TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Phương pháp dự báo nước biển dâng do bão dựa trên lập trình di truyền

Nguyễn Thị Hiền[1], Trương Tiến Phúc[2], Ngô Văn Mạnh[3], Nguyễn Thị Quyên[4], Hoàng Hải Vân[5]

[1] Học viện Kỹ thuật quân sự; nguyenthihienqn@gmail.com.
[2] Văn phòng Zalo Hà nội; truong.t.phuc@gmail.com.
[3] Trung tâm TTDL KTTV; manh.ngovan@gmail.com.
[4] Đại học Lâm nghiệp Việt Nam; quyen14121982@gmail.com.
[5] Đại học quản lý và công nghệ Hải Phòng; hoangvan041078@gmail.com.
***** Tác giả liên hệ: nguyenthihienqn@gmail.com; Tel.: +84–912092486.

Ban Biên tập nhận bài: 5/5/2020; Ngày phản biện xong: 18/7/2020; Ngày đăng bài: 25/7/2020

Tóm tắt: Nước dâng bão là hiện tượng dâng lên của mực nước biển cao hơn mực thủy triều vốn có bởi do tác động của bão vì thế việc dự báo chính xác mực nước dâng là nhiệm vụ quan trọng để tránh thiệt hại về tài sản và con người do nước dâng gây ra. Lập trình di truyền (Genetic Programming – GP) là một kỹ thuật học máy có thể giúp ta tìm được mô hình ở dạng công thức toán học. Tuy nhiên trước đây GP hầu như chưa được áp dụng triệt để cho bài toán dự báo nước biển dâng do bão cho nên trong bài báo này nhóm tác giả đề xuất phương pháp sử dụng GP để phát hiện các mô hình dự báo nước biển dâng do bão. Kết quả thực nghiệm trên dữ liệu nước biển dâng do bão tại trạm Hòn Dáu của Việt Nam cho thấy phương pháp này có thể đưa ra các mô hình dự báo nước dâng do bão chính xác hơn một số phương pháp học máy phổ biến thường sử dụng. Hơn nữa GP đưa ra mô hình dự báo dễ hiểu hơn các mô hình mà được xây dựng bằng các phương pháp khác (hộp đen) như là mạng nơ–ron. Ngoài ra mô hình dự báo do GP đưa ra sẽ giúp ta phát hiện các đặc trưng ảnh hưởng trực tiếp khi phát triển các mô hình dự báo nước biển dâng do bão.

Từ khóa: Lập trình di truyền; dự báo nước biển dâng do bão, Hòn Dáu.

1. Đặt vấn đề

Dự báo nước dâng do bão là rất quan trọng đối với quá trình ra quyết định trong quản lý ven biển để giảm rủi ro lũ lụt ở vùng trũng và đối với bài toán dự báo nước dâng do bão này người ta cần các mô hình nhanh và chính xác. Ngoài bão, sóng thần thì gió mùa mạnh cũng là nguyên nhân chính gây nước dâng vùng ven bờ. Nước dâng do bão là một thiên tai nghiêm trọng và đặc biệt nguy hiểm khi chúng xảy ra khi thủy triều lên khi đó sự kết hợp tác động của nước dâng và thủy triều.

Với hơn 600 triệu người sống ở các vùng ven biển trũng, nước dâng ven biển có thể có tác động nghiêm trọng tới xã hội. Cơn bão Katina (2005) tại Mỹ gây ra mực nước dâng cao tới 6 m, làm hơn 1000 người chết, gây thiệt hại tài sản khoảng 81,2 tỷ đô la. Cơn bão Hải Yến (11/2013) tại Philippin khiến tổng số người thiệt mạng lên đến 7000 người (chủ yếu là do nước dâng do bão). Không những thế trong tương lai các cơn bão có ảnh hưởng lớn sẽ tiếp tục xảy ra vì vậy việc dự báo nước dâng do bão chính xác sẽ làm giảm đáng kể thiệt hại về người và tài sản [1–3]. Trước đây, cách tiếp cận thông thường để dự báo nước dâng do bão là sử dụng mô hình dự báo số trị, tuy nhiên các mô hình này đòi hỏi mất nhiều năng lực

tính toán. Một cách tiếp cận khác là sử dụng các thuật toán học máy như mạng nơ–ron [1] để dự đoán các mối quan hệ giữa mực nước dâng và các đặc trưng tương ứng như là mực nước biển, gió, khí áp trên mặt biển và các đặc tính của cơn bão nhiệt đới. Người ta đã xây dựng mô hình dự báo nước dâng do bão sử dụng một số mô hình trí tuệ nhân tạo [4] để dự báo mực nước dâng cao nhất sử dụng các tham số của cơn bão nhiệt đới: áp suất tâm bão, bán kính gió lớn nhất,.. Kết quả cho thấy việc dùng mạng nơ–ron nhân tạo cho kết quả tốt hơn so với máy hỗ trợ véc–tơ. Các kết quả đã chỉ ra rằng phương pháp sử dụng trí tuệ nhân tạo và khung lưới tự do hoàn toàn đáp ứng được độ chính xác với tốc độ dự báo nhanh. So sánh với các mô hình thông thường các mô hình dựa trên mạng nơ–ron có thời gian tính toán nhanh trong khoảng 10 phút sẽ cho ra kết quả dự báo sau khi huấn luyện xong mô hình. Tuy nhiên mô hình dựa trên mạng nơ–ron này là dạng hộp đen vì vậy rất khó để giải thích chúng hơn nữa các mô hình loại này thường không đạt được khả năng ước lượng tại các cao điểm điều này rất quan trọng khi dự báo nước dâng do bão.

Lập trình di truyền (GP) là một sơ đồ tiến hóa để tìm ra lời giải bài toán. Khả năng của GP là tự học định nghĩa của một hàm từ các mẫu điều này giúp GP là một sự lựa chọn phù hợp cho việc giải bài toán hồi quy ký hiệu [5]. Chính vì vậy GP được sử dụng rộng rãi để xây dựng các mô hình hồi quy cho các ứng dụng thực tế. Chẳng hạn như mô hình dự đoán giá cổ phiếu sử dụng GP để tạo ra một chiến lược đầu tư sinh lãi [6]. Trong [7] GP được sử dụng để xây dựng mô hình dự báo sóng thời gian thực. Các kết quả của các nghiên cứu trên đã chỉ ra rằng GP là một công cụ đầy hứa hẹn cho các ứng dụng dự báo cho dữ liệu các vùng biển. Trong nghiên cứu [8] GP được sử dụng để dự báo độ xói mòn ống xảy ra ở lòng sông và kết quả cho thấy việc sử dụng GP có kết quả khả thi hơn sơ với sử dụng phương trình hồi quy và hệ thống nơ–ron nhân tạo trong việc mô hình hóa dự đoán độ sâu xói mòn xung quanh các "ống". Tuy nhiên GP đã và chưa được áp dụng trong dự báo nước dâng do bão vì vậy trong bài báo này tác giả đề xuất nghiên cứu áp dụng GP để xây dựng mô hình "hộp trắng" (một dạng mô hình dễ hiểu) cho việc dự báo nước biển dâng.

Do đó, đóng góp chính của bài viết này là chúng tôi đề xuất sử dụng GP với một số thay đổi nhỏ áp dụng cho bài toán dự báo nước biển dâng do bão và so sánh hiệu suất dự báo của nó so với các phương pháp học máy khác thường được áp dụng cho những bài toán dự báo tương tự.

Phần còn lại của bài báo này được tổ chức như sau. Phần 2 sẽ trình bày về GP bao gồm giới thiệu chung, và một số điểm riêng dùng cho bài toán dự báo nước biển dâng do bão. Phần 3 sẽ đưa ra các tham số cụ thể của GP khi chạy thực nghiệm, dữ liệu để thí nghiệm, cùng với các phương pháp học máy khác để so sánh với GP. Phần 4 trình bày kết quả của thí nghiệm đánh giá, phân tích, so sánh kết quả của các phương pháp. Cuối cùng, phần 5 kết luận lại những phát hiện và đề xuất các nghiên cứu trong tương lai.

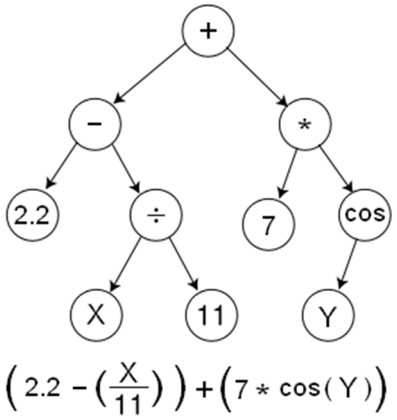
2. Phương pháp nghiên cứu

2.1 Lập trình di truyền

Lập trình di truyền (*Genetic Programming – GP*) ra đời vào năm 1992 [5] với tham vọng nhằm đưa ra một quần thể các chương trình mà chúng có thể tiến hóa một cách tự động trên những dữ liệu huấn luyện. Với nghĩa này, GP được xem như là một phần của học máy. Dựa trên lý thuyết tiến hóa của Darwinian, GP đưa ra các chương trình mã hóa dưới dạng các chuỗi di truyền thông qua quá trình tiến hóa và chọn lọc tự nhiên để tìm được chuỗi di truyền (chương trình) tốt đáp ứng được yêu cầu bài toán.

2.1.1 Biểu diễn chương trình

Chương trình trong GP được biểu diễn dưới dạng cây, trong đó mỗi nút được gán nhãn là một ký hiệu thuộc tập hàm (F) hay tập kết (T).



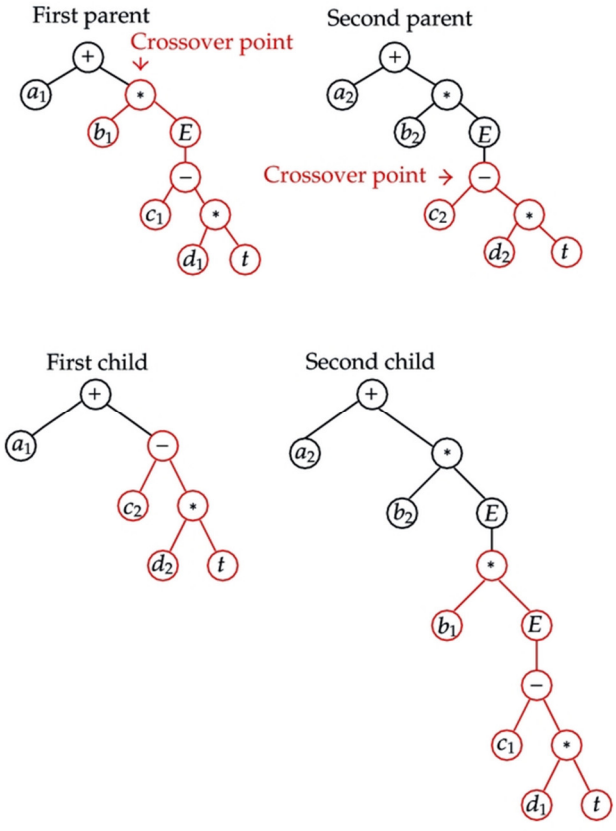
Hình 1. Biểu diễn chương trình GP.

2.1.2 Toán tử di truyền

a) Toán tử lai ghép (*crossover*)

Thể hiện quá trình trao đổi nhiễm sắc thể giữa hai cây bố mẹ. Toán tử gồm các bước sau:

- Chọn một nút ngẫu nhiên trên mỗi cây bố mẹ;
- Hoán đổi hai cây con có gốc tại hai nút vừa chọn và tráo đổi chúng cho nhau.



Hình 1. Toán tử lai ghép.

b) Toán tử đột biến (*Mutation*)

Là quá trình đột biến của một bộ nhiễm sắc thể được tạo ra. Gồm các bước sau:

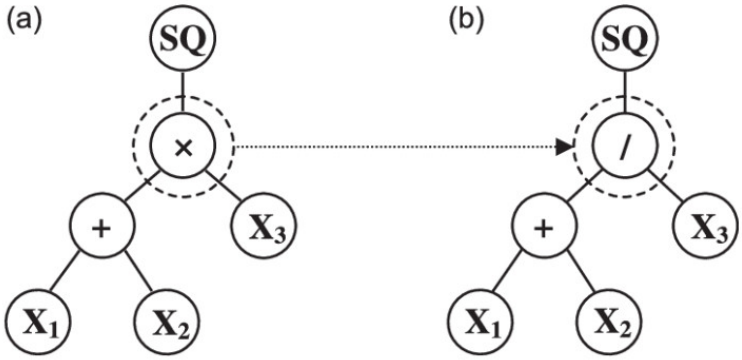
- Chọn ngẫu nhiên một nốt bất kì trên cây cha (mẹ);
- Xóa cây con thuộc nốt được chọn;
- Sinh ngẫu nhiên một cây con mới vào vị trí vừa xóa.

c) Tái sinh (*reproduction*)

Nếu một cá thể được tái sinh chúng sẽ được sao chép y nguyên vào quần thể, hay nói cách khác là sẽ có hai cá thể giống nhau trong quần thể.

2.1.3 Đánh giá độ phù hợp (*fitness*)

Mỗi một chương trình được gán một giá trị được gọi là độ phù hợp, giá trị này sẽ có ảnh hưởng quan trọng đến việc cá thể có được lựa chọn để thực hiện các toán tử di truyền hay không.



Hình 2. Toán tử đột biến (a) cây trước khi thực hiện toán tử, (b) cây sau khi thực hiện.

Như vậy các bước để chạy một thuật toán GP:

1) Khởi tạo ngẫu nhiên một quần thể (thế hệ 0) các cá thể được tạo ra từ tập hàm và tập kết.

2) Thực hiện lặp (các thế hệ) theo các bước phụ sau cho đến khi thỏa mãn điều kiện kết thúc (tìm thấy lời giải tối ưu hoặc đạt đến số thế hệ nào đó):

a) Đánh giá độ tốt của các cá thể.

b) Chọn 1 hoặc 2 cá thể từ quần thể với xác suất phụ thuộc vào độ tốt của chúng để tham gia vào các toán tử di truyền c.

c) Tạo các cá thể mới cho quần thể bằng việc áp dụng các phép toán di truyền sau với một xác suất đã định.

- Tái sinh
- Lai ghép
- Đột biến

Sau khi kết thúc quá trình tiến hóa, cá thể tốt nhất của toàn bộ quá trình chạy được coi như là kết quả của quá trình chạy.

Bên cạnh các phương pháp truyền thống: cây quyết định, tập luật quyết định, hàm thống kê và mạng nơron các nghiên cứu đã cho thấy rằng GP cũng là một phương pháp giải bài toán dự báo với độ chính xác cao bằng cách tiến hóa ra cây biểu thức. Một trong những lý do cho phép ta tin tưởng điều này là quá trình tìm kiếm của GP có kết quả tốt đối với những bài toán có không gian tìm kiếm lớn.

2.2. Lập trình di truyền cho bài toán dự báo nước biển dâng do bão

Việc sử dụng lập trình di truyền (GP) để dự báo nước biển dâng sau bão gần đây cũng đã được một số nghiên cứu áp dụng. Các tác giả trong bài báo [9] đã đề xuất sử dụng GP để dự đoán nước dâng do bão và ngập lụt do các cơn bão nhiệt đới. Các thí nghiệm được thực

hiện trên các bộ dữ liệu từ bờ biển Odisha đến tiếp giáp với Vịnh Bengal. Các kết quả đã chỉ ra rằng cả mạng nơ–ron nhân tạo (ANN) và GP đều dự báo rất tốt so với dữ liệu thực tế. Tuy nhiên, GP đã không được nghiên cứu sâu hơn nữa về các mô hình để dự báo sau khi thực hiện với thời gian dự báo khác nhau. Hơn nữa, tính linh hoạt của GP để tự động chọn các đặc trưng để xây dựng các mô hình có thể hiểu được để dự báo nước dâng do bão cũng chưa được nghiên cứu. Do đó, bài viết này tiếp tục nghiên cứu khả năng của GP để xây dựng các mô hình dự báo mức độ nước dâng sau bão.

Ở trong nước, đã có một số nghiên cứu về dự báo nước biển dâng do bão và gió mùa; tuy nhiên, chưa có nghiên cứu nào về sử dụng công cụ học máy/ trí tuệ nhân tạo để dự báo nước biển dâng.

3. Thí nghiệm

Phần này sẽ trình bày cách thiết kế thí nghiệm và các tham số của GP đã được hiệu chỉnh cho phù hợp với bài toán dự báo nước biển dâng do bão.

3.1 Tham số của GP

Bảng 1 trình bày các tham số cụ thể để chạy GP. Ở đây hàm đánh giá độ tốt của mỗi cá thể chúng tôi sử dụng hàm RMSE (*root mean square error*).

Bảng 1. Các tham số khi cài đặt GP.

Tham số	Giá trị
Tập hàm	+, –, x, /, sin, cos, ln, √
Tập kết	Biến thuộc tính
Kích thước quần thể	1000
Thuật toán khởi tạo	Ramped half–and–half
Độ cao lớn nhất của cây	15
Số thế hệ	200
Xác suất thực hiện lai ghép	0,9
Xác suất thực hiện đột biến	0,1
Phương pháp chọn lựa	Tranh đấu kích thước bằng 3

Thực hiện chạy GP 30 lần độc lập, mỗi lần chạy với giá trị khởi tạo khác nhau, sau mỗi lần chạy ta sẽ nhận được một lời giải tốt nhất. Sau 30 lần chạy ta có 30 lời giải tương ứng, sắp xếp các lời giải đó theo thứ tự tăng dần giá trị độ phù hợp, lựa chọn lời giải trung vị (*median*) của dãy đó dùng làm mô hình cuối cùng.

3.2 Dữ liệu bài toán

Dữ liệu thử nghiệm là dữ liệu nước dâng của 12 cơn bão đo tại trạm Hòn Dáu trước thời điểm nước dâng cao nhất 24h trong Bảng 2.

Bảng 2. Một số cơn bão dùng để thu thập dữ liệu nước biển dâng.

STT	Tên bão	Thời điểm bắt đầu	Thời điểm kết thúc
1.	Bão số 14 (Haiyan)	05/11/2013	11/11/2013
2.	Bão số 1	13/06/2014	17/06/2014

STT	Tên bão	Thời điểm bắt đầu	Thời điểm kết thúc
3.	Bão Rammasun	12/07/2014	21/07/2014
4.	Bão số 1 (Kujira)	19/06/2015	25/06/2015
5.	Bão số 4 (Mujigae)	01/10/2015	05/10/2015
6.	Bão số 1 (Mirinae)	25/07/2016	28/07/2016
7.	Bão số 2 (NIDA)	28/7/2016	03/08/2016
8.	Bão số 3 (DIANMU)	15/08/2016	19/08/2016
9.	Bão số 7(Sarika)	13/10/2016	19/10/2016
10.	Bão số 8 (HAIMA)	15/10/2016	23/10/2016
11.	Bão số 6 (Hato)	20/08/2017	24/08/2017
12.	Bão Talim	10/09/2017	18/9/2017

Dựa trên nghiên cứu [1], thu thập dữ liệu các tham số đầu vào bao gồm:

- Tham số khí tượng: tốc độ gió (WS) (m/s), hướng gió (WD) (độ), khí áp trên mặt biển (hPa) và độ giảm khí áp trong bão trên mặt biển (DSLP) (1013 hPa).
- Tham số hải văn: mực nước bề mặt biển (SS), thủy triều (SSL).
- Tham số theo cơn bão: kinh độ (LG), vĩ độ (LT) (độ), áp suất tâm bão (CAP) (hPa) và tốc độ gió cao nhất gần tâm bão (HWS) (m/s).

Giá trị đầu ra là giá trị nước biển dâng do bão. Các giá trị dữ liệu thu thập sẽ được chuẩn hóa theo công thức sau:

$\eta_i^t = \tilde{\eta}_i^t$ với giá trị mực nước dâng

$v_{SSL} = \tilde{v}_{SSL}$ với giá trị mực nước thủy triều

$v_{SLP} = \tilde{v}_{SLP}/1013\ hPa$ cho khí áp trên mặt biển

$v_{DSLP} = \tilde{v}_{DSLP}/100\ hPa$ cho độ giảm khí áp trong bão trên mặt biển.

$v_{WS} = \tilde{v}_{WS}/100\ m/s$ với tốc độ gió

$v_{WD} = \tilde{v}_{WD}/360\ deg$ với hướng gió

$v_{LG} = \tilde{v}_{LG}/150^oE$ với kinh độ của bão

$v_{LT} = \tilde{v}_{LT}/50^oN$ với vĩ độ của bão

$v_{CAP} = \tilde{v}_{CAP}/1013\ hPa$ với áp suất tâm bão

$v_{HWS} = \tilde{v}_{HWS}$ với tốc độ gió lớn nhất gần tâm bão.

Trong đó dấu (~) bên phải của các phương trình thể hiện giá trị gốc của các tham số.

3.3 Các kỹ thuật học máy khác để so sánh

Để so sánh GP với các kỹ thuật học máy khác khi giải quyết bài toán dự báo nước biển dâng do bão, chúng tôi lựa chọn 5 kỹ thuật học máy đưa ra mô hình dự báo chỉ dựa vào dữ liệu và có khả năng phản ánh tốt được mối quan hệ giữa các biến đầu vào và đầu ra (bài toán dự báo) mà không cần xem xét trực tiếp các quy luật vật lý của cơ chế nước biển dâng do bão. Những mô hình này hoàn toàn dựa trên thông tin có được từ việc thu thập dữ liệu. Đó là các mô hình sau:

3.3.1. Máy vec–tơ hỗ trợ (*Support Vector Machine*)

Máy véc tơ hỗ trợ hồi quy (*Support Vector Regression –SVR*) [10], là một phương pháp thành công để phạt sự phức tạp mô hình bằng cách cộng thêm giá trị này vào hàm lỗi. Để minh họa ta xem xét một mô hình tuyến tính dự báo cho bởi công thức (2):

$$f(x) = w^T x + b \quad (2)$$

Trong đó w là véc–tơ trọng số, b là độ dốc và x là véc–tơ đầu vào. Gọi x_m và y_m lần lượt là véc–tơ đầu vào, giá trị đầu ra thứ m của tập huấn luyện. Công thức tính hàm lỗi như công thức (3):

$$J = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}|y_m - f(x_m)|_\varepsilon \quad (3)$$

Số hạng thứ nhất của hàm lỗi chính là giá trị phạt độ phức tạp của mô hình, còn số hạng thứ hai là giá trị lỗi nhạy cảm với ε. Nếu hàm lỗi nhỏ hơn ε thì sẽ không phạt, đây là tham số được đưa thêm vào để điều chỉnh giảm độ phức tạp của mô hình. Chính vì vậy lời giải sẽ cực tiểu hóa hàm lỗi như công thức (4):

$$f(x) = \sum_{m=1}^{M}(\alpha_m^* - \alpha_m)x_m^T x + b \quad (4)$$

Trong đó α_m^*, α_m là nhân tử Lagrange. Véc–tơ huấn luyện đưa ra các số nhân Lagrange khác không được gọi là các véc–tơ hỗ trợ và đây là một khái niệm chính về lý thuyết SVR. Các véc–tơ không hỗ trợ không đóng góp trực tiếp vào lời giải và số lượng vectơ hỗ trợ là độ đo đo độ phức tạp của mô hình. Mô hình này được mở rộng cho trường hợp phi tuyến tính thông qua khái niệm nhân κ sinh ra công thức (5):

$$f(x) = \sum_{m=1}^{M}(\alpha_m^* - \alpha_m)\kappa(x_m^T x) + b \quad (5)$$

Trong thí nghiệm chúng tôi sẽ sử dụng nhân Gauss.

3.3.2. Cây quyết định (*Decision Tree – DCT*)

DCT [11] là một kiểu mô hình dự báo. Mỗi một nút trong của cây tương ứng với một biến; cạnh nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự báo của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là cây quyết định.

Cây quyết định có thể được học bằng cách chia tập hợp nguồn thành các tập con dựa theo một kiểm tra giá trị thuộc tính. Quá trình này được lặp lại một cách đệ quy cho mỗi tập con dẫn xuất. Quá trình đệ quy hoàn thành khi không thể tiếp tục thực hiện việc chia tách được nữa, hay khi một phân loại đơn có thể áp dụng cho từng phần tử của tập con dẫn xuất. Một bộ phân loại rừng ngẫu nhiên (*random forest*) sử dụng một số cây quyết định để có thể cải thiện tỉ lệ phân loại.

3.3.3. k–láng giềng gần nhất (*k Nearest Neighbor – kNN*)

kNN [12] là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần xếp lớp và tất cả các đối tượng trong tập dữ liệu. Một đối tượng được phân lớp dựa vào k láng giềng của nó, k là số nguyên dương được xác định trước khi thực hiện thuật toán. Người ta thường dùng khoảng cách Euclidean để tính khoảng cách giữa các đối tượng.

3.3.4. Mạng Perceptron nhiều lớp (*Multi–layer Perceptron – MLP*)

MLP [13] là mạng nơ–ron nhân tạo được gọi là perceptron nhiều lớp bởi vì nó là tập hợp của các perceptron chia làm nhiều nhóm, mỗi nhóm tương ứng với một layer. Hoạt động của chúng có thể được mô tả như sau tại tầng đầu vào các nơron nhận tín hiệu vào xử lý (tính tổng trọng số, gửi tới hàm truyền) rồi cho ra kết quả (là kết quả của hàm truyền); kết quả này sẽ được truyền tới các nơron thuộc tầng ẩn thứ nhất; các nơron tại đây tiếp nhận như là tín hiệu đầu vào, xử lý và gửi kết quả đến tầng ẩn thứ 2; quá trình tiếp tục cho đến khi các nơron thuộc tầng ra cho kết quả.

3.3.5 Rừng ngẫu nhiên (*Random Forest – RF*)

RF [14] là một tập các mô hình (*ensemble*). Mô hình RF rất hiệu quả cho các bài toán dự báo vì nó sử dụng cùng lúc rất nhiều mô hình nhỏ hơn bên trong với quy luật khác nhau để đưa ra quyết định cuối cùng. Mỗi mô hình bên trong đó có thể tốt hoặc chưa tốt khác nhau, nhưng khi tổng hợp, ta sẽ có cơ hội dự báo chính xác hơn so với khi sử dụng một mô hình đơn lẻ bất kì nào.

Rừng ngẫu nhiên (*Random Forest – RF*) cho độ chính xác dự báo khá cao khi so sánh với các thuật toán học có giám sát hiện nay bao gồm Boosting, Baging, k–láng giềng gần nhất (*k nearest neighbors*), SVM, ANN, C4.5,..

Các mô hình trên được sử dụng rất phổ biến cho các bài toán học máy và cũng đã cho thấy hiệu năng đáng kể của chúng.

4. Phân tích kết quả

Trong phần này, ta sẽ xem xét các kết quả khi chạy GP so với các thuật toán học máy điển hình. Để so sánh hiệu suất của GP với các phương pháp khác chúng tôi sử dụng hai độ đo như công thức (6, 7):

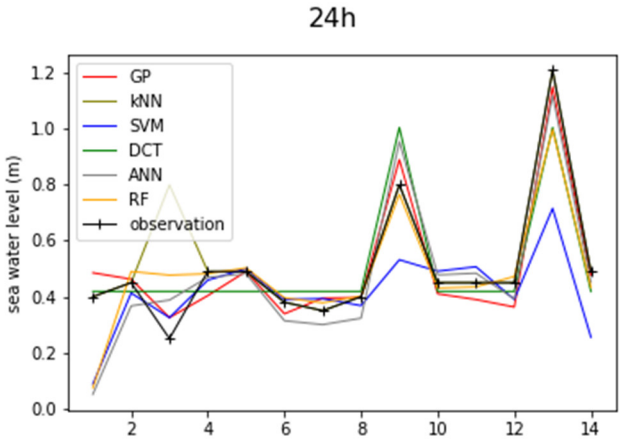
$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{obs,i}-y_{pre,i})}}{(y_{obs,max}-y_{obs,min})} \tag{6}$$

$$CC = \frac{\sum_{i=1}^{n}(y_{obs,i}-\bar{y}_{obs})(y_{pre,i}-\bar{y}_{pre})}{\sqrt{\sum_{i=1}^{n}(y_{obs,i}-\bar{y}_{obs})^2}\sqrt{\sum_{i=1}^{n}(y_{pre,i}-\bar{y}_{pre})^2}} \tag{7}$$

Trong đó NRMSE (*normal root mean squared error*) là RMSE chuẩn hóa tính theo phần trăm, CC (*correlation coefficient*) là hệ số tương quan.

Trong công thức trên n là độ lớn tập huấn luyện, $y_{pre,i}$ là giá trị dự báo của điểm mẫu i còn $y_{obs,i}$ là giá trị đo được ở điểm mẫu i.

Mục đích của GP là quá trình tiến hóa làm sao tìm cây lời giải có giá trị NRMSE nhỏ và CC lớn.

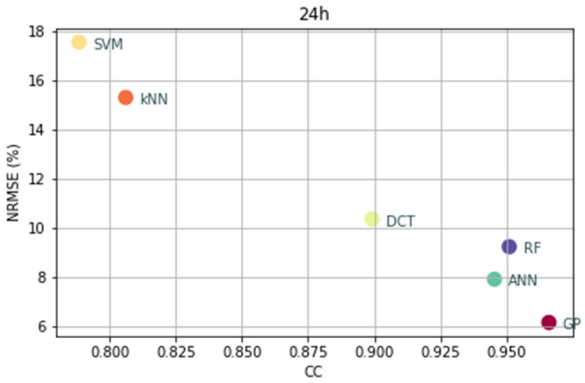


Hình 4. Giá trị dự báo theo thời gian của các mô hình dự báo và dữ liệu thực tế đo đạc được tại trạm Hòn Dáu của 12 cơn bão.

Từ Hình 4 (trong đó trục ngang thể hiện các điểm dữ liệu, trục đứng là giá trị nước dâng do bão – đơn vị đo là m) thể hiện các giá trị dự báo của 6 mô hình và giá trị thực tế ta nhận thấy mô

hình kết quả của GP (màu xanh) bám sát nhất với đường màu đen (giá trị thực tế) đặc biệt tại các điểm cao. Điều đó cho thấy rằng mô hình dự báo do GP đưa ra có khả năng đoán nhận gần đúng nhất các điểm dữ liệu thực tế.

Kết luận trên được khẳng định một lần nữa rõ ràng hơn trong Hình 5, trong đó giá trị NRMSE của 6 phương pháp dự báo nằm trong khoảng từ 6% đến 18%, còn giá trị CC nằm trong khoảng từ 0,75 đến 0,97. Và ta cũng thấy phương pháp GP vừa cho kết quả giá trị NRMSE nhỏ (sai số ít nhất) và CC lớn nhất (gần gũi với giá trị thực nhất kể cả các điểm cao) trong số 6 phương pháp.



Hình 5. Giá trị NRMSE và CC của các mô hình dự báo với dữ liệu tại Hòn Dáu.

Như vậy trên tập dữ liệu thực tế của 12 cơn bão khác nhau, GP cho mô hình dự báo tốt nhất so với các phương pháp còn lại. Kết quả khẳng định hiệu năng của GP vượt trội so với các mô hình dự báo khác.

Mô hình kết quả tiến hóa GP

Dưới đây là một cây lời giải cho bài toán dự báo nước biển dâng do bão là kết quả của quá trình tiến hóa của GP có dạng:

sqrt(sub(mul(X10,0.921279(log(mul(X3,sub(X8,X9))))),mul(X10,mul(log(add(cos(sqrt(div(sub(X5,sub(X2,log(0.299569(X10))))),X9))),add(cos(sqrt(div(sub(X5,sub(cos(mul(X2,add(X7,X10))),mul(mul(X2,X2),X10))),mul(cos(X10),X9)))),X3))),sqrt(mul(X10,cos(log(X4)))))).

Biểu thức tương ứng với cây trên là:

$$\sqrt{\begin{matrix} x_{10} \times 0.921279 \log\left(x_3 \times (x_8 - x_9)\right) \\ -\log\left(cos\sqrt{\frac{(x_5 - x_2 + log 0.299569 x_{10})}{x_9}} + \right. \\ \left. cos\sqrt{\frac{x_5 - \cos(x_2 \times (x_7 + x_{10})) - x_2^2 \times x_{10}}{cos x_{10} \times x_9}} + x_3\right) \times \\ \sqrt{\frac{x_{10}}{cos log x_4}} \times x_{10} \end{matrix}} \quad (8)$$

Với mô hình kết quả như (8) việc dự báo trở nên khá dễ dàng với các biến x_i chính là các giá trị đầu vào (trong đó x_1 là giá trị WS, x_2 là WD, x_3 là SLP, x_4 là DSLP, x_5 là SSL, x_6 là LG, x_7 là LT, x_8 là CAP, x_9 là HWS và x_{10} là SS.Và với mô hình nhận được ta nhận thấy sự phụ thuộc của kết quả vào các tham số đó cũng là một tham khảo để lựa chọn đặc trưng cho phù hợp bài toán. Đây chính là ý nghĩa hộp trắng của GP mà chỉ có mô hình DCT trong số 5 mô hình trên mới có.

5. Kết luận

Bài báo trình bày việc sử dụng GP để dự báo nước biển dâng do bão tại trạm Hòn Dáu Việt Nam, các kết quả cho thấy GP vượt trội hơn về hiệu năng so với các phương pháp dự báo khác (MLP, SVM, kNN, DCT, RF). Chính vì vậy, trong tương lai chúng tôi sẽ tiếp tục cải tiến GP để thu được kết quả dự báo tốt hơn nữa. Ngoài ra chúng tôi cũng sẽ dùng GP để áp dụng cho dữ liệu tại các trạm khác, với các cơn bão khác và với thời gian dự báo trước ngắn hơn (12h, 5h) để có được kết quả dự báo phù hợp với yêu cầu thực tế.

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: N.T.H, T.T.P, N.V.M., N.T.Q., H.H.V.; Lựa chọn phương pháp nghiên cứu: N.T.H, T.T.P; Xử lý số liệu: N.V.M; Phân tích mẫu: T.T.P, N.T.Q., H.H.V.; Lấy mẫu: N.V.M, N.T.H; Viết bản thảo bài báo: N.T.H., T.T.P.; Chỉnh sửa bài báo: N.T.H.

Lời cảm ơn: Nghiên cứu này được hỗ trợ bởi đề tài "Nghiên cứu cơ sở khoa học và giải pháp ứng dụng trí tuệ nhân tạo để nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng khí tượng thủy văn nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam", số hiệu của đề tài BĐKH.34/16–20, thuộc Chương trình Khoa học và Công nghệ ứng phó với biến đổi khí hậu, quản lý tài nguyên và môi trường giai đoạn 2016–2020, mã số BĐKH/16–20.

Lời cam đoan: Tập thể tác giả cam đoan bài báo này là công trình nghiên cứu của tập thể tác giả, chưa được công bố ở đâu, không được sao chép từ những nghiên cứu trước đây; không có sự tranh chấp lợi ích trong nhóm tác giả.

Tài liệu tham khảo

1. Kim, S.; Matsumi, Y.; Pan, S.; Mase, H. A real–time forecast modelusing artificial neural network for after–runner storm surges on the tottoricoast, Japan. *Ocean Eng.* **2016**, *122*, 44–53. https://doi.org/10.1016/j.oceaneng.2016.06.017.
2. Kim, S.W.; Lee, A.; Mun, J. A surrogate modeling for storm surgeprediction using an artificial neural network. *J. Coastal Res.* **2018**, *85*, 866–870. https://doi.org/10.2112/SI85–174.1.
3. Thuy, N.B.; Kim, S.; Chien, D.D.; Dang, V.H.; Cuong, H.D.; Wettre, C.; Hole, L.R. Assessment of storm surge along the coast of central vietnam. *J. Coastal Res.* **2016**, *33*, 518–530.
4. Lee, T.L. Prediction of storm surge and surge deviation using a neural network. *J. Coastal Res.* **2008**, *24*, 76–82.
5. Koza, John, R. Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge, MA, USA: MIT Press, 1992.
6. Kaboudan, M.A. Genetic programming prediction of stock prices. *Comput. Econ.* **2000**, *16*, 207–236.
7. Gaur, D.S.; Deo, M.C. Real–time wave forecasting using genetic programming. *Ocean Eng.* **2008**, *35*, 1166–1172. https://doi.org/10.1016/j.oceaneng.2008.04.007.
8. Azamathulla, H.M.; Ghani, A.A. Genetic Programming to Predict River Pipeline Scour. *J. Pipeline Syst. Eng. Pract.* **2010**, *1*, 127–132. https://doi.org/10.1061/(ASCE)PS.1949–1204.0000060
9. Sahoo, B.; Bhaskaran, P.K. Prediction of storm surge and inundation using climatological datasets for the indian coast using soft computing techniques. *Soft Comput.* **2019**, 23, 12363–12383. https://doi.org/10.1007/s00500–019–03775–0.
10. Smola, A.J.; Schölkopf, B.; 2004. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88
11. Rokach, L.; Maimon, O. Data mining with decision trees: Theory and applications. World Scientific Pub. Co. Inc. 2014, pp. 328.

12. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining. Inference and Prediction. New York: Springer, 2009.
13. Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington DC: Spartan Books, 1961.
14. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

A genetic programming–based storm surge for prediction

Nguyen Thi Hien[1], Truong Tien Phuc[2], Ngo Van Manh[3], Nguyen Thi Quyen[4], Hoang Hai Van[5]

[1] LeQuyDon Technical University; nguyenthihienqn@gmail.com
[2] Zalo Office, Hanoi; truong.t.phuc@gmail.com
[3] Vietnam National Hydrometeorolocical Forecasting Center Hanoi; manh.ngovan@gmail.com
[4] Vietnam National University of Forestry; quyen14121982@gmail.com.
[5] Haiphong Private University, Haiphong; hoangvan041078@gmail.com.

Abstract: Storm surge could be a genuine fiasco coming from the ocean. Therefore, an exact forecast of surges is a vital assignment to dodge property misfortunes and to decrease chance caused by tropical storm surge. Genetic Programming (GP) is an evolution–based model learning technique that can find both the functional form and the numeric coefficients for the model. From our perspective, Genetic Programming has not been enough applied to the problem of storm surge forecasting. In the reserach paper, we propose a new approach to using Genetic Programming to evolve models for storm surge forecasting. Experimental results of storm surge forecasting on HonDau station, Vietnam show that Genetic Programming could be evolved more accurate models of storm surge forecasting than other common machine learning methods tried for the problem in the literature. Moreover, the model evolved by Genetic Programming is more interpretable than the models built by other (black–box) methods such as neural networks. Additionally, Genetic Programming could automatically select relevant features when evolving storm surge forecasting models.

Keywords: Genetic Programming; Storm surge prediction; HonDau.