TIÊU ĐỀ: Ứng dụng thuật toán học máy Principal Component Analysis và Random Forest vào phân loại ảnh Viễn thám

1. Tổng quan

Thông tin về LULC (Land Use Land Cover) đóng một vai trò quan trọng trong nhiều khía cạnh của cuộc sống, từ khoa học, kinh tế đến chính trị. Thông tin về lớp phủ là cơ sở để xây dựng các ứng dụng kế thừa từ đó ảnh hưởng đến độ chính xác của tất cả các ứng dụng đó. Vì vậy nhu cầu thông tin chính xác và kịp thời về lớp phủ đất đang được đề cao. Chỉ số thể hiện những thay đổi bề ngoài của Trái đất, bất kể là loại nào, là lớp phủ đất. Các nghiên cứu gần đây đã báo cáo rằng việc sử dụng đất / thay đổi lớp phủ (LULC) đang có tác động ngày càng tiêu cực đến các khía cạnh khác nhau của bề mặt Trái đất, chẳng hạn như hệ sinh thái trên cạn, cân bằng nước, đa dạng sinh học và khí hậu. Trong số này, tác động của LULC đến hệ sinh thái trên cạn nhận được nhiều sự quan tâm nhất của các nhà nghiên cứu, vì hệ sinh thái đóng một vai trò quan trọng trong chu trình carbon toàn cầu. Tuy nhiên, do biến đổi khí hậu (ví dụ: xu hướng ấm lên, tần suất ngày càng tăng của các hiện tượng khí hậu khắc nghiệt), thay đổi lớp phủ đất, cũng như các chính sách thay đổi của chính phủ, nên hiện trạng khu Cần Giờ thay đổi nhiều. Do đó, thông tin chính xác, hiện tại và lâu dài của bản đồ sử dụng đất / bìa là yêu cầu cao ở mọi nơi, không chỉ cho sự phát triển kinh tế mà còn cho các chính sách của chính phủ, đảm bảo hệ sinh thái, …

Dữ liệu cảm biến từ xa đã được công nhận là một trong những nguồn dữ liệu quan trọng nhất cho việc lập bản đồ lớp phủ đất và để theo dõi sự thay đổi lớp phủ đất theo thời gian với Sentinel-2 là nguồn dữ liệu được sử dụng thường xuyên nhất. Các nguồn dữ liệu khác cho các nghiên cứu LULC là Satellite Pour l'Observation de la Terre (SPOT), Synthetic Aperture Radar (SAR), Máy quang phổ hình ảnh độ phân giải vừa phải (MODIS) và Landsat. Trong nhiều ứng dụng liên quan đến việc hỗ trợ ra quyết định như giám sát môi trường, biến động, đô thị hóa thì độ phân giải không gian cao và dữ liệu dài hạn là điều cần thiết. Cho đến nay, S2 là hệ thống hoạt động duy nhất có thể cung cấp độ phân giải không gian cao (10 m), độ phân giải theo thời gian (12 ngày - với một vệ tinh duy nhất và 6 ngày, nếu dữ liệu từ cả hai vệ tinh được kết hợp) và liên tục trong hơn 5 năm.

Hình : Một vài vệ tinh viễn thám. (Nguồn: NASA)

Khi lập bản đồ độ che phủ đất trên một khu vực rộng lớn, có hai thách thức chính - cần xử lý "dữ liệu lớn" và tính khả dụng của hình ảnh không có đám mây trên một khu vực rộng lớn. Sẽ rất tốn công sức nếu chúng tôi xử lý dữ liệu bằng các phương pháp truyền thống, từ tìm kiếm, lọc, tải xuống và khảm đến xử lý trước, chẳng hạn như tạo mặt nạ đám mây hoặc hiệu chỉnh khí quyển. Những "dữ liệu lớn" như vậy không chỉ sử dụng nhiều lao động mà còn yêu cầu dung lượng lưu trữ đáng kể và khả năng truy cập vào điện toán công suất cao. Hơn nữa, do mây che phủ, không dễ để đạt được hình ảnh rõ nét cho một khu vực rộng lớn trong một khoảng thời gian ngắn (ví dụ: hàng tháng). Để giảm khoảng thời gian tạo ảnh ghép không có đám mây, các ảnh có mây một phần phải được tải xuống và xử lý trước, tạo ra khối lượng công việc nhiều và khó khăn hơn.

Google Earth Engine (GEE), một nền tảng điện toán dựa trên đám mây, có thể giải quyết các vấn đề quan trọng nhất liên quan đến việc lập bản đồ lớp phủ đất của các khu vực rộng lớn. Người dùng có thể phân tích tất cả các hình ảnh được cảm nhận từ xa có sẵn bằng trình soạn thảo mã Môi trường phát triển tích hợp (IDE) dựa trên web mà không cần tải những dữ liệu này xuống máy cục bộ.

2. Phương pháp

2.1 Principal Components Analysis

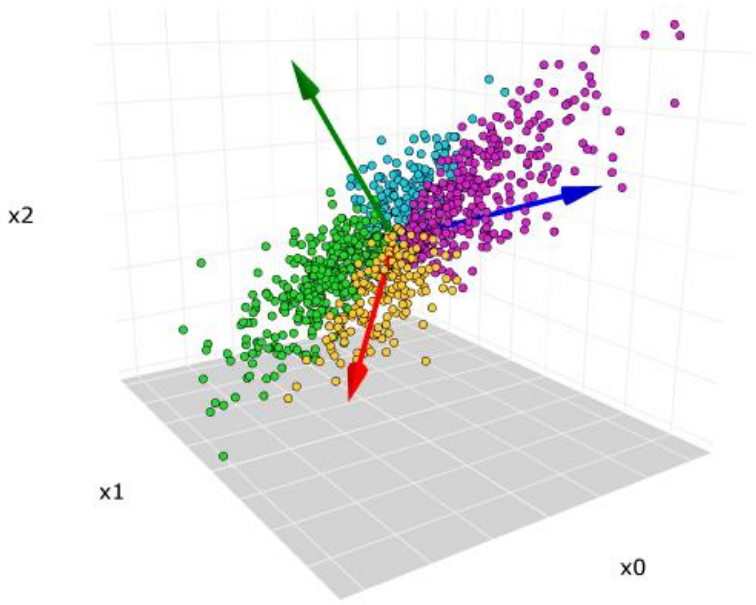
Phép phân tích thành phần chính (Principal Components Analysis - PCA) là một thuật toán thống kê sử dụng phép biến đổi trực giao để biến đổi một tập hợp dữ liệu từ một không gian nhiều chiều sang một không gian mới ít chiều hơn nhằm tối ưu hóa việc thể hiện sự biến thiên của dữ liệu. Trong bài viết này, chúng tôi sử dụng PCA để giảm số chiều dữ liệu từ 12 về 3. Phép biến đổi tạo ra những ưu điểm như: Giảm khối lượng tính toán, tối ưu hóa bộ nhớ, giảm thời gian chạy các thuật toán phân loại.

Theo dõi đoạn code sau

```
function PCA(maskedImage){
  var image = maskedImage.unmask()
  var scale = 10;
  var region = geometry;
  var bandNames = image.bandNames();
  var meanDict = image.reduceRegion({
    reducer: ee.Reducer.mean(),
    geometry: region,
    scale: scale,
    maxPixels: 1e12,
    bestEffort: true,
    tileScale: 16
  });
  var means = ee.Image.constant(meanDict.values(bandNames));
  var centered = image.subtract(means);
  var getNewBandNames = function(prefix) {
    var seq = ee.List.sequence(1, bandNames.length());
    return seq.map(function(b) {
      return ee.String(prefix).cat(ee.Number(b).int());
    });
  };


  var getPrincipalComponents = function(centered, scale, region) {
```

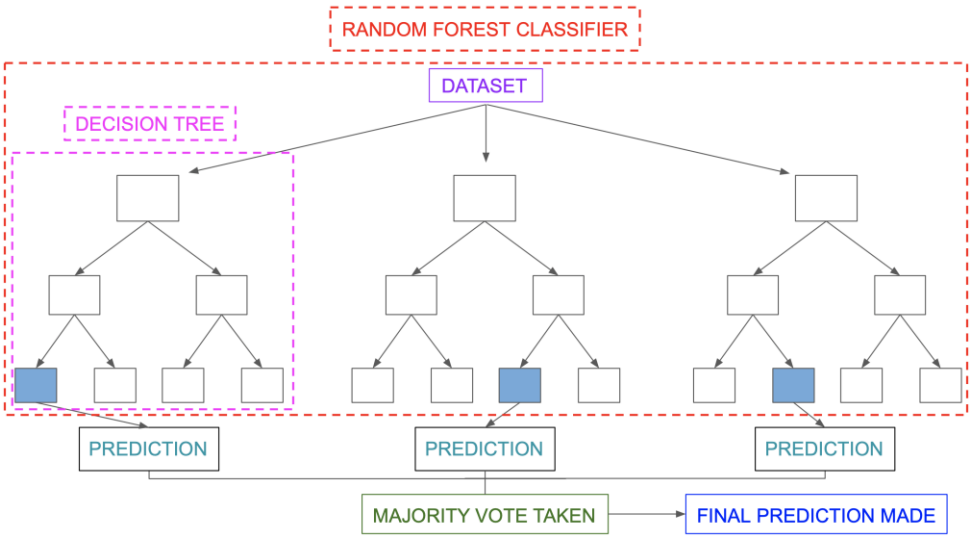
```
    var arrays = centered.toArray();
    var covar = arrays.reduceRegion({
      reducer: ee.Reducer.centeredCovariance(),
      geometry: region,
      scale: scale,
      maxPixels: 1e12,
      bestEffort: true,
      tileScale: 16
    });
    var covarArray = ee.Array(covar.get('array'));
    var eigens = covarArray.eigen();
    var eigenValues = eigens.slice(1, 0, 1);
    var eigenValuesList = eigenValues.toList().flatten()
    var total = eigenValuesList.reduce(ee.Reducer.sum())
    var percentageVariance = eigenValuesList.map(function(item) {
      return (ee.Number(item).divide(total)).multiply(100).format('%.2f')
    })
    var eigenVectors = eigens.slice(1, 1);
    var arrayImage = arrays.toArray(1);
    var principalComponents = ee.Image(eigenVectors).matrixMultiply(arrayImage);
    var sdImage = ee.Image(eigenValues.sqrt())
      .arrayProject([0]).arrayFlatten([getNewBandNames('sd')]);
    return principalComponents
      .arrayProject([0])
      .arrayFlatten([getNewBandNames('pc')])
      .divide(sdImage);
  };
  var pcImage = getPrincipalComponents(centered, scale, region);
return pcImage.mask(maskedImage.mask());
}
```



Hình : Hình minh họa mô tả về PCA (Nguồn : Towards Data Science)

2.2 Random Forest

Rừng ngẫu nhiên là một kỹ thuật học máy được sử dụng để giải quyết các vấn đề về hồi quy và phân loại. Nó sử dụng học tập đồng bộ, là một kỹ thuật kết hợp nhiều bộ phân loại để cung cấp giải pháp cho các vấn đề phức tạp. Một thuật toán rừng ngẫu nhiên bao gồm nhiều cây quyết định. “Forest” được tạo bởi thuật toán rừng ngẫu nhiên được đào tạo thông qua tổng hợp bagging hoặc bootstrap. Bagging là một siêu thuật toán tổng hợp giúp cải thiện độ chính xác của các thuật toán máy học. Thuật toán (RF) thiết lập kết quả dựa trên dự đoán của cây quyết định. Nó dự đoán bằng cách lấy giá trị trung bình hoặc trung bình của đầu ra từ các cây khác nhau. Tăng số lượng cây làm tăng độ chính xác của kết quả.

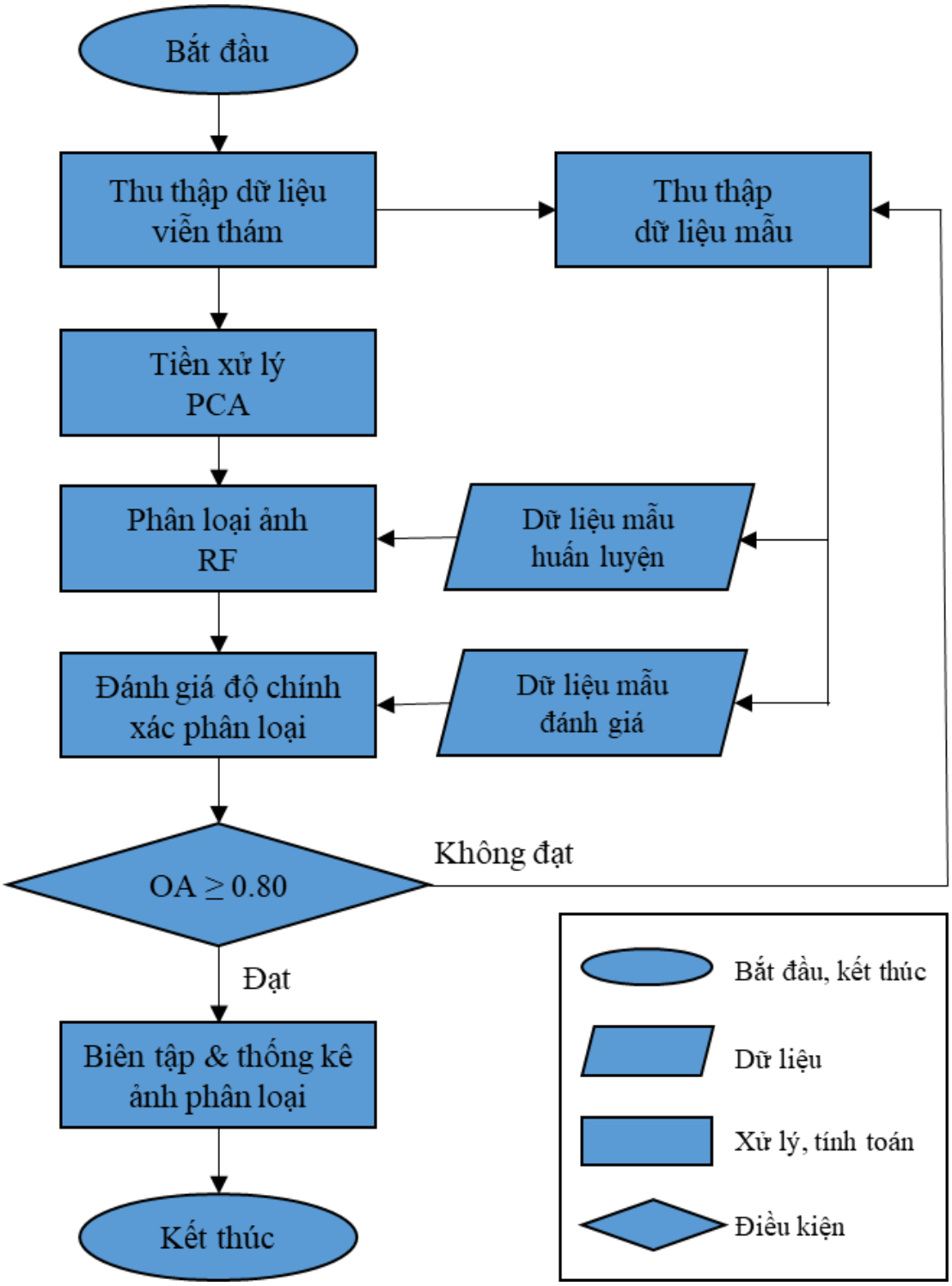


Hình: Ví dụ minh họa đơn giản thuật toán Random Forest

(Nguồn: https://miro.medium.com/)

2.3 Các bước thực hiện

Công việc chung bao gồm nhiều bước, được thực hiện trong một tập lệnh GEE (Hình dưới) gồm thành phần như: dữ liệu ban đầu, xử lý ảnh và đánh giá độ chính xác, kết quả.

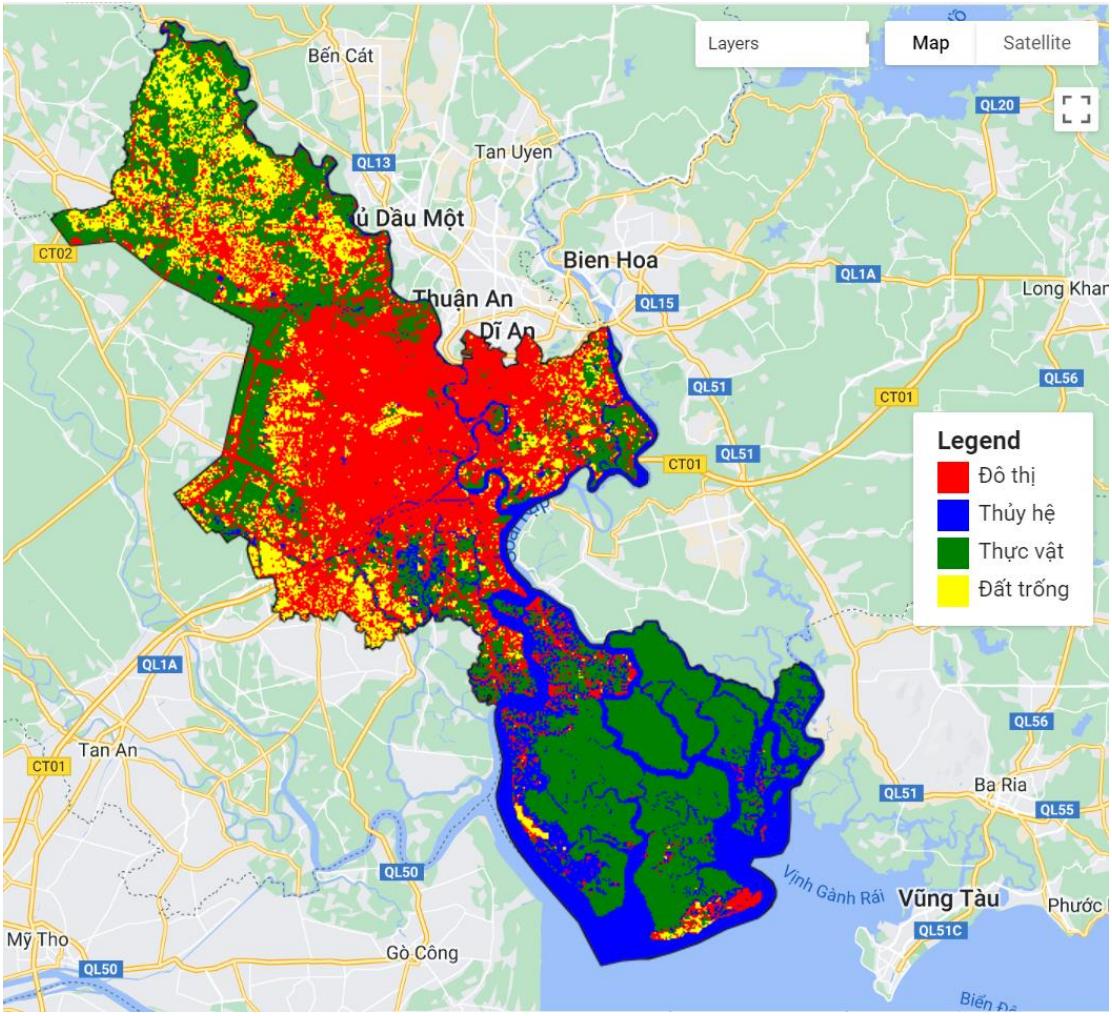


Hình : Các bước thực hiện

3. Kết quả

Dữ liệu Viễn thám ban đầu bao gồm 12 Bands ảnh thì sau khi sử dụng thuật toán PCA sẽ thành 12 thành phần sau

STT	Thành phần	%
1	PC1	82,13
2	PC2	13,68
3	PC3	2,34
4	PC4	1,20
5	PC5	0,21
6	PC6	0,14
7	PC7	0,08
8	PC8	0,08
9	PC9	0,06
10	PC10	0,04
11	PC11	0,03
12	PC12	0,02



Hình : Kết quả sau phân loại

Từ bảng phần trăm thành phần chính, ta chọn 4 thành phần chính PC1, PC2, PC3, PC4 chiếm 99,35%. Vì vậy từ 12 Bands ban đầu, ta đã giảm xuống còn 4 Bands mà đại diện gần như toàn bộ tập dữ liệu ban đầu.

Từ ma trận lỗi (confusion matrix) ta có thể đánh giá được độ chính xác. Và độ chính xác tổng thể (OA) cho đề tài là 0.9405940594059405 ~ 94%. Ta thấy sử dụng PCA kết hợp với RF là phương pháp phân loại cực kì tốt, với độ chính xác vượt trội, ở mức xuất sắc.

Kết quả phân loại được thể hiện trên nền tảng Cloud, có thể dễ dàng cho mọi người truy cập vào để xem, đánh giá, sử dụng, dowload, ... qua ứng dụng Web. Từ kết quả phân loại lớp phủ (LULC), có thể làm tiền đề để tiến hành làm các phần mềm ứng dụng viễn thám (RS), hệ thống thông tin địa lý (GIS) phục vụ công tác quản lý tài nguyên thiên nhiên, khí hậu, thiên tai…