NGHIÊN CỨU ỨNG DỤNG PHƯƠNG PHÁP HỌC MÁY (MACHINE LEARNING) TRONG PHÂN LOẠI RỪNG NGẬP MẶN TRÊN ẢNH VIỄN THÁM SPOT6 VỚI KHU VỰC THỬ NGHIỆM TẠI TỈNH CÀ MAU

PHẠM MINH HẢI⁽¹⁾, VŨ KỲ LONG⁽²⁾

⁽¹⁾Viện Khoa học Đo đạc và Bản đồ

⁽²⁾Sở Tài nguyên Môi trường Hà Nội

Tóm tắt:

Công tác lựa chọn thuật toán phù hợp là một phần quan trọng của bất kỳ ứng dụng phương pháp máy học nào và bởi vì có rất nhiều các tiêu chí và sự lựa chọn. Bên cạnh đó, sự am hiểu về điểm mạnh và điểm yếu của các thuật toán trong phương pháp học máy là điều cần thiết để đem lại hiệu quả cao. Trong phạm vi bài báo này, nhóm nghiên cứu tiến hành thử nghiệm một thuật toán của phương pháp học máy (Machine Learning) trong phân loại rừng ngập mặn trên ảnh viễn thám SPOT6 với khu vực thử nghiệm tại Tỉnh Cà Mau. Kết quả thực hiện của nghiên cứu đã đạt được 2 điểm mới đó là: ứng dụng thành công phương pháp học máy (Machine Learning) trong phân loại ảnh viễn thám, và phân loại được chi tiết theo loài thực vật của rừng ngập mặn tại khu vực thử nghiệm. Bên cạnh đó, tác giả đã khai thác được khả năng nhận diện pattern trên ảnh viễn thám của phương pháp và quá trình phân loại ảnh dựa trên các mẫu được lựa chọn đem lại kết quả có độ chính xác cao.

1. Giới thiệu chung về phương pháp học máy (Machine Learning)

Học máy cho phép con người học dữ liệu xác định chính xác các cơ hội và tạo chiến lược để cải thiện các trải nghiệm của bản thân bằng cách sử dụng thông tin ẩn trong các tập dữ liệu khổng lồ. Công tác lựa chọn thuật toán phù hợp là một phần quan trọng của bất kỳ ứng dụng phương pháp máy học nào và bởi vì có rất nhiều các tiêu chí và sự lựa chọn. Bên cạnh đó, sự am hiểu về điểm mạnh và điểm yếu của các thuật toán trong phương pháp học máy là điều cần thiết để đem lại hiệu quả cao. Trong phạm vi bài báo này, nhóm nghiên cứu tiến hành thử nghiệm một thuật toán của phương pháp học máy (Machine Learning) trong phân loại rừng ngập mặn trên ảnh viễn thám SPOT6 với khu vực thử nghiệm tại tỉnh Cà Mau.

Như đã giới thiệu về khái niệm phân tích chi tiết về thuật toán Random Forest (rừng ngẫu nhiên) trong số báo 39 tháng 3/2019 của Tạp chí Khoa học Đo đạc và Bản đồ. Đây là thuật toán phân lớp thuộc tính được phát triển bởi Leo

Breiman tại đại học California, Berkeley. Breiman cũng đồng thời là đồng tác giả của phương pháp CART (Classification and Regression Trees) được đánh giá là một trong 10 phương pháp khai phá dữ liệu kinh điển. Random Forest được xây dựng dựa trên 3 thành phần chính là: (1) CART, (2) học toàn bộ, hội đồng các chuyên gia, kết hợp các mô hình, và (3) tổng hợp bootstrap (bagging). Hình 1 dưới đây thể hiện phương pháp phân lớp random forest.

Hình 1: Sơ đồ biểu diễn các cây quyết định trong phương pháp random forest.

(<https://analyticsdefined.com/introduction-random-forests/>)

Ngày nhận bài: 06/5/2019, ngày chuyển phản biện: 09/5/2019, ngày chấp nhận phản biện: 15/5/2019, ngày chấp nhận đăng: 20/5/2019

Random forest được ứng dụng phục vụ các mục đích phân loại, tính hồi quy và các nhiệm vụ khác bằng cách xây dựng nhiều cây quyết định (Decision tree). Một cây quyết định là một cách đơn giản để biểu diễn một giao thức (Protocol). Nói cách khác, cây quyết định biểu diễn một kế hoạch, trả lời câu hỏi phải làm gì trong một hoàn cảnh nhất định. Mỗi Node của cây sẽ là các thuộc tính, và các nhánh là giá trị lựa chọn của thuộc tính đó. Bằng cách đi theo các giá trị thuộc tính trên cây, cây quyết định sẽ cho ta biết giá trị dự đoán.

2. Khu vực thực nghiệm và dữ liệu đầu vào

2.1. Khu vực thực nghiệm

Cà Mau là một tỉnh ven biển ở cực nam của Việt Nam, nằm trong khu vực Đồng bằng sông Cửu Long. Phần lãnh thổ đất liền của tỉnh Cà Mau nằm trong tọa độ từ 8°30' - 9°10' vĩ Bắc và 104°80' - 105°5' kinh Đông. Điểm cực Đông tại 105°24' kinh Đông thuộc xã Tân Thuận, huyện Đầm Dơi. Điểm cực Nam tại 8°33' vĩ Bắc thuộc xã Viên An, huyện Ngọc Hiển. Điểm cực Tây tại 104°43' kinh Đông thuộc xã Đất Mũi, huyện Ngọc Hiển. Điểm cực Bắc tại 9°33' vĩ Bắc thuộc xã Biển Bạch, huyện Thới Bình.

Hình 2: Phạm vi khu vực thực nghiệm

(<http://www.camau.gov.vn/wps/portal/bando>)

2.2. Dữ liệu đầu vào và công tác tiền xử lý ảnh

Bài báo thực hiện với dữ liệu ảnh vệ tinh

SPOT6 được chụp tại thời điểm năm 2015. Dữ liệu được cung cấp bởi Đài thu Viễn thám Quốc gia, Cục Viễn thám Quốc gia, Bộ Tài nguyên và Môi trường. Ảnh vệ tinh có độ phủ mây nhỏ hơn 10%, và được hiệu chỉnh bức xạ và khí quyển bằng công cụ ATCOR (Atmospheric correction) trong phần mềm PCI Geomatic 2018 (phiên bản dùng thử). Quá trình tiền xử lý ảnh gồm 3 bước: Hiệu chỉnh bức xạ trước đầu thu; hiệu chỉnh bức xạ ảnh hưởng mây; hiệu chỉnh ảnh hưởng khí quyển. Nhóm thực hiện nghiên cứu tiến hành công tác nắn chỉnh hình học và trộn ảnh về độ phân giải 2.5m. Ảnh được nắn về hệ tọa độ WGS84 và hệ quy chiếu UTM múi 48. Để đánh giá độ chính xác của kết quả thực hiện, nhóm thực hiện sử dụng bản đồ kiểm kê rừng tại cùng khu vực nghiên cứu được thành lập năm 2015.

3. Phân loại ảnh vệ tinh sử dụng thuật toán Random Forest

3.1. Công tác phân loại ảnh vệ tinh

Nhóm thực hiện đã tiến hành thu thập mẫu giải đoán ảnh tại khu vực nghiên cứu. Công tác thu thập mẫu được tiến hành cả bằng phương pháp thực địa và trên phần mềm Google Earth. Hệ thống chú giải được xây dựng thể hiện trên bảng 2.

Bảng 1: Hệ thống chú giải được xây dựng

STT	Mã	Chú giải
1	RNMG	Rừng ngập mặn giàu
2	RNMTB	Rừng ngập mặn trung bình
3	RNMPH	Rừng ngập mặn phục hồi
4	DTR	Đất trống, bãi cát
5	MN	Mặt nước
6	DNN	Đất NN, NTTS

Trong đó, tập dữ liệu sau khi giải đoán được chia thành 2 phần với số lượng điểm và bố trí các điểm thể hiện trên hình 3.

70% mẫu được sử dụng mẫu huấn luyện để phân loại (training data)

30% mẫu còn lại được sử dụng mẫu kiểm

chứng kết quả phân loại (testing data)

Hình 3: Phân bố điểm mẫu để phân loại và kiểm chứng

Bên cạnh đó, tác giả còn tiến hành điều tra về thành phần loài khu vực nghiên cứu. Theo kết quả của kiểm kê rừng của Bộ NN&PTNT về kiểm kê rừng tỉnh Cà Mau năm 2015 có 3 loài rừng ngập mặn chính: loài đước: chiếm trên 93,7%, loài mắm (mắm): 5%, hỗn giao giữa đước + mắm: 1% và còn lại là các loài khác.

Bảng 2: Thành phần loài cây ngập mặn tại Tỉnh Cà Mau

Loài cây	Diện tích (ha)	Tỷ lệ%
Đước	42,723	93.7%
Dừa	276	0.6%
Dừa nước	134	0.3%
Giá	39	0.1%
Mắm	2,409	5.3%
Vẹt	24	0.1%
Chà là	6	0.0%
Tổng	45,611	

Sau đó, Phương pháp Multiresolution segmentation trên phần mềm xử lý ảnh Ecognition đã được sử dụng để phân đoạn.

3.2. Phân loại ảnh vệ tinh sử dụng thuật toán Random Forest

Sử dụng thuật toán Random Forest và dữ liệu ảnh SPOT6, nhóm thực hiện nghiên cứu đã xây dựng được hiện trạng rừng ngập mặn khu vực thực nghiệm với các đối tượng cây ngập mặn được chia ra thành các loại như: mắm, đước, hỗn giao.v.v... Kết hợp hệ thống chú giải được xây dựng tại phần 3.1 cho khu vực thực nghiệm, nhóm thực hiện đã xây dựng được hiện trạng trù

lượng rừng ngập mặn theo loài tại khu vực Cà Mau thể hiện trên hình 4.

Hình 4: Kết quả phân loại ảnh sử dụng thuật toán Random Forest

3.3. Đánh giá độ chính xác kết quả phân loại ảnh

3.3.1. Đánh giá định tính

Nhằm đánh giá độ chính xác của kết quả phân loại ảnh, tác giả tiến hành so sánh bằng mắt thường về ranh giới khu vực rừng ngập mặn giữa kết quả phân loại ảnh và bản đồ kiểm kê rừng cùng thời kỳ. Kết quả cho thấy, kết quả phân loại ảnh cho độ chính xác cao hơn với mức độ chi tiết về các đường ranh giới khu vực rừng ngập mặn cao. Ngược lại, nhiều khu vực có các đường ranh giới khu vực rừng ngập mặn trên bản đồ kiểm kê rừng cùng thời kỳ đã bị khái quát hóa cao khi thành lập do đó phản ánh không chính xác vị trí ranh giới khu vực rừng ngập mặn. Kết quả so sánh được thể hiện trên hình 5. (Xem hình 5)

3.3.2. Đánh giá định lượng

Kết quả đánh giá độ chính xác khi kiểm chứng với 109 điểm kiểm chứng như sau: Độ chính xác tổng thể OA đạt 83%, trong đó kết quả phân loại rừng đước cho độ chính xác cao cả PA và OA, hỗn giao đước, mắm có OA chỉ đạt 74%, trong khi đó mắm có PA đạt 68 %do vẫn chưa tách được các lớp khác bị lẫn vào trong 1 điểm ảnh. (Xem bảng 3)

Kết luận

Nghiên cứu có cơ sở khoa học và thực tiễn

Bản đồ kiểm kê rừng 2015

Kết quả phân loại ảnh SPOT6 2015

Hình 5: Minh họa so sánh đánh giá độ chính xác giữa kết quả phân loại ảnh và bản đồ kiểm kê rừng cùng thời kỳ

Bảng 3: Ma trận đánh giá độ chính xác kết quả phân loại ảnh

Loại	Được	Hỗn giao được, mầm	Mầm	Tổng	PA	UA	OA
Được	52	5	1	58	90%	84%	83%
Hỗn giao được, mầm	6	23		29	79%	74%	
Mầm	4	3	15	22	68%	94%	
Tổng	62	31	16	109			

khi đã ứng dụng thử nghiệm thành công thuật toán Random Forest của phương pháp học máy (Machine Learning) trong phân loại ảnh vệ tinh. Kết quả thực hiện của nghiên cứu đã đạt được 2 điểm mới đó là: ứng dụng thành công phương pháp học máy (Machine Learning) trong phân loại ảnh viễn thám, và phân loại được chi tiết theo loài thực vật của rừng ngập mặn tại khu vực thực nghiệm. Bên cạnh đó, nhóm thực hiện đã khai thác được khả năng nhận diện pattern trên ảnh viễn thám của phương pháp và quá trình phân loại ảnh dựa trên các mẫu được lựa chọn đem lại kết quả có độ chính xác cao. Nghiên cứu được thực hiện và đã đạt được những kết quả nhất định, tuy nhiên nhóm nghiên cứu có một số kiến nghị cần nghiên cứu tiếp để có những kết quả có độ chính xác cao hơn. Đó là: 1). Sử dụng ảnh độ phân giải cao (ảnh quang học) để có thể phản ánh cấu trúc, kiến trúc ảnh liên quan đến thành phần loài ngập mặn; 2). Kết hợp các ảnh RADAR để có thể phản ánh cấu trúc đứng của rừng ngập mặn; 3). Sử dụng các yếu tố kiến trúc ảnh (image texture) và các chỉ số hình dạng (shape index) trong phân loại kiểu rừng; 4). Thiết kế các điểm khảo sát, OTC để xác định

loài, cấu trúc rừng, trữ lượng rừng để phân loại, kiểm chứng từ ảnh vệ tinh.○

Tài liệu tham khảo

- [1]. Apampa., P. (2016). “Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction”, Journal of International Technology and Information Management.
- [2]. Khalilia., M (2011). “Predicting disease risks from highly imbalanced data using random forest”, BMC Medical Informatics and Decision Making, 2011.
- [3]. Phạm Minh Hải (2016). “Nghiên cứu đề xuất giải pháp nâng cao độ chính xác của công tác phân loại ảnh khu vực có lớp phủ hỗn hợp-Cơ sở khoa học”, Tạp chí Khoa học Đo đạc và Bản đồ, Số 29-9/2016.
- [4]. Phạm Minh Hải, Nguyễn Ngọc Quang (2019). Khái niệm về phương pháp Random Forest trong cuộc cách mạng machine learning và định hướng sử dụng trong lĩnh vực viễn thám, Tạp chí Khoa học Đo đạc và Bản đồ, Số 39-3/2019.○

Summary

Application of Machine Learning methods in SPOT6 image satellite classification with the study area in the mangrove forest of Ca Mau Province

Pham Minh Hai

Institute of Geodesy and Cartography

The selection of suitable algorithms plays an important role in any application of machine learning methods because of their many criteria. Also, the understanding of the strengths and weaknesses of algorithms in machine learning methods is essential to bring high efficiency. Within the scope of this manuscript, the team conducted an algorithm of Machine Learning methods called Random Forest in using the SPOT6 remote sensing for the mangrove forest classification with the test area in Ca Mau Province. The results of the study have achieved two new points: successfully applying Machine Learning method in remote sensing image classification, and classification of detail species of mangrove forests in the study area. In addition, the team has exploited the potential of machine learning methods to identify patterns on remote sensing images based on the selected samples in order to extract thematic information in high accuracy. ○

ỨNG DỤNG WEBGIS TRONG GIÁM SÁT QUẢN LÝ.....

(Tiếp theo trang 16)

[2]. Quốc hội, 2013. Luật Đất đai số 45/2013/QH13.

[3]. Đặng Hùng Võ, Nguyễn Văn Thắng, 2014. Khung đánh giá quản trị đất (LGAF) và kết quả nghiên cứu ở Việt Nam. Tạp chí Kinh tế & phát triển, số 201 tháng 4/2014.

[4]. Lê Xuân Ánh, Nguyễn Đình Thông, Nguyễn Thị Thanh Tâm. 2016. Ứng dụng công nghệ WebGIS phục vụ canh tác lúa vùng đồng bằng sông Hồng. Hội thảo quốc gia về khoa học cây trồng lần thứ hai.

[5]. <http://papi.org.vn/hoi-dap>. ○

Summary

WebGIS application in supervision of management and use of land resources

Nguyen Phi Son, Institute of Geodesy and Cartography

Increasing public participation in the field of public administration has become a world trend. In land monitoring activities, people's participation will enhance the efficiency of land management and use. There are many ways for people to participate in monitoring activities, through which the Websites function of the geographic information system (WebGIS) will increase the convenience and quality of information. Not only is WebGIS conducive to raising public awareness about land law policies, citizens' rights and obligations towards national assets and resources, but it also contributes to promoting similarities. positive cooperation between public and private rights of citizens. WebGIS system is built on the basis of ASP.net programming language, spatial database is put into SQL Server 2016 management system and C# programming language library in ASP.NET environment is used. to design the website. Based on the Web interface, users can obtain information about land resources and local land use. The main function of the Web is the information channel for people and communities to provide the management agency with initial information on violations in land management and land use in the locality. The webmaster will receive the information, will synthesize and submit it to the competent authorities for resolution. This is an effective tool to support the state management of land.

The paper is built based on data and research results of the state level project under the climate change program 16-20, climate change code.10/16-20. ○