ỨNG DỤNG THUẬT TOÁN LIGHTGBM TRONG PHÂN LOẠI LỚP PHỦ HUYỆN ĐẢO LÝ SƠN, VIỆT NAM

VƯƠNG TẤN CÔNG[(1)], PHẠM HOÀNG HẢI[(2)]
[(1)]Học viện Khoa học và Công nghệ, VHL KH&CNVN
[(2)]Viện Địa lý, VHL KH&CNVN

Tóm tắt:

Hệ thống đảo của Việt Nam có phần quan trọng trong việc xây dựng những tiền đồn vững chắc để bảo vệ an ninh chính trị, độc lập chủ quyền của quốc gia trên biển và là thế bàn đạp phát triển kinh tế biển. Tại khu vực miền trung, huyện Đảo Lý Sơn có vai trò quan trọng trong phát triển kinh tế xã hội và đảm bảo an ninh quốc phòng. Nghiên cứu này đề ứng dụng mô hình phân loại LightGBM sử dụng tư liệu ảnh vệ tinh SPOT trong phân loại lớp phủ sử dụng đất. Kết quả cho thấy mặc dù với số lượng mẫu nhỏ, độ chính xác của mô hình (OA = 0,9). Mô hình huấn luyện được sử dụng cho phân loại lớp phủ sử dụng đất tại khu vực nghiên cứu, làm cơ sở cho đánh giá tổng hợp tài nguyên thiên nhiên tại huyện Đảo Lý Sơn.

Từ khóa: LightGBM, Phân loại lớp phủ, Đảo Lý Sơn

1. Giới thiệu

Vùng biển Trung Trung Bộ có 1 cụm đảo Cù Lao Chàm và 2 huyện đảo đã được công nhận là Lý Sơn và Cồn Cỏ (Nguyễn Văn Long 2019). Ở mặt tích cực, khu vực có các điều kiện tự nhiên, tài nguyên thiên nhiên, đặc biệt tài nguyên biển khá phong phú là các điều kiện thuận lợi cho phát triển sản xuất, kinh tế, có vị trí quan trọng như "cửa ngõ" của khu vực Trung Trung Bộ và Trung Bộ nói riêng và của đất nước nói chung trong giao lưu với quốc tế và khu vực, có ý nghĩa quan trọng trong đảm bảo an ninh quốc phòng, phát triển kinh tế biển (Phan Thị Thanh Hằng 2020). Ngoài ra, trừ huyện Đảo Lý Sơn, Phú Quý thì mật độ dân cư trên các đảo nói chung cho đến thời điểm hiện nay không lớn nên tài nguyên nhìn chung còn ít bị khai thác, ảnh hưởng của các quá trình và hiện tượng tự nhiên, môi trường bất lợi chưa lớn. Tuy vậy, nếu xét ở khía cạnh tiêu cực, cùng với quá trình phát triển, trong giai đoạn vừa qua cũng đã thấy nảy sinh một số vấn đề môi trường và suy thoái tài nguyên hết sức cấp bách, đó là trên một số đảo tài nguyên đất đã bị khai thác cạn kiệt, tài nguyên nước khan hiếm, vấn đề sạt lở bờ (ở Lý Sơn), vấn đề ô nhiễm môi trường cục bộ trên một số đảo (Lân 2015; Hải 2010).

Những kết quả nghiên cứu cho thấy sự phát triển của khu vực lãnh thổ này hiện nay còn ở mức thấp, trong các phương án quy hoạch tổng thể phát triển KT-XH của các địa phương nhìn chung còn chưa đánh giá đầy đủ tiềm năng, thế mạnh, chưa tương xứng với vị trí và tầm chiến lược quan trọng trong phát triển KT-XH và đảm bảo an ninh quốc phòng

Ngày nhận bài: 1/5/2023, ngày chuyển phản biện: 5/5/2023, ngày chấp nhận phản biện: 9/5/2023, ngày chấp nhận đăng: 19/5/2023

của các đảo và huyện đảo này. Vấn đề cấp bách đặt ra đối với khu vực lãnh thổ này là cần phải có một chiến lược phát triển tổng thể với những giải pháp khai thác sử dụng hợp lý tài nguyên, bảo vệ môi trường cụ thể. Trong đó bản đồ lớp phủ sử dụng đất đóng vai trò quan trọng làm cơ sở định hướng quy hoạch và đánh giá tiềm năng của khu vực.

Nghiên cứu này thử nghiệm thuật toán LightGBM trong phân loại ảnh vệ tinh SPOT, từ đó làm cơ sở để đánh giá những lợi thế và hạn chế trong phát triển KT-XH huyện đảo, và đánh giá tổng hợp để giải quyết các nhiệm vụ đặt ra cho phát triển của lãnh thổ.

2. Dữ liệu và kết quả

2.1. Dữ liệu nghiên cứu

Huyện đảo Lý Sơn - một trong 10 huyện đảo ven bờ của nước ta, phân bố nằm ở phía Đông tỉnh Quảng Ngãi trên biển Đông, cách đất liền khoảng trên 20 hải lý (38 km), bao gồm 2 đảo với tổng diện tích là 10,7 km^2, trong đó đảo Cù Lao Ré (đảo Lớn) 10 km^2, đảo Cù Lao Bờ Bãi (đảo Bé) 0,7 km^2, nằm cách nhau 4,5 km. Vùng đảo có tọa độ địa lý: 15^{0}32'04" đến 15^{0}38'14" Vĩ độ Bắc và 109^{0}05'04" đến 109^{0}14'12" Kinh độ Đông (Phan Thị Thanh Hằng 2020). Lý Sơn án ngữ con đường ra biển Đông từ khu vực kinh tế trọng điểm miền Trung qua cửa biển nước sâu Dung Quất, bao quát đường giao thông trên biển theo hướng Bắc Nam từ Vịnh Bắc Bộ đi xuống phía Nam và ngược lại. Lý Sơn còn là một điểm trên đường cơ sở, một điểm tựa chiến lược án ngữ phía Đông, đồng thời nằm kề bể dầu Nam Phú Khánh có nhiều triển vọng về dầu khí và sát với ngư trường miền Trung giàu tài nguyên biển.

Hình 1: Ảnh vệ tinh SPOT khu vực nghiên cứu. Dữ liệu hiện tại bị lỗi vạch trắng, và sẽ được hiệu chỉnh trong quá trình xử lý, và phân loại

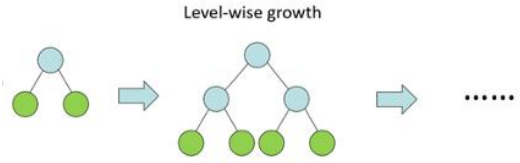
Dữ liệu huấn luyện mô hình được xây dựng thủ công, trích xuất và giải đoán trực tiếp từ ảnh. Mô hình phân loại theo phương pháp phân loại cho từng pixel (pixel-based classification). Số lượng mẫu được khoanh vùng cho 7 loại hình sử dụng đất trên huyện đảo, phục vụ mục đích đánh giá tổng hợp điều kiện tự nhiên kinh tế xã hội tại khu vực nghiên cứu. Bảy loại hình bao gồm: (1) Trảng cây bụi trên đất phong hóa từ bazan, (2) Trảng cỏ thứ sinh trên đất phong hóa từ bazan, (3) Trảng cây bụi trên cát ven biển, (4) Trảng cỏ trên cát ven biển, (5) Rừng trồng, (6) Cây trồng cạn ngắn ngày, (7) Cây trồng quanh khu dân cư. Số mẫu chi tiết được thể hiện:

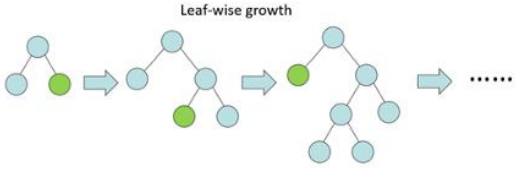
Bảng 1: Thống kê số lượng mẫu dùng để phân loại ảnh

Loài hình lớp phủ	Số lượng pixel huấn luyện	Số lượng pixel kiểm định
Trảng cây bụi trên đất phong hóa từ bazan	454	344
Trảng cỏ thứ sinh trên đất phong hóa từ bazan	616	466
Trảng cây bụi trên cát ven biển	188	143
Trảng cỏ trên cát ven biển	179	135
Rừng trồng	292	221
Cây trồng cạn ngắn ngày	623	472
Cây trồng quanh khu dân cư	1144	866

2.2. Thuật toán LightGBM

LightGBM sử dụng "**histogram-based algorithms**" thay thế cho "pre-sort-based algorithms" thường được dùng trong các boosting tool khác để tìm kiếm split point trong quá trình xây dựng tree. Cải tiến này giúp LightGBM tăng tốc độ training, đồng thời làm giảm bộ nhớ cần sử dụng. Thật ra cả xgboost và lightgbm đều sử dụng histogram-based algorithms, điểm tối ưu của lightgbm so với xgboost là ở 2 thuật toán: **GOSS (Gradient Based One Side Sampling) và EFB (Exclusive Feature Bundling)** giúp tăng tốc đáng kể trong quá trình tính toán.





Cũng giống như các thuật toán Gradient boosting khác, LightGBM có các tham số (hyper-parameters). Các tham số được điều chỉnh thủ công dựa trên tài liệu nghiên cứu trước đây và thử nghiệm dựa trên dữ liệu của nghiên cứu này. Các tham số chính bao gồm:

* num_iterations: Số lượng vòng lặp huấn luyện. Đây là tham số quan trọng để điều chỉnh độ chính xác và tốc độ học của mô hình.

LightGBM phát triển tree dựa trên **leaf-wise**, trong khi hầu hết các boosting tool khác (kể cả xgboost) dựa trên level (depth)-wise. Leaf-wise lựa chọn nút để phát triển cây dựa trên tối ưu toàn bộ tree, trong khi level-wise tối ưu trên nhánh đang xét, do đó, với số node nhỏ, các tree xây dựng từ leaf-wise thường out-perform level-wise. Các thuật toán tích hợp (ensemble), trong đó có LightGBM, đã được sử dụng trong một số nghiên cứu phân loại lớp phủ (Bui et al. 2021; Jun 2021; Jozdani, Johnson, and Chen 2019; Rahman et al. 2020; Liu et al. 2020; Machado, Karray, and Sousa 2019) và được đánh giá đem lại độ chính xác cao.

* learning_rate: Tốc độ học của mô hình, ảnh hưởng đến độ chính xác và tốc độ học của mô hình.

* num_leaves: Số lượng lá của cây quyết định, ảnh hưởng đến độ sâu của cây và khả năng phân loại của mô hình.

* max_depth: Độ sâu tối đa của cây quyết định, ảnh hưởng đến độ phức tạp của mô hình.

* min_data_in_leaf: Số lượng dữ liệu tối thiểu được yêu cầu để một lá có thể được tạo ra, ảnh hưởng đến độ chính xác và overfitting của mô hình.

* feature_fraction: Tỷ lệ số lượng đặc trưng được chọn để sử dụng trong mỗi lần huấn luyện, ảnh hưởng đến độ chính xác và khả năng phân loại của mô hình.

* bagging_fraction: Tỷ lệ dữ liệu được sử dụng trong mỗi lần huấn luyện, ảnh hưởng đến độ chính xác và overfitting của mô hình.

* bagging_freq: Số lần sử dụng bộ dữ liệu được lấy mẫu trong quá trình huấn luyện.

* lambda_l1: Tham số regularization L1, ảnh hưởng đến độ chính xác và overfitting của mô hình.

* lambda_l2: Tham số regularization L2, ảnh hưởng đến độ chính xác và overfitting của mô hình.

* objective: Hàm mục tiêu được sử dụng để tối ưu hóa mô hình, ví dụ như regression, binary classification, multiclass classification,...

* metric: Hàm đo lường hiệu suất được sử dụng để đánh giá độ chính xác của mô hình trong quá trình huấn luyện.

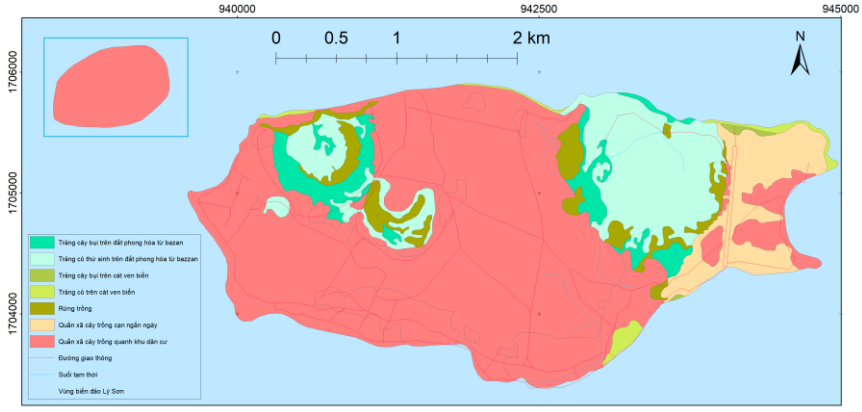
3. Kết quả và thảo luận

Độ chính xác của thuật toán LightGBM trong phân loại ảnh vệ tinh Spot phụ thuộc vào nhiều yếu tố như: số lượng dữ liệu, chất lượng dữ liệu, đặc trưng được sử dụng để huấn luyện và kiểm tra mô hình, cách xử lý dữ liệu, cấu hình mô hình, và phương pháp đánh giá kết quả. Tuy nhiên, LightGBM là một trong những thuật toán Gradient Boosting Decision Tree (GBDT) hiệu quả và phổ biến, được sử dụng rộng rãi trong các bài toán phân loại và dự đoán trên dữ liệu cấu trúc. Nó có thể xử lý được các đặc trưng phức tạp và các bộ dữ liệu lớn. Trong nghiên cứu này, LightGBM đạt được độ chính xác cao trong việc phân loại ảnh vệ tinh Spot, thông qua đánh giá bảng ma trận sai số và độ chính xác tổng thể (overall accuracy). Thông tin chi tiết trong (bảng 2)

Bảng 2: Kết quả đánh giá độ chính xác phân loại sử dụng tập dữ liệu kiểm chứng

	Trảng cây bụi trên đất phong hóa từ bazan	Trảng cỏ thứ sinh trên đất phong hóa từ bazan	Trảng cây bụi trên cát ven biển	Trảng cỏ trên cát ven biển	Rừng trồng	Cây trồng cạn ngắn ngày	Cây trồng quanh khu dân cư	**Tổng số**	**User's accuracy**
Trảng cây bụi trên đất phong hóa từ bazan	319	3	3	6	5	3	5	**344**	**0,93**
Trảng cỏ thứ sinh trên đất phong hóa từ bazan	10	392	45	6	8	3	3	**466**	**0,84**
Trảng cây bụi trên cát ven biển	1	1	124	8	2	3	5	**143**	**0,87**
Trảng cỏ trên cát ven biển	3	5	10	113	3	1	2	**135**	**0,83**
Rừng trồng	1	2	2	5	203	5	4	**221**	**0,92**
Cây trồng cạn ngắn ngày	3	5	3	6	7	435	13	**472**	**0,92**
Cây trồng quanh khu dân cư	5	5	6	7	10	26	806	**866**	**0,93**
Tổng số	**341**	**413**	**192**	**152**	**237**	**475**	**837**		
Producer's accuracy	**0,94**	**0,95**	**0,64**	**0,74**	**0,85**	**0,91**	**0,96**		**0,90**

Dựa trên kết quả đánh giá độ chính xác mô hình từ tập dữ liệu kiểm chứng cho thấy, độ chính xác tổng thể (OA) đạt 90% và độ chính xác PA, UA đều trên ngưỡng 80%, trừ một số trường hợp bị phân loại lẫn giữa trảng cây bụi trên cát ven biển (PA = 64%) và trảng cỏ trên cát (PA = 74%). Mô hình được sử dụng để phân loại cho khu vực nghiên cứu tại huyện Đảo Lý Sơn theo phương pháp Pixel-based (hình 2).



Hình 2: Bản đồ phân loại lớp phủ tại huyện Đảo Lý Sơn

Thuật toán LightGBM nói riêng và các mô hình Gradient boosting đang được sử dụng rộng rãi trong các bài toán phân loại, với cấu trúc dữ liệu đầu vào chủ yếu là dạng bảng (tabulated data). Trong nghiên cứu này, các mẫu pixels được lựa trọn ngẫu nhiên và xây dựng bộ dữ liệu dùng để huấn luyện và kiểm định mô hình, với độ chính xác đạt được 90%. Các tham số mô hình được thử nghiệm dựa trên các nghiên cứu trước, tuy nhiên các tham số có thể được tối ưu tự động dựa trên các thuật toán tối ưu hóa nhóm Bayes, hay các thuật toán meta-heuristic. Các thuật toán này phần nhiều sử dụng Root mean square error (RMSE) làm hàm mục tiêu trong quá trình huấn luyện mô hình. Trong thực tế, hiệu quả của các mô hình phân loại nhiều khi phụ thuộc vào bộ số liệu sử dụng, do đó việc tối ưu hóa tự động các tham số mô hình (thông qua việc học từ dữ liệu) có khả năng tăng thêm độ chính xác phân loại. Việc kết hợp Gradient boosting và các thuật toán tối ưu hóa gợi mở hướng nghiên cứu tiếp theo trong các bài toán phân loại lớp phủ sử dụng đất phục vụ quản lý hiệu quả tài nguyên thiên nhiên tại Việt Nam.

4. Kết luận

Nhóm các thuật toán Gradient boosting có ưu điểm về tốc độ xử lý, và là thuật toán được đánh giá là hiệu quả trong phân loại với dữ liệu dạng bảng. Trong nghiên cứu này, độ chính xác tổng thể đạt được = 0,9 với các thông số UA và PA tương ứng với mỗi lớp loại hình đều trên 0,8. Trừ một việc phân loại nhầm giữa đối tượng trảng cỏ trên các loại đất khác nhau. Mặc dù kích thước mẫu còn nhỏ, tuy nhiên với độ chính xác đạt được, thuật toán này có thể được sử dụng trong phân loại ảnh với hiệu suất tương đối cao. Kết quả bản đồ phân loại được sử dụng trong đánh giá các hoạt động kinh tế xã hội tại khu vực nghiên cứu và đánh giá tổng thể phát triển của huyện Đảo Lý Sơn.❍

Tài liệu tham khảo

[1]. Bui, Quang-Thanh, Tien-Yin Chou, Thanh-Van Hoang, Yao-Min Fang, Ching-Yun Mu, Pi-Hui Huang, Vu-Dong Pham, et al.

2021. "Gradient Boosting Machine and Object-Based CNN for Land Cover Classification." *Remote Sensing* 13 (14). doi: 10.3390/rs13142709.

[2]. Hải, Phạm Hoàng. 2010. *Các huyện đảo ven bờ Việt Nam tiềm năng và định hướng phát triển*. Hà Nội: Nxb. KHTNCN.

[3]. Jozdani, Shahab E., Brian A. Johnson, and Dongmei Chen. 2019. "Comparing Deep Neural Networks, Ensemble Classifiers, and Support Vector Machine Algorithms for Object-Based Urban Land Use/Land Cover Classification." *Remote Sensing* 11 (14). doi: 10.3390/rs11141713.

[4]. Jun, Myung-Jin. 2021. "A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area." *International Journal of Geographical Information Science*:1-19. doi: 10.1080/13658816.2021.1887490.

[5]. Lân, Trần Đình. 2015. "Lượng giá kinh tế các hệ sinh thái biển - đảo tiêu biểu phục vụ phát triển bền vững một số đảo tiền tiêu ở vùng biển ven bờ Việt Nam". In *Đề tài cấp Nhà nước mã số KC.09.08/11-15*. Hà Nội: Bộ Khoa học và Công nghệ.

[6]. Liu, H., P. Gong, J. Wang, N. Clinton, Y. Bai, and S. Liang. 2020. "Annual dynamics of global land cover and its long-term changes from 1982 to 2015". *Earth Syst. Sci. Data* 12 (2):1217-43. doi: 10.5194/essd-12-1217-2020.

[7]. Machado, M. R., S. Karray, and I. T. de Sousa. 2019. LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry. Paper presented at the 2019 14th International Conference on Computer Science & Education (ICCSE), 19-21 Aug. 2019.

[8]. Nguyễn Văn Long, Tống Phước Hoàng Sơn. Hội thảo 10 năm bảo tồn và phát triển (2009-2019). 2019. "Diễn thế các hệ sinh thái quan trọng ở Khu dự trữ sinh quyển thế giới Cù Lao Chàm - Hội An". In *Tuyển tập báo cáo Khu dự trữ sinh quyển thế giới Cù Lao Chàm - Hội An*. Hội An.

[9]. Phan Thị Thanh Hằng, Mã số: KC.09.37. Viện Địa lý, Viện HLKH&CNVN. 2020. "Cơ sở khoa học, định hướng và giải pháp phát triển kinh tế - xã hội phát triển bền vững các huyện đảo Lý Sơn và Phú Quý". In, edited by đề tài Báo cáo, Mã số: KC.09.37 Hà nội: Viện Địa lý, Viện HLKH&CNVN.

[10]. Rahman, Saifur, Muhammad Irfan, Mohsin Raza, Khawaja Moyeezullah Ghori, Shumayla Yaqoob, and Muhammad Awais. 2020. "Performance Analysis of Boosting Classifiers in Recognizing Activities of Daily Living". *International Journal of Environmental Research and Public Health* 17 (3). doi: 10.3390/ijerph17031082.❍

Summary

Application of LightGBM in classifying the landcover of Ly Son Island, Vietnam

Vuong Tan Cong

Academy of Science and Technology, Vietnam Academy of Science and Technology

Pham Hoang Hai

Institute of Geography, Vietnam Academy of Science and Technology

Vietnam's island system is essential in building solid outposts to protect the country's political security, independence, and sovereignty at sea and is a springboard for marine economic development. In the central region, the Ly Son island district plays a vital role in socio-economic development and ensures national security and defence. This study aims to apply the LightGBM classification model using SPOT satellite image data in land use cover classification. Results for the teacher, although with a small number of samples, the accuracy of the model (OA = 0.9). The training model is used for land use cover classification in the study area as a basis for the integrated assessment of natural resources in the Ly Son island district.❍

Keywords: LightGBM, Land cover classification, Ly Son island

ĐÁNH GIÁ TÁC ĐỘNG CỦA DỰ ÁN ĐẦU TƯ…..

(Tiếp theo trang 41)

Summary

Impact assessment of investment projects on biosphere reserves using machine learning algorithms and landscape metrics

Do Thi Nhung, Pham Van Manh

University of Science, Vietnam National University, Hanoi

Pham Anh Cuong, Institute of Natural Resources and Environment Development

Truong Quang Hai, Giang Van Trong

Institute of Vietnamese Studies and Development Science, Vietnam National University, Hanoi

Pham Hanh Nguyen, Ngo Xuan Quy

Department of Natural conservation and Biodiversity

Socio-economic development is one of the most influential factors to land use change that affects the living environment and threatens the landscape metrics. The managers in conservation planning need methods that can predict impacts early in the planning. This study investigates and selects landscape metrics that planners can use to assess the potential impact of habitat changes, fragmentation, and ecological connectivity resulting from intended land use changes. Unlike previous studies, this study proposes the Overall Landscape Fragmentation Index (OLFI) to assess landscape fragmentation and analyse the spatial and temporal heterogeneity in Nui Chua National Park - World Biosphere Reserve. The study used a machine learning algorithm to classify land cover/land use (LULC) with an overall accuracy of 92.84% and a Kappa coefficient of 0.90. The findings show that when there is a tourism development project, the level of impact increases significantly on the landscape structure of the Nui Chua National Park. The OLFI was developed as a new index to quantify the impact of investments on the natural landscape and to serve as a reference for conservation purposes and land use planning in other similar areas in Vietnam and different countries.❍

Keywords: Land-use/land-cover, Machine learning, Landscape fragmentation, Nui Chua National Park.