DỰ BÁO LƯỢNG MƯA TẠI MỘT SỐ TRẠM QUAN TRẮC VIỆT NAM DỰA TRÊN LẬP TRÌNH DI TRUYỀN

Nguyễn Thị Hiền[1], Nguyễn Xuân Hoài[2], Đặng Văn Nam[3], Ngô Văn Mạnh[4]

Tóm tắt: *Dự báo lượng mưa là một trong những bài toán thách thức nhất, vì nó thể hiện các đặc điểm rất độc đáo không tồn tại trong dữ liệu chuỗi thời gian khác. Hơn nữa, lượng mưa là một thành phần chính và rất cần thiết cho việc áp dụng quy hoạch tài nguyên nước. Chính vì vậy, bài viết này tập trung vào việc dự đoán lượng mưa sử dụng dữ liệu từ Cơ quan Khí tượng Việt Nam. Hiện nay trong hầu hết các nghiên cứu dự báo lượng mưa, quá trình dự báo thường bị chi phối bởi các mô hình thống kê, cụ thể là sử dụng chuỗi Markov được mở rộng với dự báo lượng mưa (MCRP). Trong bài báo này, nghiên cứu trình bày một phương pháp mới để giải quyết bài toán dự đoán lượng mưa là lập trình di truyền (Genetic Programming - GP). Đây là lần đầu tiên GP được sử dụng trong bối cảnh dự báo lượng mưa ở một số thành phố tại Việt Nam. Nghiên cứu sẽ so sánh hiệu suất của GP và các thuật toán học máy khác như SVM, MLP, DCT, kNN trên 3 bộ dữ liệu khác nhau của các thành phố tại Việt Nam và báo cáo kết quả. Mục tiêu là để xem liệu GP có khả năng dự báo tốt hơn so với các phương pháp học máy khác hay không? Các kết quả đều chỉ ra rằng nói chung GP vượt trội đáng kể so với các phương pháp học máy khác, đó là cách tiếp cận chủ đạo trong bài viết.*

Từ khóa: *Lập trình di truyền, dự báo lượng mưa.*

Ban Biên tập nhận bài: 12/09/2019 Ngày phản biện xong: 20/10/2019 Ngày đăng bài:25/11/2019

1. Đặt vấn đề

Mưa là một hiện tượng quan trọng trong hệ thống khí hậu, có bản chất hỗn loạn có ảnh hưởng trực tiếp đến quy hoạch tài nguyên nước, nông nghiệp và hệ thống sinh học. Bài toán dự báo lượng mưa đặt ra khá nhiều trở ngại, cả trong nghiên cứu và trong thực tiễn (lượng mưa là tương đối khó để đo được chính xác). Đã có khá nhiều các nghiên cứu được thực hiện để giải quyết bài toán này. Trong bài viết này nghiên cứu sẽ mô tả việc sử dụng lập trình di truyền để áp dụng cho bài toán dự báo lượng mưa tích lũy.

Mục đích bài viết này là khám phá xem GP có vượt trội hơn so với các cách tiếp cận khác thường được áp dụng trong bài toán dự báo lượng mưa hay không. GP được lựa chọn cho bài báo này chứ không phải các kỹ thuật học máy khác, bởi vì GP đưa ra lời giải bài toán ở dạng hộp trắng (giúp ta có thể hiểu được sự phụ thuộc của lời giải vào các thuộc tính đã chọn, trái ngược với mô hình hộp đen), nó cho phép ta hiểu sâu hơn về lời giải. Hơn nữa, chúng ta có thể có hiểu được phần phi tuyến trong mẫu dữ liệu mà không cần bất kỳ giả định nào liên quan đến dữ liệu. Điều này sẽ cho phép chúng ta dễ dàng đưa ra một mô hình dự báo có thể phản ánh quá trình thay đổi lượng mưa. Xa hơn nữa, người dự báo có thể nắm bắt được những sai lệch hàng năm mà hiện tại một số cách tiếp cận truyền thống không thể làm được (sử dụng chuỗi Markov để dự báo).

Do đó, đóng góp chính của bài viết này là

[1]Học viện Kỹ thuật quân sự
[2]Viện AI Việt nam
[3]Đại học Mỏ-Địa Chất
[4]Trung tâm Thông tin và Dữ liệu khí tượng thủy văn
Email: manh.ngovan@gmail.com; nguyenthihienqn@gmail.com

nghiên cứu đề xuất GP với một số thay đổi nhỏ áp dụng cho bài toán dự báo lượng mưa và so sánh hiệu suất dự đoán 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 lượng mưa. 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.

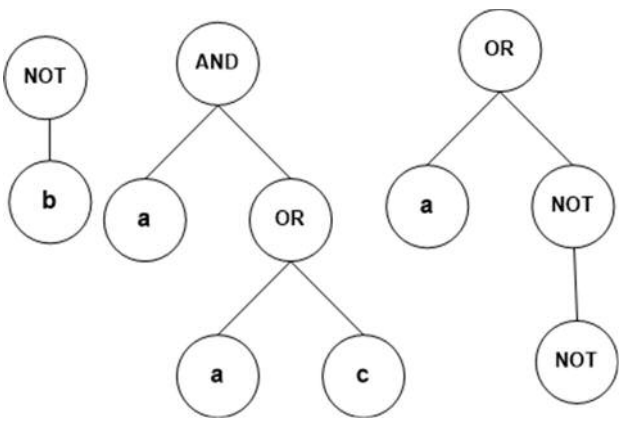
2. Phương pháp nghiên cứu

a. Lập trình di truyền

Lập trình di truyền (Genetic Programming - GP) ra đời vào năm 1992 [3] 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.

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

Toán tử di truyền

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.

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.

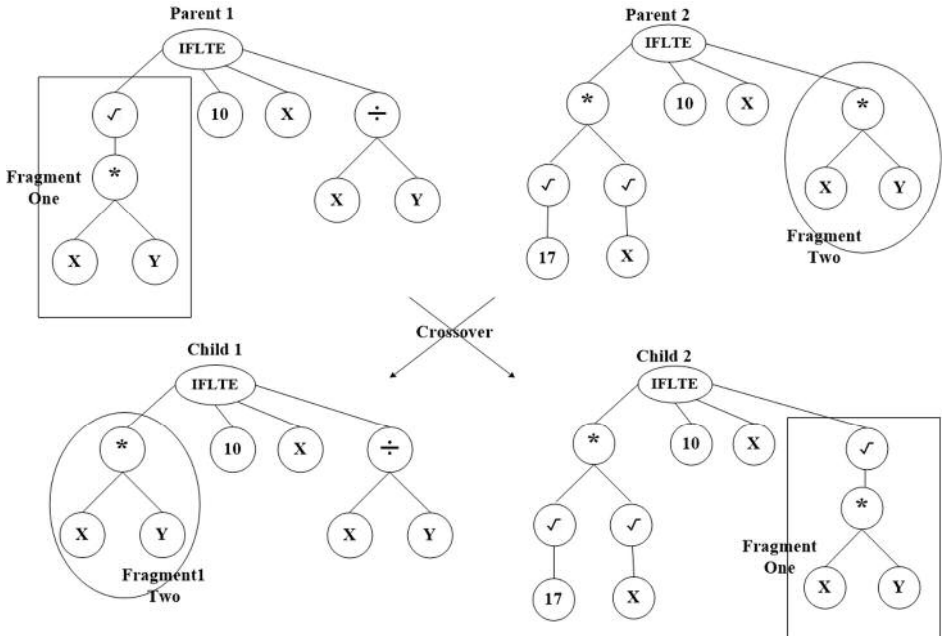
• Sinh ngẫu nhiên một cây con mới vào vị trí vừa xóa.

Tái sinh (reproduction)

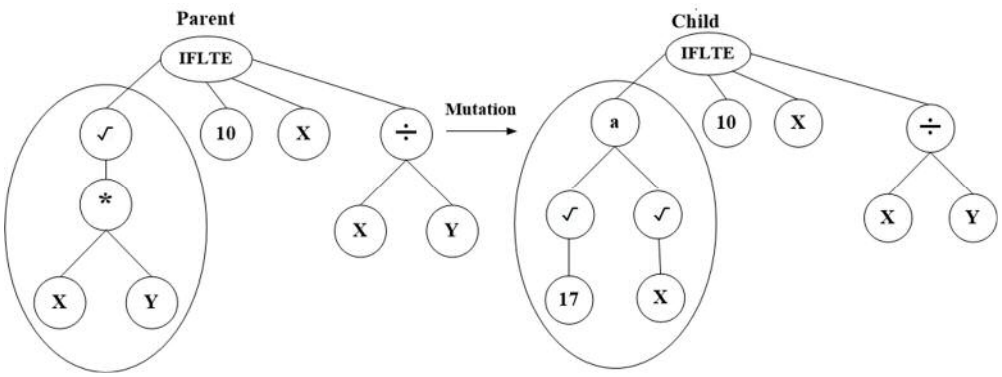
Nếu một cá thể được tái sinh chúng sẽ được sao chép y nguyên vào quần thể, hay nói cách khác là sẽ có hai cá thể giống nhau trong quần thể.

Đánh giá độ tốt (fitness)

Mỗi một chương trình được gán một giá trị được gọi là độ tốt, 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 lai ghép



Hình 3. Toán tử biến dị

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.

b. Lập trình di truyền cho dự báo lượng mưa

Việc sử dụng lập trình di truyền (GP) để dự báo lượng mưa theo thời gian thực đã mở rộng trong những năm gần đây. Madsen và cộng sự [4] đã tiến hành so sánh việc sử dụng các mô hình hồi quy tự động, AR (p), sử dụng GP và mạng nơ-ron để chỉnh sửa về lỗi dư thừa của mô hình có hiệu chỉnh. Để đánh giá ảnh hưởng của chất lượng mô hình mô phỏng lên dự báo, kỹ thuật sửa lỗi (hoặc cập nhật các giá trị đầu ra) đã được áp dụng cho cả mô hình hiệu chỉnh và mô hình không hiệu chỉnh.

Whigham & Crapper [8] đề xuất sử dụng GP dựa trên văn phạm phi ngữ cảnh. Bài báo này đã xác định tập hàm GP bao gồm các hàm số học và hàm mũ. Tập kết bao gồm một số giá trị lượng mưa trong quá khứ cũng như lượng mưa trung bình trong 5, 10, 15, 25, 30, 40, 50, 60 và 100 ngày gần nhất. Năm 2002, Liong và các cộng sự dùng GP để xác định mô hình mô tả mối quan hệ giữa lượng mưa và dòng chảy bằng cách sử dụng mã hóa cổ điển của GP giống như Koza đã định nghĩa [3] Tập hàm bao gồm các hàm số học và hàm mũ, mô hình được xây dựng như bài toán hồi quy. Hai bài báo này đưa ra mô hình nhưng không cung cấp một giải thích vật lý về hiện tượng.

Khu và các cộng sự [2] sử dụng GP và mạng nơ-ron để sinh lỗi thời gian thực dựa trên tiến hóa cập nhật chương trình để bổ sung cho mô hinhd dự báo thời gian thực gọi là WRIP (Bộ xử lý thông tin thời tiết radar) dựa trên các phép đo lượng mưa được ghi lại bằng radar.

Các thử nghiệm đã được thực hiện bằng cách sử dụng tổng lượng mưa và lượng mưa thực tế, với cả mạng nơ-ron và GP để tối ưu hóa lỗi. Đối với chức năng hiệu chỉnh, dữ liệu được ghi lại trong trận mưa tháng 12 năm 1999 ở lưu vực nông thôn ngược dòng từ Taunton, Vương quốc Anh đã được sử dụng và, trong quá trình hợp lệ, một ước tính đã được thực hiện bằng cách sử dụng dữ liệu từ tháng 4 năm 2000. Có thể cải thiện dự báo dòng chảy với mô hình WRIP sử dụng cả lập trình di truyền và một mạng nơ-ron nhân tạo bằng cách cập nhật lỗi theo thời gian thực giữa dòng chảy cần đo và giá trị mô phỏng cho tối đa năm khoảng thời gian. Phương trình là kết quả của GP có thể được xem như một dạng cải tiến của mô hình hồi quy tự động.

Trong nước hầu như chưa có nghiên cứu nào rõ ràng về bài toán dự báo lượng mưa, và hầu chắc chắn rằng chưa có nghiên cứu nào sử dụng công cụ học máy để dự báo lượng mưa.

3. Thí nghiệm

Trong phần này nghiên cứu 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 lượng mưa.

a. Tham số của GP

Bảng 1. Tham số của 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	Tranh đấu kích thước

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

GP chạy 30 lần mỗi lần với giá trị khởi tạo khác nhau, mỗi lần sẽ nhận được một lời giải tốt nhất sau đó lựa chọn lời giải trung vị (median) của dãy 30 lời giải tốt nhất đó dùng làm mô hình cuối cùng.

b. Dữ liệu bài toán

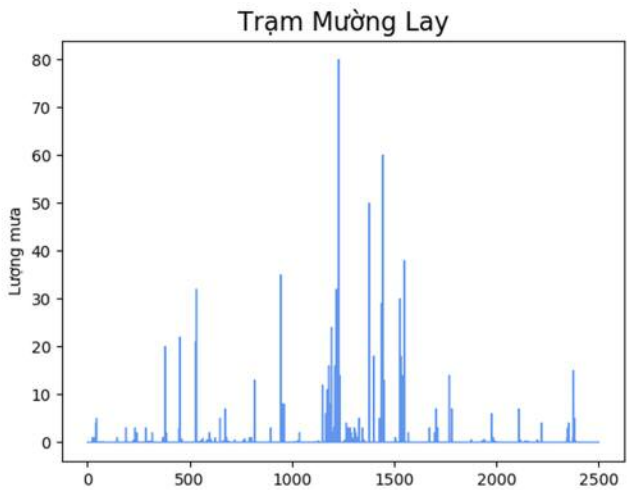
Dữ liệu thử nghiệm là dữ liệu đo được ở Mường Lay (trạm 1), Lào Cai (trạm 2), Hà Giang (trạm 3), Sơn La (trạm 4), Cao Bằng (trạm 5), Điện Biên (trạm 6), Tuyên Quang (trạm 7). Dữ liệu được lấy từ ngày 1/6/2016 đến ngày 1/1/2019, mỗi ngày gồm 8 giá trị (đo cách nhau 3 giờ).

Sau khi có dữ liệu dạng chuỗi thời gian, nghiên cứu chuyển thành dữ liệu phụ thuộc có dạng:

$$r_t = f(r_{t-1}, r_{t-2}, .., r_{t-\tau}) \quad (1)$$

Trong đó r_t là lượng mưa ở thời điểm t.

Sau khi chuyển về dữ liệu phụ thuộc và chọn $\tau = 6$, nghiên cứu lấy 5000 bản ghi làm dữ liệu huấn luyện, phần còn lại 2556 bản ghi làm dữ liệu kiểm tra.



Hình 4. Một số giá trị lượng mưa đo được tại trạm Mường Lay

c. Tổng quan các kỹ thuật học máy

Để so sánh GP với các kỹ thuật học máy khác khi giải quyết bài toán dự báo lượng mưa, nghiên cứu lựa chọn 4 kỹ thuật học máy đưa ra mô hình dự báo chỉ dựa vào dữ liệu và có khả năng phản ánh được ánh xạ 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 mưa. Những mô hình này hoàn toàn dựa trên thông tin có được từ việc thu thập dữ liệu. Đó là các mô hình sau:

Máy vec-tơ hỗ trợ (Support Vector Machine)

Máy véc tơ hỗ trợ hồi quy (Support Vector Regression -SVR) [5], là một phương pháp thành công để xử 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 này nghiên cứu sẽ sử dụng nhân Gauss.

Cây quyết định (Decision Tree - DCT) [6] 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.

k-láng giềng gần nhất (k Nearest Neighbor - kNN) [1] 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.

Mạng Perceptron nhiều lớp (Multi-layer Perceptron - MLP) [7] 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. Bốn mô hình trên được sử dụng rất phổ biến cho các bài toán học máy và cũng 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 nghiên cứu sử dụng hai độ đo như công thức (6, 7):

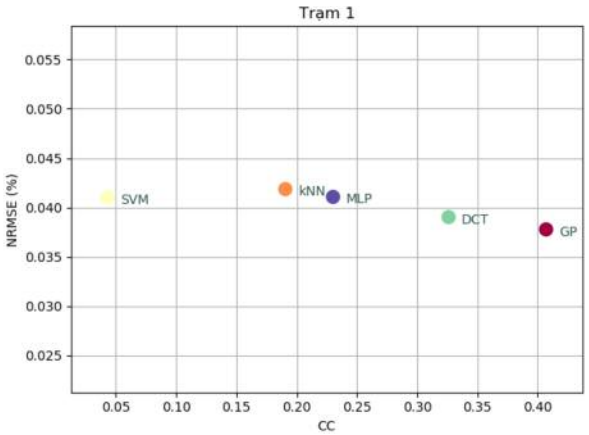
$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{obs,i} - y_{pre,i})}}{(y_{obs,max} - y_{obs,min})} \quad (6)$$

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

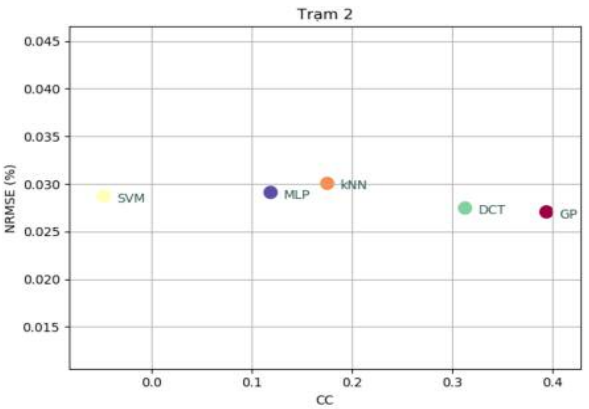
Trong đó NRMSE (normal root mean squared error) là RMSE chuẩn hóa tính theo phần trăm, CC (*correlation coefficien*) là hệ số tương quan.

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 kết quả có giá trị NRMSE nhỏ và CC lớn.



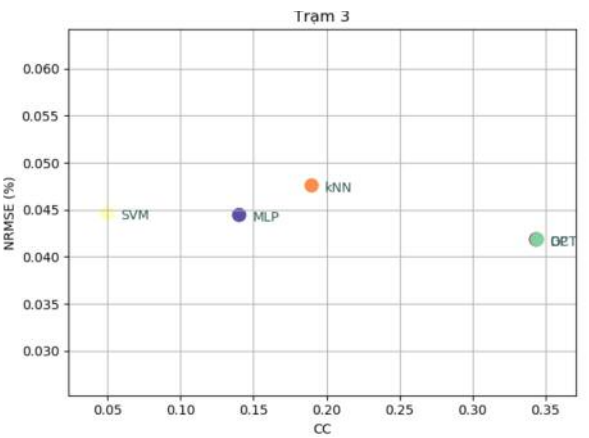
Hình 5. Giá trị NRMSE và CC của các mô hình dự báo với dữ liệu tại trạm Mường Lay



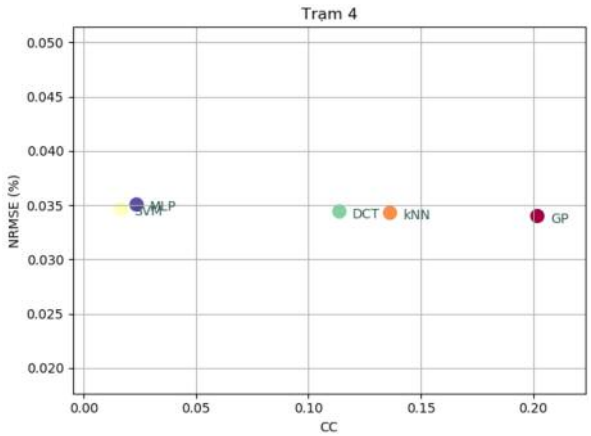
Hình 6. Giá trị NRMSE và CC của các mô hình dự báo với dữ liệu tại trạm Lào Cai

Trong hình 5, giá trị NRMSE của 5 phương pháp dự báo nằm trong khoảng từ 0,035 đến 0,045. Còn giá trị CC nằm trong khoảng từ 0,03 đến 0,4. Và ta cũng thấy phương pháp GP vừa cho kết quả giá trị NRMSE nhỏ và CC lớn nhất trong 5 phương pháp.

Trong hình 6, giá trị NRMSE của 5 phương pháp rơi vào khoảng từ 0,025 đến 0,030, trong đó MLP và SVM khá gần nhau cũng như DCT và GP. Tuy nhiên giá trị CC của SVM nhỏ hơn 0 trong khi giá trị này của các phương pháp còn lại từ 0,1 đến 0,4. Và cũng tương tự như dự báo trạm Mường Lay, lời giải GP cho kết quả tốt nhất cả về NRMSE lẫn CC.



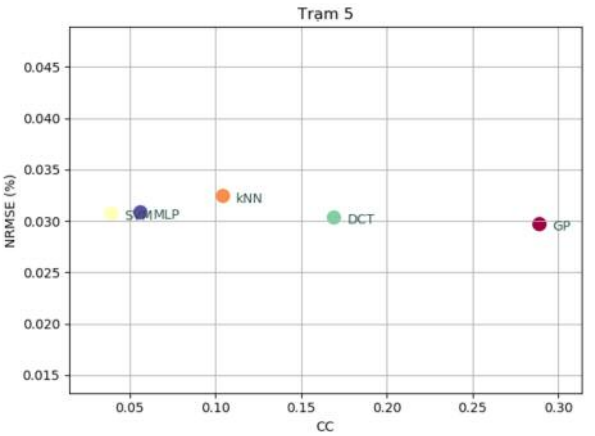
Hình 7. Giá trị NRMSE và CC của các mô hình dự báo với dữ liệu tại trạm Hà Giang



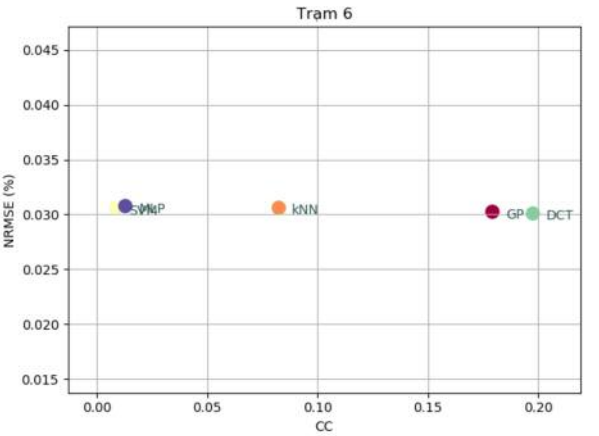
Hình 8. Giá trị NRMSE và CC của các mô hình dự báo với dữ liệu tại trạm Sơn La

Trong hình 7, giá trị NRMSE của các mô hình nằm trong khoảng 0,04 đến 0,05, trong đó giá trị này của SVM và MLP gần như bằng nhau, giá trị DCT và GP trùng nhau. Giá trị CC của các phương pháp nằm trong khoảng từ 0,05 đến 0,35 trong đó DCT và GP gần như bằng nhau. Đối với dữ liệu tại trạm này thì hiệu năng của GP và DCT là tốt nhất.

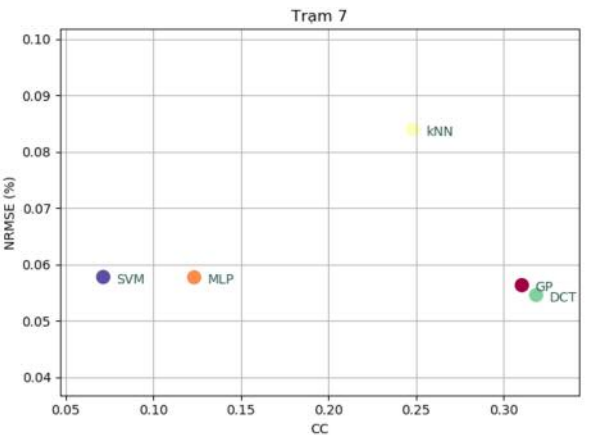
Kết quả trong hình 8 cho thấy giá trị NRMSE của các phương pháp học máy hầu như tương đồng với nhau. Chỉ có giá trị CC là khác biệt trong đó GP có giá trị CC lớn nhất, MLP, SVM là phương pháp cung cấp giá trị CC của dữ liệu dự báo kém nhất.



Hình 9. Giá trị NRMSE và CC của các mô hình dự báo với dữ liệu tại trạm Cao Bằng



Hình 10. Giá trị NRMSE và CC của các mô hình dự báo với dữ liệu tại trạm Điện Biên



Hình 11. Giá trị NRMSE và CC của các mô hình dự báo với dữ liệu tại trạm Tuyên Quang

Kết quả trong hình 9 cho thấy không có sự khác biệt lắm về kết quả của 2 mô hình SVM và MLP, trong khi đó GP vẫn chiếm ưu thế vượt trội cả về hai giá trị này.

Hình 10 cho thấy khả năng dự báo của các mô hình trên dữ liệu trạm 6 không có sự khác biệt về sai số, chỉ có giá trị CC khác nhau, trong đó DCT có giá trị này lớn nhất, GP thứ hai. SVM và MLP hầu như cũng không có khác biệt.

Kết quả trên hình 11 cho thấy kNN có giá trị sai số mô hình dự báo tồi hơn hẳn các mô hình còn lại, SVM và MLP tương đương nhau tuy nhiên CC của MLP tốt hơn. Với dữ liệu trạm 7 mô hình DCT cho kết quả tốt nhất, GP kém chút ít cả về lỗi và giá trị CC.

Như vậy trên 7 tập dữ liệu thực tế tại 7 trạm khác nhau, GP cho mô hình dự báo tốt nhất trên 5 tập dữ liệu. Trên các bài toán còn lại GP chỉ thua kém so với DCT không đáng kể về sai số của mô hình. Các 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 lượng mưa ở trạm 1 là kết quả của quá trình tiến hóa của GP có dạng:

sqrt(add(mul(mul(sqrt(mul(X4,X3)),X2),add(mul(mul(X2,X2),mul(X2,X2)),mul(add(X3,X3),X2))),mul(add(mul(add(X3,X3),sin(X3)),div(mul(X4,X5),sin(X5))),sin(sqrt(X3))))).

Biểu thức tương ứng với cây trên là:

$$\sqrt{\sqrt{X_3 \times X_4} \times X_2 \times (X_2^4 + 2X_3 \times X_2) + 2X_3 \times \sin(X_3) + X_4 \times \frac{X_5}{sinX_5} + sin\sqrt{X_3}} \quad (8)$$

Với mô hình kết quả như trên việc dự báo trở nên khá dễ dàng với các biến X_i chính là các giá trị đầu vào. 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 4 mô hình trên mới có.

5. Kết luận

Bài báo trình bày việc sử dụng GP để dự báo lượng mưa tại một số trạm quan trắc 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). Tuy nhiên giá trị CC của mô hình còn tương đối thấp, tức là kết quả dự báo vẫn còn chưa đoán được đúng xu thế của dữ liệu. Chính vì vậy, trong tương lai nghiên cứu sẽ tiếp tục cải tiến GP để thu được kết quả dự báo tốt hơn nữa. Ngoài ra tham số phụ thuộc vào số giá trị thời điểm trước cũng cần được điều chỉnh linh hoạt để có được kết quả dự báo phù hợp với thực tế.

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", mã số BĐKH.34/16-20."*

Tài liệu tham khảo

1. Hastie, T.T., (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer.

2. Khu, S.T., (2004), *An evolutionary-based real-time updating technique for an operational rainfall-runoff forecasting model*. Proceedings of the 2nd Biennial Meeting of the International Environmental Modelling and Software Society, Manno, Switzerland, 141-146.

3. Koza, J.R., (1992), *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press.

4. Madsen, H.B., (2000), *Data assimilation in rainfall-runoff forecasting*. Hydroinformatics 2000, 4th International Conference of Hydroinformatics , (pp. 1-6). Iowa, USA.

5. ¨olkop, A.J., (2004), *A tutorial on support vector regression*. Statistics and Computing, 14(3), 199-222.

6. Rokach, L., Maimon, O. (Eds). *Data mining with decision trees: theory and applications*. World Scientific Publishing Co., Inc. River Edge, NJ, USA, Series in Machine Perception and Artificial Intelligence, 81, pp. 328.

7. Rosenblatt, F., (1961), *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Arch Gen Psychiatry. 7(3), 218-219.

8. Whigham, P.A. (2001), *Modelling rainfall-runoff using genetic programming*. Mathematical and Computer Modelling, 33, (6–7), 707-721.

A GENETIC PROGRAMMING-BASED RAINFALL PREDICTION USING DATA FROM THE VIETNAM METEOROLOGICAL AGENCY

Nguyen Thi Hien[1], Nguyen Xuan Hoai[2], Dang Van Nam[3], Ngo Van Manh[4]
[1]Le Quy Don Technical University
[2]AI Academy Vietnam
[3]Hanoi University of Mining and Geology
[4]Center for Hydro-Meteorological Data and Information

Abstract: *Rainfall is one of the most challenging variables to predict, as it exhibits very unique characteristics that do not exist in other time series data. Moreover, rainfall is a major component and is essential for applications that surround water resource planning. In particular, this paper is interested in the prediction of rainfall using data from the Vietnam Meteorological Agency. Currently in the rainfall prediction literature, the process of predicting rainfall is dominated by statistical models, namely using a Markov chain extended with rainfall prediction (MCRP). In this paper we outline a new methodology to be carried out by predicting rainfall with Genetic Programming (GP). This is the first time in the literature that GP is used within the context of rainfall prediction in some city at Vietnam. We have used a GP to this problem domain and we compare the performance of the GP and SVM, MLP, DCT, kNN on 3 different data sets of cities at Vietnam and report the results. The goal is to see whether GP can outperform other machine learning methods. Results indicate that in general GP significantly outperforms other machine learning methods, which is the dominant approach in the literature.*

Keywords: *Genetic Programming, rainfall prediction.*