J. Sci. & Devel. 2015, Vol. 13, No. 2: 301-307 Tạp chí Khoa học và Phát triển 2015, tập 13, số 2: 301-307
www.vnua.edu.vn

PHƯƠNG PHÁP LẤY MẪU THUỘC TÍNH MỚI TRONG RỪNG NGẪU NHIÊN CHO PHÂN TÍCH DỮ LIỆU SNP

Nguyễn Văn Hoàng*, Phan Thị Thu Hồng, Nguyễn Thanh Tùng, Nguyễn Thị Thủy

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

Email: nvhoang@vnua.edu.vn*

Ngày gửi bài: 22.10.2014 Ngày chấp nhận: 20.12.2014

TÓM TẮT

Gần đây, các nghiên cứu liên kết mức toàn hệ gen (GWAS) đã đạt được thành công trong việc xác định một số biến thể di truyền có ảnh hưởng tương đối lớn tới một số bệnh phức tạp. Hầu hết các GWAS sử dụng các tiếp cận đơn SNP (đa hình đơn nucleotide) chỉ tập trung vào việc đánh giá sự liên hệ giữa từng SNP riêng biệt với bệnh. Tuy nhiên, trên thực tế, các bệnh phức tạp được cho là liên quan tới những nguyên nhân phức tạp bao gồm những tương tác rắc rối giữa nhiều SNPs. Do đó, cần có những cách tiếp cận khác để xác định sự ảnh hưởng của các SNP hoặc những tương tác phức tạp của các SNP tới bệnh. Phương pháp rừng ngẫu nhiên (Random Forest, RF) gần đây đã được ứng dụng thành công trong GWAS cho việc xác định một số nhân tố di truyền có ảnh hưởng lớn tới một số bệnh phức tạp. Mặc dù RF xử lý tốt trên khía cạnh chính xác dự đoán trên một số tập dữ liệu có kích cỡ trung bình, nhưng mô hình RF truyền thống có nhiều hạn chế trong việc xác định các SNPs có ý nghĩa và xây dựng các mô hình dự đoán chính xác. Trong bài báo này, chúng tôi đề xuất một phương pháp lấy mẫu hai bước để lựa chọn các đặc trưng có ý nghĩa trong việc huấn luyện mô hình rừng ngẫu nhiên. Phương pháp này cho phép chọn ra một tập nhỏ các đặc trưng có liên hệ chặt chẽ với biến đích (bệnh), do đó làm giảm số chiều và có thể xử lý tốt trên các tập dữ liệu có số chiều cao. Chúng tôi cũng tiến hành các thực nghiệm trên hai tập dữ liệu chuẩn SNP ở mức toàn bộ hệ gen để làm sáng tỏ hiệu quả của phương pháp đề xuất.

Từ khóa: Genome-wide Association Study, học máy, khai phá dữ liệu, rừng ngẫu nhiên

A New Feature Sampling Method in Learning Random Forest for SNP Data Analysis

ABSTRACT

Recently, Genome-wide association studies (GWAS) have been successful in the identification of genetic variants that have effects in some complex diseases. Most GWA studies used single SNP (single-nucleotide polymorphism) approaches that mainly focused on assessing the association between each individual SNP and the disease. However, in fact, complex diseases are thought to involve complex etiologies including complicated interactions between many SNPs. Thus, different approaches are necessary to identify SNPs that influence disease risk jointly or in complex interactions. Random Forest (RF) method recently has been successfully used in GWAS for identifying genetic factors that have effects in some complex diseases. In spite of performing well in terms of prediction accuracy in some data sets with moderate size, RF still suffers from working in GWAS for selecting informative SNPs and building accurate prediction models. In this paper, we propose a new two-stage sampling method in learning random forests. The proposed method allows to select a sub-set of informative SNPs which are most relevant to disease. Therefore, it reduces the dimensionality and can perform well with high-dimensional data sets. We conducted experiments on two genome-wide SNP data sets to demonstrate the effectiveness of the proposed method.

Keywords: Genome-wide Association Study, machine learning, data mining, random forest

1. ĐẶT VẤN ĐỀ

Công nghệ sinh học đã đạt được những bước tiến vượt bậc trong công nghệ giải mã trình tự gen. Giờ đây, toàn bộ hệ gen có thể được giải mã trình tự dễ dàng và nhanh chóng với chi phí thấp (Mardis, 2011). Hệ gen được giải mã trình tự nhanh chóng đã tạo điều kiện cho những nghiên cứu liên kết mức toàn bộ hệ gen trở nên khả thi. Thực tế là những nghiên cứu liên kết mức toàn bộ hệ gen (Genome-wide association studies - GWAS) đã giúp xác định được nhiều biến dị gen là nguyên nhân dẫn tới một số bệnh phức tạp (Wellcome Trust, 2007). Nhiều biến dị gen có liên hệ với các bệnh như bệnh tim mạch (Mohlke et al., 2008), bệnh về miễn dịch (Lettre et al., 2008), bệnh tiểu đường (Sladek et al., 2007) và nhiều bệnh ung thư khác (Easton et al., 2007; 2008) đã được xác định thông qua các nghiên cứu liên kết mức toàn bộ hệ gen. Hầu hết các GWAS đã được tiến hành sử dụng tiếp cận đơn SNP. Tiếp cận đơn SNP sử dụng chỉ xem xét ảnh hưởng của từng SNP riêng lẻ đến bệnh quan tâm. Tuy nhiên, các bệnh phức tạp được cho rằng do sự tác động kết hợp của nhiều SNP (Moore, 2005). Do đó, tiếp cận đơn SNP không xác định được nguyên nhân di truyền của những bệnh phức tạp là kết quả của sự tương tác giữa nhiều SNP. Chính vì vậy, những phương pháp nghiên cứu cho phép phát hiện ảnh hưởng cộng tác của nhiều SNP đến các bệnh là thực sự cần thiết.

Tuy nhiên, xét trên quy mô toàn bộ hệ gen số lượng SNP là vô cùng lớn. Dữ liệu SNP là dữ liệu về hàng trăm ngàn SNP được lấy mẫu từ vài nghìn, thậm chí chỉ vài trăm cá thể. Do đó dữ liệu SNP có số lượng thuộc tính lớn hơn rất nhiều so với số lượng mẫu. Như vậy, dữ liệu SNP là dữ liệu có số chiều cao và các mô hình thống kê truyền thống không còn thích hợp để phân tích. Ngoài ra, các nhà nghiên cứu sinh học tin rằng chỉ có một lượng nhỏ SNP liên quan tới một loại bệnh cụ thể nên dữ liệu SNP là dữ liệu có độ nhiễu cao. Vì vậy, việc xác định những nhóm SNP có ảnh hưởng lớn tới bệnh là một bài toán khó.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Trong mục này chúng tôi phân tích các hướng tiếp cận đã có cho bài toán phân tích dữ liệu SNP. Hướng tiếp cận đơn giản nhất là kiểm tra tất cả các tổ hợp SNP có thể. Tuy nhiên do số lượng tổ hợp là rất lớn nên đòi hỏi giá thành tính toán lớn. Tiếp cận kiểm tra tất cả các tổ hợp gồm 2 SNP đã được thực hiện và cho thấy là rất tốn thời gian, cụ thể cần tới 33 giờ để phân tích dữ liệu 1.000 trường hợp bệnh và 1.000 trường hợp đối chứng trên cluster với 10 cpu (Marchini et al., 2005). Mở rộng ra, việc kiểm tra tất cả các tổ hợp SNP sẽ trở nên không khả thi về mặt tính toán (Cordell, 2009). Một tiếp cận khác là xây dựng một tập con những SNP có khả năng liên quan tới bệnh thông qua những kiểm thử đơn biến trên mỗi SNP sau đó kiểm tra tất cả các tổ hợp SNP trên tập con SNP vừa được xây dựng. Tiếp cận này giúp giảm chi phí tính toán tuy nhiên có thể sẽ loại bỏ những SNP mà nếu đứng độc lập nó ít liên quan tới bệnh nhưng có thể ảnh hưởng lớn tới bệnh trong sự hiện diện của những SNP khác.

Random Forest (RF) 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 (Brieman, 2001). 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. Do đó, việc kết hợp kết quả của một số lượng lớn những cây quyết định độc lập có độ lệch thấp, phương sai cao sẽ giúp RF đạt được cả độ lệch thấp và phương sai thấp. 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. RF cũng đã được sử dụng trong nhiều nghiên cứu phân tích dữ liệu SNP (Bureau et al., 2005; Goldstein et al., 2010; Goldstein et al., 2011; Winham et al., 2012).

Tuy nhiên, tiếp cận cài đặt RF ban đầu của Breiman chỉ hiệu quả cho phân tích dữ liệu có số chiều thấp. Bureau và cộng sự đã cho thấy rằng RF cho kết quả tốt với dữ liệu SNP đối chứng (case-control) với cỡ chỉ 42 SNPs (Bureau et al., 2005). RF cũng có thể áp dụng trên các tập dữ liệu giả lập với số lượng SNP không quá 1.000 (Lunetta et al., 2004). Do đó tiếp cận cài đặt ban đầu của RF không thể áp dụng trên dữ liệu hàng trăm ngàn SNP. Vì vậy, RF hiếm khi được áp dụng trong phân tích dữ liệu SNP trên toàn hệ gen.

Để có thể áp dụng RF lên dữ liệu SNP trên toàn hệ gen, cần có những cải tiến thích hợp. Tiếp cận cải tiến đầu tiên là tham số *mtry*. *mtry* là cỡ của không gian con thuộc tính được lấy ngẫu nhiên từ không gian thuộc tính ban đầu để xây dựng các cây quyết định. *mtry* thường được lấy giá trị mặc định là $\log_2 M+1$ với M là số thuộc tính trong dữ liệu ban đầu. Tuy nhiên giá trị $\log_2 M+1$ chỉ thích hợp với dữ liệu có số chiều nhỏ và hoàn toàn không thích hợp cho dữ liệu có số chiều lớn, đặc biệt là dữ liệu có độ nhiễu cao như dữ liệu SNP. Trong trường hợp dữ liệu SNP, nếu *mtry* quá nhỏ thì số lượng SNP được sử dụng để tạo dựng cây quyết định sẽ ít, hơn nữa do có rất nhiều SNP không liên quan tới bệnh nên có thể sẽ dẫn tới việc chọn ra một tập con SNP mà phần lớn là các SNP không liên quan tới bệnh, điều này sẽ dẫn tới việc tạo ra những cây quyết định có chất lượng thấp, từ đó ảnh hưởng tới chất lượng dự đoán của RF. Do đó, với dữ liệu có số chiều cao và nhiễu như dữ liệu SNP thì *mtry* cần phải chọn đủ lớn để đảm bảo sự chính xác của dự đoán (Wu et al., 2012). Tuy nhiên, nếu chọn *mtry* lớn thì chi phí tính toán kèm theo sẽ lớn. Hơn nữa việc tìm kiếm giá trị tốt cho tham số *mtry* cũng không khả thi về mặt tính toán.

Một tiếp cận khác để cải tiến RF là thay đổi phương pháp sinh các không gian con thuộc tính cho xây dựng các cây quyết định. Trong cài đặt của Brieman, không gian con thuộc tính được sinh ra bằng cách lấy ngẫu nhiên có thay thế từ không gian các thuộc tính ban đầu. Việc lấy ngẫu nhiên này đã dẫn tới việc có thể sinh ra các không gian con SNP chứa đựng hầu hết các SNP không có liên quan tới bệnh và từ đó tạo ra cây quyết định có chất lượng dự đoán thấp.

3. PHƯƠNG PHÁP ĐỀ XUẤT

Như đã phân tích trong mục 2, tiếp cận cài đặt ban đầu của Breiman không phù hợp cho phân tích dữ liệu SNP có số chiều lớn vì việc lấy mẫu 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 được sinh ra. Để khắc phục nhược điểm này chúng tôi đề xuất một phương pháp lấy mẫu mới được tiến hành theo hai bước.

Ở bước đầu tiên chúng tôi cố gắng loại bỏ những thuộc tính (SNP) không có liên quan tới bệnh (biến phụ thuộc, biến đích), chúng được gọi là những thuộc tính nhiễu. Để thực hiện điều này, trước tiên chúng tôi bổ sung vào tập dữ liệu huấn luyện những thuộc tính thực sự nhiễu bằng cách sinh ngẫu nhiên. Những thuộc tính thực sự nhiễu này không có giá trị trong việc dự đoán biến đích. Sau đó RF được xây dựng từ tập dữ liệu huấn luyện đã bổ sung các thuộc tính thực sự nhiễu để ước lượng mức độ quan trọng của mỗi thuộc tính tới việc dự đoán biến đích. Ta thu thập giá trị mức độ quan trọng lớn nhất của các thuộc tính thực sự nhiễu qua mỗi lần ước lượng mức độ quan trọng của các thuộc tính để hình thành một mẫu so sánh. Cuối cùng thực hiện kiểm thử Wilcoxon cho mỗi thuộc tính để kiểm tra liệu trung bình hệ số quan trọng của thuộc tính có lớn hơn trung bình của mẫu so sánh (tức hệ số quan trọng lớn nhất của các thuộc tính thực sự nhiễu) hay không. Tất cả những thuộc tính mà kiểm thử Wilcoxon có *p*-value lớn hơn một ngưỡng cho trước è (giá trị mặc định là 0,05) được coi là những thuộc tính nhiễu, không có ý nghĩa trong việc dự đoán

thuộc tính phụ thuộc và được loại bỏ khỏi tập dữ liệu huấn luyện.

Ở bước thứ hai, tập các thuộc tính còn lại ký hiệu là $\tilde{X}$ sẽ được phân tách thành hai tập: tập các thuộc tính có ảnh hưởng mạnh tới thuộc tính phụ thuộc, ký hiệu là X_s và tập các thuộc tính có ảnh hưởng yếu tới thuộc tính phụ thuộc X_w. Để tách $\tilde{X}$ thành hai tập X_s và X_w, chúng tôi tính thực hiện kiểm thử $\div^2$ cho mỗi thuộc tính. X_s là tập tất cả những thuộc tính (SNP) sở hữu p-value nhỏ hơn hoặc bằng 0,05 thông qua kiểm thử $\div^2$ và $X_w = \tilde{X} \backslash X_s$.

Cuối cùng để sinh ra tập con thuộc tính cho xây dựng cây quyết định, các thuộc tính sẽ được chọn ngẫu nhiên và độc lập với nhau từ hai tập X_s và X_w. Số lượng thuộc tính được chọn từ mỗi tập phụ thuộc vào cỡ của không gian con thuộc tính và cỡ của hai tập X_s và X_w. Nếu cần lấy $mtry$ thuộc tính để xây dựng không gian con thuộc tính thì $mtry_s = \lceil mtry.(\|X_s\|/\|\tilde{X}\|) \rceil$ thuộc tính được lấy từ tập X_s và $gtry_w = \lfloor mtry.(\|X_w\|/\|\tilde{X}\|) \rfloor$ thuộc tính được lấy từ tập X_w, trong đó $\|A\|$ chỉ số lượng phần tử của tập hợp A. Bằng cách lựa chọn không gian con thuộc tính như vậy sẽ đảm bảo không gian con thuộc tính luôn chứa đựng những thuộc tính có ảnh hưởng lớn tới thuộc tính phụ thuộc đồng thời duy trì được việc lựa chọn ngẫu nhiên các thuộc tính.

4. KẾT QUẢ VÀ THẢO LUẬN

4.1. Các độ đo được ước lượng trong thực nghiệm

Trong phần thực nghiệm, chúng tôi đã áp dụng phương pháp đề xuất (từ đây gọi là nRF), tiếp cận cài đặt RF ban đầu của Breiman (Breiman, 2001) (từ đây gọi là RF) và wsRF (Xu et al., 2012) trên hai bộ dữ liệu đối chứng để làm sáng tỏ hiệu quả của phương pháp được đề xuất. Trong quá trình tiến hành thực nghiệm, chúng tôi sử dụng phương pháp của Breiman (Breiman, 2001) để tính toán độ đo trung bình (s), độ đo tương quan trung bình ($\bar{\text{ñ}}$) và c/s2=$\bar{\text{ñ}}/s^2$ để đo lường hiệu năng của RF. Tương quan trung bình $\bar{\text{ñ}}$ phản ánh mức độ độc lập của các cây quyết định trong rừng. Độ đo trung bình s phản ánh độ chính xác hay chất lượng của các cây quyết định trong rừng. Để có mô hình RF tốt, các cây quyết định phải có độ chính xác cao và sự tương quan giữa các cây thấp, điều này được phản ánh qua tỉ số $\bar{\text{ñ}}/s^2$, do đó c/s2 phản ánh độ chính xác tổng quát của mô hình RF.

Ngoài các độ đo trên, hai độ đo nữa cũng được sử dụng làm sáng tỏ sự chính xác và hiệu năng của mô hình RF là *Area under the curve* (AUC) và độ chính xác kiểm thử được tính như 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.

4.2. Dữ liệu thực nghiệm

Chúng tôi tiến hành thực nghiệm trên hai bộ dữ liệu SNP trên toàn bộ hệ gen với những tính chất được mô tả trong bảng 1, trong đó cột "Abbr." chỉ ra tên viết tắt của các tập dữ liệu được sử dụng trong thực nghiệm.

Tập dữ liệu đầu tiên là dữ liệu bệnh chứng cho bệnh Alzheimer (ALZ) 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 (bệnh) (Webster et al., 2009). 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 (Fung et al., 2006).

Bảng 1. Mô tả hai tập dữ liệu SNP

Tập dữ liệu	Abbr.	#SNPs	# Cases hoặc Controls	# Classes
Alzheimer	ALZ	380.157	364	2
Parkinson	PAR	408.803	451	2

4.3. Kết quả thực nghiệm

Bảng 2 cho thấy trung bình độ chính xác kiểm thử và AUC của 3 phương pháp nRF, RF và wsRF. Kết quả trong bảng 2 cho thấy nRF và wsRF luôn cho kết quả tốt với các giá trị *mtry* khác nhau. wsRF và RF cho kết quả tốt hơn khi *mtry* lớn hơn. nRF với $mtry = \sqrt{M_p}$ cho kết quả tốt hơn RF và wsRF trên cả 2 bộ dữ liệu, ở đây $M_p = \|X_s\| + \|X_w\|$ là số lượng SNP còn lại sau khi đã loại bỏ những SNP nhiễu. Như vậy, nRF thực sự tốt cho phân tích dữ liệu SNP có số chiều cao vì không đòi hỏi tham số *mtry* phải được thiết lập quá cao như hai phương pháp còn lại nhưng vẫn đạt được kết quả tốt. Như đã phân tích ở trên, việc thiết lập *mtry* quá lớn sẽ dẫn tới thời gian tính toán rất lâu, nRF thực sự đã rút ngắn đáng kể thời gian xử lý, do đó có thể áp dụng cho dữ liệu có số chiều cao.

Bảng 2. So sánh sự khác biệt giữa các phương pháp với các giá trị *mtry* khác nhau

Tập dữ liệu	Phương pháp	Mtry	Values	Acc	AUC
ALZ	nRF	$\sqrt{M_p}$	45	0,907	0,975
	wsRF	$\log_2 M$	19	0,561	0,711
	wsRF	$\sqrt{M}$	616	0,692	0,757
	RF	$\log_2 M$	19	0,530	0,623
	RF	$\sqrt{M}$	616	0,632	0,729
PAR	nRF	$\sqrt{M_p}$	22	0,895	0,959
	wsRF	$\log_2 M$	19	0,754	0,850
	wsRF	$\sqrt{M}$	638	0,837	0,917
	RF	$\log_2 M$	19	0,564	0,722
	RF	$\sqrt{M}$	638	0,799	0,848

Bảng 3. So sánh sự khác biệt trong mức độ chính xác dự đoán khi số lượng cây quyết định thay đổi

Tập dữ liệu	Phương pháp	K				
		20	50	80	100	200
ALZ	nRF	0,711	0,775	0,791	0,846	0,893
	wsRF	0,528	0,588	0,527	0,602	0,593
	RF	0,517	0,491	0,505	0,555	0,533
PAR	nRF	0,852	0,871	0,858	0,861	0,871
	wsRF	0,647	0,680	0,708	0,710	0,745
	RF	0,579	0,557	0,553	0,597	0,580

Bảng 4. So sánh sự khác biệt c/s2 khi số lượng cây quyết định thay đổi

Tập dữ liệu	Phương pháp	K				
		20	50	80	100	200
ALZ	nRF	0,711	0,775	0,791	0,846	0,893
	wsRF	0,528	0,588	0,527	0,602	0,593
	RF	0,517	0,491	0,505	0,555	0,533
PAR	nRF	0,852	0,871	0,858	0,861	0,871
	wsRF	0,647	0,680	0,708	0,710	0,745
	RF	0,579	0,557	0,553	0,597	0,580

Bảng 3 cho thấy mức độ chính xác trong dự đoán và bảng 4 cho thấy giới hạn lỗi tổng quát của các mô hình RF được sinh ra bởi cả ba phương pháp. Cả ba phương pháp đều được chạy với tham số *mtry* được nhận giá trị cố định là $\lfloor \log_2(M) + 1 \rfloor$ trong khi số lượng cây quyết định trong rừng được điều chỉnh trong mỗi lần chạy. Cụ thể chúng tôi đã thử nghiệm cả ba phương pháp với số lượng cây quyết định thay đổi từ 20 tới 200 cây. Kết quả đã cho thấy rằng nRF vượt trội RF và wsRF về sự chính xác trong dự đoán và mức độ lỗi tổng quát (c/s2) thấp hơn so với hai phương pháp còn lại.

5. KẾT LUẬN

Chúng tôi đã đề xuất một phương pháp lấy mẫu tập con thuộc tính mới dựa trên phân tích điểm yếu của phương pháp lấy mẫu trong mô hình RF truyền thống được đề xuất bởi Breiman. Phương pháp đề xuất đã đảm bảo được chất lượng của các cây quyết định khi RF được xây dựng trên tập dữ liệu có số chiều cao và độ nhiễu lớn trong khi vẫn duy trì được tính ngẫu nhiên trong RF. Kết quả thực nghiệm cho thấy phương pháp đề xuất cho một kết quả tốt hơn tiếp cận cài đặt ban đầu của Breiman cũng như một số giải thuật cải tiến của RF gần đây. Với phương pháp lấy mẫu đề xuất, RF có thể áp dụng để phân tích các dữ liệu có số chiều cao trong đó dữ liệu SNP chỉ là một trường hợp cụ thể.

TÀI LIỆU THAM KHẢO

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.

Cordell, H.J. (2009). Detecting gene–gene interactions that underlie human diseases. Nature Reviews Genetics, 10(6): 392-404.

Easton, D. et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447(7148): 1087-1093.

Easton, D. F., Eeles, R. A. (2008). Genome-wide association studies in cancer. Hum Mol Genet, 17: R109-R115.

Fung, H.C., Scholz, S., Matarin, M., Sim ´ on-S ´ anchez, J., Hernandez, D., Britton, A., Gibbs, J.R., Langefeld, C., Stiegert, M.L., Schymick, J., et al. (2006). Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. The Lancet Neurology, 5(11): 911-916.

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).Rndom Forests for Genetic Association Studies. Statistical Applications in Genetics and Molecular Biology, 10(1): 32

Lettre G., Rioux J. D. (2008). Autoimmune diseases: insights from genome-wide association studies. Hum Mol Genet, 17: R116-R121.

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

Marchini, J., Donnelly, P., Cardon, L.R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics, 37(4): 413-417.

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

Mohlke K. L., Boehnke M., Abecasis G. R. (2008). Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. Hum Mol Genet, 17: R102-R108.

Moore, J. H. (2005). A global view of epistasis. Nature Genetic, 37(1): 13-14.

Schwarz, D.F., K"onig, I.R., Ziegler, A. (2010). On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics, 26(14): 1752.

Sladek, R. et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature, 445(7130): 881-885.

Webster, J.A., Gibbs, J.R., Clarke, J., Ray, M., Zhang, W., Holmans, P., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., et al. (2009).Genetic control of human brain transcript expression in Alzheimer disease. The American Journal of Human Genetics, 84(4): 445-458.

Wellcome Trust (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447(7145): 661-678

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.

Wu, Q., Ye, Y., Liu, Y., Ng, M.K. (2012). SPN selection and classification of genome-wide snp data using stratified sampling random forests. NanoBioscience, IEEE Transactions on, 11(3): 216-227.

Xu, B., Huang, J.Z., Williams, G., Wang, Q., Ye, Y. (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.