ liệu con khác nhau dùng kỹ thuật boostrap có hoàn lại, do đó thu được những cây với lỗi bias thấp. Bên cạnh đó, mối quan hệ tương quan giữa các cây quyết định cũng được giảm thiểu nhờ việc xây dựng các không gian con thuộc tính một cách ngẫu nhiên. Do đó, việc kết hợp kết quả của một số lượng lớn những cây quyết định độc lập có bias thấp, phương sai cao sẽ giúp RF đạt được cả độ lệch thấp và phương sai thấp. Sự chính xác của RF phụ thuộc vào chất lượng dự đoán của các cây quyết định và mức độ tương quan giữa các cây quyết định.

Cho một tập dữ liệu huấn luyện (tập mẫu) chứa N mẫu dữ liệu, M thuộc tính X_j *(j=1,2,...,M)* và $Y \in \{1, 2, .., C\}$ với $C \geq 2$ là biến phụ thuộc. RF dùng chỉ số *Gini* để đo tính hỗn tạp của tập mẫu. Trong quá trình xây dựng các cây quyết định, RF phát triển các nút con từ một nút cha dựa trên việc đánh giá chỉ số Gini của một không gian con *mtry* các thuộc tính được chọn ngẫu nhiên từ không gian thuộc tính ban đầu. Thuộc tính được chọn để tách nút *t* là thuộc tính làm cực tiểu độ hỗn tạp của các tập mẫu sau khi chia. Công thức tính chỉ số *Gini* cho nút *t* như sau:

$$Gini(t) = \sum_{c=1}^{C} \Phi_c(t)[1 - \Phi_c(t)] \quad (1)$$

trong đó: $\Phi c(t)$ là tần suất xuất hiện của lớp $c \in C$ trong nút t.

Gọi *s* là một giá trị trong thuộc tính X_j tách nút *t* thành 2 nút con: nút trái t_L và nút phải t_R tùy thuộc vào $X_j \leq s$ hoặc $X_j > s$; $t_L = \{X_j \in t, X_j \leq s\}$ và $t_R = \{X_j \in t, X_j > s\}$.

Khi đó, tổng độ đo chỉ số *Gini* của 2 nút t_L và t_R sau khi dùng thuộc tính X_j tách nút t tại s là:

$$\Delta Gini(s,t) = p(t_L)Gini(t_L) + p(t_R)Gini(t_R) \quad (2)$$

Để đạt được điểm chia tốt, tại mỗi nút RF sẽ tìm tất cả các giá trị có thể của tất cả *mtry* biến để tìm ra điểm *s* có độ đo $\Delta Gini(s,t)$ nhỏ nhất làm điểm phân tách nút *t*. Thuộc tính chứa điểm phân tách nút *t* được gọi là thuộc tính tách nút t.

Gọi $IS_k(X_j)$, IS_{X_j} lần lượt là độ đo sự quan trọng của thuộc tính X_j trong một cây quyết định T_k (k=1...K) và trong một rừng ngẫu nhiên. Công thức tính $IS_k(X_j)$ và IS_{X_j} như sau:

$$IS_k(X_j) = \sum_{t \in T_k} \Delta Gini(X_j, t) \quad (3)$$

$$IS_{X_j} = \frac{1}{K} \sum_{k=1}^{K} IS_k(X_j) \quad (4)$$

Chuẩn hóa min-max để chuyển độ đo sự quan trọng thuộc tính về đoạn [0,1], theo công thức (5):

$$VI_{X_j} = \frac{IS_{X_j} - \min_{j=1}^{M}(IS_{X_j})}{\max_{j=1}^{M}(IS_{X_j}) - \min_{j=1}^{M}(IS_{X_j})} \quad (5)$$

Độ đo sự quan trọng của các thuộc tính đã chuẩn hóa theo công thức (5) được dùng để lựa chọn thuộc tính trong mô hình GRRF.

B. Rừng ngẫu nhiên có điều hướng

1. *Rừng ngẫu nhiên điều hòa*

Năm 2012 Deng và Runger [***16***] đề xuất mô hình cây điều hòa (Regularized Trees) giúp cải thiện việc lựa chọn thuộc tính trên cây quyết định. Mô hình mở rộng cho tập hợp cây và nhóm tác giả đặt là rừng ngẫu nhiên điều hòa (Regularized Random Forest- RRF).

Ý tưởng của RRF là hạn chế lựa chọn thuộc tính mới để phân tách nút. Nếu thuộc tính mới X_j có độ quan trọng tương đương với thuộc tính X'_j (X'_j là một thuộc tính đã từng được chọn để phân tách), thì RRF ưu tiên chọn thuộc tính X'_j. Thuộc tính mới X_j chỉ được chọn nếu như nó có chỉ số *Gini* nhỏ hơn tất cả các thuộc tính đã được chọn trong các nút trước (xét trong mô hình rừng).

Để thực hiện ý tưởng trên, RRF gán hệ số phạt λ cho $\Delta Gini\left(X_j, t\right)$ đối với các X_j chưa được chọn chia tập dữ liệu huấn luyện lần nào. Gọi F là tập các thuộc tính đã được sử dụng ở các lần chia trước trong mô hình rừng. Độ đo mới dùng để chọn thuộc tính phân tách nút *t* được tính như sau:

$$\Delta Gini_R\left(X_j, t\right) = \begin{cases} \lambda\, \Delta Gini\left(X_j, t\right) & với \quad X_j \notin F \\ \Delta Gini\left(X_j, t\right) & với\ X_j \in F \end{cases} \tag{6}$$

Trong đó: $\lambda \in [0,1]$ là hệ số phạt. Giá trị λ càng nhỏ, thì phạt càng cao. Tại nút gốc của cây đầu tiên *F* được gán giá trị rỗng ($F=\varnothing$). RRF sử dụng chỉ số $\Delta Gini_R\left(X_j, t\right)$ để tách nút. Thuộc tính X_j được thêm vào F nếu như nó có chỉ số $\Delta Gini_R\left(X_j, t\right)$ nhỏ hơn $min(\Delta Gini_R(X_i, t))$ với $X_i \in F$.

Bằng thực nghiệm, Deng và Runger cho thấy tiếp cận RRF làm tăng hiệu năng của RF nguyên bản [***16***] (do RRF không chỉ so độ quan trọng của một thuộc tính trong cây hiện thời mà so trên tất cả các cây đã được xây dựng trước đó để chọn thuộc tính). Vì vậy, RRF làm giảm bias trong quá trình lựa chọn thuộc tính của RF. Tuy nhiên, tại mỗi nút của cây, RRF đánh giá các thuộc tính dựa trên chỉ số *Gini* được tính toán trong một phần nhỏ của tập dữ liệu huấn luyện nhưng lại so sánh với tất cả thuộc tính đã được chọn chia trong rừng. Điều đó dẫn đến RRF có thể chọn phải những thuộc tính không tốt để dựng cây.

Năm 2013, Deng và Runger [***9***] đã thiết lập được giới hạn trên cho số giá trị Gini phân biệt trong bài toán phân lớp nhị phân có N mẫu là N(N+2)/4-1. Vì vậy, khi N nhỏ dẫn đến số giá trị Gini phân biệt nhỏ. Với bài toán chiều cao, sẽ có rất nhiều giá trị $Gini(X_j,t)$ giống nhau, nên rất khó để phân biệt thuộc tính nào là quan trọng hơn. Ví dụ, đối với bài toán phân hoạch nhị phân, tại một nút chỉ có 10 mẫu thì sẽ có khoảng 29 giá trị Gini phân biệt nhau. Trong tập dữ liệu huấn luyện, nếu có 10000 thuộc tính thì sẽ có khoảng 1000-29=971 thuộc tính đạt giá trị *Gini* giống nhau. Nếu những chỉ số *Gini* giống nhau này là những giá trị $Gini_{min}$ thì RRF sẽ chọn ngẫu nhiên một trong số các thuộc tính có chỉ số *Gini* đạt *min* để tách nút t. Như vậy, RRF có thể chọn phải những thuộc tính không hoặc ít liên quan đến biến đích để phân hoạch dữ liệu. Vì vậy, đối với các tập dữ liệu có dung lượng mẫu nhỏ, số chiều rất cao (cao hơn nhiều so với dung lượng mẫu) thì cách trích chọn thuộc tính của RRF cho hiệu quả không cao.

2. *Rừng ngẫu nhiên có điều hướng*

Để khắc phục vấn đề nêu trên của RRF, năm 2013 Deng và Runger đã đề xuất phương pháp rừng ngẫu nhiên có điều hướng (Guided Regularized Random Forests-GRRF) [***9***] áp dụng cho phân tích dữ liệu gien. Tiếp cận này sử dụng độ quan trọng thuộc tính được tạo ra bởi RF nguyên bản trên toàn bộ tập dữ liệu ban đầu làm trọng số cho các thuộc tính nên đã cải thiện được chất lượng của chỉ số Gini, các thuộc tính có độ quan trọng khác nhau sẽ có giá trị Gini khác nhau. Điều này giúp RRF có thể chọn được các thuộc tính phân tách tốt hơn trong bài toán phân tích dữ liệu mẫu nhỏ, số chiều cao, nhiều nhiễu. Thực nghiệm trên các tập dữ liệu gien, Deng và Runger cho thấy GRRF mang lại hiệu quả phân lớp tốt hơn khi so sánh với RF, RRF, varSelRF và C4.5 [***9***].

Nếu như RRF gán hệ số phạt λ bằng nhau cho tất cả các thuộc tính mới, thì GRRF căn cứ độ quan trọng của các thuộc tính dựa trên RF nguyên bản (tính theo công thức (5) từ dữ liệu *out of bag*) để gán hệ số phạt λ_j khác nhau đối với các thuộc tính khác nhau. Thuộc tính có độ quan trọng cao thì gán giá trị λ cao (phạt ít), ngược lại gán giá trị λ thấp (phạt nhiều).

Công thức tính độ quan trọng cho các thuộc tính mới tại nút t trong GRRF như sau:

$$Gain_R\left(X_j, t\right) = \begin{cases} \lambda_j Gain\left(X_j, t\right) & với\ X_j \notin F \\ Gain\left(X_j, t\right) & với\ X_j \in F \end{cases} \tag{7}$$

$\lambda_j \in (0,1]$ là hệ số phạt gán cho các X_j *(j=1,2,...,M)*. Giá trị λ_j dựa vào độ quan trọng của X_j trong RF:

$$\lambda_j = (1-\gamma)\lambda_0 + \gamma VI_{X_j} \tag{8}$$

Trong đó, $\lambda_0 \in (0,1]$ là hệ số điều khiển mức độ điều hướng, $\gamma \in [0,1]$ điều chỉnh độ quan trọng của thuộc tính đã chuẩn hóa và được gọi là hệ số quan trọng. Khi $\gamma = 0$ GRRF trở thành RRF.

Để giảm tham số cho GRRF, Deng và George Runger chọn $\lambda_0 = 1$, ta có:

$$\lambda_j = (1-\gamma) + \gamma VI_{X_j} = 1 - \gamma(1 - VI_{X_j}) \tag{9}$$

Như vậy, GRRF đã kế thừa được những ưu điểm RRF và khắc phục được phần nào hạn chế của RRF trong quá trình lựa chọn thuộc tính phân lớp tại các nút có dung lượng mẫu nhỏ. Tuy nhiên, GRRF lại sử dụng các hệ số quan trọng được tạo ra bởi RF nguyên bản trên tập dữ liệu *out-of-bag* để hướng dẫn quá trình lựa chọn thuộc tính của RRF.

Vì vậy, khi phân lớp đối với những bài toán có dung lượng mẫu nhỏ, số chiều rất cao, nhiều nhiễu như dữ liệu SNP xét trên toàn hệ gien thì GRRF bị hạn chế về độ chính xác. Để nâng cao hiệu quả của GRRF khi phân lớp dữ liệu SNP trên toàn bộ hệ gien, chúng tôi đề xuất một phương pháp tính trọng số mới cho GRRF. Tiếp cận này cải thiện độ chính xác của mô hình GRRF trong quá trình chọn SNP để dựng cây.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Như đã phân tích trong mục II, tiếp cận RF nguyên bản của Breiman cũng như RRF của Deng và Runger không phù hợp cho phân tích dữ liệu SNP trên toàn bộ hệ gien.

Để cải thiện hạn chế của GRRF cho phân tích dữ liệu chiều cao, véctơ trọng số mới được tạo ra giúp RF giảm lỗi bias trong quá trình lựa chọn thuộc tính khi dựng cây. Trọng số của các SNP được tính nhờ phương pháp lặp hoán vị kết hợp đánh giá giá trị p [***17***]. Chúng tôi đưa thêm M SNP nhiễu vào tập dữ liệu ban đầu, các SNP này được tạo ra bằng cách hoán vị giá trị từ các SNP ban đầu nhằm phá hủy quan hệ của chúng với biến đích nhưng vẫn giữ nguyên phân bố dữ liệu tại các SNP. Cách làm này giúp RF giảm xác suất lựa chọn những SNP chứa nhiều giá trị nhưng độ đo quan trọng của chúng kém, từ đó RF giảm được lỗi bias lựa chọn kiểu SNP này trong quá trình dựng cây.

Với giá trị p có được từ kiểm định t-test sau khi chạy R lần RF trên tập dữ liệu mở rộng 2M chiều để tính độ quan trọng cho cả X_j và A_j, R là tham số cho trước. Giá trị p của SNP càng nhỏ thì khả năng tham gia dự đoán của SNP đó càng lớn. Dựa vào giá trị p với một ngưỡng η cho trước (mặc định η= 0.05), véctơ trọng số được chia thành 2 tập, các SNP có giá trị p lớn hơn η sẽ được gán trọng số bằng 0, những SNP còn lại có trọng số được gán bằng trung bình cộng độ đo sự quan trọng của chúng sau R lần lặp. Các bước của phương pháp này được mô tả sau đây:

Xét bài toán SNP có 2 lớp bệnh và không bệnh tương ứng với hai nhãn {0, 1} với tập dữ liệu huấn luyện SNP $S_X = \{\{X_j\}_{j=1}^{M}, Y\}$ có N mẫu dữ liệu và M thuộc tính (SNP), trong đó X_j là các SNP, $Y \in \{1,0\}$

Bước 1: Áp dụng phương pháp lặp hoán vị kết hợp đánh giá giá trị p [***17***] để tìm các SNP không hoặc ít có liên quan tới bệnh.

1.1. Tạo thêm M SNP thực sự nhiễu A_j ứng với các X_j ($j = 1 \div M$) cho S_X bằng cách hoán vị ngẫu nhiên các giá trị của X_j trong tập dữ liệu ban đầu. Kết quả ta được tập dữ liệu SNP mở rộng:

$S_{X,A}=\{S_X,S_A\}$ với $S_A = \{A_j\}_{j=1}^{M}$

1.2. Thực hiện R lần RF trên $S_{X,A}$ để tính độ quan trọng cho tất cả các SNP thực X_j và SNP nhiễu A_j. Với mỗi lần chạy r ($r = 1 \div R$) ta tính độ quan trọng VI_X^r và VI_A^r cho các SNP và đặt chúng vào dòng thứ r của ma trận V_{Rx2M}

Sau bước 1.2 ta sẽ có một ma trận V_{Rx2M} chứa độ quan trọng của các $\{X_j\}_{j=1}^{M}$ và $\{A_j\}_{j=1}^{M}$

1.3. Chọn các giá trị VI_{A_j} lớn nhất (kí hiệu là $VI_{A_j}^{max}$) của mỗi lần lặp r và đặt nó trong mẫu so sánh VI_A^{max}

1.4. Với mỗi thuộc tính X_j ($j=1 \div M$), dùng t-test thực hiện kiểm định $\overline{VI}_{X_j} > \overline{VI}_A^{max}$ để tìm ra các giá trị p tương ứng, từ đó kết luận SNP X_j có phải là thuộc tính nhiễu hay không. Nếu SNP X_j có giá trị p lớn hơn một ngưỡng cho trước η (giá trị mặc định là 0.05) thì kết luận SNP X_j là thuộc tính nhiễu. Ngược lại kết luận SNP X_j không phải nhiễu. Mức độ quan trọng của mỗi thuộc tính tùy thuộc vào giá trị p. Giá trị p của SNP càng nhỏ thì SNP đó càng có đóng góp lớn trong quá trình phân lớp.

Bước 2: Tạo véctơ trọng số $\{\theta_1, \theta_2, \ldots, \theta_M\}$ cho các SNP trong tập dữ liệu huấn luyện S_X

2.1. Gán trọng số quan trọng bằng 0 cho các SNP được coi là nhiễu

2.2. Gán trọng số quan trọng θ_j cho các SNP được coi là mạnh (không nhiễu) theo công thức (10)

$$\theta_j = \frac{1}{R}\sum_{r=1}^{R} VI_{X_j}^r \tag{10}$$

Véctơ trọng số $\{\theta_1, \theta_2, \ldots, \theta_M\}$ này được sử dụng thay cho véctơ trọng số của GRRF để lựa chọn các SNP trong quá trình xây dựng cây. Mô hình GRRF sẽ căn cứ vào các giá trị θ_j để khởi tạo λ_j khác nhau cho các SNP.

Với phương pháp này, mặc dù dữ liệu SNP có nhiều nhiễu, chiều cao, mẫu ít nhưng GRRF vẫn tránh được các SNP có độ quan trọng kém để thực hiện phân tách nút. Do vậy, xây dựng được các cây quyết định có chất lượng dự đoán tốt, mô hình phân lớp rừng ngẫu nhiên sử dụng trọng số mới sẽ cho hiệu năng cao. Mặt khác, tiếp cận này có sử dụng kết quả trong bài báo [***17***] để đưa ra véctơ trọng số mới cho GRRF nhằm giảm lỗi bias khi tạo dựng các cây quyết định. Vì vậy, đề xuất của bài báo phù hợp cho phân tích các bài toán có dữ liệu chiều cao, nhiều nhiễu, dung lượng mẫu nhỏ, nhiều lớp mà dữ liệu SNP chuẩn có 2 lớp xét trên toàn hệ gien là một trường hợp cụ thể.

IV. THỰC NGHIỆM

A. Dữ liệu thực nghiệm

Chúng tôi tiến hành thực nghiệm trên hai bộ dữ liệu SNP chuẩn ở mức toàn bộ hệ gien để làm sáng tỏ hiệu quả của phương pháp đề xuất (từ đây gọi là iGRRF - improved GRRF). Thông tin về hai bộ dữ liệu SNP này được mô tả trong Bảng 1.

Bảng 1. Mô tả hai tập dữ liệu SNP

Tập dữ liệu	Số lượng SNPs	Số lượng cá thể	Số lớp
Alzheimer	380.157	364	2
Parkinson	408.803	451	2

Tập dữ liệu đầu tiên là dữ liệu bệnh chứng cho bệnh Alzheimer chứa đựng 380.157 SNPs được lấy mẫu từ 188 cá thể người có tình trạng thần kinh bình thường (để kiểm chứng) và 176 cá thể người mắc bệnh Alzheimer [***18***]. Tập dữ liệu thứ hai là tập dữ liệu bệnh chứng cho bệnh Parkinson chứa đựng 408.803 SNPs được lấy mẫu từ 541 cá thể, trong đó 271 trường hợp kiểm chứng và 270 trường hợp bệnh [***19***].

B. Tham số chạy mô hình và phương pháp đánh giá

Các tham số γ, *mtry* chạy mô hình GRRF và mô hình iGRRF tương ứng là: $0.1, \sqrt{M}$ (vì theo kết quả trong [***9***], khi dùng hệ số γ=0.1, GRRF cho kết quả tốt nhất, còn *mtry*=$\sqrt{M}$ là tham số tối ưu khi RF xử lý bài toán phân lớp [***5***]). Để tính véctơ trọng số, chúng tôi thực hiện 30 lần lặp RF trên tập dữ liệu mở rộng 2M chiều với mtry=10%M và số cây trong rừng là 500. Chúng tôi cũng so sánh hiệu suất của mô hình iGRRF với các mô hình RF được đề xuất những năm gần đây như: SRF của Wu và đồng nghiệp [***2***], GRRF của Deng và Runger [***9***], wsRF của Xu và đồng nghiệp [***20***] và mô hình linear kernel SVM trong gói e1071 [***21***]. Chúng tôi đặt tham số $C=2^{-5}$ cho Alzheimer và $C=2^{-2}$ cho Parkinson vì tham số này SVM đạt giá trị tốt nhất trên 2 tập dữ liệu SNP trên.

Phương pháp kiểm tra chéo *5-fold* được sử dụng để đánh giá hiệu quả của mô hình iGRRF và các mô hình đối chứng trên hai tập dữ liệu Alzheimer và Parkinson. Để đánh giá hiệu quả phân lớp của iGRRF với các mô hình rừng ngẫu nhiên khác khi số lượng cây trong rừng biến thiên, chúng tôi đặt cố định kích thước không gian con thuộc tính là *mtry*=$\sqrt{M}$ và thay đổi số lượng cây từ 20 tới 200.

Trong thực nghiệm này, hai độ đo được sử dụng để đánh giá hiệu năng của các mô hình RF trên tập dữ liệu kiểm thử D_t là *Area under the curve* (AUC) và ACC. Độ chính xác kiểm thử ACC được tính như sau:

$$Acc = \frac{1}{N_t}\sum_{i=1}^{N_t} I(Q(x_i, y_i) - max_{j \neq y_i} Q(x_i, j) > 0)$$

trong đó I(.) là hàm dấu hiệu và $Q(x_i, j) = \sum_{k=1}^{K} I(\hat{h}_k(x_i) = j)$ là số lượng cây quyết định lựa chọn x_i thuộc vào lớp j. N_t là số mẫu trong D_t.

Trong bài báo có sử dụng các gói phần mềm R mới nhất của RF [***22***] và GRRF [***23***] để tiến hành thực nghiệm trên 6 máy Linux 64-bit, mỗi máy có cấu hình như sau: IntelR XeonR CPU E5620 2.40 GHz, 16 cores, 4 MB cache, and 32 GB main memory.

C. Kết quả thực nghiệm

Bảng 2. Độ chính xác phân lớp 2 tập dữ liệu SNP của từng mô hình khi sử dụng giá trị mtry tối ưu với số cây trong rừng là 500

Tập dữ liệu	Phương pháp	Mtry	Giá trị	Acc	AUC	#SNPs
Alzheimer	iGRRF	$\sqrt{M}$	616	**0,920**	**0,976**	1997
	GRRF	$\sqrt{M}$	616	0,657	0,706	
	SRF	$(\log_2 M+1)^2$	361	0,797	0,816	
	wsRF	$\log_2 M$	19	0,561	0,711	
	wsRF	$\sqrt{M}$	616	0,692	0,757	
	SVM	C	2^{-5}	0,690	0,716	
	RF	$\log_2 M$	19	0,530	0,623	
	RF	$\sqrt{M}$	616	0,632	0,729	
Parkinson	iGRRF	$\sqrt{M}$	638	**0,860**	**0,954**	1980.8
	GRRF	$\sqrt{M}$	638	0,688	0,765	
	SRF	$(\log_2 M+1)^2$	361	0,838	0,927	

	wsRF	$\log_2 M$	19	0,754	0,850	
	wsRF	$\sqrt{M}$	638	0,837	0,917	
	SVM	C	2^{-2}	0,825	0,902	
	RF	$\log_2 M$	19	0,564	0,722	
	RF	$\sqrt{M}$	638	0,799	0,848	

Bảng 3. So sánh sự khác biệt về mức độ chính xác dự đoán khi số lượng cây quyết định thay đổi (cố định *mtry*=$\sqrt{M}$)

Tập dữ liệu	**Phương pháp**	**K**				
		20	50	80	100	200
Alzheimer	iGRRF	**0,733**	**0,772**	**0,838**	**0,832**	**0,893**
	GRRF	0,503	0,500	0,539	0,533	0,528
	wsRF	0,528	0,588	0,527	0,602	0,593
	RF	0,517	0,491	0,505	0,555	0,533
Parkinson	iGRRF	**0,850**	**0,855**	**0,861**	**0,850**	**0,854**
	GRRF	0,532	0,604	0,641	0,669	0,680
	wsRF	0,647	0,680	0,708	0,710	0,745
	RF	0,579	0,557	0,553	0,597	0,580

Trong bảng 2, cột #SNPs lưu giá trị số lượng SNP mà iGRRF đã chọn được (chia trung bình sau 5 lần chạy kiểm tra chéo *5- folds*). Điều đó cho thấy mô hình iGRRF đã tìm ra được tập con các SNP có ý nghĩa cho phân lớp. Tập #SNPs này có số chiều nhỏ hơn rất nhiều so với tập SNP trên toàn hệ gien ban đầu nhưng vẫn cho kết quả phân lớp tốt. Cột **Acc** và **AUC** trong bảng 2 thể hiện giá trị trung bình độ chính xác kiểm thử và AUC của iGRRF so với 4 phương pháp GRRF, wsRF, RF và SVM. Kết quả trong bảng 2 cho thấy, với *mtry* = $\lfloor\sqrt{M}\rfloor$, iGRRF cho kết quả phân lớp tốt hơn hẳn các phương pháp GRRF, wsRF, RF và SVM trên cả 2 bộ dữ liệu. Đặc biệt, trên bộ dữ liệu Alzheimer, iGRRF đạt chỉ số *ACC* lên đến 92%, còn chỉ số AUC đạt gần 98%, trong khi các phương pháp khác cho kết quả không cao (GRRF chỉ đạt ACC là 65.7% và AUC là 70.6%, các phương pháp còn lại cũng cho kết quả tương tự hoặc thấp hơn). Mặc dù những phương pháp đối chứng GRRF, wsRF, RF và SVM đều rất mạnh trong bài toán phân lớp và đã được chạy với các tham số tối ưu, nhưng lại cho kết quả không tốt khi phân tích trên dữ liệu SNP ở mức toàn bộ hệ gien. Kết quả iGRRF đạt được chứng tỏ rằng phương pháp đánh trọng số mới cho SNP đã đề xuất cải thiện rõ rệt cho bài toán phân lớp và lựa chọn SNP, đặc biệt là kiểu dữ liệu luôn gây khó khăn lớn cho các mô hình máy học khi số chiều rất lớn nhưng cỡ mẫu nhỏ.

Bảng 3 là kết quả sau 5 lần chạy kiểm tra chéo *5- folds* để so sánh mức độ chính xác dự đoán của iGRRF với ba mô hình rừng ngẫu nhiên GRRF, wsRF, RF khi số lượng cây quyết định thay đổi. Cả 4 phương pháp đều được chạy với tham số cố định *mtry* = $\lfloor\sqrt{M}\rfloor$ trong khi số lượng cây quyết định trong rừng được điều chỉnh trong mỗi lần chạy. Cụ thể chúng tôi đã thử nghiệm cả 4 phương pháp với số lượng cây quyết định thay đổi từ 20 tới 200 cây. Kết quả bảng 3 đã cho thấy rằng iGRRF vượt trội GRRF, wsRF và RF về sự chính xác trong dự đoán. Khi so sánh trực tiếp với mô hình GRRF, ta nhận thấy rằng iGRRF chỉ cần số lượng cây khá ít (20 cây) nhưng độ chính xác phân lớp trên cặp dữ liệu SNP tương ứng đã đạt đến 73.3% và 85%, kết quả cao hơn nhiều so với GRRF khi mô hình này chỉ đạt 50% trên bộ dữ liệu Alzheimer và 53% trên bộ Parkinson.

Như vậy, với những kết quả thực nghiệm đã liệt kê ở Bảng 2 và Bảng 3, mô hình iGRRF cho kết quả dự đoán có độ chính xác cao và khả năng trích chọn SNP hiệu quả hơn hẳn GRRF, SVM và các phương pháp cải tiến RF hiện thời trên cả hai bộ dữ liệu SNP. Những kết quả này một lần nữa chứng minh bằng thực nghiệm, phương pháp đề xuất đánh trọng số SNP mới đã cải thiện đáng kể độ chính xác phân lớp của GRRF và iGRRF là mô hình hữu hiệu có thể dùng cho phân tích dữ liệu SNP trên toàn bộ hệ gien.

V. KẾT LUẬN

Bài báo đã trình bày giải pháp sử dụng trọng số mới cải tiến mô hình rừng ngẫu nhiên có điều hướng, với mục tiêu làm tăng độ chính xác phân lớp dữ liệu SNP trên toàn bộ hệ gien. Giải pháp này đã cải thiện hiệu năng của mô hình GRRF trong quá trình chọn SNP để phân hoạch khi dựng cây, và giảm số chiều dữ liệu. Ngoài ra, số lượng SNP có độ quan trọng cao được lựa chọn theo quy tắc của GRRF trong khi vẫn duy trì được tính ngẫu nhiên của rừng. Kết quả thực nghiệm cho thấy GRRF cải tiến đạt độ chính xác phân lớp tốt hơn hẳn Guided Regularized Random Forests, Support Vector Machine, Random Forests và một số phương pháp cải tiến rừng ngẫu nhiên khác trong thời gian gần đây như: Stratified Random Forests, Weighted Subspace Random Forests.

VI. TÀI LIỆU THAM KHẢO

[1] M. K. et al. Halushka, "Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis.," *Nature Genet.*, vol. 22, pp. 239–247, 1999.

[2] Q. Wu, Y. Ye, Y. Liu, and M. K. Ng, "Snp selection and classification of genome-wide snp data using stratified sampling random forests," *The Journal of IEEE Transactions on NanoBioscience*, vol. 11, no. 3, pp. 216–227, 2012.

[3] M. Stratton, "Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls," *The Journal of Nature*, vol. 447, no. 7145, pp. 661–678, 2007.

[4] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc, 1995.

[5] L. Breiman, "Random forests," *Journal of Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[7] L. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, and B. Hayward, "Identifying snps predictive of phenotype using random forests," *Journal of Genetic epidemiology*, vol. 28, no. 2, pp. 171-182, 2005.

[8] K. L. , Hayward, L. B. , Segal, J. , Van Eerdewegh, P. Lunetta, "Screening large association study data: exploiting interactions using random forests," *The Journal of BMC genetics*, pp. 5(1): 32, 2004.

[9] H. Deng and G. Runger, "Gene selection with guided regularized random forest," *Journal of Pattern Recognition*, vol. 46, pp. 3483-3489, 2013.

[10] H. Kim and Wei-Yin Loh, "Classification trees with unbiased multi-way splits," *Journal of American Statistical Association*, vol. 96, no. 454, pp. 589-604, 2001.

[11] C. Strobl, "Statistical sources of variable selection bias in classi_cation trees based on the gini index," Sonderforschungsbereich 386, Technical report, Technical Report SFB 386 http://epub. ub. uni-muenchen. de/archive/00001789/01/paper 420. pdf, 2005.

[12] C. Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis, "Conditional variable importance for random forests," *The Journal of BMC bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[13] C. Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *Journal of BMC bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[14] C. Strobl, Anne-Laure Boulesteix, and Thomas Augustin, "Unbiased split selection for classi_cation trees based on the gini index," *The Journal of Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 483-501, 2007.

[15] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. USA: CRC press, 1984.

[16] H. Deng and G. Runger, "Feature selection via regularized trees," in *International Joint Conference on Neural Networks(IJCNN)*, 2012, pp. 1-8.

[17] E. Tuv, A. Borisov, G. Runger, and K. Torkkola, "Feature selection with ensembles artificial variables, and redundancy," *The Journal of Machine Learning*, vol. 10, pp. 1341–1366, 2009.

[18] J. A. Webster, J. R. Gibbs, J. Clarke, M. Z. Ray, and et al., "Genetic control of human brain transcript expression in Alzheimer disease," *The American Journal of Human Genetics*, vol. 84, no. 4, pp. 445-458, 2009.

[19] H. C. Fung, S. Scholz, M. Matarin, and et al, "Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data," *The Lancet Neurology*, vol. 5, no. 11, pp. 911-916., 2006.

[20] B. Xu, J. Z. Huang, G. Williams, and Q. Wang, "Classifying very high-dimensional data with random forests built from small subspaces," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 8, no. 2, pp. 44-63, 2012.

[21] K. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, "e1071: Misc functions of the department," 2012.

[22] M. Wiener and A. Liaw, "Classification and regression by randomforest," *The Journal of R news*, vol. 2, no. 3, pp. 18–22, 2002.

[23] L. Deng, "Guided random forest in the rrf package," 2013.

[24] V. Vapnik, *The Nature of Statistical Learning Theory*. USA: Springer-Verlag, 1995.

[25] Y. V Sun, Z. Cai, K. Desai, and et al., "Classification of Rheumatoid Arthritis Status with Candidate Gene and Genome-Wide Single-Nucleotide Polymorphisms Using Random Forests," in *BMC Proceedings 1*, 2007.

[26] B. Goldstein, A. Hubbard, A. Cutler, and L. Barcellos, "An application of Random Forests to a genome-wide association dataset:Methodological considerations & new findings," *BMC Genet*, vol. 11, no. 1, p. 49, 2010.

[27] D. Schwarz, I. König, and A. Ziegler, "On safari to Random Jungle: A fast implementation of Random Forests for high-dimensional data," *Bioinformatics*, vol. 26, no. 14, p. 1752, 2010.

IMPROVING THE GUIDED REGULARIZED RANDOM FORESTS METHOD FOR SINGLE NUCLEOTIDE POLYMORPHISM CLASSIFICATION

Hoang Thi Ha, Nguyen Thanh Tung

ABSTRACT *- Random forests (RF) model has showed to perform well in terms of prediction accuracy when applied to some data sets of moderate size. However, RF's performance suffers from high-dimensional data sets for selecting informative features and building an accurate prediction. Recently, Deng and Runger proposed a guided regularized RF (GRRF, Pattern Recognition-2013) model to select feature using RF, however the predictive performance of GRRF is reduced when dealing with data sets containing high dimensionality and few samples, such as genome-wide association Single Nucleotide Polymorphism (SNP) data. In this paper, we improve the prediction of accuracy of the GRRF method by using new weights to guide RF in the feature selection stage when growing trees. Our experimental results on the SNP pair data sets (Parkinson and Alzheimer disease case-control data sets comprised of 408,803 SNPs and 380,157 SNPs, respectively) demonstrated that the proposed RF model outperforms the GRRF model and some state-of-the-art machine learning models including SVM, SRF and wsRF in increasing of the prediction of accuracy.*

Keywords: *High-Dimensional Data, Machine Learning, Random Forests, Data mining*