KHÁI NIỆM VỀ PHƯƠNG PHÁP RANDOM FOREST TRONG CUỘC CÁCH MẠNG MACHINE LEARNING VÀ ĐỊNH HƯỚNG ỨNG DỤNG TRONG LĨNH VỰC VIỄN THÁM

PHẠM MINH HẢI[1], NGUYỄN NGỌC QUANG[2]

[1]Viện Khoa học Đo đạc và Bản đồ, [2]Đài Viễn thám Trung ương

Tóm tắt

Random forest là một phương pháp thống kê mô hình hóa bằng máy (machine learning statistic) dùng để 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).Random Forest cho thấy hiệu quả hơn so với thuật toán phân loại thường được sử dụng vì có khả năng tìm ra thuộc tính nào quan trọng hơn so với những thuộc tính khác.Trên thực tế, nó còn có thể chỉ ra rằng một số thuộc tính là không có tác dụng trong cây quyết định. Trong phạm vi bài báo này, nhóm nghiên cứu giới hạn phạm vi trong công tác khảo sát tính khoa học của phương pháp và định hướng việc ứng dụng phương pháp cho công tác phân loại ảnh viễn thám có kiểm định. Kết quả thử nghiệm cho thấy khả năng ứng dụng phương pháp Random forest vào trong công tác phân loại có kiểm định ảnh viễn thám là hoàn toàn khả thi.

1. Giới thiệu chung

Để chiết tách các thông tin ảnh viễn thám, việc ứng dụng các thuật toán có kiểm định như K-Nearest Neighbors (KNN) đã trở nên phổ biến. K-Nearest Neighbors phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần xếp lớp (Query point) và tất cả các đối tượng trong các bộ mẫu (Training Data). Tuy nhiên hiện nay, các nhà nghiên cứu đã và đang phát triển nhiều thuật toán mới, phức tạp, mạnh mẽ và hiệu quả hơn. Một trong những phương pháp đó là Random Forest. Đây là một một cuộc cách mạng trong công nghệ mô hình hóa bằng máy (Machine Learning). Random Forest chỉ phức tạp hơn một chút so với k-nearest neighbors, nhưng nó hiệu quả hơn nếu xét trên hiệu năng tính toán của máy tính cho kết quả chính xác hơn so với k-nearest neighbors.

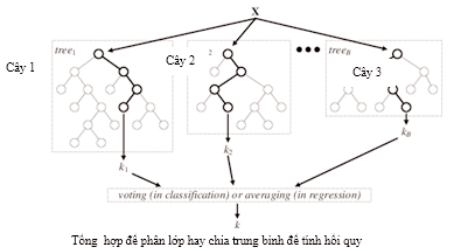
2. Khái niệm phương pháp

2.1. Định nghĩa

Random forest là một phương pháp thống kê mô hình hóa bằng máy (machine learning statistic) dùng để 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. Nhóm thuật toán cây quyết định có một điểm mạnh đó là có thể sử dụng cho cả bài toán Phân loại (Classification) và Hồi quy (Regression). Random Forest có khả năng tìm ra thuộc tính nào quan trọng hơn so với những thuộc tính khác. Trên thực tế, nó còn có thể chỉ ra rằng một số thuộc tính là không có tác dụng trong cây quyết định. *(Xem hình 1)*

Từ hình 1 chúng ta thấy rằng Random Forest được cấu thành bởi một số cây quyết định. Các cây này cùng nhận đầu vào là đối tượng x và đưa ra quyết định về danh mục thuộc tính (Attribute category) của x. Các quyết định này sẽ được tổng hợp lại lấy trung bình để chọn ra quyết định cuối cùng.

Ngày nhận bài: 01/2/2019, ngày chuyển phản biện: 12/2/2019, ngày chấp nhận phản biện: 20/2/2019, ngày chấp nhận đăng: 28/2/2019



Tổng hợp để phân lớp hay chia trung bình để tính hồi quy

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

2.2. Mô tả phương pháp random forest

2.2.1. Lựa chọn cây quyết định (decision tree learning)

Cây quyết định là một phương pháp phổ biến cho các nhiệm vụ mô hình hóa bằng máy (machine learning). Các cây quyết định được lựa chọn với các tiêu chí phù hợp để đáp ứng các yêu cầu nhiệm vụ phục vụ khai thác dữ liệu. Các cây quyết định được thiết kế với xu hướng nhận biết được cả những yếu tố bất thường: phù hợp với các mẫu có độ lệch nhỏ nhưng phương sai lớn.

2.2.2. Thuật toán mô hình bằng máy

Thuật toán lấy mẫu cho phương pháp random forest ứng dụng cho các phương pháp sử dụng thuật toán mô tả thống kê để ước lượng số lượng từ một mẫu dữ liệu (bagging). Ví dụ như một một tập mẫu X = x_1, ..., x_n với các câu trả lời Y = y_1, ..., y_n, lấy giá trị trung bình (B lần), chọn một mẫu ngẫu nhiên từ bộ mẫu phù hợp với cây quyết định:

Lặp *b = 1,..., B:*

n mẫu từ giá trị tọa độ (X, Y); gọi là (X_b, Y_b);

lớp dữ liệu hay kết quả hồi quy f_b của biến X_b, Y_b;

Sau khi lấy mẫu, các phép tính toán cho các mẫu là ẩn số x' có thể được thực hiện bằng cách lấy trung bình các giá trị nội suy từ tất cả các cây hồi quy riêng lẻ của biến x' hoặc lấy giá trị từ đa số của các mẫu trong cây quyết định:

$$f = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Phương pháp thống kê này ước lượng một giá trị trung bình từ số lượng mẫu dữ liệu. Chúng ta cần rất nhiều mẫu từ tập dữ liệu, tính giá trị trung bình. Sau đó, tính trung bình tất cả các giá trị trung bình của các tập dữ liệu trong các cây quyết định thành phần để tính toán được tốt hơn giá trị trung bình thật. Kết quả dẫn đến hiệu suất mô hình được tính toán sẽ tốt hơn vì nó làm giảm phương sai của mô hình, mà không làm tăng độ lệch. Điều này có nghĩa là khi thiết kế nhiều cây quyết định trong một tập các mẫu được lấy sẽ đưa ra sự tương quan tốt hơn của các cây quyết định với nhau.

2.2.3. Từ thuật toán mô hình bằng máy đến Random forest

Các bước 2.2.1 và 2.2.2 đã mô tả cách thực hiện thuật toán thống kê để ước lượng giá trị trung bình từ số lượng các cây quyết định của tập mẫu dữ liệu (bagging). Phương pháp random forest khác cơ bản so với phương pháp thống kê trên là chúng sử dụng thuật toán xử lý theo các cây quyết định (tree learning algorithm). Tại mỗi phần tử ở trong quy trình này được gán ngẫu nhiên các tập con thuộc tính của các mẫu. Lý do thực hiện quy trình này là sự tương quan của các cây quyết định thành phần trong một thuật toán thống kê để ước lượng giá trị trung bình từ số lượng các cây quyết định thông thường: nếu một hoặc một vài thuộc tínhlà các yếu tố dự báo mạnh cho biến đầu ra, các tính năng này sẽ được chọn trong nhiều cây B, chúng sẽ trở nên tương quan.

Random forest có thể sắp xếp sự quan trọng của các biến trong các bài toán phân loại hay hồi quy. Các phương pháp sắp xếp có thể được mô tả trong các nghiên cứu của Breiman. Bước đầu tiên để xác định các biến quan trọng trong 1 tập dữ liệu là làm phù hợp phương pháp random forest với tập dữ liệu:

$$D_n = \{(X_n, Y_n)\}^n \text{với } i = 1$$

Trong quá trình này, các lỗi dự báo xảy ra (out-of-bag error) tại mỗi điểm xử lý được ghi lại và tính giá trị trung bình. Để xác định được tính

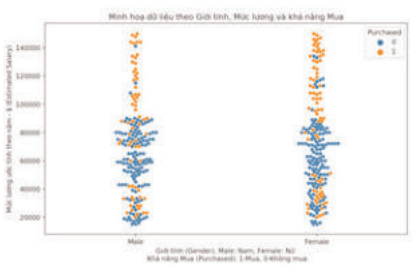
quan trọng của đối tượng thứ i sau khi lấy mẫu, các giá trị của mẫu i được hoán vị trong tập mẫu và các lỗi dự báo được tính toán lại trong tập dữ liệu. Độ quan trọng của đối tượng được tính bằng điểm, các điểm được tính toán bằng cách lấy trung bình của độ chênh lệch giữa các lỗi dự báo trước và sau khi hoán vị. Các đối tượng có giá trị lớn được xếp quan trọng hơn các điểm có giá trị nhỏ.

3. Giới thiệu một số ứng dụng

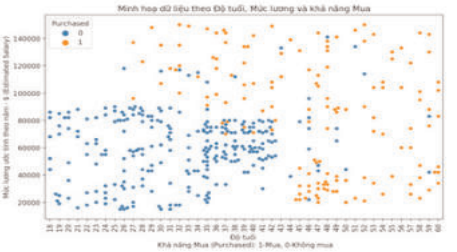
Như đã trình bày ở trên, phương pháp random forest có thể ứng dụng cả trong phân loại lẫn hồi quy, có thể thực thi với số lượng lớn các đặc trưng của đối tượng và thật sự hữu ích trong việc dự báo đánh giá các biến được xem là quan trong trong tập dữ liệu đưa vào mô hình. Dưới đây là một thử nghiệm thực tiễn, ứng dụng phương pháp phân loại KNN và random forest. Bài toán tìm và phát hiện các nhóm khách hàng tiềm năng sử dụng xe ô tô để giúp cho việc tương mại hóa sản phẩm được tập trung vào đúng đối tượng có nhu cầu nhằm tăng tối đa hiệu quả bán hàng và giảm thiểu lãng phí về mặt chi phí cũng như thời gian dành cho việc quảng cáo. Để thực hiện, tác giả sử dụng thư viện Scikit-Learn (scikit-learn.org) và ngôn ngữ lập trình Python.

Về dữ liệu

Thử nghiệm này sử dụng dữ liệu từ nguồn *superdatascience.com*, đây là dữ liệu thống kê (dạng CSV) thu thập thông tin của khách hàng nhằm quảng cáo bán xe đa dụng. Dữ liệu chứa các thông tin về ID (User ID), Giới tính (Gender), Độ tuổi (Age), Mức lương ước tính theo năm (EstimatedSalary) và khả năng Mua (Purchased) của 400 người ở Mỹ.



Hình 2: Minh họa về dữ liệu khách hàng được phân chia theo giới tính



Hình 3: Minh họa về dữ liệu khách hàng được phân chia khả năng mua hàng

Phương pháp phân loại

Phân loại theo KNN và random forest được sử dụng từ thư viện Scikit-Learn thông qua ngôn ngữ lập trình Python. Với KNN dữ liệu được đưa vào mô hình với 03 trường thông tin về Độ tuổi, Mức lương, và Khả năng mua.

Trong đó tệp dữ liệu mẫu (Training data) và dữ liệu kiểm tra (Test data) được lựa chọn với tỉ lệ với thứ tự lần lượt là 75:25, nghĩa là sẽ có 300 cho dữ liệu mẫu và 100 cho dữ liệu kiểm tra. Với trường thông tin về độ tuổi và mức thu nhập có sự chênh lệch về mặt giá trị quá lớn nên sẽ phải quy đổi giá trị theo tỉ lệ phù hợp thông qua hàm StandardScaler.

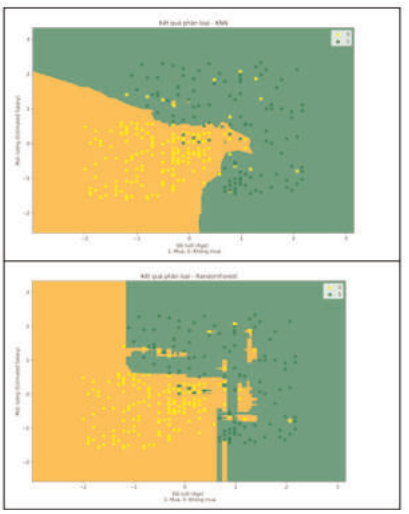
Phương pháp phân loại KNN sử dụng mô hình KNeighborsClassifier với tham số n_neighbors =5, kiểu "Minkowski" với p=2 phù hợp với trị đo Euclidean tiêu chuẩn).

Về phương pháp phân loại random forest, khâu chuẩn bị và tiền xử lý dữ liệu giống phương pháp KNN, giá trị dữ liệu đều được quy đổi theo tỉ lệ tiêu chuẩn. Ở đây, phương pháp sử dụng hàm đo chất lượng chia nhánh cây quyết định là "Entropy" nhằm tăng lượng thông tin chính xác và lựa chọn số cây tham gia chạy mô hình là 50 cây, ở đó mỗi lớp được phân ra sẽ có sự tổng hợp từ 50 cây để chọn ra lớp chiếm đa số là kết quả cuối cùng.

Kết quả và bàn luận

Từ phân tích tập dữ liệu này ở hình 4 có thể thấy có một xu thế mua xe ô tô ở độ tuổi 45 trở ra với thu nhập trải dài từ cận dưới (~20,000$) đến cận trên (~140,000$), có thể vì đây là đối

tượng đã có gia đình và cả nhà thường có những chuyến đi xa,...nên lựa chọn dòng xe tiện lợi phù hợp, và một xu hướng khác là giới trẻ ở độ tuổi từ 28 có thu nhập từ cận giữa trở lên (~70,000$) có thể chọn những dòng xe ô tô đắt. Để việc công tác thương mại hóa sản phẩm tập trung đúng đối tượng và tăng hiệu quả bán hàng cần phải phân loại dữ liệu có độ chính xác cao. Dưới đây là kết quả phân loại theo KNN và random forest.



Hình 4: Kết quả phân loại sức mua ô tô theo độ tuổi và mức lương được thực hiện bởi phương pháp random forest

Ở thử nghiệm này có thể thấy rõ kết quả ưu việt mà phương pháp random forest mang lại với việc phân loại gần như chính xác tuyệt đối đối với lớp không mua với sai số 5%. Có thể thấy so với KNN thì phương pháp random forest mang lại kết quả tương đối ấn tượng và rõ ràng.

Với đặc điểm sử dụng các cây quyết định với nhiều mẫu được lựa chọn, giá trị cuối cùng được đưa ra sau khi xem xét giá trị trung bình của các giá trị trung bình trong các cây quyết định thành phần đã tạo ra sản phẩm phân loại có độ chính xác cao.

Với khả năng thực tếmà phương pháp phân loại học máy nói chungmang lại, việc nghiên cứu thử nghiệm ứng dụng phương pháp random forest trong lĩnh vực viễn thám mà cụ thể là phân loại ảnh là hoàn toàn có thể áp dụng được. Công việc này cần được quan tâm và thúc đẩy triển khai thực tế nhanh để nâng cao hiệu quả trong việc phân tách chính xác các đối tượng trên ảnh viễn thám nhằm tạo ra nhiều thông tin cũng như sản phẩm giá trị gia tăng có ý nghĩa với kinh tế-xã hội.

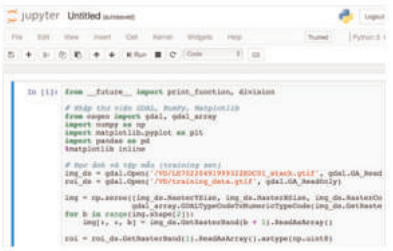
4. Định hướng ứng dụng trong lĩnh vực viễn thám

Để định hướng sử dụng thuật toán phục vụ công tác phân loại học máy (machine learning) cho ảnh viễn thám, nhóm nghiên cứu đã tiến hành khảo sát ứng dụng thư viện để chạy bài toán phân loại random forest. Công tác khảo sát cho thấy phương pháp Random forest được thiết kế trong bộ công cụ scikit-learn. Đây là thư viện ứng dụng học máy được phát triển sử dụng cho ngôn ngữ lập trình Python, được sử dụng trong nhiều mục đích phân loại giải đoán dữ liệu, trong đó xử lý ảnh cũng được áp dụng.

Trong lĩnh vực xử lý ảnh có thể sử dụng bộ thư viện hỗ trợ như GDAL, OpenCV,… thông qua nền tảng ứng dụng nổi tiếng Anaconda với một số trình biên dịch Python phổ biến hiện nay như Jupyter Notebook, Spyder… Chức năng cũng như độ tuỳ biến của các trình biên dịch này rất rộng bao gồm: làm sạch dữ liệu và chuyển đổi, mô phỏng dữ liệu, mô hình thống kê, xử lý, phân tích dữ liệu.v.v.

Hình 5: Minh họa Jupiter Notebook

Ở phạm vi nghiên cứu này, nhóm thực hiện đã sử dụng trình biên dịch Jupyter Notebook để thực nghiệm nhập dữ liệu đầu vào, đánh giá tệp "training data" và đưa vào mô hình phân loại học máy RF sử dụng thư viện Sklearn.



Hình 6: Minh hoạ nhập thư viện và dataset đầu vào

Minh họa đánh giá tệp mẫu, gán nhãn và lựa chọn số lớp cần phân loại:

Minh họa đưa vào mô hình phân loại học máy RF.

Việc thử nghiệm chạy các tập dữ liệu ảnh đầu vào, đánh giá xác định tệp mẫu và gán nhãn lớp dữ liệu cần phân loại đã được thực hiện. Việc bổ sung các tệp mẫu và để chạy ra kết quả bằng mô hình phân loại học máy ảnh viễn thám theo phương pháp RF hiện tiếp tục được thực hiện, kèm với đánh giá kiểm chứng sẽ hứa hẹn mang lại sự chính xác và hiệu quả trong ứng dụng phục vụ mục tiêu phát triển kinh tế-xã hội.

Kết luận

Hiện nay có rất nhiều nghiên cứu về ứng dụng viễn thám trong giám sát và quản lý tài nguyên thiên nhiên và môi trường. Tuy nhiên hiện nay, các phương pháp phân loại truyền thống vẫn đang được sử dụng rộng rãi. Trong từng trường hợp cụ thể, các sản phẩm của các phương pháp phân loại truyền thống có độ chính xác chưa cao do bị ảnh hưởng vẫn đề nhiễu điểm ảnh (Phạm Minh Hải, 2016). Qua công tác nghiên cứu khảo sát cho thấy, việc nghiên cứu ứng dụng phương pháp Randon Forest trong phân loại học máy cho ảnh viễn thám hoàn toàn khả thi và dự báo sẽ đem một phương pháp tương đối mới và mang lại độ chính xác cao với chi phí thấp hơn với các phương pháp phân loại có kiểm định truyền thống trong phần mềm thương mại đang sử dụng. Trong bài báo tiếp theo, nhóm nghiên cứu sẽ trình bày cụ thể ứng dụng phương pháp random forest trong công tác phân loại học máy áp dụng cho ảnh vệ tinh có độ phân giải vừa và nhỏ.❍

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]. Hai., P.M (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.❍

Summary

An introduction of Random forest in the machine learning revolution and the application in satellite image classification

Pham Minh Hải, Nguyen Ngoc Quang

Random forest is a machine learning statistic method for satellite image classfication, regression by using multiple decision trees. Random Forest shows that it is more efficient than the commonly image classificaton methods because it is possible to find which attributes are more important than others in the decision tree. In fact, it may also indicate that some attributes are ineffective. The main objective of this manuscript isto investigate the method and direct the methodto apply insatellite image classification.❍