ĐÁNH GIÁ HIỆU QUẢ PHÂN LỚP DỮ LIỆU GENE CHIỀU CAO DỰA TRÊN RỪNG NGẪU NHIÊN, SVM VÀ KẾT HỢP PHƯƠNG PHÁP CHỌN ĐẶC TRƯNG RỪNG NGẪU NHIÊN ĐIỀU HƯỚNG

Hoàng Thị Hà

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

Email: htha@vnua.edu.vn

Ngày gửi bài: 15.02.2017 Ngày chấp nhận: 08.01.2018

TÓM TẮT

Phân lớp dữ liệu là phương pháp phổ biến được sử dụng để tìm kiếm các tri thức tiềm ẩn từ cơ sở dữ liệu lớn. Trong số nhiều mô hình phân lớp dữ liệu, các mô hình Rừng ngẫu nhiên và SVM nổi lên là những công cụ phân lớp rất hiệu quả với dữ liệu có số chiều cao. Hiện nay, có nhiều phiên bản của Rừng ngẫu nhiên đã được đề xuất. Tuy nhiên, khi phân tích dữ liệu gene cỡ hàng nghìn đặc trưng, các phương pháp dựa trên mô hình Rừng ngẫu nhiên và SVM vẫn hạn chế. Lý do là dữ liệu gene chứa rất nhiều nhiễu. Chính vì vậy, kết hợp phương pháp trích chọn đặc trưng với các thuật toán phân lớp dữ liệu sẽ cho kết quả cao hơn. Phương pháp lựa chọn đặc trưng của GRF dựa trên RF được đề xuất bởi Deng và Runger được đánh giá là phương pháp trích chọn đặc trưng cho độ chính xác cạnh tranh so với GRRF, RRF, varSelRF, LASSO. Bài báo này, chúng tôi tóm tắt các thuật toán phân lớp dữ liệu dựa trên mô hình Rừng ngẫu nhiên, SVM và đánh giá hiệu quả phân lớp dữ liệu chiều cao của các thuật toán này. Tiếp theo, chúng tôi kết hợp chọn đặc trưng của GRF với các bộ phân lớp RF, WSRF, RUF, SVM. 7 tập dữ liệu gene được sử dụng để đánh giá các thuật toán. Kết quả thực nghiệm cho thấy, việc kết hợp này không những làm tăng độ chính xác mà còn giảm thời gian thực hiện của các thuật toán.

Từ khóa: Phân lớp dữ liệu, phân lớp dữ liệu chiều cao, rừng ngẫu nhiên, trích chọn đặc trưng.

Evaluation of Data Classification Methods Base on Random Forest, SVM before and after Using Feature Selection Method Guided Random Forest for High Dimensional Gene Data

ABSTRACT

Data classification is a common method used for mining the potential knowledge from large databases. Several methods have been proposed such as AdaBoost, Support Vector Machine (sVM), Neural Network, random forest (RF), C45… Among all these classifiers, Random forest and SVM provided more accurate and efficient result for high-dimensional data. However, the performance of these algorithms may be affected/degraded when working on gene data with thousands of features. This is because gene data usually contain many redundant features which are uninformative to classification. Therefore, using the subsets of selected genes may give a better performance than using all the features. In this study, we summarized the data classification algorithms based on the random forest models, SVM, and evaluate the effectiveness of the algorithms for classifying high-dimensional data. Next, we evaluated the combination the features selection using GRF and other classifiers such as RF, WSRF, RUF, SVM. Seven gene data sets were used to evaluate the methods. Experimental results showed that the combination not only increased the accuracy but also decreased the execution time of the algorithms

Keywords: Classification, high dimensional data, machine learning, random forest, feature selection.

1. ĐẶT VẤN ĐỀ

Cùng với sự ra đời của nhiều công cụ thu thập dữ liệu tự động, một lượng dữ liệu khổng lồ cả về số bản ghi và số thuộc tính đã được tạo ra (Rea, 1995). Đặc biệt, công nghệ sinh học đã đạt được những bước tiến vượt bậc trong công nghệ giải mã trình tự gene. Giờ đây, toàn bộ hệ gene có thể được giải mã

trình tự dễ dàng và nhanh chóng với chi phí thấp (Mardis, 2011). Việc tìm ra tri thức tiềm ẩn trong cơ sở dữ liệu khổng lồ đang là nhiệm vụ đặt ra cho lĩnh vực khai phá dữ liệu nói chung và học máy nói riêng. Trong đó, bài toán phân lớp dữ liệu chiều cao đang là một thách thức và nhiệm vụ chính hiện nay trong lĩnh vực học máy (Xu, 2012.)

Nhiều kỹ thuật phân lớp dữ liệu đã được đề xuất như Naïve Bayes, Cây quyết định (IC3, C4.5, C50), Mạng noron, Máy vector hỗ trợ SVM, Rừng ngẫu nhiên,… Trong các phương pháp đó, lớp thuật toán dựa trên mô hình Rừng ngẫu nhiên và SVM 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), Neural Network, C45,... (Manuel *et al*., 2014). Tuy nhiên, khi phân tích dữ liệu gene cỡ hàng nghìn đặc trưng, các thuật toán dựa trên Rừng ngẫu nhiên và SVM vẫn cho kết quả hạn chế. Nguyên nhân là do dữ liệu gene chứa nhiều biến nhiễu, dung lượng mẫu ít, điều đó sẽ dẫn đến các mô hình Rừng ngẫu nhiên chọn phải các thuộc tính không chứa nhiều thông tin đến lớp để phân lớp. Chính vì vậy, việc kết hợp phương pháp trích chọn đặc trưng với các thuật toán phân lớp dữ liệu hiệu quả sẽ cho kết quả cao hơn.

Theo Deng (2013), kết quả phân lớp của phương pháp GRF (Guided Random Forest) cho độ chính xác cạnh tranh hơn GRRF (Guided Regularized Random Forest). Hơn nữa, lựa chọn đặc trưng của GRRF cho độ chính xác cao, cao hơn so với RRF(RF), varSelRF(RF), LASSO(RF). Điều đó chứng tỏ trong quá trình phân lớp, GRF đã chọn được những tập thuộc tính tốt, tốt hơn GRRF để phân lớp.

Vì vậy, mục tiêu của bài báo này là đánh giá hiệu quả các phương pháp phân lớp dữ liệu dựa trên mô hình Rừng ngẫu nhiên như RF, WSRF, RUF và SVM cho bài toán phân lớp dữ liệu chiều cao. Sau đó kết hợp phương pháp trích chọn đặc trưng của GRF với các bộ phân lớp RF, WSRF, RUF, SVM.

Bài báo trình bày tổng quan các phương pháp phân loại dữ liệu dựa trên mô hình rừng ngẫu nhiên, phương pháp SVM, phương pháp trích chọn đặc trưng GRF. Sau đó đánh giá hiệu quả phân lớp của các thuật toán RF, WSRF, RUF, SVM trước và sau khi sử dụng phương pháp trích chọn đặc trưng GRF. Dữ liệu thực nghiệm là 7 tập dữ liệu gene có chiều rất cao, nhiều nhiễu và dung lượng mẫu ít.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Mục này trình bày vật liệu nghiên cứu và tóm tắt các phương pháp phân lớp dữ liệu dựa trên mô hình rừng ngẫu nhiên, phương pháp SVM và phương pháp kết hợp trích chọn đặc trưng GRF với các bộ phân lớp khác.

2.1. Vật liệu và thực nghiệm

2.1.1. Dữ liệu thực nghiệm

Dữ liệu chúng tôi chạy thử nghiệm là 7 tập dữ liệu gene thực có số chiều rất lớn lấy tại địa chỉ *http://www.gems system.org* được chúng tôi liệt kê trong bảng 1 theo chiều tăng dần của số thuộc tính.

Bảng 1. Tóm tắt 7 tập dữ liệu genes

Tập dữ liệu	Số mẫu	Số gene	Số lớp
Leukemia1	72	5327	3
Brain-tumor1	90	5920	5
Prostate-tumor	102	10,509	2
Leukemia2	72	11,225	3
Lung-cancer	203	12,600	5
11-tumors	174	12,533	11
GCM	190	16,063	14

2.1.2. Thực nghiệm

Môi trường thực nghiệm: Chúng tôi tiến hành thực nghiệm trên máy tính Windows 64-bit, Intel, Core™ i5-4460, CPU@3.20GHz

Các gói Randomforest, RUF, wsRF, SVM (https://cran.r-project.org) 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 độ lệch chuẩn.

2.2. Rừng ngẫu nhiên

Rừng ngẫu nhiên (Random forest - RF) (Breiman, 2001) là mô hình học tập thể của nhiều cây quyết định không cắt nhánh.

Với bài toán phân lớp*:* Cho một tập dữ liệu huấn luyện $D = \{(d_i)\}_{i=1}^{N} = \{(x_i, y_i)\}_{i=1}^{N}$, với x_i là vector M chiều, $y_i \in Y$ (trong đó, Y là lớp, giả sử có C nhãn lớp $Y \in \{1, 2, .., C\}$ ($C \geq 2$). Ý tưởng chính của mô hình RF là lựa chọn ngẫu nhiên 2 lần (ngẫu nhiên mẫu và ngẫu nhiên thuộc tính) trong suốt quá trình xây dựng cây như sau:

Bước 1: Từ tập dữ liệu ban đầu D, sử dụng kỹ thuật boostrap (lấy mẫu ngẫu nhiên có hoàn lại) để tạo ra t tập dữ liệu con $S = \{S_1, S_2..., S_t\}$.

Bước 2: Trên mỗi tập dữ liệu S_j, xây dựng một cây quyết định h_j. Mô hình Rừng ngẫu nhiên là mô hình $h = \{h_j\}_{j=1}^{t}$. Thay vì sử dụng tất cả các biến là biến ứng cử để lựa chọn điểm chia tốt nhất, tại mỗi nút RF chọn ngẫu nhiên một không gian tập con M' thuộc tính từ M thuộc tính ban đầu (M'<<M). Bên cạnh đó, cây quyết định trong mô hình RF là cây quyết định không cắt nhánh.

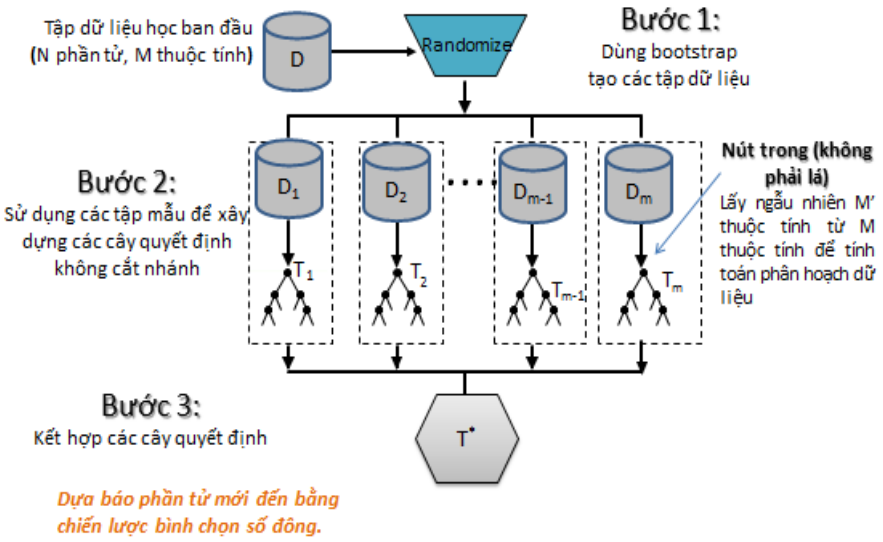
Bước 3: RF dự đoán nhãn lớp của phần tử mới đến bằng chiến lược bình chọn số đông của các cây quyết định.

Ưu điểm của RF là xây dựng cây không thực hiện việc cắt nhánh từ các tập dữ liệu con khác nhau, do đó thu được những cây với lỗi bias thấp. Bên cạnh đó, mối tương quan giữa các cây quyết định cũng được giảm xuống nhờ việc xây dựng các không gian con thuộc tính một cách ngẫu nhiên. 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 trong rừng.

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 *M'* 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 có điểm cắt làm cực tiểu độ hỗn tạp của các tập mẫu sau khi chia. Công thức tính chỉ số *Gini* cho nút *t* như sau:

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

$\text{Gini}(t) = \sum_{c=1}^{C} \Phi_c(t)[1 - \Phi_c(t)]$ trong đó: *Φc(t)* là tần suất xuất hiện của lớp $c \in C$ trong nút *t*.



Hình 1. Mô hình Rừng ngẫu nhiên

Gọi *s* là một giá trị của thuộc tính X_j. Giả sử tách nút *t* thành 2 nút con: nút trái t_L và nút phải t_R tại *s*. Tùy thuộc vào $X_j \le s$ hoặc $X_j > s$ ta có 2 nút con: $t_L = \{X_j \in t, X_j \le s\}$ 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:

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

Để đạt được điểm chia tốt, tại mỗi nút RF sẽ tìm tất cả các giá trị phân biệt của tất cả *M'* thuộc tính để tìm ra điểm phân tách nút *t* (điểm *s* có độ đo ΔGini(s, t) nhỏ nhất). 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÷m) và trong một rừng ngẫu nhiên. Công thức tính $IS_k(X_j)$ và IS_{X_j} như sau:

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

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

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

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

Kết quả dự đoán của mô hình rừng ngẫu nhiên là kết hợp kết quả của một số lượng lớn những cây quyết định có mối tương quan thấp (do RF lấy ngẫu nhiên mẫu và xây dựng các không gian con thuộc tính cũng ngẫu nhiên) nên RF đạt được cả độ lệch thấp và phương sai thấp. Trong thực tế RF đã trở

thành một công cụ tin cậy cho phân tích dữ liệu chiều cao. Tuy nhiên, tiếp cận cài đặt ban đầu, RF chỉ cho kết quả tốt trên các dữ liệu có số chiều vừa phải và giảm đáng kể hiệu năng khi xử lý dữ liệu có số rất chiều cao cỡ hàng nghìn thuộc tính, nhiều nhiễu, dung lượng mẫu ít (bài toán phân tích dữ liệu gene là một trường hợp cụ thể). 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.

Chính vì vậy, đã có nhiều đề xuất cho việc cải tiến mô hình Rừng ngẫu nhiên. Dưới đây sẽ trình bày tóm tắt một số phương pháp cải tiến mô hình Rừng ngẫu nhiên.

2.3. Một số mô hình rừng ngẫu nhiên cải tiến

2.3.1. Mô hình WSRF

WSRF(Weighted Subspace Random Forest) (Xu, 2012) là phương pháp mở rộng mô hình RF nhằm giải quyết bài toán phân lớp dữ liệu chiều rất cao (cỡ hàng nghìn thuộc tính) và nhiều nhiễu. Điểm khác biệt của wsRF so với RF nguyên bản là việc lựa chọn không gian con thuộc tính M' tại mỗi nút. Với WSRF, tại mỗi nút các thuộc tính được đánh trọng số giúp cho việc lựa chọn không gian con M' của RF được tốt hơn.

Phương pháp đánh trọng số của wsRF được trình bày như sau:

Bước 1: WSRF sử dụng một trong 2 phương pháp *thống kê chi-square(CS)* hoặc *information gain ratio(IGR)* để đo độ lợi thông tin giữa các biến X_j với lớp Y. Giá trị *thống kê chi-square* của biến X_j với lớp Y được ký hiệu: $Corr_{CS}(X_j,Y)$. Gía trị *information gain ratio* của biến X_j với lớp Y được ký hiệu: $Corr_{IGR}(X_j,Y)$. Giá trị $Corr_{CS}(X_j,Y)$ hoặc $Corr_{IGR}(X_j,Y)$ càng lớn thì chứng tỏ biến X_j càng chứa nhiều thông tin cho việc phân tách lớp Y.

Bước 2: Chuẩn hóa $Corr_{CS}(X_j,Y)$ hoặc $Corr_{IGR}(X_j,Y)$ theo công thức dưới ta được trọng số của biến X_j.

$$w_j = \frac{\sqrt{corr(X_j,Y)}}{\sum_{i=1}^{M}\sqrt{corr(X_j,Y)}} \quad (6)$$

Trong đó, $corr(X_j, Y)$ có thể là $Corr_{IGR}(X_j,Y)$ hoặc $Corr_{CS}(X_j,Y)$. Các w_j này được xem như xác suất để lựa chọn thuộc tính vào một không gian con M' thuộc tính.

Tại mỗi nút, wsRF sử dụng phương pháp "weighted sampling" để lựa chọn ngẫu nhiên M' thuộc tính. Những thuộc tính có trọng số cao sẽ có cơ hội được chọn nhiều hơn.

Với cách tiếp cận trên, độ chính xác của các mô hình Rừng ngẫu nhiên có thể được cải thiện, kết quả ổn định hơn mà không cần tăng kích thước không gian thuộc tính.

2.3.2. Rừng ngẫu nhiên điều hòa

Năm 2012 Deng và Runger (Deng & Runger, Feature selection via regularized trees, 2012) 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).

Bằng thực nghiệm, Deng và Runger cho thấy tiếp cận RRF trên các tập dữ liệu có dung lượng mẫu cao, số chiều nhỏ cho kết quả rất tốt. 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 đối với các tập dữ liệu chiều cao, nhiều nhiễu.

2.3.3. Mô hình RUF

RUF (Random Uniform Forest) (Ciss, 2015) là một hướng tiếp cận khác của mô hình RF. Không giống RF, trong quá trình dựng cây quyết định, tại mỗi nút RUF vẽ ra vùng cắt ngẫu nhiên sử dụng phân phối đều liên tục. Điểm cắt được chọn là một điểm tối ưu trong tập các điểm cắt ngẫu nhiên. RUF hướng tới giảm mối tương quan giữa các cây trong rừng.Vì vậy, tiếp cận này có thể giảm bias cho mô hình Rừng ngẫu nhiên.

Ưu điểm của RUF:

- Tiếp cận RUF hướng tới mục tiêu giảm mối tương quan giữa các cây rừng trong khi vẫn giữ được những ưu điểm của mô hình RF nguyên bản.

- Mô hình RUF thường có trung bình lỗi bình phương thấp hơn so với RF.

- RUF ít nhạy cảm với "overfitting" mà không làm tăng phương sai trung bình của các cây.

- Hội tụ nhanh hơn mô hình RF nguyên bản.

2.3.4. Mô hình GRF

GRF (Guided Random Forest) (Deng H. , 2013) là mô hình phân lớp dựa trên rừng ngẫu nhiên. Trong GRF cách tính độ quan trọng của mỗi thuộc tính đã có sự thay đổi. GRF 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, ngược lại gán giá trị λ thấp.

Với cách tính này, GRF chọn được ít thuộc tính hơn RF. Hơn nữa, nếu áp dụng RF trên các thuộc tính được trích chọn bởi GRF thì RF cho kết quả cao hơn RF nguyên bản.

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

$$\Delta Gini_R(X_j, t) = \lambda_j \Delta Gini(X_j, t) \quad (7)$$

$\lambda_j \in (0,1]$ 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:

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

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.

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

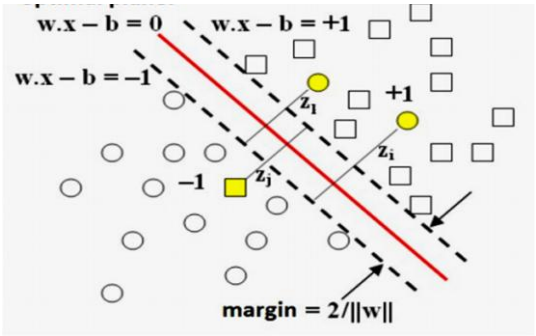
$$\lambda_j = (1-\gamma) + \gamma VI_{X_j} = 1 - \gamma(1 - VI_{X_j}) \quad (9)$$

Để có được tập thuộc tính trích chọn nhỏ nhất, George Runger chọn $\gamma = 1$.

Với phương pháp này GRF có thể chọn được những thuộc tính có nghĩa, chứa nhiều thông tin liên quan đến lớp.

2.4. Máy học véc tơ hỗ trợ

SVM (support Vector machine) (Vapnik, 1995) là mô hình hiệu quả và phổ biến cho bài toán phân dữ liệu có số chiều lớn. Bài toán cơ bản của SVM là phân loại 2 lớp: cho trước tập



Hình 2. Minh họa mô hình phân lớp SVM cơ bản (Đỗ Thanh Nghị, 2013)

dữ liệu huấn luyện $D=\{(d_i)\}_{i=1}^N=\{(x_i, y_i)\}_{i=1}^N$, với x_i là vector M chiều, $y_i \in Y$ (trong đó, Y là nhãn lớp, giả sử có 2 nhãn lớp $Y \in \{+1, -1\}$)

Mục đích của giải thuật SVM là cố gắng tìm một siêu phẳng (hyperplane) tách các điểm này thành hai lớp riêng biệt tương ứng với nhãn của chúng. Có thể giải thích SVM rõ hơn như sau: trong không gian M chiều cần tìm siêu phẳng H: y = w.x - b = 0 và hai siêu phẳng H1, H2 hỗ trợ song song với H và có cùng khoảng cách đến H. Trong đó, H chia đôi tập dữ liệu D thành 2 phần, mỗi phần gồm toàn các đối tượng chung một nhãn +1 hay -1 và không có phần tử nào của tập mẫu nằm giữa H1 và H2, khi đó:

w.x - b >= +1 với y = +1

w.x - b >= -1 với y = -1

Kết hợp hai điều kiện trên ta có:

y(w.x - b) >= 1

Phân lớp phần tử $\boldsymbol{x}$ dựa vào dấu của *(w.x - b).* Nếu giá trị biểu thức *(w.x- b) > 0* thì gán nhãn cho $\boldsymbol{x}$ là lớp dương (lớp +1), ngược lại thì gán nhãn cho $\boldsymbol{x}$ là lớp âm (lớp -1).

predict(x) = sign(w.x - b)

Khoảng cách giữa hai siêu phẳng hỗ trợ gọi là biên (margin) và bằng $2/|w|$, trong đó $|w|$ là độ lớn (2-norm) của pháp vector *w.* Trường hợp dữ liệu không khả tách tuyến tính (linearly separable), ta xem mỗi phần tử nằm sai phía so với mặt phẳng hỗ trợ tương ứng với lớp của chúng là lỗi, khoảng cách từ phần tử lỗi đến siêu phẳng hỗ trợ được kí hiệu z_i ($z_i >= 0$). Vì thế, SVM phải đồng thời cực đại hoá lề (margin) và cực tiểu hoá lỗi.

Biến thể của giải thuật SVM sử dụng các hàm phân lớp khác nhau. Để có thể có hàm phân lớp khác, ta không cần thay đổi giải thuật mà chỉ cần thay đổi hàm nhân tuyến tính bằng các hàm nhân khác. Bằng cách này ta thu được các mô hình phân lớp dựa trên các véc tơ hỗ trợ khác nhau. Một số hàm nhân thường dùng được cho trong bảng 1.

2.5. Kết hợp phương pháp trích chọn đặc trưng của GRF với các bộ phân lớp RF, WSRF, RUF, SVM

Lựa chọn đặc trưng giúp ta có thể chọn được những biến chứa nhiều thông tin cho việc phân lớp. Đặc biệt, với những tập dữ liệu có số chiều rất lớn, mà đa phần các biến không có ý nghĩa cho việc phân lớp như dữ liệu gene, nếu không trích chọn đặc trưng thì

không những tốn thời gian thực hiện mà độ chính xác cũng sẽ không cao. Do đó, thông thường, đối với những tập dữ

Bảng 2. Một số hàm nhân thường dùng

Type of kernel function	Formular
Linear kernel	K<x,y>=x.y
Polynomial kernel	$K<x,y>=(x.y+1)^d$
Radial basis function (Gaussian) kernel	$K(x,y) = e^{\frac{-\|x-y\|^2}{2\sigma^2}}$

Phương pháp GRF_T (T là tên một thuật toán phân lớp)
Đầu vào:
- Tập dữ liệu huấn luyện D có N mẫu và M thuộc tính
- Thuật toán phân lớp T dùng kết hợp với GRF
Đầu ra:
- Bộ phân lớp kết hợp **GRF_T** có khả năng phân loại các phần tử mới.
Bắt đầu
- Từ tập dữ liệu ban đầu, sử dụng GRF chọn ra M' thuộc tính có độ quan trọng cao, từ đó ta tạo 1 tập dữ liệu con D' có N mẫu và M' thuộc tính.
- Sử dụng thuật toán phân lớp T trên tập dữ liệu huấn luyện D' vừa tìm được để thu được bộ phân lớp đối với D'.
- Dùng bộ phân lớp thu được ở trên (bước 2), để dự đoán nhãn lớp cho phần tử mới.
Kết thúc

liệu chiều cao, nhiều nhiễu như dữ liệu gene, việc chọn các biến tốt cho quá trình phân tách lớp trước khi sử dụng các bộ phân lớp là thực sự cần thiết.

Như đã trình bày trong mục 2.3.4, GRF cho kết quả cạnh tranh hơn GRRF (Deng, 2013). Điều đó chứng tỏ GRF có thể chọn được những tập thuộc tính tốt hơn GRRF cho việc phân lớp.

Để làm rõ hơn hiệu quả GRF với vai trò là phương pháp trích chọn đặc trưng và đánh giá các bộ phân lớp RF , WSRF, RUF, SVM trên dữ liệu chiều cao, nhiều nhiễu, chúng tôi tiến hành thử nghiệm kết hợp phương pháp trích chọn đặc trưng của GRF với các thuật toán phân lớp RF, wsRF, RUF, SVM trên 7 tập dữ liệu gene cỡ hàng chục nghìn thuộc tính (đã được tóm tắt trong Bảng 1). Phương pháp kết hợp được mô tả như sau:

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

Các tham số mtry (số thuộc tính chọn ngẫu nhiên tại mỗi nút của các mô hình Rừng ngẫu nhiên), ntree (số cây trong rừng) để chạy mô hình RF, WSRF, RUF, GRF_M tương ứng là: $\sqrt{M}$, 500. Trong đó, $mtry = \sqrt{M}$, ntree = 500 là những tham số đã được đề xuất bởi (Breiman, 2001); Tham số γ (tham số điều khiển của mô hình GRF) lần lượt được chạy với các giá trị 0,1; 0,5; 0,8 và 1. Trong đó, $\gamma = 1$ là tham số được đề xuất bởi (Deng, 2013) để GRF có thể chọn được tập đặc trưng có nghĩa nhỏ nhất cho phân lớp. Hàm nhân sử dụng cho SVM là hàm Poly. Mỗi tập dữ liệu đầu vào được chia ngẫu nhiên thành 2 phần: 70% cho phần training và 30% cho testing.

Với mỗi tập dữ liệu thử nghiệm được chạy 30 lần để tạo ra 30 tập dữ liệu con (chỉ chứa các thuộc tính đã được trích chọn đặc trưng). Các thuật toán RF, WSRF, RUF, SVM sử dụng 7 tập dữ liệu gene khác nhau đã được trích chọn thuộc tính và 7 tập dữ liệu gene gốc để đánh giá hiệu quả thực hiện của các mô hình. Sau khi chạy các thuật toán, độ đo *Acc* được sử dụng để đánh giá độ chính xác của các mô hình trên tập các dữ liệu kiểm thử. Độ chính xác kiểm thử *Acc* được tính theo công thức:

$$Acc = \frac{\sum_{i=1}^{k} d_i}{N} \qquad (10)$$

Trong đó, d_i là đối tượng được thuật toán phân lớp đúng, N là tổng số mẫu trong tập dữ liệu huấn luyện.

Chúng tôi sử dụng phương pháp paired t-test, mức ý nghĩa 0,05 để so sánh sự khác nhau về độ chính xác của các kết quả phân lớp trước và sau khi sử dụng GRF.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Các kết quả của phương pháp phân lớp GRF khi tham số γ thay đổi

Bảng 3 cho thấy khi giá trị γ tăng dần (0,1; 0,5; 0,8; 1) độ chính xác phân lớp của phương

Bảng 3. Trung bình (mean ± std-dev %) độ chính xác phân lớp của GRF khi tham số γ thay đổi, giá trị tốt nhất được tô đậm

Dữ liệu	Trung bình độ chính xác (mean ± std-dev %)			
	GRF			
	GRF(γ=0,1)	GRF(γ=0,5)	GRF(γ=0,8)	GRF(γ=1)
Leukemia1	0,944 ± 0,039	0,922 ± 0,049	0,989 ± 0,025	0,844 ± 0,057
Brain-tumor1	0,835 ± 0,019	0,826 ± 0,043	0,823 ± 0,038	0,814 ± 0,032
Prostate-tumor	0,992 ± 0,017	1,000 ± 0,000	1,000 ± 0,000	0,988 ± 0,018
Leukemia2	0,856 ± 0,034	0,800 ± 0,030	0,789 ± 0,025	0,711 ± 0,053
Lung-cancer	0,899 ± 0,009	0,890 ± 0,010	0,890 ± 0,011	0,884 ± 0,006
11-tumors	0,882 ± 0,019	0,877 ± 0,026	0,840 ± 0,019	0,840 ± 0,030
GCM	0,688 ± 0,000	0,663 ± 0,022	0,667 ± 0,029	0,600 ± 0,025

pháp GRF giảm dần và GRF với tham số $\gamma = 1$ cho độ chính xác thấp nhất. Thực nghiệm cũng cho thấy, khi γ càng tăng, tập thuộc tính mà GRF chọn để phân tách lớp càng nhỏ.

3.2. So sánh các kết quả của RF, WSRF, RRF, GRF, RUF, SVM trên các tập dữ liệu gene ban đầu

Bảng 4 thể hiện kết quả các thuật toán RF, WSRF, RRF, GRF, RUF, SVM trên 7 tập dữ liệu gốc.

Trong bảng 4, cột "**Dữ liệu**" chứa danh sách các tập dữ liệu thực nghiệm. Các cột RF,WSRF, GRF, RRF, RUF, SVM chứa trung bình độ chính xác phân lớp của mỗi tập dữ liệu đã đạt được khi sử dụng các phương pháp phân lớp trên dữ liệu gốc. Ứng với mỗi tập dữ liệu, kết quả có độ chính xác cao nhất được tô đậm.

Kết quả của Bảng 4 cho thấy, mô hình RUF đạt kết quả phân lớp *cao hơn* ba mô hình: RF, GRF($\gamma = 1$) và WSRF với tỷ lệ 6/7 tập dữ liệu. Nguyên nhân là do mô hình RUF thiết lập được mối tương quan giữa các cây trong rừng thấp hơn mô hình RF. Mặt khác, RUF vẫn duy trì được ưu điểm của mô hình RF. Mô hình RRF tỏ ra không hiệu quả, kém hơn RF. Mô hình GRF khi chạy với tham số $\gamma = 1$ với vai trò là bộ phân lớp cho kết quả không khác nhiều so với RF. Nhưng nếu để GRF với vai trò là phương pháp trích chọn đặc trưng kết hợp với các bộ phân lớp khác sẽ cho kết quả đáng ngạc nhiên (xem mục 3.2.)

Bảng 4. Trung bình (mean ± std-dev %) độ chính xác phân lớp của các phương pháp Rừng ngẫu nhiên (tham số mặc định) và phương pháp SVM, giá trị tốt nhất được tô đậm

Dữ liệu	Trung bình độ chính xác (mean ± std-dev %)					
	RF	wsrf	RRF	GRF(γ=1)	ruf	SVM
Leukemia1	0,824 ± 0,045	0,852 ± 0,040	0,778 ± 0,068	0,844 ± 0,057	0,944 ± 0,076	0,944 ± 0,00
Brain-tumor1	0,803 ± 0,024	0,820 ± 0,018	0,660 ± 0,089	0,814 ± 0,032	0,876 ± 0,032	0,889 ± 0,039
Prostate-tumor	0,991 ± 0,017	0,997 ± 0,010	0,713 ± 0,078	0,988 ± 0,018	0,998 ± 0,010	0,941 ± 0,023
Leukemia2	0,734 ± 0,033	0,651 ± 0,050	0,733 ± 0,060	0,711 ± 0,053	0,872 ± 0,032	0,971 ± 0,042
Lung-cancer	0,886 ± 0,008	0,879 ± 0,008	0,727 ± 0,097	0,884 ± 0,006	0,900 ± 0,009	0,941 ± 0,000
11-tumors	0,853 ± 0,018	0,851 ±0,023	0,695 ± 0,082	0,840 ± 0,030	0,841 ± 0,077	0,932 ± 0,000
GCM	0,684 ± 0,018	0,638 ± 0,019	0,567 ± 0,058	0,600 ± 0,025	0,688 ± 0,101	0,708 ± 0,000

Bảng 5. Thời gian thực hiện các các thuật toán phân lớp RF, wsRF, RRF, GRF, RUF, SVM

Tập dữ liệu	Thời gian thực hiện của các thuật toán					
	RF	WSRF	RRF	GRF	RUF	SVM
Leukemia1	2,854 (s)	16,681 (s)	2,872 (s)	6,623 (s)	17,512 (s)	27,717 (mins)
Brain-tumor1	3,833 (s)	28,733 (s)	4,257 (s)	8,810 (s)	24,940 (s)	33,047 (mins)
Prostate-tumor	8,362 (s)	51,222 (s)	8,793 (s)	20,401 (s)	38,271 (s)	1,325 (hours)
Leukemia2	6,303 (s)	32,904 (s)	6,397 (s)	14,313 (s)	34,666 (s)	1,434 (hours)
Lung-cancer	22,310 (s)	2,760 (mins)	24,599 (s)	54,000 (s)	1,384 (mins)	2,297 (hours)
11-tumors	27,704 (s)	3,313 (mins)	27,210 (s)	1,027 (mins)	1,693 (mins)	3,082 (hours)
GCM	42,052 (s)	5,482 (mins)	41,520 (s)	1,547 (mins)	4,512 (mins)	6,514 (hours)

Ghi chú: s - giây; mins - phút; hours - giờ

Bảng 4, ta còn thấy kết quả tô đậm chủ yếu tập trung ở cột SVM, điều đó có nghĩa là mô hình SVM cho kết quả dự đoán cao. Tuy nhiên, thời gian thực hiện của mô hình SVM chậm hơn rất nhiều so với các mô hình khác. Kết quả trong bảng 5 sau đây thể hiện điều đó.

Bảng 5 cho biết, trong các mô hình trên thì mô hình RF nguyên bản thực hiện nhanh hơn tất cả các

mô hình còn lại. Tiếp đó đến RRF, GRF, RUF, WSRF và cuối cùng là SVM. Các mô hình dựa trên RF thực hiện nhanh hơn rất nhiều so với mô hình SVM .

3.3. So sánh các kết quả của RF, WSRF, RUF, SVM trên các tập dữ liệu gene khi kết hợp với phương pháp trích chọn đặc trưng của GRF

Bảng 6 thể hiện số thuộc tính mà GRF trích chọn được trên mỗi tập dữ liệu và trung bình độ chính xác phân lớp của mỗi phương pháp RF, WSRF, RUF, SVM trên các tập dữ liệu sau khi trích chọn đặc trưng.

Trong bảng 6, cột **"Dữ liệu"** chứa danh sách các tập dữ liệu thực nghiệm. Cột **"Số thuộc tính"** chứa số thuộc tính trong tập dữ liệu ban đầu và sau khi sử dụng GRF trích chọn đặc trưng. Các cột GRF_RF,GRF_WSRF, GRF_RUF, GRF_SVM chứa *trung bình độ chính xác phân lớp* ± *độ lệch chuẩn* của phương pháp phân lớp tương ứng trên các tập dữ liệu sau khi dùng GRF trích chọn đặc trưng. Với mỗi tập dữ liệu, kết quả có độ chính xác cao nhất được tô đậm. Và như vậy, phương pháp GRF_SVM có độ chính xác phân lớp cạnh tranh hơn so với các phương pháp còn lại.

Bảng 6. Trung bình (mean ± std-dev %) độ chính xác phân lớp của các phương pháp Rừng ngẫu nhiên và SVM sau khi sử dụng GRF lựa chọn đặc trưng, giá trị tốt nhất được tô đậm

Dữ liệu	Số thuộc tính		Trung bình độ chính xác (mean ± std-dev %)			
	Ban đầu	Sau khi dùng GRF (γ=1)	GRF_RF	GRF_WSRF	GRF_RUF	GRF_SVM
Leukemia1	5.327	330	1,00 ± 0,000	0,999 ± 0,006	0,998 ± 0,011	0,998 ± 0,011
Brain-tumor1	5.920	372	0,866 ± 0,012	0,834 ± 0,020	0,882 ± 0,019	0,928 ± 0,049
Prostate-tumor	10.509	468	0,999 ± 0,004	0,999 ± 0,004	0,998 ± 0,009	0,974 ± 0,025
Leukemia2	11.225	432	0,875 ± 0,024	0,868 ± 0,036	0,882 ± 0,020	0,928 ± 0,049
Lung-cancer	12.600	558	0,907 ± 0,013	0,924 ± 0,009	0,912 ± 0,017	0,961 ± 0,019
11-tumors	12.533	653	0,910 ± 0,016	0,920 ± 0,013	0,862 ± 0,074	0,964 ± 0,012
GCM	16.063	776	0,707 ± 0,022	0,718 ± 0,022	0,665 ± 0,107	0,720 ± 0,011

Bảng 7. So sánh thời gian thực hiện của phương pháp Rừng ngẫu nhiên và phương pháp SVM trên các tập dữ liệu trước và sau khi sử dụng GRF trích chọn đặc trưng

Tệp dữ liệu	RF			WSRF			RUF			SVM		
	Ban đầu	Sau chọn đặc trưng	GRF_RF	Ban đầu	Sau chọn đặc trưng	GRF_ WSRF	Ban đầu	Sau chọn đặc trưng	GRF_ RUF	Ban đầu	Sau chọn đặc trưng	GRF_ SVM
Leukemia1	2,96 (s)	0,58 (s)	6,16 (s)	16,68 (s)	1,95 (s)	7,537 (s)	17,51 (s)	4,05 (s)	9,638 (s)	27,72 (mins)	6,94 (mins)	421,85 (s)
Brain-tumor1	3,83 (s)	0,69 (s)	8,09 (s)	28,73 (s)	3,93 (s)	11,33 (s)	24,94 (s)	9,15 (s)	16,553 (s)	33,05 (mins)	3,56 (mins)	222,32 (s)
Prostate-tumor	8,73 (s)	1,31 (s)	19,30 (s)	51,23 (s)	6,10 (s)	24,08 (s)	38,27 (s)	8,71 (s)	26,70 (s)	1,33 (hours)	5,30 (minus)	336,27 (s)
Leukemia2	6,75 (s)	0,44 (s)	13,16 (s)	32,90 (s)	2,09 (s)	14,81 (s)	34,66 (s)	5,12 (s)	17,84 (s)	1,43 (hours)	7,90 (secs)	486,82 (s)
Lung-cancer	22,32 (s)	2,55 (s)	53,47 (s)	2,76 (mins)	16,41 (s)	67,327 (s)	1,38 (mins)	13,06 (s)	63,98 (s)	2,3 (hours)	6,40 (mins)	431,38 (s)
11-tumors	27,71 (s)	7,08 (s)	68,10 (s)	3,31 (mins)	44,52 (s)	105,612 (s)	1,69 (mins)	27,97 (s)	89,06 (s)	3,08 (hours)	21,39 (mins)	22,40 (mins)
GCM	42,78 (s)	12,88 (s)	106,39 (s)	5,48 (mins)	1,57 (mins)	3,13 (s)	2,87 (mins)	58,29 (s)	151,0 (s)	6,51 (hours)	39,13 (mins)	40,58 (mins)

Chúng tôi sử dụng phương pháp paired t-test, mức ý nghĩa 0,05 để so sánh độ chính xác của các kết quả phân lớp trước và sau khi sử dụng GRF trích chọn đặc trưng của từng thuật toán. Kết quả cho thấy giá trị trung bình độ chính xác của cả 4 phương pháp phân lớp: RF, WSRF, RUF, SVM trước và sau khi sử dụng GRF là có sự khác biệt. Các thuật toán phân lớp sau khi dùng GRF trích chọn đặc trưng hầu như đều cao hơn khi dùng toàn bộ thuộc tính ban đầu. Điều đó chứng tỏ sự kết hợp GRF với các

phương pháp phân lớp RF, WSRF, RUF, SVM thực sự hiệu quả.

Chúng tôi cũng sử dụng phương pháp paired t-test, mức ý nghĩa 0,05 so sánh kết quả của GRF (Bảng 4) với kết quả của GRF_RF, GRF_WSRF, GRF_RUF, GRF_SVM và thấy rằng các phương pháp GRF_RF, GRF_WSRF, GRF_RUF, GRF_SVM cho độ chính xác tốt hơn GRF.

Để làm rõ hơn đặc điểm của các thuật toán và hiệu quả của sự kết hợp phương pháp dùng GRF trích chọn đặc trưng, chúng tôi đánh giá thêm thời gian thực hiện. Thời gian thực hiện của các phương pháp được thể hiện trong bảng 7.

Trong bảng 7, các cột **"Ban đầu"** và **"Sau chọn đặc trưng"** lần lượt chứa thời gian thực hiện của các thuật toán RF, WSRF, RUF, SVM khi chạy trên các tập dữ liệu gốc và tập dữ liệu sau chọn đặc trưng. Các cột **GRF_RF, GRF_WSRF, GRF_RUF, GRF_SVM** lần lượt chứa thời gian thực hiện của các thuật toán RF, WSRF, RUF, SVM khi sử dụng phương pháp GRF_T (được nêu trong mục 2.5).

Kết quả bảng 6 và 7 cho thấy: Việc kết hợp phương pháp trích chọn đặc trưng của GRF với các mô hình phân lớp WSRF, RUF, SVM không những làm *tăng độ chính xác* của kết quả dự đoán mà còn *giảm thời gian thực hiện* của các thuật toán. Tuy nhiên, khi kết hợp GRF với RF nguyên bản thì thuật toán chạy chậm hơn nhưng độ chính xác cao hơn. Vì vậy, đối với dữ liệu gene chiều rất cao, ít mẫu, nhiều thuộc tính rác thì việc việc kết hợp phương pháp trích chọn đặc trưng của GRF với các phương pháp phân lớp (RF hoặc SVM) là sự lựa chọn tốt. Đặc biệt, kết hợp GRF_SVM cho độ chính xác phân lớp cao hơn hẳn so với các phương pháp khác mà thời gian thực hiện tương đối nhanh (chuyển từ giờ sang phút).

4. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày các phương pháp phân lớp dữ liệu chiều cao gồm: Rừng ngẫu nhiên nguyên bản (RF), Rừng ngẫu nhiên điều hòa (RRF), Rừng ngẫu nhiên có điều hướng (GRF), Rừng ngẫu nhiên không gian con thuộc tính có trọng số(wsRF), Rừng ngẫu nhiên đều (RUF) và SVM. Tiếp đến, chúng tôi đánh giá hiệu quả thực hiện của các phương pháp này trên hai phương diện: độ chính xác phân lớp và thời gian thực hiện. Sau cùng, đề xuất kết hợp phương pháp trích chọn đặc trưng của GRF với các mô hình phân lớp của RF và SVM. Kết quả thực nghiệm cho thấy sự kết hợp này mang lại hiệu quả rõ rệt, đặc biệt tốt khi kết hợp GRF với SVM cho bài toán phân lớp dữ liệu chiều rất cao cỡ hàng nghìn thuộc tính, nhiều thuộc tính rác, dung lượng mẫu ít như dữ liệu gene.

LỜI CẢM ƠN

Chúng tôi xin bày tỏ lời cảm ơn chân thành đến TS. Nguyễn Thanh Tùng (Khoa Công nghệ thông tin - Trường đại học Thủy Lợi) đã chia sẻ nguồn dữ liệu thử nghiệm và góp ý cho chúng tôi những ý kiến có giá trị.

TÀI LIỆU THAM KHẢO

Baoxun Xu, Joshua Zhexue Huang, Graham Williams, Qiang Wang and Yunming Ye (2012). Classifying very high-dimensional data with random forests built from small subspaces. International Journal of Data Warehousing and Mining (IJDWM), 8(2): 44-63.

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

Ciss, S. (2015). Variable Importance in Random Uniform Forests. https://hal.archives-ouvertes.fr/hal-01104340/file/RandomUniform Forests.pdf.

Deng, H. (2013). Guided random forest in the rrf package. arXiv preprint arXiv:1306.0237.

Deng, H., & Runger, G. (2012). Feature selection via regularized trees. International Joint Conference on Neural Networks (IJCNN), pp. 1-8.

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

Đỗ Thanh Nghị, P. N. (2013). So sánh các mô hình dự báo lượng mưa cho thành phố Cần Thơ. Tạp chí Khoa học, Trường đại học Cần Thơ, tr. 80-90.

Manuel, F.-D., Eva, C., & Senén, B. (2014). Do we need hundreds of classifiers to solve. The Journal of Machine Learning Research, 15(1): 3133-3181.

Mardis, E. R. (2011). A decade's prespective on DNA sequencing technology. Nature, 470(7333): 198-203.

Rea, A. (1995). Data Mining - An Introduction. Nor of The Queen's University of Belfast.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. USA: Springer-Verlag.