ĐÁNH GIÁ CÁC KỸ THUẬT LỰA CHỌN ĐẶC TRƯNG CHO BÀI TOÁN PHÂN LOẠI BIỂU HIỆN GEN

Phan Thị Thu Hồng*, Nguyễn Thị Thủy

Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam

Email: hongptvn@gmail.com*

Ngày gửi bài: 11.08.2015 Ngày chấp nhận: 08.03.2016

TÓM TẮT

Xác định các gen có khả năng gây bệnh là một thách thức lớn trong nghiên cứu về biểu hiện gen. Nhiều phương pháp lựa chọn gen chỉ tập trung vào việc đánh giá sự liên hệ của từng gen riêng biệt với bệnh. Thực tế, một trong những nguyên nhân gây ra các bệnh được cho là liên quan tới những tương tác phức tạp giữa các gen. Phương pháp rừng ngẫu nhiên (RF) gần đây đã được ứng dụng thành công trong việc xác định một số nhân tố di truyền có ảnh hưởng lớn tới một số bệnh. Tuy nhiên mô hình này chỉ hiệu quả đối với một số tập dữ liệu có cỡ trung bình nhưng lại hạn chế trong việc xác định các gen có ý nghĩa và xây dựng các mô hình dự đoán chính xác cho dữ liệu có số chiều lớn. Trong bài báo này chúng tôi tập trung vào các phương pháp rừng ngẫu nhiên cải tiến cho phép chọn ra một tập nhỏ các đặc trưng có liên hệ chặt chẽ với biến đích, do đó làm giảm số chiều và có thể xử lý tốt trên các tập dữ liệu có số chiều cao. Hiệu năng của các mô hình này được phân tích để tìm ra phương pháp phân lớp hiệu quả với từng mục tiêu như độ chính xác hay tập các gen có ý nghĩa dựa vào kết quả thử nghiệm trên 8 tập dữ liệu biểu hiện gen được lấy từ ngân hàng dữ liệu y sinh (Kent Ridge) và tin sinh (Bioinformatics).

Từ khóa: Dữ liệu biểu hiện gen, lựa chọn đặc trưng, phân loại, rừng ngẫu nhiên, rừng ngẫu nhiên điều hòa, rừng ngẫu nhiên điều hòa có điều hướng, rừng ngẫu nhiên có điều hướng.

Evaluation of Feature Selection Methods for Gene Expression Data Classifcation

ABSTRACT

Selection of relevant genes that have effects in some diseases is a challenging task in gene expression studies. Most gene selection studies focused on assessing the association between individual gene and the disease. In fact, diseases are thought to involve a complex etiology including complicated interactions between many genes and the disease. Random Forest (RF) method has recently been successfully used for identifying genetic factors that have effects in some complex diseases. In spite of performing well in some data sets with moderate size, RF still suffers from working for selecting informative genes and building accurate prediction models. In this paper, we investigated some methods in learning advanced random forests that allow one to select a sub-set of informative genes (most relevant to disease). The method can therefore reduce the dimensionality and can perform well in prediction high-dimensional data sets. The performance of these methods has been analyzed for finding the robust one for each interest objective (the accuracy of the prediction model or the smallest possible set of relevant genes) based on experiments results on 8 available public data sets of gene expression from the repository of biomedical data sets (Kent Ridge) and bioinformatics data sets (Bioinformatics).

Keywords: Classification, gene expression data, feature selection, Random forest, Regularized Random Forest, Guided Regularized Random Forests.x

x

x

1. ĐẶT VẤN ĐỀ

Lựa chọn đặc trưng là việc lựa chọn từ một tập hợp các đặc trưng đầu vào để đưa ra một tập nhỏ các đặc trưng có ý nghĩa nhất. Xét một vector đặc trưng đầu vào có d biến $X = \{X_1,..., X_d\}$ và $Y = \{1, 2,..., C\}$ là giá trị đầu ra có thể dự đoán từ vector đặc trưng X. Nhiệm vụ lựa chọn đặc trưng chính là việc tìm ra các đặc trưng X_i có liên quan nhất đến dự đoán giá trị Y. Những phương pháp phân lớp bị phụ thuộc rất lớn vào yếu tố đầu vào, khả năng phân lớp của thuật toán có xu hướng giảm khi X chứa các biến không có ý nghĩa. Khi dữ liệu có số lượng đặc trưng lớn, việc tìm kiếm tập các đặc trưng tối ưu là rất khó. Lựa chọn đặc trưng có tầm quan trọng rất lớn đặc biệt là đối với bài toán phân lớp dữ liệu gen, trong đó vectơ đặc trưng có rất ít các phần tử dữ liệu có ý nghĩa nhưng số chiều rất lớn và có nhiễu. Đây là một trong mười vấn đề khó của cộng đồng khai phá dữ liệu (Yang and Wu, 2006). Lựa chọn các gen có liên quan để phân loại mẫu (ví dụ, để phân biệt giữa các bệnh nhân mắc và không mắc bệnh ung thư) là một nhiệm vụ đang rất được quan tâm trong hầu hết các nghiên cứu biểu hiện gen (Lee *et al.*, 2005; Yeung *et al.*, 2005; Jirapech-Umpai and Aitken, 2005; Hua *et al.*, 2005; Li *et al.*, 2005; Díaz-Uriarte, 2005). Khi thực hiện lựa chọn những gen ảnh hưởng đến bệnh, các nhà nghiên cứu y sinh học thường quan tâm tới một trong hai mục tiêu sau đây:

(1) Xác định các gen có liên quan để phục vụ cho các nghiên cứu tiếp theo; kết quả của quá trình này là một tập hợp các gen liên quan đến biến đích (có thể là một tập gồm nhiều gen) và tập này có thể chứa các gen có chức năng tương tự và có tương tác chặt chẽ với nhau;

(2) Xác định một tập nhỏ các gen mà chúng có thể được sử dụng cho mục đích chẩn đoán lâm sàng hay điều chế dược phẩm; quá trình này thu được một tập nhỏ nhất có thể các gen mà kết quả dự đoán vẫn có thể đạt hiệu quả tốt (các gen "dư thừa" không được chọn).

Trong bài báo này chúng tôi tập trung vào mục tiêu (2): thử nghiệm với các phương pháp lựa chọn đặc trưng khác nhau, phân tích đánh giá các phương pháp này tùy theo mục đích bài toán để tìm ra được tập đặc trưng tốt nhất hay để đạt được kết quả dự đoán cao. Phần còn lại bài báo được bố trí như sau: Phần 2 giới thiệu các nghiên cứu liên quan. Phần 3 trình bày về các phương pháp rừng ngẫu nhiên cải tiến. Phần 4 đề cập đến dữ liệu thực nghiệm và phương pháp đánh giá. Phần 5 trình bày một số kết quả thực nghiệm nhằm kiểm chứng khả năng phân loại của phương pháp lựa chọn đặc trưng cho bài toán phân lớp dữ liệu biểu hiện gen. Phần cuối cùng là kết luận.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Hiện nay phần lớn các phương pháp lựa chọn gen là thực hiện kết hợp việc xếp thứ hạng các gen (ví dụ, sử dụng các phương pháp kiểm thử thống kê F-ratio hoặc Wilcoxon) với một bộ phân loại cụ thể (ví dụ, K- hàng xóm gần nhất). Lựa chọn một số lượng đặc trưng tối ưu để thực hiện phân loại là công việc rất khó khăn và phức tạp, mặc dù đã có các hướng dẫn cơ bản dựa trên nghiên cứu mô phỏng (Hua *et al.*, 2005). Thông thường quyết định số gen được giữ lại là tùy ý, ví dụ 50 gen với xếp hạng tốt nhất (Lee *et al.*, 2005; Dudoit *et al.*, 2002); 150 gen (Li *et al.*, 2004). Cách tiếp cận này có thể thích hợp với mục tiêu phân loại mẫu nhưng không phải thích hợp nhất nếu để tìm ra tập hợp nhỏ nhất có thể của các gen có ảnh hưởng đến biến đích và những gen này cho phép dự đoán tốt. Một cách tiếp cận phổ biến khác nữa (van't Veer *et al.*, 2002; Roepman *et al.*, 2002; Furlanello *et al.*, 2003) là loại bỏ dần dần các gen từ tập ban đầu theo các lần lặp (loại bỏ gen dựa trên thứ hạng các gen được đánh giá theo các phương pháp thống kê hay dựa trên tỷ lệ lỗi dự đoán khi thực hiện loại bỏ từng gen) cho đến khi mục tiêu được thỏa mãn (tỷ lệ lỗi nhỏ nhất trong tất cả các bộ gen đã thử). Tuy nhiên với phương pháp này có thể sẽ loại bỏ gen nếu

đứng độc lập vì nó ít liên quan đến bệnh (dựa trên sự xếp hạng đơn biến, hoặc dựa trên tương tác các cặp gen (Bø and Jonassen, 2002) nhưng có thể ảnh hưởng lớn tới bệnh nếu có sự hiện diện của những gen khác.

Mặt khác, vấn đề chọn lọc gen thường gặp khó khăn hơn khi thực hiện phân lớp với những bộ dữ liệu đa lớp (có nhiều hơn hai lớp) (Yeung *et al.*, 2005; Li *et al.*, 2004). Do đó, các thuật toán phân lớp cung cấp các độ đo thuộc tính quan trọng như sự ảnh hưởng của các thuộc tính đến việc phân loại là những phương pháp rất được quan tâm để thực hiện lựa chọn gen, đặc biệt là các thuật toán phân lớp có thể đáp ứng được với tập dữ liệu có số chiều rất lớn nhưng số phần tử lại rất ít.

Năm 2001, Brieman đề xuất phương pháp Random Forest (RF), đây là một phương pháp phân lớp và hồi quy dựa trên việc kết hợp kết quả dự đoán của một số lượng lớn các cây quyết định. Trong mô hình RF truyền thống mỗi cây quyết định được xây dựng từ tập dữ liệu được lấy ngẫu nhiên từ tập dữ liệu ban đầu và việc phát triển các nút con từ một nút cha dựa trên thông tin trong một không gian con các thuộc tính được chọn ngẫu nhiên từ không gian thuộc tính ban đầu. Do đó, RF xây dựng các cây quyết định từ một tập con những thuộc tính được lựa chọn ngẫu nhiên và tổng hợp kết quả dự đoán của các cây để tạo ra kết quả dự đoán cuối cùng. Các cây quyết định được xây dựng sử dụng thuật toán CART (Brieman, 1984) mà không thực hiện việc cắt tỉa do đó thu được những cây với độ lệch 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. Như vậy, 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.

Trong thực tế RF đã trở thành một công cụ tin cậy cho phân tích dữ liệu, đặc biệt là dữ liệu tin sinh học (Bureau *et al.*, 2005; Goldstein *et al.*, 2010; Goldstein *et al.*, 2011; Winham *et al.*, 2012). Tuy nhiên, tiếp cận RF ban đầu của Breiman chỉ hiệu quả cho phân tích dữ liệu có số chiều thấp (Bureau *et al.*, 2005; Lunetta *et al.*, 2004). Mô hình RF truyền thống không thể áp dụng trên dữ liệu có số chiều lớn, có thể lên đến hàng ngàn hay trăm ngàn gen. Nguyên nhân là trong quá trình xây dựng cây quyết định, tại mỗi nút, RF sử dụng một tập con những thuộc tính được lựa chọn ngẫu nhiên từ tập thuộc tính ban đầu. Vì vậy khi xử lý với các dữ liệu nhiều chiều như dữ liệu gen, RF có thể lựa chọn ngẫu nhiên những gen không có ảnh hưởng đến biến đích và từ đó tạo ra cây quyết định có chất lượng dự đoán thấp.

Gần đây, một số phương pháp rừng ngẫu nhiên cải tiến đã được đề xuất để thực hiện lựa chọn các thuộc tính giúp cải thiện quá trình lựa chọn thuộc tính và tăng hiệu quả dự đoán với các bộ dữ liệu nhiều chiều và nhiều nhiễu như phương pháp rừng ngẫu nhiên điều hòa (Regularized Random Forest- RRF) (Deng and Runger, 2012), rừng ngẫu nhiên điều hòa có điều hướng (Guided Regularized Random Forests-GRRF) (Deng and Runger, 2013) và phương pháp rừng ngẫu nhiên có điều hướng (Guided Random Forest) (Deng, 2013). Vì vậy trong phạm vi nghiên cứu của bài báo này, chúng tôi tập trung vào các phương pháp phân lớp rừng ngẫu nhiên cải tiến cho phép tìm ra một tập nhỏ các gen có ảnh hưởng lớn đến bệnh, do đó làm giảm số chiều và có thể xử lý tốt trên các tập dữ liệu có số chiều cao. Chúng tôi tiến hành các thử nghiệm trên 8 tập dữ liệu biểu hiện gen được lấy từ ngân hàng dữ liệu y sinh (Kent Ridge) và tin sinh (Bioinformatics). Sau đó chúng tôi phân tích hiệu năng của các kỹ thuật trên cũng như số đặc trưng được lựa chọn của từng phương pháp từ đó đưa ra đề xuất sử dụng phương pháp phân lớp với từng mục đích cụ thể (lựa chọn các đặc trưng ảnh hưởng đến biến đích hay yêu cầu hiệu quả dự đoán cao).

3. CÁC PHƯƠNG PHÁP RỪNG NGẪU NHIÊN CẢI TIẾN

3.1. Rừng ngẫu nhiên điều hòa

Như đã phân tích ở trên, RF nguyên bản của Breiman không phù hợp cho phân tích dữ liệu biểu hiện gen có số chiều lớn, vì việc lấy

mẫu trong không gian con thuộc tính có thể dẫn tới việc chọn phải những mẫu không tốt và kết quả là nhiều cây quyết định có chất lượng thấp, dẫn đến giảm khả năng dự đoán của RF. Để khắc phục nhược điểm này năm 2012 Deng và Runger đề xuất mô hình rừng ngẫu nhiên điều hòa, RRF). Cụ thể các tác giả đã thay đổi cách tính độ đo cho mỗi thuộc tính để giảm số thuộc tính mới được chọn cho việc thực hiện phân tách nút tại bước xây dựng cây. Nếu thuộc tính mới X_i và X_j có độ quan trọng là như nhau mà thuộc tính X_j đã từng được chọn để phân tách nút thì RRF ưu tiên chọn thuộc tính X_j. Thuộc tính mới X_i chỉ được chọn khi chỉ số *gain* của X_i lớn hơn chỉ số *gain* của tất cả các thuộc tính đã được chọn trong các nút trước. Gọi F là tập các thuộc tính đã được sử dụng ở các lần chia trước trong mô hình rừng. Độ đo mới của các thuộc tính được tính như sau:

$$gain_R(X_i) = \begin{cases} \lambda . gain(X_i)\ X_i \notin F \\ gain(X_i)\ X_i \in F \end{cases}$$

Ở đây $\lambda \in [0, 1]$ là hệ số phạt; λ càng nhỏ thì phạt càng lớn đối với những thuộc tính không thuộc tập *F*. RRF sử dụng $gain_R(\cdot)$ để tách nút.

3.1.1. Rừng ngẫu nhiên điều hòa có điều hướng (GRRF)

Trong phương pháp rừng ngẫu nhiên điều hòa, Deng *et al.* (2012) đã thay đổi cách tính độ đo quan trọng của mỗi thuộc tính do đó RRF làm giảm độ lệch (bias) so với RF nguyên bản. Tuy nhiên các chỉ số đo độ quan trọng thuộc tính này được đánh giá dựa trên một phần của dữ liệu huấn luyện tại mỗi nút của cây so với tất cả các thuộc tính đã được chọn để xây dựng cây trong rừng. Mặt khác đối với các tập dữ liệu có số mẫu nhỏ, số chiều lớn thì có rất nhiều các thuộc tính có cùng độ đo. Với N mẫu thì số lượng tối đa các thuộc tính có các chỉ số *Gini* khác nhau trong bài toán phân lớp nhị phân là (N(N + 2)/4)-1 (Deng and Runger, 2013). Ví dụ ta có 30 mẫu có số chiều là 3.000, như vậy có lớn nhất là 239 thuộc tính có độ đo khác nhau và 3.000-239 = 2.761 thuộc tính cùng độ đo. Chính vì vậy RRF phải chọn ngẫu nhiên một trong các thuộc tính đó để tách nút. Các thuộc tính này có thể là những thuộc tính không tốt (không hoặc ít có liên quan đến biến đích) dẫn đến khả năng dự đoán của rừng RRF không cao.

Xuất phát từ lý do trên, Deng *et al.* (2013) đã đề xuất phương pháp rừng ngẫu nhiều điều hòa có điều hướng (Guided Regularized Random Forests, GRRF) để khắc phục nhược điểm của RRF. Ở phương pháp GRRF các tác giả tính độ quan trọng thuộc tính dựa trên độ quan trọng thuộc tính được tạo ra bởi RF gốc trên toàn bộ tập dữ liệu ban đầu. Do vậy chỉ số *Gini* của các thuộc tính có độ quan trọng khác nhau sẽ có giá trị khác nhau. Khi đó với các bài toán có số mẫu nhỏ, số chiều lớn như dữ liệu gen, GRRF sẽ chọn được các thuộc tính chia nút tốt hơn và kết quả phân lớp cũng tốt hơn (Deng and Runger, 2013).

Nếu như RRF gán hệ số phạt như nhau cho tất cả các thuộc tính mới thì GRRF sử dụng những thuộc tính có độ quan trọng lớn từ RF truyền thống để "*hướng dẫn*" quá trình lựa chọn thuộc tính mới phân tách nút. Thuộc tính có độ quan trọng cao thì được gán giá trị λ cao, ngược lại thuộc tính có độ quan trọng thuộc tính thấp được gán giá trị λ thấp. Công thức tính độ quan trọng cho các thuộc tính mới tại nút *v* trong GRRF như sau:

$$Gain_R(X_i, v) = \begin{cases} \lambda_i Gain_R(X_i, v)\ X_i \notin F \\ Gain_R(X_i, v)\ X_i \in F \end{cases}$$

Với $\lambda_i \in (0, 1]$ là hệ số phạt của X_i và λ_i được tính như sau:

$$\lambda_i = (1-\gamma)\lambda_0 + \gamma Imp_i' ;$$
$$Imp_i' = \frac{Imp_i}{max_{j=1}^{P} Imp_j}$$

Trong đó $\lambda_0 \in (0, 1]$ là hệ số điều khiển mức độ điều hướng (trong mô hình RRF). Còn hệ số $\gamma \in [0, 1]$ điều khiển độ quan trọng của một thuộc tính (đã được chuẩn hóa). Khi $\gamma = 0$ thì GRRF chính là RRF. Một thuộc tính có độ quan trọng lớn sẽ bị phạt ít hơn. Để thay đổi kích thước tập con thuộc tính được chọn ta có thể thay đổi các giá trị của λ_0 và γ và để giảm tham số cho mô hình *GRRF* các tác giả chọn $\lambda_0 = 1$. Khi đó, ta có:

$$\lambda_i = (1-\gamma) + \gamma Imp_i' = 1 - \gamma(1 - Imp_i')$$

3.1.2. Rừng ngẫu nhiên có điều hướng (Guided Random Forest, GRF)

Tương tự như phương pháp lựa chọn thuộc tính GRRF, Deng *et al.* (2013) đã đề xuất phương pháp rừng ngẫu nhiên có điều hướng bằng cách sử dụng các độ đo độ quan trọng thuộc tính từ RF nguyên bản. Tuy nhiên, các cây trong GRRF được xây dựng một cách tuần tự, liên quan chặt chẽ và không cho phép tính toán song song, trong khi các cây trong GRF được xây dựng một cách độc lập và có thể được thực hiện song song. Phương pháp này cũng cho phép sử dụng các chỉ số đo độ quan trọng khác độ đo độ thuộc tính từ phương pháp rừng ngẫu nhiên gốc (các chỉ số có thể được cung cấp bởi chính người dùng thông qua chỉ số λ_i).

Ý tưởng chính của GRF là tăng trọng số gain(X_i) dựa vào độ đo độ quan trọng thuộc tính từ RF nguyên bản

$$gain_G(X_i) = \lambda_i gain(X_i),$$

Trong đó, $gain(X_i)$ là độ đo *Gini* của thuộc tính X_i để thực hiện tách nút và λ_i được tính như sau:

$$\lambda_i = 1 - \gamma + \gamma \frac{Imp_i}{Imp^*}$$

Với Imp_i, Imp^* là độ đo thuộc tính và giá trị lớn nhất của độ đo thuộc tính từ phương pháp RF nguyên bản. $Imp/Imp^* \in [0, 1]$ là hệ số chuẩn hóa độ quan trọng thuộc tính, $\gamma \in [0, 1]$ là hệ số quan trọng. Ở phương pháp GRF, các thuộc tính có độ quan trọng nhỏ hơn sẽ bị phạt nhiều hơn và độ phạt tăng khi γ tăng (GRF trở thành RF khi $\gamma = 0$).

Từ các trình bày của các phương pháp ở trên, chúng ta thấy sự khác biệt căn bản giữa GRF và GRRF là: các thuộc tính được sử dụng để xây dựng các cây trước trong đó của rừng GRRF có thể tiếp tục được sử dụng (ảnh hưởng) để xây dựng cây hiện tại, nhưng ngược lại cách xây dựng cây của GRF những thuộc tính đã được sử dụng xây dựng cây trước sẽ không được sử dụng lại (không ảnh hưởng) để xây dựng cây hiện tại. Các thuộc tính được sử dụng trong mô hình GRRF là có liên quan đến biến đích và không lựa chọn lặp lại (những gen có chức năng tương tự) trong khi các đặc trưng được sử dụng trong một mô hình GRF là có liên quan đến biến đích và có thể lựa chọn lặp lại (các gen có thể được chọn lại hoặc chứa các gen có chức năng tương tự).

4. DỮ LIỆU THỰC NGHIỆM VÀ PHƯƠNG PHÁP ĐÁNH GIÁ

4.1. Dữ liệu thực nghiệm

Để đánh giá hiệu quả của các phương pháp đã đề cập ở trên chúng tôi tiến hành thực nghiệm trên 8 bộ dữ liệu gen được thu thập từ ngân hàng dữ liệu y sinh (Kent Ridge) và tin sinh (Bioinformatics). Bảng 1 mô tả các bộ dữ liệu gen bao gồm bộ dữ liệu về ung thư máu (ALL-AML_Leukemia, MLL_Leukemia), ung thư vú (Breast Cancer), ung thư đại tràng (Colon Tumor), ung thư phổi (LungCancer-Harvard (dữ liệu lấy từ trường y Havard) và Lung Cancer-Michigan (dữ liệu cung cấp bởi trường đại học Michigan)), khối u phôi ở hệ thần kinh trung ương (Nervous System), và ung thư buồng trứng.

4.2. Phương pháp đánh giá

Trong bài báo này chúng tôi xây dựng rừng với số cây cố định $ntree = 500$ và $mtry = \sqrt{M}$ (M là số thuộc tính của từng bộ dữ liệu) cho cả 4 mô hình RF truyền thống, GRF, RRF và GRRF (tham số *mtry* là tham số tối ưu theo (Breiman, 2001)). Với mô hình GRRF, chúng tôi lần lượt kiểm thử với tham số gamma lần lượt là $\gamma = 0,5$, và $\gamma = 0,1$. Còn mô hình GRF, chúng tôi sử dụng hệ số phạt tối đa tức là $\gamma = 1$ để thu được một số lượng nhỏ nhất các thuộc tính có thể. Phương pháp tiến hành kiểm thử được liệt kê trong cột "Phương pháp kiểm thử" ở bảng 1. Cụ thể với 4 bộ dữ liệu ALL-AML_Leukemia, MLL_Leukemia, Breast Cancer, Lung Cancer- Harvard

Bảng 1. Mô tả các tập dữ liệu gen

Tên tập dữ liệu	Số phần tử	Số chiều	Số lớp	Phương pháp kiểm thử

ALL-AML_Leukemia	72	7.129	2 (ALL, AML)	Train-Test
MLL_Leukemia	72	12.582	3 (ALL, MLL, AML)	Train-Test
Breast Cancer	97	24.481	2 (Relapse, non-relapse)	Train-Test
Colon Tumor	62	2.000	2 (Negative, positive)	Hold-out (OOB)
Lung Cancer-Harvard	181	12.533	2 (ADCA, Mesothelioma)	Train-Test
Lung Cancer-Michigan	96	7.129	2 (Normal, Tumor)	Hold-out (OOB)
Nervous System	60	7.128	2 (Class0, Class1)	Hold-out (OOB)
Ovarian-PBSII-061902	255	15.154	2 (Cancer, Normal)	Hold-out (OOB)

là những bộ dữ liệu có sẵn tập huấn luyện và tập thử, chúng tôi dùng tập huấn luyện để xây dựng mô hình với các tham số đã nêu ở trên. Sau đó, dùng mô hình thu được để phân lớp tập thử. Những bộ dữ liệu còn lại không có sẵn tập học và tập thử, chúng tôi sử dụng phương pháp hold-out: 2/3 tập dữ liệu để huấn luyện và 1/3 dữ liệu còn lại để kiểm thử. Để so sánh hiệu năng của các phương pháp chúng tôi sử dụng độ chính xác Acc được tính bởi công thức sau:

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

Trong đó, I(.) là *indicator function* và $Q(d_i, j) = \sum_{k=1}^{K} I(\hat{h}_k(d_i) = j)$ là số lượng cây quyết
định lựa chọn d_i thuộc vào lớp j.

Chúng tôi tiến hành thực nghiệm trên máy tính IntelR Core i7 3.40 GHz, bộ nhớ chính 32GB với các gói phần mềm RF, GRRF phiên bản mới nhất được cài đặt trên môi trường R. Mỗi thử nghiệm được chạy 30 lần sau đó lấy trung bình độ chính xác và trung bình số lượng thuộc tính được chọn để xây dựng cây.

5. KẾT QUẢ VÀ THẢO LUẬN

Bảng 2 chỉ ra số lượng các thuộc tính được lựa chọn tương ứng của từng bộ dữ liệu với các mô hình phân lớp RF khác nhau (số lượng các thuộc tính (các gen) được chọn chia trung bình sau 30 lần chạy) khi với tham số $\gamma = 0{,}1$ (GRRF), $\gamma = 1$ (GRF). Cũng từ kết quả của bảng 2 cho thấy, khi chúng ta quan tâm đến độ chính xác phân lớp thì phương pháp GRF cho kết quả tốt hơn trên 7/8 bộ dữ liệu, đặc biệt có những bộ dữ liệu GRF cho kết quả phân lớp chính xác 100% (bộ dữ liệu số 6), nhưng phương pháp RRF thì chỉ đạt 83,56%, và 85,33% với phương pháp GRRF. Ngược lại, số thuộc tính được lựa chọn của GRF lại nhiều hơn đáng kể so với mô hình RRF và mô hình GRRF. Tuy nhiên khi chúng ta so sánh với số chiều ban đầu của các bộ dữ liệu thì số gen được chọn để xây dựng cây trong rừng GRF nhỏ hơn rất nhiều. Với phương pháp GRRF (khi chọn tham số $\gamma = 0{,}1$) thì số lượng gen được lựa chọn lớn hơn số lượng thuộc tính được lựa chọn của mô hình GRF, nhưng kết quả phân lớp của GRRF tốt hơn trên tất cả các tập dữ liệu so với của phương pháp GRF.

Bảng 3 là kết quả trung bình của 30 lần chạy kiểm tra để so sánh mức độ chính xác dự đoán của cả bốn mô hình rừng ngẫu nhiên RF, GRF, RRF và GRRF khi thay đổi hệ số điều khiển độ quan trọng thuộc tính γ ($\gamma = 0{,}5$) (tất cả các phương pháp đều được chạy với tham số cố định $mtry = \sqrt{M}$, $ntree = 500$ cây). Kết quả bảng 3 cho thấy rằng GRF vượt trội về độ chính xác trong dự đoán (7/8 bộ dữ liệu). Khi so sánh trực tiếp với mô hình RRF, ta nhận thấy rằng GRRF sử dụng số lượng thuộc tính rất ít để xây dựng cây, nhưng độ chính xác phân lớp vẫn tốt hơn 6/8 bộ dữ liệu. Từ những kết quả thực nghiệm đã liệt kê ở bảng 2 và bảng 3, khi hiệu chỉnh tham số γ ($\gamma = 0{,}1$) nhỏ thì mô hình GRRF có độ phân lớp chính xác cao hơn so với phương pháp

Bảng 2. Độ chính xác phân lớp dữ liệu biểu hiện gen và số lượng thuộc tính lựa chọn được (#Gen) để xây dựng từng mô hình với số cây trong rừng là 500 và $\gamma = 0{,}1$ (GRRF), $\gamma = 1$ (GRF)

Số chiều	RF	GRF	RRF	GRRF

	Acc	#Gen	Acc	#Gen	Acc	#Gen	Acc	#Gen
7.129	78,63%	813	85,78%	302	79,90%	5	89,71%	7
12.582	78,95%	2814	82,10%	708	61,05%	27	67,37%	64
24.481	80,10%	977	80,70%	213	77,87%	11	79,26%	23
2.000	99,33%	635	99,33%	334	89,15%	3	91,54%	5
12.533	97,07%	551	98,39%	264	93,79%	3	97,78%	4
7.129	100%	1681	100%	461	83,56%	8	85,33%	12
7.128	59,51%	1443	59,13%	375	56,85%	15	58,77%	36
15.154	98,40%	2092	98,32%	532	94,58%	7	97,86%	8

Bảng 3. Độ chính xác phân lớp dữ liệu biểu hiện gen và số lượng thuộc tính lựa chọn được (#Gen) để xây dựng từng mô hình với số cây trong rừng là 500 và γ = 0,5 (GRRF), γ = 1 (GRF)

Số chiều	RF		GRF		RRF		GRRF	
	Acc	#Gen	Acc	#Gen	Acc	#Gen	Acc	#Gen
7.129	78,43%	816	85,88%	302	81,18%	6	87,16%	4
12.582	81,75%	2818	80,88%	711	61,23%	27	57,19%	7
24.481	77,85%	991	80,32%	212	74,84%	11	77,59%	5
2.000	99,33%	625	99,40%	332	89,73%	3	90,54%	3
12.533	96,42%	566	98,09%	266	92,46%	3	97,08%	3
7.129	99,78%	1673	100,00%	464	85,11%	8	84,89%	5
7.128	61,24%	1459	61,11%	373	57,04%	16	57,77%	6
15.154	98,17%	2101	97,81%	529	95,73%	7	96,53%	4

RRF nhưng số lượng thuộc tính được lựa chọn của mô hình GRRF lại lớn hơn so với số lượng thuộc tính được chọn bởi mô hình RRF. Khi hiệu chỉnh tham số γ tăng lên (γ = 0,5) thì độ chính xác của mô hình GRRF đồng thời số lượng thuộc tính lựa chọn để xây dựng rừng cũng giảm đi.

Như vậy, từ phân tích ở trên chúng ta nhận thấy rằng các mô hình rừng ngẫu nhiên cải tiến RRF, GRRF đã tìm ra được tập con các gen có ý nghĩa cho việc phân lớp. Tập #Gen này có số chiều nhỏ hơn rất nhiều so với tập gen ban đầu nhưng mô hình GRRF vẫn cho kết quả phân lớp khá tốt, kết quả này cho thấy rằng phương pháp này phù hợp với các kiểu dữ liệu có số chiều lớn nhưng số mẫu nhỏ. Nhưng khi chúng ta quan tâm đến độ chính xác của mô hình phân lớp hơn với việc tìm ra tập gen có ý nghĩa thì mô hình GRF là lựa chọn tối ưu.

6. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày các phương pháp rừng ngẫu nhiên cải tiến (rừng ngẫu nhiên điều hòa, rừng ngẫu nhiên điều hòa có điều hướng, rừng ngẫu nhiên có điều hướng. Những phương pháp phân lớp này phù hợp với bài toán có số chiều cao nhưng cỡ mẫu lại nhỏ hơn rất nhiều so với số chiều, đây chính là đặc thù của các bài toán phân loại dữ liệu biểu hiện gen. Kết quả thực nghiệm trên 8 bộ dữ liệu gen khác nhau cho chúng ta thấy tùy thuộc vào mục tiêu bài toán mà chúng ta chọn phương pháp phân lớp cho thích hợp: Khi chúng ta quan tâm độ chính xác phân lớp của mô hình hơn tập gen có ý nghĩa thì GRF là giải pháp phù hợp; ngược lại trong trường hợp chúng ta mong muốn tìm ra những gen có ảnh hưởng đến biến đích với số lượng ít nhất thì GRRF là mô hình phù hợp hơn cả trong các mô hình được đề cập ở trên.

TÀI LIỆU THAM KHẢO

Bioinformatics Research Group, http://eps.upo.es/bigs/datasets.html.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole

Advanced Books & Software. ISBN 978-0-412-04841-8.

Breiman L. (2001). Random forests. Machine Learning, 45(1): 5-32.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P., Van Eerdewegh, P. (2005). Identifying snps predictive of phenotype using random forests. Genetic epidemiology, 28(2): 171-182.

Bø TH., Jonassen I. (2002). New feature subset selection procedures for classification of expression profiles. Genome Biology, 3(4): 0017.1-0017.11.

Deng H. and G. Runger (2013). Gene selection with guided regularized random forest. Journal of Pattern Recognition, 46: 3483-3489.

Deng H and G. Runger (2012). Feature selection via regularized trees. International Joint Conference on Neural Networks (IJCNN).

Deng H. (2013). Guided random forest in the RRF package, http://arxiv.org/abs/1306.0237.

Díaz-Uriarte R. (2005). Supervised methods with genomic data: a review and cautionary view. In Data analysis and visualization in genomics and proteomics. Edited by Azuaje F, Dopazo J. New York: Wiley, pp.193-214.

Dudoit S, Fridlyand J, Speed TP (2002). Comparison of discrimination methods for the classification of tumors suing gene expression data. J Am Stat Assoc., 97(457): 77-87.

Furlanello C, Serafini M, Merler S, Jurman G: An accelerated procedure for recursive feature ranking on microarray data. Neural Netw, 16: 641-648.

Goldstein B. A., Hubbard, A. E., Cutler, A., Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: Methodological considerations and new findings.BMC Genetics, 11: 49.

Goldstein B. A., Polley, E. C. Briggs, Farren B. S. (2011). Random Forests for Genetic Association Studies. Statistical Applications in Genetics and Molecular Biology, 10(1): 32.

Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER (2005). Optimal number of features as a function of sample size for various classification rules. Bioinformatics, 21: 1509-1515.

Kent Ridge Bio-medical Dataset, http://datam.i2r.a-star.edu.sg/datasets/krbd/

Jirapech-Umpai T, Aitken S (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. BMC Bioinformatics, 6: 148.

Lee JW, Lee JB, Park M, Song SH (2005). An extensive evaluation of recent classification tools applied to microarray data. Computation Statistics and Data Analysis, 48: 869-885.

Lunetta, K.L., Hayward, L.B., Segal, J., Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. BMC genetics, 5(1): 32.

Li Y, Campbell C, Tipping M (2002). Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics, 18: 1332-1339.

Li T, Zhang C, Ogihara M (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20: 2429-2437.

Roepman P, Wessels LF, Kettelarij N, Kemmeren P, Miles AJ, Lijnzaad P, Tilanus MG, Koole R, Hordijk GJ, van der Vliet PC, Reinders MJ, Slootweg PJ, Holstege FC (2005). An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. Nat Genet, 37: 182-186.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415: 530-536.

Yang Q. and X. Wu (2006). Challenging Problems in Data Mining Research. Journal of Information Technology and Decision Making 5(4): 597-604.

Yeung KY, Bumgarner RE, Raftery AE (2005). Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics, 21: 2394-2402.

Wiener M. and A. Liaw (2002). "Classification and regression by randomforest," The Journal of R news, 2(3): 18-22.

Winham, S.J., Colby, C. L., Freimuth, R., Wang, X., Andrade, M., Huebner, M., Biernacka, J. M. (2012). SNP interaction detection with Random Forests in high-dimensional genetic data. BMC Bioinformatics, 13: 164.