|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI



HỌ VÀ TÊN: NguyỄn HẢI ĐÌNH

**SỬ DỤNG MÔ HÌNH BÁN GIÁM SÁT PHÁT HIỆN VÙNG SẠT LỞ TỪ ẢNH VỆ TINH**

ĐỒ ÁN TỐT NGHIỆP

HÀ NỘI, NĂM 2023

|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI

HỌ VÀ TÊN: NGUYỄN HẢI ĐÌNH

**SỬ DỤNG MÔ HÌNH BÁN GIÁM SÁT PHÁT HIỆN VÙNG SẠT LỞ TỪ ẢNH VỆ TINH**

|  |  |
| --- | --- |
| Ngành: | Công nghệ thông tin |
| Mã số: |  |

|  |  |
| --- | --- |
| NGƯỜI HƯỚNG DẪN | 1. TS. TRẦN MẠNH TUẤN |

HÀ NỘI, NĂM 2023

**GÁY BÌA ĐỒ ÁN TỐT NGHIỆP, KHÓA LUẬN TỐT NGHIỆP**

**HỌ VÀ TÊN: NGUYỄN HẢI ĐÌNH ĐỒ ÁN/KL TỐT NGHIỆP HÀ NỘI, NĂM: 2023 ……**

|  |  |
| --- | --- |
|  | CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM  **Độc lập - Tự do - Hạnh phúc**  --------------------  **NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP** |

|  |  |
| --- | --- |
| Họ tên sinh viên: **Nguyễn Hải Đình** | Hệ đào tạo: **Đại học chính quy** |
| Lớp: **61HT** | Ngành: **Công nghệ thông tin** |
| Khoa: **Công nghệ thông tin** |  |

1. TÊN ĐỀ TÀI:

- SỬ DỤNG MÔ HÌNH BÁN GIÁM SÁT PHÁT HIỆN VÙNG SẠT LỞ TỪ ẢNH VỆ TINH

1. CÁC TÀI LIỆU CƠ BẢN:

1. NỘI DUNG CÁC PHẦN THUYẾT MINH VÀ TÍNH TOÁN:

Chương 1: Giới thiệu (20%)

* + Lý do chọn đề tài
  + Mục tiêu đề tài
  + Đối tượng và phạm vi nghiên cứu
  + Khu vực nghiên cứu

Chương 2: Tiếp cận cơ sở lý thuyết (20%)

Chương 3: Ứng dụng phương pháp và xây dựng mô hình (60%)

1. GIÁO VIÊN HƯỚNG DẪN TỪNG PHẦN

|  |  |
| --- | --- |
| **Phần** | **Họ tên giáo viên hướng dẫn** |
| Chương 2. Tiếp cận cơ sở lý thuyết | TS. Trần Mạnh Tuấn |
| Chương 3. Ứng dụng phương pháp và xây dựng mô hình | TS. Trần Mạnh Tuấn |

1. NGÀY GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Ngày .... tháng …. năm 2023.

|  |  |  |
| --- | --- | --- |
| **Trưởng Bộ môn**  *(Ký và ghi rõ Họ tên)* |  | **Giáo viên hướng dẫn chính**  *(Ký và ghi rõ Họ tên)* |

Nhiệm vụ Đồ án tốt nghiệp đã được Hội đồng thi tốt nghiệp của Khoa thông qua.

|  |
| --- |
| Ngày …. tháng …. năm 2023.  **Chủ tịch Hội đồng**  *(Ký và ghi rõ Họ tên)* |

Sinh viên đã hoàn thành và nộp bản Đồ án tốt nghiệp cho Hội đồng thi.

|  |
| --- |
| Ngày .... tháng …. năm 2023.  **Sinh viên làm Đồ án tốt nghiệp**  *(Ký và ghi rõ Họ tên)* |

**TÓM TẮT ĐỀ TÀI**

Trong thời đại kỹ thuật số hiện nay, trí tuệ nhân tạo (Artificial Intelligence - AI) đang trở thành một trong những chủ đề nóng bỏng nhất. AI là khả năng của máy tính và hệ thống chương trình thuật toán chứa kiến thức lý luận để thực hiện hành động phức tạp mà trước đây chỉ có con người mới làm được. Công nghệ AI đang được sử dụng rộng rãi trong các lĩnh vực như y tế, tài chính, sản xuất, giáo dục và nhiều lĩnh vực khác đặt biệt là vấn đề môi trường.

Hiện nay, trái đất chúng ta đang chịu nhiều những thiên tai do ảnh hưởng của sự biến đổi khi hậu. Sự tác động của thiên tai như bão, lũ, hạn hán, … đã gây ra rất nhiều những hậu quả nghiêm trọng về vật chất cũng như tinh thần. Việc khắc phục những hậu quả do thiên tai gây ra cũng vô cùng khó khăn đặt biệt là sạt lở bờ biển

Đáp ứng giải quyết vấn đề trên nên em chọn làm đề tài tốt nghiệp, sử dụng một vài mô hình bán giám sát để dự báo những vùng sạt lở dựa trên vùng sạt lở đã xảy ra.

**CÁC MỤC TIÊU CHÍNH**

Với đề tài “Sử dụng mô hình bán giám sát phát hiện vùng sạt lở từ ảnh vệ tinh ”, em đặt ra một số mục tiêu chính như sau:

* Mục tiêu 1: Tìm hiểu về bài toán phát hiện vùng sạt lở.
* Mục tiêu 2: Thu thập dữ liệu ảnh vệ tinh vùng sạt lở.
* Mục tiêu 3: Tìm hiểu một số mô hình bán giám sát.
* Mục tiêu 4: Cài đặt mô hình bán giám sát phát hiện vùng sạt lở từ ảnh vệ tinh .
* Mục tiêu 5: So sánh kết quả dự đoán với một số mô hình với nhau..

**KẾT QUẢ DỰ KIẾN**

Kết quả dự kiến của đề tài là:

**1.** **Lý thuyết**

● Nắm được những kiến thức đã tìm hiểu trong quá trình tìm hiểu.

● Nắm vững kiến thức của các công nghệ sử dụng.

**2.** **Kỹ năng**

● Kỹ năng sử dụng các công nghệ Python, sklearn.

● Kỹ năng quản lý dự án.

**3.** **Sản phẩm**

● Hoàn thành báo cáo đồ án.

● Xây dựng thành công ứng dụng dự báo vùng sạt lở từ ảnh về tinh.

**LỜI CAM ĐOAN**

Tác giả xin cam đoan đây là Đồ án tốt nghiệp của bản thân tác giả. Các kết quả trong đồ án tốt nghiệp này là trung thực, và không sao chép từ bất kỳ một nguồn nào và dưới bất kỳ hình thức nào. Việc tham khảo các nguồn tài liệu (nếu có) đã được thực hiện trích dẫn và ghi nguồn tài liệu tham khảo đúng quy định.

|  |
| --- |
| **Tác giả ĐATN**  **NGUYỄN HẢI ĐÌNH** |

**LỜI CẢM ƠN**

Lời đầu tiên, em xin chân thành cảm ơn các giảng viên Trường đại học Thủy Lợi và đặc biệt là các thầy cô bộ môn khoa Công nghệ thông tin đã truyền dạy kiến thức cần thiết về ngành học cũng như trong đồ án . Đặc biệt về TS. Trần Mạnh Tuấn, em xin chân thành cảm ơn thầy đã giúp đỡ, bổ sung cho em những kiến thức và cho em những lời khuyên, gợi ý để em có thể hoàn thành đồ án một cách nhanh chóng và hiệu quả nhất.

**EM XIN CHÂN THÀNH CẢM ƠN**

*Sinh viên*

*Nguyễn Hải Đình*

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc136952525)

[1.1 Lý do chọn đề tài 1](#_Toc136952526)

[1.2 Mục tiêu đề tài 1](#_Toc136952527)

[1.3 Đối tượng và phạm vi nghiên cứu 1](#_Toc136952528)

[1.3.1 Đối tượng nghiên cứu 1](#_Toc136952529)

[1.3.2 Phạm vi nghiên cứu 1](#_Toc136952530)

[1.3.3 Khu vực nghiên cứu 1](#_Toc136952531)

[CHƯƠNG 2: TIẾP CẬN CƠ SỞ LÝ THUYẾT 2](#_Toc136952532)

[2.1 Học máy (Machine Learning) 2](#_Toc136952533)

[2.1.1 Tổng quan học máy 2](#_Toc136952534)

[**2.1.1.1 Học có giám sát (Supervised Learning)** 3](#_Toc136952535)

[**2.1.1.2 Học không giám sát (Unsupervised Learning)** 4](#_Toc136952536)

[**2.1.1.3 Học bán giám sát** 5](#_Toc136952537)

[2.2 Lý thuyết mờ : 5](#_Toc136952538)

[2.3 Tổng quan về phân cụm bán giám sát : 5](#_Toc136952539)

[2.3.1 Thuật toán cơ sở Fuzzy C-mean ( FCM ) : 6](#_Toc136952540)

[2.3.2 Thông tin bổ trợ trong phân cụm bán giám sát : 6](#_Toc136952541)

[2.3.3 Các thuật toán phân cụm bán giám sát được kết hợp với thông tin bổ trợ : 6](#_Toc136952542)

[CHƯƠNG 3: ỨNG DỤNG PHƯƠNG PHÁP VÀ XÂY DỰNG MÔ HÌNH 6](#_Toc136952543)

CHƯƠNG 1: GIỚI THIỆU

1.1 Lý do chọn đề tài

Hiện nay, do trái đất đang chịu rất nhiều thiệt hại về biến đổi khi hậu đặc biệt là nước ta đang phải chịu hậu quả của thiên tai khá nhiều về tinh thần cũng như vật chất. Để giảm hậu quả ít nhất có thể vậy nên việc dự báo vùng sạt lở trong tương lai qua ở gần vùng sạt lở đã xảy ra, thì sử dụng mô hình Bán giám sát là một phương pháp hiệu quả khi dựa trên ảnh.

Với những lý do trên, đề tài nghiên cứu của đồ án: “Sử dụng mô hình bán giám sát phát hiện vùng sạt lở ” được đề xuất là hết sức cần thiết.

1.2 Mục tiêu đề tài

Dựa trên lý do chọn đề tài, trong nghiên cứu này đồ án hướng đến các mục tiêu chính sau:

* Mục tiêu 1: Tìm hiểu về bài toán phát hiện vùng sạt lở.
* Mục tiêu 2: Thu thập dữ liệu ảnh vệ tinh vùng sạt lở.
* Mục tiêu 3: Tìm hiểu một số mô hình bán giám sát.
* Mục tiêu 4: Cài đặt mô hình bán giám sát phát hiện vùng sạt lở từ ảnh vệ tinh .
* Mục tiêu 5: So sánh kết quả dự đoán với một số mô hình với nhau.

1.3 Đối tượng và phạm vi nghiên cứu

1.3.1 Đối tượng nghiên cứu

Xây dựng mô hình bán giám sát phát hiện vùng sạt lở

1.3.2 Phạm vi nghiên cứu

Với đề tài lựa chọn là “ phát hiện vùng sạt lở ”, em tập trung vào ảnh vệ tinh có xuất hiện vùng sạt lở ít nhất một lần

1.3.3 Khu vực nghiên cứu

Nghiên cứu phát hiện vùng sạt lở tập trung sâu vào vùng ven bờ biển đã xảy ra sạt lở

CHƯƠNG 2: TIẾP CẬN CƠ SỞ LÝ THUYẾT

2.1 Học máy (Machine Learning)

2.1.1 Tổng quan học máy

Máy học (ML - Machine learning) là một nhánh của ngành trí tuệ nhân tạo AI (Artificial Intelligence) và khoa học máy tính, lấy việc thu thập dữ liệu, phân tích và sử dụng dữ liệu vào thuật toán để bắt chước cách con người học, dần dần cải thiện độ chính xác của nó.

Máy học đóng vai trò quan trọng trong xã hội hiện nay, là công nghệ này đang được nghiên cứu sâu, phát triển rộng với nhiều ứng dụng hữu ích trong thực tiễn đời sống.

Máy học có sử dụng thuật toán hiệu để học hỏi từ dữ liệu thay vì học hỏi thông tin từ con người .



Hình 2.1 Hình ảnh mô tả về học máy

Máy học hiện đang được sử dụng rộng rãi trong nhiều lĩnh vực nên đã là thuật ngữ chỉ các chương trình có khả năng tự hoàn thiện về năng xuất theo thời gian dựa trên dữ liệu

Học máy là một thành phần quan trọng của lĩnh vực khoa học dữ liệu đang phát triển. Dựa vào việc sử dụng phương pháp thống kê, các thuật toán có thể phân loại, dự đoán và khám phá những thông tin chi tiết trong các mô hình khai thác dữ liệu.

Do tốc độ sản sinh ra dữ liệu càng lớn vậy nên máy học vẫn cần con người trong việc tìm hiểu dữ liệu cơ sở và lựa chọn các kĩ thuật phù hợp xử lí thông tin, đánh giá mô hình. Đồng thời, trước khi sử dụng, dữ liệu phải được làm sạch, không có sai lệch và không có dữ liệu giả.

Hiện nay, sự xuất hiện và phát triển của công nghệ Dữ liệu lớn (Big data) đã giúp đáp ứng cho thuật toán machine learning lượng dữ liệu đủ lớn để cải thiện độ chính xác của mô hình và dự đoán.

Theo các phương pháp học tập, học máy có thể được chia thành: học có giám sát “supervised learning”, học không giám sát “unsupervised learning”, học bán giám sát “semi-supervised learning”, học tăng cường “reinforcement learning”.

**2.1.1.1 Học có giám sát (Supervised Learning)**

**Học có giám sát** là một kĩ thuật của ngành học máy để xây dựng một hàm (*function*) từ dữ liệu huấn luyện. Dữ liệu huấn luyện bao gồm các cặp gồm đối tượng đầu vào (thường dạng vec-tơ), và đầu ra mong muốn. Đầu ra của một hàm có thể là một giá trị liên tục (gọi là hồi qui), hay có thể là dự đoán một nhãn phân loại cho một đối tượng đầu vào (gọi là phân loại). Nhiệm vụ của chương trình học có giám sát là dự đoán giá trị của hàm cho một dữ liệu đầu vào bất kỳ, sau khi đã xem xét một số ví dụ huấn luyện (nghĩa là, các cặp đầu vào và đầu ra tương ứng).

Một số ví dụ sử dụng phương pháp học có giám sát: dự báo mức độ nặng nhẹ của bệnh nhân dựa vào chỉ số đã đo được, phân loại thư rác, dự đoán giá của sản phẩm

**Hàm mất mát (Loss Function)**

Hàm mất mát là một hàm để tiến hành đánh giá một mô hình máy học xem mô hình có vận hành tốt hay tệ, hàm mất mát đối chiếu giá trị trả về của hàm mục tiêu đối với tập dữ liệu huấn luyện đầu vào. Theo quy ước, giá trị của hàm mất mát càng thấp thì độ chính xác của mô hình trả về càng tốt.

Hàm bình phương sai số (Squared error) là hàm mất mát có độ phổ biến nhất cho giá trị số, hàm entropy chéo (Cross entropy) với giá trị là nhãn.

Để mô hình không bị quá khớp khi huấn luyện ta thường chia dữ liệu đầu vào thành hai phần: tập dữ liệu huấn luyện để học mô hình, tập dữ liệu kiểm định để đánh giá mô hình và cho ra lỗi kiểm thử.

**Ưu điểm**

Cho phép thu thập thông tin hoặc tạo ra đầu ra với kinh nghiệm được tiếp thu trước đó.

Giải quyết vấn đề tính toán trong thời gian thế giới.

Phương pháp đơn giản dễ ứng dụng.

Kết quả đầu ra có thể cho ra kết quả với độ chính xác khá cao.

**Nhược điểm**

Dữ liệu cần để mô hình đào tạo phải đạt nhiều yêu cầu cần thiết: đủ nhãn cho từng lớp, bị nhiễu ít, đầu ra phải chuẩn xác.

Phân loại dữ liệu lớn có thể là một thử thách lớn.

Cần nhiều thời gian và công nghệ để tính toán

**2.1.1.2 Học không giám sát (Unsupervised Learning)**

Đây là phương pháp học máy mà ta không biết được mục tiêu của bài toán đang nhắm đến hay là nhãn có trong bài toán mà chỉ có dữ liệu đầu vào. **Học không có giám sát** là một phương pháp của ngành học máy nhằm tìm ra một mô hình mà phù hợp với các quan sát. Nó khác biệt với học có giám sát ở chỗ là đầu ra đúng tương ứng cho mỗi đầu vào là không biết trước. Trong học không có giám sát, một tập dữ liệu đầu vào được thu thập. Học không có giám sát thường đối xử với các đối tượng đầu vào như là một tập các biến ngẫu nhiên. Sau đó, một mô hình mật độ kết hợp sẽ được xây dựng cho tập dữ liệu đó. Học không giám sát sử dụng những thông tin chưa được gán nhãn để suy ra được mối quan hệ. Phương pháp này thường được ứng dụng để tìm ra được cấu trúc của tập dữ liệu đầu vào. Tuy nhiên lại thiếu đi phương pháp đánh giá cấu trúc được tìm ra là đúng hay sai.

**Ưu điểm**

Tìm ra được mối liên kết ẩn trong dữ liệu.

Tìm ra những đặc trưng để phân thể loại dữ liệu.

Xử lý trong khoảng thời gian thực, dữ liệu đầu vào được phân tích và phân loại ngay.

Dễ thu thập thông tin chưa gán nhãn hơn là gán nhãn.

**Nhược điểm**

Cần một tập dữ liệu đầu vào đủ lớn để có thể phân loại chính xác.

Chẳng thể đo lường độ tin cậy của kết quả đưa ra.

Không biết được số lớp cho trước.

#### **2.1.1.3 Học bán giám sát**

- Đây là phương thức học máy sử dụng cả dữ liệu đã gán nhãn và chưa gán nhãn để huấn luyện - điển hình là một lượng nhỏ dữ liệu có gán nhãn cùng với lượng lớn dữ liệu chưa gán nhãn.

- Trong phân cụm bán giám sát, một phần của dữ liệu được gán nhãn và sử dụng để xây dựng mô hình. Các nhãn này có thể được sử dụng như các điểm dữ liệu đã biết trong quá trình phân cụm. Các điểm dữ liệu còn lại, không được gán nhãn, được sử dụng để tìm ra cấu trúc ẩn trong dữ liệu và phân cụm chúng.

- Nó có những ưu điểm của 2 loại trên đồng thời có thể giảm những nhược điểm của cả 2

## 2.2 Logic mờ :

- Là phương pháp lập luận, có khả năng lập luận khá giống với con người

- kết quả lập luận là 1 tỷ lệ xấp xỉ thay vì một chính xác cụ thể vậy nên hay bị nhầm lẫn với “ Xác suất ”

- Ví dụ như Nam đang đứng ở vị trí cửa nối giữa phòng học và hành lang vậy thì Nam đấy sẽ đứng cả trong phòng học và hành lang. Nếu có 1 ngón chân của Nam ở trong lớp thì ta có thể nói là Nam đứng ở trong lớp 1% và ngoài hành lang 99%. Vậy nên chỉ cần Nam đứng ở cửa thì sẽ không có 1 biến cố nào chắc chắn rằng Nam đang hoàn toàn ở trong lớp hay Nam đang hoàn toàn ở ngoài hành lang.

- Vậy nên các tập mờ đưa ra kết quả không dựa vào sự ngẫu nhiên.

- Độ liên thuộc của Logic mờ có giá trị trong giới hạn từ 0 đến 1

## 2.3 Tổng quan về phân cụm bán giám sát :

### 2.3.1 Thuật toán cơ sở Fuzzy C-mean ( FCM ) :

Thuật toán phân cụm mờ được Bezdek [2] đề xuất dựa trên độ thuộc ukj của phần tử dữ liệu Xk từ cụm j. Hàm mục tiêu được xác định như sau:

(1)



+ m là số mờ hóa

+ C là số cụm, N là số phần tử dữ liệu, r là số chiều của dữ liệu.

+ là độ thuộc của phần tử dữ liệu từ cụm j.

+ là phẩn tử thứ k của .



+ là tâm của cụm j.

Khi đó ràng buộc của (1) là :

(2)



Sử dụng phương pháp Lagrange, xác định được tâm của cụm dựa vào (3) và độ thuộc dựa vào (4) từ hàm mục tiêu (1):

(3)



(4)



Khi đó, các bước thực hiện thuật toán phân cụm mờ bán giám sát thuật toán Fuzzy C-means như sau

|  |  |
| --- | --- |
| Input | Tập dữ liệu X gồm N phần tử trong không gian r chiều; số cụm C; mờ hóa m; ngưỡng ԑ; số lần lặp lớn nhất MaxStep>0. |
| Output | Ma trận U và tâm cụm V. |
| FCM | |
| 1 | t=0 |
| 2 | Random thỏa mãn điều kiện (2) |
| 3 | Repeat |
| 4 | t=t+1 |
| 5 | Tính bởi công thức (3) |
| 6 | Tính bởi công thức (4) |
| 7 | Cho đến khi or t > maxStep thì dừng |

### 2.3.2 Thông tin bổ trợ trong phân cụm bán giám sát :

- Các thuật toán bán giám sát được xây dựng dựa trên các thuật toán phân cụm mờ và thông tin bổ trợ từ phía người dùng. Mục đích của thông tin bổ trợ là hướng dẫn, giám sát, điều khiển quá trình phân cụm

- Thông tin bổ trợ thường được xây dựng trên 3 loại sau :

+ Các ràng buộc Must-link và Cannot-link: Ràng buộc Must-link yêu cầu 2 phần tử phải thuộc vào cùng 1 cụm, ngược lại ràng buộc Cannot-link chỉ ra 2 phần tử không thuộc cùng 1 cụm (mà phải thuộc 2 cụm khác nhau).

+ Các nhãn lớp của một phần dữ liệu: Một phần của dữ liệu được gán nhãn và phần còn lại không được gán nhãn.

+ Độ thuộc được xác định trước.

### 2.3.3 Các thuật toán phân cụm bán giám sát được kết hợp với thông tin bổ trợ :

**2.3.3.1 : Semi-supervised standard fuzzy clustering ( SSSFC )** :

- Được xây dựng dựa trên các thuật toán phân cụm mờ kết hợp với thông tin bổ trợ được người dùng cung cấp

- Thông tin bổ trợ cho phân cụm cho phân cụm bán giám sát có 3 dạng cơ bản gồm các ràng buộc Must-link , Cannon-link và 1 phần dữ liệu được gắn nhãn và độ thuộc được xác định trước

- Yasunori et al. [] đã đề xuất một thuật toán phân cụm bán giám sát mờ với thông tin bổ trợ là hàm phụ thuộc được bổ sung vào hàm mục tiêu FCM ( Fuzzzy C-Mean ) nhằm cải thiện hiệu quả trong quá trình phân cụm.

- Công thức hàm mục tiêu :

Trong đó có :

+ m là số mờ hóa

+ C là số cụm, N là số phần tử dữ liệu, r là số chiều của dữ liệu

+ là độ thuộc của phần tử vào dữ liệu là từ cụm j

+ là phần tử thứ k của

+ là tâm là cụm j

- Điều kiên ràng buộc : thì khi đó hàm độ thuộc bổ trợ của phần tử với cụm là đồng thời thỏa mãn , ,

- Khi đó dựa vào điều kiện ràng buộc và hàm mục tiêu trên thì tâm cụm được xác định như sau :

với

- Các giá trị của là :

+ Với m > 1 :

+ Với m = 1 :

Các bước thực hiện thuật toán phân cụm mờ bán giám sát SSSFC ( Semi-supervised standard fuzzy clustering ) :

|  |  |
| --- | --- |
| Input | Tập dữ liệu X gồm N phần tử, số cụm C, ma trận độ phụ thuộc , ngưỡng , số lần lặp tối đa maxStep > 0 |
| Output | Ma trận U và tâm cụm V |
| **SSSFC** | |
| 1 : | t = 0 |
| 2 : | Khởi tạo ngẫu nghiên |
| 3 : | Bước lặp ( Repeat ) |
| 4 : | t = t + 1 |
| 5 : | Tính theo các trường hợp trên của m |
| 6 : | Tính như trên |
| 7 : | Cho đến khi or t > maxStep thì dừng |

**2.3.3.2 Semi-supervised entropy regularized fuzzy clustering ( eSSFC )** :

- Được đề xuất bởi Yasunori và cộng sự [29] vào năm 2009

- Đến năm 2012, Yin [30] có đề xuất hiệu chỉnh hệ số Entropy và khi đó thuật toán đã

sử dụng độ thuộc bổ trợ để tăng hiệu suất phân cụm với điều kiện :

- Với tâm cụm ban đầu :

- ma trận phương sai của các mẫu được tính như sau :

- Khi đó, công thức tính khoảng cách Mahalanobis ( với )

- Hàm mục tiêu của eSFCM là :

- Công thức xác định ma trận độ phụ thuộc :

- Trong đó và tâm cụm

Các bước thực hiện thuật toán phân cụm mờ bán giám sát eSSFC ( Semi-supervised entropy regularized fuzzy clustering ) :

|  |  |
| --- | --- |
| Input | Tập dữ liệu X gồm N phần tử, số cụm C, độ phụ thuộc bổ trợ , ngưỡng , số lần lặp tối đa maxStep > 0 |
| Output | Ma trận U và tâm cụm V |
| eSFCM | |
| 1: | Tính ma trận P theo công thức với ma trận độ phụ thuộc đã cho và các tâm cụm ban đầu |
| 2: | t = 1 |
| 3: | Bước lặp lại |
| 4: | t = t + 1 |
| 5: | Tính bởi công thức trên |
| 6: | Tính |
| 7: | Cho đến khi or t > maxStep |

CHƯƠNG 3: ỨNG DỤNG PHƯƠNG PHÁP VÀ XÂY DỰNG MÔ HÌNH

**3.1 : Bài toán ứng dụng** :

- Sạt lở vùng biển là hiện tượng mất mát đất liền và đáy biển do các nguyên nhân khác nhau như sự động đất, thay đổi môi trường, khí hậu và hoạt động của con người. Sạt lở vùng biển gây ra những thiệt hại nghiêm trọng cho đời sống của cộng đồng và môi trường. Nó có thể làm giảm diện tích đất, tăng mực nước biển, phá hủy địa điểm sinh sống và gây ra sự cố môi trường. Những hậu quả của sạt lở vùng biển rất nghiêm trọng cả về vật chất và tinh thần.

- Vì vậy nên việc áp dụng mô hình để dự báo về vùng sạt lở từ những vùng sạt lở đã xảy ra giúp cho chúng ta có thể kịp thời đưa ra các giải pháp để có thể giảm thiểu tối đa được những thiệt hại nghiêm trọng

**3.2 Ngôn ngữ lập trình Python và thư viện** :

**3.2.1 Giới thiệu python** :

Python là một ngôn ngữ lập trình hướng đối tượng bậc cao, mạnh mẽ được đưa ra bởi Guido van Rossum. Nó rất dễ học và nhanh chóng trở thành một trong những ngôn ngữ lập trình nhập môn tối ưu nhất cho người mới bắt đầu. Python được sử dụng bộ nhớ cấp pháp động và tạo kiểu động. Python là ngôn ngữ lập trình bậc cao nhưng lại có cách tiếp cận đơn giản, hiệu quả để lập trình hướng đối tượng. Với cú pháp đơn giản của python mà vẫn linh hoạt, rõ ràng, dễ hiểu để viết kịch bản và phát triển ứng dụng trong nhiều ngành nghề, lĩnh vực của cuộc sống.

Python là một ngôn ngữ lập trình nổi bật với các tính năng:

* Ngôn ngữ bậc cao nhưng lại đơn giản và dễ hiểu: Python với cú pháp đơn giản và gọn gàng. Nó dễ học và lập trình hơn nhiều so với nhiều ngôn ngữ lập trình khác như C++, Java, C#. Ngôn ngữ python giúp cho việc lập trình trở nên thú vị, cho phép người lập trình có những giải pháp tốt hơn trong các lĩnh vực của cuộc sống.
* Mã nguồn mở là miễn phí: Ngôn ngữ được sử dụng tự do không tính phí, ngay cả cho mục đích thương mại. Bởi vì nó là mã nguồn mở, nên có thể sửa đổi một cách đơn giản. Ngôn ngữ lập trình python có một cộng đồng lớn đang liên tục được cải thiện sau mỗi bản cập nhật.
* Khả năng lập trình trên nhiều hệ điều hành: Các mã chương trình python có thể di chuyển từ nền tảng này sang nền tảng khác mà không thay đổi cấu trúc của chương trình. Nó chạy được hầu hết trên các hệ điều hành phổ biến như Windows, macOs, Linux.
* Khả năng nhúng và mở rộng của chương trình: Python dễ dàng kết hợp các đoạn mã C, C++ và những ngôn ngữ lập trình khác vào đoạn mã python. Điều này giúp ứng dụng có thêm nhiều tính năng tốt hơn mà những ngôn ngữ lập trình đơn lẻ khác khó có thể làm được.
* Ngôn ngữ thông dịch cấp cao: Khác với C/C++, python không khó khăn trong việc quản lý bộ nhớ. Vì khi chạy mã nó tự động chuyển mã lập trình thành ngôn ngữ máy giúp hoàn thành công việc. Bởi vì khi lập trình với nhiều
* Thư viện tiêu chuẩn lớn và phổ biến: Python có một lượng lớn các thư viện tiêu chuẩn giúp việc lập trình đơn giản hơn rất nhiều vì không trực tiếp cần phải tự viết tất cả các đoạn mã. Mà chỉ việc hiểu và ứng dụng các thư viện có sẵn được liên kết với python.
* Hướng đối tượng: Là ngôn ngữ lập trình hướng đối tượng (OOP) giúp giải quyết nhiều vấn đề phức tạp một cách trực quan. Với các vẫn đề phức tạp OPP giúp chia nhỏ tạo ra các đối tượng.

Hiện nay python vẫn là ngôn ngữ lập trình được sử dụng phổ biến trong các lĩnh vực của đời sống như: Lập trình ứng dụng web, khoa học tính toán, dự báo thiên nhiên, chẩn đoán trong y học, phát triển trò chơi. Ngoài những tính năng nổi bật như trên python còn có một số hạn chế đáng kể:

* Tốc độ tính toán, thực thi chậm.
* Sử dụng nhiều không gian bộ nhớ.
* Hạn chế trong việc phát triển ứng dụng trên thiết bị di động.
* Hạn chế trong việc thiết kế ứng dụng.

**3.2.2 Những thư viện sử dụng trong mô hình** :

**3.2.2.1 Numpy** :

- Là một thư viện hỗ trợ tính toán số học và xử lý mảng đa chiều một cách hiệu quả

- Được phát triển vào năm 2005, Numpy đã trở thành 1 công cụ phổ biến trong cộng

đồng khoa học dữ liệu và đóng góp đáng kể vào sự phát triển của Python trong lĩnh vực

này.

- Dưới đây là một số điểm nổi bật về thư viện NumPy :

+ Mảng đa chiều: NumPy định nghĩa kiểu dữ liệu ndarray (n-dimensional array) để lưu trữ dữ liệu trong các mảng đa chiều. Đây là một cấu trúc dữ liệu mạnh mẽ và linh hoạt, cho phép lưu trữ và xử lý dữ liệu số học trong các chiều khác nhau. Mảng NumPy có khả năng lưu trữ dữ liệu liên tục trong bộ nhớ, điều này làm cho việc truy cập và thao tác trên dữ liệu trở nên nhanh chóng và hiệu quả.

+ Tính toán vector hóa: NumPy cung cấp các phép toán vector hóa, cho phép thực hiện các phép toán số học và logic trên toàn bộ mảng một cách nhanh chóng và dễ dàng. Thay vì viết vòng lặp để thao tác trên từng phần tử của mảng, bạn có thể áp dụng các phép toán trực tiếp lên mảng, giúp tăng tốc đáng kể quá trình tính toán.

+ Broadcasting: Broadcasting là một tính năng quan trọng trong NumPy, cho phép bạn thực hiện các phép toán giữa các mảng có hình dạng khác nhau mà không cần phải tạo ra các mảng trung gian. NumPy tự động mở rộng kích thước các mảng để phù hợp với nhau và thực hiện phép toán một cách tự động.

+ Thao tác trên mảng: NumPy cung cấp nhiều hàm và phương thức cho phép bạn thao tác trên mảng một cách dễ dàng. Bạn có thể thực hiện các phép cắt (slicing), chuyển vị (transpose), thay đổi hình dạng (reshape), ghép nối (concatenate) và phân tách (split) mảng một cách linh hoạt

+ Các hàm toán học và thống kê: NumPy cung cấp nhiều hàm toán học và thống kê cho việc tính toán trên mảng. Bạn có thể áp dụng các hàm như sin, cos, exp, log, sqrt, mean, max, min, v.v. trực tiếp lên mảng của bạn một cách dễ dàng

+ Tích hợp với các thư viện khác: NumPy tích hợp tốt với các thư viện khoa học dữ liệu và máy học khác như SciPy, Pandas và Scikit-learn. Sự tích hợp này giúp cho việc phân tích dữ liệu và xây dựng mô hình trở nên thuận tiện hơn và mạnh mẽ hơn

+ Hiệu suất và tối ưu hóa: NumPy được viết bằng ngôn ngữ C, cho phép nó hoạt động nhanh chóng và hiệu quả. Ngoài ra, NumPy cũng cung cấp các công cụ tối ưu hóa để giảm thiểu việc sao chép dữ liệu và tăng tốc các phép toán trên mảng.

Trên đây chỉ là một tổng quan sơ lược về thư viện NumPy. Thư viện này rất phong phú và mạnh mẽ, cho phép bạn thực hiện nhiều tác vụ phức tạp trong tính toán số học và xử lý dữ liệu.

**3.2.2.2 Pandas** :

- Thư viện pandas là một thư viện mã nguồn mở phát triển dành cho Python, chuyên về xử lý và phân tích dữ liệu. Nó được xây dựng trên nền tảng của NumPy và cung cấp một cấu trúc dữ liệu mạnh mẽ và dễ sử dụng gọi là DataFrame, cho phép xử lý dữ liệu có cấu trúc và không có cấu trúc một cách linh hoạt.

- Dưới đây là một số điểm nổi bật về thư viện pandas :

+ DataFrame: DataFrame là cấu trúc dữ liệu chính trong pandas. Nó là một bảng hai chiều, tương tự như một bảng trong SQL hoặc một spreadsheet trong Excel. DataFrame cho phép bạn lưu trữ và làm việc với dữ liệu có cấu trúc theo cách dễ dàng và linh hoạt. Nó cung cấp các phương thức và chức năng để thực hiện các hoạt động như truy cập, lọc, xử lý, ghép nối, phân tích và biến đổi dữ liệu.

+ Xử lý dữ liệu: pandas cung cấp một số công cụ và chức năng để xử lý dữ liệu. Bạn có thể thực hiện các hoạt động như lọc, sắp xếp, đánh chỉ mục, xóa trùng lắp và chuẩn hóa dữ liệu. Nó cũng hỗ trợ xử lý các giá trị bị thiếu và tách dữ liệu thành các nhóm dựa trên các tiêu chí cụ thể.

+ Thao tác trên dữ liệu: pandas cung cấp nhiều phương thức và chức năng để thực hiện các thao tác trên dữ liệu. Bạn có thể thực hiện các phép cắt (slicing), ghép nối (merging), ghép (joining), và nhóm (grouping) dữ liệu một cách linh hoạt. Nó cũng hỗ trợ các phép toán trên hàng loạt dữ liệu và cho phép bạn thực hiện các phép toán số học và logic trên dữ liệu theo cách dễ dàng

+ Tích hợp dữ liệu: pandas cho phép tích hợp dữ liệu từ nhiều nguồn khác nhau. Bạn có thể đọc và ghi dữ liệu từ các định dạng như CSV, Excel, SQL, JSON, HTML và nhiều định dạng dữ liệu khác. Nó cũng hỗ trợ kết nối và truy vấn cơ sở dữ liệu SQL

+ Tính toán và thống kê: pandas cung cấp các chức năng tính toán và thống kê để phân tích dữ liệu. Bạn có thể tính toán các thống kê mô tả, như mean, median, mode, và standard deviation. Nó cũng cung cấp khả năng thực hiện các phép toán tổng hợp, như sum, count, min, max, và nhiều hơn nữa

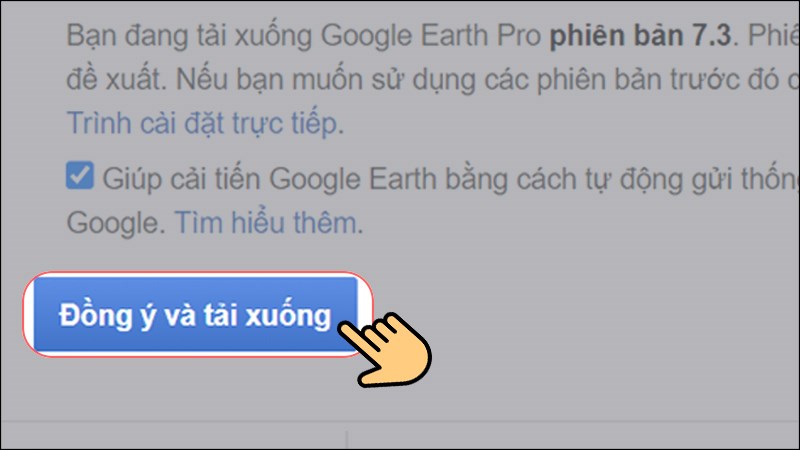
**3.3 Lược đồ tổng quát** :

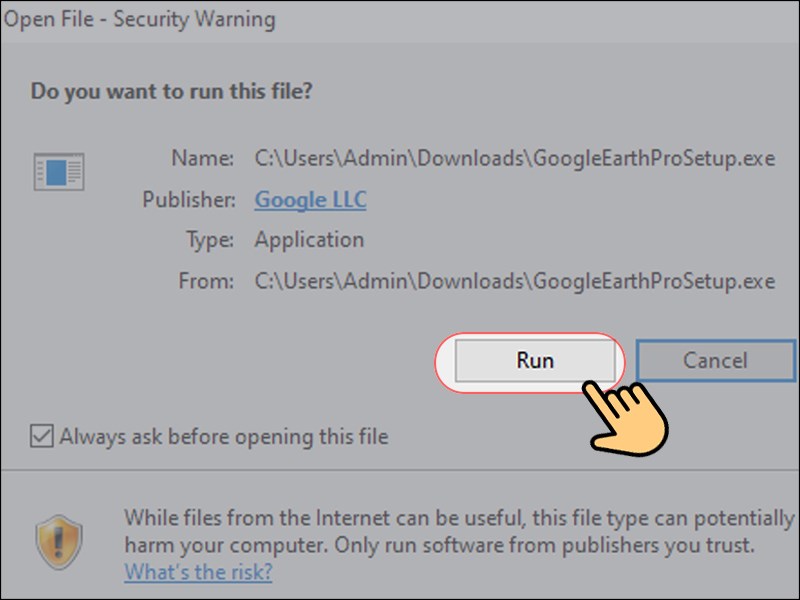
3.3.1 Thu thập dữ liệu :

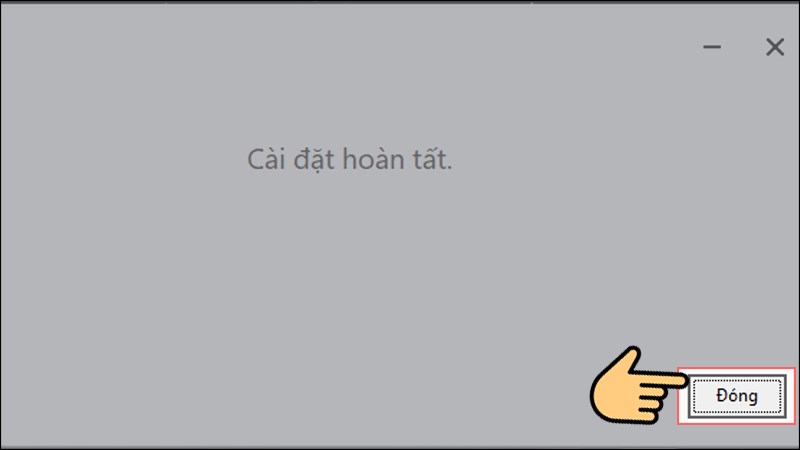
- Dữ liệu đầu vào bao gồm :

+ 1 tập được xây dựng trên bộ dữ liệu thực tế gồm 9 ảnh vệ tinh về vùng sạt lở bờ biển. Ảnh đều có kích thước 306 x 230 với Horizortal resolution là 96dpi và vertical resolution là 96dpi. Cách lấy ảnh vệ tinh về sạt lở vùng biển :

Bước 1 : Cài đặt ứng dụng Google Earth tại <https://www.google.com/intl/vi/earth/versions/> về máy tính



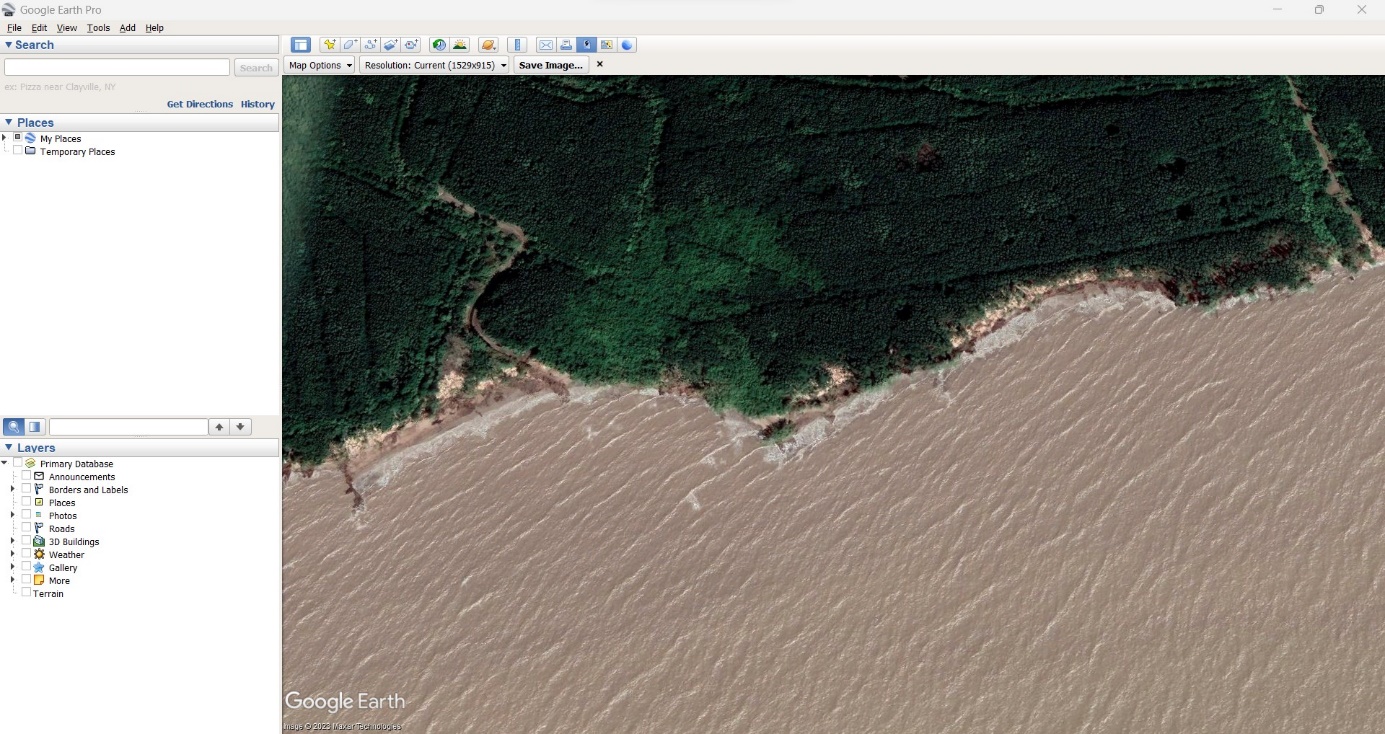




Bước 2 : Tìm và chọn lọc ra khu vực bị sạt lở vùng biển trên ứng dụng Google Earth pro và lưu ảnh

+ Giữ chuột trái và xoay chuột để di chuyển đến vị trí sạt lở vùng biển cần tìm

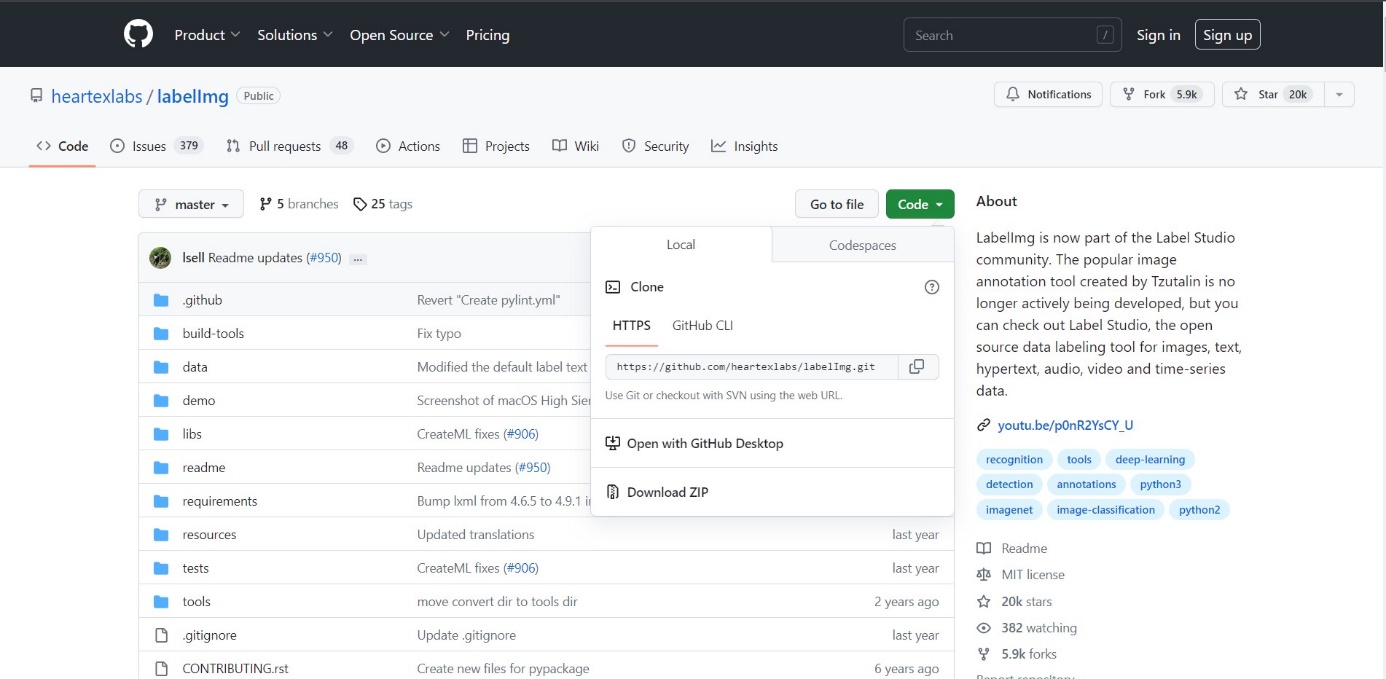
+ Nhấn Save image, tắt những Map option không cần thiết và chọn đô phân giải của ảnh rồi nhấn Save Image:



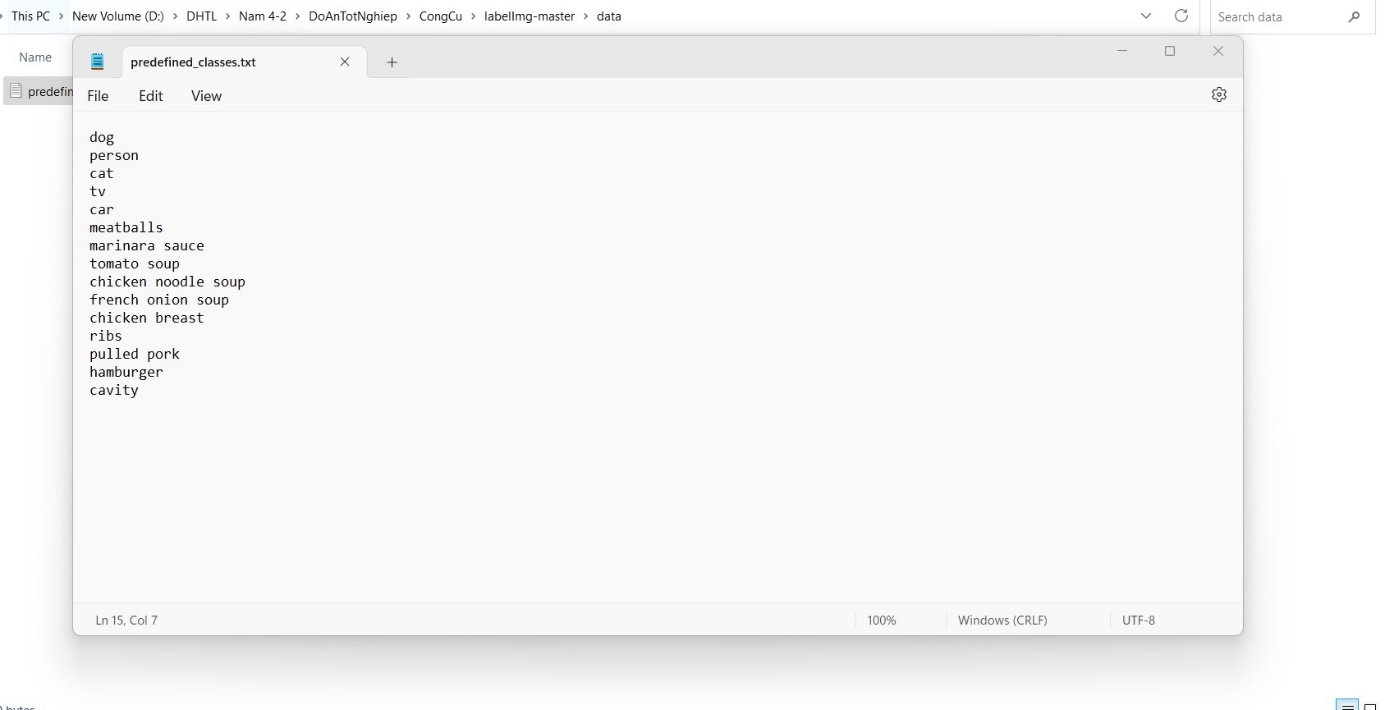
+ 1 file text chứa dữ liệu vùng đã xác định theo từng cụm được tạo ra từ ứng dụng labelImg-master sau khi khoang vùng từ ảnh trên

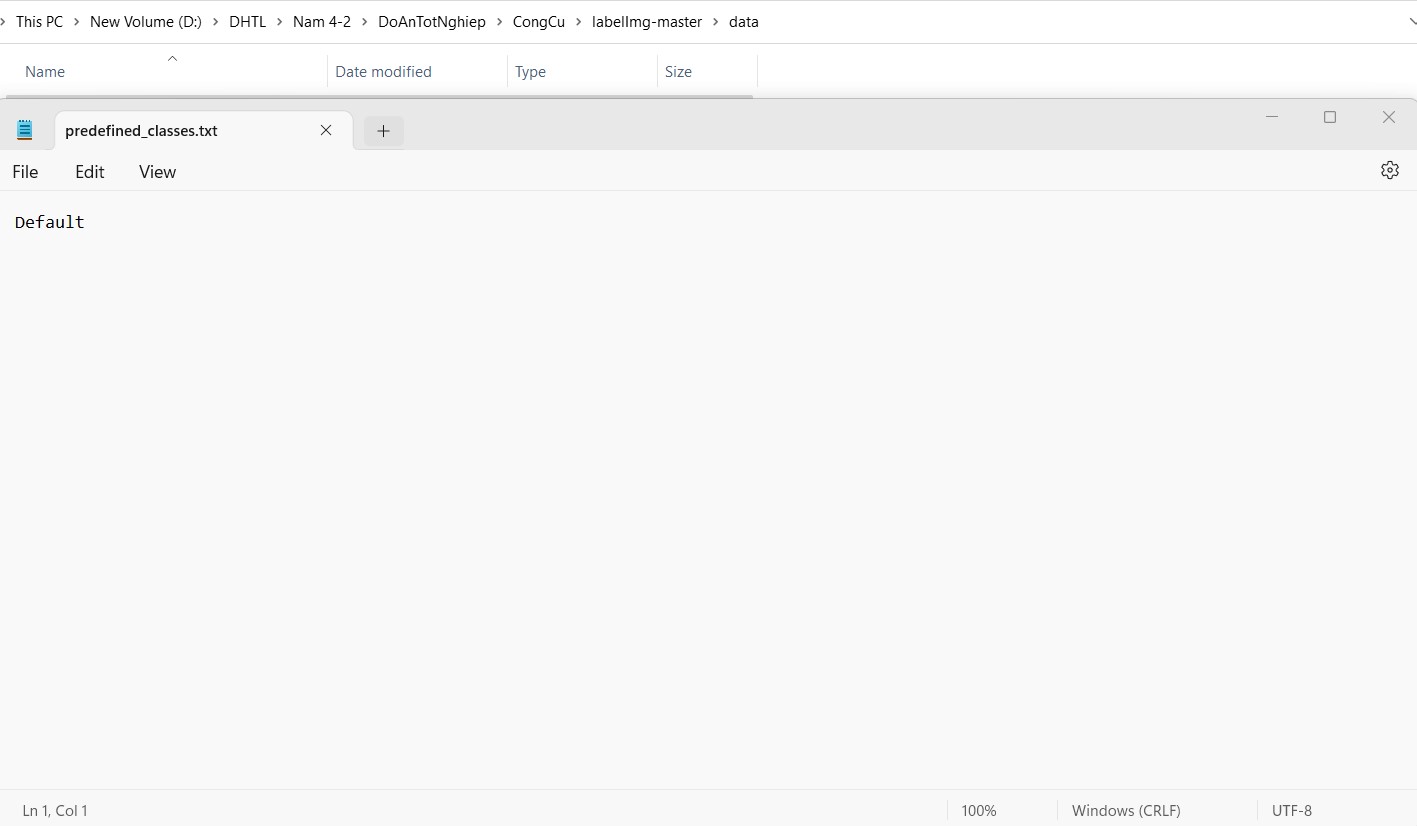
Bước 1 : Vào <https://github.com/tzutalin/labelImg> và tải về giải nén

“ Nhấn Code -> Download zip -> Giải nén ra thư mục tự chọn ”



Bước 2 : Vào thư mục vừa tải tại thư mục “ data ” xóa tất cả có sẵn và thêm chữ Default vào file “ predefined\_classes.txt ”





Bước 3 : Vào file “ labelimg.py ” để bắt đầu gán nhãn theo “ Nhấn Create React Box ( hoặc phím W ) -> Đóng khung khu vực cần gán nhãn -> Ghi tên nhãn -> Ok ”



Bước 4 Lưu ảnh bằng cách nhấn vào Save trên giao diên ( hoặc nhấn tổ hợp phím Ctrl + S )

**3.3.2 Lược đồ về thuật toán** :

- Tư tưởng của thuật toán phân cụm bán giám sát là :

+ với mỗi ảnh vệ tinh về sạt lở bờ biển là mỗi 1 bức ảnh và 1 file text từ ứng dụng labelimg đầu vào sẽ tạo ra được 1 ma trận độ phụ thuộc bổ trợ .

+ Sau đó sẽ sử dụng các phương pháp phân cụm bán giám sát với thông tin bổ trợ từ ma trận độ phụ thuộc đã tạo ở trên. Kết quả của thuật toán sẽ xác định được ma trận độ phụ thuộc U và các tâm cụm V từ các ảnh đầu và file text đã cho

+ Cuối cùng, hiệu năng của thuật toán sẽ được đánh giá dựa trên các độ đo DB, PBM, IFV với công thức tính giá trị như sau :

**3.3.2.1 Độ đo Davies Bouldin ( DB )** :

- Liên quan đến tiêu chuẩn tỉ số phương sai (VRC) trong đó, dựa trên tỉ số giữa các khoảng cách trong nhóm và ngoài nhóm. Đặc biệt là độ đo xác định chất lượng của phân hoạch theo công thức sau :

- Trong đó có :

- Với là các khoảng cách trong nhóm trung bình của các cụm thứ j và thứ m tương ứng, còn là khoảng cách giữa các cụm này. Với công thức như sau :

- Khi các thuật toán được cài đặt thực nghiệm, kết quả độ đo DB nhận được càng nhỏ càng tốt.

**3.3.2.2 Độ đo IFV** :

- Được mô tả bởi phương trình sau :

trong đó :

Giá trị của độ đo IFV trong thực nghiệm tỉ lệ thuận với hiệu năng của thuật toán được cài đặt

**3.3.2.3 Độ đo PDM** :

- cũng dựa trên các khoảng cách trong cụm và khoảng cách giữa các cụm và được tính theo công thức

- Trong đó :

- Độ đo PBM càng cao thì hiệu năng của thuật toán càng cao. Do đó phân hoạch tốt nhất sẽ cho biết khi nào PBM đạt cực đại, điều này có nghĩa là D\_K đạt cực đại trong khi E\_k đạt cực tiểu.

- Sơ đồ

**3.3 Thử nghiệm và đánh giá mô hình thuật toán** :

3.3.3.1 : Kết quả thuật toán :

|  |  |  |
| --- | --- | --- |
| Phương pháp | SSSFC | eSFCM |
| Ảnh 1 | | |
| DB |  |  |
| IFV |  |  |
| PBM |  |  |
| Ảnh 2 | | |
| DB |  |  |
| IFV |  |  |
| PBM |  |  |
| Ảnh 3 | | |
| DB |  |  |
| IFV |  |  |
| PBM |  |  |
| Ảnh 4 | | |
| DB |  |  |
| IFV |  |  |
| PBM |  |  |
| Ảnh 5 | | |
| DB |  |  |
| IFV |  |  |
| PBM |  |  |
| Ảnh 6 | | |
| DB |  |  |
| IFV |  |  |
| PBM |  |  |
| Ảnh 7 | | |
| DB |  |  |
| IFV |  |  |
| PBM |  |  |
| Ảnh 8 | | |
| DB |  |  |
| IFV |  |  |
| PBM |  |  |
| Ảnh 9 | | |
| DB |  |  |
| IFV |  |  |
| PBM |  |  |

**Tài liệu tham khảo** :

- [1] : Bouchachia, A., & Pedrycz, W. (2006). Data clustering with partial supervision.Data Mining and Knowledge Discovery, 12(1), 47-78.

- [2] : Hu, C., Meng, L., & Shi, W. (2008). Fuzzy clustering validity for spatial data.Geo-spatial information science, 11(3), 191-196.

- [3] : Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

- [4] : Vendramin, L., Campello, R. J., & Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining: The ASA Data Science Journal, 3(4), 209-235.

- [5] : Yasunori, E., Yukihiro, H., Makito, Y., & Sadaaki, M. (2009, August). On semi-supervised fuzzy c-means clustering. In *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on* (pp. 1119-1124). IEEE.

- [6] : Yin, X., Shu, T., & Huang, Q. (2012). Semi-supervised fuzzy clustering with metric learning and entropy regularization. *Knowledge-Based Systems*, *35*, 304-311.

- [7] : Zhang, H., & Lu, J. (2009). Semi-supervised fuzzy clustering: A kernel-based approach. Knowledge-Based Systems, 22(6), 477-481.