# 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 phân cụm 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

- Mô hình phân cụm bán giám sát mờ

- Ảnh vệ tinh

### 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 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 hai loại trên đồng thời có thể giảm những nhược điểm của cả hai

## 2.2 Logic mờ :

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

- kết quả lập luận là một 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 Phân cụm :

- Phân cụm là một thuật toán trong lĩnh vực học máy không giám sát được sử dụng để nhóm các đối tượng hoặc mẫu dữ liệu tương tự vào các nhóm khác nhau. Mục tiêu của phân cụm là tạo ra các nhóm có tính chất tương tự bên trong và khác biệt với các nhóm khác.

- Thuật toán phân cụm giúp chúng ta khám phá cấu trúc tự nhiên của dữ liệu mà không cần có thông tin trước về nhãn hay kết quả mong đợi. Nó cho phép chúng ta phân loại dữ liệu dựa trên sự tương đồng hoặc sự khác biệt giữa các mẫu dữ liệu mà không cần biết trước các nhóm.

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

Thuật toán phân cụm mờ được Bezdek [1] 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)



Bảng 1 Bước thực hiện Fuzzy C-means

|  |  |
| --- | --- |
| 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 [6] 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. [4] 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 [4] :

(5)

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 (2) và hàm mục tiêu (5) trên thì tâm cụm được xác định như sau :

với (6)

- Các giá trị của là :

+ Với m > 1 :

(7)

+ Với m = 1 :

(8)

Bảng 2 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 công thức (7) nếu m > 1 hoặc công thức (8) nếu m = 1 |
| 6 : | Tính theo công thức (6) |
| 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 [4] vào năm 2009

- Đến năm 2012, Yin [5] 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 :

(9)

- Với tâm cụm ban đầu :

(10)

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

(11)

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

(12)

- Hàm mục tiêu của eSFCM là :

(13)

- Công thức xác định ma trận độ phụ thuộc :

(14)

- Trong đó và tâm cụm

(15)

Bảng 3 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 (11)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 (14) |
| 6: | Tính bởi công thức (15) |
| 7: | Cho đến khi or t > maxStep |

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

## 3.1 : Bài toán ứng dụng :

### 3.1.1 Giới thiệu bài toán :

- 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

### 3.1.2 Ảnh đầu vào :

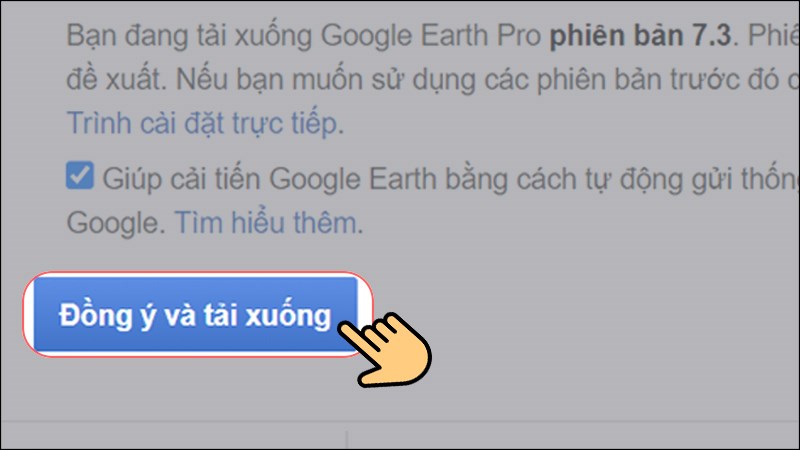
- Ảnh vệ tinh là hình ảnh được chụp từ các thiết bị vệ tinh trong không gian, nhằm quan sát và ghi lại hình ảnh của Trái Đất hoặc các vùng khác trong không gian. Công nghệ này đã thay đổi cách chúng ta nhìn nhận và nghiên cứu về Trái Đất, cung cấp thông tin quan trọng về địa hình, khí hậu, môi trường và các sự kiện tự nhiên.

- Các ảnh vệ tinh được chụp từ máy ảnh hoặc cảm biến gắn trên vệ tinh quay quanh Trái Đất. Có nhiều loại vệ tinh được sử dụng để chụp ảnh vệ tinh, bao gồm các vệ tinh quan sát Trái Đất, vệ tinh thời tiết và vệ tinh khoa học.

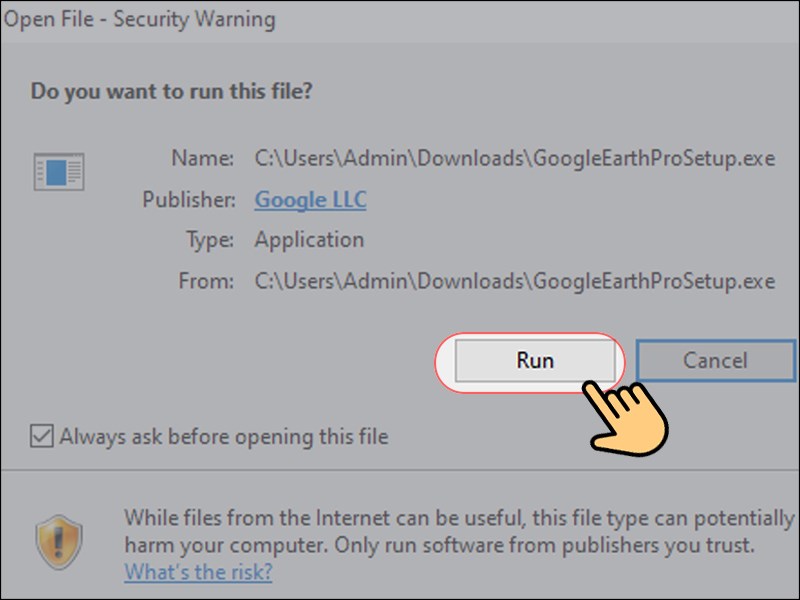
- Ảnh vệ tinh cung cấp thông tin đa dạng về Trái Đất. Chúng có thể cho thấy các đặc điểm địa hình như núi, sông, đồng cỏ và rừng rậm. Ngoài ra, chúng cũng có thể theo dõi sự biến đổi của môi trường, bao gồm quá trình đô thị hóa, thay đổi sử dụng đất, bạo lực tự nhiên như động đất, bão và cháy rừng. Ảnh vệ tinh cũng được sử dụng để nghiên cứu khí hậu, dự báo thời tiết và đánh giá tài nguyên tự nhiên. Vậy nên đối tượng đầu vào của thuật toán là ảnh vệ tinh có sạt lở bờ biển

- Cách lấy được ảnh đầu vào :

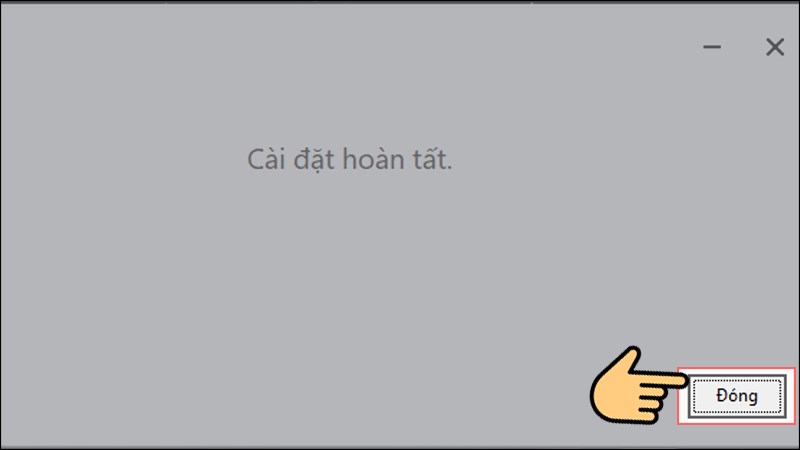
Bước 1 : Cài đặt ứng dụng Google Earth tại <https://www.google.com/intl/vi/earth/versions/> về máy tính



Hình **2** Ảnh Demo tải ứng dụng Google Earth thứ 1



Hình 3 Ảnh Demo tải ứng dụng Google Earth thứ 2

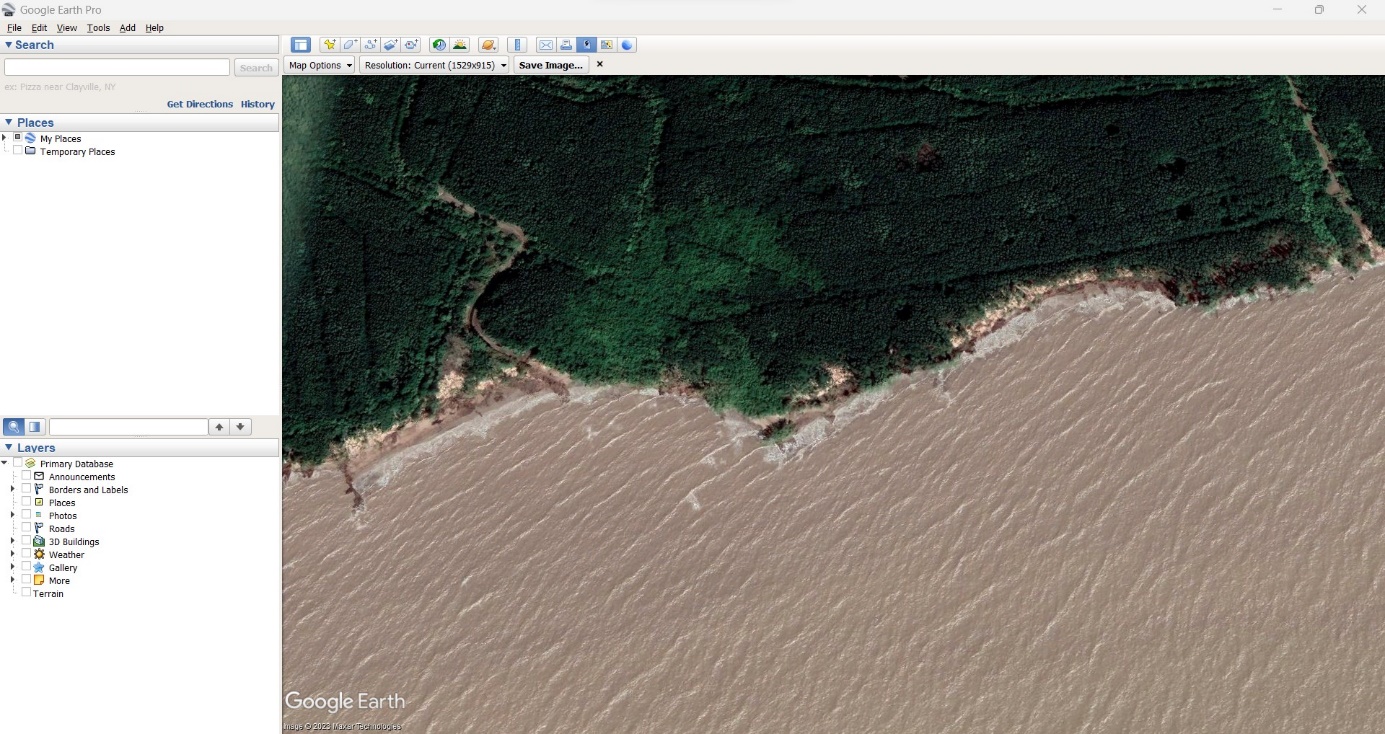


Hình 4 Ảnh Demo tải ứng dụng Google Earth thứ 3

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:



Hình 5 Ảnh Demo tải ứng dụng Google Earth thứ 5

### 3.1.3 Vùng gán nhãn :

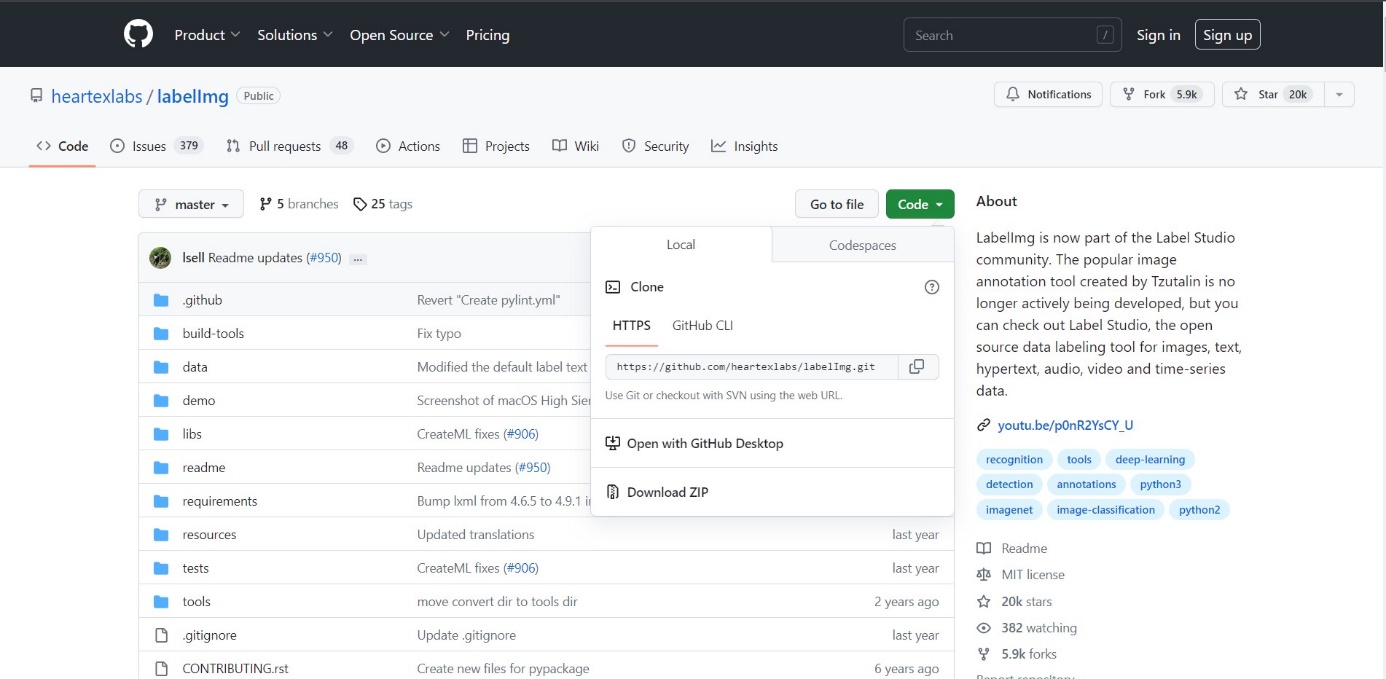
- Vùng gán nhãn (hay còn gọi là vùng đánh dấu) là quá trình xác định và đánh dấu các vùng đặc biệt trên hình ảnh hoặc video để tạo dữ liệu huấn luyện cho các mô hình học máy và trí tuệ nhân tạo. Việc gán nhãn vùng cung cấp thông tin cụ thể về đối tượng, vị trí và các thuộc tính của nó ( ví dụ như vùng sạt lở ), từ đó giúp mô hình nhận biết và phân loại đối tượng trong các tác vụ như nhận dạng khuôn mặt, phát hiện đối tượng, phân đoạn ảnh, và nhiều ứng dụng khác.

- Quá trình gán nhãn vùng thường được thực hiện bằng cách sử dụng các công cụ và phần mềm đồ họa, nơi người gán nhãn chọn và vẽ các khung chứa đối tượng hoặc các vùng quan tâm khác trên hình ảnh hoặc video ( ví dụ như labelImg-master ). Các nhãn có thể được gán bằng các đường viền, hình chữ nhật, hay các đường cong tùy thuộc vào mục đích cụ thể của tác vụ.

- Cách tạo file text về vùng gán nhãn từ ứng dụng labelImg-master

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

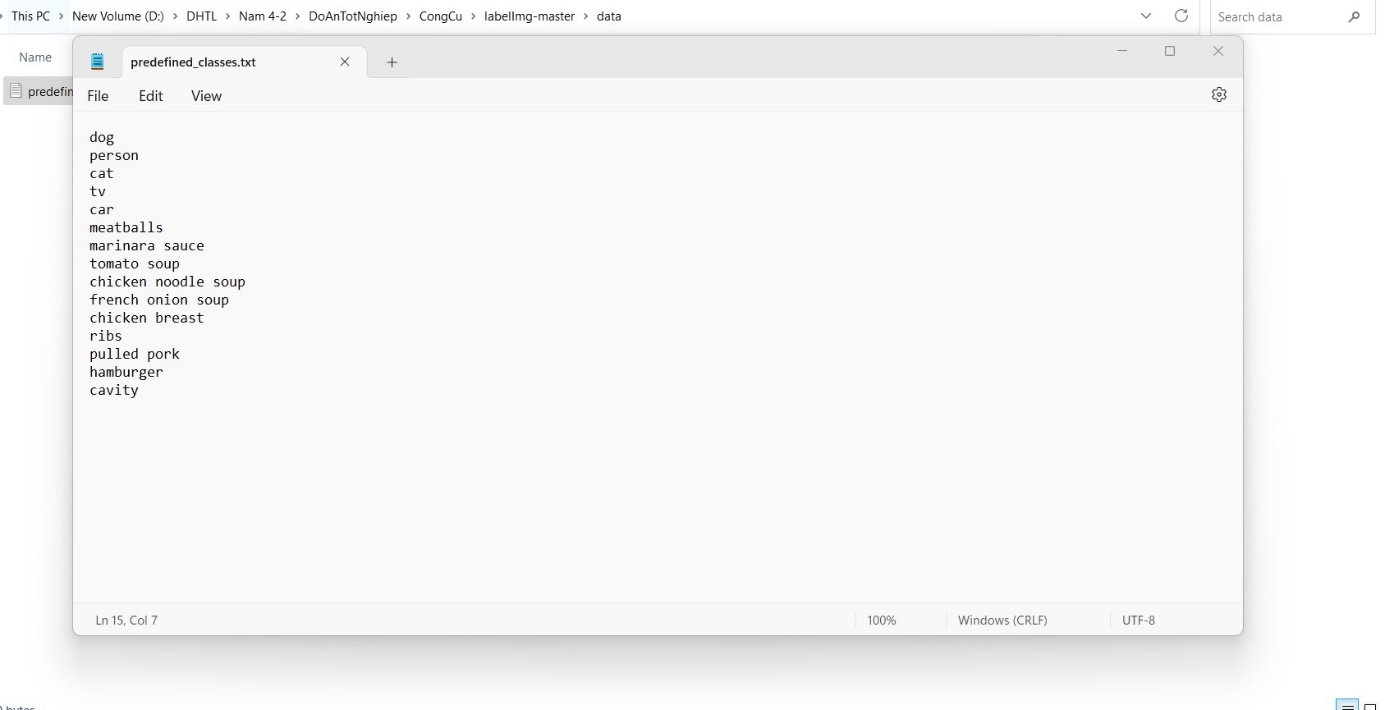
“ Nhấn Code -> Download zip -> Giải nén ra thư mục tự chọn ”



Hình 6 Demo tải ứng dụng LabelImg 1

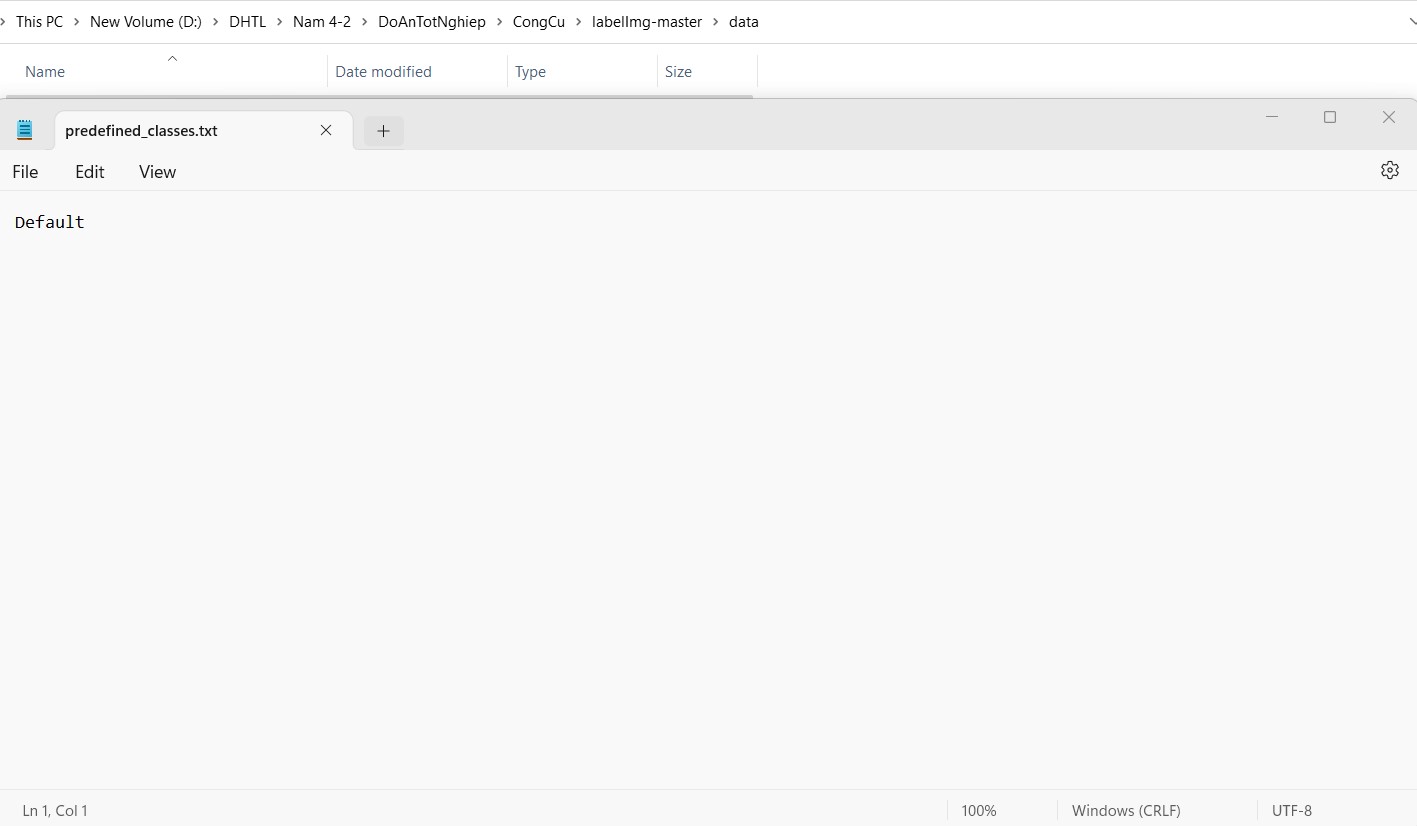
Bước 2 : Vào thư mục vừa tải tại thư mục “ data ” xóa tất cả có sẵn và thêm chữ Default vào file “ predefined\_classes.txt ”

+ Trước khi thay đổi file



Hình 7 Demo tải ứng dụng LabelImg 2

+ Sau khi thay đổi file :



Hình 8 Demo tải ứng dụng LabelImg 3

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 ”



Hình 9 Demo tải ứng dụng LabelImg 4

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

### 3.1.4 Ảnh đầu ra :

- Ảnh đầu ra là ảnh được tạo nên bởi kết quả quá trình phân cụm bán giám sát.

- Với mỗi điểm dữ liệu có giá trị điểm ảnh là kết quả của công thức :

(16)

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

### 3.2.1 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 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 :

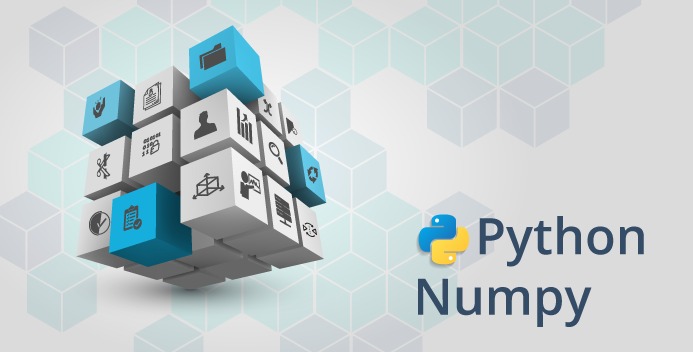
#### 3.2.2.1 Numpy :

- 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.



Hình 10 Hình minh họa của Numpy

- 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 vài 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.



Hình 11 Hình minh hoa pandas

- Cấu trúc dữ liệu của pandas có 3 loại :

+ Series là một mảng một chiều có khả năng chứa các loại dữ liệu khác nhau. Series có thể được xem như một cột trong bảng dữ liệu hoặc một mảng nhãn, với các chỉ mục (index) tương ứng với từng phần tử trong mảng.

+ DataFrame là mảng hai chiều đại diện cho một bảng dữ liệu hai chiều, tương tự như một bảng trong SQL hoặc một spreadsheet trong Excel. DataFrame bao gồm các hàng và cột, và mỗi cột có thể chứa các loại dữ liệu khác nhau.

+ Panel là một đối tượng dữ liệu ba chiều trong thư viện pandas của Python. Nó được sử dụng để lưu trữ và làm việc với dữ liệu có cấu trúc theo ba chiều, tương tự như một bảng dữ liệu ba chiều. Tuy nhiên, từ phiên bản pandas 0.20.0 trở đi, Panel đã được gắn thẻ là "không khuyến nghị sử dụng" và khuyến khích sử dụng các cấu trúc dữ liệu khác như DataFrame hoặc xarray thay thế.

#### 3.2.2.3 Tkinter :

- Thư viện Tkinter là một thư viện đồ họa phổ biến trong Python, được sử dụng để xây dựng giao diện người dùng đồ họa (GUI). Tkinter là giao diện Python cho thư viện Tk, một toolkit đa nền tảng được viết bằng ngôn ngữ C.



Hình 12 Hình minh hoa Tkinter

- Dưới đây là một tổng quan về các tính năng và chức năng chính của thư viện Tkinter:

+ Xây dựng giao diện người dùng: Tkinter cung cấp các phương thức và lớp để tạo ra các thành phần giao diện người dùng như cửa sổ, nút, hộp văn bản, hộp thoại, danh sách, và hình ảnh.

+ Điều khiển sự kiện: Tkinter cho phép xử lý các sự kiện từ người dùng, như bấm nút, nhập liệu từ bàn phím, di chuột, và nhiều sự kiện khác. Bạn có thể gắn các hàm xử lý sự kiện vào các thành phần giao diện người dùng để thực hiện các hành động tương ứng.

+ Định dạng và tùy chỉnh giao diện: Tkinter cho phép bạn tùy chỉnh giao diện bằng cách thay đổi màu sắc, phông chữ, kích thước và kiểu chữ, định dạng văn bản, các hiệu ứng đồ họa và nhiều thuộc tính khác của các thành phần giao diện.

+ Cài đặt layout: Bạn có thể sử dụng Tkinter để xây dựng giao diện người dùng theo các cấu trúc layout như grid, pack và place. Các cấu trúc này giúp bạn tổ chức các thành phần giao diện và điều chỉnh vị trí, kích thước và sự tự động điều chỉnh của chúng.

+ Hỗ trợ đa nền tảng: Tkinter là một phần của thư viện chuẩn của Python, nghĩa là nó được cài đặt trên hầu hết các phiên bản Python và hỗ trợ trên nhiều nền tảng, bao gồm Windows, macOS và Linux.

+ Tích hợp các thành phần khác: Tkinter có khả năng tích hợp với các thư viện và công cụ khác trong hệ sinh thái Python, cho phép bạn sử dụng các tính năng và chức năng mở rộng từ các thư viện khác như NumPy, Matplotlib và Pillow.

- Tkinter là một thư viện mạnh mẽ và dễ sử dụng để phát triển các ứng dụng GUI đơn giản trong Python. Nó cung cấp một cách nhanh chóng và tiện lợi để tạo ra các giao diện người dùng tương tác và tùy chỉnh theo ý muốn.

#### 3.2.2.4 Pil :



Hình 13 Hình minh hoa Pillow

- Thư viện PIL (Python Imaging Library) là một thư viện mã nguồn mở được sử dụng rộng rãi trong Python để xử lý và chỉnh sửa hình ảnh. Thư viện này cung cấp các công cụ mạnh mẽ để thao tác với các tệp hình ảnh, thực hiện các phép biến đổi, chỉnh sửa, tạo ra các hiệu ứng đồ họa và thực hiện nhiều tác vụ xử lý hình ảnh khác.

- Dưới đây là một tổng quan về các tính năng và chức năng chính của thư viện PIL:

+ Định dạng hỗ trợ: PIL hỗ trợ nhiều định dạng hình ảnh phổ biến như JPEG, PNG, GIF, TIFF và BMP. Nó cung cấp các công cụ để đọc và ghi dữ liệu hình ảnh từ và đến các định dạng này.

+ Xử lý hình ảnh cơ bản: PIL cung cấp các phương pháp để điều chỉnh kích thước, cắt, xoay, lật, chuyển đổi màu sắc, chỉnh sửa độ tương phản và sáng tối của hình ảnh. Bạn có thể thực hiện các thao tác này để cắt, chỉnh sửa và cải thiện hình ảnh theo ý muốn.

+ Các phép biến đổi hình ảnh: PIL cho phép bạn thực hiện các phép biến đổi hình ảnh phức tạp như xoay, co giãn, kéo dãn và biến đổi các hình ảnh theo các phép biến đổi hình học khác.

+ Các hiệu ứng và bộ lọc: PIL cung cấp các bộ lọc hình ảnh như mờ, nhòe, lọc nhiễu và làm sắc nét hình ảnh. Bạn có thể áp dụng các hiệu ứng này để tạo ra các hiệu ứng đồ họa đặc biệt trên hình ảnh.

+ Ghép ảnh và tạo đồ họa: PIL cho phép bạn ghép nhiều hình ảnh lại với nhau, thêm văn bản và hình vẽ lên hình ảnh, và tạo ra các biểu đồ và đồ họa động đơn giản.

+ Xử lý dữ liệu hình ảnh: PIL cung cấp các công cụ để truy cập và xử lý dữ liệu pixel trong hình ảnh. Bạn có thể truy cập và chỉnh sửa giá trị pixel, thực hiện các toán tử logic và tính toán trên dữ liệu hình ảnh.

- PIL là một thư viện rất phổ biến và có độ ổn định cao trong việc xử lý hình ảnh trong Python. Nó dễ sử dụng và linh hoạt, cho phép bạn thực hiện nhiều tác vụ xử lý hình ảnh phức tạp và tạo ra các hiệu ứng đồ họa đa dạng.

### 3.2.3 Các hàm được sử dụng trong đồ án :

- abs() : Hàm tính giá trị tuyệt đối

- float() : Convert dữ liệu sang kiểu float

- int() : Convert dữ liệu sang kiểu int

- len() : Trả về chiều dài của dữ liệu

- list() : Khởi tạo hoặc convert dữ liệu về kiểu list

- max() : Trả về giá trị lớn nhất

- min() : Trả về giá trị nhỏ nhất

- pow() : Hàm tính số mũ

- range() :

- round() : Trả về giá trị đã làm tròn

- sum() : Tính tổng dữ liệu

- str() : Convert dữ liệu sang kiểu string

## 3.3 Lược đồ tổng quát :

### 3.3.1 Thu thập dữ liệu :

- Dữ liệu đầu vào gồm :

+ 1 tập ảnh vệ tinh có vùng sạt lở bờ biển :

Bảng 4 Thuộc tính của ảnh đầu vào

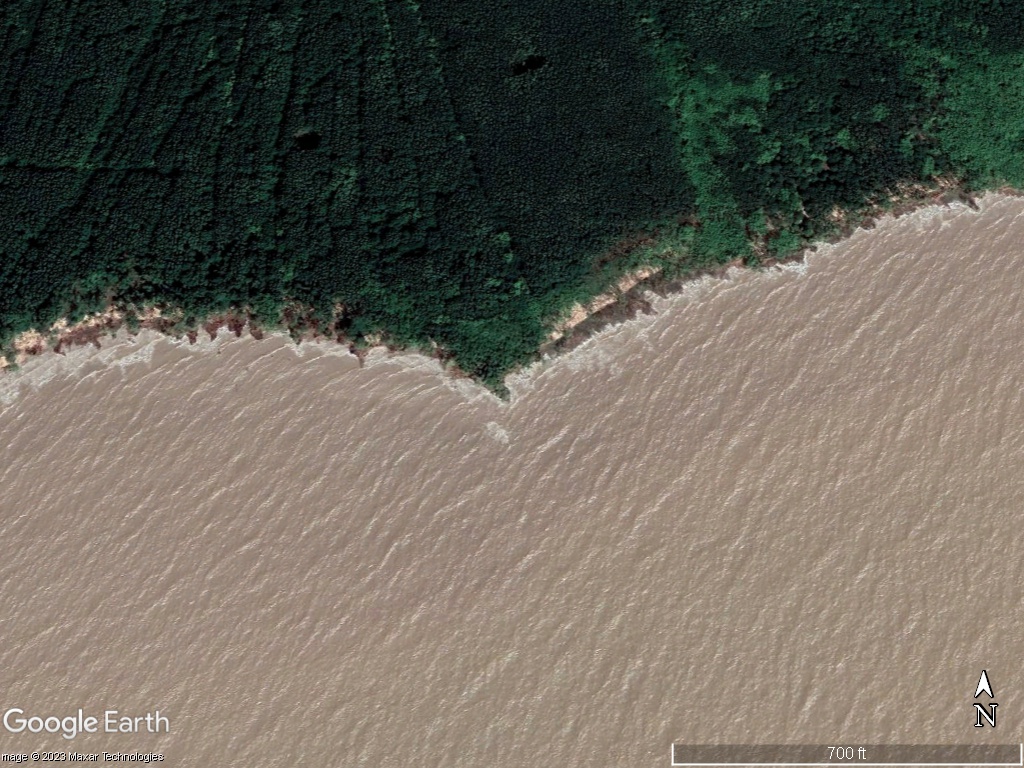
|  |  |  |  |
| --- | --- | --- | --- |
| Ảnh | Kích thước ( width x height ) | Độ phân giải ( dpi ) | Dung lượng ( KB ) |
| [img\_1](UngDung/Data/Image/1.jpg) | 1024 x 768 | 96 | 348 |
| [img\_2](UngDung/Data/Image/2.jpg) | 1024 x 768 | 96 | 364 |
| [img\_3](UngDung/Data/Image/3.jpg) | 1024 x 768 | 96 | 352 |
| [img\_4](UngDung/Data/Image/4.jpg) | 1024 x 768 | 96 | 355 |
| [img\_5](UngDung/Data/Image/5.jpg) | 1024 x 768 | 96 | 345 |



Hình 14 img\_1



Hình 15 img\_2



Hình 16 img\_3



Hình 17 img\_4



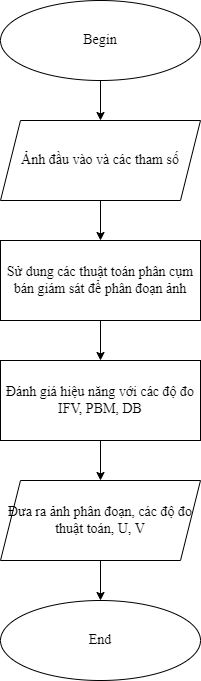
Hình 18 img\_5

+ từ đó có thể tạo ra được file text chứa thông tin để xác định được vùng sạt lở từ ảnh trên như sau :

Bảng 5 Thuộc tính của file gán nhãn của ảnh tương ứng

|  |  |  |
| --- | --- | --- |
| File | Ảnh | Dung lượng ( KB ) |
| [txt\_1](UngDung/Data/Image/1.txt) | img\_1 | 1 |
| [txt\_2](UngDung/Data/Image/2.txt) | img\_2 | 1 |
| [txt\_3](UngDung/Data/Image/3.txt) | img\_3 | 1 |
| [txt\_4](UngDung/Data/Image/4.txt) | img\_4 | 1 |
| [txt\_5](UngDung/Data/Image/5.txt) | img\_5 | 1 |

### 3.3.2 Lược đồ về thuật toán :



Hình 19 Biểu đồ sơ lược thuật toán

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

+ Đầu tiên, chúng ta sẽ sử dụng xử lý dữ liệu đầu vào đã cho là ảnh vệ tinh về vùng sạt lở vùng biển và file text về cụm người dùng đã xác định trước từ labelimg. Thông tin bổ trợ từ ma trận độ phụ thuộc sẽ được sử dụng để cung cấp thông tin bổ sung cho quá trình phân cụm.

+ Sau đó, ảnh vệ tinh trên và ma trận độ phụ thuộc đã tạo ở trên sẽ trở thành tham số đầu vào của thuật toán phân cụm bán giám sát để xử lý dữ liệu. Thông tin bổ trợ từ ma trận độ phụ thuộc sẽ được sử dụng để cung cấp thông tin bổ sung cho quá trình phân cụm

+ Kết quả của thuật toán phân cụm sẽ là ma trận độ phụ thuộc U và các tâm cụm V từ các ảnh đầu và file text đã được cung cấp. Ma trận độ phụ thuộc U thể hiện mức độ tương đồng giữa các mẫu trong quá trình phân cụm. Các tâm cụm V đại diện cho các điểm trung tâm của từng nhóm phân cụm.

+ 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 ) :

- DB [3] 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 :

(16)

- Trong đó có :

(17)

(18)

- 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 :

(19)

(20)

- 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 :

- IFV [2] được mô tả bởi phương trình sau :

(21)

trong đó :

(22)

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 :

- PDM [3] cũng dựa trên các khoảng cách trong cụm và khoảng cách giữa các cụm và được tính theo công thức

(23)

- Trong đó :

(24)

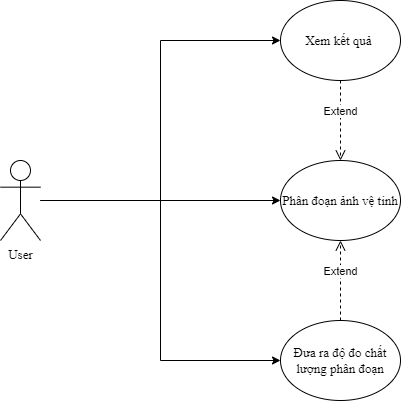
(25)

(26)

- 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.

### 3.3.3 Thiết kế hệ thống :

- Hệ thống cho người dùng phân đoạn ảnh vệ tinh kết hợp với vùng gán nhãn từ đó người dùng có thể xem chi tiết kết quả cũng như các độ đo chất lượng thuật toán



Hình 20 Biểu đồ usecase hệ thống

#### 3.3.3.1 Chức năng phân đoạn ảnh :

- Tác nhân : Người dùng

- Input : Ảnh vệ tinh cần phân đoạn

- Output : Ảnh đã phân đoạn

- Mô tả chi tiết :

+ Người dùng chọn ảnh cần phân đoạn

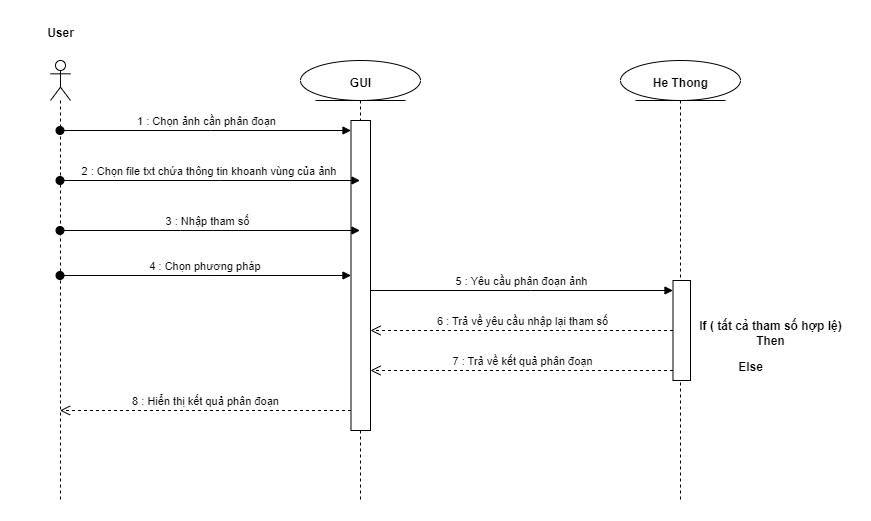
+ Người dùng chọn thêm file txt chứa thông thi vùng gán nhãn

+ Người dùng nhập các tham số

+ Người dùng chọn phương pháp phân đoạn ảnh

+ Hệ thống kiểm tra tham số, yêu cầu nhập lại nếu không thỏa mãn , thỏa mãn thì thực hiện phân đoạn và trả lại kết quả

- Biểu đồ trình tự :



Hình 21 Biểu đồ trình tự chức năng phân đoạn ảnh

#### 3.3.3.2 Chức năng xem kết quả :

- Tác nhân : Người dùng

- Input : Sau khi ảnh đã phân đoạn và người dùng chọn xem kết quả phân cụm

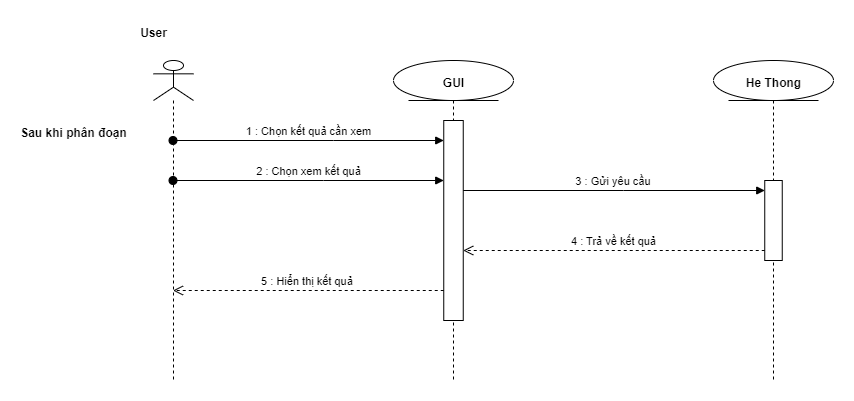
- Output : Hiển thị kết quả phân cụm

- Mô tả chi tiết :

+ Người dùng chọn chức năng xem kết quả

+ Hệ thống sẽ xử lý và hiển thị kết quả phân cụm

- Biểu đồ trình tự :



Hình 22 Biểu đồ trình tự chức năng xem kết quả

#### 3.3.3.3 Chức năng xem độ đo chất lượng thuật toán :

- Tác nhân : Người dùng

- Input : Sau khi ảnh đã phân đoạn và người dùng chọn xem độ đo chất lượng thuật toán

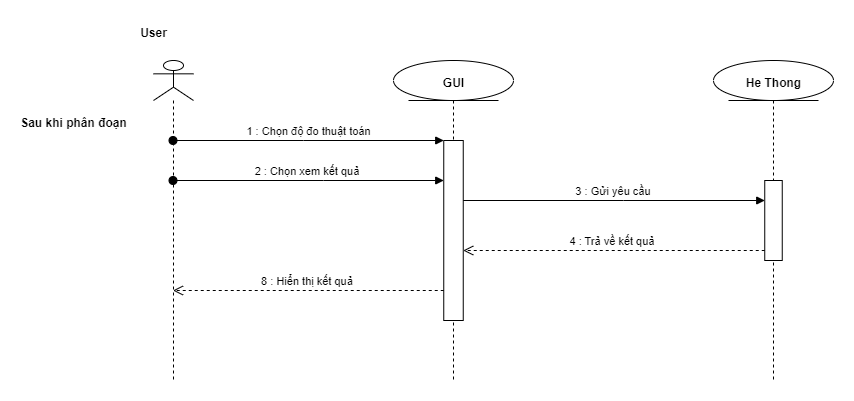
- Output : Kết quả độ đo chất lượng thuật toán được hiển thị

- Mô tả chi tiết :

+ Người dùng chọn chức năng xem độ đo chất lượng thuật toán

+ Hệ thống sẽ xử lý và hiển thị ra màn hình độ đo chất lượng thuật toán

- Biểu đồ trình tự :



Hình 23 Biểu đồ trình tự chức năng xem độ đo chất lượng thuật toán

## 3.4 Thử nghiệm và đánh giá mô hình thuật toán :

### 3.4.1 : Kết quả thuật toán :

Bảng 6 Kết quả chạy thuật toán

|  |  |  |
| --- | --- | --- |
| Phương pháp | SSSFC ( m = 2 ) | eSFCM ( lamda = 0.01) |
| img\_1 : | | |
| DB | 0.00846 | 0.214846 |
| IFV | 19.276578 | 91173.560344 |
| PBM | 110569744422618.69 | 181078438923907.2 |
| img\_2 : | | |
| DB | 0.007867 | 0.009524 |
| IFV | 21.180398 | 26879.980966 |
| PBM | 239993268239350000 | 226710096633.68073 |
| img\_3 : | | |
| DB | 0.000049 | 0.009098 |
| IFV | 19.921778 | 19103.111717 |
| PBM | 312576853297989.94 | 95612450578569.05 |
| img\_4 : | | |
| DB | 0.005589 | 0.208008 |
| IFV | 20.735304 | 96275.612114 |
| PBM | 1056928855775179600 | 185429230412122.66 |
| img\_5 : | | |
| DB | 0.003302 | 0.009601 |
| IFV | 21.104778 | 11948.317382 |
| PBM | 157604899429145540 | 263749331199.0855 |

#### 3.3.3.2 Đánh giá thuật toán :

- Sau khi chạy với bộ dữ liệu ảnh, từ giá trị đo được từ các độ đo trên có thể dễ dàng thấy hiệu năng của từng thuật toán được đánh giá một cách cụ thể và có sự khác biệt cụ thể giữa các độ đo :

Bảng 7 Kết quả trung bình

|  |  |  |
| --- | --- | --- |
|  | SSSFC | eSFCM |
| DB | 0.0050534 | 0.0902154 |
| IFV | 20.4437672 | 49076.1165046 |
| PBM | 355534724032856825.926 | 150315909549454.32 |



Hình 24 Ảnh img\_1 kết hợp với vùng gán nhãn txt\_1



Hình 25 Ảnh kết quả phân cụm SSSFC



Hình 26 Ảnh kết quả phân cụm eSFCM

# CHƯƠNG 4: KẾT LUẬN :

- Đồ án đã nghiên cứu và triển khai một số phương pháp phân cụm bán giám sát với mục tiêu là xây dựng một hệ thống có khả năng tính toán và xác định được vùng đất có nguy cơ bị sạt lở bờ biển từ điểm ảnh của ảnh vệ tinh.

- Đồ án đã thực hiện được những nội dung như sau :

+ Tìm hiểu về những kiến thức tổng quan về ảnh vệ tinh, phân đoạn ảnh từ đó đặt ra được “ Sử dụng phân cụm bán giám sát phát hiện vùng sạt lở từ ảnh vệ tinh ”

+ Tìm hiểu các phương pháp phân đoạn ảnh vệ tinh và đưa ra thuật toán, đánh giá và các ví dụ về từng thuật toán. Từ đó, có thể hiểu hơn về các thuật toán phân cụm ảnh vệ tinh hiện có

+ Cài đặt thuật toán SSSFC, eSFCM để phân đoạn ảnh vệ tinh về vùng sạt lở bờ biển và các độ đo DB, PBM, IFV để đánh giá chất lượng thuật toán. Từ đó có thể làm rõ hơn về tác dụng của phân cụm mờ trong việc phân đoạn ảnh vệ tinh về vùng sạt lở bờ biển

- Dựa vào những nội dung, kết quả đạt được thì tương lai đề tài có thể sẽ có hướng phát triển như sau :

+ Tiết tục nâng cấp, xây dựng thêm phương pháp để tăng hiểu quả phân đoạn cũng như thời gian xử lý

+ Tìm ra được các để hiện rõ hơn về vùng sạt lở bờ biển trong ảnh

**Tài liệu tham khảo** :

- [1] Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers.

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

- [3] 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.

- [4] 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.

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

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