Journal of educational equipment: Applied research, Volume 2, Issue 301 (November 2023)
ISSN 1859 - 0810

So sánh thuật toán học máy về phân loại lớp phủ bề mặt từ ảnh vệ tinh Sentinel 2 trên nền tảng Google Earth Engine

Đặng Thanh Tùng*, Tạ Minh Ngọc*

** Trường Đại học Tài nguyên và Môi trường Hà Nội*
Received: 15/9/2023; Accepted: 29/9/20223; Published: 10/10/2023

Abstract: *The Google Earth Engine cloud computing platform has proven highly effective in land cover classification. In this study, we utilized the Classification and Regression Tree (CART) and Random Forest (RF) algorithms to classify land cover in Sentinel-2 satellite images. The results in the study area showed significant variations between the two algorithms. Specifically, the CART algorithm achieved an overall accuracy (OA) of 0.92 and a Kappa coefficient of 0.85, while the RF algorithm had an OA of 0.89 and a Kappa coefficient of 0.86.*

Keywords: *Landsat, Land cover, Google Earth Engine, Cart, RF algorithms*

1. Đặt vấn đề

Hiện nay, có nhiều thuật toán được ứng dụng một cách hiệu quả trong việc phân loại lớp phủ bề mặt đất từ ảnh vệ tinh.Trong đó, các thuật toán Cart, RF là các thuật toán đã áp dụng cho việc học máy được sử dụng nhiều trong phân loại lớp phủ từ dữ liệu ảnh vệ tinh. Hiện nay, tại Việt Nam và trên thế giới đã cho thấy tính hiệu quả của việc khai thác các thuật toán Cart hoặc RF và cũng đánh giá được độ tin cậy của từng thuật toán nêu trên [1, 2, 3]. Các kết quả phân loại lớp phủ bề mặt đất từ ảnh vệ tinh phụ thuộc nhiều yếu tố như điều kiện tự nhiên, vị trí địa lý, khí hậu ... của khu vực nghiên cứu, chất lượng ảnh, thời điểm thu nhận ảnh, các loại ảnh, độ phân giải không gian và công tác lấy mẫu, đặc biệt là các thuật toán sử dụng trong công tác xử lý, phân loại ảnh.

Nghiên cứu này tiến hành khai thác dữ liệu ảnh vệ tinh trực tuyến trên nền tảng điện toán đám mây GEE, sử dụng ngôn ngữ JavaScript xây dựng chương trình tính toán và so sánh kết quả phân loại ảnh khi sử dụng các thuật toán Cart và RF tại khu vực quận Long Biên, Hà Nội.

2. Nội dung nghiên cứu

2.1. Dữ liệu và khu vực nghiên cứu

Long Biên, Hà Nội, có diện tích khoảng 60.38 km^2, dân số 271.500 người. Trong nghiên cứu này, nhóm tác giả sử dụng tư liệu ảnh vệ tinh Sentinel 2 với chất lượng hình ảnh rõ ràng, độ phủ mây thấp. Kết quả đã lựa chọn ảnh Sentinel 2 thu nhận trong tháng 9 năm 2023. Đây là dữ liệu với độ phủ mây rất thấp, khoảng 1.0%, các thông tin vật lý của ảnh đảm bảo chất lượng để tiến hành nghiên cứu. Hình 2.1 thể hiện ảnh vệ tinh khu vực nghiên cứu.

2.2. Phương pháp nghiên cứu

Nhóm tác giả lựa chọn 6 lớp phủ để phân loại lần lượt bao gồm: 1) lớp phủ Đất trống, 2) lớp phủ Mặt nước, 3) lớp phủ Cây lâu năm, 4) lớp phủ Cây hàng năm, 5) lớp phủ Dân cư, 6) lớp phủ Giao thông. Các bước phân loại được tiến hành: Thu thập dữ liệu ảnh Sentinel 2 (level 1T) từ GEE; Lọc và lựa chọn ảnh có độ phủ mây là thấp nhất; Lấy mẫu theo các vị trí để phục vụ phân loại; Phân loại theo các thuật toán Cart và RF; Thu nhận kết quả ảnh phân loại theo thuật toán Cart và RF; Đánh giá độ chính xác của ảnh sau phân loại theo các thuật toán trên; So sánh kết quả ảnh sau phân loại của các thuật toán.

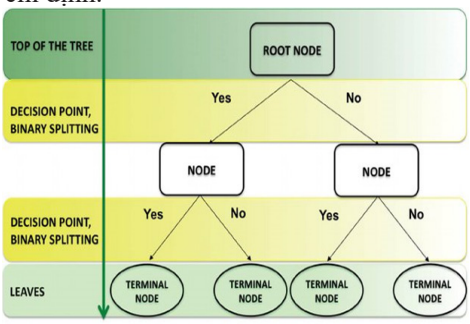
Các thuật toán sử dụng trong phân loại của nghiên cứu này bao gồm Cart, RF và SVM được trình bày theo các nội dung dưới đây:

a) Thuật toán Cart: Thuật toán Cart là một thuật toán học máy có giám sát trong hệ thống phân loại dựa trên cây quyết định (Decision tree) và sử dụng các mẫu huấn luyện để xác định, nhận dạng, phân loại đối tượng trên ảnh viễn thám Cart được sử dụng rộng rãi để phân loại viễn thám, nó còn được gọi là cây phân loại và hồi quy [4]. Thuật toán Cart chia không gian n chiều thành các hình chữ nhật không chồng lên nhau bằng phép đệ quy. Đầu tiên, một biến độc lập x_i được chọn, và sau đó xác định một giá trị ui tương ứng. Không gian n chiều được chia thành hai phần. Một số điểm thỏa mãn $x_i \leq u_i$, và những điểm khác thỏa mãn $x_i > u_i$. Đối với một biến không liên tục, chỉ có hai giá trị là bằng hoặc không bằng nhau. Trong quá trình xử lý đệ quy, hai phần này dựa vào bước đầu tiên để chọn lại một thuộc tính và tiếp tục phân vùng cho đến khi chia hết không gian n chiều. Các thuộc tính có giá trị hệ số GINI tối thiểu được sử dụng làm chỉ mục phân vùng. Đối với tập dữ liệu D, hệ số GINI được xác định theo công thức (1) như sau:

$$GINI*(D)=\sum_{i=1}^{k} p_i*(1-p_i)=1-\sum_{i=1}^{k} p_i^2 \quad (1)$$

Trong đó k là số loại mẫu và p_i biểu thị xác suất một mẫu được xếp vào loại i. Giá trị GINI càng nhỏ có nghĩa là chất lượng của mẫu càng cao và hiệu ứng phân loại càng tốt.

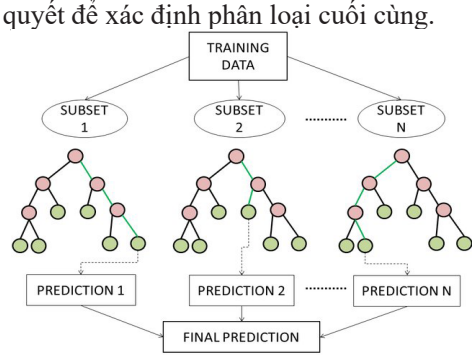
Cây quyết định bao gồm các nút nhiều cấp và nhiều lá. Các nút tối đa đề cập đến số lượng lá tối đa trên mỗi cây và quần thể lá tối thiểu là số lượng nút tối thiểu chỉ được tạo cho tập huấn luyện. Để xây dựng một cây phù hợp, phải tạo đủ các nút và nhánh. Giá trị nút tối đa là không giới hạn nếu nó không được chỉ định.



Hình 2.1. Mô hình phân loại theo thuật toán Cart.

b) Thuật toán RF**:** RF là một thuật toán học tích hợp có thể tích hợp nhiều cây quyết định và sau đó tạo thành một khu rừng. Thuật toán kết hợp các tính năng ngẫu nhiên để tạo ra một cây. Phương pháp đóng bao được sử dụng để tạo các mẫu huấn luyện và mỗi tính năng đã chọn được rút ngẫu nhiên bằng cách thay thế N (kích thước của tập huấn luyện ban đầu). Sau đó, kết quả dự đoán cuối cùng thu được bằng cách kết hợp nhiều cây quyết định [5]. Công thức (2) thực hiện quyết định phân loại cuối cùng như sau: $H_{(x)}=\text{argmax}Y\sum_{i=1}^{k} I(h_{i(x)}=Y)$ (2)

Trong đó H(x) là mô hình kết hợp, hi là mô hình phân loại của cây quyết định đơn lẻ, Y là biến đầu ra (hoặc biến mục tiêu) và I (·) là hàm chỉ báo. Công thức cho thấy rằng RF sử dụng đa số các quyết định biểu quyết để xác định phân loại cuối cùng.



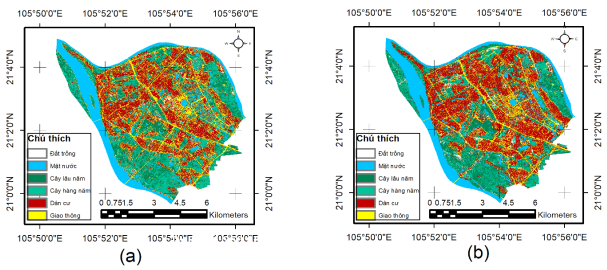
Hình 2.2. Mô hình phân loại theo thuật toán RF.

Tham số điều chỉnh của thuật toán RF là số lượng cây và số lượng cây được chọn theo kinh nghiệm. Trong các bài toán phân lớp dữ liệu thì thuật toán RF được sử dụng phổ biến. Thuật toán RF được đánh giá cao bởi tính chính xác của mô hình. Nhược điểm chính của thuật toán RF là khối lượng tính toán lớn.

c) Phương pháp đánh giá độ chính xác: Ma trận nhầm lẫn (Confusion Matrix) là phương pháp quan trọng và phổ biến được sử dụng để đánh giá độ chính xác, có thể mô tả độ chính xác của phân loại và chỉ ra sự nhầm lẫn giữa các lớp đối tượng. Các thống kê cơ bản cho ma trận nhầm lẫn bao gồm: Sai số tổng thể (Overall Accuracy - OA), Sai số người dùng (User's Accuracy - UA), Sai số nhà sản xuất (Producer's Accuracy - PA) và hệ số Kappa. Trong đó hệ số Kappa có giá trị từ 0.4 đến 0,6 được đánh giá là đạt kết quả trung bình, giá trị từ lớn hơn 0.6 đến 0.8 là tốt và hơn 0.8 đến 1.0 là rất tốt.

2.3. Kết quả nghiên cứu và thảo luận

Kết quả của nghiên cứu bao gồm ba sản phẩm ảnh sau khi phân loại theo các thuật toán Cart và RF. Mỗi một ảnh sau phân loại bao gồm 6 lớp phủ được thể hiện tại Hình 2.3.

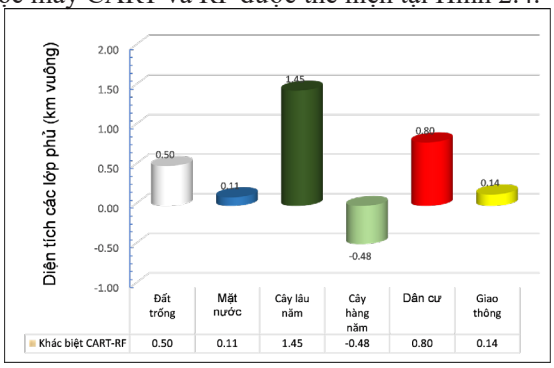


Hình 2.3. Ảnh sau phân loại: a. Phân loại theo Cart; b. Phân loại theo RF

Các lớp phủ bề mặt khu vực quận Long Biên sau phân loại được thể hiện rõ ràng theo từng thuật toán đã sử dụng. Trong đó lớp phủ Dân cư được xác định là tập trung, có mật độ cao phân bố tương đối đều trên khu vực nghiên cứu. Lớp Mặt nước chiếm diện tích đa số là mặt nước sông Hồng. Lớp Đất trống tập trung không cao, nằm rải rác xen kẽ nhau, trong khi đó các lớp thực vật phân bố nhiều ở các khu vực xung quanh của quận Long Biên. Sự phân bố các lớp phủ theo khu vực như trên phù hợp với đặc điểm tự nhiên, phân bố dân cư, tập quán canh tác và phát triển kinh tế xã hội tại quận Long Biên, Hà Nội.

Kết quả phân loại theo các thuật toán Cart, RF có sự khác biệt về diện tích đối với lớp là Đất trống, lớp Cây hàng năm. Các giá trị khác biệt về diện tích của các lớp trên là 05 và -0.48 km^2 tương ứng với 0.78%

và -0.74%. Đối với lớp là Cây lâu năm, sự khác biệt của kết quả giữa hai thuật toán có giá trị cao nhất với diện tích khác biệt 1.45 km^2 tương ứng 2.25%. Lớp Dân cư có khác biệt diện tích là 1.24% tương ứng diện tích 0.8 km^2. Lớp Giao thông có khác biệt 0.22% tương đương 0.14 km^2. Khác biệt nhỏ nhất là lớp Mặt nước với 0.11 km^2 ương đương 0.17%. Các sự khác biệt về kết quả phân loại giữa hai thuật toán học máy CART và RF được thể hiện tại Hình 2.4.



Hình 2.4. Kết quả phân loại: a. So sánh kết quả phân loại các lớp phủ; b. Tỷ lệ phần trăm theo diện tích tự nhiên

Việc đánh giá độ chính xác của sản phẩm phân loại dựa trên ma trận nhầm lẫn (Confsion matrix). Tỷ lệ điểm kiểm tra và tổng số điểm lấy mẫu lần lượt là 30 % và 70%. Kết quả các độ chính xác sau phân loại thể hiện trong Bảng 2.1. Theo các đánh giá về giá trị của hệ số Kappa đạt từ trên 0.41 đến 0.60 là có độ chính xác trung bình; Kappa từ 0.61 đến 0.80 là có độ chính xác tốt; Kappa từ 0.81 đến 1.0 là rất tốt, còn dưới 0.40 là độ chính xác kém. Như vậy độ chính xác phân loại theo cả hai thuật toán Cart và RF đều đạt loại tốt.

Bảng 2.1. Độ chính xác phân loại ảnh

	Cart	RF
Overall Accuracy (OA)	0.92	0.89
Kappa	0.85	0.86

Phương pháp lấy mẫu sử dụng trong nghiên cứu này được lấy mẫu theo các vùng đặc trưng của từng lớp phủ trên ảnh vệ tinh. Chất lượng mẫu huấn luyện được sử dụng là một trong những yếu tố quan trọng ảnh hưởng đến khả năng phân loại ảnh. Nghiên cứu này cho thấy độ chính xác sau phân loại của thuật toán Cart và RF là gần như tương đương, sự khác biệt không quá lớn. Kết quả của nghiên cứu cũng tương đồng với một số nghiên cứu khác trên thế giới, tuy nhiên cũng có nghiên cứu đưa ra báo cáo ngược lại. Điều này còn phụ thuộc vào chất lượng mẫu, chất lượng dữ liệu ảnh đầu vào, đặc điểm phân bố các lớp phủ của từng khu vực nghiên cứu và nhiều nguyên nhân khác.

3. Kết luận và đề xuất

Trong lĩnh vực Quản lý đất đai, việc ứng dụng công nghệ trí tuệ nhân tạo với các thuật toán học máy và khai thác nguồn dữ liệu ảnh vệ tinh trên nền tảng điện toán đám mây đã đem lại hiệu quả cao, đảm bảo độ tin cây và chi phí rất thấp. Dữ liệu ảnh Sentinel-2 sử dụng trong nghiên cứu là nguồn tài nguyên miễn phí và được coi là dữ liệu đa thời gian, liên tục gần thời gian thực đã tạo ra sản phẩm là hiện trạng các lớp phủ bề mặt tại quận Long Biên, Hà Nội năm 2023. Độ chính xác của công tác phân loại theo các thuật toán học máy CART và RF đạt yêu cầu cao và có độ chính xác phân loại gần như tương đương. Nghiên cứu này mới chỉ phân loại ảnh với 6 lớp phủ từ nguồn dữ liệu ảnh vệ tinh miễn phí. Để có những kết quả chi tiết hơn, các nghiên cứu sau có thể phân loại nhiều lớp phủ chi tiết hơn. Có thể sử dụng nhiều hơn hai thuật toán ngoài CART và RF để phân tích đánh giá hiệu quả của từng thuật toán đối với các khu vực nghiên cứu khác nhau.

Kết quả nghiên cứu có khả năng đóng góp nhất định cho việc phân tích, sử dụng dữ liệu ảnh vệ tinh đầu vào và các thuật toán học máy một cách hiệu quả, có độ tin cậy cao cho các công tác nghiên cứu về phân tích, theo dõi biến động lớp phủ bề mặt, lớp phủ sử dụng đất trong quản lý đất đai, quản lý môi trường tại những khu vực khác có điều kiện tương tự.

(Nghiên cứu này được sự hỗ trợ của Ttrường Đại học Tài nguyên và Môi trường Hà Nội, Khoa Quản lý đất đai trong chương trình nghiên cứu của đề tài mã số 13.01.23.M.03.)

Tài liệu tham khảo

[1] Vũ Hữu Long và cộng sự (2019). *Ứng dụng công nghệ xử lý ảnh viễn thám trên nền tảng điện toán đám mây (GEE) trong theo dõi biến động đường bờ sông – Thí điểm tại sông Cửu Long.* Tạp chí Khoa học Tự nhiên và Công nghệ. 16, 38.

[2] Bùi Thị Hồng Thắm, Trịnh Thị Thu (2020). *Phân loại đối tượng chiết tách lớp phủ bề mặt tại khu vực công viên địa chất toàn cầu Non nước Cao Bằng dựa trên nền tảng điện toán đám mây.* Tạp chí Khoa học Tài nguyên và Môi trường. 31, 65.

[3] Nguyen B. Luong (2020). *Land cover change detection in northwestern Vietnam using Landsat images and Google Earth Engine.* Journal of Water and Land development. 46, 162.

[4] Breiman L. and R. Ihaka (1984), *Nonlinear discriminant analysis via scaling and ACE.* Department of Statistics, University of California, Technical Report. 40, 1.