Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

# SO SÁNH CÁC MÔ HÌNH DỰ BÁO LƯỢNG MƯA CHO THÀNH PHỐ CẦN THƠ

Đỗ Thanh Nghị[1], Phạm Nguyên Khang[1], Nguyễn Nhị Gia Vinh[2] và Văn Phạm Đăng Trí[3]

[1] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*
[2] *Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ*
[3] *Khoa Môi trường & Tài nguyên Thiên nhiên, Trường Đại học Cần Thơ*

***Thông tin chung:***
*Ngày nhận: 03/09/2013*
*Ngày chấp nhận: 21/10/2013*

***Title:***
*A comparision of rainfall forecast models for Can Tho city - Vietnam*

***Từ khóa:***
*Dự báo lượng mưa, hồi qui tuyến tính, k láng giềng, cây quyết định, bagging, rừng ngẫu nhiên, máy học véc-tơ hỗ trợ*

***Keywords:***
*Rainfall forecast, linear regression, k nearest neighbors, decision trees, bagging, random forests, support vector machines*

**ABSTRACT**

*In recent years, climate change is one of the environmental problems that needs to be studied in the Mekong Delta of Vietnam, especially those in conjunction with temperature and rainfall. As temperature and rainfall changes directly affect agriculture and aquaculture activities - driving factors of the delta's development, the raising question is if such changes could be forecasted with acceptable level of uncertainties. This paper presents algorithms and models of adjusting the forecasted rainfall data obtained from climate data of the SEA-START. A comparison of these forecast models is conducted by forecast error analysis. A case study is experimented by using rainfall data in Can Tho city - Vietnam. The results show that the linear regression model has the greatest forecast error while the non-linear forecast models give better results. The diversity of these forecast models can be applied to solve environmental problems in practice.*

**TÓM TẮT**

*Trong những năm gần đây, biến đổi khí hậu là một trong những vấn đề môi trường cần được nghiên cứu ở vùng đồng bằng sông Cửu Long - Việt Nam, đặc biệt là những vấn đề liên quan đến các yếu tố nhiệt độ và lượng mưa. Do sự thay đổi nhiệt độ và lượng mưa ảnh hưởng trực tiếp đến các hoạt động nông nghiệp và nuôi trồng thủy sản - những yếu tố chính dẫn đến sự phát triển của vùng đồng bằng Sông Cửu Long, câu hỏi được đặt ra là liệu những thay đổi về nhiệt độ và lượng mưa có thể được dự báo với độ không chắc chắn ở mức có thể chấp nhận được hay không. Bài báo này trình bày các giải thuật và mô hình dự báo lượng mưa từ nguồn dữ liệu khí hậu của SEA-START. Các mô hình dự báo này được so sánh với nhau bằng phương pháp phân tích lỗi dự báo. Các kết quả trong bài báo này cho thấy mô hình hồi qui tuyến tính có lỗi dự báo cao nhất trong khi các mô hình dự báo phi tuyến cho kết quả dự báo tốt hơn. Tính đa dạng của những mô hình dự báo này có thể được ứng dụng để giải các bài toán môi trường trong thực tiễn.*

## 1 GIỚI THIỆU

Hệ thống khí hậu trái đất bao gồm bốn thành phần: khí quyển, đại dương, khu vực đóng băng và sinh quyển. Các tiến trình vật lý như bức xạ, tuần hoàn và mưa phản ứng với các tiến trình sinh học như tiến trình hấp thu carbon do trồng cây, các biến đổi hóa học để hình thành nên hệ thống khí hậu

biến đổi phức tạp (McKuffie *et al*., 2005). Những biến đổi phức tạp này tác động mạnh mẽ đến sản xuất nông nghiệp ở các nước trên thế giới, đặc biệt là các nước ở vùng nhiệt đới. Ở nước ta, các ảnh hưởng của biến đổi khí hậu đối với nguồn tài nguyên nước lên các lĩnh vực nông nghiệp và thủy sản là mối quan tâm hàng đầu của các nhà nghiên cứu thủy văn học. Theo báo cáo của (Bộ Tài nguyên và Môi trường, 2011) thì vùng đồng bằng sông Cửu Long (ĐBSCL) là vùng đất thấp ven biển của Việt Nam và là khu vực bị tác hại nặng nề nhất do biến đổi khí hậu gây ra. Thành phố Cần Thơ nằm ở trung tâm ĐBSCL với đặc điểm là nắng nhiều và nhiệt độ cao quanh năm. Mùa mưa kéo dài từ tháng 5 đến tháng 10, mùa khô từ tháng 11 đến tháng 4 năm sau. Ngoài ra do nằm cạnh sông Hậu nên Cần Thơ có mạng lưới sông, kênh, rạch khá chằng chịt. Vùng tứ giác Long Xuyên có địa hình thấp trũng và chịu ảnh hưởng lũ trực tiếp hàng năm. Theo báo cáo của Bộ Tài nguyên và Môi trường năm 2011 thì trị số phổ biến của lượng bức xạ tổng cộng trung bình năm là 150-170 kcal/cm2 và trị số phổ biến về lượng mưa trung bình năm khoảng 1600 đến 2000 mm (Bộ Tài nguyên và Môi trường, 2011). Lượng mưa ngày lớn nhất ở Thành phố Cần Thơ khoảng 150-350 mm. Cả mùa mưa có từ 4 đến 6 tháng mưa trên 200 mm/tháng. Việc biến đổi khí hậu sẽ làm thiệt hại cho sản xuất nông nghiệp do đất đai bị bạc màu và nhiễm mặn, hạn hán bất thường, lũ lụt không theo qui luật và nhiều dịch bệnh mới hình thành,...

Nghiên cứu về tác động của biến đổi khí hậu đối với tài nguyên nước đòi hỏi phải chi tiết hóa (downscaling) lượng mưa hằng ngày từ các dự báo cấp khu vực (Regional Climate Model - RCM). Bài báo này đề xuất một phương pháp downscaling hai bước để dự báo lượng mưa hằng ngày. Bước đầu tiên thực hiện việc dự báo một ngày nào đó có mưa hay không. Bước thứ hai sẽ dự báo lượng mưa nếu như ngày đó được dự báo là có mưa ở bước một.

Các phần tiếp theo của bài báo này như sau: phần 2 trình bày ngắn gọn về các nghiên cứu liên quan đến mô hình dự báo lượng mưa, phần 3 trình bày các mô hình dự báo lượng mưa. Phần 4 trình bày các kết quả thực nghiệm, tiếp theo sau đó là phần kết luận và hướng phát triển.

## 2 NGHIÊN CỨU LIÊN QUAN

Các chuyên gia đã sử dụng mô hình tuần hoàn tổng quát (GCM - General Circulation Model) để thiết kế mô hình và mô phỏng các tiến trình biến đổi khí hậu trong phạm vi toàn cầu (Ghosh *et al*., 2008). Mô hình GCM sử dụng các biến thời tiết có độ phân giải thấp để dự báo các biến đổi khí hậu dài hạn và trung hạn cho các vùng với phạm vi rộng lớn, do đó làm cho các chuyên gia khó khăn trong việc dự báo ảnh hưởng của biến đổi khí hậu đối với nguồn tài nguyên nước tại các vùng có phạm vi nhỏ. Việc biến đổi kết quả đầu ra của mô hình GCM để dự báo biến đổi khí hậu tại các vùng có phạm vi nhỏ hơn (như: cấp xã, ấp, cánh đồng) là một bài toán khó vì mô hình GCM không đề cập đến các tiến trình cơ bản xảy ra ở các vùng có phạm vi nhỏ (ví dụ: tiến trình bốc hơi nước, hấp thụ nước, phân bố lượng mưa).

Các phương pháp downscaling đã được phát triển để tạo sự liên hệ giữa kết quả đầu ra của mô hình GCM có độ phân giải thấp với các biến thời tiết có độ phân giải cao hơn ở các vùng có phạm vi nhỏ. Các phương pháp downscaling có thể được phân thành hai nhóm chính: downscaling thống kê và downscaling động. Phương pháp downscaling thống kê có thể được chia thành bốn nhóm: phân loại thời tiết (weather typing method) (Bárdossy *et al*., 1992; Von Storch *et al*., 1993; Bárdossy , 1997), bộ sinh dữ liệu thời tiết ngẫu nhiên (stochastic weather generator) (Selker and Haith, 1990; Tung and Haith, 1995; Yu *et al*., 2002), phương pháp lấy mẫu lại (resampling method) (Murphy, 2000; Buishand and Brandsma, 2001; Palutikor *et al*., 2002) và phương pháp hồi quy (regression method).

Phương pháp hồi quy thiết lập một hàm tuyến tính hoặc phi tuyến thực nghiệm giữa các biến thời tiết ở cấp độ vùng có phạm vi nhỏ (cấp độ địa phương-local scale) và các biến ở cấp độ toàn cục (global scale) của mô hình GCM. Phương pháp này thường được sử dụng vì dễ cài đặt. Ngoài ra, hàm hồi quy cho downscaling có thể được xây dựng bằng mạng nơ-ron (Neural network) (Hewitson and Crane, 1996; Olsson *et al*., 2001; Dibike and Coulibaly, 2006), phân tích tương quan chính tắc (Burger, 1996; Menzel and Burger, 2002; Chu *et al*., 2008) hay máy học véc-tơ hỗ trợ (Support vector machine) (Tripathi *et al*., 2006; Anamdhi *et al*., 2008). Nghiên cứu của (Chen *et al*., 2010) đề xuất kết hợp mô hình phân lớp (mưa hay không mưa) và mô hình hồi quy sử dụng máy học véc tơ hỗ trợ.

Nhiều mô hình và phần mềm downscaling đã được hình thành và phát triển. Nhưng mô hình SDSM (Statistical downscaling model) của Wilby *et al*. (2002) được sử dụng nhiều nhất. Ví dụ như, Wilby *et al*. (2006) đã kết hợp SDSM với một mô hình cân bằng nước và mô hình chất lượng nước cân bằng để nghiên cứu đánh giá ảnh hưởng của

biến đổi khí hậu và sự không chắc chắn trong các dòng chảy của sông. Thêm vào đó, SDSM thường được so sánh với các phương pháp downscaling thống kê.

Trong nghiên cứu dự báo lượng mưa, chúng tôi trước tiên sử dụng mô hình hồi quy tuyến tính. Tiếp đến, nghiên cứu tập trung vào hướng tiếp cận dựa trên các mô hình máy học tự động như: k láng giềng (k Nearest Neighbors) (Fix and Hodges, 1952), cây quyết định (Decision Trees) (Breiman *et al*., 1984), bagging (Breiman, 1996), rừng ngẫu nhiên (Random Forests) (Breiman, 2001) và máy học véc tơ hỗ trợ (Support Vector Machines) (Vapnik, 1995). Chúng tôi cũng đề xuất mô hình học phân cấp kết hợp giữa mô hình phân lớp và mô hình hồi quy dựa trên rừng ngẫu nhiên và máy học véc tơ hỗ trợ.

## 3 MÔ HÌNH DỰ BÁO

### 3.1 Mô hình hồi quy tuyến tính (linear regression - LM)

Hồi quy là phương pháp toán học được áp dụng thường xuyên trong thống kê để phân tích mối liên hệ giữa các hiện tượng kinh tế xã hội. Hồi quy tuyến tính được sử dụng rộng rãi trong thực tế do tính chất đơn giản hóa của hồi quy.

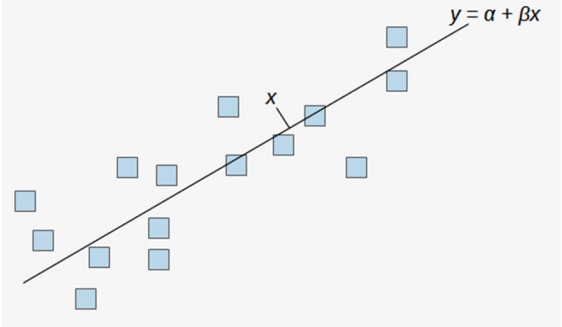
Phân tích hồi quy là phân tích thống kê để xác định mối quan hệ giữa biến phụ thuộc ***y*** với một hay nhiều biến độc lập ***x***. Mô hình hồi quy đơn giản nhất là hàm tuyến tính (bậc 1) dùng để mô tả mối quan hệ giữa biến phụ thuộc và biến độc lập là tuyến tính. Mô hình hồi quy tuyến tính có dạng:

$$y = \alpha + \beta x \quad (1)$$

với α là chặn (intercept), β là độ dốc (slope)

Các tham số *α, β* của mô hình được ước lượng từ dữ liệu quan sát. Xét tập dữ liệu gồm ***m*** phần tử $x_1, x_2, \ldots, x_m$ trong không gian ***n*** chiều (biến độc lập, thuộc tính), có giá trị tương ứng của biến phụ thuộc (cần dự báo) là $y_1, y_2, \ldots, y_m$. Các tham số *α, β* của mô hình được ước lượng bằng phương pháp bình phương bé nhất (least squares):

$$Min\left(\sum_{i=1}^{m}\left[y_i - (\alpha + \beta x_i)\right]^2\right) \quad (2)$$



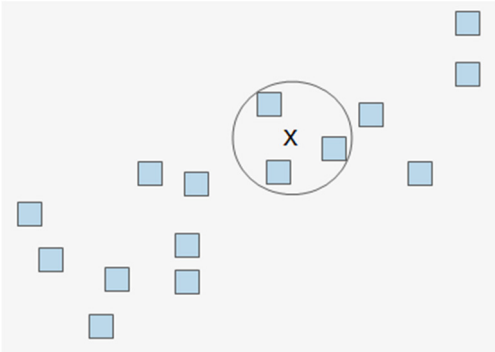

**Hình 1: Hồi quy tuyến tính**

Giá trị dự báo cho phần tử mới ***x*** dựa vào công thức (3):

$$\hat{y} = \alpha + \beta x \quad (3)$$

### 3.2 k láng giềng (k Nearest Neighbors - kNN)

Giải thuật k láng giềng (kNN) được Fix và Hodges đề xuất từ những năm 1952. Đây là phương pháp rất đơn giản nhưng cũng cho hiệu quả cao trong khai mỏ dữ liệu (Hastie *et al*., 2009; Wu and Kumar, 2009). Giả sử có tập dữ liệu bao gồm ***m*** phần tử $x_1, x_2, \ldots, x_m$ trong không gian ***n*** chiều, có giá trị tương ứng của biến phụ thuộc là $y_1, y_2, \ldots, y_m$.




**Hình 2: Giải thuật k láng giềng**

Giải thuật kNN không có quá trình học. Khi dự đoán giá trị biến phụ thuộc của phần tử dữ liệu ***x*** mới đến, giải thuật đi tìm ***k*** láng giềng (*k=1, 2, ...*) của ***x*** từ tập dữ liệu học là các phần tử $\{(x_1,y_1), \ldots, (x_k,y_k)\}$, sau đó thực hiện:

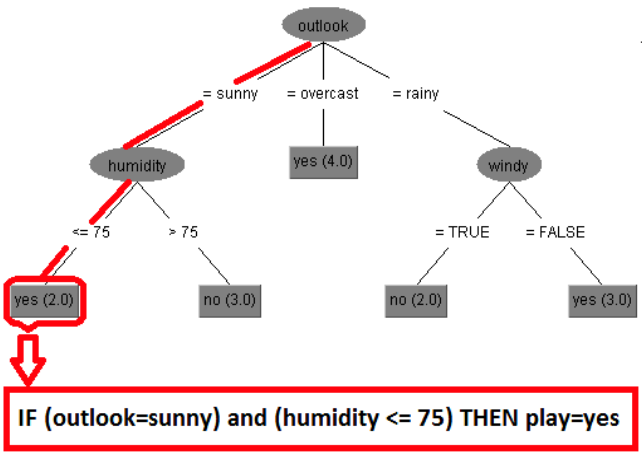
– Phân lớp với bình chọn số đông trong các giá trị $\{y_1, \ldots, y_k\}$,

– Hồi quy với giá trị trung bình của các $\{y_1, \ldots, y_k\}$.

Quá trình tìm k láng giềng của $\boldsymbol{x}$ thường sử dụng khoảng cách (distance) hay độ tương tự (similarity).

### 3.3 Cây quyết định (Decision Trees - DT)

Cây quyết định đề xuất bởi (Breiman *et al.*, 1984; Quinlan, 1993) là mô hình máy học tự động sử dụng rất nhiều trong khai mỏ dữ liệu (Wu and Kumar, 2009) do tính đơn giản và hiệu quả. Hình 3 minh họa một ví dụ của cây quyết định thu được bằng cách học từ tập dữ liệu, để dự đoán chơi Golf (ŷ = yes / no?) từ các biến (thời tiết, nhiệt độ, độ ẩm, gió). Mô hình rất dễ hiểu bởi vì chúng ta có thể rút trích luật quyết định tương ứng với nút lá có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá. Các luật quyết định dễ hiểu với người sử dụng.




| outlook | temp. | hum. | windy | Play |
|---|---|---|---|---|
| sunny | 85 | 85 | false | No |
| sunny | 80 | 90 | true | No |
| overcast | 83 | 78 | false | Yes |
| rain | 70 | 96 | false | Yes |
| rain | 68 | 80 | false | Yes |
| rain | 65 | 70 | true | No |
| ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... |

**Hình 3: Cây quyết định học từ dữ liệu cho phép dự báo chơi Golf**

Xét tập dữ liệu bao gồm $\boldsymbol{m}$ phần tử $\boldsymbol{x_1, x_2, \ldots, x_m}$ trong không gian $\boldsymbol{n}$ chiều, có giá trị tương ứng của biến phụ thuộc là $\boldsymbol{y_1, y_2, \ldots, y_m}$. Giải thuật học từ dữ liệu là quá trình xây dựng cây bắt đầu từ nút gốc đến nút lá. Đây là giải thuật đệ quy phân hoạch tập dữ liệu theo các biến độc lập thành các phân vùng chữ nhật rời nhau mà ở đó các phần tử dữ liệu $\boldsymbol{x_i}$, $\boldsymbol{x_j}$, …, $\boldsymbol{x_k}$ của cùng phân vùng (nút lá) có các $\boldsymbol{y_i}$, $\boldsymbol{y_j}$, …, $\boldsymbol{y_k}$ là thuần khiết:

– Giống nhau trong vấn đề phân lớp,

– Tương tự nhau trong vấn đề hồi quy.

Giải thuật học mô hình cây quyết định từ dữ liệu gồm 2 bước lớn: xây dựng cây, cắt nhánh để tránh học vẹt. Quá trình xây dựng cây được làm như sau:

– Bắt đầu từ nút gốc, tất cả các dữ liệu học ở nút gốc,

– Nếu các phần tử dữ liệu tại 1 nút là thuần khiết thì nút đang xét được cho là nút lá, giá trị dự báo của nút lá cho vấn đề phân lớp với bình chọn số đông trong các giá trị $\{\boldsymbol{y_i, \ldots, y_k}\}$, cho vấn đề hồi quy với giá trị trung bình của các $\{\boldsymbol{y_i, \ldots, y_k}\}$.

– Nếu dữ liệu ở nút quá hỗn loạn (các giá trị $\{\boldsymbol{y_i, \ldots, y_k}\}$ rất khác nhau) thì nút được cho là nút trong, tiến hành phân hoạch dữ liệu một cách đệ quy bằng việc chọn 1 biến để thực hiện phân hoạch tốt nhất có thể.

Một biến được cho là tốt được sử dụng để phân hoạch dữ liệu sao cho kết quả thu được cây nhỏ nhất. Việc lựa chọn này dựa vào các heuristics: chọn biến sinh ra các nút thuần khiết nhất. Hiện nay có 2 giải thuật học cây quyết định tiêu biểu là C4.5 của (Quinlan, 1993), CART của (Breiman *et al.*, 1984).

Để đánh giá và chọn biến khi phân hoạch dữ liệu, Quinlan đề nghị sử dụng độ lợi thông tin (chọn biến có độ lợi thông tin lớn nhất) và tỉ số độ lợi dựa trên hàm entropy của Shannon. Độ lợi thông tin của một biến được tính bằng: độ đo hỗn loạn trước khi phân hoạch trừ cho sau khi phân hoạch. Giả sử $\boldsymbol{p_c}$ là xác suất mà phần tử trong dữ liệu $\boldsymbol{D}$ thuộc lớp $\boldsymbol{y_c}$ (c =1 , C), độ đo hỗn loạn thông tin trước khi phân hoạch được tính theo công thức entropy (4) như sau:

$$Info(D) = -\sum_{c=1}^{C} p_c \log_2 p_c \qquad (4)$$

Độ đo hỗn loạn sau khi sử dụng biến ***A*** phân hoạch dữ liệu ***D*** có ***m*** phần tử thành ***v*** phân vùng kích thước tương ứng là $m_1$, $m_2$, ..., $m_v$ được tính bởi (5):

$$Info_A(D) = -\sum_{j=1}^{v} \frac{m_j}{m} Info(D_j) \quad (5)$$

Độ lợi thông tin khi chọn biến ***A*** phân hoạch dữ liệu ***D*** thành ***v*** phần được tính bởi công thức (6):

$$\text{Gain(A)} = \text{Info(D)} - \text{Info}_A\text{(D)} \quad (6)$$

Giải thuật CART của Breiman và các cộng sự sử dụng chỉ số Gini để phân hoạch dữ liệu trong quá trình xây dựng cây. Giả sử $p_c$ là xác suất mà phần tử trong dữ liệu ***D*** thuộc lớp $y_c$ (c =1, C), chỉ số Gini được tính theo công thức (7):

$$Gini\ (D) = 1 - \sum_{c=1}^{k} p_c^2 \quad (7)$$

Hàm Gini nhỏ nhất khi lớp trong ***D*** bị lệch. Nếu sử dụng biến ***A*** phân hoạch ***D*** kích thước ***m*** thành 2 tập con $D_1$ (kích thước $m_1$) và $D_2$ (kích thước $m_2$), hàm Gini được tính như công thức (8). Biến được chọn phân hoạch dữ liệu là biến cho giá trị chỉ số Gini nhỏ nhất.

$$Gini_A(D) = \frac{m_1}{m} Gini(D_1) + \frac{m_2}{m} Gini(D_2) \quad (8)$$

Cho vấn đề hồi quy, độ đo hỗn loạn thông tin tại phân vùng ***D*** dựa trên độ lệch chuẩn như trong (9) với $\mu$ là giá trị trung bình của các giá trị ***y*** trong ***D***.

$$S(D) = \sum_{i=1}^{k} \frac{(y_i - \mu)^2}{k} \quad (9)$$

Nếu sử dụng biến ***A*** phân hoạch ***D*** kích thước ***m*** thành 2 tập con $D_1$ (kích thước $m_1$) và $D_2$ (kích thước $m_2$), độ hỗn loạn sau khi phân hoạch được tính như công thức (10).

$$S_A(D) = \frac{m_1}{m} S(D_1) + \frac{m_2}{m} S(D_2) \quad (10)$$

Biến được chọn phân hoạch dữ liệu là biến cho giá trị độ hỗn loạn trước khi phân hoạch trừ cho độ hỗn loạn sau khi phân hoạch là nhỏ nhất.

Mô hình cây quyết định sau khi xây dựng thường không mạnh với nhiễu và dễ dẫn đến học vẹt. Tức là mô hình có tính tổng quát thấp, chỉ cần dữ liệu kiểm tra có thay đổi một ít so với dữ liệu học thì cây quyết định dự báo sai. Để khắc phục khuyết điểm này, Quinlan cũng đề nghị các chiến lược cắt nhánh trong giải thuật C4.5. Có 2 lựa chọn hoặc postpruning (cắt nhánh cây sau khi xây dựng cây) hay prepruning (dừng sớm quá trình phân nhánh). Trong thực tế, postpruning được sử dụng nhiều hơn prepruning. Tuy nhiên độ phức tạp của việc cắt nhánh sau khi xây dựng cây rất phức tạp, sử dụng các chiến lược để ước lượng lỗi sinh ra bởi mô hình sau khi cắt nhánh.

### 3.4 Mô hình Bagging (BagDT)

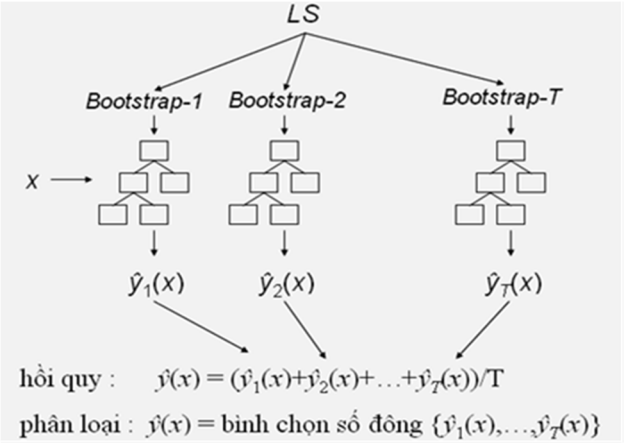
Từ những năm 1990, cộng đồng máy học đã nghiên cứu cách để kết hợp nhiều mô hình phân loại yếu thành mô hình tập hợp phân loại mạnh cải thiện độ chính xác cao hơn so với chỉ một mô hình phân loại đơn yếu. Trong phân tích thành phần lỗi của giải thuật học, Breiman đã chỉ ra trong (Breiman, 1996), lỗi bao gồm 2 thành phần là bias và variance. Thành phần lỗi bias là khái niệm về lỗi của mô hình học (không liên quan đến dữ liệu học) và thành phần lỗi variance là lỗi do tính biến thiên của mô hình so với tính ngẫu nhiên của các mẫu dữ liệu học. Mục đích của các mô hình tập hợp là làm giảm variance và/hoặc bias của các giải thuật học. Dựa trên cách phân tích hiệu quả của giải thuật học dựa trên thành phần lỗi bias và variance, Breiman đã đề xuất giải thuật học Bagging (Bootstrap AGGregatING) nhằm giảm lỗi variance của giải thuật học nhưng không làm tăng lỗi bias quá nhiều. Giải thuật có thể được tóm tắt như sau:

– Từ tập dữ liệu học ***LS*** có ***m*** phần tử, xây dựng ***T*** mô hình cơ sở độc lập nhau.

– Mô hình thứ ***t*** được xây dựng trên tập mẫu Bootstrap thứ ***t*** (lấy mẫu ***m*** phần tử có hoàn lại từ tập học ***LS***).

– Kết thúc quá trình xây dựng ***T*** mô hình cơ sở, dùng chiến lược bình chọn số đông để phân lớp một phần tử ***x*** mới đến hoặc giá trị trung bình cho bài toán hồi quy.

Trong thực tế, giải thuật Bagging cải thiện rất tốt các mô hình đơn không ổn định như cây quyết định và thường có thành phần lỗi variance cao. Hình 4 là ví dụ của giải thuật Bagging được áp dụng cho mô hình cơ sở là cây quyết định.




**Hình 4: Giải thuật Bagging của cây quyết định**

### 3.5 Rừng ngẫu nhiên (Random Forests - RF)

Tiếp cận rừng ngẫu nhiên do (Breiman, 2001) đưa ra là một trong những phương pháp tập hợp mô hình thành công nhất. Giải thuật rừng ngẫu nhiên tạo ra một tập hợp các cây quyết định không cắt nhánh, mỗi cây được xây dựng trên tập mẫu bootstrap (như Bagging), tại mỗi nút phân hoạch tốt nhất được thực hiện từ việc chọn ngẫu nhiên một tập con các thuộc tính. Lỗi tổng quát của rừng phụ thuộc vào độ chính xác của từng cây thành viên trong rừng và sự phụ thuộc lẫn nhau giữa các cây thành viên. Giải thuật rừng ngẫu nhiên xây dựng cây không cắt nhánh nhằm giữ cho thành phần lỗi bias thấp và dùng tính ngẫu nhiên để điều khiển tính tương quan thấp giữa các cây trong rừng. Tiếp cận rừng ngẫu nhiên cho độ chính xác cao khi so sánh với các thuật toán học có giám sát hiện nay. Như Breiman đã đề cập trong (Breiman, 2001), rừng ngẫu nhiên học nhanh, chịu đựng nhiễu tốt và không bị tình trạng học vẹt. Giải thuật rừng ngẫu nhiên sinh ra mô hình có độ chính xác cao đáp ứng được yêu cầu thực tiễn cho vấn đề phân loại, hồi qui. Giải thuật rừng ngẫu nhiên (Hình 5) có thể được trình bày ngắn gọn như sau:

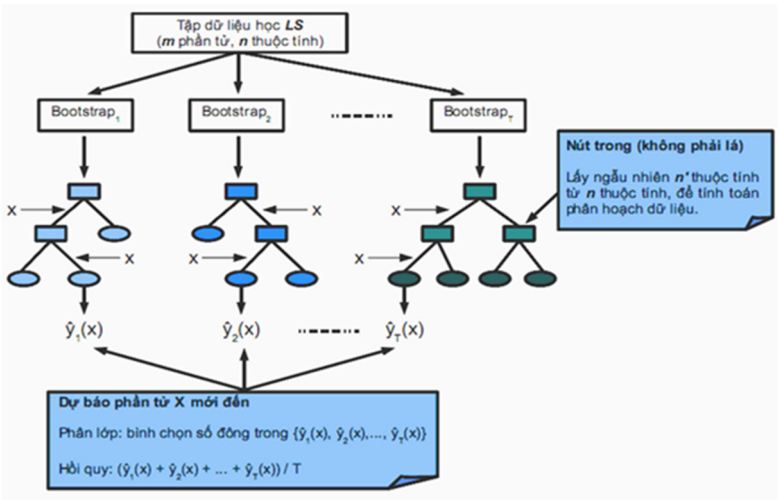
– Từ tập dữ liệu học LS có m phần tử và n biến (thuộc tính), xây dựng T cây quyết định một cách độc lập nhau.

– Mô hình cây quyết định thứ t được xây dựng trên tập mẫu Bootstrap thứ t (lấy mẫu m phần tử có hoàn lại từ tập học LS).

– Tại nút trong, chọn ngẫu nhiên n' biến (n'<<n) và tính toán phân hoạch tốt nhất dựa trên n' biến này.

– Cây được xây dựng đến độ sâu tối đa không cắt nhánh.

Kết thúc quá trình xây dựng T mô hình cơ sở, dùng chiến lược bình chọn số đông để phân lớp một phần tử mới đến hoặc giá trị trung bình cho bài toán hồi quy.

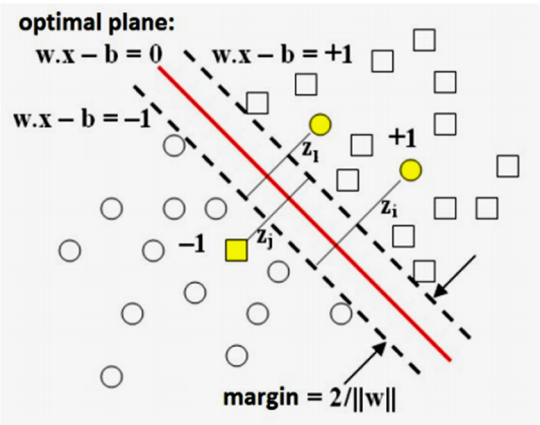



**Hình 5: Giải thuật rừng ngẫu nhiên**

### 3.6 Máy học véc tơ hỗ trợ (Support Vector Machines - SVM)

Máy học véc tơ hỗ trợ (SVM) được đề xuất bởi (Vapnik, 1995) là mô hình hiệu quả và phổ biến cho vấn đề phân lớp, hồi quy những tập dữ liệu có số chiều lớn.

Xét ví dụ phân lớp nhị phân tuyến tính (hình 6) với $m$ phần tử $x_1, x_2, \ldots, x_m$ trong không gian $n$ chiều, có nhãn (lớp) của các phần tử là $y_1, y_2, \ldots, y_m$ có giá trị $1$ hoặc $-1$. $y_i = 1$, nếu $x_i$ thuộc lớp $+1$ (lớp dương, lớp chúng ta quan tâm), $y_i = -1$, nếu $x_i$ thuộc lớp $-1$ (lớp âm hay các lớp còn lại). SVM tìm siêu phẳng tối ưu (xác định bởi véc tơ pháp tuyến $w$ và độ lệch của siêu phẳng $b$) dựa trên 2 siêu phẳng hỗ trợ của 2 lớp. Các phần tử lớp +1 nằm bên phải của siêu phẳng hỗ trợ cho lớp +1, các phần tử lớp -1 nằm phía bên trái của siêu phẳng hỗ trợ cho lớp -1. Những phần tử nằm ngược phía với siêu phẳng hỗ trợ được coi như lỗi. Khoảng cách lỗi được biểu diễn bởi $z_i \geq 0$ (với $x_i$ nằm đúng phía của siêu phẳng hỗ trợ của nó thì khoảng cách lỗi tương ứng $z_i = 0$, còn ngược lại thì $z_i > 0$ là khoảng cách từ điểm $x_i$ đến siêu phẳng hỗ trợ tương ứng của nó).




**Hình 6: Phân lớp tuyến tính với máy học véc tơ hỗ trợ**

Khoảng cách giữa 2 siêu phẳng hỗ trợ được gọi là lề. Siêu phẳng tối ưu (nằm giữa 2 siêu phẳng hỗ trợ) tìm được từ 2 tiêu chí là cực đại hóa lề (lề càng lớn, mô hình phân lớp càng an toàn) và cực tiểu hóa lỗi. Vấn đề dẫn đến việc giải bài toán quy hoạch toàn phương (11):

$$\min \Psi(w, b, z) = (1/2)\,||w||^2 + c\sum_{i=1}^{m} z_i \;\; s.t \qquad (11)$$

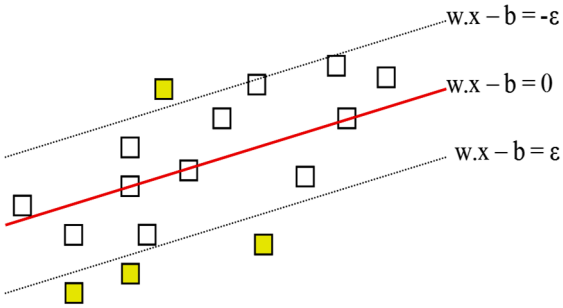
$$y_i(w.x_i - b) + z_i \geq 1$$

$$z_i \geq 0 \;\; (i=1, 2, \ldots, m)$$

với hằng c > 0 được sử dụng để chỉnh độ rộng lề và lỗi.

Giải bài toán quy hoạch toàn phương (11), thu được (*w, b*). Phân lớp phần tử $x$ dựa vào dấu của *(w.x - b)* như trong (12). Nếu giá trị biểu thức *(w.x - b) > 0* thì gán nhãn cho $x$ là lớp dương (+1), ngược lại thì gán nhãn cho $x$ là lớp âm (-1).

$$predict(x) = sign(w.x - b) \qquad (12)$$




**Hình 7: Hồi quy với máy học véc tơ hỗ trợ**

Máy học SVM cũng có thể xử lý bài toán hồi quy. Trong vấn đề hồi quy như Hình 7, SVM tìm siêu phẳng đi qua tất cả các phần tử dữ liệu với độ lệch chuẩn là $\varepsilon$. Huấn luyện máy học SVM cho xử lý vấn đề hồi quy dẫn đến việc giải bài toán quy hoạch toàn phương (13) như sau:

$$\min \Psi(w, b, z^*, z) = (1/2)\,||w||^2 + c\sum_{i=1}^{m}(z_i^* + z_i) \;\; s.t. \qquad (13)$$

$$w.x_i - b - y_i - z_i^* \leq \varepsilon$$

$$w.x_i - b - y_i + z_i \geq -\varepsilon$$

$$z_i^*, z_i \geq 0 (i=1, 2, \ldots, m)$$

với hằng c > 0 được sử dụng để chỉnh độ rộng lề và lỗi.

Giải bài toán quy hoạch toàn phương (13) sẽ thu được siêu phẳng hồi quy ***(w, b)*** của SVM. Dự báo cho phần tử mới đến $x$ dựa trên siêu phẳng ***(w, b)*** được tính theo công thức (14):

$$predict(x) = (w.x - b) \qquad (14)$$

Giải thuật SVM có thể thay thế các tích vô hướng trong các công thức (11-14) bởi hàm nhân (kernel functions), sẽ cho phép giải quyết một số lớn các bài toán phân lớp và hồi quy phi tuyến. Không có bất kỳ một thay đổi nào cần thiết về mặt

giải thuật, việc làm duy nhất là thay thế các tích vô hướng của hai véc tơ trong các công thức bởi một trong các hàm nhân cơ bản được dùng phổ biến như:

– Đa thức bậc *d*: $K(u, v) = (u.v + c)^d$ (15)

– Radial Basis Function (RBF):

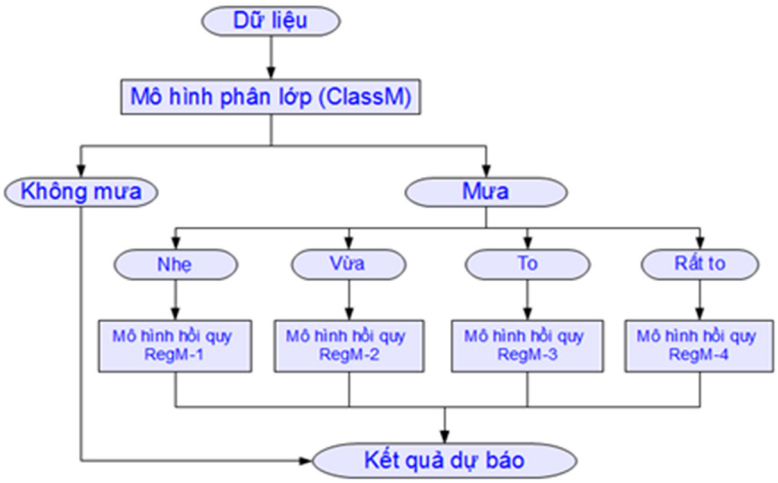
$K(u, v) = exp(-\gamma||u - v||^2)$ (16)

### 3.7 Mô hình hồi quy phân cấp

Chúng ta có thể sử dụng trực tiếp các mô hình hồi quy vừa được trình bày để dự báo lượng mưa trong ngày. Mỗi mô hình đều có ưu điểm và khuyết điểm khác nhau. Chẳng hạn mô hình hồi quy tuyến tính thì rất đơn giản, thời gian xây dựng mô hình và dự báo nhanh, điều tất yếu là độ chính xác cũng không cao. Riêng mô hình kNN cũng đơn giản, chỉ sử dụng duy nhất tham số là ***k = 1, 2, ...*** là số láng giềng, tuy nhiên thời gian dự báo lâu hơn do phải tìm kiếm láng giềng của phần tử cần dự báo. Mô hình cây quyết định chỉ cần duy nhất tham số ***minobj = 1,2, ...*** là số phần tử tối thiểu tại mỗi nút lá, thời gian xây dựng mô hình và dự báo nhanh, đạt được độ chính xác tương đối cao so với kNN và hồi quy tuyến tính. Bagging và rừng ngẫu nhiên thì cần thêm tham số là số lượng cây ***T = 50, 100,...*** riêng, rừng ngẫu nhiên còn sử dụng thêm tham số là số biến ngẫu nhiên sử dụng cho phân hoạch (trong khoảng $\left[\sqrt{n}, \frac{n}{2}\right]$ với ***n*** là số biến của dữ liệu). Cho dù phức tạp, nhưng Bagging, rừng ngẫu nhiên vẫn đơn giản hơn khi so với máy học SVM. Xây dựng mô hình SVM cần thiết ba tham số là hằng số ***c > 0*** (để chỉnh độ rộng lề và lỗi), độ lệch chuẩn là $\boldsymbol{\varepsilon}$ và tham số của hàm nhân. Thời gian xây dựng mô hình và dự báo rất cao (ít nhất là bậc 2 so với số lượng phần tử). Mặc dù phức tạp, nhưng Bagging, rừng ngẫu nhiên và SVM là mô hình phi tuyến, nên xử lý tốt cho các vấn đề phi tuyến, đặc biệt là dự báo lượng mưa đang xét ở đây.

Hình 8 minh họa mô hình hồi quy phân cấp. Dữ liệu được phân lớp (ClassM) vào một trong năm lớp như: không mưa (lượng mưa = 0), mưa nhẹ (lượng mưa: 0-2,5 mm), mưa vừa (lượng mưa: 2,5-7,6 mm), mưa to (lượng mưa: 7,6-50mm), rất to (lượng mưa trên 50 mm). Tương ứng với từng lớp, một mô hình hồi quy được xây dựng cho phép dự báo tốt các phần tử thuộc lớp đó (RegM-i).

Xét về độ phức tạp, xử lý vấn đề phân lớp đơn giản hơn rất nhiều so với bài toán hồi quy. Hơn nữa, quá trình xây dựng mô hình hồi quy càng phức tạp hơn khi cần dự báo lượng mưa từ tập dữ liệu, có mối quan hệ phi tuyến giữa biến phụ thuộc (lượng mưa) với nhiều biến độc lập (bức xạ mặt trời, hướng gió, tốc độ gió, nhiệt độ). Từ phân tích trên, chúng tôi đề xuất mô hình hồi quy phân cấp, kết hợp giữa mô hình phân lớp và nhiều mô hình hồi quy cục bộ để nâng cao hiệu quả xử lý của dự báo lượng mưa.




**Hình 8: Mô hình phân cấp (phân lớp + hồi quy)**

## 4 KẾT QUẢ THỰC NGHIỆM

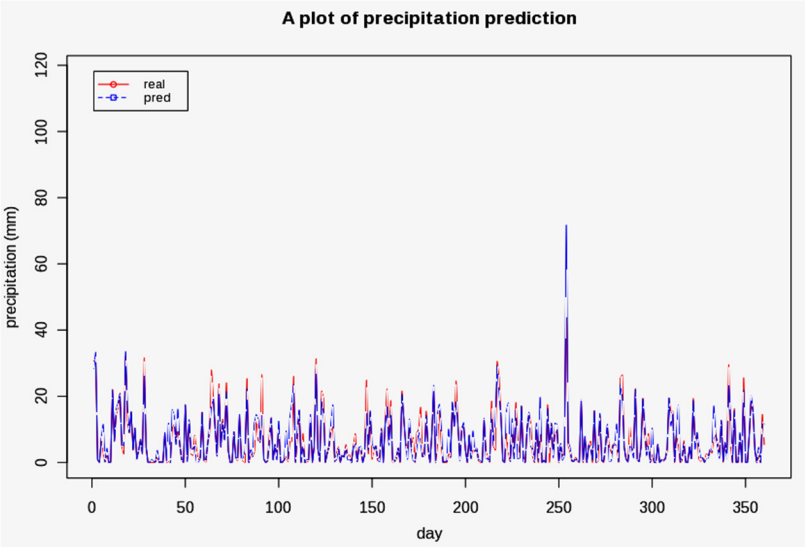
Để tiến hành đánh giá hiệu quả của các mô hình dự báo lượng mưa, chúng tôi tiến hành cài đặt tất cả các chương trình dự báo bằng ngôn ngữ R (Ihaka and Gentleman, 1996) có sử dụng các gói thư viện FNN, rpart, ipred, randomForest, e1071. Chương trình bao gồm các mô hình: Hồi quy tuyến tính (LM), k láng giềng (kNN), Cây quyết định (DT), Bagging (BagDT), Rừng ngẫu nhiên (RF), Máy học véc tơ hỗ trợ cho hồi quy SVR, Mô hình phân cấp: RF phân lớp và RF hồi quy (RFC-RFR), Mô hình phân cấp: SVC phân lớp và SVR hồi quy (SVC-SVR) để dự báo lượng mưa.

**Bảng 1: Kết quả dựa báo lượng mưa của các mô hình**

| Phương pháp | MSE | MAE |
|---|---|---|
| hồi quy tuyến tính (lm) | 34.161274 | 4.406217 |
| k láng giềng (k=5) | 19.815441 | 2.593839 |
| cây quyết định (leaf-size=5) | 15.508522 | 2.052494 |
| Bagging (#trees=100) | **8.970759** | 1.481438 |
| rừng ngẫu nhiên (#trees=100, #randim=3) | 9.131517 | ***1.545566*** |
| máy học SVR (RBF, $\gamma$=0.01, $\varepsilon$=0.1, C= $10^4$) | 17.006430 | 2.455377 |
| mô hình phân cấp RFC-RFR | ***10.412143*** | **1.469576** |
| mô hình phân cấp SVC-SVR | 20.074542 | 2.405764 |

Chúng tôi sử dụng tập dữ liệu (gồm 24 tập con) từ SEA-START RC có địa chỉ website là http://cc.start.or.th. Đây là hệ thống phân phối dữ liệu biến đổi khí hậu nằm trong chương trình hợp tác giữa trung tâm START khu vực Đông Nam Á và ESRI của Thái Lan. Tập dữ liệu thu được từ kết quả của quá trình mô phỏng phức tạp cho từng ngày với kích thước lưới là 20 x 20 km cho toàn bộ khu vực sông Mêkong trong khoảng từ năm 1980 đến năm 2006. Chúng tôi chỉ sử dụng dữ liệu ở lưới gần Cần Thơ (có kinh độ LON = 105.8 và vĩ độ LAT = 10.2). Tập dữ liệu có 9360 dòng (ngày), mỗi dòng có 6 giá trị thuộc tính là nhiệt độ (tmax, tmin), bức xạ mặt trời (solar radiation), hướng gió (wind-dir), tốc độ gió (wind-speed) và lượng mưa (rainfall). Vấn đề chúng ta cần kiểm thử là xây dựng các mô hình dự báo sử dụng tập dữ liệu có được để dự báo lượng mưa (rainfall) từ 5 thuộc tính còn lại. Chúng tôi sử dụng nghi thức kiểm thử hold-out bằng cách lấy ngẫu nhiên 2/3 tập dữ liệu (6240 dòng) làm tập huấn luyện các mô hình dự báo và 1/3 còn lại (3120 dòng) làm tập kiểm tra kết quả dự báo. Kết quả dự báo được đánh giá trên tiêu chí trung bình bình phương lỗi (Mean Square Error - MSE) và trung bình lỗi tuyệt đối (Mean Absolute Error - MAE). Chúng tôi chỉ sử dụng tập huấn luyện để điều chỉnh các tham số của các mô hình. Các tham số này được lựa chọn sao cho đạt tiêu chí lỗi thấp nhất.

Kết quả thu được từ các mô hình dự báo (với các tham số tối ưu) được trình bày trong Bảng 1. Ở hai cột **MSE** và **MAE**, kết quả dự báo với lỗi thấp nhất được in đậm, lỗi thấp thứ hai được in gạch dưới và lỗi thấp thức ba được in đậm và nghiêng.




**Hình 9: Kết quả dự báo 360 ngày của mô hình phân cấp RFC-RFR**

Không có gì ngạc nhiên khi mô hình hồi quy tuyến tính cho lỗi dự báo cao nhất. Trong khi các mô hình dự báo phi tuyến chứng tỏ nhiều ưu thế hơn. Mặc dù vậy, mô hình máy học véc tơ hỗ trợ cho hồi quy SVR và cả mô hình phân cấp SVC-SVR vẫn chỉ thắng thế khi so sánh với kNN và hồi quy tuyến tính. Trong khi đó, mô hình cây quyết định đơn giản cũng cho kết quả rất khả quan khi so sánh với tất cả các mô hình còn lại. Tuy nhiên, hiệu quả nhất vẫn là phương pháp tập hợp mô hình như Bagging, rừng ngẫu nhiên và mô hình phân cấp RFC-RFR, cho phép dự báo rất chính xác lượng mưa (lỗi dự báo thấp). Mô hình Bagging dự báo

với trung bình bình phương lỗi nhỏ nhất trong khi mô hình phân cấp RFC-RFR có thể dự báo với trung bình lỗi tuyệt đối là nhỏ nhất.

Đồ thị về kết quả dự báo lượng mưa của 360 ngày của mô hình phân cấp RFC-RFR được trình bày trong Hình 9. Quan sát đồ thị này, chúng ta có thể thấy rằng mô hình phân cấp RFC-RFR dự báo hiệu quả lượng mưa.

## 5 KẾT LUẬN VÀ ĐỀ XUẤT

Nghiên cứu này đã so sánh các mô hình dự báo theo phương pháp phân tích lỗi dự báo. Phương pháp downscaling hai bước được đề xuất trong bài báo này nhằm dự báo lượng mưa hằng ngày và cho thấy khả năng ứng dụng của các mô hình dự báo lượng mưa trong thực tế. Nghiên cứu ngày đã áp dụng các phương pháp Hồi quy tuyến tính, k láng giềng, Cây quyết định, Bagging, Rừng ngẫu nhiên (RF), Máy học véc tơ hỗ trợ cho hồi quy SVR, Mô hình phân cấp: RF phân lớp và RF hồi quy (RFC-RFR), Mô hình phân cấp: SVC phân lớp và SVR hồi quy (SVC-SVR) để dự báo lượng mưa từ tập dữ liệu của SEA-START ở lưới gần Thành phố Cần Thơ. Kết quả thực nghiệm cho thấy rằng mô hình hồi quy tuyến tính không phù hợp cho dự báo lượng mưa trong khi các mô hình dự báo khác như Bagging, rừng ngẫu nhiên và mô hình phân cấp RFC-RFR dự báo chính xác hơn.

Trong tương lai, chúng tôi sẽ áp dụng các mô hình dự báo này vào dữ liệu thực tế của Thành phố Cần Thơ ngay khi thu thập và tiền xử lý dữ liệu. Chúng tôi có thể nghiên cứu áp dụng cho các vấn đề dự báo tương tự như dự báo mực nước, dự báo lưu lượng cuộc gọi điện thoại,... do các mô hình trong bài là tổng quát cho các vấn đề về dự báo.

## TÀI LIỆU THAM KHẢO


1. P. Aksornsingchai, C. Srinilta. Statistical Downscaling for Rainfall and Temperature Prediction in Thailand. *Proc. of the Intl. MultiConference of Engineers and Computer Scientists*, pp. 356-361, (2011).
2. A. Anandhi, V.V. Srinivas, R.S. Nanjundiah, D.N. Kumar. Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. *International Journal of Climatology*, vol. 28(3):401–420, (2008).
3. A. Bárdossy and E.J. Plate. Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research 28: doi: 10.1029/91WR02589*. ISSN: 0043-1397, (1992).
4. A. Bárdossy. Downscaling from GCMs to local climate through stochastic linkages. *Journal of Environmental Management*, vol. 49(1): 7-17, (1997).
5. T.A. Buishand and T. Brandsma. Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. *Journal Water Resources Research*, Vol.37(11):2761-2776, (2001).
6. Bộ Tài nguyên và Môi trường. Kịch bản biến đổi khí hậu và nước biển dâng cho Thành phố Cần Thơ. *Báo cáo kỹ thuật, Bộ Tài nguyên và Môi trường, Hà Nội*, (2011).
7. L. Breiman, J.H. Friedman, R.A. Olshen and C. Stone. Classification and Regression Trees. *Wadsworth International*, (1984).
8. L. Breiman. Bagging predictors. *Machine Learning* vol. 24(2):123–140, (1996).
9. L. Breiman. Random forests. *Machine Learning* vol. 45(1):5–32, (2001).
10. C.C. Chang and C.J. Lin. LIBSVM - a library for support vector machines. (2011).
11. S.T. Chen, P.S. Yu, Y.H. Tang. Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *Journal of Hydrology* 385:13–22, (2010).
12. C.T. Dhanya, D.N. Kumar. Multivariate nonlinear ensemble prediction of daily chaotic rainfall with climate inputs. *Journal of Hydrology, Elsevier*, vol.403(3-4):292-306, (2011).
13. C.T. Dhanya, D.N. Kumar. Data Mining for Evolving Fuzzy Association Rules for Predicting Monsoon Rainfall of India. *Journal of Intelligent Systems, Freund & Pettman, UK*, vol.18(3):193-209, (2010).
14. K. McKuffie and A. Henderson-Sellers, A Climate Modeling Primer, *John Wiley & Sons Ltd., UK*, ISBN 0-470-85750-1, (2005).
15. E. Fix, J. Hodges. Discriminatoiry Analysis: Small Sample Performance. *Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, USA*, (1952).
16. S. Ghosh, P.P. Mujumdar. Statistical downscaling of GCM simulations to

streamflow using relevance vector machine. *Advances in Water Resources*, vol. 31(1):132-146, (2008).

17. S. Ghosh. SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output. *Journal of Geophysical Research: Atmospheres*, vol.115(D22):1984-2012, (2010).
18. M.K. Goyal, C.S.P. Ojha. Evaluation of Various Linear Regression Methods for Downscaling of Mean Monthly Precipitation in Arid Pichola Watershed. *Natural Resources*, vol.1(1):11-18, (2010).
19. M.Z. Hashmi, A.Y. Shamseldin, B.W. Melville. Statistical downscaling of precipitation: state-of-the-art and application of bayesian multi-model approach for uncertainty assessment. *Hydrology and Earth System Sciences Discuss.* (6):6535-6579, (2009).
20. R. Ihaka, R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, vol.5(3):299-314, (1996).
21. J. Murphy. Predictions of climate change over Europe using statistical and dynamical downscaling techniques. *Intl Journal of Climatology*, Vol.20(5):489-501, (2000).
22. J.P. Palutikof, C.M. Goodess, S.J. Watkins and T. Holt. Generating Rainfall and Temperature Scenarios at Multiple Sites: Examples from the Mediterranean. *Journal of Climate*, Vol.15(24): 3529-3548, (2002).
23. A. Pasini. Neural NetworkModeling in Climate Change Studies. *In Artificial Intelligence Methods in the Environmental Sciences, S. E. Haupt et al. (eds.*), pp. 235-254, (2009).
24. J.R. Quinlan. C4.5: Programs for Machine Learning. *Morgan Kaufmann*, (1993).
25. D. Raje, P.P. Mujumdar. A comparison of three methods for downscaling daily precipitation in the Punjab region. *Hydrological Processes*, vol.25(23):3575–3589, (2011).
26. J.S. Selker and D.A. Haith. Development and testing of single-parameter precipitation distributions. *Water Resources Research 26: doi: 10.1029/90WR01648.* ISSN: 0043-1397, (1990).
27. S. Tripathi, V.V. Srinivasa, R.S. Nanjundiahb. Downscaling of precipitation for climate change scenarios: a support vector machine approach. Journal of Hydrology 330:621–640, (2006).
28. C.P. Tung and D.A. Haith. Global-warming effects on New York streamflows. *Journal of Water Resources Planning and Management*, 121(2), pp. 216-225, (1995).
29. V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, (1995).
30. H. Von Storch, E. Zorita and U. Cubasch. Downscaling of climate change estimates to regional scales: An application to winter rainfall in the Iberian Peninsula. Journal of Climate 6: 1161-1171, (1993).
31. X. Wu and V. Kumar. Top 10 Algorithms in Data Mining. Chapman & Hall/CRC, (2009).
32. H. Yu, S.C. Liu and R.E. Dickinson. Radiative effects of aerosols on the evolution of the atmospheric boundary layer. *Journal of Geophysical Research: Atmospheres*, 107(D12), 4142, doi:10.1029/2001JD000754, (2002).