DỰ BÁO MỰC NƯỚC TRÊN SÔNG MÊ-KÔNG DÙNG MÔ HÌNH HỒI QUY PHI TUYẾN RANDOM FORESTS

Nguyễn Thanh Tùng[1], Nguyễn Khắc Tiên Phước[2]
[1]*Trường Đại học Thủy lợi. Email: tungnt@wru.vn*

1. GIỚI THIỆU

Random Forests (RF) [2], một trong những phương pháp phi tuyến nổi trội được dùng phổ biến trong lĩnh vực học máy và khai thác dữ liệu trên thế giới những năm gần đây, RF hoạt động tốt trên các bài toán phân loại và hồi quy [1-5] và ít bị ảnh hưởng bởi nhiễu (noise). Thông thường, mô hình hồi quy được viết ở dạng tổng quát như sau:

$$Y = f(x) + \varepsilon \quad (1)$$

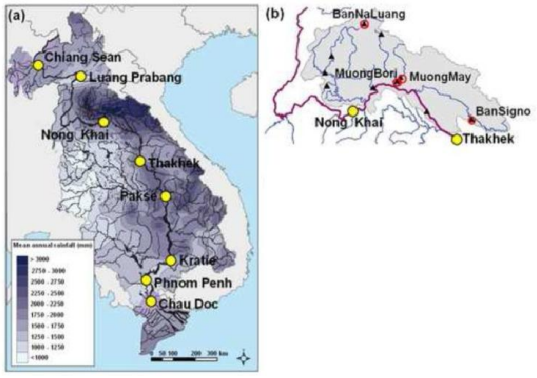
trong đó ε là lỗi của mô hình, E(ε) = 0, Var(ε) = σ_ε^2.

Mục tiêu của bài toán hồi quy là tìm mô hình mà giá trị ước lượng của nó được dự đoán bởi hàm f(·) có lỗi càng nhỏ càng tốt. Trong biểu thức (1), X ∈ R^M và Y ∈ R^1 là các *biến ngẫu nhiên* với xác suất $\wp$, cụ thể, $\wp(X = x, Y = y)$ là xác suất mà các biến ngẫu nhiên X, Y nhận các giá trị x và y. Ở đây, M là số chiều của tập dữ liệu đầu vào. Mô hình hồi quy RF được dùng như 1 hàm phi tuyến (nonparametric) f : $R^M \rightarrow R^1$ ước lượng giá trị y ∈ Y tương ứng với dữ liệu đầu vào x ∈ R^M.

Trong bài báo này, chúng tôi cải tiến mô hình hồi quy phi tuyến RF và ứng dụng mô hình này dự báo mực nước trên sông Mê-kông tại vị trí trạm đo Thakhek, nơi thể hiện sự đóng góp quan trọng của lưu vực thuộc phía trung-nam nước Cộng hòa Dân chủ Nhân dân Lào cho dòng chính sông Mê-kông. Vị trí nghiên cứu được minh họa trong Hình 1. Trong bài toán này, mối quan hệ giữa các biến đầu vào-đầu ra trong mô hình hồi quy phi tuyến RF như sau:

$H_{Thakhek}(t+5) = f\{H_{Thakhek}(t), H_{Thakhek}(t-1), H_{Thakhek}(t-2), H_{NongKhai}(t), H_{NongKhai}(t-1), H_{NongKhai}(t-2), P_{w3}(t), P_{w5}(t), P_{w7}(t)\}$. (2)

Ở biểu thức (2), biến đầu ra HThakhek(t+5) là mực nước dự báo cho 5 ngày tới tại trạm Thakhek. HThakhek(t), HThakhek(t-1) và HThakhek(t-2) lần lượt là mực nước đo được trong ngày hiện tại và hai ngày trước. HNongKhai(t), HNongKhai(t-1) và HNongKhai(t-2) lần lượt là mực nước đo được trong ngày hiện tại và hai ngày trước tại trạm Nông Khai. Đây là trạm đo ở phía trên của Thakhek, cách



Hình 1. *Vị trí lưu vực nghiên cứu.*
(a) Hạ lưu sông Mê kông, (b) tiểu lưu vực 1: 16.906 (km²).

Thakhek 300 km về phía thượng nguồn sông Mê-kông. $P_{w3}(t)$, $P_{w5}(t)$ và $P_{w7}(t)$ lần lượt là lượng mưa trung bình trên lưu vực gia nhập khu giữa Nông Khai và Thakhek cho các thời đoạn 3, 5 và 7 ngày gần đây nhất. Các số liệu mực nước và mưa được chuẩn hóa tuyến tính theo công thức:

$$x_{Norm} = FMIN + \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})} \times (FMAX - FMIN). \quad (3)$$

Trong đó: FMIN và FMAX là các giá trị nhỏ nhất và lớn nhất trong miền chuẩn hóa được chọn tương ứng là 0.1 và 0.9; x_{Norm} là giá trị chuẩn hóa, x_{min} và x_{max} là các giá trị nhỏ nhất và lớn nhất trong chuỗi số liệu đo, x_i là giá trị thực đo trước khi chuẩn hóa.

Ngoài những ưu điểm về xử lý số liệu chiều cao (hàng triệu biến đầu vào), mô hình RF các ưu điểm khác như : *i) Không cần những giả định về phân bố như những mô hình hồi quy tuyến tính; ii) Dễ dàng cài đặt và xử lý dữ liệu trên hệ thống nguồn mở R ; iii) Dễ dàng điều chỉnh các tham số của mô hình để đạt kết quả tối ưu*. Để mô hình RF có hiệu suất tốt hơn, chúng tôi cải tiến cách xây dựng cây hồi quy và cách tổng hợp kết quả dự đoán trong RF. Breiman [1-3] đề xuất dùng phương pháp bình phương tối thiểu (least squares) làm điều kiện tách mỗi nút của cây nhị phân hồi quy trong RF, tiếp đó là lấy giá trị trung bình của các cây trong RF khi tổng hợp kết quả cuối cùng [2, 3]. Phương pháp này có sai số cao

khi: *i) dữ liệu có nhiều phần tử ngoại lai (Outliers), ii) phân bố dữ liệu đầu vào bị kéo lệch, không theo Gaussian*. Những nguyên nhân này làm hiệu suất của mô hình phi tuyến RF bị giảm đáng kể. Trong bài báo này, chúng tôi thay thế phương pháp tách nút của cây nhị phân hồi quy bằng độ lệch tuyệt đối (least absolute deviation) và tổng hợp kết quả dự báo dựa trên giá trị trung vị (median). Các kết quả thực nghiệm cho thấy mô hình cải tiến vượt trội mô hình hồi quy phi tuyến của Breiman và các biến thể RF được đề xuất gần đây dựa trên sai số trung bình và chỉ số Nash-Sutcliffekhi ứng dụng vào bài toán dự báo mực nước trên sông Mê-kông.

2. PHƯƠNG PHÁP

Trong mục này, chúng tôi trình bày thuật toán hồi quy phi tham số RF cải tiến dựa trên mô hình RF của Breiman [2]. Từ dữ liệu đầu vào $L=(X_i, Y_i)_{1\le i\le N}$, trong đó N là cỡ mẫu đo đạc được, $X=\{H_{Thakhek}(t), H_{Thakhek}(t-1), H_{Thakhek}(t-2), H_{NongKhai}(t), H_{NongKhai}(t-1), H_{NongKhai}(t-2), P_{w3}(t), P_{w5}(t), P_{w7}(t)\}$ và $Y=H_{Thakhek}(t+5)$. RF lấy K *mẫu ngẫu nhiên có hoàn lại* (bagged samples) [1] từ tập L, ta được $L_1, L_2, .., L_k$ ($1\le k\le K$), mỗi cây nhị phân hồi quy T_k được xây dựng trên từng tập mẫu tương ứng L_k. RF được tổng hợp từ K cây nhị phân hồi quy và kết quả dự đoán được tính toán dựa trên việc lấy kết quả trung bình của các cây nhị phân trong RF [2-5].

Trong mô hình RF cải tiến, chúng tôi sử dụng độ lệch tuyệt đối làm điều kiện để tách nút trong cây nhị phân, ý tưởng này đã trình bày trong [4] (trang 258) có chứng minh lý thuyết, tuy nhiên Breiman không đưa giải pháp và cách cài đặt. Giải thuật cải tiến này chúng tôi cài đặt bằng ngôn ngữ C++, gọi các hàm tương ứng từ môi trường R. Thuật toán gồm một số bước chính như sau:

- Bước 1 : Lấy K mẫu $L_1, L_2, .., L_k$ có hoàn lại từ L.
- Bước 2 : Với mỗi mẫu L_k, xây dựng cây nhị phân hồi quy với các tiêu chí sau :

+ Tại nút *t*, chọn ngẫu nhiên *mtry* biến, tính độ lệch tuyệt đối tại mỗi biến tương ứng và tách nút *t* thành 2 nút con [3-6]. Gọi *v(t)* là giá trị phân vị của phân bố biến đầu ra Y_t tại nút *t*, ta cực tiểu hóa biểu thức :

$$\sum_{X_I\in t_2} |Y_1 - v(t_2)| \sum_{X_I\in t_R} |Y_1 - v(t_R)| \qquad (4)$$

trong đó $v(t_L)$, $v(t_R)$ là giá trị phân vị của phân bố biến đầu ra tại nút con bên trái và bên phải tại nút *t*.

+ Mỗi cây được xây dựng không cắt nhánh cho đến khi thỏa mãn điều kiện dừng n_{min}.

- Bước 3 : Cho tập dữ liệu đầu vào mới $X=x_{new}$, gọi $\hat{H}^k$ là giá trị dự báo mực nước của cây hồi quy T^k. Giá trị dự báo mực nước của RF là phân vị của K cây hồi quy, ta tính như sau: $\hat{H}^L = v(\hat{H}^k)$.

3. KẾT QUẢ THỬ NGHIỆM

3.1. Tham số mô hình

Số liệu trong mùa lũ của các năm 1994-1997 được dùng làm tập số liệu xây dựng mô hình hồi quy, tập dữ liệu này gồm 9 biến đầu vào và 1 biến đầu ra, tổng số bản ghi là 1071. Số liệu trong mùa lũ của các năm 1998-2000 gồm 459 bản ghi được dùng để đánh giá hiệu quả mô hình hồi quy phi tuyến RF. Số lượng cây nhị phân hồi quy trong RF được đặt mặc định là 1000, số lượng biến chọn ngẫu nhiên để tách nút trong cây hồi quy mtry=3, số lượng phần tử tại mỗi nút lá của cây nmin=20.

3.2. Phương pháp đánh giá

Chúng tôi dùng sai số trung bình (mean absolute error-*MAE*[1]) để đánh giá hiệu quả mô hình hồi quy. Giá trị *MAE* càng thấp thì khả năng dự báo của mô hình càng cao. Theo khuyến nghị của Ủy ban sông Mê-Kông[2] [6], giá trị *MAE* chấp nhận được cho dự báo trước 5 ngày tại vị trí Thakhek là nhỏ hơn 0.75 mét. Bên cạnh đó chỉ số Nash-Sutcliffe để đánh giá hiệu quả mô hình hồi quy cũng được sử dụng.

$$CE = 1 - \frac{\sum_{i=1}^{N}\left(H_i - \hat{H}_i\right)^2}{\sum_{i=1}^{N}\left(H_i - \overline{H}\right)^2} \qquad (5)$$

Trong đó $\overline{H}$ là giá trị trung bình của mực nước thực đo.Khi CE = 1, có nghĩa là giá trị dự báo hoàn toàn khớp với giá trị đo. Nếu CE ≤ 0 sẽ tương ứng với việc hiệu quả của mô hình rất kém, vì nó không bằng sử dụng chính giá trị đo trung bình làm giá trị dự báo.

Trong phần thử nghiệm, chúng tôi so sánh giải thuật đề xuất với các mô hình random forests được sử dụng phổ biến hiện nay. Các gói phần mềm R phiên bản mới nhất *randomForests* (RF), *quantregForest* quantile regression forests (QRF) và *cForest* conditional random forests (cRF) được dùng để thực hiện các thử nghiệm. Các gói phần mềm này được cung cấp trên CRAN[3]. Các mô hình RF đều sử dụng chung tham số cài đặt và chung dữ liệu.

[1] http://en.wikipedia.org/wiki/Mean_absolute_error
[2] http://ffw.mrcmekong.org/accuracy.htm
[3] http://cran.r-project.org

3.2. Kết quả

Trong mục này, chúng tôi trình bày tóm tắt kết quả đạt được từ việc thực hiện thử nghiệm các gói phần mềm R trên bộ số liệu đo đạc trên sông Mê-kông tại vị trí nghiên cứu như đã trình bày ở mục 1 và 3.1.

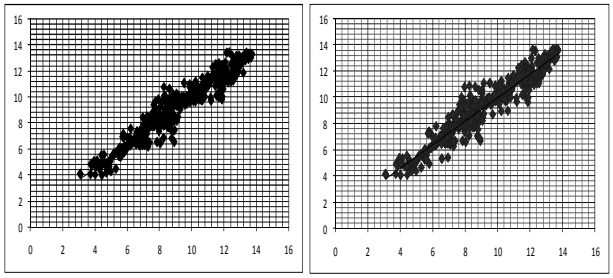
Bảng 1 liệt kê kết quả các mô hình hồi quy phi tuyến ứng dụng dự báo mực nước trên sông Mê-kông. Bốn mô hình hồi quy phi tuyến được kiểm thử với số liệu mùa lũ từ năm 1998-2000, sau đó tính *MAE* và *CE*. Chúng ta có thể thấy sai số trung bình *MAE* (mét) của mô hình đề xuất (RF.median) cho kết quả thấp nhất, tương tự với giá trị CE đạt kết quả cao hơn so với các mô hình thông dụng RF, QRF và cRF. Kết quả này chứng minh mô hình đề xuất dự báo chính xác hơn (*MAE* nhỏ hơn) và chỉ số Nash- Sutcliffe cao hơn (gần hơn với giá trị thực đo) so với các mô hình nổi tiếng dựa trên RF.

Kết quả dự đoán được trình bày trực quan hơn ở các Hình 2, các hình này hiển thị quan hệ giữa giá trị thực đo và giá trị dự đoán. Trục hoành trình bày các số liệu thực đo $H_{Thakhek}(t+5)$, trục tung là giá trị dự đoán khi sử dụng các mô hình hồi quy random forests. Tập hợp các điểm trên hình vẽ được chia làm 3 phần, các giá trị dự đoán bằng với giá trị thực đo sẽ nằm trên đường thẳng (*solid line*), những điểm nằm bên trái hoặc bên phải đường thẳng là những giá trị dự đoán thấp hơn hoặc cao hơn giá trị thực đo. Ta có thể thấy ở Hình 2, giá trị dự đoán của mô hình hồi quy đề xuất RF.median có sai số ít hơn so với RF mật độ phân bố của các điểm sát nhau hơn và nằm dọc theo đường thẳng quan hệ.

Bảng 1: Kết quả dự báo mực nước trên sông Mê-kông

Mô hình/ Đánh giá	RF.median	RF	QRF	cRF
MAE (m)	**0,5391**	0,5539	0,5629	0,5653
CE	**0,9287**	0,9156	0,9182	0,9171

* Giá trị in đậm là kết quả tốt nhất.



(a) RF.median *(b) RF*

Hình 2: *Quan hệ tuyến tính giữa giá trị thực đo và giá trị dự đoán của mô hình RF mới và Breiman đề xuất.*

4. KẾT LUẬN

Ứng dụng mô hình hồi quy trong công tác dự báo luôn được các nhà khoa học quan tâm nghiên cứu và áp dụng. Chúng tôi đã trình bày mô hình cải tiến hồi quy phi tuyến random forests, một trong những mô hình học máy được dùng phổ biến hiện nay trên thế giới. Trong giải thuật đề xuất, chúng tôi đã cải tiến cách tách nút khi xây dựng cây nhị phân, khi tổng hợp kết quả chúng tôi sử dụng giá trị phân vị của các kết quả ở các cây hồi quy thay vì cách tính lấy giá trị trung bình trong random forests truyền thống. Kết quả thử nghiệm dự đoán mực nước sông Mê-kông trên bộ dữ liệu mùa lũ các năm từ 1994-2000 cho thấy, giải thuật đề xuất cho kết quả chính xác hơn, chỉ số *CE* cao hơn và sai số *MAE* ít hơn so với các mô hình nổi tiếng như state-of-the-art RF, QRF và cRF.

Lời cảm ơn. Bài báo được hỗ trợ bởi đề tài "Nghiên cứu xây dựng hệ thống phần mềm tác nghiệp quản lý đê điều và các công trình trên đê, phục vụ phòng, tránh và giảm nhẹ thiên tai cho vùng Hà Nội", mã số 01C-07/01-2012-2.

TÀI LIỆU THAM KHẢO

[1]. Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140.

[2]. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[3]. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. BocaRaton: CRC Press.

[4]. N.T.Tung, Joshua Z.Huang, Thuy, N.T. "Two Level QRF for Bias Correction in Range Prediction", In Machine Learning, Springer, 2014.

[5]. N.T.Tung, Joshua Z.Huang, Thuy, N.T. "Unbiased Feature Selection in Learning Random Forests for High Dimensional Data", The Scientific World Journal, vol. 2014, Article ID 471371.

[6]. Nguyen, Phuoc Khac-Tien, Lloyd Hock-Chye Chua, and L.H. Son. "Flood forecasting in large rivers with data-driven models." Natural hazards 71.1 (2014): 767-784.